The Journal of Open Source Software

FGTpartitioner: A rapid method for parsimonious delimitation of ancestry breakpoints in large genome-wide SNP datasets

Tyler K. Chafin[1]

1 Tyler K. Chafin, Ph.D Candidate, University of Arkansas

DOI: 10.21105/joss.02030

Software

- Review
- Repository
- Archive

Editor: Lorena Pantano

Reviewers:

- @dangeles
- @YichaoOU

Submitted: 07 January 2020
Published: 06 February 2020

Summary

Partitioning large (e.g. chromosomal) alignments into ancestry blocks is a common step in phylogenomic analyses (Springer & Gatesy, 2018). However, current solutions require complicated analytical assumptions, or are difficult to implement due to excessive runtimes. Multiple approaches have been proposed for delimiting ancestry blocks in genomes (i.e. establishing recombination breakpoints), which generally fall into one of two categories: those which require dense or phased genotypic data (Liu et al., 2013); and those with complex analytical assumptions which require the definition of informative prior probability distributions and are computationally intensive (Dutheil et al., 2009). Both conditions are problematic for genome-scale studies of non-model species, where large-scale resequencing and phased reference data are unavailable, and genomes are often sequenced at low coverage.

I here describe a solution, `FGTpartitioner`, which is specifically designed for use with non-model genomic data without the need for high-quality phased reference data or dense population-scale sampling. `FGTpartitioner` delimits chromosome scale alignments using a fast interval-tree approach which detects pairwise variants which violate the four-gametes assumption (Hudson & Kaplan, 1985), and rapidly resolves a most parsimonious set of recombination events to yield non-overlapping intervals which are both unambiguously defined and consistent regardless of processing order. These sub-alignments are then suitable for separate phylogenetic analysis, or as a 'first pass' which may facilitate parallel application of finer-resolution (yet more computationally intensive) methods.

After parsing user-inputs, the workflow of `FGTpartitioner` is as follows:

(1) For each SNP, perform four-gamete tests sequentially for rightward neighboring records, up to a maximal physical distance (if defined) and stopping when a conflict (='interval') is found. Intervals are stored in a self-balancing tree. When using multiprocessing, daughter processes are each provided an offset which guarantees a unique pairwise SNP comparison for each iteration
(2) Merge interval trees of daughter processes (if using optional parallel computation)
(3) Assign rank k per-interval, defined as the number of SNP records (indexed by position) spanned by each interval
(4) Order intervals by k; starting at min(k), resolve conflicts as follows: For each candidate recombination site (defined as the mid-point between SNPs), compute the depth d of spanning intervals. The most parsimonious breakpoint is that which maximizes d

These algorithm choices have several implications: indexing SNPs by physical position guarantees that the same recombination sites will be chosen given any arbitrary ordering of SNPs; and defining breakpoints as physical centerpoints between nodes means that monomorphic sites will be evenly divided on either side of a recombination event. Because monomorphic sites by definition lack phylogenetic information, they cannot be unambiguously assigned to any particular ancestry block, thus my solution is to evenly divide them. Heterozygous sites in diploid genomes are dealt with in multiple ways. By default, `FGTpartitioner` will randomly resolve haplotypes. The user can select an alternate resolution strategy which will either treat a SNP pair as failing if any resolution meets the four-gamete condition, or as passing if any possible resolution passes (i.e. the 'pessimistic' and 'optimistic' strategies of Wang et al. (2010)).

In conclusion, `FGTpartitioner` has several advantages over similar methods: 1) algorithmic and performance enhancements allow it to perform orders of magnitude faster, thus extending application to larger genomes; and 2) the flexibility of diploid resolution strategies precludes the need for haplotype phasing a priori. Validation using empirical data indicated the suitability of `FGTpartitioner` for highly distributed work on high-performance computing clusters, with parallelization easily facilitated by built-in options in the command-line interface. Additionally, runtime and memory profiling indicate its applicability on modern desktop workstations as well, when applied to moderately sized datasets. Thus, it provides an efficient and under-friendly solution to alignment pre-processing for phylogenomic studies, or as a method of breaking up large alignments in order to efficiently distribute computation for more rigorous recombination tests.

Acknowledgements

I would like to thank VonHoldt et al. (2016) and Kukekova et al. (2018) for making their raw datasets available via the NCBI SRA, and the Arkansas High Performance Computing Center and my Ph. D. advisors Drs. Michael and Marlis Douglas for access to computational resources which I used for benchmarking this program.

References

Dutheil, J. Y., Ganapathy, G., Hobolth, A., Mailund, T., Uyenoyama, M. K., & Schierup, M. H. (2009). Ancestral population genomics: The coalescent hidden Markov model approach. *Genetics*, *183*(1), 259–274. doi:10.1534/genetics.109.103010

Hudson, R. R., & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, *111*(1), 147–164.

Kukekova, A. V., Johnson, J. L., Xiang, X., Feng, S., Liu, S., Rando, H. M., Kharlamova, A. V., et al. (2018). Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours. *Nature Ecology & Evolution*, *2*(September). doi:10.1038/s41559-018-0611-6

Liu, Y., Nyunoya, T., Leng, S., Belinsky, S. A., Tesfaigzi, Y., & Bruse, S. (2013). Softwares and methods for estimating genetic ancestry in human populations. *Human Genomics*, *7*(1), 1–7. doi:10.1186/1479-7364-7-1

Springer, M. S., & Gatesy, J. (2018). Delimiting Coalescence Genes (C-Genes) in Phylogenomic Data Sets. *Genes*, *9*(123), 1–19. doi:10.3390/genes9030123

VonHoldt, B. M., Cahill, J. A., Fan, Z., Gronau, I., Robinson, J., Pollinger, J. P., Shapiro, B., et al. (2016). Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. *Science Advances*, *2*(7), e1501714–e1501714. doi:10.1126/sciadv.1501714

Wang, J., Moore, K. J., Zhang, Q., Villena, F. P.-M. de, Wang, W., & McMillan, L. (2010). Genome-wide compatible SNP intervals and their properties. *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, 43–52. doi:10.1145/1854776.1854788