FSTest: an efficient tool for estimation and visualization of cross-population fixation index on variant call format files

Siavash Salek Ardestani[1], Seyed Milad Vahedi[2*]

[1]Department of Animal Science, Science and Research Branch, Islamic Azad University,Tehran 1477893855, Iran.

[2]Department of Animal Science and Aquaculture, Dalhousie University, Truro, NS, B2N 5E3, Canada

[*]**Corresponding author:** smvahedi@dal.ca

Summary

Fixation index (Fst) statistics provide critical insights into the evolutionary processes affecting the structure of genetic variation within and among populations. The Fst statistics have been widely applied in population and evolutionary genetics to identify regions of the genome targeted by selection pressures. The FSTest software was developed to estimate and visualize four Fst statistics of Hudson, Weir and Cockerham, Nei, and Wright using high-throughput genotyping or sequencing data. This package provides Z-transformation of results and can estimate FST using a sliding window approach in which the window size is determined by the number of markers. Therefore, overestimation and underestimation of FST values are minimized. Publication-quality visualization of the results using Manhattan plots was also included in this software. FSTest is freely available at https://github.com/similab/FSTest.

Keywords: fixation index, signatures of selection, genetic variation

1. Introduction

Fixation index (Fst) includes a group of statistics that quantify levels of genetic differentiation between populations based on allele frequencies (Biswas and Akey, 2006). Moreover, they can be used to identify loci at which a beneficial allele has recently gone to fixation (Holsinger and Weir, 2009). The Fst has been conceptually defined in many ways (Hudson et al., 1992; Wright, 1949), among which the Weir and Cockerham (1984) and Nei (1973) estimations are the most popular statistics. Two widely used software, including VCFtools (Danecek et al., 2011) and PLINK (Purcell et al., 2007) have provided the functions to compute the Weir and Cockerham and Wright Fst, respectively. However, the Hudson et al.'s (1992) is preferable statistic among other Fst estimators since it shows lower sensitivity to sample sizes and does not systematically overestimate (Bhatia et al., 2013). Additionally, except for VCFtools, a limited number of tools are currently available to produce Fst statistics from variant call format (VCF) data. Given that data transformation between different softwares is a time-consuming process, a single package that completely covers well-known estimates of Fst and visualizes the results would be an asset.

2. Description

The FSTest can generate cross-population Fst values either per single nucleotide polymorphism (SNP) or as an average per genomic windows harboring multiple SNPs. In a sliding window analysis, Fst statistics are calculated for a small window of the data; however, windows can be determined based on a fixed number of base pairs (bp) or a fixed number of SNPs. While VCFtools provides sliding window modes only based on base pairs, PLINK lacks this feature. However, sliding window analysis on a fixed number of base pairs, particularly when low coverage genotypic data is utilized, leads to overestimation or underestimation of statistics since the number of SNPs inside windows would be inconsistent (Gusnanto et al., 2014).

Implementation of sliding window analysis using a fixed number of SNPs is available using FSTest.

FSTest is implemented in Python and accepts VCF files as input, which can be generated by PLINK software. The FSTest automatically manages missing genotypes and supports phased and unphased genotypic data. Inputs include two files: (i) a genotypic file and (ii) a grouping file in which individuals belonging to each of two populations are determined.

Between-population Fst estimators of Hudson et al. (1992), Weir and Cockerham (1984), Nei (1973), and Wright (1949) can be calculated as follows:

$$Fst_{Hudson} = \frac{(\tilde{p}_1-\tilde{p}_2)^2 - \frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1-1} - \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2-1}}{\tilde{p}_1(1-\tilde{p}_2)+\tilde{p}_2(1-\tilde{p}_1)},$$

$$Fst_{Weir\ and\ Cockerham} = 1 - \frac{2\frac{n_1 n_2}{n_1+n_2}\left(\frac{1}{n_1+n_2-2}\right)[n_1\tilde{p}_1(1-\tilde{p}_1)+n_2\tilde{p}_2(1-\tilde{p}_2)]}{\frac{n_1 n_2}{n_1+n_2}(\tilde{p}_1-\tilde{p}_2)^2+\left(2\frac{n_1 n_2}{n_1+n_2}-1\right)\frac{1}{n_1+n_2-2}[n_1\tilde{p}_1(1-\tilde{p}_1)+n_2\tilde{p}_2(1-\tilde{p}_2)]},$$

$$Fst_{Nei} = \frac{(\tilde{p}_1-\tilde{p}_2)^2}{(\tilde{p}_1+\tilde{p}_2)\left(1-\frac{\tilde{p}_1+\tilde{p}_2}{2}\right)},$$

$$Fst_{Wright} = \frac{(\tilde{p}_1+\tilde{p}_2)\left(1-\frac{\tilde{p}_1+\tilde{p}_2}{2}\right)-\left[\left(\frac{n_1}{n_1+n_2}\right)2\tilde{p}_1(1-\tilde{p}_1)+\left(\frac{n_2}{n_1+n_2}\right)2\tilde{p}_2(1-\tilde{p}_2)\right]}{(\tilde{p}_1+\tilde{p}_2)\left(1-\frac{\tilde{p}_1+\tilde{p}_2}{2}\right)},$$

where n_i is the sample size, and $\tilde{p}_i$ is the sample allele frequency in population i for $i \in \{1, 2\}$. All Fst estimations range between 0 and 1, and values less than zero are denoted zero by the software as they do not have a biological interpretation.

For sliding window analysis, the average Fst values across SNPs will be taken for each window. Users can also implement Z-transformation on all Fst estimates for SNP- or window-based approaches to get unbiased Z_{Fst} values as given:

$$Z_{FST_i} = \frac{Fst_i - \overline{Fst}}{S},$$

where Fst_i is the Fst value of ith SNP or ith window, $\overline{FST}$ and S represent the average value and standard deviation of Fst between two populations estimated from all SNPs or windows. We have also added the functionality of creating publication-quality Manhattan plots for the results of SNP and sliding window analyses.

In the software documentation, we demonstrated an example case by analyzing 46,819 SNP markers from 133 genotyped samples (50K SNP array). The analysis, including Fst estimation per SNP and using a 20 SNP sliding window with a step size of 5 SNP, along with Manhattan plots visualization (Fig. 1), could be completed within minutes on a standard desktop computer.

<<Figure 1 around here>>

3. Conclusion

FSTest software contains extra pairwise Fst estimation and visualization methods that do not present in the VCFtools and PLINK software, allowing direct analysis of VCF files with a small amount of code. All functions in FSTest were designed to be user-friendly with attention to the quality of visualizations.

Funding

We have received no specific funding for the current Study.

Data availability

The data underlying this article are available at FSTest GitHub repository (https://github.com/similab/FSTest).

Author contributions

SSA and SMV developed the original FSTest software. SMV wrote the manuscript. All authors read and approved the final manuscript.

References

Bhatia, G., Patterson, N., Sankararaman, S., Price, A.L., 2013. Estimating and interpreting FST: the impact of rare variants. Genome research 23, 1514–1521.

Biswas, S., Akey, J.M., 2006. Genomic insights into positive selection. TRENDS in Genetics 22, 437–446.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158.

Gusnanto, A., Taylor, C.C., Nafisah, I., Wood, H.M., Rabbitts, P., Berri, S., 2014. Estimating optimal window size for analysis of low-coverage next-generation sequence data. Bioinformatics 30, 1823–1829.

Holsinger, K.E., Weir, B.S., 2009. Genetics in geographically structured populations: defining, estimating and interpreting F ST. Nature Reviews Genetics 10, 639–650.

Hudson, R.R., Slatkin, M., Maddison, W.P., 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132, 583.

Nei, M., 1973. Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences 70, 3321–3323.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics 81, 559–575.

Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. evolution 1358–1370.

Wright, S., 1949. The genetical structure of populations. Annals of eugenics 15, 323–354.

Figures

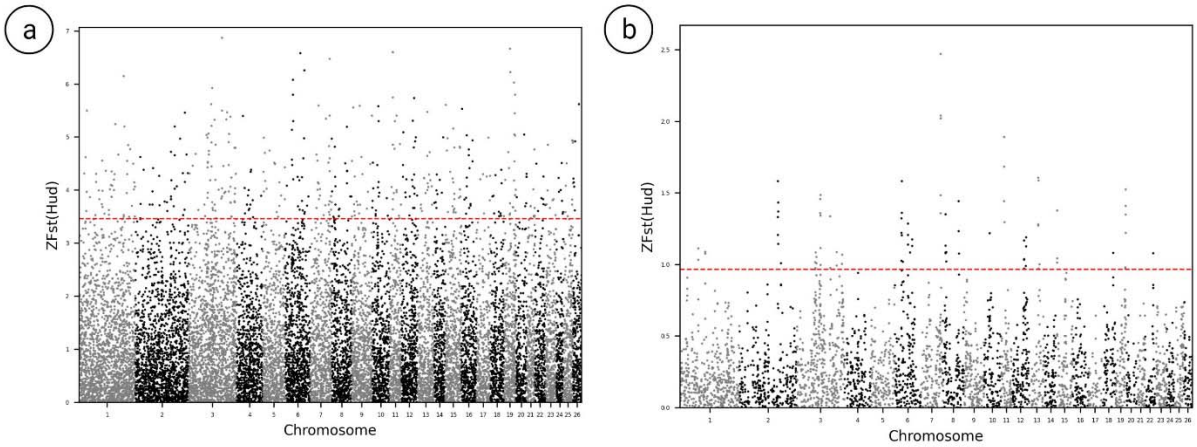
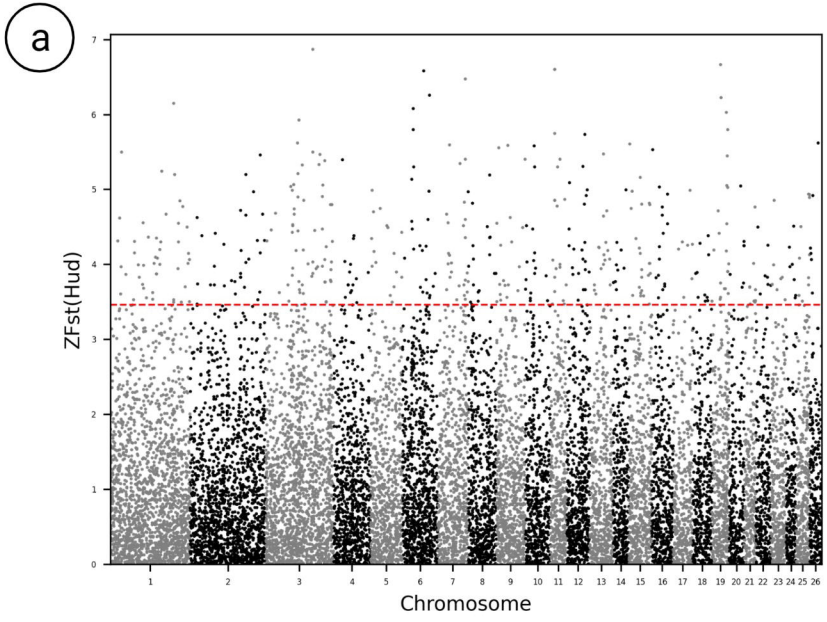
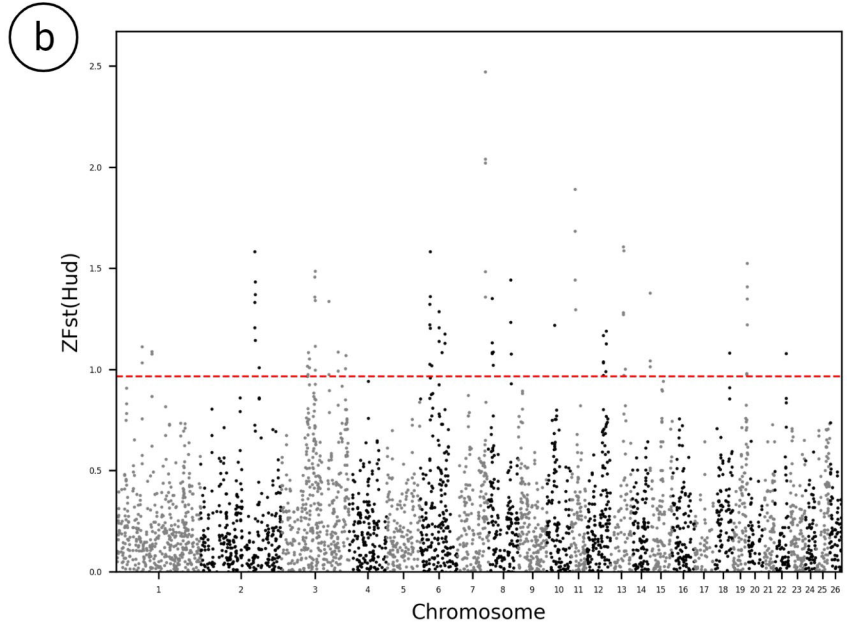


Figure 1. Example figures generated by the FSTest package. Manhattan plots (a) and (b) depict Z-transformed Hudson's Fst values estimated per SNP-based and using sliding window (sliding window size of 20 SNP and step size of 5 SNP) approach, respectively. The significance threshold is set at the top 1% of the Z(Fst) values distribution tail.


a
ZFst(Hud)
Chromosome


b
ZFst(Hud)
Chromosome