Final Report for the ESF GENOMIC-RESOURCES - Exchange Grant 4216

Title of the project: "Impact of sample size, marker density and population structure on the estimation of effective population size from the SNP chip information: a simulation approach".

Applicant: Dr. Stamatina Trivizaki, from the Institute of Animal Genetic Improvement of Nea Mesimvria, Thessaloniki (Greece).

Host group leader: Dr. Ino Curik, at the University of Zagreb (Croatia)

Visit duration: 01/04/2013-29/06/2013 (13 weeks)

1. Purpose of the visit

The conservation of genetic diversity is a priority factor in rare small populations (Meuwissen, 2001). The effective size (Ne) is one of the most important issues in population genetics, given its usefulness as a measure of the long-term performance of the population regarding both diversity and inbreeding and, therefore, to characterize the risk status of livestock breeds (Duchev et al., 2006; FAO, 1998). Recently, the average of effective size across populations has been proposed to assess the genetic diversity in livestock (Villanueva et al., 2010). Ne helps to explain the observed extent and pattern of genetic variation in a population from a retrospective point of view and to predict the loss of genetic variation and the survival of small breeding populations from a prospective point of view (Wang 2005). Ne is also necessary in the application of genomic selection of populations, since the accuracy of the breeding values depends on the linkage disequilibrium between the QTL and SNP, and a higher Ne means higher number of markers is needed (Meuwissen et al., 2001). However, estimates of Ne vary with the methodology used to asses it, thus limiting its ability to characterize the risk status of a population or to predict the precision of the genomic selection. Therefore, assessing the reliability of the estimates of Ne is a very important challenge.

The basic methods used to estimate Ne can be divided into demographic, pedigree-based or marker-based approaches. Though, demographic methods provide crucial Ne estimates in some situations, they are based on simplified population models and use limited population data. In pedigree-based methods, inbreeding rate is estimated from pedigree records, which in turn is used to estimate Ne. However, these methods require reliable and complete pedigree data over several generations, which is often lacking even for breeds in developed countries (Konig and Simianer, 2006). Since 2007, assays to generate dense genome-wide single nucleotide polymorphism (SNP) data became available for cattle. These developments offered new possibilities regarding the application of marker-based methods for estimating Ne with linkage disequilibrium (LD) information (Flury et al., 2010, Qanbari et al., 2010).

Sved (1971) and Hill (1981) proposed that LD would come exclusively from genetic drift for neutral unlinked loci in an isolated population with random mating. This phenomenon could be used to estimate Ne by exploiting that the variance of LD or the correlation of gene frequencies r between loci is a known function of the population size (Hill, 1981). Based on simulations and real data, the same relation between Ne, segment length c and expected LD was confirmed for a multi locus measure of LD called chromosome segment homozygosity (CSH) (Hayes et al., 2003). Generally, the Ne is estimated $(2c)^{-1}$generation ago where c is either distance (in Morgans) between two markers for which LD is estimated (Hill, 1981) or the chromosomal segment length for CSH (Hayes et al. 2003).

With many breeding programs now incorporating genomic information at great expense, simulation provides a useful tool for providing information about the potential that different analysis methods have to increase the accuracy of estimating breeding values and to compare the alternative structures of breeding programs, at low cost while overcoming difficulties that origin from data accessibility and availability. The application of

simulation for LD estimation of Ne and the effect of a series of population characteristics and inputs to the estimation accuracy is a preliminary approach that would provide information about the best methodology to be further applied in real population data.

The aim of this short visit is multi-dimensional as it targets not only in acquiring useful results according to the proposal but also to create a solid research background that could be further developed through future collaboration and invest in the expansion of the scientific horizon of both the applicant and the host group. The first step to this direction is to investigate the possibilities and capacities of bioinformatics tools, mathematic equations and genomic background information in estimating Ne from LD. The genomic dimension of LD approach of Ne is quite innovative and still unexplored in many aspects. Several steps have been made to the direction of estimating Ne in historical populations, while several formulas have been proposed for this reason from Sved (1971), Weir and Hill (1994), Waples (2006) and Ober et al. (2013)(Table 2). In this short visit research plan the main focus was different than the approach attempted by previous scientific groups on the estimation of Ne. In our case study, the interest is focusing on recent historical population, examining Ne estimation in the previous 100 generations (recent past).To this extend the effect of 1) sample size and (2) marker density on LD estimation of Ne were examined.

2. Description of the work carried out during the visit;

Simulated Data

The simulation method, implemented in a software package called AlphaDrop (Hickey and Gorjanc, 2012). This package is used to simulate genomic data and phenotypes with flexibility in terms of the historical population structure, recent pedigree structure, distribution of quantitative trait loci effects, and with sequence and single nucleotide polymorphism-phased alleles and genotypes. The system is calling a combination of coalescent and gene drop methods to simulate sequence, SNP, and QTL and it is packaged in a Fortran 95 program called AlphaDrop, which calls the Markovian Coalescence Simulator (MaCS) (Chen et al. 2009). In order to perform simulation according to the project requirements, collaboration with Assistant Professor Gregor Gorjanc (University of Ljubljana) was established. Dr. Gorjanc demonstrated the capacities of the program in terms of the parameters included and modifications availability due to the needs of the project.

Three different bovine populations simulated having the same characteristics but scanned for BeadChips of different size. Each individual in the population is represented by 29 chromosomes. AlphaDrop is setting up data structures in terms of SNP chips and pedigree. It then calls MaCS, which simulates a sample of haplotypes with sequence information for each chromosome according to the specified ancestral population, mutation and recombination rates that in this case were both 10^{-8} (approximately one mutation and one recombination event per Morgan). Pedigree was internally created in AlphaDrop with each generation consisting of 50 sires, one dam per sire and two offsprings per dam, holding the population size per generation constant to 100 individuals. The population simulated for effective population size of 200.

The BeadChips used are: (a) BovineSNP50 v2 DNA Analysis BeadChip that contains 54.609 highly informative SNPs uniformly distributed across the entire genome of major cattle breed types, (b) BovineHDBeadChip with more than 777.000 SNPs that uniformly span the entire bovine genome and (c) IlluminaBovineLD Genotyping BeadChip with 6.909 SNP across bovine genome. Considering the chromosomal length at 1 Morgan the distribution of SNP by each BeadChip is presented in the following table (1).

Table 1. SNP distribution in total bovine genome by different BeadChips provided.

BeadChip	Number of SNP in the entire bovine genome	Number of SNP per chromosome	
		approximately	used in analysis
IlluminaBovineLD Genotyping BeadChip	6.909	238	200
BovineSNP50 v2 DNA Analysis BeadChip	54.609	1.883	2.000
BovineHDBeadChip	777.000	26.793	20.000

The full sequence and phased data were additionally programmed to be outputted.
Haplotype data were simulated in AlphaDrop for 100 generations ago.
The size of the output files is 2,1G for BovineLD, 24G for BovineSNP50 and 78G for BovineHD. In order to shrink output file size and adjust to available computational capacities, data was retracted using Linux commands for the most remote generation 90 to 100 generations ago to be further analysed.

Pedigree estimation of Ne

The software package, GRain v2.1 (Baumung et al.,) have been demonstrated and used in the assistance of Professor Curik. This software intended to enable and promote testing of various hypotheses with respect to purging and heterogeneity of inbreeding depression.
The program withdraw information from the pedigree output file created internally in AlphaDrop and estimate the inbreeding coefficients that were applied to the estimation of inbreeding Ne per generation according to (1).

(1)

Where ΔF = the increment in inbreeding per generation (the rate of inbreeding)

Analysis of LD data and estimation of Ne

An R script has been accommodated in collaboration to Dr. Gorjanc for the estimation of both LD and Ne among different BeadChip sizes either from haplotypes.
In order to approach this estimation all correlations between SNPs involved were estimated. According to Ne theory, physical SNP positioning in the genome could be translated to genomic distance according the formula (2).

(2)

Where Gen= generations ago, c= genetic distance in Morgan (Hill, 1981).
The position of SNP in the genome is provided by AlphaDrop output positioning file. As the main interest focuses on the recent past, we targeted on physical distances between 500 and 40.000 Kbp which corresponds to 1,25 generation up to 100 generation ago. The physical distance has been cut into bins of size 500 Kbp. All the LD values estimated from simulated data were assigned to the corresponding bin and the average, median and CV of

LD within each bin was estimated. In order to approach Ne three different mathematic formulas were tested (Table 2). Physical distances are transcribed into genetic distances and into their relevant count of generations.

Table 2. Mathematic equations for the LD estimate of Ne.

Literature reference	Equation
1. Sved (1971)	(3)
2. Weir and Hill (1994)	(4)
3. Ober et al. (2013)	(5)
$E(r^2)$= expectation correlation between SNPs, Ne= effective population size, c= genetic distance (in Morgans), S= census population sample size	

While the first three formulas are approaching Ne using haplotype data, the last derivation is approaching Ne from genotypic data. The precursor formula presented by Sved although it had been revised and updated it is still widely used in Ne research.

Specific routines and loops were added to the R script in order to estimate Ne in different simulated single generation or in sequential series of generations of AlphaDrop. Moreover, R script was modified to estimate Ne retrieving information from the total available population or a random sample of the population per generation. Additionally a routine for excluding Minor Allele Frequencies (MAF) SNPs modified R script according to literature suggestions. SNPs with MAF smaller than 0,01 were excluded from the calculations.

After setting different options, three simulated sets of haplotype data created by scanning with different density markers and analysed by taking into consideration the exclusion of MAF, total population (100 individuals) or subsets of 50 and 25 individuals and estimates in generations 90 to 100 of AlphaDrop. Ne was approached by all three formulas referred above.

Bioinformatic tools and additional knowledge acquired

During the short visit in the University of Zagreb the applicant had the chance to acquire basic knowledge in the use of Linux operating system and commands. The use of Linux is approached in PC terminal but also in University server. Additional computational skill has been practiced by the introduction to R statistical computing and furthermore the use of more sophisticated routines that required for data processing. Dr. Trivizaki explored the capacities of MaCS and AlphaDrop simulation programs and practiced estimations using VanRad and GRain programs for inbreeding coefficients.

Moreover, the applicant had the opportunity to attend a short course in artificial neutral network theory lectured by Dr. Hayrettin Okut, an introduction course to Bayesian statistics by Dr. Gregor Gorjanc and a short term course in Conservation Genetics by Dr. Ino Curik.

3. Description of the main results obtained;

Simulated Data analysis

R script application and estimation of Ne in simulated data created by HD BeadChips couldn’t be accessed because of lack in computer capacity to allocate 20000 SNP matrix in data reading.

Pedigree estimation of inbreeding Ne

For the simulated data population inbreeding Ne approximately 200 was estimated by the application of GRain.

LD estimation of Ne

Considering Ober formula for estimating Ne from LD, that is the most recent proposed for estimation in remote past generations, it is clear that Ne is overestimated in the recent 200 past generations. According to this formula Ne is "indicating" values of 300 with increasing variation around the mean value for each bin as the sample size taken into account is getting smaller. Though when applying the bigger chip (2000 SNP/chr) instead of 200 SNP/chr, Ne values are shrinking around the mean in each bin, demonstrating less dispersion.

In the 100 more recent past generations Sved and Weir-Hill formulas show better fitting to the expected value of Ne 200. The use of total population information gives higher precision to the Ne values, while use of smaller divisions of the population underestimate Ne values for Sved and slight overestimate for Weir-Hill. These results are more visible using denser molecular marker that contributes to the elimination of Ne values around the mean. Actually, Weir-Hill formula seems to have the most accurate approach of Ne value estimated from inbreeding (pedigree) for this time frame. (Table 3)

All formulas demonstrate important amount of outliers in recent generations in the past when genome is scanned with 200 SNP/chr. creating a lot of noise and make it difficult to identify any existing trend of Ne.

Another point of interest that is evident in processing these data sets is the influence of bin size selected. It is obvious that while addressing to more recent generations in the past the bin values of reference are more in number and are gradually decreasing when moving to most remote generations. This suggests that more recent generation LD values and the relevant Ne estimated from them, will be more detailed analysed. For example, 30 bins equal 30 intervals in genomic distances that refers to Ne values when approaching the first two generations in the past while only few bins are capturing Ne information for generations 28 to 200.

4. Future collaboration with host institution (if applicable);

The visit to University of Zagreb, Department of Animal Science had been a great opportunity to establish collaboration with Professor Curik and the members of his scientific team. The present project is very promising to its contribution to genetic conservation perspectives though it is estimated to be at its primitive stages. The three month collaboration is considered very fruitful as it established the foundation to future collaboration while both sides had agreed in expanding their research in a common direction that would contribute to scientific achievements. Moreover, a working group in recent past estimation of effective population size from molecular data is being formed with the contribution of Assistant Professor of the University of Ljubljana, Dr. Gorjanc.

5. Projected publications / articles resulting or to result from the grant (ESF must be acknowledged in publications resulting from the grantee's work in relation with the grant);

Based on the feedback of the current study and future research results an article on the topic will be submitted to a high quality peer reviewed scientific journal. Acknowledgement on the funding institution will be expressed in the coming manuscripts and adhere to other principles as required by the ESF GENOMIC-RESOURCES programs.

6. Other comments (if any).

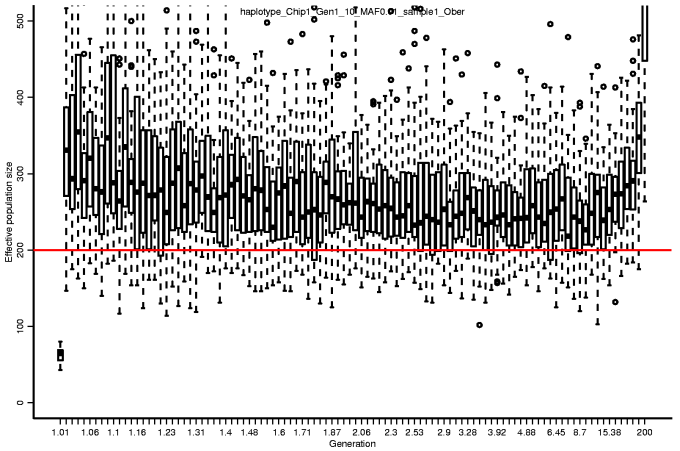
The aim of this study as it was primarily proposed for ESF grant was an opportunity to assess the effect of several factors (marker density, sample size and population structure) on genomic data (LD) estimation of Ne. While starting accommodating this features the applicant, Dr. Trivizaki and host Professor, Dr. Curik realized the importance of exploring the area of estimating recent past Ne. Observations in this field as it is presented in section 3 revealed issues in question that hadn't been taken into consideration initially like evaluating behaviour of mathematic formulas in recent past Ne and computational restrictions. Taking into account possibilities of further exploring the area of Ne estimated by SNP data information, the scientific team is very much concentrated in delivering high quality results, especially since the current attempt is funded by an organization as reputed as ESF. The scientific team is already working on the direction proposed to ESF taking into consideration the key points attributed by this short visit.

On behalf of the applicant, I would like to express my gratitude to ESF for supporting the current study. This opportunity contributed in my involvement to the field of conservation genetics but also to the establishment of scientific network. I would like to thank Dr. Gorjanc for his assistance and teaching guidance in several bioinformatics tools. Furthermore I would also like to thank the scientific team working in the Department of Animal Science at University of Zagreb for their assistance in minor and major importance issues not only in the science but also to everyday life. Last but not least, I would like to thank Professor Curik for giving me the opportunity to participate in his team and get involved in such a crucial topic as Ne. His guidance and support cannot be appreciated more. I hope that our future collaboration will bring me in a position to compensate for the knowledge, the assistance, the opportunities and the personal contribution.

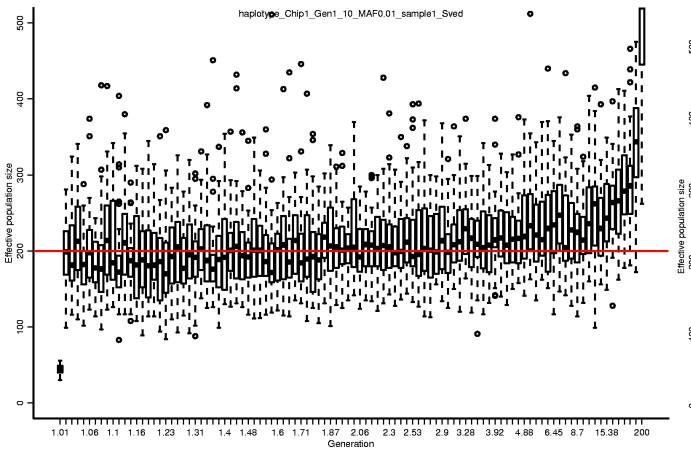
Table 3. LD estimation of Ne in the past 200 generation according to different mathematic formulas. LD estimation of Ne in sample size of total (100) and in sample of 50 and 25 individuals in population has been calculated.

(a.1) Chip size : 200 SNP/chromosome, Sample size : 100 individuals

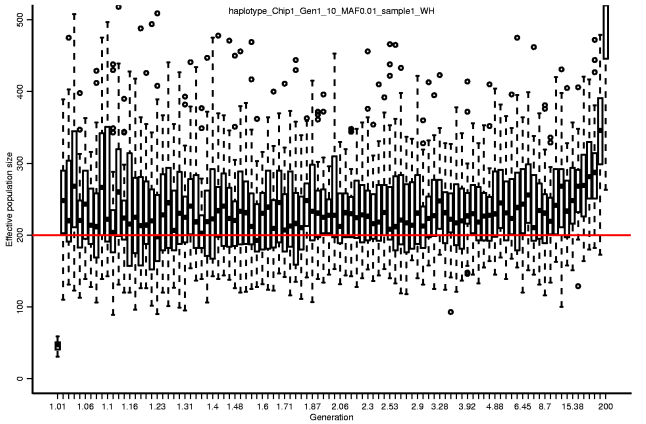
Ober formula



Sved formula



Weir-Hill formula

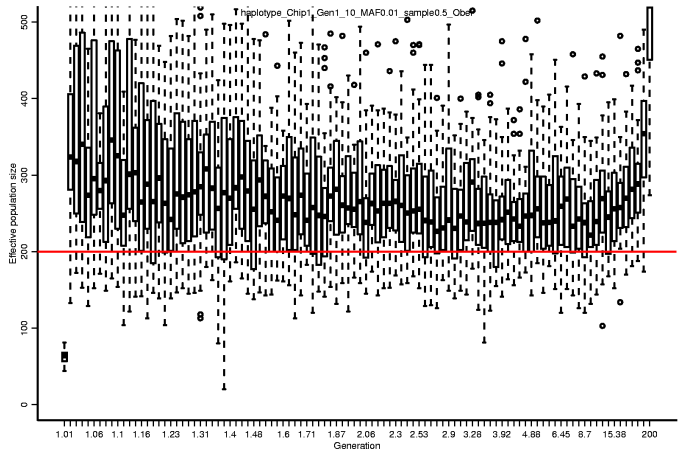


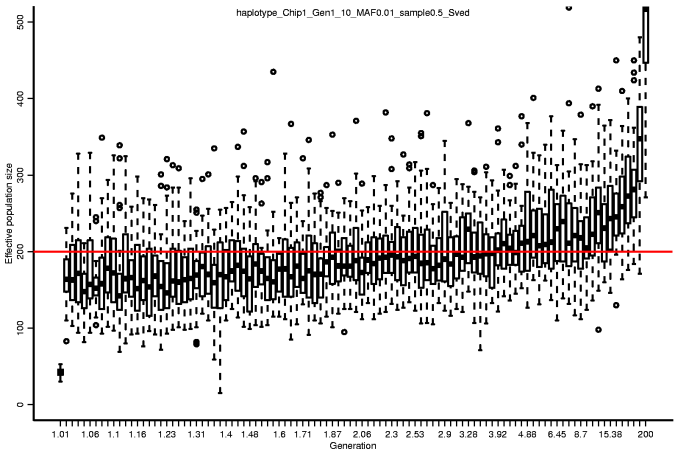
(a.2) Chip size : 200 SNP/chromosome, Sample size : 50 individuals

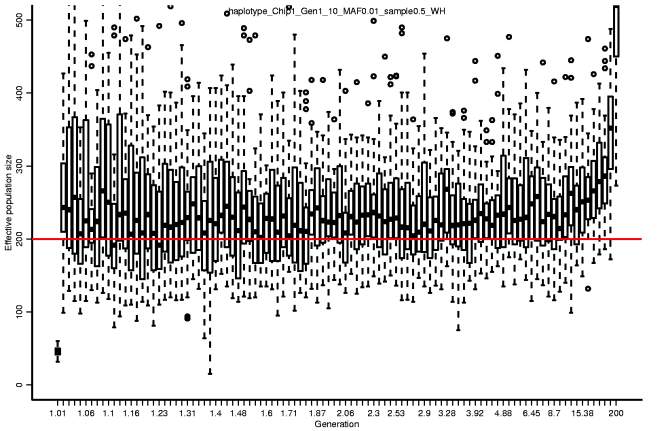
Ober formula

Sved formula

Weir-Hill formula

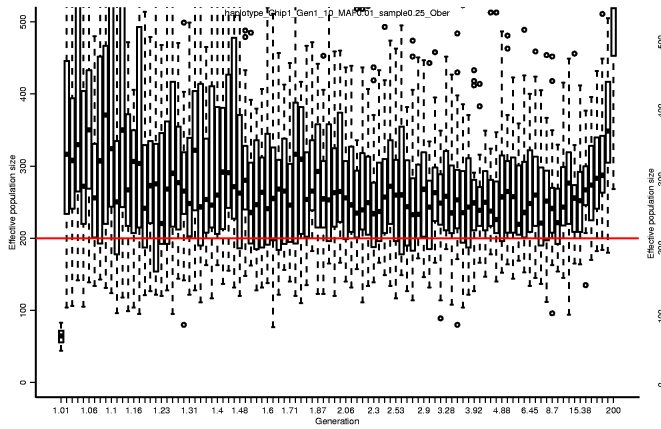


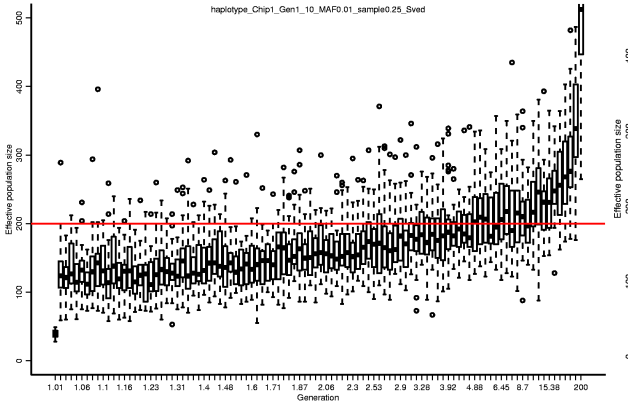


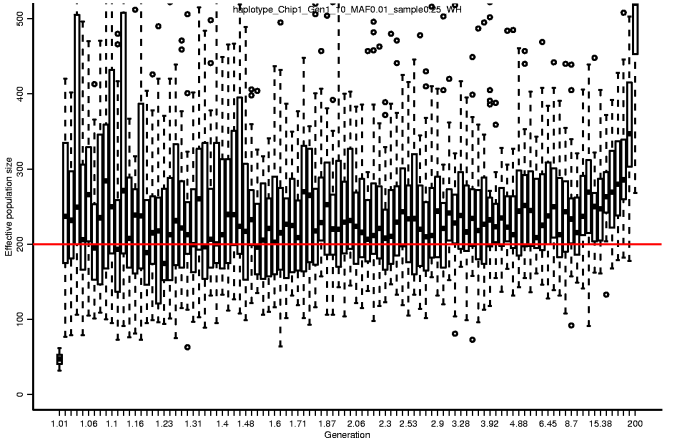


(a.3) Chip size : 200 SNP/chromosome, Sample size : 25 individuals

Ober formula Sved formula Weir-Hill formula

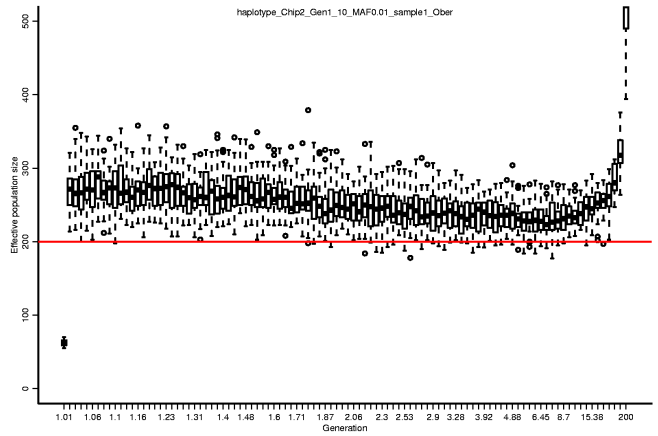


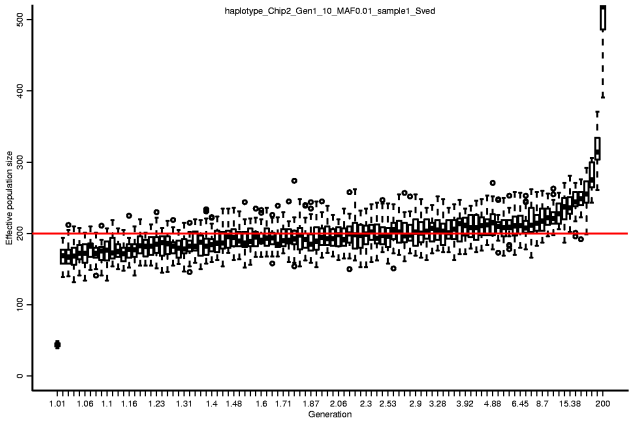


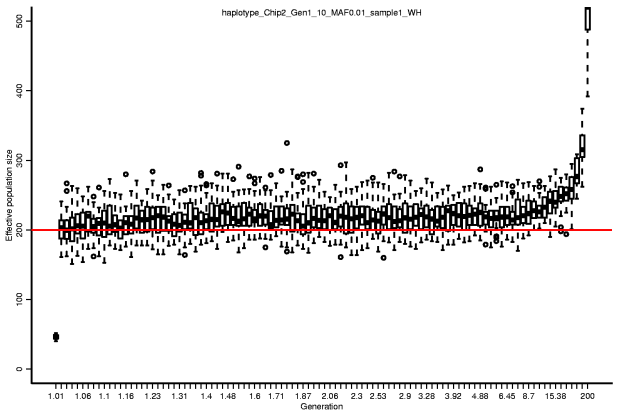


(b.1) Chip size : 2000 SNP/chromosome, Sample size : 100 individuals

Ober formula Sved formula Weir-Hill formula

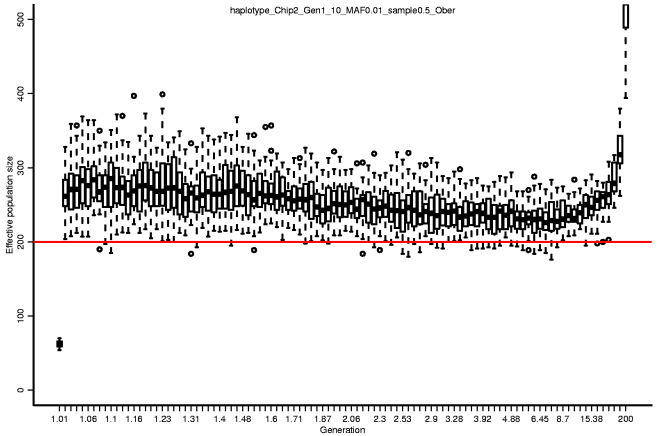

haplotype_Chip2_Gen1_10_MAF0.01_sample1_Ober
Effective population size
0
100
200
300
400
500
1.01 1.06 1.1 1.16 1.23 1.31 1.4 1.48 1.6 1.71 1.87 2.06 2.3 2.53 2.9 3.28 3.92 4.88 6.45 8.7 15.38 200
Generation


haplotype_Chip2_Gen1_10_MAF0.01_sample1_Sved
Effective population size
0
100
200
300
400
500
1.01 1.06 1.1 1.16 1.23 1.31 1.4 1.48 1.6 1.71 1.87 2.06 2.3 2.53 2.9 3.28 3.92 4.88 6.45 8.7 15.38 200
Generation

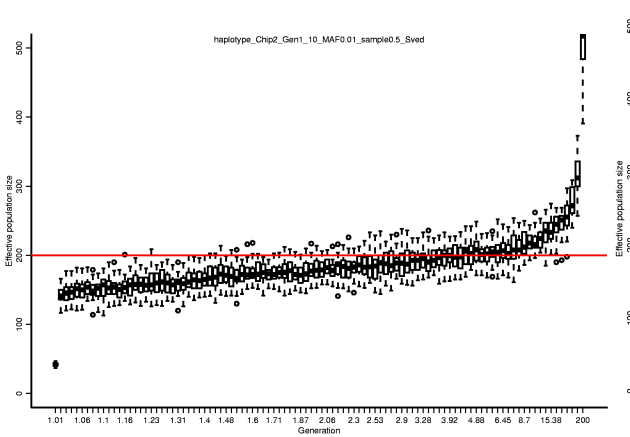

haplotype_Chip2_Gen1_10_MAF0.01_sample1_WH
Effective population size
0
100
200
300
400
500
1.01 1.06 1.1 1.16 1.23 1.31 1.4 1.48 1.6 1.71 1.87 2.06 2.3 2.53 2.9 3.28 3.92 4.88 6.45 8.7 15.38 200
Generation

(b.2) Chip size : 2000 SNP/chromosome, Sample size : 50 individuals

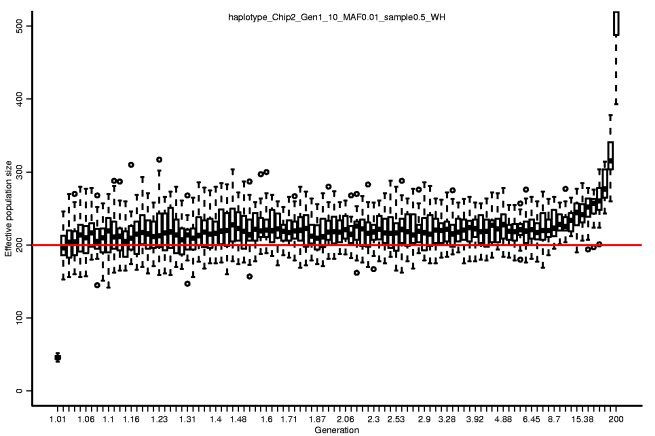
Ober formula



Sved formula

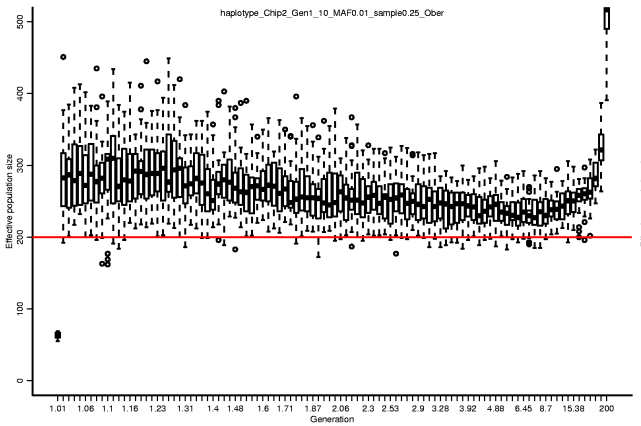


Weir-Hill formula

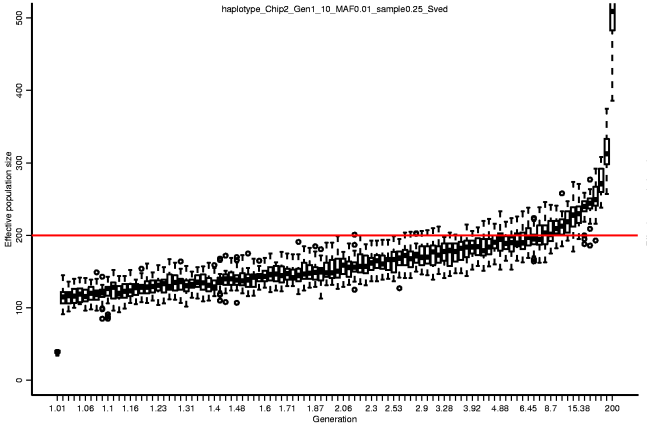


(b.3) Chip size : 2000 SNP/chromosome, Sample size : 25 individuals

Ober formula



Sved formula



Weir-Hill formula