Model-based assessments of differential introgression and linked natural selection during divergence and speciation

Arun Sethuraman, Vitor Sousa, Jody Hey

September 27, 2019

Abstract

Demographic changes such as fluctuating population size and differential introgression can mask the effects of natural selection, and affect rates of genome evolution, local adaptation, reproductive isolation, and eventual speciation. Besides identifying differentially introgressing genes (and genomic regions) that are “labeled” to be retroactively causal to adaptive evolution and speciation, there is significant impetus to understand, and perhaps estimate the underlying demography that affects current genomic diversity. Using model-based likelihood methods to directly estimate, and decouple the effects of differential introgression and demography across genomic loci offers an ideal solution to detect differential introgression, and population demography and build hypotheses around its underlying evolutionary processes. We describe a computationally efficient parallelized implementation of mixture-model based isolation with migration (IM) analyses to assign loci to classes based on their shared coalescent histories (population sizes, or migration rates). We apply this method to several genomic data sets (great apes - chimpanzees and bonobos, *Anopheles* mosquitoes, threespine sticklebacks, Mullerian mimics of *Heliconius* butterflies, mice, European rabbits, and fruitflies), that have been previously characterized (perhaps erroneously) using genome-wide scans of differentiation. We show that we cannot reject a model of differential introgression, or linked selection across a majority of species analyzed, with two species

showing the combined effects of differential introgression and linked natural selection across multiple, non-independent genomic loci.

In recent years, a number of studies have reported findings of the hidden hand of natural selection in contributing to species formation, as revealed by signals differential introgression in which some parts of genomes are exchanged between diverging populations while others are not. (e.g. Duranton et al. [8], Feulner et al. [10], Geraldès et al. [13], Kronforst et al. [16], Martin et al. [18], Ravinet et al. [24], Roux et al. [25], Tavares et al. [33], Turner and Hahn [34], Turner et al. [35]). Although the evidence is indirect, the argument is straightforward; for if it is true that only some portions of the genome are exchanged between diverging populations, then it follows that some selective force is preventing exchange at complementary portions of the genome (see Ravinet et al. [24], Roux et al. [25], Sousa and Hey [28]). A process of speciation that includes natural selection against introgression, is one in which selection at the genic level is acting indirectly as a diversifying force at the species level, and it is a far more complex process than one of simple allopatric divergence. Consequently the many reports of differential introgression during divergence have been used to support the case that our overall understanding of the speciation process is changing in important ways Abbott et al. [1], Feder et al. [9], Seehausen et al. [26], Sousa and Hey [28].

A potential difficulty with interpretations of differential introgression lies in the fact that such conclusions are themselves indirect, as they are usually based upon static patterns of variation observed within and between populations. A typical approach is to perform a genome scan, measuring divergence in windows along the genome, and identifying regions of markedly low or high divergence Cruickshank and Hahn [7], Payseur and Rieseberg [22]. The difficulty in this context is that divergence, by any measure, can be affected by other evolutionary factors besides gene flow. In particular, as some have pointed out Cruickshank and Hahn [7], Noor and Bennett [21], variation in divergence can arise even in the absence of gene exchange as a byproduct of natural selection within populations. For example, genetic divergence between populations is frequently measured using Wright's F_{ST} Wright [40, 41],

which is the proportion of variation that lies between populations. However, as a proportion, F_{ST} can be low either because the ‘between’ portion of the variation is low or because the ‘within’ portion is high. Loci that experience forces that lessen variation within populations, but do not otherwise affect divergence, will exhibit a high F_{ST} , as will loci that have typical polymorphism within populations, but also have an elevated level of divergence. A simpler measure of divergence is the average number of base pair differences between sequences from different populations (frequently referred to as d_{XY} Nei and Miller [19]). Although not a proportion, this measure includes both variation within and between populations and can be elevated or reduced because of changes in either.

Our approach is to adapt a demographic model-based framework to allow for variation in the demographic parameters across the genome, in a way that it can account for the effects of natural selection on a subset of loci in the analysis Barton and Bengtsson [2], Charlesworth [5], Galtier et al. [11], Petry [23], Storz et al. [30]. We use an isolation-with-migration model (Fig. 1) which includes parameters for genetic drift (i.e. effective population size, gene flow, and time of population separation Latter [17], Nielsen and Wakeley [20], Takahata [32], Wakeley and Hey [38]). We extend the method of Sousa et al. [29] that allows for multiple sets of demographic parameters in a mixture model in which loci are grouped on the basis of their inferred demographic parameters, to deal with genome-wide data and perform likelihood ratio tests to infer if variation in introgression is required to explain the data. To improve computational efficiency, we developed a method that distributes the proposed Metropolis Coupled Markov Chain Monte Carlo (MC3) method across processors, providing a nearly linear speed-up in computational times, thus making it possible to perform inference using larger data sets (of the order of hundreds of loci) than previously feasible. This is implemented in the program MigSelect. Analyses of genomic data from seven paradigmatic species pairs across a variety of taxa under this framework reveals evidence of differential introgression and linked selection effects across several species, including new discoveries of differential introgression between species of chimpanzees (*Pan troglodytes verus*

78 and *Pan troglodytes troglodytes*).

79 Results

80 Inference Framework

81 Following the framework developed by Sousa et al. [29], we propose a series of three tests
 82 that together can decouple the effects of differential introgression and linked selection across
 83 loci. For a given data set the process begins by sampling from an MCMC simulation under
 84 a full IM model in which there are two sets of population size parameters *and* two sets of
 85 migration parameters (see Fig. 1). Then the sampled genealogies are used to construct a
 86 function that is an estimate of the joint posterior density for all of these parameters Hey
 87 and Nielsen [15]. Following Sousa et al. [29] we obtained the joint parameter estimate by
 88 optimizing this function while effectively integrating over both assignment vectors (i.e. for
 89 population size parameter sets, and for migration rate parameter sets). Furthermore, under
 90 a uniform prior on demographic parameters, this joint posterior density is proportional to
 91 the likelihood, and the estimated surface can be used for log likelihood ratio (LLR) tests of
 92 nested models. Specifically, we develop three LLR tests - (1) Test of Differential Introgression,
 93 wherein there are two sets of loci with estimably different migration rates (2) Test of Linked
 94 Selection, wherein there are two sets of loci with different effective population sizes, and (3)
 95 Test of Independence of Loci under differential introgression versus linked selection effects.

96 Choice of Species Pairs

97 We analyzed genomic data from a variety of taxa (see Table 1), exhibiting several inter-
 98 esting natural history characteristics, including naturally non-hybridizing species (e.g. *P.*
 99 *paniscus verus* and *Pan troglodytes troglodytes*), species that freely hybridize in the wild
 100 (eg. *Anopheles gambiae* - *A. coluzzii*, formerly M subtype, and *A. gambiae sensu stricto*,
 101 formerly S subtype), and species that exhibit some levels of noted hybrid incompatibility

(eg. *Mus musculus* - hybrid males showing lower fertility, or inviability). Species analyzed also form complexes under a variety of ecological speciation mechanisms - allopatry (eg. *P.p.verus*, *P.t.troglodytes*, and parapatry (e.g. *O.c.cuniculus*, *O.c.algirus*, *Gasterosteus* river versus lake forms). Ten replicate runs of IMA2p (Sethuraman and Hey [27]) and MigSelect were performed, discarding 100,000 MCMC iterations as burn-in, and sampling 10,000 genealogies each. The 100,000 genealogies sampled thus across replicate runs were then pooled and used in computing the marginal posterior density distributions of population mutation rates, migration rates, and divergence times, and in Likelihood Ratio Tests (LRT's) of linked selection, differential introgression, and a test of independence of loci in nested models.

Tests of Significant Migration

Tests of significant migration under two species IM models using IMA2p could not reject the null hypothesis of no significant migration in six out of the eight species pairs analyzed (Table 2). Out of these, sampled loci across *Heliconius* butterflies, *Oryctolagus* rabbits, and *Anopheles* mosquitoes showed signatures of significant bidirectional migration across all loci, whereas the great ape species pairs (*P.paniscus* vs *P.t.troglodytes*, and *P.t.troglodytes* vs *P.t.verus*), and *D. pseudoobscura* vs *D. persimilis* estimate significant unidirectional gene flow. Interestingly, despite previous reports of significant migration between natural populations (see Feulner et al. [10], Harr [14]), our IMA2p analyses do not estimate any significant migration between these species pairs.

Tests of Differential Introgression

Five out of eight species pairs analyzed using MigSelect failed to reject the null hypothesis of no differential introgression effects across their genomes at a significance level of 0.01 (Table 3).

Analyses of the functions of loci enriched for 'low' and 'high' migration rates in each species pair discovered genes that encoded for LIM binding proteins, sodium ion channel

proteins, 3L meiotic protein, and AGAP010317, AGAP010315, which have all been reported to be present on predicted divergence islands (White et al. [39]) in *Anopheles*.

Similar analyses of functional genes in sticklebacks showed that loci classified as having ‘low’ migration rates include large regions between 193kb - 216kb, 281kb - 302kb (including the *slu7*, *kinesin*, *gpcr* genes), 313kb - 379kb (including the highest divergence interval, as identified by Supple et al. [31]), 405kb - 455kb (including the *optix* genic region, which was also previously identified as containing high divergence in the Peruvian Hybrid Zone of *H. melpomene*), and a small 1kb region between 553kb - 554kb of the B/D interval around the *optix* gene (see Fig. 4 of Supple et al. [31]). Our analyses also classified a few loci as displaying ‘high’ migration rates. These include a large region between 16kb - 27kb, a 1kb region between 92kb - 93kb, a large region between 454kb - 532kb, and 555kb - 560kb. The latter two regions are immediately downstream of regions that contain loci that are enriched for “low” migration around the *optix* gene - potentially containing regulatory elements, congruent with the hypothesis of natural selection on color pattern variation in these hybrid zones.

Functional annotation enrichment analyses of loci that are putatively under differential introgression effects in great ape species pairs (79 out of 200 loci in *P. paniscus* vs *P. t. troglodytes*) discovered several ‘hits’ to GO terms mapped to the *hg38* human genome annotation (see Supplemental Table:...). In *D. pseudoobscura* vs *D. persimilis*, most loci classified as having ‘high’ migration rates occurred around a chromosome 2 inversion, suggesting support for gene flow for loci within this inversion, as also reported by Machado et al. 2007.

Tests of Linked Selection

Tests of linked natural selection also failed to reject the null hypothesis of no linked selection effects (i.e. two different groups of loci with different effective population sizes) in six of eight species pairs analyzed (Table 4) at a significance level of 0.01.

Across *Oryctolagus* species pair loci, MigSelect analyses show a clear differentiation of loci into three classes of loci - those with high population size, and high migration rate (A31-GBE,A37-EXT1,A38-LUM), loci with high migration rate, and low population sizes (X3-MAOA,X16-DIAPH2), and with low migration rates, and low population sizes (X23-SHOX,X19-G6PD,X18-FMR1,X15-AMOT,X17-F9,A43-STAG1), consistent with the observations of Carneiro et al. [4] and Sousa et al. [29].

Analyses of linked selection effects in *Gasterosteus* species pair loci discovered 20 out of the 200 loci to be classified as having ‘higher’ effective population size than others (see Supplemental Table ...). Among these, was Locus2 (GroupI:2150001-2160000), which is about 100kb upstream of the *ATP1a1* gene, which is known to be gene involved in osmotic regulation (and previously identified as an F_{ST} outlier by Feulner et al. [10] across all populations analyzed. As reported by Feulner et al. [10], the *Eda* gene or its immediate neighboring regions were not enhanced for linked selection effects. Also of note are two nearly consecutive regions on linkage groupXX(17730001-17740000,17760001-17770000) which are also classified as being under the effects of linked natural selection effects. These regions contain exons belonging to the *Lin28* and *MTF1* genes, which are implicated in *IGF2* translational regulation, and in metal homeostasis respectively.

In the *Mus musculus* species pair, seven out of the fourteen loci analyzed (*Ppp1r3b*, *Sec63*, *Nr1d2*, *Wdt2*, *Bnc1*, *XP_620246*, *Ggh*, *Melk*) were classified as loci with higher effective population sizes, than the others. These loci had a wide range of variability in estimated F_{st} (0.13 (*Ggh*)-0.72(*Sec63*) (see Supplementary Table 1 of Cruickshank and Hahn [7]), further exhibiting the ‘confounding’ effects of demography on estimates of differentiation or divergence.

Analyses of *P.t.verus* versus *P.t.troglodytes*, which estimated signatures of linked natural selection effects identified 14 loci (out of 200) that were classified to have greater effective sizes than others, and thus potentially under the effects of linked natural selection. Identifiable human orthologs to these ‘outlier’ loci include those involved in G-protein receptor functions,

fukutin, and membrane protein synthesis.

In the *Drosophila* species pair, loci classified as being enriched for ‘low’ effective population size occurring around a chromosome 2 inversion (9Mb-17Mb) as reported by Machado et al. 2007 and McGaugh and Noor 2012. Additional loci with ‘low’ effective sizes were also identified on chromosomes 3 and 4.

Tests of Independence of Genomic Loci

A Fisher’s exact test of independence between loci that were classified as under differential introgression effects versus those under linked selection effects failed to reject the null hypothesis of independence in two species pairs - *Anopheles* and *Drosophila* (Table 5). In the *Anopheles* complex, one locus (cytochrome P450 gene) on the X chromosome was determined to enriched for low population size, and high migration rate, indicative of possible adaptive introgression on the X chromosome (also shown to have low differentiation - $\pi < 0.001$ by Turner et al. [35]).

In the *Drosophila* species complex, we see a consistent overlap between loci that are enriched for ‘high’ migration rates and those with ‘low’ effective population size around a chromosome 2 inversion, suggesting support for gene flow for loci within this inversion, as also reported by Machado et al. 2007. Loci classified as having ‘low’ effective sizes on chromosomes 3 and 4 were also enhanced for ‘high’ migration rates. Several loci on chromosome 4 (group 3) were enhanced for ‘high’ effective sizes, which contained loci overlapping with having been enhanced for ‘low’ migration rates. We hypothesize that these loci are potentially under the effects of balancing selection in each species, and could potentially represent speciation islands.

Discussion

Adaptive evolutionary history post-isolation is often complicated by the effects of differential migration and linked natural selection between diverged populations of a species. Tradi-

tional methods of quantifying diversity and differentiation across genomic regions, including genome-wide scans of F_{ST} , d_{XY} , and similar statistics fail to capture these differences. As pointed out by Cruickshank and Hahn [7], all measures of divergence and differentiation can be affected by other evolutionary factors besides gene flow. For example, background selection, or negative selection on deleterious alleles (perhaps from the migrant population) at genomic loci can result in reduced absolute or relative diversity within populations, and thus increased divergence or differentiation between populations (Charlesworth et al. [6], Noor and Bennett [21], Cruickshank and Hahn [7]). Similarly, positive selection and selective sweeps can cause reduction in absolute or relative genomic diversity at linked neutral sites within populations, and increased relative divergence between populations (Via [36], Via and West [37]). There is as yet no consensus on which measure (relative or absolute) is more powerful in detecting loci experiencing either differential migration or linked selection effects (Cruickshank and Hahn [7], Geneva et al. [12]). Numerous lines of evidence also point to the role of recombination in maintenance of genomic diversity in the face of gene flow post-divergence (Noor and Bennett [21], Burri et al. [3]). Regions of low recombination (eg. pericentromeric, or telomeric regions) can harbor low genomic diversity due to hitchhiking, or background selection, thus portending greater detected differentiation and divergence between populations. Similarly, regions of chromosomal rearrangements exhibit lower diversity owing to incomplete lineage sorting of rearranged sections (Noor and Bennett [21]). It is thus difficult to identify the cause of variation in genomic diversity or divergence using a genome scan of summary statistics.

We present a model-based approach and a series of tests of linked selection, differential introgression, and a test of independence between the two effects to help tease these effects apart. We demonstrate its utility by analyzing simulated and empirical genomic data sets of varying sizes and natural histories. Implemented using a parallelized framework (using `OpenMPI`), `MigSelect` is fast (data not shown, but see Sethuraman and Hey [27], and scales easily to hundreds of genomic loci with computing power. The software is now freely available

for academic use at www.github.com/arunsethuraman/MigSelect/.

Methods

Test of Linked Selection Effects

If directional selection (positive or negative) is systematically shaping polymorphism and divergence patterns at some loci, then these loci will appear to have altered effective population sizes, relative to loci that are not similarly affected, and sampled assignment vectors for population size parameters will tend to show a grouping of these loci. Finally a comparison of log likelihoods, for the full model (that includes two sets of population size parameters) against the nested model in which all loci share the same population size parameters, is expected to reveal a large difference.

We denote the full, or alternate model, H_a , as the full set of parameters:

$H_a = \{N_0^A, N_1^A, N_a^A, N_0^B, N_1^B, N_a^B, m_{0 \rightarrow 1}^C, m_{1 \rightarrow 0}^C, m_{0 \rightarrow 1}^D, m_{1 \rightarrow 0}^D\}$, where A and B are the two sets of population size parameters, and C and D are the two sets of migration rate parameters. And we denote the null model H_0 as having just a single set of population size parameters: $H_0 = \{N_0, N_1, N_a, m_{0 \rightarrow 1}^C, m_{1 \rightarrow 0}^C, m_{0 \rightarrow 1}^D, m_{1 \rightarrow 0}^D\}$. Then from a single set of genealogies sampled under the full model, we calculate the LLR statistic, $L(X)$, as twice the difference between maximized log values from two joint density functions.

$$L(X) = -2 \times \Delta$$

$$\approx -2 \times \left(\log_{max} \left(\frac{1}{n} \sum_{i=1}^n \frac{\pi_0(G^i | H_0)}{\pi_a(G^i)} \right) - \log_{max} \left(\frac{1}{n} \sum_{i=1}^n \frac{\pi_a(G^i | H_a)}{\pi_a(G^i)} \right) \right)$$

Here G^i is the i^{th} genealogy sampled from the posterior, $G^i \sim P(X|H_a)$, and π_a denotes a prior probability calculated by analytic integration over the demographic parameters in H_a (Hey and Nielsen [15]).

Test of Differential Introgression

If the fate of genes that introgress from one population to another, varies systematically, such that some portions of the genome experience different effective rates of gene flow than others, then we expect to observe a large difference in log likelihoods for models that do, and do not, include multiple sets of migration parameters. The test of differential introgression is very similar to that for linked selection effects, with the full model, H_a , being the same. However in the case of differential introgression the null model includes only one set of migration parameters: $H_0 = \{N_0^A, N_1^A, N_a^A, N_0^B, N_1^B, N_a^B, m_{0 \rightarrow 1}, m_{1 \rightarrow 0}\}$, where A and B are the two sets of population size parameters. The LLR statistic $L(x)$ is then similarly computed using Eqn. .

Test of independence of loci affected by differential introgression and linked selection effects

We consider a third test, which determines if there are signatures of evolutionary processes that simultaneously affect population sizes, and migration rates (and hence linked selection, and differential introgression effects) between sampled populations. Consider the null hypothesis (H_0) that the groups of loci that are differentially introgressing are independent of loci that exhibit linked selection effects. Alternately, consider the H_a that loci that are differentially introgressing and the loci that have linked selection effects are non-independent. We use a Fisher's exact test of independence with contingency tables of assignments of loci, computed across all sampled assignment vectors in the method of Sousa et al. [29].

Empirical Datasets

Data from all previous studies that utilized targeted sequencing (< 100 loci) were aligned, phased, pruned to exclude recombinant, and invariant segments, prior to demographic analyses. In data sets where whole-genomic data was available, variant calls were extracted, genome-wide estimates of relative differentiation (F_{ST}) were computed, filtered by genotyp-

ing/sequencing quality, and loci sampled uniformly from across this distribution by ranking F_{ST} values. Sampled loci were then subject to the same IM model assumption checks as before, and for most pairs of species, 200 representative loci were used in tests of linked selection and differential introgression.

All data sets, and scripts to produce them have been deposited with Dryad (DOI:...).

Acknowledgments

This work was supported by NSF Grant:1564659 to AS and JH, NIH Grant:R01GM078204 to JH.

Figure 1: (A) Depiction of a typical Isolation with Migration (IM) model of Nielsen and Wakeley [20], which describes an ancestral population of size N_a splitting into two populations of sizes N_1 and N_2 at time t_s . The two populations subsequently continue to exchange genes at rates m_{12} and m_{21} from population 1 into 2 and vice versa respectively. (B) This model describes a scenario where there are two groups of loci, 1 and 2, with different migration rate distributions ($m_{12,1}, m_{21,1}, m_{12,2}, m_{21,2}$). Note that this model only contains one group of effective population sizes, N_1, N_2 , and N_a . (C) This depicts a scenario where there are two groups of loci with different effective population size distributions, 1 and 2 ($N_{1,1}, N_{2,1}, N_{a,1}$ and $N_{1,2}, N_{2,2}, N_{a,2}$). This model also only parametrizes one group of migration rates, m_{12} and m_{21} . A full model which nests all models (A, B, and C) would in turn have two groups of loci for both effective population sizes, and migration rates, and is shown in panel D.

Figures and Tables

References

- [1] Abbott, R., D. Albach, S. Ansell, J. W. Arntzen, S. J. E. Baird, N. Bierne, J. Boughman, A. Brelsford, C. A. Buerkle, R. Buggs, R. K. Butlin, U. Dieckmann, F. Eroukhmanoff, A. Grill, S. H. Cahan, J. S. Hermansen, G. Hewitt, A. G. Hudson, C. Jiggins, J. Jones, B. Keller, T. Marczewski, J. Mallet, P. Martinez-Rodriguez, M. Möst, S. Mullen, R. Nichols, A. W. Nolte, C. Parisod, K. Pfennig, A. M. Rice, M. G. Ritchie, B. Seifert, C. M. Smadja, R. Stelkens, J. M. Szymura, R. Vaimola, J. B. W. Wolf and D. Zinner.

2013. Hybridization and speciation. *Journal of Evolutionary Biology* **26**:229–246. URL
<http://dx.doi.org/10.1111/j.1420-9101.2012.02599.x>.
- [2] Barton, N. and B. O. Bengtsson. 1986. The barrier to genetic exchange between hybridizing populations. *Heredity* **56**:357–376. Models multilocus gene flow across a hybrid barrier; ; ; xerox,.
- [3] Burri, R., A. Nater, T. Kawakami, C. F. Mugal, P. I. Olason, L. Smeds, A. Suh, L. Dutoit, S. Bureš, L. Z. Garamszegi et al. 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome research* gr-196485.
- [4] Carneiro, M., J. A. Blanco-Aguilar, R. Villafuerte, N. Ferrand and M. W. Nachman. 2010. Speciation in the european rabbit (*Oryctolagus cuniculus*): islands of differentiation on the x chromosome and autosomes. *Evolution* **64**:3443–60. URL
<http://www.ncbi.nlm.nih.gov/pubmed/20666840>.
- [5] Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**:195–205. URL <http://dx.doi.org/10.1038/nrg2526>.
10.1038/nrg2526.
- [6] Charlesworth, B., M. Nordborg and D. Charlesworth. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetics Research* **70**:155–174.
- [7] Cruickshank, T. E. and M. W. Hahn. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol* **23**:3133–57. URL <http://www.ncbi.nlm.nih.gov/pubmed/24845075>.
- [8] Duranton, M., F. Allal, C. Fraïsse, N. Bierne, F. Bonhomme and P.-A. Gagnaire. 2018. The origin and remolding of genomic islands of differentiation in the european sea bass. *Nature communications* **9**:2518.

- [9] Feder, J. L., S. M. Flaxman, S. P. Egan, A. A. Comeault and P. Nosil. 2013. Geographic mode of speciation and genomic divergence. *Annual Review of Ecology, Evolution, and Systematics* **44**:73–97.
- [10] Feulner, P. G., F. J. Chain, M. Panchal, Y. Huang, C. Eizaguirre, M. Kalbe, T. L. Lenz, I. E. Samonte, M. Stoll, E. Bornberg-Bauer, T. B. Reusch and M. Milinski. 2015. Genomics of divergence along a continuum of parapatric population differentiation. *PLoS Genet* **11**:e1004966. URL <http://www.ncbi.nlm.nih.gov/pubmed/25679225>.
- [11] Galtier, N., F. Depaulis and N. H. Barton. 2000. Detecting bottlenecks and selective sweeps from dna sequence polymorphism. *Genetics* **155**:981–987. URL <http://www.genetics.org/content/genetics/155/2/981.full.pdf>.
- [12] Geneva, A. J., C. A. Muirhead, S. B. Kingan and D. Garrigan. 2015. A new method to scan genomes for introgression in a secondary contact model. *PLoS One* **10**:e0118621. URL <http://www.ncbi.nlm.nih.gov/pubmed/25874895>.
- [13] Geraldès, A., P. Basset, B. Gibson, K. L. Smith, B. Harr, H. T. Yu, N. Bulatova, Y. Ziv and M. W. Nachman. 2008. Inferring the history of speciation in house mice from autosomal, x-linked, y-linked and mitochondrial genes. *Mol Ecol* **17**:5349–63. URL <http://www.ncbi.nlm.nih.gov/pubmed/19121002>.
- [14] Harr, B. 2006. Genomic islands of differentiation between house mouse subspecies. *Genome Res* **16**:730–7. URL <http://www.ncbi.nlm.nih.gov/pubmed/16687734>.
- [15] Hey, J. and R. Nielsen. 2007. Integration within the felsenstein equation for improved markov chain monte carlo methods in population genetics. *Proc Natl Acad Sci U S A* **104**:2785–90. URL <http://www.ncbi.nlm.nih.gov/pubmed/17301231>.
- [16] Kronforst, M. R., M. E. Hansen, N. G. Crawford, J. R. Gallant, W. Zhang, R. J. Kulathinal, D. D. Kapan and S. P. Mullen. 2013. Hybridization re-

veals the evolving genomic architecture of speciation. Cell Rep **5**:666–77. URL <http://www.ncbi.nlm.nih.gov/pubmed/24183670>.

[17] Latter, B. D. 1973. The island model of population differentiation: a general solution. Genetics **73**:147–57. 0016-6731 Journal Article Maybe first paper on isolation with migration model thought does not use the term.

[18] Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, M. Blaxter, A. Manica, J. Mallet and C. D. Jiggins. 2013. Genome-wide evidence for speciation with gene flow in heliconius butterflies. Genome Research **23**:1817–1828.

[19] Nei, M. and J. C. Miller. 1990. A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. Genetics **125**:873–879.

[20] Nielsen, R. and J. Wakeley. 2001. Distinguishing migration from isolation: a markov chain monte carlo approach. Genetics **158**:885–96. URL <http://www.ncbi.nlm.nih.gov/pubmed/11404349>.

[21] Noor, M. A. F. and S. M. Bennett. 2009. Islands of speciation or mirages in the desert[quest] examining the role of restricted recombination in maintaining species. Heredity **103**:439–444. URL <http://dx.doi.org/10.1038/hdy.2009.151>. Corrigendum that appeared a year later Heredity (2010) 104, 418; doi:10.1038/hdy.2010.13 Correction to: Heredity (2009) 103 , 434–444; doi:10.1038/hdy2009.151 Since the publication of this paper, the authors have noticed that they included a citation to Turner et al . (2005) on page 440, and then said on page 441 ‘each of the empirical studies cited above specifically documented this pattern at least in part using relative divergence measures such as F_{ST} and D_a and interpreted in the context of regions of low recombination facilitating species divergence in the presence of gene flow.’ In fact, the Turner et al. (2005) study used a different metric, and it does further test directly for

the type of artifact we described in our review. We apologize for incorrectly listing our paper in this context.

[22] Payseur, B. A. and L. H. Rieseberg. 2016. A genomic perspective on hybridization and speciation. *Molecular Ecology* **25**:2337–2360. URL <http://dx.doi.org/10.1111/mec.13557>.

[23] Petry, D. 1983. The effect on neutral gene flow of selection at a linked locus. *Theoretical Population Biology* **23**:300–313. URL <http://www.sciencedirect.com/science/article/pii/0040580983900205>.

[24] Ravinet, M., K. Yoshida, S. Shigenobu, A. Toyoda, A. Fujiyama and J. Kitano. 2018. The genomic landscape at a late stage of stickleback speciation: High genomic divergence interspersed by small localized regions of introgression. *PLoS genetics* **14**:e1007358.

[25] Roux, C., C. Fraisse, J. Romiguier, Y. Anciaux, N. Galtier and N. Bierne. 2016. Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS biology* **14**:e2000234.

[26] Seehausen, O., R. K. Butlin, I. Keller, C. E. Wagner, J. W. Boughman, P. A. Hohenlohe, C. L. Peichel, G.-P. Saetre, C. Bank, A. Brannstrom, A. Brelsford, C. S. Clarkson, F. Eroukhmanoff, J. L. Feder, M. C. Fischer, A. D. Foote, P. Franchini, C. D. Jiggins, F. C. Jones, A. K. Lindholm, K. Lucek, M. E. Maan, D. A. Marques, S. H. Martin, B. Matthews, J. I. Meier, M. Most, M. W. Nachman, E. Nonaka, D. J. Rennison, J. Schwarzer, E. T. Watson, A. M. Westram and A. Widmer. 2014. Genomics and the origin of species. *Nat Rev Genet* **15**:176–192. URL <http://dx.doi.org/10.1038/nrg3644>.

[27] Sethuraman, A. and J. Hey. 2015. Ima2p - parallel mcmc and inference of ancient

demography under the isolation with migration (im) model. Mol Ecol Resour URL
<http://www.ncbi.nlm.nih.gov/pubmed/26059786>.

[28] Sousa, V. and J. Hey. 2013. Understanding the origin of species with
genome-scale data: modelling gene flow. Nat Rev Genet **14**:404–14. URL
<http://www.ncbi.nlm.nih.gov/pubmed/23657479>.

[29] Sousa, V. C., M. Carneiro, N. Ferrand and J. Hey. 2013. Identifying loci under selec-
tion against gene flow in isolation-with-migration models. Genetics **194**:211–33. URL
<http://www.ncbi.nlm.nih.gov/pubmed/23457232>.

[30] Storz, J. F., M. A. Beaumont and S. C. Alberts. 2002. Genetic evidence for long-
term population decline in a savannah-dwelling primate: Inferences from a hier-
archical bayesian model. Molecular Biology and Evolution **19**:1981–1990. URL
<http://mbe.oxfordjournals.org/content/19/11/1981.abstract>.

[31] Supple, M. A., H. M. Hines, K. K. Dasmahapatra, J. J. Lewis, D. M. Nielsen, C. Lavoie,
D. A. Ray, C. Salazar, W. O. McMillan and B. A. Counterman. 2013. Genomic archi-
tecture of adaptive color pattern divergence and convergence in heliconius butterflies.
Genome research gr. 150615.112.

[32] Takahata, N. 1995. A genetic perspective on the origin and history of humans. Annual
Review of Ecology and Systematics **26**:343–372. Review applications of coalescent and
mixed model estimates of historical effective population sizes. Lots of estimates.; ; ; xrx,
discuss both gene flow and divergence time in divergence process a good early reference
for IM model (though does not say Isolation with Migration).

[33] Tavares, H., A. Whibley, D. L. Field, D. Bradley, M. Couchman, L. Copsey, J. Elleouet,
M. Burrus, C. Andalo, M. Li et al. 2018. Selection and gene flow shape genomic islands
that control floral guides. Proceedings of the National Academy of Sciences **115**:11006–
11011.

- 416 [34] Turner, T. L. and M. W. Hahn. 2007. Locus- and population-specific selection and
417 differentiation between incipient species of *Anopheles gambiae*. *Mol Biol Evol* **24**:2132–
418 8. URL <http://www.ncbi.nlm.nih.gov/pubmed/17636041>.
- 419 [35] Turner, T. L., M. W. Hahn and S. V. Nuzhdin. 2005. Genomic is-
420 lands of speciation in *Anopheles gambiae*. *PLoS Biol* **3**:e285. URL
421 <http://www.ncbi.nlm.nih.gov/pubmed/16076241>.
- 422 [36] Via, S. 2012. Divergence hitchhiking and the spread of genomic isolation during ecolog-
423 ical speciation-with-gene-flow. *Philos Trans R Soc Lond B Biol Sci* **367**:451–60. URL
424 <http://www.ncbi.nlm.nih.gov/pubmed/22201174>.
- 425 [37] Via, S. and J. West. 2008. The genetic mosaic suggests a new role
426 for hitchhiking in ecological speciation. *Mol Ecol* **17**:4334–45. URL
427 <http://www.ncbi.nlm.nih.gov/pubmed/18986504>.
- 428 [38] Wakeley, J. and J. Hey. 1998. Testing speciation models with DNA sequence data,
429 157–175. Birkhäuser Verlag, Basel. Early reference for isolation with migration models.
- 430 [39] White, B. J., C. Cheng, F. Simard, C. Costantini and N. J. Besansky. 2010. Genetic
431 association of physically unlinked islands of genomic divergence in incipient species of
432 *Anopheles gambiae*. *Molecular Ecology* **19**:925–939.
- 433 [40] Wright, S. 1922. Coefficients of inbreeding and relationship. *American Naturalist*
434 **56**:330–338. Early reference for F statistics though it does not use that notation have
435 pdf Fst.
- 436 [41] ———. 1950. Genetical structure of populations. *Nature* **166**:247–49.

Species	Pop 1	Pop 2	Loci	Citation	F_{st}^{Low}	F_{st}^{High}
<i>Oryctolagus cuniculus</i>	<i>cuniculus</i>	<i>algirus</i>	44	Sousa et al. 2013	-0.04	0.92
<i>Anopheles coluzzii</i>	<i>coluzzii</i>	<i>gambiae</i>	36	Turner and Hahn 2007, Turner et al. 2005	0.02	0.92
<i>Gasterosteus aculeatus</i> - Germany	Lake	Riverine	200	Feulner et al. 2015	-0.08	0.73
<i>Heliconius melpomene</i>	<i>aglaope</i>	<i>amaryllis</i>	200	Supple et al. 2013	-0.12	0.95
<i>Mus musculus</i>	<i>musculus</i>	<i>domesticus</i>	14	Harr 2006	0.18	0.95
<i>Pan</i>	<i>paniscus</i>	<i>trogodytes</i> <i>trogodytes</i>	200	Prado-Martinez et al. 2013	-0.09	0.46
<i>Pan</i>	<i>trogodytes</i> <i>trogodytes</i>	<i>trogodytes</i> <i>verus</i>	200	Prado-Martinez et al. 2013	-0.09	0.82
<i>Drosophila</i>	<i>pseudoobscura</i>	<i>persimilis</i>	200	Machado, Kingan and Garigan (unpublished)	-0.04	0.85

Table 1: Species pairs analyzed using MigSelect. Where genomic data was available, 200 random intergenic, unlinked, non-recombinant loci were sampled based on binned 10 kbps F_{ST} windows across the genome. Also shown are the ranges of pairwise F_{ST} values, computed across windows using the estimator of Weir and Cockerham (1984). Non-genomic data-sets were curated from GenBank, or Dryad, aligned, pruned, phased, and used in IM analyses.

Species	$m_{0 \rightarrow 1}$ (LLR)	$m_{1 \rightarrow 0}$ (LLR)	Decision
<i>Heliconius melpomene</i> vs <i>aglaope</i>	0.29(p<0.001)	0.48(p <0.001)	Reject H_0
<i>amaryllis</i>			
<i>Oryctolagus cuniculus</i> vs <i>algius</i>	0.37(p < 0.001)	0.36(p < 0.001)	Reject H_0
<i>Anopheles coluzzii</i> vs <i>gambiae</i>	9.944(p<0.001)	5.419(p<0.001)	Reject H_0
<i>Pan paniscus</i> vs <i>troglodytes</i>	0.129(ns)	0.178(p < 0.001)	Reject H_0 for $m_{1 \rightarrow 0}$ (Bonobo→Central)
<i>troglodytes</i>			
<i>Pan troglodytes</i> vs <i>troglodytes</i>	0.000(ns)	2.05 (p < 0.001)	Reject H_0 for $m_{1 \rightarrow 0}$ (Central→Western)
<i>troglodytes</i> vs <i>troglodytes</i>			
<i>Drosophila pseudoobscura</i> vs <i>persimilis</i>	641.585(p<0.001)	0.000(ns)	Reject H_0 for $m_{0 \rightarrow 1}$ (pers.→pseudo.)
<i>Gasterosteus aculeatus</i> Germany Lake vs Riverine	0.06(ns)	0.005(ns)	Cannot reject H_0
<i>Mus musculus</i> vs <i>musculus</i>	0.000(ns)	0.17(ns)	Cannot reject H_0
<i>domesticus</i>			

Table 2: LLR Tests of significant migration between sampled populations using IMA2p, under the H_0 that there is no significant migration between sampled populations, and H_A that there is significant migration (adjudged at a P-value of 0.05). Significant migrations are shown in red.

Species	LLR	P value	Decision
<i>Heliconius melpomene</i> vs <i>aglaope</i>	9.84	0.01	Reject H_0 , accept H_a
<i>Oryctolagus cuniculus</i> vs <i>algericus</i>	23.50	<0.001	Reject H_0 , accept H_a
<i>Pan troglodytes</i> vs <i>trogodytes verus</i>	16.35	<0.001	Reject H_0 , accept H_a
<i>Anopheles coluzzii</i> vs <i>gambiae</i>	22.49	<0.001	Reject H_0 , accept H_a
<i>Drosophila pseudoobscura</i> vs <i>persimilis</i>	12.75	<0.01	Reject H_0 , accept H_a
<i>Gasterosteus aculeatus</i> Germany Lake vs Riverine	2.29	0.32	Cannot reject H_0
<i>Mus musculus</i> vs <i>domesticus</i>	0.25	0.88	Cannot reject H_0
<i>Pan paniscus</i> vs <i>trogodytes</i>	0.47	0.79	Cannot reject H_0

Table 3: Tests of differential introgression using two groups of loci for migration rates, using the H_0 that there is only one group of loci for migration rates versus the H_A that there are two groups of loci. Significance was adjudged at a P-value of 0.05, and significant tests are indicated in red.

Species	LLR	P value	Decision
<i>Gasterosteus aculeatus</i> Germany Lake vs Riverine	15.92	<0.01	Reject H_0 , accept H_a
<i>Mus musculus</i> vs <i>musculus</i> vs <i>domesticus</i>	13.64	<0.01	Reject H_0 , accept H_a
<i>Pan paniscus</i> vs <i>troglodytes</i>	241.50	<0.001	Reject H_0 , accept H_a
<i>Oryctolagus cuniculus</i> vs <i>algericus</i>	128.50	<0.001	Reject H_0 , accept H_a
<i>Anopheles coluzzii</i> vs <i>gambiae</i>	14.25	<0.01	Reject H_0 , accept H_a
<i>Drosophila pseudoobscura</i> vs <i>persimilis</i>	1099.00	<0.001	Reject H_0 , accept H_a
<i>Heliconius melpomene</i> vs <i>aglaope</i>	7.06	0.07	Cannot reject H_0
<i>Pan troglodytes</i> vs <i>troglodytes</i> vs <i>trogodytes verus</i>	5.78	0.12	Cannot reject H_0

Table 4: Tests of linked selection effects using two groups of loci for effective population sizes using the H_0 that there is only one group of loci, versus H_A that there are two groups of loci for effective population sizes. Significance was adjudged at a P-value of 0.05. Species pairs that are significant for tests of linked selection effects are shown in color (red - indicates species that were not significant for tests of differential introgression, blue - species that were significant for tests of differential introgression).

Species	P value	Decision
<i>Anopheles coluzzii</i> vs <i>gambiae</i>	<0.001	Reject H_0
<i>Drosophila pseudoobscura</i> vs <i>persimilis</i>	0.04	Reject H_0
<i>Oryctolagus cuniculus</i> vs <i>algericus</i>	0.06	Cannot reject H_0

Table 5: Fisher’s exact tests of independence of loci that are affected by both differential introgression and linked selection effects. Results are thus only shown for species pairs that were significant for two groups of analyzed loci for migration, and effective population sizes. Significance was adjudged at a P-value of 0.05, and indicated in red.