Association of adaptive phenotypic traits to causal variants using the extended genotype homozygosity matrix.

May 20, 2025

Abstract

TO DO ...

Introduction

The detection of the effect of adaptive phenotypic traits is fundamental for understanding the number of affected variants and the strength of selection for these adaptive traits. The detection of the association between an adaptive phenotype and the genotype can be difficult, depending on the effect of each associated variant and the effect of the adaptation over the positions around it. A number of studies, which the aim of understanding the patterns of frequencies for adaptive traits (*e.g.*, EYRE-WALKER, 2010; CABALLERO *et al.*, 2015; LOURENÇO *et al.*, 2011; CONNALLON and CLARK, 2015), shown that the genome architecture for a trait depends on the (mathematical) relationship of the trait versus fitness (that is, if the fitness is directly correlated with the trait or, alternatively, fitness and traits are correlated distributions). Simons et al. (2018) avoided this problem by inferring the relationship between trait and fitness from empirical data and estimated the genetic architecture

of the trait, assuming pleiotropy and stabilizing selection. Also, BEISSINGER *et al.* (2018) developed a test to detect the effect of selection on a quantitative complex trait using the whole information on effect sizes.URICCHIO *et al.* (2019) developed a new method to detect the signals of polygenic selection using the derived allele effect size from GWAS analysis and the derived allele frequency of the associated variants. All of these methods can be applied to observe the effects of environmental changes on the genome for specific traits related to adaptation. Recently ZENG *et al.* (2019) developed a bayesian algorithm (SBayesS) that requires GWAS summary statistics and functional genomic annotation to detect the effect of selection on complex polygenic traits.

Here, we are aimed to (i) discover if a given trait is under a recent adaptive process and (ii) to detect strong and medium adaptive effects with higher precision, based on the association of the trait versus phenotype. We propose to use the signals of linkage disequilibrium produced by causal adaptive variants to increase the precision in locating associate variants as well as distinguishing between adaptive versus non-adaptive variants. The reasoning is the following: the causal variants of an adaptive trait will left signals of adaptation (*e.g.,*, increase of homozygosity around the causal variants, more linkage disequilibrium, reduction of variation) around them. If a trait is under adaptive pattern, we expect that the associated variants exhibit signals of adaptation, while other non-adaptive traits will not have such signals on associated variants. Additionally, the signals of adaptation may increase the precision to locate the causal variants, together with the genotype data.

We propose to study the association between the phenotype and the genotype considering additionally (as additional genotype information) a matrix containing the extended genotype homozygosity. This matrix would be used in the analysis of association to detect the effect of adaptive selection of the studied phenotype. Our hypothesis is that this approach can have promising results to detect the causal mutations in panmictic populations that have experimented stronger selective events. A consistent simulation study is required in this project to evaluate different sce-

narios.

The resolution of this kind of statistics is expected to be in the range of genetic breeding improvement up to domestication process (approximately from some dozens to hundreds of generations (less than N_e generations, (SABETI *et al.*, 2002). This range may allow to detect critical traits and the effect of these traits on the genome under crucial events such as the domestication. In this work, the statistics used are calculated at genome level , which are analyzed with the phenotypes of individuals.

Methods

Considering a single population and polarized alleles (ancestral and derived)

Following SABETI *et al.* (2002), we define the statistic EHH_d as the proportion of homozygote individuals of the derived allele from position i (the target position) to position j, in relation to the number of homozygote individuals with derived variants at the position i (note that derived variants are obtained by comparison to an/several outgroup species). That is:

$$EHH_{d_{ij}} = \frac{\sum_{k=1}^{n} I_{d_{k,ij}}}{\sum_{l_d=1}^{n} I_{d_{l,i}}}, \tag{1}$$

where n is the number of individual samples, and $I_{d_{k,ij}}$ counts 1 if the individual k is homozygote for derived variant at position i and is still homozygote from position i to j, otherwise counts 0. In the same way, $I_{d_{l,i}}$ counts 1 if the individual l is homozygote for derived at position i, otherwise counts 0. The statistic EHH_a is calculated in the same way than EHH_d but considering ancestral variant instead of derived variant. Note that this value can not be calculated in case there are not derived (ancestral) homozygous individuals.

In order to quantify the effect of the neighbouring homozygosity at a given studied position (i), all values for EHH_{ij} are summed from the position i to their right and and to their left sides considering their distance (physical or recombinant),

and until reaching a threshold arbitrary value of EHH_{ij} (let's say 0.1). That is:

$$iES_{d_i} = \sum_{j=x+1}^{y} \frac{(EHH_{d_{ij-1}} + EHH_{d_{ij}})}{2}(Pos_j - Pos_{j-1}), \tag{2}$$

x and y are the positions (at the right and at the left) where $EHH_{d_{ij}}$ becomes below the threshold or it is too far (by the presence of large gaps) from the central position (so the area out of x and y is considered unimportant), and Pos is the physical or the recombinant position (from a linkage map). The same is for the ancestral statistic (iES_{a_i}).

The integrated relative Extended homozygosity statistic is then calculated as:

$$iRES_{da_i} = \frac{iES_{d_i}}{iES_{a_i}}. \tag{3}$$

A value higher than 1 (if the genotype extension statistic for derived variants is larger than for ancestral variants) suggest the effect of positive selection favouring a derived variant at position i.

The homozigosity at each position per individual

In case of doing a GWAS analysis, usually each position and genotype is evaluated independently in relation to each individual phenotype. That means that it is crucial to distinguish the genetic information between individuals at this position. Here we propose to include the information concerning to the homogeneity of the individual at neighbouring regions from the focus position (that is, a way to consider the linkage disequilibrium between the two chromosome copies of an individual at this region) in addition to the genotype at the focus position. That is, it is calculated a statistic related to the neighbouring homogeneity for every position and individual. This statistic is obtained by slicing the $iRES_{da_i}$ statistic given the contribution of each individual (k) to the total:

$$iES_{dk_i} = \sum_{j=x+1}^{y} (EHH_{d_{k,ij-1}} + EHH_{d_{k,ij}})(Pos_j - Pos_{j-1})/2, \text{ and}$$
$$iRES_{da_{k,i}} = \frac{iES_{d,k_i}}{iES_a}, \tag{4}$$

where x and y are the positions (at the right and at the left boundaries for derived) and $EHH_{d_{k,ij}} = I_{d_{k,ij}}/(\sum_{l_d=1}^{n} I_{d_{l,i}})$.

Quantifying homogeneity with no outgroup and unphased data

To quantify the homogeneity with no outgroup and unphased data, it is convenient to use the framework developed by TANG *et al.* (2007) to study the degree of homogenization of a given position (using unphased genotype information) as is explained below.

Define candidate positions

In a first step, it is useful to define those candidate positions in relation to their maximum local extension of their homozygosity. This step is optional, as other criteria for defining candidate positions can be used. Following (TANG *et al.*, 2007), we define $EHHS_{ij}$ as the proportion of homozygote individuals from position i (the position of interest) to position j, in relation to the number of homozygote individuals at the position i. That is:

$$EHHS_{ij} = \frac{\sum_{k=1}^{n} I_{k,ij}}{\sum_{l=1}^{n} I_{l,i}}, \tag{5}$$

where n is the number of individual samples, and $I_{k,ij}$ counts 1 if the individual k is homozygote from position i to j (*i.e.*, using the genotype nomenclature, all variants have the values 0 or 2 at this region), otherwise counts 0. In the same way, $I_{l,i}$ counts 1 if the individual l is homozygote at position i, otherwise counts 0. This statistic is calculated from position i to any position (left or right) until

this proportion becomes enough small to be considered negligible. This threshold value, although is somewhat arbitrary, it has been considered 0.1 in the original work (TANG *et al.*, 2007) and here it is used the same criteria. The $EHHS_{ij}$ values for the position i are kept and used to calculate the next statistic iES_i (TANG *et al.*, 2007).

The following calculation of the iES_i statistic pretends to quantify the effect of the neighbouring homozygosity at a given studied position. Having all values for $EHHS_{ij}$, we count the total area of homozigosity around the position i, that is, having the position i as the center, we sum, at their right and and their left, all the contributions of $EHHS_{ij}$, considering their distance (physical or recombinant). that is:

$$iES_i = \sum_{j=x+1}^{y} \frac{(EHHS_{ij-1} + EHHS_{ij})}{2}(Pos_j - Pos_{j-1}), \qquad (6)$$

where x and y are the positions (at the right and at the left) where $EHHS_{ij}$ becomes bellow the threshold or it is too far (by the presence of large gaps) from the central position (so the area out of x and y is considered unimportant), where Pos may be the physical or the recombinant position (from a linkage map).

The homogeneity at each position per individual using unphased and no outgroup data

As explained above, for GWAS analyses it is convenient to have information of the homogeneity per individual (to be contrasted with phenotype data). The desired statistic is obtained by dividing the iES_i statistic given the contribution of each individual to the total:

$$iES_{k,i} = \frac{1}{2\sum_{l=1}^{n} I_{l,i}} \sum_{j=a+1}^{b} (Pos_j - Pos_{j-1})(I_{k,ij-1} + I_{k,ij}), \qquad (7)$$

where $\sum_{k=1}^{k=n} iES_{k,i} = iES_i$ (considering the same threshold values -a and b- for the each of the samples, lke iES_i statistic). Here the only differential term between

individuals (in relative terms) is the last sum.

Using a reference population: The quotient between the extension of homozygosity in target individuals from a population versus a reference population

Following the same reasoning, it is possible to estimate the effect of the extension of homozygosity per position and per individual in relation to the effect in a reference population (note that a reference population is not an outgroup species). This is useful in case we are considering the effect of selection in the target population while we assume no selection in the reference population. Those position that have high iES_i at both populations would be considered nuisance given by other factors, like genomic effect caused by the genetic architecture of the genome. Therefore, following (TANG *et al.*, 2007), we define the statistic derived from Rsb_i:

$$Rsb_{k,i} = \frac{iES_{k,i}}{iES_i^{popRef}} \tag{8}$$

where $\sum_{k=0}^{k=n} Rsb_{k,i} = Rsb_i$.

The study of association phenotype-genotype and the information related to the homogeneity of individuals in neighbouring regions

A multivariate regression model has been used to detect the association of a position with the phenotype, considering only, the genotype information, only the extended genotype homozygosity information or both. The probability of association is calculated with the linear model function in R (lm), considering the vector of phenotypes (Y), obtained by summing up (additive model) all the effects per variant position, considering their dominance parameter, in relation to the genotype matrix and/or the iES_k matrix (with a dimension of SNPs x individuals). All these vectors and matrix were normalised previously. The r^2 regression value was obtained for each SNP, and the p-value was estimated assuming a F distribution. False discovery rate and Bonferroni methods (BENJAMINI and HOCHBERG, 1995)

that correct for multiple testing were applied to raw P-values.

To evaluate the capacity that we have to detect real beneficial mutations affecting the studied phenotype we measured the specificity (True Negative Rate) and sensitivity (True Positive Rate or statistical power) for the used matrices (the iES_k, the genotype counts and both together).

Simulations

Forward simulations of a population have been performed with Slim v4 (HALLER and MESSER, 2019). A population with 5000 diploid individuals is simulated with a mutation rate of 6e-8 and a recombination rate of 1e-7. A fragment of one million positions is simulated. Mutations type "m2" can be located in segments of 1Kb in regions separated 100Kb positions. After 40K generations, the population split in two populations with same indiviuals (i.e., $N_e = 5000$) and no migration between them. All mutations start to be neutral, but m2 type mutations start to have an additive effect over the fitness after 50,000 generations. Additionally, once simulations are finished, a non-adaptive phenotype is constructed from additive effects of random mutations across the genome. This additive effect can be dependent of frequency, such as effect depending on (i) low heterozygosity, on (ii) high heterozygosity and also on (iii) high number of low frequency variants within the population). The model considers heritability effect on phenotype of h2=0.75 on both adaptive and non-adaptive phenotypes. Populations are evolved until the optimal phenotype value in the population is achieved (an arbitrary small standard deviation of 0.1 close to the optimal) or until arriving 60K generations.

Once simulations are finished, sequences with all variants (in ms format posteriorly translated to genotypes) and phenotypes (in a tabulated format) were kept into files to run Rsbki code, which includes the statistics here developed. The statistics per individual (i.e., iESk, iRESk, Rsbk per individual and position) and the genotype data were used to perform GWA analysis versus individual phenotypes. We expect to identify (some) QTL regions (that is, not all QTLs contain

mutations in enough frequency and/or with enough effect to be identified) using standard GWA analysis with genotype data for phenotypes under adaptive effect but also with non-adaptive effect. Nevertheless, it is expected that in case using EHH statistics, only QTLs with phenotypes under adaptation would be identified. This is because the differential signal would only be observed in positions with adaptation signals.

Results

Simulation of data under different scenarios and Validation

Figure 1 shows an example of the GWA study using genotype versus EHH statistics per individual. For phenotypes associated to adaptive processes affecting the fitness, if the effect is quite strong to affect the linkage patterns and the frequencies of the surrounding positions, a signal in the EHH statistics will be present, and will be also present in individuals suffering this effect of adaptation. On the other hand, phenotypes not affected by relative strong selection will not present this pattern in their sequences. GWA studies using genotype sequence should be agnostic to this pattern because each variant is independently analyzed. Therefore, GWA studies using genotype will be able to detect adaptive and non-adaptive association. This is observed in the example of Figure 1A and C. Instead, GWA studies using EHH statistics instead genotype are able to differentiate between adaptive and non-adaptive phenotypes (Figure 1B and D). Statistics based on comparative patterns between populations (Rsbk) seem to be more robust to false positive signals than statistics based on a single populations (results to show here). iRESk seem to be less discriminant than the other two statistics (results to show here). Preliminar simulations before performing a statistical power analyses suggest that this expectation is true. Nevertheless, when the strength of the adaptive process affects regions that are also linked to regions affecting non-adaptive phenotypes, then the signal will be also observed, which may confound the interpretation of the

results.

Include here the study of statistical power, sensibility... using different parameters.

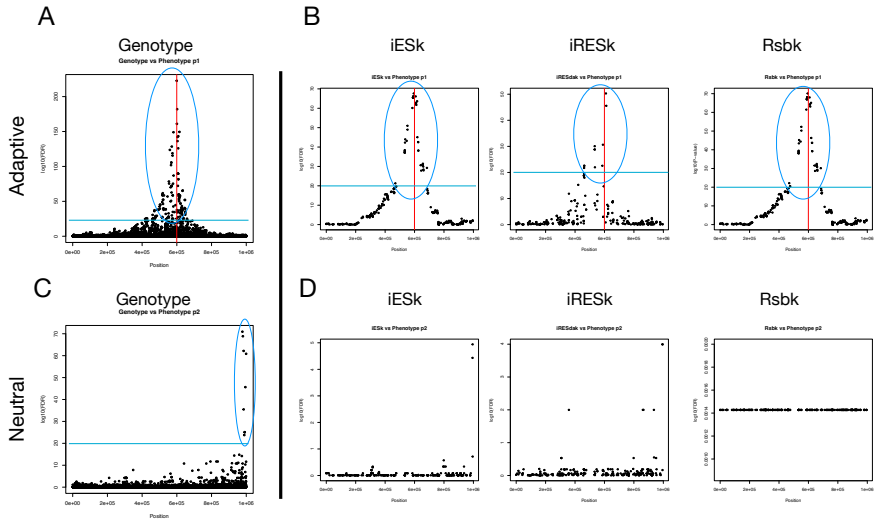

Figure 1: GWA analysis for genotype (A, C) and for EHH statistics (B, D). Two phenotypes are analyzed, one with adaptive effect (A, B) and one with neutral effect (C, D). The horizontal blue line and the blue circle sorrounding points indicate the positions with significant associations. Red line indicate the regions with mutations having additive effect on fitness at frequency higher than 0.05.

Real Data Analysis

Discussion

We provide a methodology to discriminate in one step the phenotypes that are affected by adaptive processes versus neutral. Advantages and disadvantages. Nevertheless, when the strength of the adaptive process affects regions that are also linked to regions affecting non-adaptive phenotypes, then the signal will be also observed, which may confound the interpretation of the results. Statistical power of each statistic. Code to run Develop an R function?

References

BEISSINGER, T., J. KRUPPA, D. CAVERO, N.-T. HA, M. ERBE, *et al.*, 2018 A simple test identifies selection on complex traits. Genetics **209**: 321–333.

BENJAMINI, Y., and Y. HOCHBERG, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B : 289–300.

BOYKO, A. R., S. H. WILLIAMSON, A. R. INDAP, J. D. DEGENHARDT, R. D. HERNANDEZ, *et al.*, 2008 Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet **4**: e1000083.

CABALLERO, A., A. TENESA, and P. D. KEIGHTLEY, 2015 The nature of genetic variation for complex traits revealed by gwas and regional heritability mapping analyses. Genetics **201**: 1601–13.

CONNALLON, T., and A. G. CLARK, 2015 The distribution of fitness effects in an uncertain world. Evolution **69**: 1610–1618.

EYRE-WALKER, A., 2010 Evolution in health and medicine sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc Natl Acad Sci U S A **107 Suppl 1**: 1752–6.

HALLER, B. C., and P. W. MESSER, 2019 Slim 3: Forward genetic simulations beyond the wright-fisher model. Mol Biol Evol **36**: 632–637.

LOURENÇO, J., N. GALTIER, and S. GLÉMIN, 2011 Complexity, pleiotropy, and the fitness effect of mutations. Evolution **65**: 1559–71.

SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. P. LEVINE, D. J. RICHTER, *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419**: 832–7.

TANG, K., K. R. THORNTON, and M. STONEKING, 2007 A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol **5**: e171.

URICCHIO, L. H., H. C. KITANO, A. GUSEV, and N. A. ZAITLEN, 2019 An evolutionary compass for detecting signals of polygenic selection and mutational bias. Evol Lett **3**: 69–79.

WANG, J., A. CABALLERO, P. D. KEIGHTLEY, and W. G. HILL, 1998 Bottleneck effect on genetic variance. a theoretical investigation of the role of dominance. Genetics **150**: 435–47.

ZENG, J., A. XUE, L. JIANG, L. R. LLOYD-JONES, Y. WU, *et al.*, 2019 Bayesian analysis of gwas summary data reveals differential signatures of natural selection across human complex traits and functional genomic categories. bioRxiv .