Variation in recombination rate affects detection of outliers in genome scans under neutrality

Tom R. Booker^{a,b}, Sam Yeaman^c, 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

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.

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

Abstract

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 mixed sampling distributions of mean F_{ST} ($\widehat{F}_{ST}$), a common population genetic summary statistic. In particular, we show that in regions of low recombination the distribution of $\widehat{F}_{ST}$ estimates have more variance and a longer tail than in more highly recombining regions. Determining outliers from the genome-wide distribution without taking local recombination rate into consideration 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. Similar patterns are observed for other summary statistics used to capture variation in the coalescent process. Linked selection, particularly background selection, is often invoked to explain heterogeneity in $\widehat{F}_{ST}$ across the genome, but here we point out that even under neutrality, statistical artefacts can arise due to variation in recombination rate. Our results highlight a flaw in the design of genome scan studies and suggest that without estimates of local recombination rate, interpreting the genomic landscape of any summary statistic that captures variation in the coalescent process will be very difficult.

45 Introduction

46

47 Genome scans have become a standard analysis in population genetics to identify particular regions
 48 that exhibit a significant departure from the background patterns observed in the rest of the
 49 genome. Many different statistics are used in such scans, and finding an outlier in the distribution
 50 of the test statistic of interest can indicate the presence of a causal variant affecting a phenotype,
 51 driving a selective sweep, or contributing to local adaptation, among other things (Cruickshank &
 52 Hahn, 2014; Haasl & Payseur, 2016; Hoban et al. 2016; Wolf & Ellegren, 2017). While some
 53 approaches are based on single-locus measures, a common practice is to aggregate statistics across
 54 loci within windows of the genome. As closely linked loci tend to share the same evolutionary
 55 history and therefore exhibit similar statistical signatures, this can potentially increase the power
 56 and decrease the influence of stochastic noise. However, because recombination rate can vary
 57 across the genome, some windows will tend to share evolutionary history among loci more strongly
 58 than other windows, and failure to account for this may result in bias in the estimation of the null
 59 distribution and testing procedure using window-based approaches. Here, we explore these issues
 60 using F_{ST} as an example, which is a statistic commonly used to search for signatures of local
 61 adaptation in the genome. There has been considerable discussion about how other processes,
 62 such as background or positive selection, can affect F_{ST} , and a general acknowledgement of the
 63 importance of recombination variation for these alternative selection models (Burri, 2017;
 64 Cruickshank & Hahn, 2014; Nachman & Payseur, 2012), but little attention has been paid to the
 65 effects of recombination rate variation on the distribution of the neutral model. Given that all
 66 approaches that discuss selection and F_{ST} do so with reference to departures from a null model
 67 under drift, here we explore how recombination variation can also have critically important effects
 68 on the behaviour of the null model. This note emphasizes how failure to incorporate recombination
 69 rate into the null model leads to significant errors in our inferences and suggests more generally

that this will also be important for any window-based approach to genome scans using other statistics.

Wright (1937) examined the behaviour of allele frequencies in metapopulations and defined F_{ST} as the correlation of allelic states between individuals from the same population. 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 (1991) showed that for a neutrally evolving metapopulation F_{ST} reflects the pattern of coalescence times of those genealogies. In particular, $F_{ST} = \frac{\bar{T} - T_0}{\bar{T}}$, where T_0 is the average coalescence time 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 F_{ST} across linked sites in an attempt to account for these sources of variation (Bhatia et al. 2013; Holsinger & Weir, 2009). Furthermore, the underlying demographic process is also stochastic and will generate different histories across the genome giving rise to heterogeneity in F_{ST} .

In the context of genome scans, $\widehat{F_{ST}}$ is often calculated in analysis windows of a fixed physical size or number of single nucleotide polymorphisms (SNPs), and the genome-wide distribution is examined under the implicit assumption that all windows share the same sampling distribution. This assumption is reflected by the discrete thresholds that are often applied to genome-wide distributions of $\widehat{F_{ST}}$ to identify outliers in empirical studies (Harpak et al., 2020; Lai et al., 2019; Ma et al., 2017; Olofsson et al., 2019; Talla et al., 2019; Zhou et al., 2019). 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.

95 Thus, the sampling distribution of $\widehat{F_{ST}}$, and indeed any summary statistic that captures aspects of
 96 coalescent history, will likely covary with the rate of recombination.
 97
 98 Consider a 10,000bp analysis window in a neutrally evolving genomic region that experiences
 99 virtually no recombination (for example, the mitochondrial genome or non-pseudoautosomal
 100 regions of the Y-chromosome in humans). Every nucleotide within the window would share a single
 101 genealogy, and every polymorphism present would reflect exactly the same coalescent history. If
 102 there were a large number of polymorphisms in the window, the combined $\widehat{F_{ST}}$ across all of them
 103 may provide a very precise estimate of their shared history. If one had a large number of these
 104 analysis windows and examined the distribution of their averages, it should resemble the
 105 coalescent distribution of F_{ST} , which, for a population conforming to the island model, can be
 106 approximated using the χ^2 distribution (Lewontin & Krakauer, 1973), which has a long upper tail.
 107
 108 Now consider the alternate extreme, 10,000bp of freely recombining nucleotides. In this case, all
 109 sites would have a distinct evolutionary history, each representing a quasi-independent
 110 instantiation of the coalescent process. As a result, every polymorphism present would carry
 111 information on a different genealogy, and each of these may have a distinct F_{ST} . The observed $\widehat{F_{ST}}$
 112 across all polymorphisms in this set provides an estimate of the mean of the coalescent distribution
 113 of F_{ST} . If one had a large number of these sets of loci, the values of $\widehat{F_{ST}}$ should be tightly distributed
 114 about the mean, and central limit theorem predicts that the distribution should be approximately
 115 Gaussian.
 116
 117 Obviously, 10,000 contiguous nucleotides do not experience free recombination, but eukaryotes do
 118 exhibit wide variation in recombination rates across their genomes. For example, humans (Kong et
 119 al. 2010), collared flycatchers (Kawakami et al. 2014) and *Drosophila melanogaster* (Comeron et al.

2012) all exhibit recombination rate variation over at least an order of magnitude across substantial proportions of their genomes (Supplementary Figure 1). If the sampling distribution of $\widehat{F}_{ST}$ had a longer tail in regions of the genome with low recombination and was narrower in more highly recombining ones, heterogeneous landscapes may arise as a statistical artefact under neutrality. Determining outliers from the genome-wide distribution of summary statistics may, therefore, result in false positives and negatives depending on the local recombination rate. In this study, we demonstrate that the sampling distribution of $\widehat{F}_{ST}$ 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 $\widehat{F}_{ST}$ 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 $\widehat{F}_{ST}$ arose under different recombination rates (Figure 1). The theoretical expectation of F_{ST} was the same in all of the simulations shown in Figure 1, and the average $\widehat{F}_{ST}$ was very close to the expectation in all cases. However, low recombination rates led to longer-tailed distributions of $\widehat{F}_{ST}$ and higher rates led to distributions that were tighter about the mean (Figure 1). Similar patterns are seen under a variety of demographic models (Supplementary Figure 2).

A typical method for determining outliers in genome scans, used by all the empirical studies cited in the introduction, 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 $\widehat{F}_{ST}$ to be enriched for analysis windows in low recombination regions under neutrality (Figure 1). Indeed, the

144 recombination rates shown in Figure 1 obviously lead to different distributions of $\widehat{F}_{ST}$, but the
 145 extent to which this will affect the analysis of real organisms will depend on the pattern of variation
 146 in recombination rates across their genomes.
 147

148
 149 **Figure 1.** The distribution of F_{ST} calculated in 10 kbp windows for an island population with $N_e = 100$
 150 haploids per deme. F_{ST} was averaged using the method of Weir and Cockerham (1984). The dashed
 151 vertical line is the theoretical expectation of F_{ST} for the simulated population, and the coloured
 152 vertical lines indicate the means for each set of simulations.
 153

154 We examined the relationship between $\widehat{F}_{ST}$ and recombination rate across the *D. melanogaster*
 155 genome using simulated and empirical datasets. We simulated the entire *D. melanogaster* genome
 156 under a model of isolation-with-migration incorporating recombination rate variation using
 157 *stdpopsim* (Adrion et al. 2019) and performed genome scans on the resulting data (Figure 2). We
 158 performed a similar genome scan on the population genomic data from North American *D.*
 159 *melanogaster* populations generated by Reinhardt et al. (2014). In both cases, we found a
 160 significant enrichment of outliers at low recombination rates compared to high recombination rates

when we defined outliers based on the 95th percentile of $\widehat{F}_{ST}$ (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 $\widehat{F}_{ST}$ and recombination rate (Figure 2C; simulated data, Kendall's $\tau = -0.746$, p -value $< 10^{-15}$; empirical data, Kendall's $\tau = -0.193$, p -value = 0.00445). We found qualitatively similar patterns in our simulated data when analysis windows were based on a fixed number of SNPs (Supplementary Figure 3). Of course, the genomic landscape of $\widehat{F}_{ST}$ in North American *D. melanogaster* has likely been shaped by processes other than just migration and genetic drift. Indeed, Figure 2A shows at least one $\widehat{F}_{ST}$ outlier at high recombination rates, which our simulations suggest would be unlikely under neutrality (Figure 2B). In summary, our findings suggest that applying fixed thresholds to the genome-wide distribution of $\widehat{F}_{ST}$ may enrich for false positives in low recombination regions and be overly conservative in highly recombining regions of the genome.

Figure 2. The plot of $\widehat{F}_{ST}$ calculated in 10,000bp analysis windows against local recombination rate for *Drosophila melanogaster* from A) the North American populations of *Drosophila melanogaster* analysed by Reinhardt et al. (2014) and B) coalescent simulations of the genome under a model of isolation with migration. C) Shows the variance in $\widehat{F}_{ST}$ against recombination rate for both empirical and simulated datasets. The horizontal lines in A and B are the 95th percentiles of the genome-wide distribution in each case

Previous studies have shown greater heterogeneity in F_{ST} in regions of low recombination rate across the genome, but these have interpreted their findings purely in terms of linked selection

(e.g. Duranton et al. 2018; Gagnaire et al. 2018; Samuk et al. 2017). Indeed, the effects of selection at linked sites, particularly background selection, have been invoked to explain the role of recombination in generating patterns of variation in F_{ST} , and other summary statistics across the genome, in particular the so-called “islands of differentiation” involved in speciation (Adrion et al., 2019; Burri, 2017; Cruickshank & Hahn, 2014; Nosil et al., 2009). In this study, we are suggesting that genome-scans using analysis windows should incorporate recombination rate variation into the null model used to interpret the data, even in the absence of selection. Obviously, linked selection, which has the greatest impact when recombination rates are low, should be considered as well, but without considering the statistical artefacts that arise under neutrality, it will be difficult to distinguish pattern and process from genome-scan data.

It has long been known that the variance of summary statistics such as the number of segregating sites responds to variation in recombination rate (Wakeley, 2009), so the difficulty of comparing analysis windows of discrete physical size across varying recombination rates will extend to other data summaries that capture variation in the coalescent process. For example, genome-scan studies may analyse nucleotide diversity between populations (D_{XY}), which is proportional to T_B , the coalescence times between populations, within population diversity (π_w), Tajima’s D (Tajima, 1989) or the haplotype based statistics developed by Garud et al. (2015), for instance H12. We calculated each of those statistics from our simulated data and, as expected, the variance of each increased with recombination rate (Supplementary Figure 4). Genome-scan studies often examine combinations of summary statistics that respond differently to various evolutionary processes, but even under neutrality extreme values for multiple summary statistics may occur (Supplementary Figure 5). Without an understanding of what constitutes an extreme value of a given summary statistic for a particular recombination rate, interpreting the genome-wide distribution may be difficult.

209

210 There are obviously cases, however, where the heterogeneity in genome-scans introduced by
 211 recombination rate variation would not affect the ability to determine the loci involved in
 212 evolutionary processes. When there is a monogenic or oligogenic architecture of adaptation or
 213 speciation, the regions involved may clearly stand out from the genomic landscape of summary
 214 statistics. For instance, Jones et al. (2018) performed a genome-scan comparing snowshoe hares
 215 with white and brown winter coats and found an analysis window containing the *Agouti* gene had
 216 by far the highest between-population divergence genome-wide. However, in their study, Jones et
 217 al. (2018) also drew on orthogonal evidence from a genome-wide association study and gene
 218 expression data to support their conclusions about that region's involvement in adaptation. In cases
 219 where processes such as local adaptation involve a highly polygenic architecture, the effects on
 220 genome-wide variation may be more subtle and difficult to interpret (Latta, 1998; Le Corre &
 221 Kremer, 2012; Yeaman, 2015).

222

223 The results from this study should emphasize the need to incorporate recombination into the null
 224 model used to interpret genome scans. Generating recombination maps is a large undertaking, and
 225 there are numerous systems in which it is not feasible to obtain them, but new methods for
 226 estimating recombination rate may greatly facilitate this (Dreau et al. 2019; Sun et al. 2019). In
 227 cases where accurate recombination rate estimates are not available, one could perhaps
 228 determine analysis windows based on patterns of average linkage disequilibrium across sites.
 229 Alternatively, if accurate recombination rate estimates were available, outlier thresholds could be
 230 determined for discrete recombination rate bins, or analysis windows could be designed using fixed
 231 genetic, rather than physical, distances. Further work is required to develop and test analysis
 232 methods that are robust to recombination rate variation. Without accounting for recombination

233 rate variation in some way, however, it will be difficult to interpret variation in the genomic
234 landscape of summary statistics.

235

236 Materials and Methods

237 Coalescent simulations of structured populations were performed using *msprime* v0.7.3 (Kelleher et
238 al. 2016). We simulated an island model of 100 demes, each with $N_e = 100$ haploid individuals.

239 Migration was constant between all demes and was set to $M = N_e m = 2$, giving an expected F_{ST} of $\frac{1}{9}$.

240 After simulating a particular population, we added mutations to the resulting tree sequence at a
241 rate of 10^{-6} . We extracted 100 individuals each from two demes then calculated the following
242 summary statistics using *scikit-allel* v1.2.1: the weighted average of F_{ST} using the method of Weir
243 and Cockerham (1984); between population genetic diversity (D_{XY}); Tajima's D (Tajima, 1989); and
244 Garud's H_{12} (Garud et al. 2015). For Garud's H_{12} , we only examined the first 400 SNPs present in
245 the simulated regions. We added mutations and calculated summary statistics 100 times for each
246 simulated population and took the average across replicates. We performed 1,000 such simulations
247 for each recombination rate tested. Smoothed density plots were generated with default settings in
248 *ggplot2* (Wickham, 2016).

249

250 We re-analysed allele frequency data from the North American *Drosophila* populations generated
251 by Reinhardt et al. (2014). In that study, pooled sequencing was used to estimate derived allele
252 frequencies for pools of 16 isofemale lines each from Maine and Florida. To obtain accurate allele
253 frequencies we excluded any SNP that was reported at a depth greater than or less than one
254 standard deviation from the mean coverage.

255

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

262

263 We analysed the genome-wide distribution of F_{ST} in the empirical and simulated *Drosophila*
 264 *melanogaster* datasets. Because the Reinhardt et al. (2014) data was generated using pooled-
 265 sequencing, we calculated F_{ST} for each polymorphism and averaged across sites using the formulae
 266 for haploids given in Weir (1990). $\widehat{F}_{ST}$ was calculated in 10,000 bp non-overlapping windows using
 267 the ratio-of-averages approach. Recombination rates from the Comeron et al. (2012) map were
 268 obtained using the *Drosophila* recombination rate calculator ([http://petrov.stanford.edu/cgi-](http://petrov.stanford.edu/cgi-bin/recombination-rates_updateR5.pl)
 269 [bin/recombination-rates_updateR5.pl](http://petrov.stanford.edu/cgi-bin/recombination-rates_updateR5.pl); Fiston-Lavier et al. 2010). For both the simulated and
 270 empirical *D. melanogaster* data, we only examined the “normally recombining” regions of the
 271 *Drosophila* genome, following Reinhardt et al. (2014). We divided the genome into 75 equally sized
 272 bins based on recombination rate. We calculated Kendall's correlation between the variance of $\widehat{F}_{ST}$
 273 and the mean recombination rate in each bin. We classified analysis windows as outliers if $\widehat{F}_{ST}$ was
 274 greater than the 95th percentile genome-wide. We examined the analysis windows with the 1,000
 275 largest and smallest recombination rate estimates and used Fisher's exact test to determine
 276 whether there was a significant enrichment of outliers in low recombination regions. We excluded
 277 analysis windows from the simulated data that had recombination rates of 0.

278

Acknowledgements

Thanks to Sally Otto, Andra Thomaz and Wouter van der Bijl for helpful discussions and to Reto Burri, Katie Lotterhos and Matt Pennell for comments on the manuscript. Thanks to Andrew Kern for insights and for providing the *Drosophila* data and to Takeshi Kawakami for providing the flycatcher recombination rate estimates. This study is part of the CoAdapTree project which is funded by Genome Canada (241REF), Genome BC and 16 other sponsors (<http://coadapttree.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.

Bibliography

- Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., ... Kern, A. D. (2019). A community-maintained standard library of population genetic models. *BioRxiv*, 2019.12.20.885129.
- Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, 23(9), 1514–1521.
- Burri, R. (2017). Interpreting differentiation landscapes in the light of long-term linked selection. *Evolution Letters*, 1(3), 118–131.
- Comeron, J. M., Ratnappan, R., & Bailin, S. (2012). The Many Landscapes of Recombination in *Drosophila melanogaster*. *PLoS Genetics*, 8(10), e1002905.
- Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.*, 23(13), 3133–3157.
- Duranton, M., Allal, F., Fraïsse, C., Bierne, N., Bonhomme, F., & Gagnaire, P. A. (2018). The origin and remolding of genomic islands of differentiation in the European sea bass. *Nature Communications*, 9(1), 1–11.
- Fiston-Lavier, A. S., Singh, N. D., Lipatov, M., & Petrov, D. A. (2010). *Drosophila melanogaster* recombination rate calculator. *Gene*, 463(1–2), 18–20.
- Gagnaire, P.-A., Lamy, J.-B., Cornette, F., Heurtebise, S., Dégremont, L., Flahauw, E., ... Lapègue, S. (2018). Analysis of Genome-Wide Differentiation between Native and Introduced Populations of the Cupped Oysters *Crassostrea gigas* and *Crassostrea angulata*. *Genome Biology and Evolution*, 10(9), 2518–2534.
- Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet*, 11(2), e1005004.
- Haas, R. J., & Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, 25(1), 5–23.
- Harpak, A., Garud, N., Rosenberg, N. A., Petrov, D. A., Combs, M., Pennings, P. S., & Munshi-South, J. (2020). Genetic Adaptation in New York City Rats. *BioRxiv*, 2020.02.07.938969. <https://doi.org/10.1101/2020.02.07.938969>
- Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., ... Whitlock, M. C. (2016). Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. *Am. Nat.*, 188(4), 379–397.
- Holsinger, K. E., & Weir, B. S. (2009, September). Genetics in geographically structured populations: Defining, estimating and interpreting FST. *Nature Reviews Genetics*, Vol. 10, pp. 639–650.
- Jones, M. R., Scott Mills, L., Alves, P. C., Callahan, C. M., Alves, J. M., Lafferty, D. J. R., ... Good, J. M. (2018). Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science*, 360(6395), 1355–1358.
- Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., Mugal, C. F., ... Ellegren, H. (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.
- Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Comput. Biol.*, 12(5), e1004842.
- Kong, A., Thorleifsson, G., Gudbjartsson, D. F., Masson, G., Sigurdsson, A., Jonasdottir, A., ... Stefansson, K. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319), 1099–1103.
- Lai, Y. T., Yeung, C. K. L., Omland, K. E., Pang, E. L., Hao, Y., Liao, B. Y., ... Li, S. H. (2019). Standing genetic variation as the predominant source for adaptation of a songbird. *Proceedings of the*

- 352 *National Academy of Sciences of the United States of America*, 116(6), 2152–2157.
- 353 Latta, R. G. (1998). Differentiation of allelic frequencies at quantitative trait loci affecting locally
- 354 adaptive traits. *The American Naturalist*, 151(3), 283–292.
- 355 Le Corre, V., & Kremer, A. (2012). The genetic differentiation at quantitative trait loci under local
- 356 adaptation. *Molecular Ecology*, 21(7), 1548–1566.
- 357 Lewontin, R. C., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the
- 358 selective neutrality of polymorphisms. *Genetics*, 74(1), 175–195.
- 359 Ma, T., Wang, K., Hu, Q., Xi, Z., Wan, D., Wang, Q., ... Liu, J. (2017). Ancient polymorphisms and
- 360 divergence hitchhiking contribute to genomic islands of divergence within a poplar species
- 361 complex. *Proceedings of the National Academy of Sciences of the United States of America*,
- 362 115(2), E236–E243.
- 363 Nachman, M. W., & Payseur, B. A. (2012, February 5). Recombination rate variation and speciation:
- 364 Theoretical predictions and empirical results from rabbits and mice. *Philosophical Transactions*
- 365 *of the Royal Society B: Biological Sciences*, Vol. 367, pp. 409–421.
- 366 Nosil, P., Funk, D. J., & Ortiz-Barrientos, D. (2009, February 1). Divergent selection and
- 367 heterogeneous genomic divergence. *Molecular Ecology*, Vol. 18, pp. 375–402.
- 368 Olofsson, J. K., Dunning, L. T., Lundgren, M. R., Barton, H. J., Thompson, J., Cuff, N., ... Christin, P. A.
- 369 (2019). Population-Specific Selection on Standing Variation Generated by Lateral Gene
- 370 Transfers in a Grass. *Current Biology*, 29(22), 3921–3927.e5.
- 371 Reinhardt, J. A., Kolaczowski, B., Jones, C. D., Begun, D. J., & Kern, A. D. (2014). Parallel geographic
- 372 variation in *Drosophila melanogaster*. *Genetics*, 197(1), 361–373.
- 373 Samuk, K., Owens, G. L., Delmore, K. E., Miller, S. E., Rennison, D. J., & Schluter, D. (2017). Gene flow
- 374 and selection interact to promote adaptive divergence in regions of low recombination.
- 375 *Molecular Ecology*, 26(17), 4378–4390.
- 376 Slatkin, M. (1991). Inbreeding coefficients and coalescence times. *Genetical Research*, 58(2), 167–
- 377 175.
- 378 Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by polymorphism.
- 379 *Genetics*, 123, 585–595.
- 380 Talla, V., Johansson, A., Dincă, V., Vila, R., Friberg, M., Wiklund, C., & Backström, N. (2019). Lack of
- 381 gene flow: Narrow and dispersed differentiation islands in a triplet of *Leptidea* butterfly
- 382 species. *Molecular Ecology*, 28(16), 3756–3770.
- 383 Wakeley, J. (2009). *Coalescent theory : an introduction*. Roberts & Co. Publishers.
- 384 Weir, B S. (1990). Genetic data analysis. Methods for discrete genetic data. *Sinauer Association*,
- 385 *Sunderland, Massachusetts*, 193.
- 386 Weir, Bruce S, & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population
- 387 structure. *Evolution*, 38(6), 1358–1370.
- 388 Wickham, H. (2016). ggplot2: elegant graphics for data analysis. In *Journal of the Royal Statistical*
- 389 *Society: Series A (Statistics in Society)*.
- 390 Wolf, J. B. W., & Ellegren, H. (2017, February 1). Making sense of genomic islands of differentiation
- 391 in light of speciation. *Nature Reviews Genetics*, Vol. 18, pp. 87–100.
- 392 Wright, S. (1937). The Distribution of Gene Frequencies in Populations. *Proceedings of the National*
- 393 *Academy of Sciences*, 23(6), 307–320.
- 394 Yeaman, S. (2015). Local Adaptation by Alleles of Small Effect. *The American Naturalist*, 186 Suppl 1,
- 395 S74–89.
- 396 Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., ... Gaut, B. S. (2019). The
- 397 population genetics of structural variants in grapevine domestication. *Nature Plants*, 5(9),
- 398 965–979.
- 399

Supplementary Material

Supplementary Figure 1

Recombination rate variation in three animal species calculated for 200 kbp genomic windows. The human map comes from Kong et al. (2010)(data downloaded from <https://www.decode.com/addendum/>). The collared flycatcher data come from Kawakami et al. (2014), and the *Drosophila melanogaster* data come from Comeron et al. (2012) and were obtained using the *Drosophila* recombination rate calculator (http://petrov.stanford.edu/cgi-bin/recombination-rates_updateR5.pl; Fiston-Lavier et al. 2010)

Supplementary Figure 2

The distribution of F_{ST} calculated in 10 kbp windows for four differently structured metapopulations. F_{ST} was averaged using the method of Weir and Cockerham. Note, the 100-deme island model is represented in Figure 1 in the main text.

Supplementary Figure 3

The plot of $\widehat{F}_{ST}$ and the variance in $\widehat{F}_{ST}$ against local recombination rate for coalescent simulations of the genome under a model of isolation with migration. The panels show $\widehat{F}_{ST}$ calculated in analysis windows of either a fixed physical size (10,000 bp) or fixed number of SNPs (250 or 50).

Supplementary Figure 4

The distribution of five population genetic summary statistics often used in genome scan studies. All statistics were calculated in 10 kbp windows except for H12 which was calculated using a window of 400 SNPs. F_{ST} was averaged using the method of Weir and Cockerham. Populations were simulated according to a 100-deme island model. Vertical lines indicate the mean of a given distribution.

Supplementary Figure 5

Pairwise comparison of five population genetic summary statistics often used in genome scan studies under of four recombination rates. Each panel shoes the results from 1,000 independent coalescent simulations. All statistics were calculated in 10 kbp windows except for H12 which was calculated using a window of 400 SNPs. F_{ST} was averaged using the method of Weir & Cockerham, (1984). π and D_{XY} are expressed per base pair. Populations were simulated according to a 100-deme island model.