MINES ParisTech
Center for Computational Biology

institut Curie

A way to analyse GWAS data using exons

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

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 - June 8, 2023

Abstract

Genome Wide Association Studies aim to analyse the link between frequency of single nucleotide polymorphism and phenotype by comparing to a reference population. The final goal is to find relevant biomarkers involved in predisposition to diseases.

The first step of this analyse is evaluating how the p-values of SNPs resulting from comparison are transferred to genes. The commonly used methods consider genes in totality and base their computing on χ^2 statistics. However, in the genome, genes are divided into exons and are built by a mechanism where splicing is a key step. The SNPs close to splice sites, which are near exons, disturb the structure of the concerned genes. The distances from SNPs to exons are a factor influencing the effect of SNPs.

Using summary statistics of GWAS, the method is based on the statistical relation between positions of SNPs and positions of exons. The obtained weights are introduced in the Stouffer's Z-score formula. So, the p-values transferred from SNPs to genes become a function of distance from SNPs to exons.

Moreover, a simple geometric analysis of the couple distance and Z-score shows an area of extreme values. The analysis can be directed to a subset of SNPs which probably plays a role in difference of genotype.

Keywords: GWAS, SNP, exon, intron, gene, splicing, gamma law, linkage disequilibrium

Aim and principle

Genome Wide Association Studies allow analysis of the link between frequency of single nucleotide polymorphism and phenotype (for instance predisposition to specific disease) by comparing to a reference population. Their results are synthesized in the summary statistics used here. They provide a p-value measuring the gap between the frequency of every SNP in the studied population and the reference population whose features are similar but without the phenotype. Their goal is to find biomarkers which are presumed to be involved in the phenotype.

The aim of this study is to find a function giving p-values of genes from p-values of SNPs based on the consistency of genes and where they are along the genome. Genes are divided into exons which are transformed in messenger RNA (mRNA) by the mechanism of splicing. So, this method is very different from methods usually used softwares such as VEGAS2 [15] or MAGMA [5] based on statistical comparisons using χ^2 statistic (see Comparison to MAGMA).

The elements of human genome on which are focused are the coding sequences (a tiny part less than 2% of genome) and, between them, the introns, interfering in gene expression by splicing (about 30% of genome). The other parts of the genome: regulatory DNA sequences, repetitive DNA sequences and mobile genetic elements are neglected. Notably, other effects are possible by the regulation of expression by regulatory DNA sequences which is not yet well known. Taking them into consideration is difficult.

Base positions of exons are extracted by chromosomes from ***ncbi.nlm.nih.gov***(acces by [23]). The computed number and length are globally:

- 20,108 genes divided in 1 to 190 exons (repartition fig 1)
- 193,532 exons covering 1,1 % of genome
- exons and introns covering 34 % of genome
- only about 12% of exons include one gene

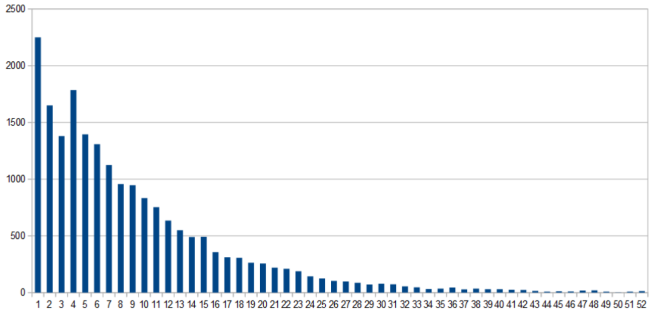

Figure 1: Number of exons by genes, until 52 exons

Cellular machinery builds proteins from messenger RNA. A primary transcript of DNA is transformed in mRNA by removing introns between exons and binding exons. This process is splicing. An exon can be used to make several genes. Here the analysis is not about the detailed mechanism but is based on statistics about relative position of mutations and genes.

SNPs are punctual and some studies are about repartition of SNPs along the genome [7] and [10], but they do not take into consideration the position of other elements involved in building genes.

The relative positions of SNPs in respect of exons imply different ways of influencing gene expression. With the reasonable hypothesis: the effect of SNPs is not able to jump over neighboring exons, six cases are distinguished:

1. inside an exon: effect on thin structure of the gene;
2. between 2 exons of the same gene: effect on global structure of the gene by disturbance of splicing;
3. between 2 exons of 2 different genes and included in the 2 genes: effect on structure of 2 genes by splicing;
4. between 2 exons of 2 genes and included only in 1 gene: effect on structure of the 1 gene by splicing, no effect on the other;
5. between 2 exons and outside any gene: no effect;
6. before the first exon and after the last exon: no effect.

The effect of SNPs to genes is different according to the relative position, confirmed by some publication of effect by SNPs on splicing ([12] [16] and [13]). For the case 'no direct effect', other effects are possible by the regulation of expression by regulatory DNA sequences which are neglected here.

Observations about Relative Positions of SNPs and Exons

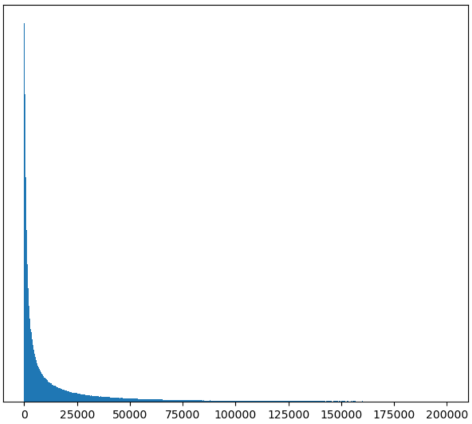

Figure 2: Histogram showing the numbers of SNPs by bins of minimal distances from SNPs to exons (GCST006719 from ebi [22])

The repartition of SNPs appears not uniform along the genome. It can be observed visually and reported in [11]. Further the proximity of exons seems to increase the density of SNPs. The proximity of exons plays a role in its variations: more SNPs are near exons, more they are numerous. So, the chosen variable for evaluating these variations is the minimal distance from SNPs to exons. Exons are considered as anchors in the genome. The histogram 2 confirms this observation.

The profile of this relation is specified by drawing the normalised cumulative number of SNPs in function of the minimal distance from SNPs to exons (fig 3). Obviously, these inside genes are nearer to exons than these outside genes. In the following, only the SNPs inside the genes will be taken into account.

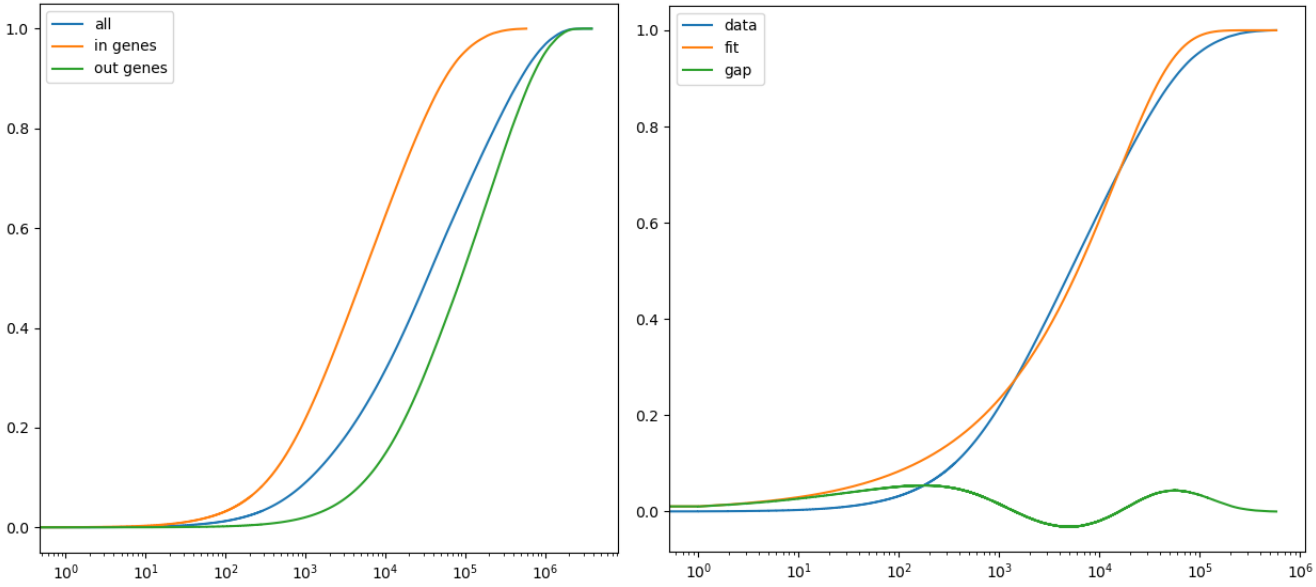

Figure 3: On the left, position of SNPs relative to exons (orange in genes, green out genes, blue all), normalized cumulative number in function of minimal distance of SNPs to exons, logarithmic scale on the abscissa.
On the right, fitting the normalized cumulative number in function of minimal distance of SNPs to exons by γ distribution (logarithmic scale on the abscissa).
(source: GCST006719 from ebi [22])

This distribution can be fitted by a γ distribution whose probability density function (pdf) is:

$$\gamma.pdf(x, shape, scale) = \frac{1}{\Gamma(shape) \, * \, scale^{shape}} \; * \; x^{shape \, - \, 1} \; * \; e^{-\frac{x}{scale}}$$

The fitted curve (fig 3) is the cumulative density function (cdf) of γ law whose parameters are explained below .

This distribution can be interpreted as a heterogeneous spatial Poisson process. The intensity λ, average number of points by length, depends on the location. The inhomogeneous Poisson process is a counting process n(t) in a time interval where the intensity λ gives the number of events in an interval by:

$$E(n(t+h) - n(t)) = \int_t^{t+h} \lambda(\tau) d\tau$$

$\lambda(\tau)$ is the probability density function multiplied by the number of SNPs and t is the minimal distances from SNPs to exons: $\lambda(t) = N * \gamma.pdf(t)$. The intensity is used next as weight, the term N can be removed without changing the result.

Therefore the intensity λ is the parameter of Poisson law which is not constant contrary to the well known law. This can be pictured by cars driving along a straight line where there are villages which slow the traffic (traffic light, speed limit). Exons are villages, SNPs are cars at a moment and intensity in the instantaneous car flow. Finding a γ distribution is not surprising, it is common in nature [6]. γ distribution is the maximum entropy distribution for a positive variable, given its mean value and the mean value of its logarithm.

The fitting normalized cumulative number in function of minimal distance of SNPs to exons (fig 3) gives these values of parameters for particular data: shape = 0.44924 scale = 32738. This observation does not concern only this example as the above array shows for different sources [22] [20] (only SNPs inside genes). The parameters obtained are close to each other:

data source	shape	scale	mse*	phenotype
CIDR_AFRO	0.44718	34050	$8.6\ 10^{-4}$	breast cancer
GCST004988	0.44924	32738	$8.9\ 10^{-4}$	breast cancer
GCST006719	0.45382	34094	$8.2\ 10^{-4}$	breast cancer
GCST007236	0.44931	33066	$8.8\ 10^{-4}$	breast cancer
GCST90011804	0.44956	32842	$8.9\ 10^{-4}$	breast cancer
GCST90011808	0.44960	32863	$8.9\ 10^{-4}$	prostate cancer
GCST90011811	0.44956	32839	$8.9\ 10^{-4}$	colon cancer

* mse: mean square error

Shape is the critical parameter for the link between intensity and distance: if shape = 1, γ is exponential, if shape < 1, its form is hyperbolic, if shape > 1, γ increase after that decrease (confer γ distribution for different values of shape [21]). Here the shape is far less than 1, so SNPs tend to agglomerate preferably close to exons. It leads to notice accumulation of SNPs near exons as the histogram 2 shows.

Discussion about these observations

These observations raise two questions. Could this distribution be the consequence of a statistical effect ? Can the fit be improved?

The statistical effect can be analysed by simulation. The SNPs are almost evenly distributed. Their repartition respects a gamma law with a significant dispersion of values (shape = 0.75 - 0.82 and scale = 230 - 650). The uniform distribution corresponds to shape equals to 1.0. The law of intervals between values uniformly distributed is exponential.

The distribution of SNPs seems to be the result of linkage disequilibrium. The correlation between loci decreases with the distance between loci as measures of LD show [18] [1]. The consequence is that the variants contributing to an analog effect tend to agglomerate as shown by a simulation. Their repartition is near a γ distribution (see supplementary information Result of Simulation).

On this basis, SNPs are drawn along a female genome with shape = 0.80 and scale = 640. To simplify, exons are taken as punctual at their middle. The minimal distances from SNPs to exons is computed in both cases: this random distribution and the found law γ with shape = 0.45 and scale = 33400. A ratio between the both is computed:

minimal distance	10^1	10^2	10^3	10^4	10^5	10^6
ratio	22.6	3.8	2.1	1.7	1.2	1.1

These values confirm the accumulation of SNPs close to exons.

Improving fit is possible by using the generalized gamma distribution:

$$gen\gamma.pdf(x,a,c) = \frac{c}{\Gamma(a)} * x^{ca-1} * e^{-x^c}$$

In addition to the parameters a and c, loc (localization) and scale are used to shim and scale the function. The fit result is better (fig 4), the mean square error is divided by 30, above the detail for the breast cancer GWAS:

data src	a	c	loc	scale	mse
GCST006719	4.53286	0.236350	59.2710	12.65112	2.8E-05
CIDR_AFRO	5.13095	0.219329	51.4059	4.16039	2.8E-05
GCST004988	5.91680	0.204237	49.2646	1.13130	2.5E-05
GCST007236	5.75427	0.207206	50.3181	1.48760	2.6E-05
GCST90011804	5.80228	0.206401	50.4562	1.37293	2.6E-05

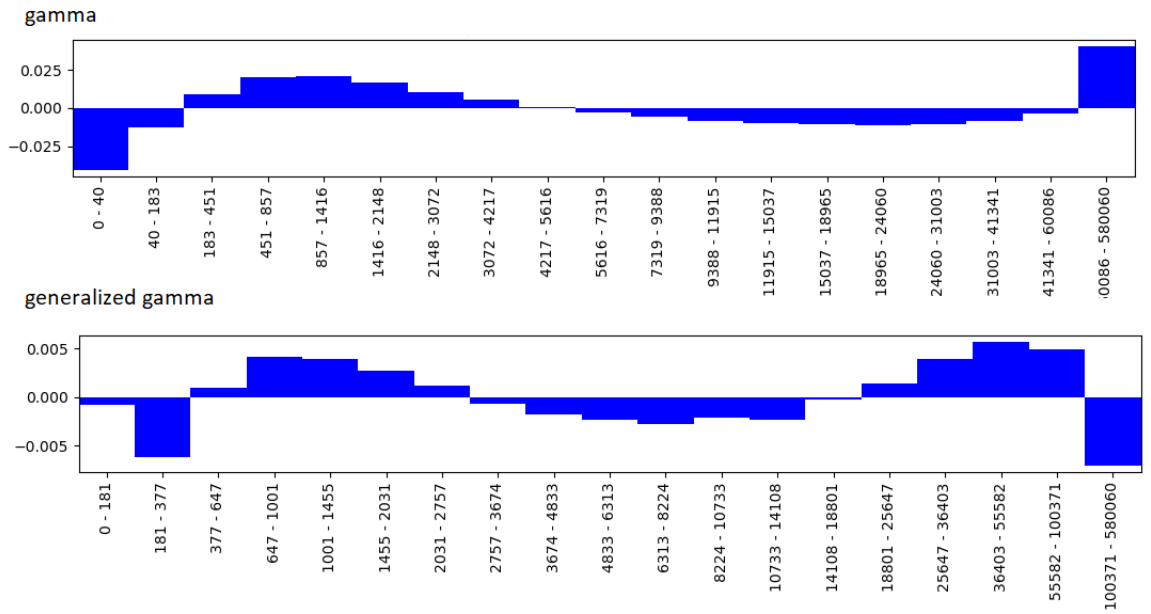

Figure 4: Quantile quantile plot as slice of 5% of theoretical number of gamma and generalized gamma law, the usual qq plot by SNPs gives a distorted vision because the density of points decreases dramatically along the value axis (GCST006719 from ebi [22])

The dispersion of parameters is greater than fit by γ law. The relative standard deviations of the parameters are 9.7%, 5.6%, 7.0%, 105% for generalized gamma law and 0.5%, 1.8% for γ law. This is the result of an overfitting which complicates the method without providing real better precision.

Also intervals between exons follow a generalized gamma distribution whose parameters are a = 4.223377, c = -0.313196, loc = -89.94722, scale = 144093.2 and mean square error 8.2E-05. It shows the generalized gamma distribution is a very versatile law.

The statistical power of the simple γ law is sufficient. The maximal deviation between cumulative functions is between 0.053 and 0.054. The Pearson correlation coefficient between data and fit is between 0.995 and 0.999. The power computed through the Fisher z-transformation is 0.999. So, only the simple γ law is used in the following.

Method

The accumulation of SNPs close to exons is a real phenomenon. The SNPs located preferentially inside genes and near exons according to a γ law with a scale much lower than 1, dramatically decreasing with the distance, which suggest that effect of SNPs on genes decreases also when the distances from SNPs to exons increases. The relative positions of SNPs and exons play a leading role in the effect of SNPs on genes. So, the method to evaluate the quantitative link between SNPs and genes is based on the following rules:

- Only the direct effect is kept for SNPs inside genes, the indirect regulation is neglected
- The scope of the effect by SNPs is limited to the neighboring exons and does not jump over exons
- The resulting p-values of genes are computed from the weighted mean of Z-score
- The weights used in Z-score formula are the probability density function of gamma law

It follows that using the probability density function to transfer the p-value from SNPs to genes seems natural. Z-scores are used to place this problem in a linear space. The vector of Z-scores at loci is conditioned by the presence of the SNPs. The resulting Z-score is equal to the conditional mean of Z-scores worn by SNPs. The conditional means is the projection of the value vector onto the probability vector (the linear space has the inner product).

With vector of $\vec{Z} = \begin{bmatrix} z_1 & \cdots & z_n \end{bmatrix}$ and vector of probabilities of attendance of $\vec{P} = \begin{bmatrix} p_1 & \cdots & p_n \end{bmatrix}$, the conditional mean is: $E(\vec{Z}) = \frac{\vec{Z}.\vec{P}}{||\vec{P}||}$. Attendances of SNP_i and values of Z_i are not independent. Along the genome, loci are correlated by the effect of linkage disequilibrium. This used as computation basis by [15] and [8]. So, the norm $||\vec{P}||$ is not Pythagorean. The space is distorted by the non independence of $\vec{Z}$ and the unit vectors are not orthogonal. The cosine of their angles are the correlation coefficients. If R is the linkage disequilibrium correlation matrix, $||\vec{P}||^2$ is the bilinear form PRP^t. The formula becomes:

$$E(Z) = \frac{PZ^t}{\sqrt{PRP^t}}$$

This formula results also from the Stouffer's Z-score method ([25]) with a particular choice of weight coefficients.

$$Z = \frac{\sum_i w_i * z_i}{\sqrt{\sum_i w_i^2 + 2 * \sum_{i<j} r_{ij} * w_i * w_j}}$$

The simple formula of Z-score does not contain the term $2 * \sum_{i<j} r_{ij} * w_i * w_j$. It is commonly used in meta analysis considering analysis statistically independent. But in this case this term must be taken into account. Indeed r_{ij} is the correlation coefficient between the data used for computing the p-values of SNP_i and SNP_j. This correlation coefficient must be not null because the frequency of alleles at two loci are not independent due to the linkage disequilibrium. The terms of the correlation matrix R are $r_{ij} = Linkage\ Disequilibrium\ Matrix(SNP_i, SNP_j)$ (same references than above [15] and [8]). Noting $r_{ij} = r_{ji}$ and $r_{ii} = 1$ and $W = \begin{bmatrix} w_1 & \cdots & w_n \end{bmatrix}$, the Stouffer's formula becomes:

$$Z_{snp} = \frac{WZ^t}{\sqrt{WRW^t}}$$

The Stouffer's Z-score method is outside the usual scope of use. Every measure of p-value is considered as an analysis. So, the Z-score due to SNPs is estimated by this formula. The weights must be determined and they are the subject of discussion [3]. The choice of Mosteller and Bush (1954) is the size of samples. The density of probability is a number of SNPs by length unit. So, it is assimilated to a differential size of samples. In this frame, the both formulas: conditional probability and Stouffer are the same with $W = P$.

In conclusion, the formula chosen is $Z_{snp} = \frac{WZ^t}{\sqrt{WRW^t}}$ where W is the $\gamma.pdf()$ using the estimated shape and scale. The Z-score of every gene is computed from the Z-score of SNP_i in the scope. A SNP located between two exons influences the two genes that contain them. Z_i and w_i can be repeated, so the same goes for r_{ij}. The weights w_i are the probability density function at the distance between SNP_i and exons of this gene. They are one or two according to the case because the effect of SNPs does not jump over genes.

When the SNPs are inside exons, the w_i go to infinity in numerator and denominator. Z is computed with the same formula where $W = \begin{bmatrix} 1 & \cdots 1 \end{bmatrix}$.

The effects of SNPs outside exons and inside genes are computed separately. This computing allows getting two lists of genes sorted by decreasing computed Z in two cases: SNPs inside exons and SNPs inside genes but not inside exons. Their effects may be added for genes influenced in two ways by the Z-score method (sum divided by $\sqrt{2}$). But the kind of effect, necessary for analysing mechanisms, is lost. The contributions of every SNP to the Z-score of a gene can be detailed by every term of the Z-score formula. These detailed results show how a set of SNPs can reinforce its effects.

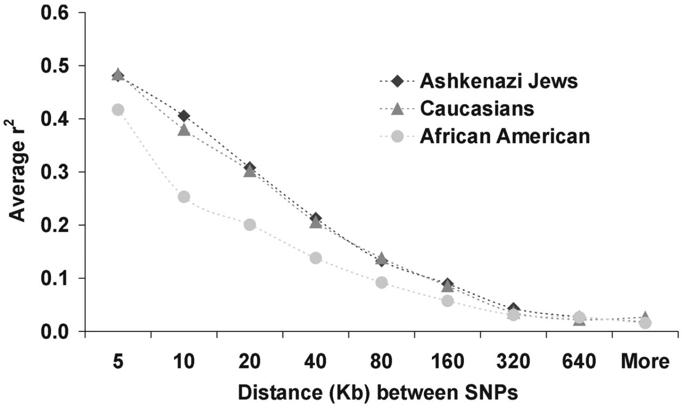

Figure 5: Linkage disequilibrium correlation coefficient in function of distance between loci [19]. This graph is fitted by $r^2 = \frac{1}{1+0.017*\sqrt{distance}}$

On a practical level, the computing of the whole matrix R is out of reach, but it is possible to get a submatrix on the subset of SNPs influencing a gene. It can be done by a software such as LDlinkR [14]. The linkage disequilibrium decreasing regularly with the distance of loci, the approximate formula can be used, which speeds up the calculation (fig 5). If the detailed statistics are available, the LD correlation matrix R can be computed from them. Sometimes the p-values take extreme values. Python cannot compute the Z-score below $5.55 * 10^{-17}$. The values below this threshold must be discarded as outliers. This is the case of GCST006719 and GCST007236 [22].

Focus on extreme values

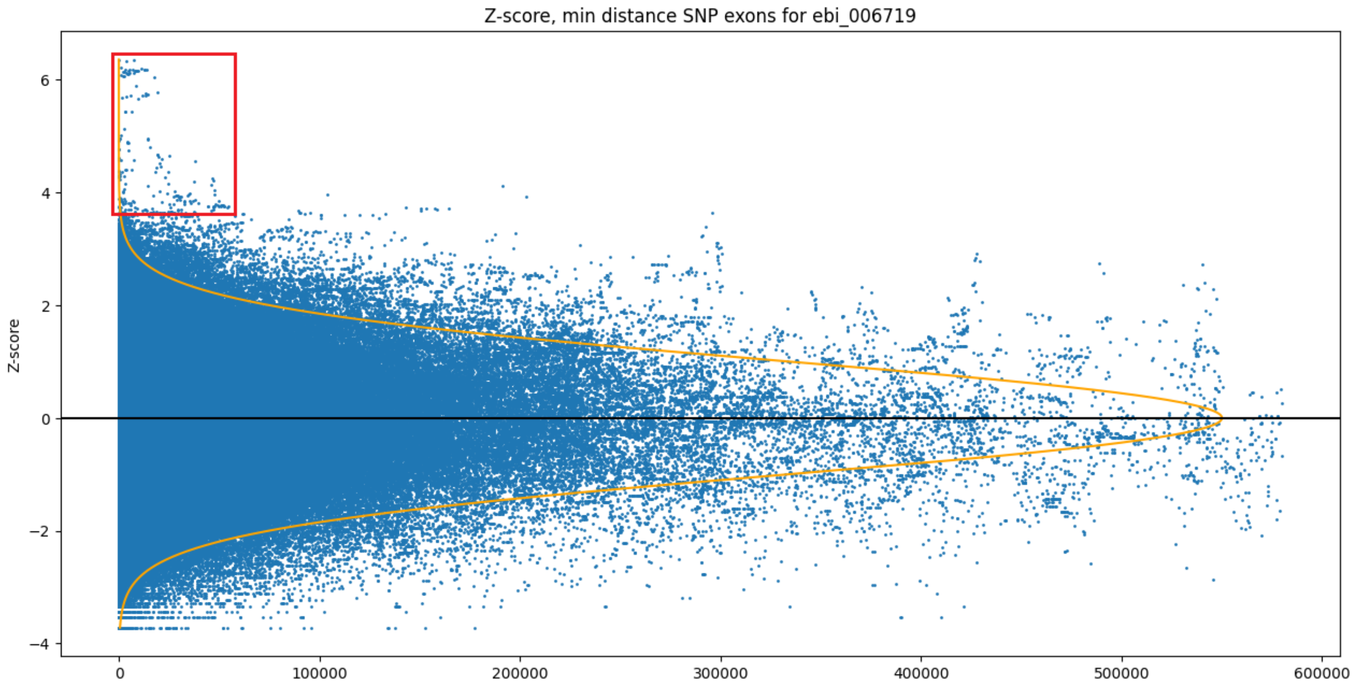

Figure 6: Z-score of SNPs inside scopes of genes in function of their minimal distances to exons and an approximate gaussian distribution in orange (GCST006719 from ebi [22])

The figure 6 is a scatter plot of Z-score of SNPs inside scopes of genes in function of minimal distance of SNPs to exons. So only SNPs having an effect are taken into account. It confirms the non-uniformity of the distribution of SNPs quantified by the γ function.

The qualitative observation of extreme values framed in red on the figure gives a localized information. So the analysis of the effect of SNPs in genes can be focused on this subset of SNPs. The weighting formula quantitatively completes this analysis. An incidental remark: the contribution by a subset of SNPs may exceed 100% due to possible negative Z-scores. This analysis can be used for starting an accurate biological study.

Comparison to MAGMA

MAGMA [5] is based on χ^2 statistics, a commonly used method. The method of MAGMA is of the same type as those implemented in PLINK [17] and VEGAS [15]. There are some small differences between them. MAGMA has the advantage of being computationally performant and easy to be used.

The results by MAGMA are far from these by SNP exons method. The difference of results is confirmed by the two graphs: gap for every gene sorted by Z-score and the histogram of gaps fig 7. The histogram shows a non completely gaussian difference. But, these differences upset the ranking of genes. To number these differences, the Euclidean distance between the resulting Z-score is equal to 66.

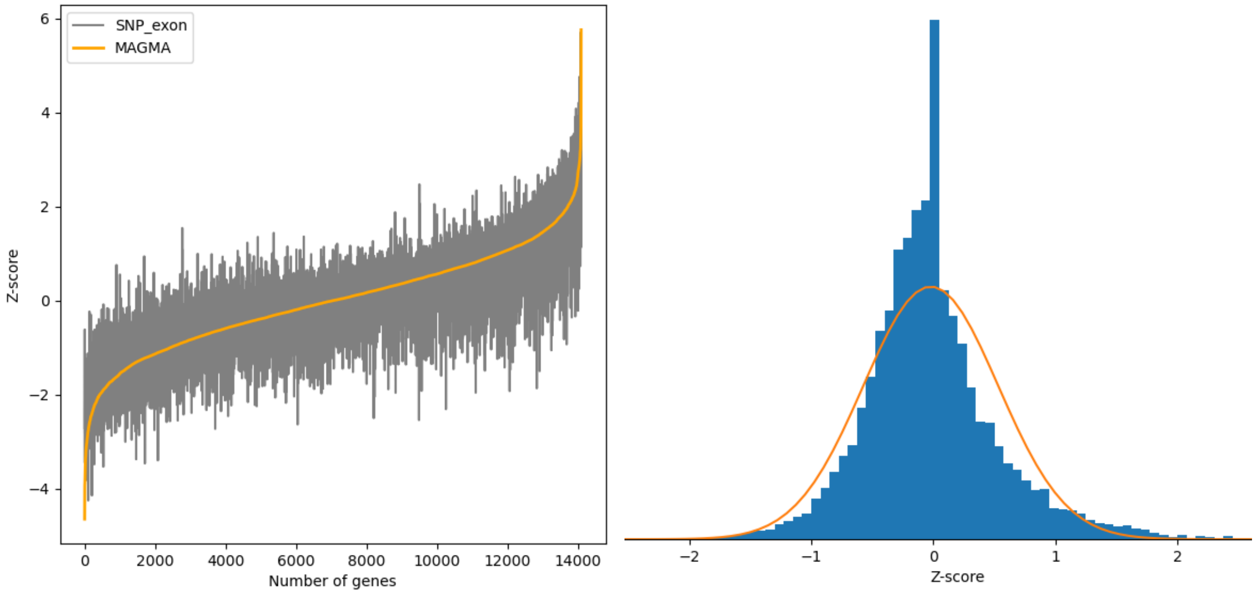

Figure 7: Cumulative Z-score of genes and histogram show the difference of outcome between MAGMA and SNP exon method, MAGMA is reference (GCST006719 from ebi [22])

A surprising ascertainment: the moving average of a hundred gene Z-score almost coincides with the result, which can be explained by a similar process of distributing the Z-score of SNPs to genes. But this ascertainment has a weak interest to explain these differences.

To identify the origin of differences, a way is checking the sensitivity of results to data by playing with both variables: p-value and base position. So, artificial data is created to evaluate the sensitivity to data. Firstly all p-values are made equal to the geometric mean of p-values. Secondly, SNPs are distributed uniformly along the segments of chromosomes common to genes, so SNPs stay inside genes where they were and variations of distances are canceled. Genes taken from first exon to last exon may cover each other, certainly it is infrequent. The Z-scores of genes obtained with these artificial data are compared to the Z-score obtained with real data.

At the sight of these histograms 8, the sensitivities to p-values of MAGMA and SNP exon method are very closed, but MAGMA is insensitive to positions of SNPs except to these inside the margin of proximity. Regardless of the position of SNPs inside a gene, MAGMA gives the same Z-score to the gene. On the

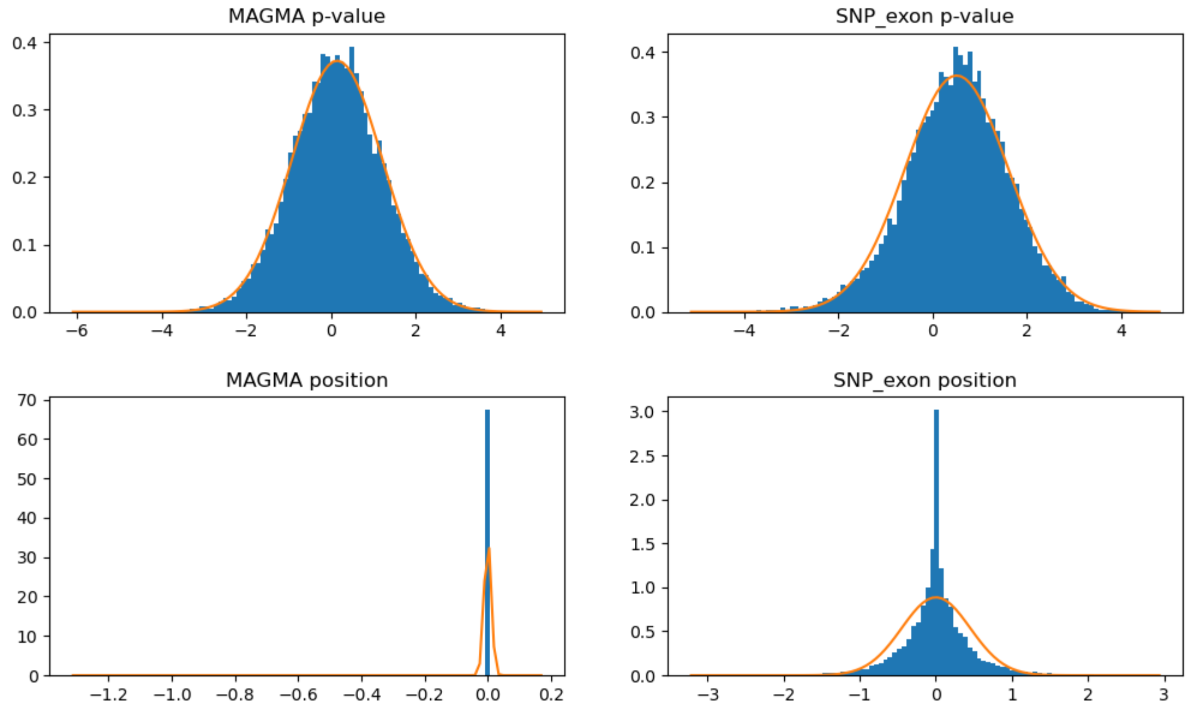

Figure 8: On the left Z-scores of genes got by MAGMA with same p-value and homogeneous position compared to result from real data as histogram, on the right same comparison for SNP exon method(GCST006719 from ebi [22])

contrary, the SNP exon method modulates the effect of SNPs by the position measured by the distance from SNPs to exons. The question is: is the effect of SNP on genes independent of the position of SNPs inside genes ?

Biological knowledge

This observation must be confronted with knowledge in biology. The presence of a mutation within an exon can modify the shape of the active site or the protein folding. This obviously leads to modifications or disappearances of functions as [12].

Likewise several publications show the effect of SNPs on the mechanism of splicing as [9] and [24]. That last publication identifies an area of exonic splicing enhancers located at less than 69 base position. In this area, the density is multiplied by 7 according to the above simulation. These considerations underline the proximity between effective SNPs and exons.

A gene consists of multiple small exons (fig 1) separated by exons by a distance much bigger than the sizes of exons. The splice machinery has the task of finding small right exons among much longer exons. This is the role of splice sites. Splice site sequences that drive exon recognition are located at the begin and the end of introns [2]. So SNPs in or the near splice site can disturb splicing and affect building of genes.

Positions of SNPs in exons or in introns are not neutral to the phenotype.

Results

This method is applied to the five GWAS of breast cancer. The goal is to find some genes common to several GWAS data involved in predisposition of breast cancer. The selection is based on the sum of Z-scores of a given number of data (2 to 5). The case of one is treated in the corresponding publications. The genes whose sum of Z-score is greater than 0.8 * maximum are selected, which gives about thirty genes. The analysis of the role of genes is based on Uniprot[4]

The selected genes are classified into three categories: possible direct link to cancer process, possible indirect link through regulation and no obvious link to cancer process.

In the first category, ATXN7L3 remodels chromatin and deubiquinates histones, CD300LG binds lymphocytes, CXCR3 is a receptor in angiogenic and apoptotic process, DUSP3 is a negative regulator of T cell, GRN acts as regulator in inflammation and proliferation, HDAC5 is responsible for the deacetylation of histones, HROB promotes DNA repair synthesis, LSM12 is a RNA binding protein, UBTF activates transcription mediated by RNA polymerase.

In the second category, ASB16 is involved in protein modification, CCDC91 in protein transport , DNAJB7 acts as co-chaperone, MPP3 participate to PI3K pathway involved in several processes of cell, NAGS plays a role in amino-acid biosynthesis, SLC25A39 is a mitochondrial transporter, SLC4A1 is a transporter and a structural protein.

In the third category, the genes are involved in various processes as metabolism, intra-cell transport and growth of other organs: CCDC194, TXNDC8, SOST, FAM171A2, G6PC3, GPR89A, MRPS30, PPY, PYY, TMUB2, TMEM101.

These results are not probative. The GWAS were not performed on similar populations. But this method could show other tracks to find genes promoting a phenotype. The link of causality with the position of SNPs can be established.

Discussion

The main drawback is that the results of the SNP exon method are far from those obtained by the commonly used method (based on χ^2 statistics). The differences are difficult to explain, except for the anecdotal coincidence of smoothed curves. Because their logics are far from each other. The χ^2 statistics method ignores the effect of positions of SNPs inside genes.

The SNP exon method has some avantages. The Z-scores or p-values of genes are computed from Z-scores or p-values of SNPs by a simple function. All parameters are computed from available data being rid of outlier values. No hyper-parameter is necessary. Here the method is a way to create a bridge between a global approach in order to be nearer biological mechanisms.

The criticizable choice of the SNP exon method is the use of Stouffer Z-score method where used weights are the probability density function γ. the use of linkage disequilibrium correlation matrix complicates and slows down the calculation. The weighting is the simplest method to use these observations. It probably needs to be deepened in a more rigorous way.

But, trying to get closer to biological reality, this method ignores some important phenomenons. The indirect mechanisms of regulation by regulatory DNA are neglected. SNPs in these areas probably disturb the expression of genes. Here the method is based on the linear distances in the genome as 23 segments and so it obscures the spatial storage of folded DNA. It goes part of the way to bring closer the statistical approach and the biological mechanisms.

The χ^2 method and SNP exon method are rather complementary. Facing their both results must enlarge searching the cause of phenotype. In addition, the focus on extreme values in Z-score of SNPs in function of minimal distances from SNPs to exons allows to identify a suspected set of SNPs. These analyses completed by positioning in pathways must enlighten the biological mechanism involved in the studied phenotype.

Tools

The Python scripts for data analysis and gene p-values with the manual are available at
https://github.com/drovera/SNP_exon (except scripts about comparison and simulation)

References

[1] Swetlana Berger, Martin Schlather, Gustavo de los Campos, Steffen Weigend, Rudolf Preisinger, Malena Erbe, and Henner Simianer. A scale-corrected comparison of linkage disequilibrium levels between genic and non-genic regions. *PLOS one*, October 2015.

[2] Susan M. Berget. Exon recognition in vertabrate splicing. *The Journal of Biological Chemistry*, February 1995.

[3] Zhongxue Chen and Saralees Nadarajah. On the optimally weighted z-test for combining probabilities from independent study. *ScienceDirect*, 70, February 2014.

[4] The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, January 2023.

[5] Christiaan A. de Leeuw, Joris M. Mooij, Tom Heskes, and Danielle Posthuma. Magma: Generalized gene-set analysis of gwas data. *PLoS Computational Biology*, April 2015.

[6] Steven A. Frank. The common patterns of nature. *J Evol Biol*, August 2009.

[7] Norio Gouda, Yuh Shiwa, Motohiro Akashi, Hirofumi Yoshikawa, Ken Kasahara, and Mitsuru Furusawa. Distribution of human single-nucleotide polymorphisms is approximated by the power law and represents a fractal structure. *Genes to Cells*, March 2016.

[8] Jianfei Huang, Kai Wang, Peng Wei, Xiangtao Liu, Xiaoming Liu, Kai Tan, Eric Boerwinkle, James B. Potash, , and Shizhong Han. Flags: A flexible and adaptive association test for gene sets using summary statistics. *Genetics*, 202, March 2016.

[9] Jeremy Hull, Susana Campino, Kate Rowlands, Man-Suen Chan, Richard R. Copley, Martin S. Taylor, Kirk Rockett, Gareth Elvidge, Brendan Keating, Julian Knight, and Dominic Kwiatkowski. Identification of common genetic variation that modulates alternative splicing. *PLoS Genetics*, June 2007.

[10] Chang-Yong Lee. A model for the clustered distribution of snps in the human genome. *arXiv*, May 2016.

[11] Chang-Yong Lee. A model for the clustered distribution of snps in the human genome. *Cornell University arXiv*, May 2016.

[12] Younghee Lee, Seonggyun Han, Dongwook Kim, Emrin Horgousluoglu, Shannon L. Risacher, Andrew J Saykin, and Kwangsik Nho. Genetic variation affecting exon skipping contributes to brain structural atrophy in alzheimer's disease. *AMIA joint Summits on Tanslational Science*, May 2018.

[13] Yang I. Li, Bryce van de Geijn, Anil Raj, David A. Knowles, Allegra A. Petti, David Golan, Yoav Gilad, , and Jonathan K. Pritchard. Rna splicing is a primary link between genetic variation and disease. *Science*, April 2016.

[14] Mitchell J. Machiela and Stephen J. Chanock. Ldlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, 31, July 2015.

[15] Aniket Mishra and Stuart Macgregor. Vegas2: Software for more flexible gene-based testing. *Cambridge University Press*, February 2015.

[16] Eliseos J. Mucaki, Ben C. Shirley, and Peter K. Rogan. Expression changes confirm genomic variants predicted to result in allele-specific, alternative mrna splicing. *frontiers in Genetics*, March 2020.

[17] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. Plink: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, september 2007.

[18] Sagiv Shifman, Jane Kuypers, Mark Kokoris, Benjamin Yakir, and Ariel Darvasi. Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetics*, 2003.

[19] Sagiv Shifman, Jane Kuypers, Mark Kokoris, Benjamin Yakir, and Ariel Darvasi. Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetic*, 2003.

[20] web site. Cidr-gwas of breast cancer in the african diaspora. *www.omicsdi.org/dataset/dbgap/phs000383*.

[21] web site. en.wikipedia.org/wiki/gamma distribution.

[22] web site. The nhgri-ebi catalog of human genome-wide association studies. *www.ebi.ac.uk/gwas/summary-statistics*.

[23] web site. www.ncbi.nlm.nih.gov/genome/51. in column refseq click on chromosome. in box send to choose coding sequence, file, format fasta nucleotide.

[24] XianMing Wu and Laurence D. Hurst. Determinants of the usage of splice-associated cis-motifs predict the distribution of human pathogenic snps. *Mol. Biol. Evol.*, October 2015.

[25] Dmitri V. Zaykin. Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology*, August 2011.

Supplementary information
Toy Simulation of Linkage Disequilibrium

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.

Repartition of SNPs

the repartition of SNPs respects a γ (fig 9) 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.

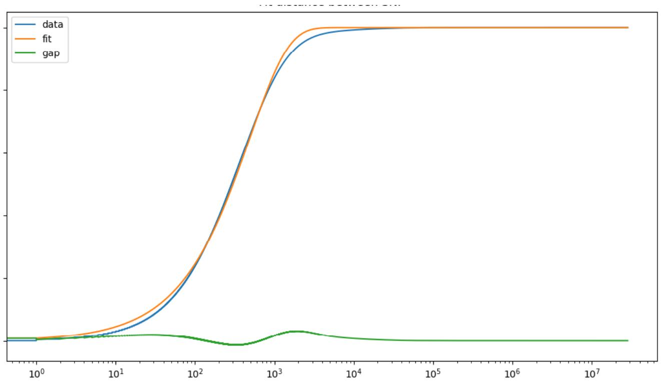

Figure 9: Graph of fitting distances between SNPs (GCST006719 from www.ebi.ac.uk/gwas/summary-statistics)

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 10. 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.

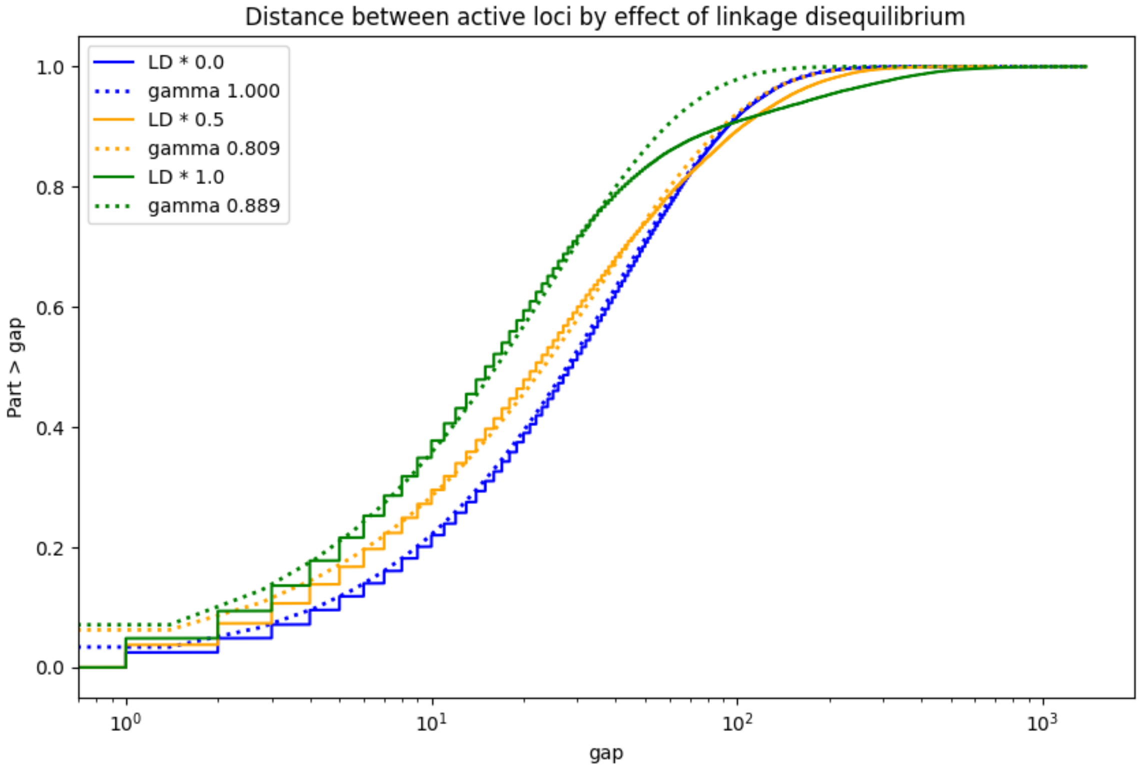

Figure 10: Cumul of distances between active loci for different values of linkage desiquilibrium and fit by γ distribution