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 are dragged along ("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. 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 detecting 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 by simulating scenarios along the entire dimensional space (frequency range and population size ratios between the test and null populations) 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 and applied to any diploid organism.

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 ($\overline{H}_S$), 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:

$$BLL_{ij} = ln\left\{\binom{n_T}{k_T}(1-p_N)^{k_T}p_N^{n_T-k_T}\right\} \tag{3}$$

where BLL_{ij} is the probability to observe the test allele distribution at locus j on chromosome i in the null population. The smaller BLL_{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 (for instance 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 the lack of symmetry with BLL, a possibility is to combine the likelihood of BLL1 and BLL2 (T vs N and N vs T). A natural approach to combining probabilities is through multiplication. Therefore the multiplicative log-likelihood was defined as the geometric mean of the two binomial log-likelihoods:

$$MLL_{ij} = \sqrt{BLL_{ij(TvsN)} * BLL_{ij(NvsT)}} \tag{4}$$

where MLL_{ij} is the multiplicative log-likelihood at locus j on chromosome i and $BLL_{ij(TvsN)}$ and $BLL_{ij(NvsT)}$ are the two binomial log-likelihoods (BLL1 and BLL2).

Hypergeometric log-likelihood (HGLL)
An alternative approach is to look at the test population as a random sample from the reference population and calculate the probability of obtaining the observed allele counts. Under the null hypothesis that $p_N = p_T$, the two populations can be combined into one and such probability would follow a hypergeometric distribution:

$$HGLL_{ij} = \frac{\binom{n_{A_N}+n_{A_T}}{n_{A_T}} \cdot \binom{(N_N+N_T)-(n_{A_N}+n_{A_T})}{N_T-n_{A_T}}}{\binom{N_N+N_T}{N_T}} \quad (5)$$

where n_{A_N}, n_{A_T}, N_N and N_T are the number of A alleles and the total number of alleles in the null and test population respectively. HGLL is symmetric with respect to the population chosen as test or reference ([9]), and assumes that sampling is without replacement, which may be closer to the allele sampling process involved in the biological differentiation of populations.

Distribution of the differences (DIFF)
Instead of comparing the allele frequency observed in the test population against that of the null population, one could look directly at the difference between the two allele frequencies. First, the absolute difference between allele frequencies is calculated: $d = |p_1 - p_2|$. In order to obtain a likelihood value for d, both an analytical (*DIFF1*) and numerical (*DIFF2*) approach were adopted. In *DIFF1*, d was standardized ($z = \frac{d}{\sqrt{\hat{p}(1-\hat{p})(1/N_N+1/N_T)}}$) and compared against a normal distribution. In *DIFF2*, genotypes were randomly permuted between the null and test population, and the difference between the allele frequencies calculated each time. A distribution of values of d under H_0 ($d = 0$) was thus produced and its empirical probability distribution function used to obtain a likelihood value for the allele frequency difference between the original populations, *d*. Different numbers of permutations were tested in terms of results obtained and computation time.

Simulated scenarios

For all methods for the detection of signatures of selection, different scenarios were simulated, in order to evaluate the statistical behaviour and computational requirements for chainging values of the parameters used. Specifically, the methods were evaluated against different values of allele frequency and different sample sizes in the null and test populations. The whole range of allele frequency (from 0 to 1) was simulated, and three different population size ratios were considered: balanced design (equally sized null and test population), mildly unbalanced design (1:2 population size ratio), and extremely unbalanced design (1:10 population size ratio).

"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 ([10]). 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 ?

Sliding windows

When analysing actual data, F_{ST} and D^2 values, and the likelihoods obtained with all the methods presented, were combined in sliding windows based on the base-pair distance (bps) between SNPs, in order to reduce the influence of spurious signals. A fixed sliding window of 200 kbps (check this! Maybe better 500 kb: see [11]) was used, and composite log-likelihoods (CLL) were thus obtained:

$$CLL_{i\bar{j}} = \frac{1}{w} \sum_{j=1}^{j+w-1} LL_{ij} \tag{6}$$

where $CLL_{i\bar{j}}$ is the composite log-likelihood at position $\bar{j}$ (midpoint of the sliding window) on chromosome i, and LL_{ij} are the log-likelihood values calculated for all the SNPs between SNP j and 200 kbps ahead of it.

Software

Code in C/C++ (check with Nelson!) was written to implement all of the statistical methods to detect signatures of selection presented above. The R programming environment and Octave ([12]) for statistical analysis were used to prototype some of the tests and produce graphical visualization of the results.

Results and discussion

All methods were evaluated in terms of their statistical properties (sensitivity to detect signatures of selection, symmetry around population subdivision, robustness to unbalanced data, behaviour at the margin of the dimensional space) and computation requirements.

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

BLL1 and BLL2 go to $+\infty$ at the boundaries of the dimensional space (when the null and test populations are fixed for alternate alleles at the given locus). This leads to numerical problems and also to visualization problems when values of different

methods are plotted together for comparison (BLL1 and BLL2 are shrunken towards the bottom of the plot area). This can also partly explain why BLL1 (or BLL2) and MLL appear to follow the same pattern: since $MLL = \sqrt{BLL1 * BLL2}$, when one or both of the BLL go to infinity, the MLL is dragged along with it.

Is Fst obsolete?

Use haplotypes (e.g. EHH)?

Numerical illustration

Given its behaviour at the boundaries of the dimensional space (no numerical problems for p_1 or p_2 equalling either 0 or 1), F_{ST} is powerful to detect fixation (same allele or alternate alleles) ([13]). Other methods may be more appropriate to detect ongoing selection, balancing selection, loci or haplotypes at moderate frequencies.

Boundedness

BLL1, *BLL2* and the derived method *MLL* have no upper bound: they go to $+\infty$ when alternative alleles are fixed in the test and null populations

Computation time (for permutations only)

[probably this will not be a paragraph bat will go together with the discussion on permutations]

Cmmputation times for increasing number of permutations were investigated. The dimensional space of $d = p_1 - p_2$ was explored for permutations going from 1000 to 50 000, with steps of 1000, for a population of 50 diploid individuals (25 each in the test and null population) and 1 locus; at each step, the whole range of frequency comparisons between the test and null populations was tested. The relationship between n. of iterations and computation time is clearly linear, as expected for a first-order problem. From these results, the computation time required for larger numbers of permutations can be extrapolated, and the computation time for a single frequency comparison and a given n. of permutations derived. For instance, on a *2.0GHz AMD Opteron*, if we wished to explore the entire dimensional space with 100 000 permutationas at each step (as it's been done in the numerical illustrations for the paper), it would take approximately 9 hours and 20 minutes. Or, if one wished to make a one-locus comparison between a test and a null experimental populations (which is usually the case when looking for signatures of selection) it would take less than 6 minutes with 200 000 permutations and a sample size of 50. Increasing the sample size would linearly increase the computation time. With 500 individuals (250 each in the test and null population if balanced, any other combination thereof otherwise) the empirical analysis of frequency difference with 100 000 permutations would take approximately $\frac{1}{2}$ hour. The same linear relationship holds for increasing number of loci. With the same 500 individuals and 100 000 permutations, if the comparison would involve 1000 loci the analysis would take 500 hours (20 days!). Parallelization could help reduce the computation time needed (with 4 cores the last example would go down to 5 days).

A different permutation strategy (check with Nelson and describe) might be more efficient and computation times.

Application to real data

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 ([14]), and specifically in the field of genetic studies ([15, 16]). 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.

A classical approach to controlling the number of false positives is to apply the Bonferroni correction ([17]) and reject H_0 when $p - value \leq \alpha/m$ (i.e. family-wise error rate, FWER), where α is the chosen significance threshold and m is the number of tests performed. The Bonferroni correction aims to make the number of false positives as close as possible to zero, and is justified when there is an implicit prior assumption that the probability that all tests are null is not small ([18, 19]), but tends to be overly conservative in many practical applications. For example in our case ... Can the Bonferroni correction be applied to Fst? And D^2? The same holds for FDR.

- 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 ([19])
- FDR
- Permutation test

Conclusions

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section . . .

Acknowledgements
Data (Marras ...)? . . .
The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n. 276699 ("NEUTRADAPT")
"NEXTGEN" (how?)

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. Jantosciak, J., Barnier, W.: Duality and symmetry in the hypergeometric distribution. Mathematics magazine **75**(2), 135–143 (2002)
10. 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)
11. Qanbari, S., Gianola, D., Hayes, B., Schenkel, F., Miller, S., Moore, S., Thaller, G., Simianer, H.: Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. BMC genomics **12**(1), 318 (2011)
12. Eaton, J., Bateman, D., Hauberg, S.: Gnu octave manual (network theory limited, bristol, uk) (2002)
13. Biswas, S., Akey, J.M.: Genomic insights into positive selection. TRENDS in Genetics **22**(8), 437–446 (2006)
14. Berry, D.A.: The difficult and ubiquitous problems of multiplicities. Pharmaceutical Statistics **6**(3), 155–160 (2007)
15. Lander, E.S., Schork, N.J., et al.: Genetic dissection of complex traits. SCIENCE-NEW YORK THEN WASHINGTON-, 2037–2037 (1994)
16. 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)
17. Hochberg, Y.: A sharper bonferroni procedure for multiple tests of significance. Biometrika **75**(4), 800–802 (1988)
18. Westfall, P.H., Johnson, W.O., Utts, J.M.: A bayesian perspective on the bonferroni adjustment. Biometrika **84**(2), 419–427 (1997)
19. 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.

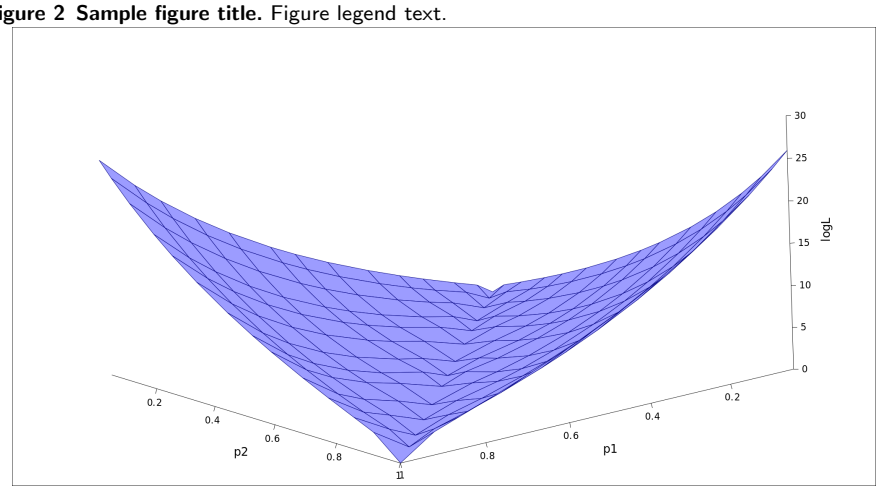

Figure 3 The l-o-n-g caption for all the subfigures (FirstFigure through FourthFigure) goes here.

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	..	.	.