# Speciation as a Sieve for Ancestral Polymorphism

Rafael F. Guerrero[1,*] and Matthew W. Hahn[1,2]

[1] Department of Biology, Indiana University, 1001 E Third St, Bloomington, IN 47405, USA

[2] School of Informatics and Computing, Indiana University, 1001 E Third St, Bloomington, IN 47405, USA

* Corresponding author: rafguerr@indiana.edu

Keywords: balancing selection, genomic divergence, allopatric speciation, differential gene flow

Running title: Balanced polymorphisms sieved by speciation

Studying the process of speciation using patterns of genomic divergence between species requires that we understand the determinants of genetic diversity within species. Because sequence diversity in an ancestral population determines the starting point from which divergent populations accumulate differences (Gillespie & Langley 1979), any evolutionary forces that shape diversity within species can have a large impact on measures of divergence between species. These forces include those both decreasing (e.g., selective sweeps (Begun *et al.* 2007; Cruickshank & Hahn 2014) or background selection (Phung *et al.* 2016)) and increasing variation (e.g., balancing selection (Charlesworth 2006)). Selection can increase diversity by favoring the maintenance of polymorphism via overdominance, frequency dependence, and heterogeneous selection. Nevertheless, balanced polymorphisms are considered rare in nature, and such loci are often overlooked as major contributors to genome-wide variation in levels of sequence diversity and divergence.

Here, we argue that speciation can act as a "sieve" that will reveal otherwise elusive balancing selection by sorting ancestral balanced polymorphisms unequally across descendent lineages. By sorting alternative alleles between different species, this process uncovers the existence of balancing selection in the ancestral population as regions of higher-than-expected divergence. Sorted ancestral polymorphism may also be responsible for many of the observed peaks of genomic divergence between closely related taxa, confounding studies of differential gene flow and "islands of speciation."

## The effect of balancing selection on genetic diversity

Balancing selection encompasses various, rather disparate, mechanisms that favor the maintenance of polymorphism. Most forms of balancing selection involve heterogeneous or variable selective forces: polymorphism is favored when selection varies across space or time (reviewed in Felsenstein (1976)), between the sexes (reviewed in Otto *et al.* (2011)), or as a function of allele frequency (negative frequency-dependent selection; reviewed in Ayala & Campbell (1974)). Balanced polymorphisms can also occur under constant selective pressures when heterozygotes have a fitness advantage over homozygotes (overdominance; (Wright 1931)).

Despite clear differences in their underlying mechanisms, all forms of balancing selection have qualitatively similar long-term effects on linked neutral variation. Genomic regions closely linked to loci under balancing selection are expected to show increased divergence between the allelic classes defined by the balanced polymorphism. This increase in divergence—the extent of which depends on population size, recombination rate, and strength of balancing selection—results from a population subdivision imposed by the balanced polymorphism. When two (or more) balanced alleles persist for a long time in a population, closely linked regions tend to accumulate differences between the allelic classes. Thus, the genealogies of samples linked to balanced polymorphisms resemble those of structured populations, with allelic classes acting as subpopulations and recombination allowing “gene flow” between these subpopulations. However, when the balanced alleles at a locus are not known, samples cannot be partitioned by allelic class. This makes it difficult to quantify the increased

54 divergence between allelic classes directly (for example, using the statistic $F_{ST}$).
55 Instead, the only observable signals of balanced polymorphism are an excess of
56 intermediate frequency alleles (detected using Tajima's *D*), or overall increases in
57 diversity (detected using the average number of pairwise differences, or π).
58 Unfortunately, the large amount of variance in both *D* and π associated with even
59 neutrally evolving loci means that loci under balancing selection are hard to detect
60 (Simonsen *et al.* 1995).

61 The potential difficulty in detecting balanced polymorphism is somewhat alleviated in
62 cases of local adaptation. Because alternative balanced alleles differ in frequency
63 between the populations where they are individually advantageous, measures of
64 differentiation between populations can be used as a proxy for divergence between
65 allelic classes. Moreover, actual population structure can lower the effective
66 recombination rate between allelic classes, exacerbating their divergence. Perhaps due
67 to its increased detectability, spatially varying selection is reported in natural populations
68 much more frequently than other forms of balancing selection (Asthana *et al.* 2005;
69 Charlesworth 2006; Fan *et al.* 2016). Indeed, while overdominance and negative
70 frequency-dependence continue to be considered rare, local adaptation is considered
71 pervasive, and is even a compulsory first step in some models of speciation (Nosil
72 2012).

73 **The sieve: ancestral lineage sorting after speciation**

74 Consider a simple case of allopatric speciation: a single population is split in two by
75 vicariance (*e.g.*, the rise of a mountain range or the construction of a thousand-mile

wall). For neutral biallelic polymorphisms in the ancestral population, drift will fix alternative alleles in the two nascent species at half of all such loci (the rest of the time both species will fix the same allele). In the presence of balancing selection, expectations can differ. Balancing selection may favor the maintenance of polymorphism in both nascent species (resulting in 'trans-specific polymorphism'; (Muirhead *et al.* 2002)), or increase the chance that alternative alleles are fixed. For instance, when selection varies across space, a geographic barrier is likely to create two unequal ranges (*i.e.*, areas with different proportions of habitats driving local adaptation)—these unequal habitat ranges may favor dramatically different equilibrium frequencies at the locally adapted loci in the nascent species. This could result in selection favoring the fixation of opposite alleles in each species, sieving ancestral alleles in the descendant lineages (Figure 1A).

Sieved balanced polymorphisms carry the signature of selection in the form of increased sequence divergence between the descendant species. Immediately after the split, the level of genetic divergence is largely determined by the diversity present in the ancestral population. Formally, this can be seen in the expectation of absolute divergence: $E(d_{XY})=2\mu t+\theta_{Anc}$ (Gillespie & Langley 1979). As the time since the species split ($t$) approaches zero, $E(d_{XY})$ becomes approximately equal to the ancestral level of diversity, $\theta_{Anc}$ ($=4N_e\mu$ for diploids, where $N_e$ is the effective population size and $\mu$ is the neutral mutation rate). Initial levels of divergence are therefore strongly affected by forces that affect levels of ancestral neutral diversity, such as balancing selection. In regions linked to sieved balanced polymorphism, measures of divergence between

98 species (such as $d_{XY}$ or $F_{ST}$) reflect the divergence accumulated between allelic classes
99 both in the ancestor and since the lineages split, so sieved polymorphisms maintained
100 in the ancestor for a long time can appear as regions of elevated divergence between
101 nascent species.

102 Interestingly, this implies that balancing selection could be more readily detected
103 after speciation (as a sieved polymorphism with elevated $d_{XY}$) than in the ancestor when
104 the causal alleles are unknown (using $D$ or π; Figure 1B). Indeed, the signature of
105 balancing selection is expected to be twice as strong on $d_{XY}$ than on π at loci with alleles
106 maintained at equal frequencies in the ancestor (Figure 1C; see Supplement). As either
107 allele becomes rarer, the effect is more severe: when the minor allele frequency was
108 10% in the ancestor, the increase in $d_{XY}$ is expected to be almost ten times larger than
109 the increase in π. Therefore, by separating the relevant haplotypes, speciation can
110 dramatically increase our power to find loci under balancing selection.

111 **Relevance of sieved polymorphisms during recent genomic divergence**

112 The prevalence of sieved polymorphisms in nature is unknown (and is probably low),
113 but the importance of this phenomenon for observed patterns of genomic divergence
114 does not stem from its frequency. The amount of balanced polymorphism sieved by
115 speciation is proportional to the fraction of loci under balancing selection in the ancestor
116 and the probability of fixing alternative alleles at those loci (which depends on the mode
117 of selection operating at each locus). If balancing selection is as rare as usually
118 assumed (Charlesworth 2006), sieved polymorphisms are likely to be uncommon and
119 would not drastically elevate average levels of divergence. Instead, the few instances of

sieved polymorphisms will have subtle but significant repercussions: these regions will tend to appear at the top of the distribution of divergence across the genome, 'fattening' its upper tail and potentially affecting further inferences. The potential consequences for the distribution of divergence depend not only on the fraction of loci sieved, but also on parameters specific to each polymorphism (namely, its age, strength of selection, and recombination rate). If, for instance, a genome carries only one sieved polymorphism, it will likely appear as a divergence outlier. On the other hand, a higher fraction of sieved regions—caused, for instance, by numerous locally adapted alleles differentiated between populations prior to speciation—will considerably fatten the upper tail in the divergence distribution. Such an observation could be interpreted as evidence of a period of differential gene flow following an initial split (*cf*. Yang et al. (2017)). This latter case highlights the fact that sieved polymorphism can mimic the signature of other evolutionary processes, and distinguishing among these may be challenging without additional pieces of evidence (see below).

Recent findings suggest that sieved polymorphisms do play an important role in shaping patterns of genomic divergence. Multiple studies have found regions with levels of divergence so high that differentiation at these loci almost certainly started before speciation, consistent with ancestral balanced polymorphism. In these cases, the timing of speciation—or at least a bound on the timing—can be independently estimated, highlighting the mismatch between species divergence and genetic divergence.

In the radiation of Darwin's finches, speciation events happened roughly between 50 to 500 thousand years ago (Lamichhaney *et al.* 2015), yet some genomic regions seem

142 to have started differentiating well before then (up to one million years ago; (Han *et al.*
143 2017)). At least two of these genomic regions, linked to loci associated with beak shape
144 and size (genes *ALX1* and *HMGA2*), are likely sieved polymorphisms. Across the nine
145 species of tree and ground finches studied, these loci have two distinct haplotype
146 classes (*i.e.*, there is high divergence between classes and reduced divergence within
147 class, even between species) that are responsible for marked phenotypic differences
148 (blunt vs. pointed beaks, and small vs. large beaks). As a result, species pairs that have
149 fixed different haplotype classes show 'islands of divergence' at these loci. As
150 hypothesized by Han *et al.* (2017), however, this beak polymorphism was probably
151 balanced in the ancestor (perhaps under negative frequency-dependent selection) and
152 was later sieved across the Galapagos (Han *et al.* 2017).

153 In the freshwater threespine sticklebacks of western North America, the *Eda* locus
154 represents a clear example of how ancestral polymorphisms can emerge as
155 conspicuous peaks of genomic divergence. Polymorphism at *Eda* has been maintained
156 in marine populations for approximately two million years, and the minor allele has been
157 selected repeatedly during colonization events of glacier lakes around ten thousand
158 years ago (Colosimo *et al.* 2005). Expectedly, *Eda* shows dramatic differentiation
159 between marine and freshwater populations, but most of this divergence happened
160 before the invasion of the glacier lakes and is unrelated to recent processes.

161 Other types of polymorphism can also be sieved. Among these, chromosome
162 rearrangements (*e.g.*, inversions, fusions) are of special interest for their role during
163 local adaptation. Rearrangements can evolve by capturing locally adapted alleles

(Kirkpatrick & Barton 2006; Guerrero & Kirkpatrick 2014), and established rearrangements can promote further local adaptation (Navarro & Barton 2003). Moreover, balanced rearrangements (especially inversions) are usually conspicuous in population genomic data (*e.g.*, (Cheng *et al.* 2012; Kapun *et al.* 2016)), as they typically cause a dramatic reduction in recombination, which in turn leads to much stronger population subdivision compared to other balanced polymorphisms (Guerrero *et al.* 2012; Guerrero & Kirkpatrick 2014). Due to their role in local adaptation and their large genomic footprint, rearrangements are thought to be key players in the buildup of differentiation that can lead to speciation. Some chromosome inversions have in fact been linked to speciation processes (*e.g.*, in cactophilic *Drosophila* (Lohse *et al.* 2015), *Mimulus* (Fishman *et al.* 2013)). In other cases, however, locally adapted inversions are maintained as polymorphisms within a species or as trans-specific polymorphisms – without being involved in speciation. In the *Anopheles gambiae* species complex, inversion *2La* arose in the ancestor of six species (well before the most recent speciation events) and it is still polymorphic in two of these (Fontaine *et al.* 2015). This inversion has been sieved at least two times, such that comparisons between species fixed for alternative arrangements show increased divergence across many megabases of sequence.

## Implications for inferences of speciation islands and speciation-with-gene-flow

Many of the patterns described above for sieved polymorphisms mimic predictions of models of speciation with gene flow. In the most common version of this model (which is conceptually similar to sympatric and parapatric speciation models; reviewed in (Bush 1975; Via 2001)), populations have uninterrupted exchange of migrants, but achieve total reproductive isolation gradually by the accumulation of locally adapted loci that limit effective migration (*cf.* Charlesworth *et al.* (1997)). At the genomic level, differential effective migration is expected to leave a clear signature: regions linked to loci under divergent selection (*a.k.a.* local adaptation) accumulate higher divergence compared to the rest of the genome, appearing as 'genomic islands of speciation' (Turner *et al.* 2005). Variation in divergence across the genome has therefore been attributed to differential migration among loci, with individual loci showing much higher levels of divergence implicated as being causal in the speciation process. Several studies initially reported finding such islands using relative measures of divergence (such as $F_{ST}$), which can be affected by selection in the sampled populations (*e.g.*, (Turner *et al.* 2005; Geraldes *et al.* 2011; Ellegren *et al.* 2012)). Because absolute measures of divergence (such as $d_{XY}$) are unaffected by current levels of polymorphism, it was suggested that these would be preferred in identifying regions that are truly resistant to introgression (Cruickshank & Hahn 2014). It has therefore become more common to search for islands using $d_{XY}$ and related statistics, and some researchers have reported finding these important loci (*e.g.*, (Malinsky *et al.* 2015; Marques *et al.* 2016)).

Using absolute measures of divergence does not obviate the problem of variation in levels of diversity in the ancestral population. As discussed above, regions of elevated $d_{XY}$ can be produced by balanced polymorphisms in ancestral populations, and variance in levels of divergence across the genome can be driven by variance in diversity in these ancestral populations. It is simply not true that all loci start out equally diverged at speciation, or that differential migration is the only force that can produce variation in $d_{XY}$ beyond that expected from neutral coalescent variation in the ancestor.

How would one distinguish between true islands and sieved balanced polymorphisms? One commonality shared by the clearest examples of sieved polymorphisms given above is that independent estimates exist for the earliest time when speciation could have started. Glacial lakes that could not have existed prior to the retreat of the glaciers, radiations onto geological features (such as oceanic islands) that recently appeared on the landscape, or simply the date of an earlier divergence from a more distantly related species, all limit the maximum time pairs of focal species could have been separated. Given such limits, we can then contrast hypotheses of speciation-with-gene-flow with those involving sieved polymorphisms (Figure 2). In fact, it takes quite a long time for loci resistant to gene flow to appear as divergence outliers under models of speciation-with-gene-flow (Figure 2A; also see Fig. B1 in Cruickshank and Hahn 2014). By contrast, sieved balanced polymorphisms can be detected almost immediately by both relative and absolute measures of divergence, and ironically these signals are stronger in $F_{ST}$ than $d_{XY}$ (Figure 2B).

We can apply these ideas to assess the likely causes of "islands of divergence" in previously published examples. In the cichilds of Lake Massoko, for instance, levels of divergence might stem from ancestral polymorphism (Malinsky *et al.* 2015). The speciation process in the focal pair started some 3500 generations ago (Malinsky *et al.* 2015), which would allow enough time for the accumulation of divergence on the order of $d_{XY}$=1.05x10$^{-4}$ (assuming $\mu$=1.5x10$^{-8}$ and $\theta_{Anc}$=0). However, $d_{XY}$ observed in the most highly diverged regions is considerably higher than this expectation (mean $d_{XY}$=9x10$^{-4}$; https://twitter.com/millanek1/status/758209899964862465), suggesting that a large fraction of the observed differentiation is due to ancestral diversity. This rough calculation ignores many factors, such as variation in mutation rate, that can contribute to the observed patterns. However, it allows us to emphasize that the genomes of extant populations give us a glimpse into ancestral processes that transcend the most recent speciation event.

Similarly, in the threespine stickleback of Lake Constance, subspecies show several regions of divergence that most likely predate the current process of local adaptation (which started about 150 generations ago), and for which standing variation has been invoked as a probable source (Marques *et al.* 2016). In this case, high differentiation was inferred in 37 genomic regions based on allele frequency differences among populations (using SNPs obtained via RAD-seq). The observed diversity levels in these regions are not significantly reduced, suggesting that—while current selection may be driving allele frequency divergence—the accumulation of divergent SNPs is not the

247 result of a recent sweep. Rather, many of these “islands” are likely ancestral balanced
248 haplotypes currently being sorted.

249 Confounding sieved polymorphism with islands of speciation can lead to an
250 additional erroneous inference. The observation of large amounts of variation in
251 divergence time among loci may lead to the conclusion that gene flow has occurred,
252 when none has. To some extent, these inferences follow from the observation of
253 islands—if there are loci resistant to gene flow, then it follows that there must have been
254 gene flow. But this false signal can also affect methods for inferring gene flow that
255 assume that there is no selection, and therefore interpret the excess variance observed
256 as due to migration. It has recently been recognized that modeling the effects of
257 selection on the levels of sampled polymorphism is important in controlling such false
258 positives (Roux *et al.* 2016). The implication here is that variation in levels of ancestral
259 polymorphism must also be considered, since they can lead to false inference of gene
260 flow and current selection (*i.e.*, labeling sieved regions, which may be neutral now, as
261 resistant to gene flow).

262 ## Conclusions

263 Sieved polymorphism—in conjunction with factors such as population structure,
264 assortative mating, background selection, or variation in mutation and recombination
265 rates—contributes to heterogeneity in genomic divergence levels. Due to the complexity
266 of the divergence distribution, inferences that rely solely on its outliers (*e.g.*, taking an
267 arbitrary upper quantile of $d_{XY}$ as speciation islands) can yield misleading results by
268 selecting regions, such as sieved polymorphisms, that are unrelated to speciation.

Model-based analyses are necessary but not sufficient, because extremely similar patterns of genomic divergence can be generated by alternative models of speciation. In fact, biologically significant differences between speciation models are occasionally irrelevant from a theoretical standpoint (*e.g.*, cessation of gene flow is modeled identically [$m$=0] regardless of the mechanism behind it, whether due to hybrid inviability, a geographic barrier, or other). For this reason, independent lines of evidence are critical to disentangle the multiple forces at play. For instance, having a lower bound on the time since speciation can allow us to determine how much ancestral polymorphism is expected, since its effect is strongest in recent speciation events (*i.e.*, $t < 2N_e$, when $\theta_{Anc}$ accounts for more than half of $E(d_{XY})$). Using current levels of polymorphism as a proxy for ancestral levels in model-based analyses will also be a useful starting point in trying to understand the causes of variation in divergence levels. Finally, a search for the signals of a speciation sieve applied across the many new whole-genome datasets being produced may cause us to reconsider the frequency with which balancing selection occurs.

## References

Asthana S, Schmidt S, Sunyaev S (2005) A limited role for balancing selection. *Trends in Genetics* **21**, 30-32.

Ayala FJ, Campbell CA (1974) Frequency-dependent selection. *Annual Review of Ecology and Systematics* **5**, 115-138.

Begun DJ, Holloway AK, Stevens K*, et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology* **5**, e310.

Bush GL (1975) Modes of animal speciation. *Annual Review of Ecology and Systematics* **6**, 339-364.

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetical Research* **70**, 155-174.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics* **2**, 379-384.

Cheng CD, White BJ, Kamdem C*, et al.* (2012) Ecological genomics of *Anopheles gambiae* along a latitudinal cline: A population-resequencing approach. *Genetics* **190**, 1417-1432.

Colosimo PF, Hosemann KE, Balabhadra S*, et al.* (2005) Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles. *Science* **307**, 1928-1933.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology* **23**, 3133-3157.

Ellegren H, Smeds L, Burri R*, et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**, 756-760.

Fan S, Hansen ME, Lo Y, Tishkoff SA (2016) Going global by adapting local: A review of recent human adaptation. *Science* **354**, 54-59.

Felsenstein J (1976) The theoretical population genetics of variable selection and migration. *Annual Review of Genetics* **10**, 253-280.

Fishman L, Stathos A, Beardsley PM, Williams CF, Hill JP (2013) Chromosomal rearrangements and the genetics of reproductive barriers in *Mimulus* (monkey flowers). *Evolution* **67**, 2547-2560.

Fontaine MC, Pease JB, Steele A*, et al.* (2015) Mosquito genomics. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* **347**, 1258524.

Geraldes A, Basset P, Smith KL, Nachman MW (2011) Higher differentiation among subspecies of the house mouse (*Mus musculus*) in genomic regions with low recombination. *Molecular Ecology* **20**, 4722-4736.

Gillespie JH, Langley CH (1979) Are evolutionary rates really variable? *Journal of Molecular Evolution* **13**, 27-34.

Guerrero RF, Kirkpatrick M (2014) Local adaptation and the evolution of chromosome fusions. *Evolution* **68**, 2747-2756.

Guerrero RF, Rousset F, Kirkpatrick M (2012) Coalescent patterns for chromosomal inversions in divergent populations. *Philosophical Transactions of the Royal Society London B* **367**, 430-438.

Han F, Lamichhaney S, Grant BR*, et al.* (2017) Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Research* doi: 10.1101/gr.212522.116.

Kapun M, Fabian DK, Goudet J, Flatt T (2016) Genomic evidence for adaptive inversion clines in *Drosophila melanogaster*. *Molecular Biology and Evolution* **33**, 1317-1336.

Kirkpatrick M, Barton N (2006) Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419-434.

Lamichhaney S, Berglund J, Almen MS*, et al.* (2015) Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371-375.

Lohse K, Clarke M, Ritchie MG, Etges WJ (2015) Genome-wide tests for introgression between cactophilic *Drosophila* implicate a role of inversions during speciation. *Evolution* **69**, 1178-1190.

Malinsky M, Challis RJ, Tyers AM*, et al.* (2015) Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science* **350**, 1493-1498.

Marques DA, Lucek K, Meier JI*, et al.* (2016) Genomics of rapid incipient speciation in sympatric Threespine Stickleback. *PLoS Genetics* **12**, e1005887.

Muirhead CA, Glass NL, Slatkin M (2002) Multilocus self-recognition systems in fungi as a cause of trans-species polymorphism. *Genetics* **161**, 633-641.

Navarro A, Barton NH (2003) Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes. *Science* **300**, 321-324.

Nosil P (2012) *Ecological speciation* Oxford University Press, Oxford ; New York.

Otto SP, Pannell JR, Peichel CL*, et al.* (2011) About PAR: the distinct evolutionary dynamics of the pseudoautosomal region. *Trends in Genetics* **27**, 358-367.

Phung TN, Huber CD, Lohmueller KE (2016) Determining the effect of natural selection on linked neutral divergence across species. *PLoS Genetics* **12**, e1006199.

Roux C, Fraisse C, Romiguier J*, et al.* (2016) Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biology* **14**, e2000234.

Simonsen KL, Churchill GA, Aquadro CF (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**, 413.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology* **3**, 1572-1578.

Via S (2001) Sympatric speciation in animals: the ugly duckling grows up. *Trends in Ecology & Evolution* **16**, 381-390.

Wright S (1931) Evolution in Mendelian populations. *Genetics* **16**, 0097-0159.

Yang M, He Z, Shi S, Wu CI (2017) Can genomic data alone tell us whether speciation happened with gene flow? *Molecular Ecology* doi: 10.1111/mec.14117.

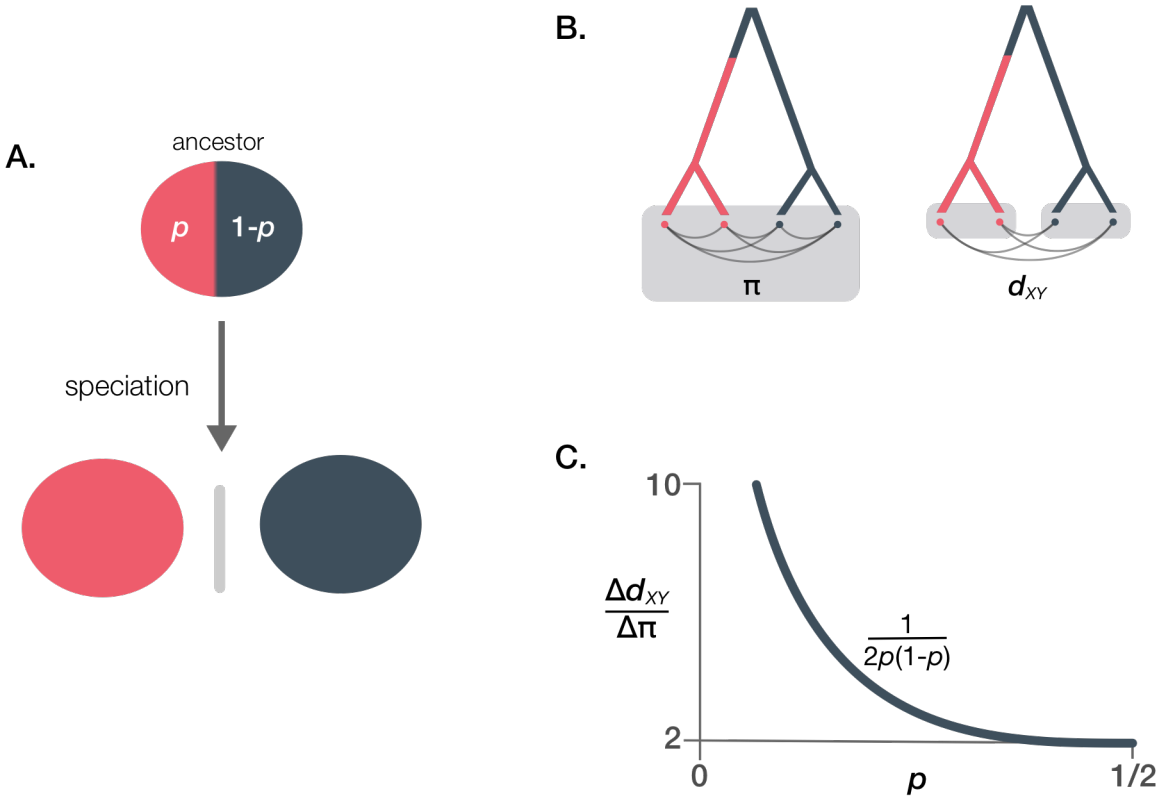


**Figure 1.** A) Schematic of a sieved polymorphism. In the ancestor, two alleles are balanced at frequencies *p* and 1-*p*, and after speciation different alleles fix in descendant lineages. B) The increased power to detect balancing selection stems from the partitioning of the ancestral population imposed by the sieve. While π is a measure of all pairwise distances in a population (left), $d_{XY}$ compares only samples from different allelic classes (right). C) The ratio of the effect of a balanced polymorphism on diversity ($\Delta\pi = \pi - \pi_o$, where $\pi_o$ is the baseline diversity) and on divergence ($\Delta d_{XY} = d_{XY} - d_o$, and $d_o = \pi_o$) increases as the minor allele becomes rare. As the time since the origin of the polymorphism increases, the ratio converges to the inverse of the expected heterozygosity, *i.e.*,$1/2p(1-p)$.

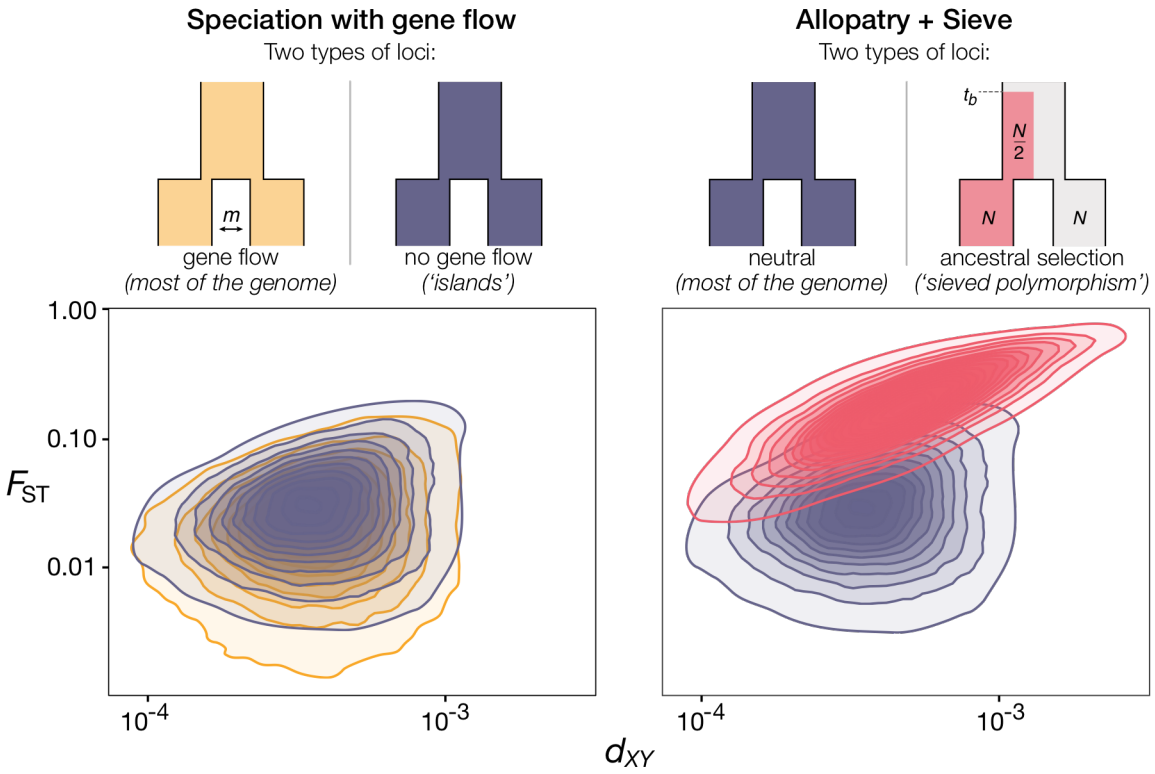


**Figure 2.** Distributions of absolute and relative divergence ($d_{XY}$ and $F_{ST}$) for genomes under two scenarios of recent speciation (1000 generations ago, population size $N$=$10^4$ for each species). In each scenario, genomes have two types of regions. On the left, regions experience differential gene flow since the split: while in most of the genome (in light orange, $m$=0.0001) gene flow prevents divergence, in "speciation islands" (in purple, $m$=0) there is a slight increase in $F_{ST}$. On the right, there is no gene flow after speciation, but some regions are tightly linked to a sieved polymorphism (in pink; balanced locus is at $r$=$10^{-5}$ from the simulated region, originated $t_b$ = $10^4$ generations ago, stable at frequency of ½ in ancestor, alternative alleles are fixed in descendants). We simulated the genealogy for a sample of 20 chromosomes drawn from each species ($10^5$ coalescent simulations for each type of genomic region, a 10 Kb non-recombining segment with μ=$10^{-8}$).