Toy Simulation of Linkage Disequilibrium
Supplementary information of
a way to analyse GWAS data using exons

Daniel Rovera (daniel.rovera@gmail.com) supervised by Chloé-Agathe Azencott, June 16, 2022

Institut Curie, PSL Research University, F-75005 Paris, France, INSERM, U900, F-75005 Paris, France
CBIO-Centre for Computational Biology, Mines ParisTech, PSL Research University, 75006 Paris, France

Abstract

The repartition of SNPs along the genome respect a γ law where the shape parameter is less than 1. So they tend to agglomerate. This distribution seems be due to the linkage disequilibrium. A toy simulation reinforce this assertion.

Keywords: *SNP, linkage disequilibrium, gamma law*

Repartition of SNPs

the repartition of SNPs respects a γ (fig 1) law with a significant dispersion of values:
shape = 0.75 - 0.82
scale = 230 - 650

The variants contributing to an analog effect on phenotype tend to agglomerate.

This law about SNPs seems to be the result of linkage disequilibrium, indeed the correlation between loci decreases with the distance between loci as measures reported in publication of LD show. This simulation belows intended to confirm this interpretation.

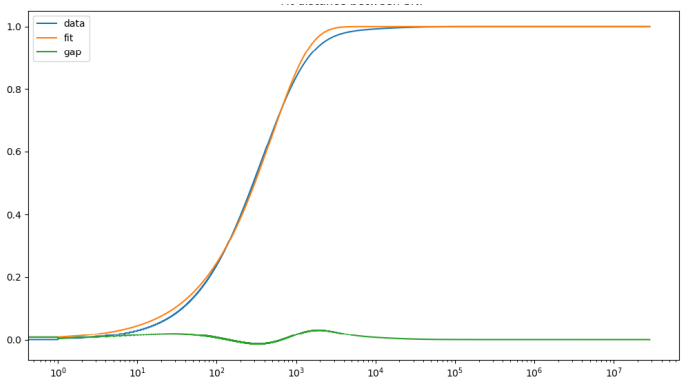

Figure 1: Graph of fitting distances between SNPs (GCST006719 from www.ebi.ac.uk/gwas/summary-statistics)

Process of Simulation

Loci are located randomly along a segment according to an uniform distribution, distances between loci following an exponential law.

Along a fictitious chromosome, the state of the next locus is drawn successively according to the state of the previous locus and the conditional probability from the formula of linkage disequilibrium.

The usual notation is: allele A occurs with frequency P_A, allele B with frequency P_B. So, $LD = P_{AB} - P_A * P_B$. The conditional probabilities linking successive states are:

$$P(A1/B1) = \frac{P_A * P_B + D}{P_B} \qquad P(A1/B0) = \frac{P_A * (1-P_B) + D}{(1-P_B)}$$

$$P(A0/B1) = \frac{(1-P_A) * P_B + D}{P_B} \qquad P(A0/B0) = \frac{(1-P_A) * (1-P_B) + D}{(1-P_B}$$

The linkage disequilibrium is measure as correlation coefficient between indicator variables for the presence of alleles following Bernouilli's law, it varies as LD:

$$R^2 = \frac{(P_{AB} - P_A * P_B)^2}{P_A * (1-P_A) * P_B * (1-P_B)}.$$

The ceiling of LD named LD_{max} is the minimum of $(1 - P_A) * P_B$ and $P_A * (1 - P_B)$ (probability is between 0 and 1). LD varies from LD_{min} to LD_{max} when the distance between loci varies from the length of chromosome to zero. The value of LD is approximated by a decreasing exponential function of the distance between loci:

$$LD(d) = LD_{max} * e^{\frac{1-d}{s}}, \text{ d: distance between loci, } s = \frac{1 - LenghtOfChr}{ln(LD_{max}) - ln(LD_{min})}$$

Then LD is multiplied by a factor between 0 and 1 to evaluate the effect of LD. $\frac{LD_{min}}{LD_{max}}$ is taken to 0.01. P_A and P_B are taken to 0.5, they stay approximately to this value during simulation.

Result of Simulation

It follows that the repartition of intervals between active loci shows a trend to accumulate each other when the multiplicative factor of D increases from 0 to 1 fig 2. The maximal distance between them tends to increase too.

Without linkage disequilibrium (ld *0), the graph is an exponential. The uniform distribution generates a Poisson process where the distribution between events is exponential. The deformation by variation of LD makes it look like a γ distribution.

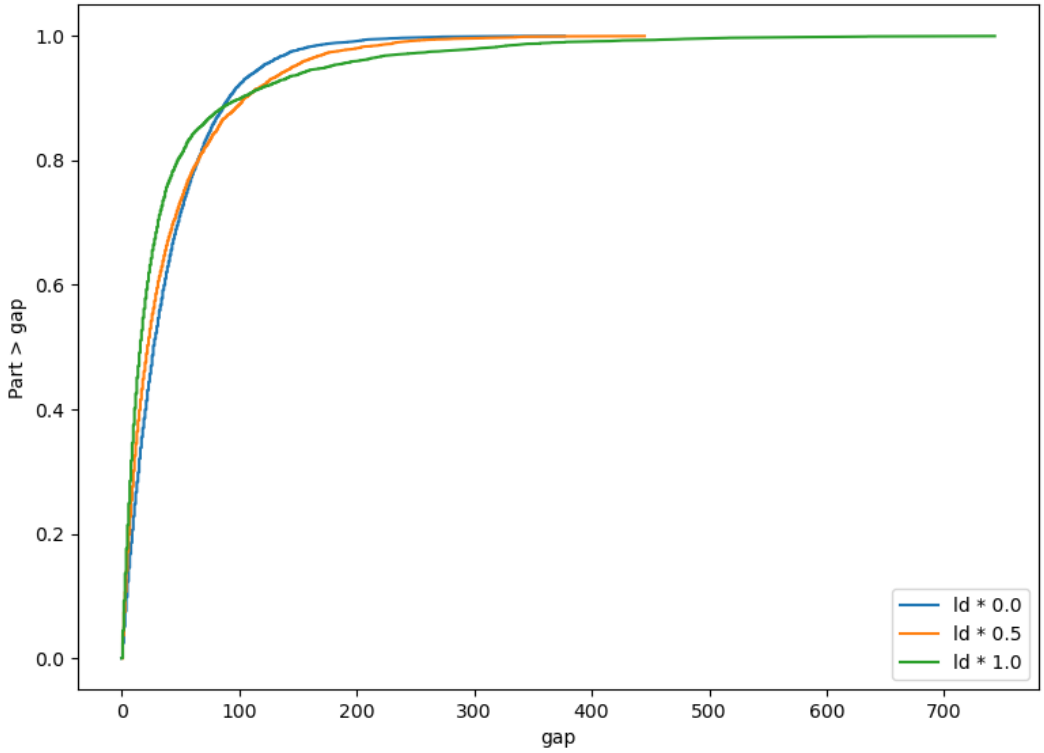

Figure 2: Cumul of distances between active loci for different values of linkage desiquilibrium