**Validation of the simulation of genomic data exhibiting realistic MAF distribution and LD pattern with Evogen**

**Objective:** compare LD simulated with Evogen with predictions under genetic models and real data

Questions

Why number selected males and females > number to be selected (here 25)?

This is not due to evogen but select\_perents(…) python-implemented function (now is fixed).

A black screen with white text

AI-generated content may be incorrect.

What is the mutation rate (per haploid or diploid)?

--> see below: the distribution of allele freq indicates that the mutation rate is high relative to Ne

-->  and hence the r2 profile does not match expectations under a simple genetic model

How to tune the mutation rate?

**Measures of LD**

The two most common measures of linkage disequilibrium, D and r2, are respectively equivalent to the covariance and squared correlation between alleles at 2 different loci. Considering two biallelic loci l and m, with alleles A and a at the l locus and B and b at the m locus, there are four possible haplotypes AB, Ab, aB, and ab with respective frequencies pAB, pAb, paB and pab.

The LD measure D is equal to the covariance between alleles at the two loci:

D = pAB - pApB

Where pA and pB denote the respective frequencies of alleles A and B

r2 is a standardized measure in the range 0-1 and is calculated as D2 / pA(1-pA)pB(1-pB)

**LD estimated in livestock populations using real genomic data (SNP arrays)**

As a starting point, the shape of LD decay with inter-marker distance is illustrated for cattle and pig populations using real SNP genotyping data. Some differences in r2 values on short and long distances are observed and might be explained both by genetic diversity before domestication (wild population size), at the time of domestication (founder effect), changes in population size, natural selection, and more recently with the development of modern breeding programs.

The LD between close markers (200kb) is much lower in cattle (~0.15) than in pigs (>0.40). The persistence of LD is higher in pigs at intermediate distances (<5Mb) but the long-range LD is ~0.1 in both species.

Beef cattle breeds – Porto-Neto et al (2015)

A graph of different colored lines

AI-generated content may be incorrect.

Danish pig breeds (Wang et al., 2014)A graph of a number of different colored lines

AI-generated content may be incorrect.

**Prediction of LD**

The LD observed today results from combined effects of multiple processes and factors that make it difficult to simulate with simple genetic models. However, it is important to check that the software can simulate LD expected under simple genetic models to validate that mutation and recombination processes are correctly implemented in the program.

We start by predicting LD in a simple mutation-drift model in which allele frequencies are only affected by mutation and drift. Data will be simulated with evogen, and possibly QMSim, and compared to expected LD measures under well-known models.

**Prediction of LD in a mutation-drift model**

In theory, a simulated population under the mutation-drift model reaches an equilibrium state that depends on the value of 4Neu where Ne is the effective population size and u is the mutation rate.

If 4Neu < 1 then drift overrules mutation, and the probability of losing new alleles obtained by mutation is high over several generations. Hence, allele frequencies follow a U-shape distribution, with most of remaining SNPs segregating at very low frequencies. In this scenario, we expect higher LD possibly on long stretches of genomes if the effective population size is small. For a given effective population size, LD is expected to be stronger for low mutation rates.

If 4Neu = 1 then effects of drift and mutation balance each other. Allele frequencies freely evolve in the range [0 – 1] and SNP follow a uniform distribution. LD will be much small on short distances, and the resulting LD decay much flatter.

If 4Neu > 1 then allele frequencies follow a normal distribution because mutations bring in diversity and keep most of allele frequencies towards 0.5. this will result in low LD both on short and long distances.

If we consider that the effect of mutation is negligible 4Neu<<<1, Sved (1971) derived a simple formula for estimating r2 as a function of effective population size and the recombination fraction between loci (c).

A math equation with numbers and symbols

AI-generated content may be incorrect.

If c is small enough, then the term between brackets at the denominator will be close to 1. Then, this equation simplifies to:



Considering Haldane’s mapping function, the recombination fraction can be estimated as:

C = 0.5(1 – e-2d)

Where d is the genetic distance between loci in M. In mammals we consider as a rule of thumbs that 1cM ~ 1 Mb.

A graph of different colored lines

AI-generated content may be incorrect.

**Simulations with Evogen**

In the following we simulated the genome of a population whose size and number of generations have been varied as follows:

* Ne : 50, 100, 500
* Number of generations: 1000, 5000

The genome was made up of 1 chromosome comprising 100 000 000 base pairs, and we retained 1 SNP every 10 000 bp, meaning that the SNP panel is made up of 10 000 markers. This is close to a standard chromosome in mammals and the SNP density of a medium-density SNP array. At start of simulation, the allele frequency was assumed 0.5 and SNPs were in linkage equilibrium.

The effect of these parameters is plotted in the graphs below considering SNP with MAF>5% to get insight in the allele frequency distribution and LD decay.

The distribution of allele frequencies indicate that mutation rate might be high with respect to effective population sizes. Indeed, the distribution of allele frequencies lies somewhere between a uniform and a U-shaped distribution when Ne=50, and it is a gaussian for Ne>=100.

As a result, r2 values obtained in evogen simulations are much lower than expected under the mutation drift equilibrium. For example, when Ne=50, we estimated r2~0.05 for SNP that are 1Mb apart vs 0.32 under the mutation drift model. No gain in LD was observed when increasing the number of simulated generations.

These LD values are much lower than observed in real populations.

Ne=50, Ngen=1000

Distribution of allele frequencies

A graph of a number of bars

AI-generated content may be incorrect.

A graph with a line drawn on it

AI-generated content may be incorrect.

Ne=50, Ngen=5000

A graph of a number of bars

AI-generated content may be incorrect.

A graph with a line drawn on it

AI-generated content may be incorrect.

Ne=100, Ngen=1000

A graph with a line drawn on it

AI-generated content may be incorrect.

Ne=100, Ngen=5000

A graph of a number of bars

AI-generated content may be incorrect.

A graph with a line drawn on it

AI-generated content may be incorrect.

Ne=500, Ngen=1000

A graph with a line

AI-generated content may be incorrect.

Ne=500, Ngen=5000

A graph of a number of black and white bars

AI-generated content may be incorrect.A graph with numbers and lines

AI-generated content may be incorrect.

**Example with QMSim**

Below are simulations obtained with QMsim, for a population with an effective size of 200 made up of 100 males and 100 females mating at random for 1000 generations, each female having 2 offspring. The LD is presented as an average for all pairs of SNP situated at a given distance and for 3 values of 4Neu (0.1, 1 and 10).

A graph of data with numbers and lines

AI-generated content may be incorrect.

In this example, the simulation with 4Neu=0.1 has a LD pattern much more similar to what is observed in livestock populations. However, the distribution of allele frequencies is U-shaped with fewer SNPs segregating as the number of generations is increased.

A graph of a number of people

AI-generated content may be incorrect.

To simulate realistic ancestral populations, there is a compromise to find to be able to simulate the right LD structure and have sufficient SNPs segregating so that we can create a SNPs with distribution of allele frequencies similar to what is used in practice (~uniform distribution). To approach this case, it is needed to simulate many SNPs and adjust the mutation rate accordingly.

In the simulations above, I only considered a recurrent mutation rate.

With QMSim, it took 1 to 2 seconds to simulate 1 replicate of a population for 1000 generations.

With ADAM, it takes 2 days per replicate to simulate 1000 generations but considering default parameters that are very different form above. Hopefully the simulation of my 5 replicates is finished by Monday.

In AlphaSimR, simulation of genomic data is carried out with MaCS (chen et al., 2009), a genome simulation software grounded on coalescent models with recombination.

**Examples of LD decay with intermarker distance for livestock populations.**

LD decay in German Holstein (Qanbari et al., 2010)

A screenshot of a graph

AI-generated content may be incorrect.

A graph of a number of different colored lines

AI-generated content may be incorrect.

A good review paper on estimation of LD in livestock by Qanbari (2020) and below the effect of the ascertainment bias induces by the pre-selection of SNPs put on SNP chips

A graph of a function

AI-generated content may be incorrect.