Variation in recombination rate affects detection of F_{ST} outliers under neutrality

Tom R. Booker[a,b,1], **Samuel Yeaman**[c], **and Michael C. Whitlock**[b,d]

[a]Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, Canada; [b]Biodiversity Research Centre, University of British Columbia, Vancouver, Canada; [c]Department of Biological Sciences, University of Calgary, Calgary, Canada; [d]Department of Zoology, University of British Columbia, Vancouver, Canada

Genome scans can potentially identify genetic loci involved in evolutionary processes such as local adaptation and gene flow. Here, we show that recombination rate variation across a neutrally evolving genome gives rise to different sampling distributions of mean F_{ST} ($\hat{F_{ST}}$), a common population genetic summary statistic. In particular, we show that in low recombination regions the distribution of $\hat{F_{ST}}$ estimates may have a longer tail than in more highly recombining regions. Determining outliers from the genome-wide distribution without accounting for local recombination rate may therefore increase the frequency of false positives in low recombination regions and be overly conservative in more highly recombining ones. We perform genome-scans on simulated and empirical *Drosophila melanogaster* datasets and, in both cases, find patterns consistent with this neutral model. Our results highlight a flaw in the design of genome scan studies and suggest that without estimates of local recombination rate, it is very difficult to interpret the genomic landscape of population genetic summary statistics.

Population structure | Genome scan | F_{ST} | Coalescence

Genome scans have become a standard analysis in population genetics used to identify genomic regions involved in important evolutionary processes. With a reference genome and DNA re-sequencing data, the genome can be divided into analysis windows of a fixed physical size and the distribution of summary statistics across those windows can be examined. Outliers to the genome-wide distribution of a particular summary statistic may then be investigated for their roles in processes such as adaptation and speciation.

A particularly important summary statistic used in genome scans is Wright's F_{ST}. Wright (1) examined the behaviour of allele frequencies in metapopulations and referred to the correlation of frequencies among demes as F_{ST}. The evolutionary history of a population can be thought of as a series of genealogies describing the relationships between every individual at a certain point in the genome, and Slatkin (2) showed that for a neutrally evolving metapopulation F_{ST} can be expressed in terms of the coalescence times of those genealogies. In particular, $F_{ST} = \frac{\bar{T}-\bar{T_0}}{\bar{T}}$, where $\bar{T_0}$ is the average coalescence times for a pair of alleles drawn from the same deme and $\bar{T}$ is the average coalescence time for a pair of alleles drawn from the metapopulation as a whole. F_{ST} varies across the genome; the mutations that happened to occur in the population and the individuals that happened to be sampled will give rise to variation in F_{ST} at different points in the genome. There are numerous methods to average across linked sites in an attempt to account for these sources of variation (reviewed in (3)). Furthermore, the underlying demographic process is also stochastic and will generate different histories across the genome generating heterogeneity in F_{ST}.

In the context of genome scans, $\hat{F_{ST}}$ is calculated in analysis windows, and the genome-wide distribution is examined under the implicit assumption that all windows share the same sampling distribution. However, recombination breaks down associations across genetically linked sites allowing coalescence times to vary across the genome causing some windows to reflect more or fewer coalescent histories than others. Thus, the sampling distributions of summary statistics that capture aspects of coalescent history likely covary with the rate of recombination.

Consider a 10,000bp analysis window in a neutrally evolving genomic region that experiences virtually no recombination (for example, the mitochondrial genome or non-pseudoautosomal regions of the Y-chromosome in humans). Every nucleotide within this window would share a single genealogy, and every polymorphism present would reflect the exact same coalescent history. If there were a large number of polymorphisms in the window, the combined $\hat{F_{ST}}$ across all of them may provide a very precise estimate of their shared history. If one had a large number of these analysis windows and examined the distribution of their averages, it should resemble the coalescent distribution of F_{ST}, which, for a population conforming to the island model, can be approximated using the χ^2 distribution (4), which has a long upper tail.

Now consider the alternate extreme, 10,000bp of freely recombining nucleotides. In this case, all sites would have a distinct evolutionary history, each representing a quasi-independent instantiation of the coalescent process. As a result, every polymorphism in the window would carry information on a different genealogy, and each of these may have a distinct F_{ST}. The observed $\hat{F_{ST}}$ across all polymorphisms in this set provides an estimate of the mean of the coalescent distribution of F_{ST}. If one had a large number of these sets of loci, the values of $\hat{F_{ST}}$ should be tightly distributed about the mean, and central limit theorem predicts that the distribution should be approximately Gaussian.

Obviously 10,000 contiguous nucleotides do not experience free recombination, but eukaryotes do exhibit wide variation in recombination rates across their genomes. For example, humans (5), collared flycatchers (6) and *Drosophila melanogaster* (7) all exhibit recombination rate variation over at least an order of magnitude across substantial proportions of their genomes. If the sampling distribution of $\hat{F_{ST}}$ had a longer tail in regions of the genome with low recombination and was narrower in more highly recombining ones, heterogeneous landscapes of F_{ST} may arise under neutrality. Determining outliers from the genome-wide distribution of summary statistics may, therefore, result in false positives and negatives depending on

Author contributions: T.R.B., S.Y., and M.C.W. designed research; T.R.B. performed research; and T.R.B., S.Y., and M.C.W. wrote the paper.

The authors declare no competing interests

[1]To whom correspondence should be addressed. E-mail: booker@zoology.ubc.ca

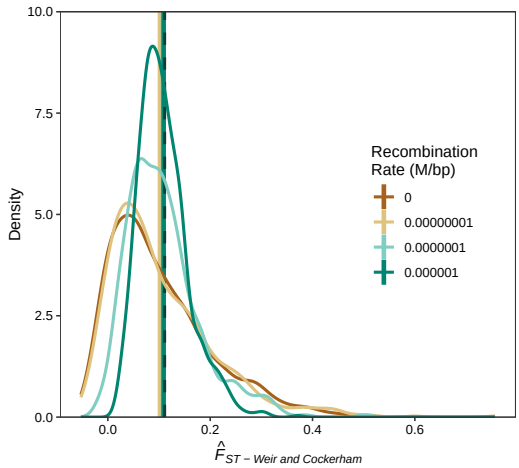

Fig. 1. The distribution of F_{ST} calculated in 10 Kbp windows. F_{ST} was averaged using the method of Weir and Cockerham. The dashed vertical line is the theoretical expectation of F_{ST} for the simulated population and the coloured vertical lines indicate the means for each set of simulations.

the local recombination rate. In this study, we demonstrate that the sampling distribution of $\hat{F_{ST}}$ does indeed depend on the local recombination rate and show how this may affect genome-scans using simulated and empirical datasets.

Results and Discussion

To determine whether the sampling distribution of $\hat{F_{ST}}$ is affected when recombination rates vary over orders of magnitude, we simulated genomic data under an island model of population structure. Different distributions of mean $\hat{F_{ST}}$ arose under different recombination rates (Fig. 1). The theoretical expectation of F_{ST} was the same in all of the simulations shown in Figure 1, and the average $\hat{F_{ST}}$ was very close to the expectation in all cases. However, low recombination rates led to longer-tailed distributions of $\hat{F_{ST}}$ and higher rates led to distributions that were tighter about the mean (Fig 1).

A typical method for determining outliers in genome scans is to examine the top n^{th} percentile of a particular summary statistic. If one were performing a genome scan in a species that exhibited recombination rate variation even over a single order of magnitude, one would expect the tail of the genome-wide distribution of $\hat{F_{ST}}$ to be enriched for analysis windows in low recombination regions under neutrality (Fig. 1). Indeed, the recombination rates shown in Figure 1 obviously lead to different distributions of $\hat{F_{ST}}$, but the extent to which this will affect the analysis of real organisms will depend on the pattern of variation in recombination rates across their genomes.

We examined the relationship between $\hat{F_{ST}}$ and recombination rate across the *D. melanogaster* genome using simulated and empirical datasets. We simulated the entire *D. melanogaster* genome under a model of isolation-with-migration incorporating recombination rate variation using *stdpopsim* (8) and performed genome scans on the resulting data (Fig. 2). We performed a similar genome scan on the population genomic data from North American *D. melanogaster* populations generated by Reinhardt et al (9). In both cases, we found a significant enrichment of outliers at low recombination rates compared to high recombination rates when we defined outliers based on the 95th percentile of $\hat{F_{ST}}$ (simulated data Fisher's exact test *p-value* $< 10^{-15}$; empirical data Fisher's exact test *p-value* $= 0.00265$). Furthermore, both the simulated and empirical datasets exhibited significant negative correlations between the variance of $\hat{F_{ST}}$ and recombination rate (Fig. 2C; simulated data, Kendall's τ = -0.746, *p-value* $< 10^{-15}$; empirical data, Kendall's τ = -0.193, *p-value* $= 0.00445$). Of course, the genomic landscape of $\hat{F_{ST}}$ in North American *D. melanogaster* has likely been shaped by processes other than just migration and genetic drift. Indeed, Figure 2B shows at least one $\hat{F_{ST}}$ outlier at high recombination rates, which our simulations suggest would be unlikely under neutrality (Fig. 2A). In summary, our findings suggest that applying fixed thresholds to the genome-wide distribution of $\hat{F_{ST}}$ may enrich for false positives in low recombination regions and be overly conservative in highly recombining regions of the genome.

The effects of selection at linked sites, particularly background selection, have been invoked to explain patterns of variation in F_{ST} across the genome (10, 11). Linked selection is most effective when recombination rates are low, so may cause heterogeneity in F_{ST} across the genome. Closely related species that have correlated recombination rate landscapes and syntenic genomes may experience similar background selection effects and thus exhibit similar variation in $\hat{F_{ST}}$ (11). However, recent work has shown that background selection may have quite limited effects on F_{ST} in many cases (12), so other explanations are perhaps required to explain genome-wide variation in $\hat{F_{ST}}$. Notably, species with correlated recombination landscapes may be prone to exhibiting high F_{ST} in similar regions. Under neutrality, $\hat{F_{ST}}$ does not systematically vary as a function of recombination rate, but the potential to be an outlier does. Overlapping F_{ST} outliers may then be mistaken for convergent adaptive evolution.

We have focussed on F_{ST} in this study, but it has long been known that the variance of summary statistics such as the number of segregating sites responds to variation in recombination rate (13), so the difficulty of comparing analysis windows of discrete physical size across varying recombination rates likely extends to other data summaries as well. For example, absolute nucleotide diversity between populations (D_{XY}) is often used in conjunction with F_{ST} in genome scans. We calculated D_{XY} from our simulated data and found that the variance of that statistic was also negatively correlated with recombination rate (*data not shown*).

The results from this study should emphasise the need to incorporate recombination into the null model used to interpret genome scans. Without estimates of the local recombination rate it will be difficult to determine whether a particular statistical outlier is driven by adaptive or neutral processes.

Methods

Coalescent simulations of structured populations were performed using *msprime* v0.7.3 (14). We simulated an island model of 100 demes, each with $N_e = 100$ haploid individuals. Migration was constant between all demes and was set to $M = N_e m = 2$, giving an expected F_{ST} of $0.1\dot{1}$. After simulating a particular population, we added mutations to the resulting tree sequence at a rate of 10^{-6} then sampled 10 diploid individuals each from two demes and calculated the weighted average of F_{ST} using the method for diploids given in Weir (15) as implemented in *scikit-allel* v1.2.1. We repeated this procedure 100 times for each simulated population and took the average

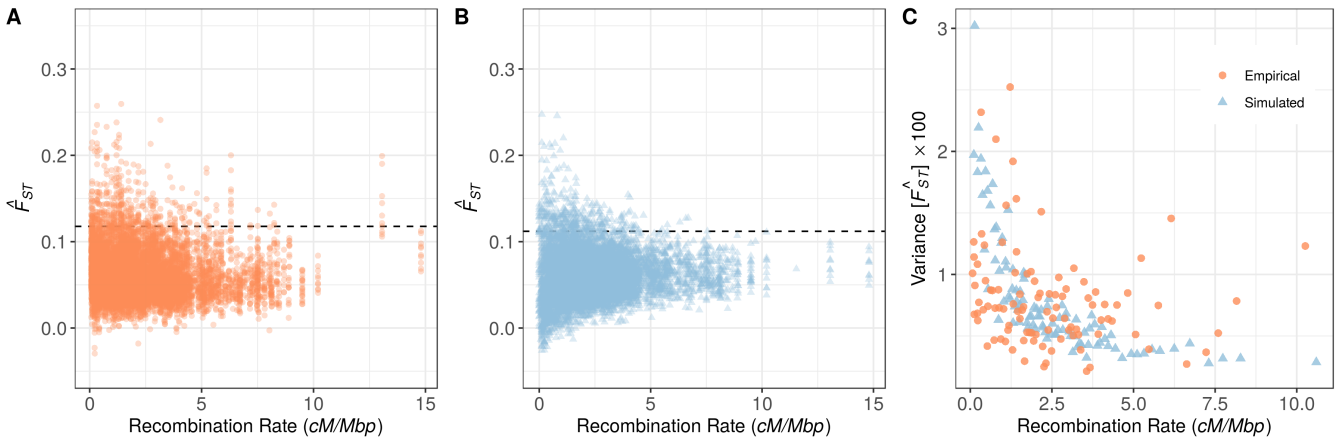

Fig. 2. The plot of $\hat{F_{ST}}$ calculated in 10,000bp analysis windowsd against local recombination rate for *Drosophila melanogaster* from A) the North American populations of *Drosophila melanogaster* analysed by Reinhardt et al (9) and B) coalescent simulations of the genome under a model of isolation with migration. C) Shows the variance in $\hat{F_{ST}}$ against recombination rate for both empirical and simulated datasets. The horizontal lines are the 95th percentiles of the genome-wide distribution in each case

across replicates. We performed 1,000 such simulations for each recombination rate tested.

We analysed allele frequency data from the North American *Drosophila* populations analysed by Reinhardt et al (9). In that study, pooled sequencing was used to estimate derived allele frequencies for pools of 16 isofemale lines each from Maine and Florida. To obtain accurate allele freuqencies we excluded any single nucleotide polymorphism that was reported at a depth greater than or less than one standard deviation from the mean coverage.

We simulated the entire *Drosophila melanogaster* genome under a model of isolation with migration model in *stdpopsim* v0.1.1. An ancestral population with effective size $2N_e = 344,120$ individuals, split into two demes of N_e individuals N_e generations in the past. Symmetrical migration of $m = \frac{3}{2N_e}$ migrants per generation occurred after the split. The migration rate was chosen match F_{ST} with the empirical data. We sampled 20 haploid individuals from each deme.

We analysed the genome-wide distribution of F_{ST} in the empirical and simulated *Drosophila melanogaster* datasets. Because the Reinhardt et al (9) data was generated using pooled-sequencing, we calculated F_{ST} for each polymorphism and averaged across sites using the formulae for haploids given in Weir (15). $\hat{F_{ST}}$ was calculated in 10,000bp non-overlapping windows using the ratio-of-averages approach. Recombination rates were extracted from the Comeron et al (7) map. For both the simulated and empirical *D. melanogaster* data, we examined the relationship between the variance of $\hat{F_{ST}}$ and recombination rate by dividing the recombining genome into 75 equally sized bins. We calculated Kendall's correlation between the variance of $\hat{F_{ST}}$ and the mean recombination rate in these bins. We classified analysis windows as outliers if $\hat{F_{ST}}$ was greater than the 95th percentile genome-wide. We examined the analysis windows with the 1,000 largest and smallest recombination rate estimates and used Fisher's exact test to determine whether there was a significant enrichment of outliers in low recombination regions. We excluded analysis windows from the simulated data that had recombination rates of 0.

Data availability. Python scripts implementing the simulations shown as well as additional population models are available at https://github.com/TBooker/Recombination_Fst.

ACKNOWLEDGMENTS. Thanks to Sally Otto, Andréa Thomaz and Wouter van der Bijl for helpful discussions and to Matt Pennell for comments on the manuscript. Thanks to Andrew Kern for insights and for providing the *Drosophila* data. This project is part of the CoAdapTree project which is funded by Genome Canada (241REF), Genome BC and 16 other sponsors (http://coadaptree.forestry.ubc.ca/sponsors/). MCW is supported by a Discovery Award from NSERC. SY is supported by a Discovery Award from NSERC and a research chair from Alberta Innovates.

References.

1. Wright S (1931) Evolution in Mendelian Populations. *Genetics* 16.
2. Slatkin M (1991) Inbreeding coefficients and coalescence times. *Genetical Research* 58(2):167–175.
3. Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: Defining, estimating and interpreting FST.
4. Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74(1):175–195.
5. Kong A, et al. (2010) Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467(7319):1099–1103.
6. Kawakami T, et al. (2014) A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Molecular Ecology* 23(16):4035–4058.
7. Comeron JM, Ratnappan R, Bailin S (2012) The Many Landscapes of Recombination in Drosophila melanogaster. *PLoS Genetics* 8(10):e1002905.
8. Adrion JR, et al. (2019) A community-maintained standard library of population genetic models. *bioRxiv*.
9. Reinhardt JA, Kolaczkowski B, Jones CD, Begun DJ, Kern AD (2014) Parallel geographic variation in Drosophila melanogaster. *Genetics* 197(1):361–373.
10. Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* 23(13):3133–3157.
11. Burri R (2017) Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters* 1(3):118–131.
12. Matthey-Doret R, Whitlock MC (2019) Background selection and f_{ST}: consequences for detecting local adaptation. *Molecular ecology* 28(17):3902–3914.
13. Wakeley J (2009) *Coalescent theory : an introduction.* (Roberts & Co. Publishers), p. 326.
14. Kelleher J, Etheridge AM, McVean G (2016) Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.* 12(5):e1004842.
15. Weir BS (1990) *Genetic data analysis. Methods for discrete population genetic data.* (Sinauer Associates, Inc. Publishers).