RESEARCH

Methods to compute the composite log-likelihood (CLL) of allelic frequencies for the detection of signatures of selection in diploid genomes

Filippo Biscarini*, Nelson Nazzicari and Alessandra Stella

*Correspondence:
filippo.biscarini@tecnoparco.org
Department of Bioinformatics,
PTP, Via Einstein - Loc. Cascina
Codazza, Lodi, Italy
Full list of author information is
available at the end of the article

Abstract

First part title: Text for this section.

Second part title: Text for this section.

Keywords: composite log-likelihood; signatures of selection; diploid genomes

Background

Selection, both natural and artificial, is one of the major forces that can shape the genome of living organisms and change allele frequencies. A mutation that is beneficial for the adaptation of an organism to its environment, or that is of agricultural or industrial interest, tends to increase in frequency in the population, together with neighbouring genomic regions which tend to be hitch-hiked in the process ([1]). Through the last decade, the on-going genomic revolution has been making available hundreds of thousands of genetic markers for several animal, microbial and plant species. By looking at the allele frequency at marker loci along the genome of populations experiencing different selective pressures, it is possible to identify genomic regions -and ultimately genes- involved in processes such as domestication, adaptation, evolution and artificial selection. There are a number of methods based on allele frequencies that have been developed to detect such signatures of selection. A popular method is Wright's F_{ST} ([2, 3]) that has been applied to studies in humans ([4]), plants ([5]) and animals ([6]). Alternatively, the likelihood of the difference between allele frequencies in different populations can be estimated and used to detect the presence of signatures of selection ([7, 8]). However, there are several ways in which such likelihoods can be computed, and these might differ in computation requirements and statistical properties, such as the sensitivity to detect signals of selection or the behaviour along the margins of the dimensional space. (add examples?) It may therefore be of interest to investigate the statistical properties of different estimators for the likelihood of the difference between allele frequency, and assess how well they are capable of detected actual signals of selection.

In this study we evaluated 4 different ways to estimate the likelihood of the difference in allele frequency between populations. The logarithm of the likelihoods thus calculated were then computed and combined across sliding windows along the genome (CLL, composite log-likelihood) in order to detect signatures of selection. The proposed CLL approaches were compared among them and with approaches based on F_{ST} and on simple squared distances between genotypes. All methods were first analysed numerically along the entire dimensional space (frequencies in analysed populations ranging from 0 to 1) and then tested with real data where strong signals of selection are known to be present. The methods hereby presented have been developed having in mind biallelic genetic markers (namely SNPs), but can in principle be adapted to any kind of markers (discuss this in "Results and discussion"?).

Methods

Let's assume we have a test population T (the population on which we want to detect signatures of selection) and a null or reference population N (the contrasting population(s)). On these populations F_{ST}, the squared difference of the allelic frequencies, and four different likelihood measures of allele frequency differences have been computed and evaluated in terms of their ability to detect signatures of selection. First, we present numerical details of the different calculations, then examples with real data on which the different estimators were tested.

Detecting signatures of selection

F_{ST}

The F_{ST} (Wright's F statistics for *Subpopulation* vs *Total*), a.k.a. *fixation index*, is a measure of the genetic differentiation between (sub)populations (e.g. the test and reference populations in our case). For any given locus, F_{ST} was calculated as described by Nei ([3]); partitioning the total expected heterozygosity (H_T, or gene diversity) into the *interpopulation* diversity (D_{ST}) and the *intrapopulation* diversity ($\overline{H}_S$, average of the expected heterozygosity -H_S- in the subpopulations weighted by their size), the F_{ST} is defined as the proportion of gene diversity due to differentiation among subpopulations:

$$F_{st} = \frac{D_{st}}{H_T} = \frac{H_T - \overline{H}_S}{H_T} = 1 - \frac{\overline{H}_S}{H_T} \tag{1}$$

where $\overline{H}_S = \sum_{i=1}^{k} w_i H_{S_i}$ is the weighted average of the expected heterozygosities within k subpopulations each accounting for a proportion w_i of the total population size ($H_{S_i} = 1 - \sum_{j=1}^{m} p_j^2$ is the expected heterozygosity within subpopulation i -with m alleles at the given locus); $H_T = 1 - \sum_{i=1}^{m} \overline{p}_i^2$ is the expected heterozygosity in the overall population, with $\overline{p_i} = \sum_{i=1}^{k} w_k p_k$, the average allele frequency at the given locus across k subpopulations each accounting for a proportion w_i of the total sample size.

F_{ST} measures the differentiation between the null and reference populations, and was used in this study as primary benchmark against which other methods were evaluated.

Squared difference (D^2)

A very simple and naive approach for the estimation of the genetic distance between populations is the squared difference of their allele frequencies. For any given locus, the allele frequencies of the first allele in the test and null populations (p_1 and

p_2) were calculated, and their squared difference used to measure the degree of differentiation between the two populations:

$$D^2 = (p_1 - p_2)^2 \tag{2}$$

While F_{ST} was chosen as benchamrk due to its long-standing theoretical development and its wide-spread use in the detection of population substructure, D^2 has been used as additional naive benchmark against which more sophisticated methods can be compared.

Binomial log-likelihood (BLL)
Allele frequencies at a given locus (A/a) between the test and null populations are compared. Let p_N be the frequency of allele A in the null population, and n_T and k_T the total number of alleles and the number of A alleles in the test population respectively. Under the null hypothesis that $p_T = p_N$, the allele count in the test population can be thought of as a random sample from the same reference population. The following binomial function measures the likelihood that such a sample actually comes from the reference population:

$$LL_{ij} = \binom{n}{k}(1-q)^k q^{n-k} \tag{3}$$

where LL_{ij} is the probability to observe the test allele distribution at locus j on chromosome i in the null population. The smaller LL_{ij} the less likely it is that the two populations are one and the same, and the more justifiable it is to accept the alternative hypothesis that the test and null populations are actually different (maybe because of some genetic processs such as natural or artificial selection).

Unless the test and null populations have the same size and their allele frequencies are complementary ($p_N = p_T$), the binomial log-likelihood is not symmetric. Therefore, three cases can be envisaged: 1) comparison of N vs T (*BLL1*); 2) comparison of T vs N (*BLL2*); 3) comparison of T vs N+T (*BLL3*). The latter is what was implemented in [8], and works well with multiple comparisons, when one population has to be compared against many other populations: it has the advantage of creating fewer numerical problems (lack of symmetry, falling outside of the computational space etc ...), but has the shortcoming of shrinking the power of the comparison (since $T \subset N$ -the test population is included in the null population). Additionally, it is often the case that two populations have to be compared against each other. For the reasons above, all three scenarios (*BLL1, BLL2, BLL3*) were evaluated in the present study.

Multiplicative log-likelihood (MLL)
To overcome the issue of lack of symmetry with BLL, a possibility is to combine the likelihood of BLL1 and BLL2 (T vs N and N vs T). Being probabilities, a natural approach is to multiplicate them: $MLL = LL_{TvsN} * LL_{NvsT}$

Hypergeometric log-likelihood (HGLL)
Distribution of the differences (DIFF)
Two cases: parametrical test, resampling method

"Ballons d'essai"

Data from a population of xxx Holstein-Frisian (zzz males and xxx females) and yyy Piedmontese (males and females?) cattle were available. The first breed has been long selected for dairy production, the latter for meat production. All animals were genotyped with the Bovine 54k SNP-chip. Genotypes were edited for individual and marker call-rate (> 95% ?) and for MAF (> 0.05 ?). Remaining missing genotypes were imputed using (check this!). Xxx SNPs were eventually available for analysis. We selected the zzz SNPs on chromosome 3 (BTA-3) as working example to test the different estimators of the CLL of the allele frequency difference between populations and the reference methods (F_{ST} and Euclidean distances) for the detection of signatures of selection. BTA-3 is known to host, within the gene *SLC35A3* (position: 43400328-43445390 bps, approximately halfay along the chromosome), the point mutation responsible for CVM (Complex Vertebral Malformation) in Holstein-Frisian cattle ([9]). CVM is a recessive inherited disorder that frequently causes abortion or perinatal death in Holstein-Frisian calves. Other cattle breeds -including Piedmontese- are not affected by the condition. A strong signal is therefore expected to be found at this site, making this an ideally suited comparison to test different methods for the detection of signatures of selection.

¿ Also DGAT1, Myostatin, Caseins ?

Working example

Text for this sub-sub-heading . . .

Sub-sub-sub heading for section sasad

Sliding windows

$$CLL_{ij} = \frac{1}{w} \sum_{j=1}^{j+w-1} LL_{ij} \tag{4}$$

Testing significance

When analysing tens or hundreds of thousands of markers the problem of multiple comparisons is incurred. This is a known statistical problem in general ([10]), and specifically in the field of genetic studies ([11, 12]). As the number of independent tests increases, the probability of obtaining at least one false positive, under H_0, approaches one. This probability is a function of the number of tests n and of the chosen significance threshold α: $P(false\_positive) = 1 - (1 - \alpha)^n$. When, for instance, 50000 SNP markers are tested along the genome in search of signatures of selection, for $\alpha = 0.01$, 500 false positive associations are expected on averege, just by mere chance.

- Bonferroni correction: very strict; implicit prior assumption that the probability that *all* tests are null (H_0 is true) is not small. If we believe that all tests could be null, then aiming to make the number of false positives zero is justifiable ([13])
- FDR
- Permutation test

Software

Code in C/C++

Results and discussion

For biallelic loci, D^2 is intrinsically symmetric: $(p_1 - p_2)^2 = ((1 - p_1) - (1 - p_2))^2$

Numerical illustration

Application to real data

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Text for this section . . .

Acknowledgements

Data (Marras ...)? . . .

References

1. Braverman, J.M., Hudson, R.R., Kaplan, N.L., Langley, C.H., Stephan, W.: The hitchhiking effect on the site frequency spectrum of dna polymorphisms. Genetics **140**(2), 783–796 (1995)
2. Wright, S.: The genetical structure of populations. Annals of eugenics **15**(1), 323–354 (1949)
3. Nei, M.: F-statistics and analysis of gene diversity in subdivided populations. Annals of human genetics **41**(2), 225–233 (1977)
4. Akey, J.M., Zhang, G., Zhang, K., Jin, L., Shriver, M.D.: Interrogating a high-density snp map for signatures of natural selection. Genome research **12**(12), 1805–1814 (2002)
5. Zhao, K., Wright, M., Kimball, J., Eizenga, G., McClung, A., Kovach, M., Tyagi, W., Ali, M.L., Tung, C.-W., Reynolds, A., *et al.*: Genomic diversity and introgression in o. sativa reveal the impact of domestication and breeding on the rice genome. PLoS One **5**(5), 10780 (2010)
6. Kijas, J.W., Lenstra, J.A., Hayes, B., Boitard, S., Neto, L.R.P., San Cristobal, M., Servin, B., McCulloch, R., Whan, V., Gietzen, K., *et al.*: Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS biology **10**(2), 1001258 (2012)
7. Nielsen, R., Williamson, S., Kim, Y., Hubisz, M.J., Clark, A.G., Bustamante, C.: Genomic scans for selective sweeps using snp data. Genome research **15**(11), 1566–1575 (2005)
8. Stella, A., Ajmone-Marsan, P., Lazzari, B., Boettcher, P.: Identification of selection signatures in cattle breeds selected for dairy production. Genetics **185**(4), 1451–1461 (2010)
9. Thomsen, B., Horn, P., Panitz, F., Bendixen, E., Petersen, A.H., Holm, L.-E., Nielsen, V.H., Agerholm, J.S., Arnbjerg, J., Bendixen, C.: A missense mutation in the bovine slc35a3 gene, encoding a udp-n-acetylglucosamine transporter, causes complex vertebral malformation. Genome research **16**(1), 97–105 (2006)
10. Berry, D.A.: The difficult and ubiquitous problems of multiplicities. Pharmaceutical Statistics **6**(3), 155–160 (2007)
11. Lander, E.S., Schork, N.J., et al.: Genetic dissection of complex traits. SCIENCE-NEW YORK THEN WASHINGTON-, 2037–2037 (1994)
12. Risch, N., Merikangas, K., *et al.*: The future of genetic studies of complex human diseases. Science-AAAS-Weekly Paper Edition **273**(5281), 1516–1517 (1996)
13. Wakefield, J.: Reporting and interpretation in genome-wide association studies. International journal of epidemiology **37**(3), 641–653 (2008)

Figures

Figure 1 Sample figure title. A short description of the figure content should go here.

Figure 2 Sample figure title. Figure legend text.

Tables

Additional Files

Additional file 1 — Sample additional file title

Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title

Additional file descriptions text.

Table 1 Sample table title. This is where the description of the table should go.

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.