

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 6 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 (Euclidean?) 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 results of the study are hereby presented and discussed.

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 six 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.

Numerical illustration

F_{ST}

The F_{ST} (Wright's F statistics for *Subpopulation* vs *Total*), a.k.a. *fixation index*, is a measure of the genetic substructure in a joint population (or of the genetic differentiation between (sub)populations). 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 (\bar{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:

$$F_{st} = \frac{D_{st}}{H_T} = \frac{H_T - \bar{H}_S}{H_T} = 1 - \frac{\bar{H}_S}{H_T} \quad (1)$$

$$H_T = D_{st} + \bar{H}_S \quad (2)$$

\bar{H}_S is the average expected heterozygosity within subpopulations. Let H_S be the expected heterozygosity within a single subpopulation (with L number of loci and m number of alleles within locus):

$$H_S = 1 - \frac{1}{L} \sum_{l=1}^L \sum_{i=1}^m p_i^2 \quad (3)$$

The average expected heterozygosity within the k subpopulations is then:

$$\bar{H}_S = \frac{1}{k} \sum_{s=1}^k H_S \quad (4)$$

We shall now consider the expected heterozygosity in the overall population; let the average allele frequency at locus i across k subpopulations be the weighted average of individual allele frequencies: $\bar{p}_i = \frac{1}{\sum_{s=1}^k n_s} \sum_{s=1}^k n_s p_{is}$, with n_s the subpopulation size, or, using proportions, $\bar{p}_i = \sum_{i=1}^k w_k p_k$, with w_k the proportion of

subpopulation k on the total. The expected heterozygosity in the total population is then:

$$H_T = 1 - \frac{1}{L} \sum_{l=1}^L \sum_{i=1}^m \bar{p}_{il}^2 \quad (5)$$

F_{st} measures the deviation of heterozygosity from Hardy-Weinberg equilibrium in the total population which is due to the differentiation among subpopulations. The comparison between the heterozygosity expected under Hardy-Weinberg equilibrium and the heterozygosity observed in the population gives indications with respect to phenomena such as inbreeding, introgression, and admixture.

Squared difference (D^2)

Binomial log-likelihood (BLL)

Three cases: A vs B, B vs A and A vs A+B

Multiplicative log-likelihood (MLL)

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

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 : $P(\text{false_positive}) = 1 - (1 - \alpha)^n$. 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 average, 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

Results and discussion

Text ...

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)

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

Figure 2 Sample figure title. Figure legend 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

Figures

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.