zalpha: an R package for the identification of regions of the genome under selection

Clare Horscroft[1, 3], **Reuben J Pengelly**[1, 3], **Timothy J Sluckin**[2], and **Andrew Collins**[1, 3]

1 Genetic Epidemiology and Bioinformatics, Faculty of Medicine, University of Southampton **2** Mathematical Sciences, University of Southampton **3** Institute for Life Sciences, University of Southampton

DOI: 10.21105/joss.02638

Software

- Review
- Repository
- Archive

Editor: Lorena Pantano

Reviewers:

- @juanvillada
- @marcrr

Submitted: 08 August 2020
Published: 05 December 2020

Summary

Detecting evidence of selection and evolution in population genomes is crucial to understanding the history and the selective pressures experienced by a population. While there are many statistics for identifying regions of the genome under selection, there is a need for software to enable reproducible, standardised results. The statistics implemented in the `zalpha` R package use the relationships and correlations in genetic variation to find patterns that could be indicative of a selective sweep.

The methods contained within this R package are a development of the statistics published by Jacobs et al. (2016). This package allows users to run a range of selection statistics on genetic data, which previously were not made publicly available in software. The software is designed to be flexible to allow users to efficiently combine statistics and is open source.

The package also allows users to utilise a linkage disequilibrium (LD) profile, taking into account expected relationships between alleles, ultimately increasing the power of the statistics. This is important as LD varies immensely along the genome, with recombination the biggest contributor to LD fluctuations (Jeffreys et al., 2001).

Statement of Need

The purpose of the `zalpha` package is to:

- Allow users to accurately apply the Z_α statistic to find candidate regions of the genome for a selective sweep
- Refine Z_α results by adjusting for expected correlations between genetic variants
- Further characterise sweeps as ongoing or near fixation using the Z_β statistic
- Generate results that are reproducible
- Be user-friendly and accessible by using R

Software and Methodology

The `zalpha` package examines correlations between single nucleotide polymorphisms (SNPs) along a chromosome. If SNPs are highly correlated in a region of a chromosome in relation to the rest of the genome, this could indicate the presence of a selective sweep (Vitti et al., 2013).

Correlation, in the context of genetics, is the ability to predict the value of one SNP, given the value of another. An example is given in Figure 1A. The metric used by these statistics to measure correlation is r^2 (Cutter, 2019).

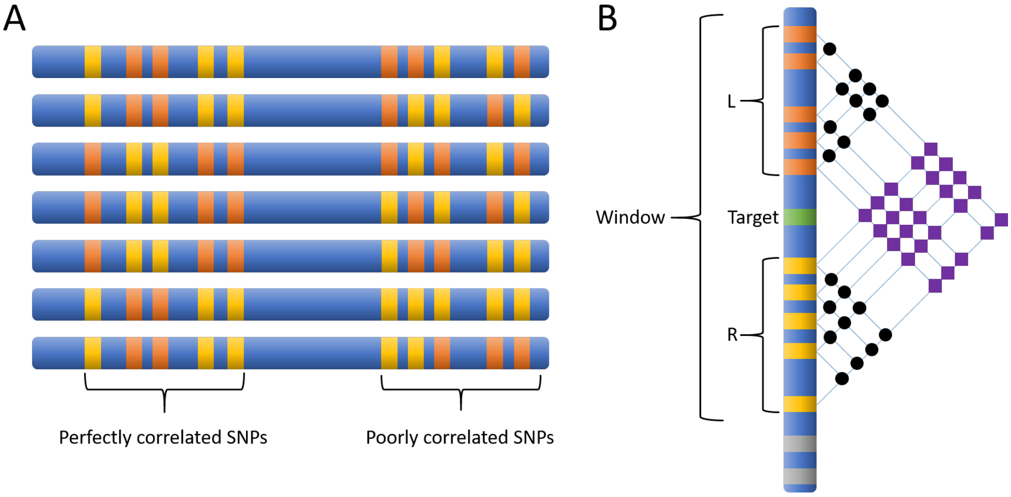

Figure 1: (A) This figure shows a population of seven chromosomes with 10 biallelic SNPs highlighted in orange and yellow to indicate the two alleles of the SNP. The cluster of five SNPs on the left are perfectly correlated. The second cluster on the right are poorly correlated: SNPs cannot be used to predict each other. (B) This figure shows a chromosome with SNPs highlighted. A statistic is being calculated for the target locus in green. The window is defined, and all SNPs falling within the window to the left and right of the target locus are assigned to sets L and R respectively. The correlations between pairs of SNPs are evaluated and represented by the black circles for SNPs within sets L and R, and as purple squares for pairs between the sets.

When a selective sweep occurs, the locus under selection becomes more frequent in the population, as individuals possessing the beneficial allele are more likely to survive and reproduce. When this happens, variants nearby the selected locus will also sweep, a phenomenon known as "hitchhiking" (Maynard Smith & Haigh, 1974). This creates a region of the genome that is highly correlated. Eventually recombination will erode away these correlations.

`zalpha` allows the user to apply a range of statistics to genetic data. Figure 1B shows a target locus with a window, the size of which is set by the user, centred on the locus. Any SNPs either side that fall within the window to the left and right of the target locus are contained within sets L and R respectively. The statistic Z_α, after which the package is named, is defined as:

$$Z_\alpha = \frac{\binom{|L|}{2}^{-1} \sum_{i,j \in L} r^2_{i,j} + \binom{|R|}{2}^{-1} \sum_{i,j \in L} r^2_{i,j}}{2} \quad (1)$$

|L| and |R| are the number of SNPs in each set, and $r^2_{i,j}$ is the correlation between two SNPs i and j. Figure 1B shows these r^2 values as black circles.

The other base statistic supplied in the `zalpha` package is Z_β, as defined as follows:

$$Z_\beta = \frac{\sum_{i \in L, j \in R} r^2_{i,j}}{|L||R|} \quad (2)$$

In Figure 1B the r^2 values for Z_β are represented as purple squares.

Typically, a user will want to find the maximum Z_α statistic in a region of a chromosome, and compare this to other regions, to find possible evidence of selection for that region.

The package is designed to be as user-friendly as possible and is reflected in the flexibility of the input requirements. The basic statistics only require three elements:

- vector of physical locations of each SNP,
- a window size, and
- a matrix of SNP values where the rows are SNPs and the columns are haplotypes. This matrix could be binary, where the 0s represent ancestral alleles and the 1s derived, or it could be nucleotides (i.e. As, Cs, Gs, and Ts), or any other biallelic labelling system.

One of the benefits of this package is the ability to calculate multiple statistics simultaneously. During a selective sweep, the correlations between alleles near to the selected locus increase. This means both Z_α and Z_β should be higher than in other areas of the genome not experiencing selective pressure. Towards the end of a selective sweep however, the correlations between the sets of alleles on the left and the right of the target locus are expected to diminish (Kim & Nielsen, 2004). This suggests at the end of a sweep, Z_α should remain high, but Z_β will reduce. Thus, it is advantageous to calculate and combine the different statistics to ascertain the strength and stage of sweeps. Z_α/Z_β is a simple way to achieve this.

Recombination is a process that has the effect of breaking down the relationship between alleles. However, it is known that recombination does not occur uniformly across the genome. It is therefore imperative to consider recombination when calculating statistics based on LD measures. This package allows the user to supply a population LD profile, providing information on the expected relationships between alleles given the genetic distances between them. Supplying these data increases the power of the statistics and creates more opportunities for combinations and comparisons between statistics. Users can specify whatever units they wish for genetic distance (for example centimorgans (cM)), derived from an appropriate data source. The software contains a function for creating an LD profile from the data. Ideally, an LD profile would be created from a neutral data source without selection, for example from a simulation with relevant population parameters. However, this is not always possible, so creating an LD profile from the same data being analysed is sufficient.

There are many statistics included in the package for adjusting for expected r^2 using the LDprofile and genetic distances between SNPs. It is recommended the user runs all the statistics using the `Zalpha_all()` function and then chooses the ones they are interested in, perhaps even creating their own. For example, $Z_\alpha/Z_\alpha^{E[r^2]}$ performs well as a simple way to adjust for expected r^2. If it is known that the r^2 values for each genetic distance are normally distributed, Z_α^{Zscore} is appropriate, otherwise $Z_\alpha^{BetaCDF}$ may be useful. For more details of how they are derived see the paper by Jacobs et al. (2016). This paper also shows how the different statistics perform under a range of demographic scenarios.

The output of the functions is in list format. The SNP positions and the values of the statistic(s) are stored in vectors of equal length in the list. Users can then identify outlying SNPs in their data that are candidate regions for selection.

There are a few other R packages that can be used for single population selection scans, although none utilise the Z_α statistics described here. `PopGenome` has the advantage of being able to read in multiple standard genomic data formats and perform other analyses as well as selection scans (Pfeifer et al., 2014). The scans it performs are the CL and CLR methods by Nielsen et al. (2005). It can also calculate Kelly's Z_{nS} (Kelly, 1997) and Rozas' ZA/ZZ (Rozas et al., 2001) for LD, which are similar to the basic Z_α statistic; however, neither use an LD profile to correct for expected correlation. `rehh` is another package used for selection scans, and for single populations uses extended haplotype homozygosity (EHH) and related statistics (Gautier et al., 2017). These are popular and easy to interpret methods; however, they are not as effective for soft sweeps (Liebert et al., 2017). Variations in local recombination rate are accounted for by implementing the cross-population EHH (XP-EHH) (Sabeti et al., 2007), which requires another population to compare with (Vitti et al., 2013). All these packages identify evidence of selective sweeps in different ways, and each have merits. It would be advisable to use a combination of approaches to achieve the most accurate and valid results.

Conclusion

This new package allows researchers to calculate the Z_α suite of selection statistics efficiently using the free, open source R platform. These statistics had previously not been publicly available in software. The package's flexibility allows the user to adjust the statistics for the expected r^2 value via an LD profile in a variety of ways, and enables the adjustment of the base statistics to create new and novel methods.

Acknowledgements

The authors acknowledge the use of the IRIDIS High Performance Computing Facility in the completion of this work. The authors would like to acknowledge Guy Jacobs whose work on these statistics was the foundation for this package.

References

Cutter, A. D. (2019). *A Primer of Molecular Population Genetics*. Oxford University Press.

Gautier, M., Klassmann, A., & Vitalis, R. (2017). rehh2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Molecular Ecology Resources*, *17*(1), 78–90. https://doi.org/10.1111/1755-0998.12634

Jacobs, G. S., Sluckin, T. J., & Kivisild, T. (2016). Refining the Use of Linkage Disequilibrium as a Robust Signature of Selective Sweeps. *Genetics*, *203*(4), 1807–1825. https://doi.org/10.1534/genetics.115.185900

Jeffreys, A. J., Kauppi, L., & Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, *29*, 217. https://doi.org/10.1038/ng1001-217

Kelly, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics*, *146*(3), 1197–1206.

Kim, Y., & Nielsen, R. (2004). Linkage Disequilibrium as a Signature of Selective Sweeps. *Genetics*, *167*(3), 1513–1524. https://doi.org/10.1534/genetics.103.025387

Liebert, A., Lopez, S., Jones, B. L., Montalva, N., Gerbault, P., Lau, W., Thomas, M. G., Bradman, N., Maniatis, N., & Swallow, D. M. (2017). World-wide distributions of lactase persistence alleles and the complex effects of recombination and selection. *Human Genetics*, *136*(11-12), 1445–1453. https://doi.org/10.1007/s00439-017-1847-y

Maynard Smith, J., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research*, *23*(1), 23–35. https://doi.org/10.1017/S0016672300014634

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Research*, *15*(11), 1566–1575. https://doi.org/10.1101/gr.4252305

Pfeifer, B., Wittelsbuerger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution*, *31*, 1929–1936. https://doi.org/10.1093/molbev/msu136

Rozas, J., Gullaud, M., Blandin, G., & Aguadé, M. (2001). DNA Variation at the rp49 Gene Region of Drosophila simulans: Evolutionary Inferences From an Unusual Haplotype Structure. *Genetics*, *158*(3), 1147–1155.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., & Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918. https://doi.org/10.1038/nature06250

Vitti, J. J., Grossman, S. R., & Sabeti, P. C. (2013). Detecting Natural Selection in Genomic Data. In *Annual review of genetics* (Vol. 47, pp. 97–120). Annual Reviews. https://doi.org/10.1146/annurev-genet-111212-133526