Detection of the effect of selected phenotypic traits using additional statistics in GWAS.

July 15, 2019

We propose to use the quantity of linkage disequilibrium associated to a given position as a ponderation for each of the individuals at each given studied position. This statistic will be used in GWAS analysis of association to detect the effect of adaptive selection of the studied phenotype. If we consider a single population that have suffered a strong selective event, those causal positions that strongly determined the selective effect will increase rapidly their frequency, and thus will have larger linkage disequilibrium with the closest sequences than other regions not affected by selective events. With this idea it is pretended to take advantage of the stronger footprint of a sweep associated to a selective event to give more weight to those individuals in whose the studied position contains a determinant allele, as well as longer linkage in close regions caused by the effect of the selective sweep. Our hypothesis is that this approach can have promising results to detect the causal mutations in panmictic populations that have experimented stronger selective events. On the other hand, we think that can not give additional information in GWAS studies designed to distinguish between divergent populations (or control-disease populations). A consistent simulation study is required in this project to evaluate different scenarios.

The GWAS analysis with a single panmictic population is based on the association analysis of a single variant position for each of the individuals (usually

genotype information, that is, a variable with the discrete values 0, 1 and 2, indicating the number of copies of a variant - reference or alternative -), considering the relationship between individuals and other variables, versus the observed phenotype at each of the individuals. Thus, it is crucial to define all the genetic features at the studied position for each individual. One of this genetic features is the degree of homozygosity/heterozygosity in neighbouring regions to this position at each of the individuals. The presence of causal variants close to a given position and the observation of homogeneity of the individual at this region will, presumably, be associated with the phenotype. The larger homogeneity on the focus position with their neighbouring region may indicate a causal effect.

The resolution of this k,ind of statistics is expected to be in the range of the domestication process (not too short, approximately from several hundreds of generations to thousands of generations, REFERENCE). This range may allow to detect critical traits and the effect of these traits on the genome under crucial events such as the domestication.

Methods

Here, we will use the framework developed by TANG *et al.* (2007) to study the degree of homogenization of a given position (using unphased genotype information).

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:

$$EHHS_{ij} = \frac{\sum_{k=1}^{n} I_{k,ij}}{\sum_{l=1}^{n} I_{l,i}}, \quad (1)$$

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:

$$iES_i = \sum_{j=a+1}^{b} \frac{(EHHS_{ij-1} + EHHS_{ij})}{2}(Pos_j - Pos_{j-1}), \quad (2)$$

where a and b 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 a and b is considered unimportant), where Pos may be the physical or the recombinant position (from a linkage map).

Quantifying homogeneity at each position and individual

In case of doing a GWAS analysis, usually each position is evaluated independently in relation to the 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 new statistic related to the neighbouring homogeneity for every position and individual. This statistic is obtained by dividing the iES_i statistic given the contribution of each individual to the total:

$$iES_{k,i} = \frac{1}{2\sum_{l=1}^{n} I_{l,i}} \sum_{j=a+1}^{b} (Pos_j - Pos_{j-1})(I_{k,ij-1} + I_{k,ij}), \qquad (3)$$

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. That is, the objective here is the difference between individuals and not the absolute value. Therefore:

$$iES'_{k,i} = (\frac{1}{2\sum_{l=1}^{n} I_{l,i}} \sum_{j=a+1}^{b} (Pos_j - Pos_{j-1})(I_{k,ij-1} + I_{k,ij}))/iES_i, \qquad (4)$$

and $\sum_{k=1}^{k=n} iES'_{k,i} = 1$.

Considering two populations: The quotient between the extension of homozigosity 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 homozigosity per position and per individual in relation to the effect in a reference population. 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 nui-

sance 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:

$$Rsb_{k,i} = \frac{iES_{k,i}}{iES_i^{popRef}} \tag{5}$$

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

To perform a study of association genotype-phenotype, we can include the information provided by the statistic $iES_{k,i}$ and $Rsb_{k,i}$ in two different ways: (a) by adding the matrix iES (or $Rsbi$) as a new random factor or (b) using this matrix as a ponderation of the genotype matrix. The two approaches have different meanings and thus different interpretations. In the first case, it is assumed that the homogeneity along the contiguous regions of a position in a given individual has the same importance than the genotype of this individual, which suggest that regions close to the causal position (not necessarily causal) would be associated. In the second case, the method would stress only those genotypes with an associated phenotype that additionally would have strong ponderation (large homogeneity in the surrounding region).

A mixed model is used for relating the phenotype of each individual (a vector y of phenotypes) with the genetic features of the individual at each position (that is, a vector g containing the genotypes and the homogeneity versus surrounding positions). For each SNP we relate the phenotype using this general expression:

$$y = Xb + Zg + e, \tag{6}$$

where X is a matrix that relates the observed with the parameters for fixed effects defined in the b vector and Z is a matrix that relates the observed data with

the parameters related to the genotype data. e is the residual error. DEFINE ASSUMPTIONS, DISTRIBUTIONS AND SCENARIOS.

Results

Simulation of data under different scenarios and Validation

-A population suffering an strong selective sweep on a trait of interest.

-A population NOT suffering an strong selective sweep on a trait of interest.

-Different scenarios with violations of the assumptions of a panmictic population or with violation of the effect of selection.

Real Data Analysis

-Real data from humans?

-Real data from bovine?

Discussion

References

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.