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 - March 25, 2022

Abstract

Genome Wide Association Studies allow 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 disease.

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.

Here method is based on the statistical relation between positions of SNPs and positions of exons. It follows weights used by the Stoufffer's Z-score method. So, a function of distance from SNPs to exons gives the p-values transferred to genes. Moreover, an area of extreme values allows targeting a set of SNPs which probably plays a role in difference of genotype.

Keywords: GWAS, SNP, exon, intron, gene, splice

Aim and principle

Genome Wide Association Studies allow to analyse the link between frequency of single nucleotide polymorphism and phenotype 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 studied population and a reference population. Thier goal is to find biomarkers which are presumed to be involved in predisposition to disease.

The aim of study is to find a function giving p-values of genes from p-values of SNPs based on the consistency of genes and how 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 [11] or MAGMA [3] based on statistical comparisons using χ^2 statistic.

The elements of human genome on which are focused are the coding sequences (a tiny part less than 2%) 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 position of exons are extracted by chromosomes from ***ncbi.nlm.nih.gov***(acces by [19]). The computed number end 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 genes include one gene

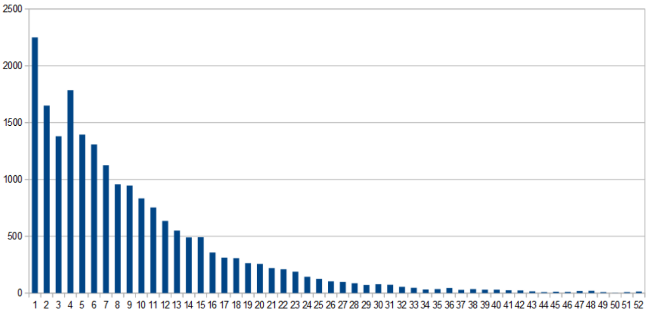

Figure 1: Number of exons by genes, until 52 exons

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 [5] and [7], but they do not approach the other components of the genome.

The relative position of SNPs in respect of exons implies 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 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 ([9] [12] and [10]). 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

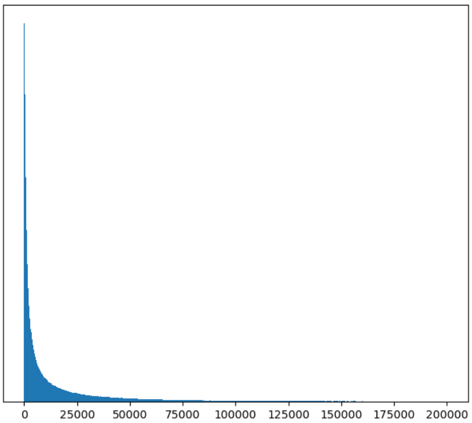

Figure 2: Histogram showing the numbers of SNPs by bins of minimal distances from SNPs to exons (GCST006719 from ebi [18])

The repartition of SNPs appears not uniform along the genome. It can be observed visually and reported in [8]. 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.

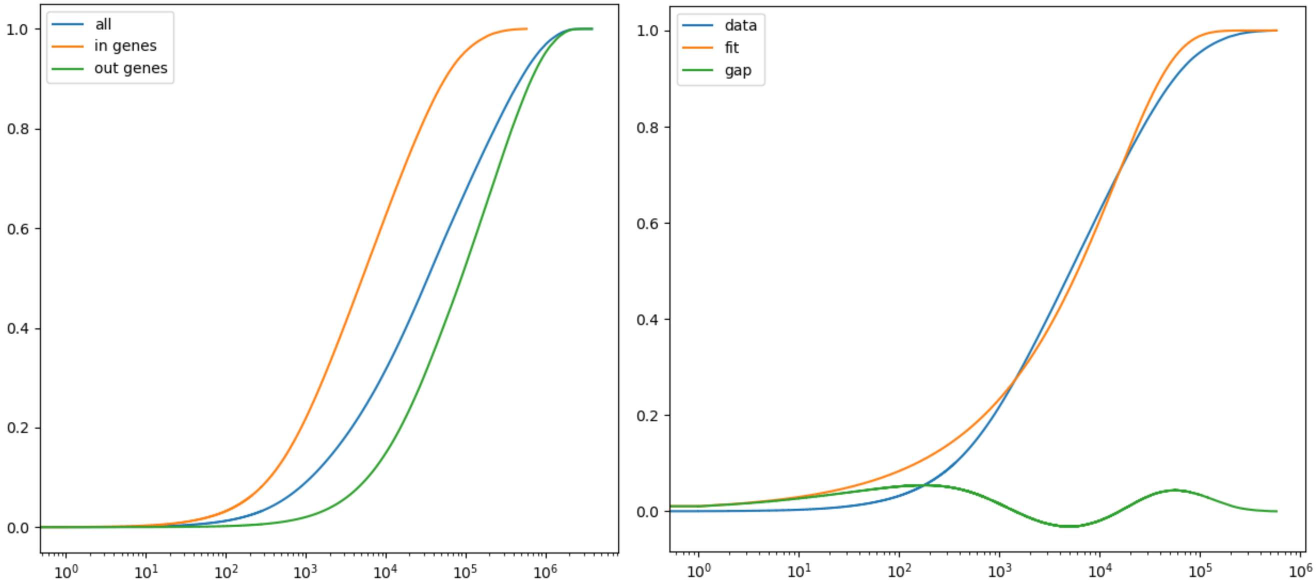

Figure 3: On the left, position of SNPs relative to exons (orange in genes, green out genes, blue all), normalized cumulative number in function of minimal distance of SNPs to exons, logarithmic scale on the abscissa.
On the right, fitting a normalized cumulative number in function of minimal distance of SNPs to exons, logarithmic scale on the abscissa.
(GCST006719 from ebi [18])

This distribution can be fitted by a γ distribution (annex About 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 is the cumulative density function (cdf) of γ law (fig **??**) .

This distribution can be interpreted as a heterogeneous spatial Poisson process. The intensity λ, average number of points by length, depends on the location. It is shown (annex Intensity of inhomogeneous Poisson process):

$$\lambda(d) = \frac{\gamma.pdf(d)}{\gamma.cdf(d)}.$$

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.

The fitting normalized cumulative number in function of minimal distance of SNPs to exons (fig 3) gives these values of parameters: shape = 0.44924 scale = 32738. This observation does no concern only this example as the above array show for different sources [18] [16] (only SNPs inside genes):

data src	in shape	in scale	in std	phenotype
CIDR_AFRO	0.44718	34050	0.00086	breast cancer
GCST004988	0.44924	32738	0.00089	breast cancer
GCST006719	0.45382	34094	0.00082	breast cancer
GCST007236	0.44931	33066	0.00088	breast cancer
GCST90011804	0.44956	32842	0.00089	breast cancer
GCST90011808	0.44960	32863	0.00089	prostate cancer
GCST90011811	0.44956	32839	0.00089	colon cancer

Shape is the critical parameter for the link between intensity and distance (confer γ distribution for different values of shape [17]). Its value is less than one and so the curve shows a hyperbolic form. It leads to notice accumulation of SNPs near exons as the histogram 2 shows.

This observation questions. The repartition of exons does not respect any well known statistical law. However, the repartition of SNPs respects a gamma law with a significant dispersion of values (shape = 0.75 - 0.82 and scale = 230 - 650). The statistical effect caused by a difference of random variables following a γ law must be excluded. The attraction of SNPs by exons is a phenomenon observed. This law about SNPs seems to be the result of linkage disequilibrium. The correlation between loci decreases with the distance between loci as measures of LD show [15] [1]. The consequence is that the variants contributing to an analog effect tend to agglomerate (annex Toy Simulation of Linkage Disequilibrium).

Following these observations: 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.

So, the method to evaluate the quantitative link between SNPs and genes is based on the following choices:

- 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 by the Stouffer's Z-score method
- The weights used in Z-score formula are the probability density function of gamma law

Method

It follows that using the probabilty density function as weights to transfer the p-value to genes seems natural. The resulting p-values of genes are computed by the Stouffer's Z-score method ([21]) (Z-score is the number of standard deviations between value and the mean value):

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

The Z-score of every gene is computed from the Z-score of SNP_i in the scope. 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.

The simple formula of Z-score does not contain the term $2 * \sum_{i<j} r_{ij} * w_i * w_j$. This formula is commonly used in meta analysis. 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. The SNPs contribute to the same phenotype, so the correlation coefficient is not null and must be evaluated.

Computing every r_{ij} is out of reach. Only one value is assigned to the correlation coefficient r_{ij}. It is computed by linear adjustment of log(p-values) and log(cumulative number) (fig 4), which is justified by the beta uniform mixture model (annex Correlation and distribution of p-values).

To evaluate the effect of SNPs inside exons, the weights are taken infinite, which cancels the effect of any SNPs outside exons (annex Formula of Z-score). 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 is lost.

In addition, the contributions of every SNP are detailed, every term of the Z-score formula. This detailed result shows how a set of SNPs can reinforce its effects. this is restricted to a sublist of genes due to the size of the output

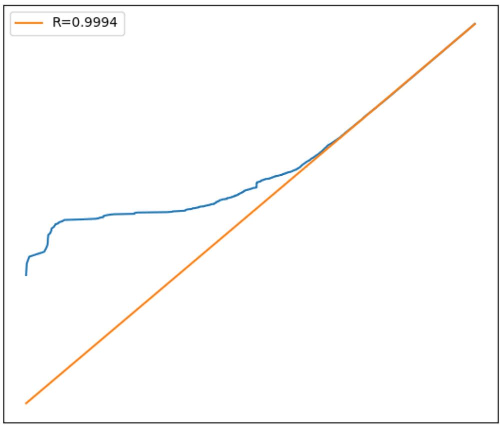

Figure 4: Linear log log fitting of normalized cumulative number of SNPs and their p-values (GCST006719 from ebi [18])

Extreme values of Z-score and minimal distances

The Z-score of SNPS in function of minimal distance of SNPs to exons (fig 5) shows a feature of extreme values: the higher the Z of SNP, the closer SNP is to the exon.

This interesting observation of extreme values provides a confirmation that the distance of SNPs to exons plays a role in the effect of SNPs.

This analysis of effect of SNPs on genes can be focused on this this subset of SNPs and be deepened by an accurate biological study. An incidental remark: the partial use of the weighting formula may give surprising results about genes because the evaluation of effect by not extreme SNPs can be negative (contribution by a set of SNPs > 100%).

Comparison to MAGMA

MAGMA [3] is based on χ^2 statistics, a commonly used method. The method of MAGMA is of the same type as those implemented in PLINK [14] and VEGAS [11]. 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 6. 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 and Kendall's τ coefficient

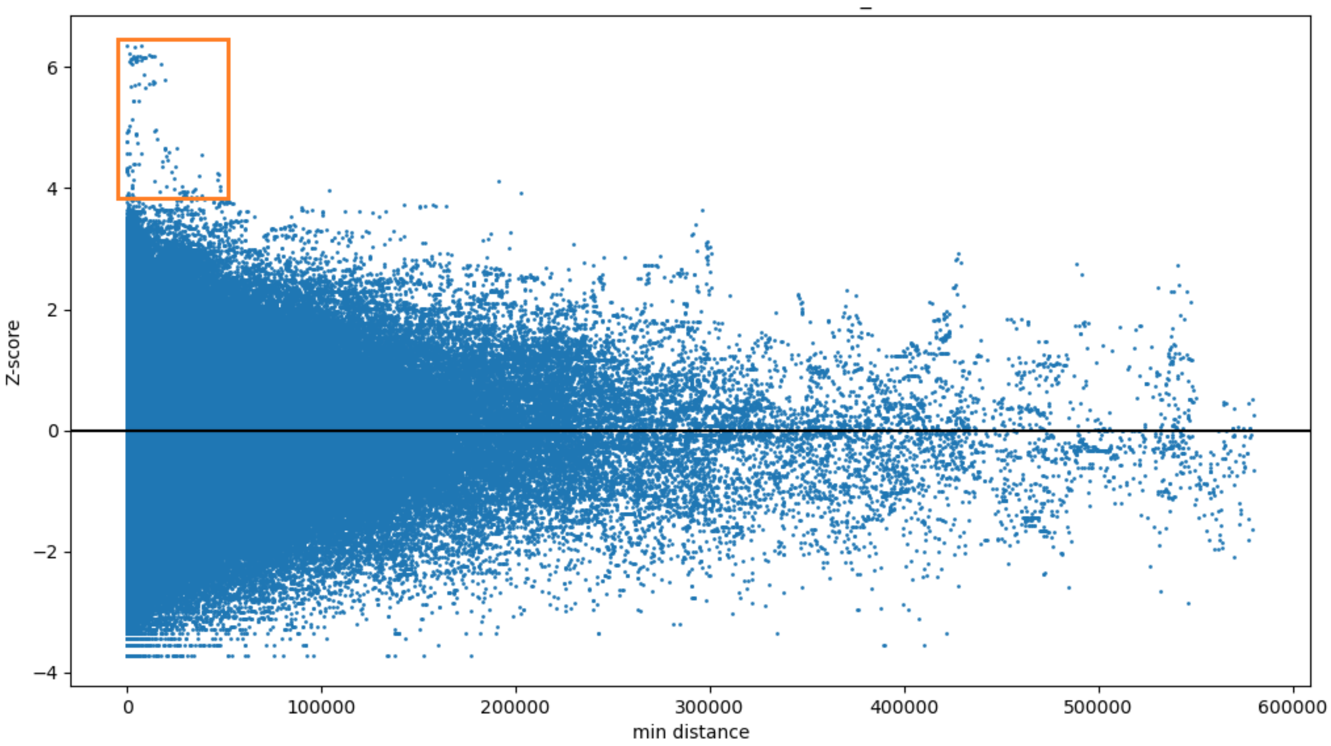

Figure 5: Z-score of SNPs in function of their minimal distances to exons (GCST006719 from ebi [18])

is equal to 0.70.

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 7, 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 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 ?

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 [9]. Likewise several publications show the effect of SNPs on the mechanism of splicing as [20] and [6]. They 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.

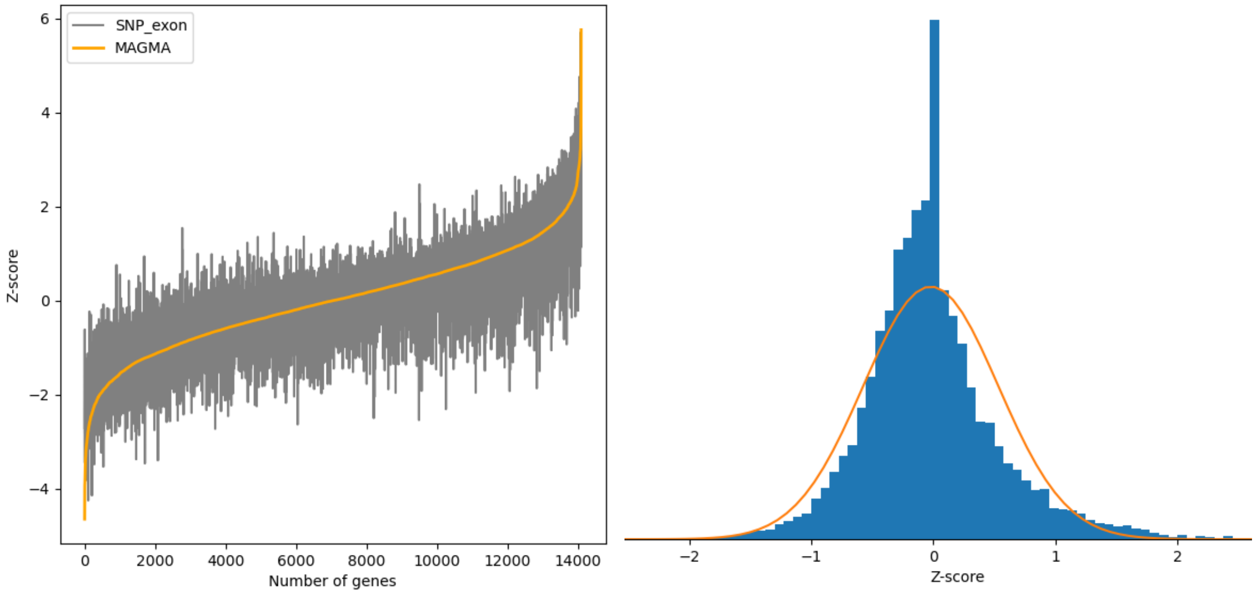

Figure 6: Cumulative Z-score of genes and histogram show the difference of outcome between MAGMA and SNP exon method, MAGMA is reference (GCST006719 from ebi [18])

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

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 trom 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 continuous 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 a nearer biological mechanism.

The main critique of the SNP exon method is about the use of Stouffer Z-score method where used weights are the probability density function γ. It is a too simple mode to use these observations. It 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 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 supposedly set of SNPs. These analyses completed by positioning in pathways must enlighten the biological mechanism involved in the studied phenotype.

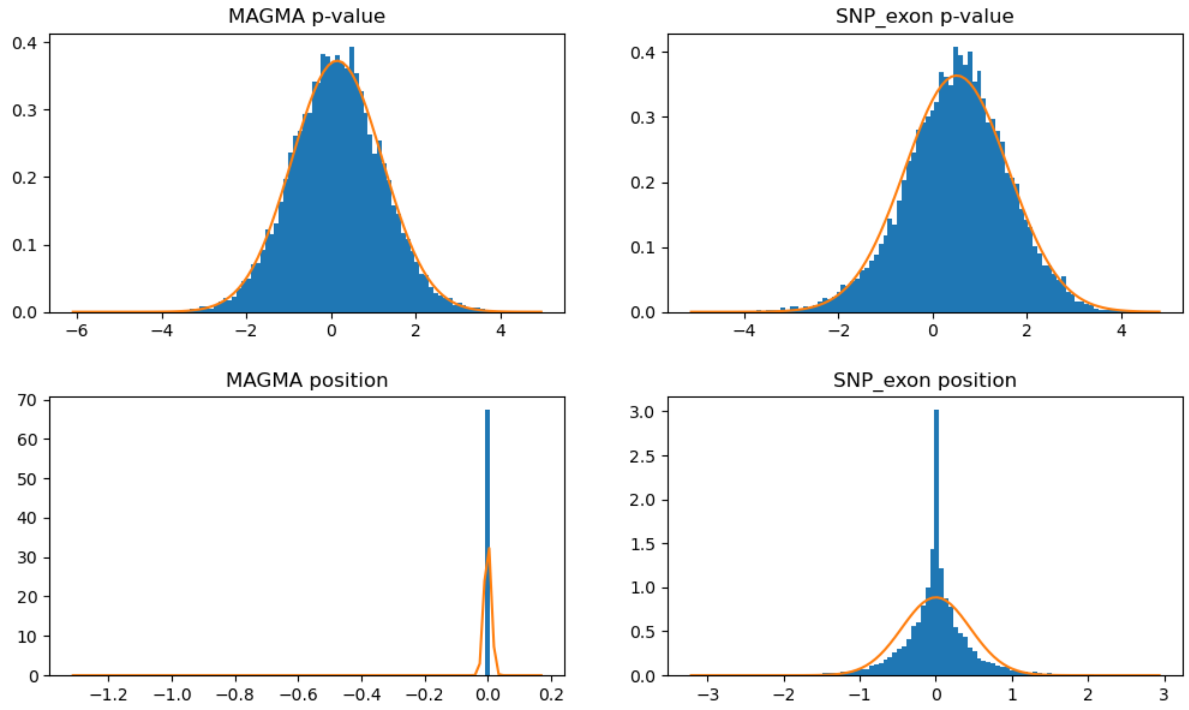

Figure 7: On the left Z-scores of genes got by MAGMA with same p-value and homogeneous position compared to result from real data as histogram, on the right same comparison for SNP exon method(GCST006719 from ebi [18])

Annex

Outliers p-values

Python cannot compute the Z-score below $5.55 * 10^{-17}$. So the p-value between $6.7 * 10^{-133}$ and $3.7 * 10^{-17}$ can be taken into account. They are considered as outlier values. Other outlier values are also eliminated in tails to enter in calculable p-values by Python. This is the case of GCST006719 and GCST007236 [18].

Intensity of inhomogeneous Poisson process

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:

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

This expression with h → 0 is brought closer to the law from observation (N number of SNPs):

$$\frac{dn(t)}{dt} = \frac{1}{N} * \gamma.pdf(t)$$

Considering t as the minimal distances from SNPs to exons, the expression of λ is:

$$\lambda(t) = \frac{1}{N} * \gamma.pdf(t)$$

This function is used only as weight, the term $\frac{1}{N}$ can be removed without changing the result.

Correlation and distribution of p-values

The distribution of p-values can be described by a $\beta(a, 1)$ law to which the noise is added (beta uniform mixture model [13]). The expression of β is:

$$\beta.pdf(x) = a * x^{a-1} \ and \ \beta.cdf(x) = x^a$$

The formula with noise λ is:

$$bum.pdf(x) = \lambda + (1 - \lambda) * a * x^{a-1}$$

The noise ruins the correlation between p-values. So, the linear fitting log-log on cumulative normalized values gives the correlation coefficient between p-values.

Formula of Z-score

The formula of Z-score method for n Z-score to be weighted with R as the common correlation coefficient:

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

When the weights are infinity, it becomes, :

$$Z = \frac{\sum_i w_i * z_i}{\sqrt{n + R * n * (n \ - \ 1)}}$$

About γ distribution

Finding a γ distribution is not surprising, it is common in nature [4]. γ distribution is the maximum entropy distribution for a positive variable, given its mean value and the mean value of its logarithm.

A Q-Q plot precisely shows how the minimal distance from SNPs inside genes to exons follows a γ distribution with shape = 0.44924 scale = 32738 (fig 8).

Shape is the critical parameter for the relation between effect and distance:

- if shape = 1, γ is exponential
- if shape < 1, its form is hyperbolic
- if shape > 1, γ increase after that decrease

Here the shape is far less than 1, so SNPs tend to agglomerate preferably near exons.

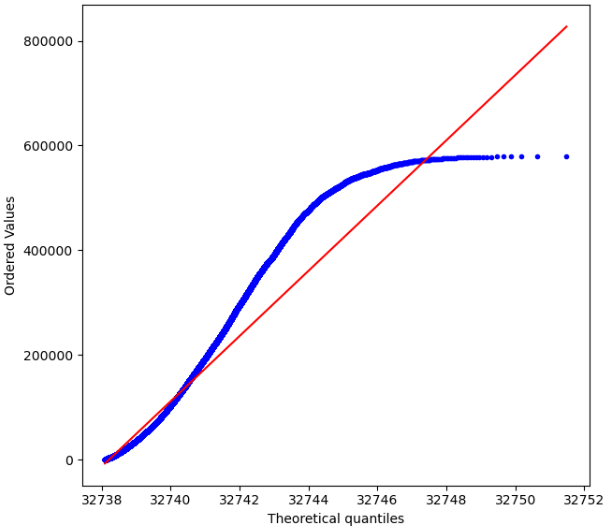

Figure 8: Q-Q plot minimal distance from SNPs inside genes to exon - γ distribution (GCST006719 from ebi [18])

Toy Simulation of Linkage Disequilibrium

This intuition is reinforced by a toy simulation. Loci are located randomly along a segment. State of the next locus is drawn successively according to a conditional probability from the formula of linkage disequilibrium.

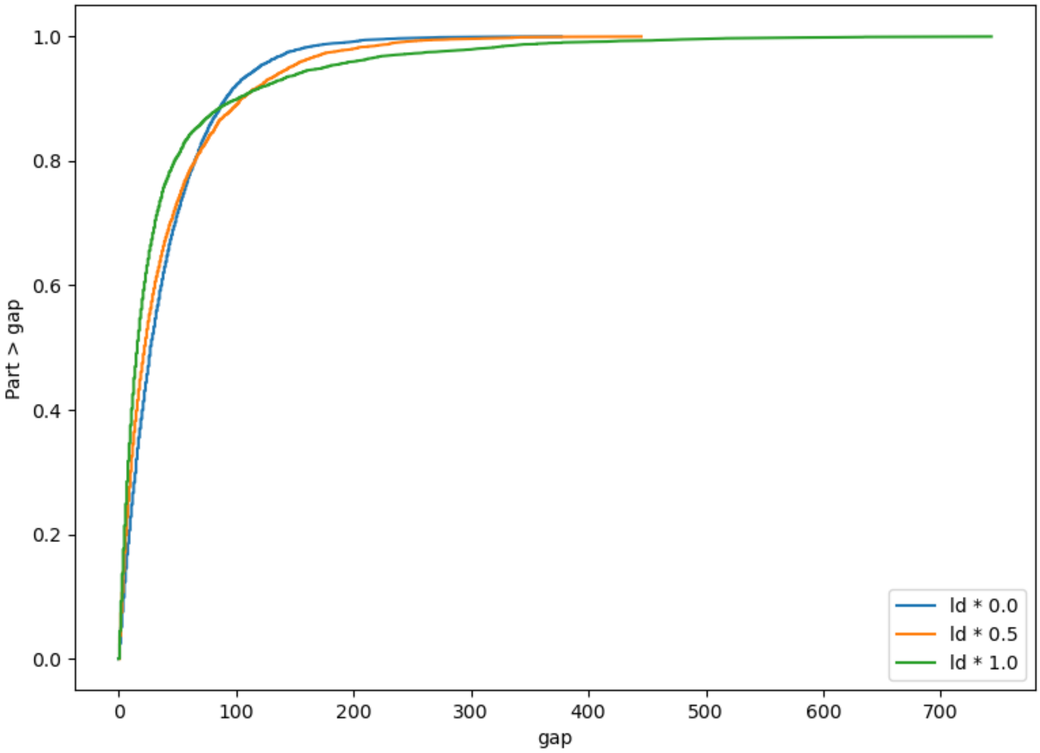

Figure 9: Toy simulation, cumul of distances between active loci for different values of linkage desiquilibrium

According to the usual notation: $D = P_{AB} - P_A * P_B$, the conditional probabilities are

$$P(A1/B1) = \frac{P_A * P_B + D}{P_A} \text{ and 7 analog formulas.}$$

The ceiling of D is the minimum of $(1 - P_A) * P_B$ and $P_A * (1 - P_B)$.

The maximum value of D is approximated by a decreasing exponential function of the distance between loci. Their correlation decreases if distances between loci increases [15] [1]. D is multiplied by a factor between 0 and 1 to evaluate the effect of LD. P_A and P_B are arbitrarily taken to 0.5. They stay approximately to this value during 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 9. The maximal distance between them tends to increase too. For ld equal to 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. A simulation is not a proof.

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 which look like trials, if interested, contact author)

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] 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.

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

[5] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

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

[12] 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.

[13] Stan Pounds and Stephan W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *BIOINFORMATICS*, January 2003.

[14] 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.

[15] Sagiv Shifman, Jane Kuypers, Mark Kokoris, Benjamin Yakir, and Ariel Darvasi. Linkage disequilibrium patterns of the human genome across populations sagiv shifman 1,2 , jane kuypers 3 , mark kokoris 3 , benjam. *Human Molecular Genetics*, 2003.

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

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

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

[19] 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.

[20] 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.

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