Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*

Daniel R. Schrider[*,†,1], Julien Ayroles[§], Daniel R. Matute[‡], Andrew D. Kern[*,†]

[*]Department of Genetics, and [†]Human Genetics Institute of New Jersey, Rutgers University, Piscataway, New Jersey 08554

[§]Ecology and Evolutionary Biology Department; Lewis Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, 08540

[‡]Biology Department, University of North Carolina, Chapel Hill, NC 27510.

[1]Corresponding author. Current address: Department of Genetics, University of North Carolina at Chapel Hill, 120 Mason Farm Rd, Chapel Hill, NC 27514. E-mail: drs@unc.edu

ABSTRACT

Hybridization and gene flow between species appears to be common. Even though it is clear that hybridization is widespread across all surveyed taxonomic groups, the magnitude and consequences of introgression are still largely unknown. Thus it is crucial to develop the statistical machinery required to uncover which genomic regions have recently acquired haplotypes via introgression from a sister population. We developed a novel machine learning framework, called FILET (Finding Introgressed Loci via Extra-Trees) capable of revealing genomic introgression with far greater power than competing methods. FILET works by combining information from a number of population genetic summary statistics, including several new statistics that we introduce, that capture patterns of variation across two populations. We show that FILET is able to identify loci that have experienced gene flow between related species with high accuracy, and in most situations can correctly infer which population was the donor and which was the recipient. Here we describe a data set of outbred diploid *Drosophila sechellia* genomes, and combine them with data from *D. simulans* to examine recent introgression between these species using FILET. Although we find that these populations may have split more recently than previously appreciated, FILET confirms that there has indeed been appreciable recent introgression (some of which might have been adaptive) between these species, and reveals that this gene flow is primarily in the direction of *D. simulans* to *D. sechellia*.

AUTHOR SUMMARY

Understanding the extent to which species or diverged populations hybridize in nature is crucially important if we are to understand the speciation process. Accordingly numerous research groups have developed methodology for finding the genetic evidence of such introgression. In this report we develop a supervised machine learning approach for uncovering loci which have introgressed across species boundaries. We show that our method, FILET, has greater accuracy and power than competing methods in discovering introgression, and in addition can detect the directionality associated with the gene flow between species. Using whole genome sequences from *Drosophila simulans* and *Drosophila sechellia* we show that FILET discovers quite extensive introgression between these species that has occurred mostly from *D. simulans* to *D. sechellia.* Our work highlights the complex process of speciation even within a well-studied system and points to the growing importance of supervised machine learning in population genetics.

INTRODUCTION

Up to 10% of animal [1] and plant [2] species have the ability to hybridize with other species. Our recent ability to collect large-scale genomic data has confirmed that hybridization is common in nature. Indeed the ubiquity of hybridization upon secondary contact raises the question of how large a role hybridization plays in the emergence or collapse of new lineages [3].

Three general patterns have emerged from recent efforts to search for introgression in genomic data. First, whole-genome sequencing has shown that introgression occurs in all taxa for which its signature has been systematically sought (primates reviewed in [4], plants in [5, 6], fungi [7] and oomycetes in [8]). In general, genetic exchange between species through fertile hybrids might be common between closely related species [9-13] but can also occur between divergent species [14-17].

Second, introgression is heterogeneously distributed across the genome. For instance, mitochondrial genome exchange is surprisingly common (e.g., [18-20] among many) between species, whereas sex chromosomes are less likely to cross species-boundaries, perhaps due to their disproportionate role in hybrid incompatibilities [17, 21-24]. Generally it seems that functional regions of the genome might be less likely to participate in introgression. This is perhaps best known from the case of Neanderthal hybridization with non-African human populations [25, 26], which has left modern human genomes distinct gradients of introgression across different functional compartments of the genome.

Finally, the mode and intensity of natural selection acting on introgressed DNA can vary substantially. Loci that contribute to reproductive isolation, and as such to the persistence of species in the face of hybridization, should be less likely to be introgressed [27] as a result of purifying selection in hybrids. Additionally, introgressed haplotypes containing mildly deleterious variants may be purged after migrating into a population with a larger effective size where selection is more effective [28, 29]. On the other hand, much of the genome may be porous to introgression between closely related species if the net effect of such introgressed variation is fitness neutral. Of course genetic exchange between populations can also provide a source of adaptive alleles that may facilitate adaptation to new environments (reviewed in ref. [30]). Introgressions have indeed been shown to be involved in adaptation in animals (e.g. [31-33]), plants (e.g. [34]) and fungi [35]. For instance, adaptation to high altitude in Tibetans appears to have been caused by introgression of alleles from an archaic Denisovan-like source into modern humans [36]. Another particularly well-studied system of adaptive introgression comes from *Heliconius* butterflies where gene exchange has facilitated the origin and maintenance of mimetic rings [32] and even of hybrid species [37, 38].

Clearly, hybridization and introgression play an important role in shaping the landscape of genetic variation, thus if we wish to fully understand its evolutionary role a reliable framework for the inference of introgressed alleles is needed. Approaches to detect introgression in the genome fall into a few different camps. Genome-wide approaches can identify whether

admixture has occurred in a set of populations. These include clustering methods which seek to infer which individuals are admixed and to assign a proportion of admixture to each individual without previous knowledge of the parental populations [39-41]. Some genome-wide approaches instead attempt to infer the directionality of introgression by examining allele frequency differences among populations [25, 42]. The main limitation of this class of methods is that they cannot identify which regions of the genome are likely to have crossed species boundaries.

On the other hand, locus-specific ancestry approaches (e.g. [43-47]) seek to uncover the mosaic of ancestry for each sampled haplotype, and thus can also identify portions of haplotypes that have been introgressed between species or populations. These fine-resolution approaches are powerful but often have assumptions and requirements that cannot be fulfilled in many taxa which range from the need of phased haplotypes to recombination maps. The main limitation of these approaches is that many require a set of reference haplotypes—individuals known to be unadmixed representatives of either population—in order to properly infer the origin of each allele in each (non-reference) sample haplotype.

The last family of approaches designed to uncover introgressed loci has focused on the use of relative and absolute levels of divergence measured in genomic windows. Largely such methods have consisted of new summary statistics that capture elements of the expected coalescent genealogy under a model of recent introgression between species. These approaches have the advantage that no donor or recipient populations must be identified a priori. Among the measurements of divergence, F_{ST} [48] is most commonly used. Another popular point of departure has been the d_{xy} statistic of Nei and Li [49] which measures the average pairwise distance between alleles sampled from two populations. For instance, Joly et al. [50], Geneva et al. [51] and Rosenzweig et al. [52] use the minimum rather than the mean of these pairwise divergence values, termed d_{min}. d_{min} is sensitive to abnormally short branch lengths between alleles drawn from two populations, as would be expected when introgression is recent. Each of these statistics has attractive properties and adequate power in some instances, however no one statistic has perfect sensitivity in every scenario.

Here we introduce a new method for finding introgressed loci based on supervised machine learning that we call FILET (Finding Introgressed Loci using Extra Trees Classifiers). FILET combines a large number of summary statistics (Materials and Methods) that provide complementary information about the shape of the genealogy underlying a region of the genome. These summary statistics include both previously developed statistics (including, but not limited to, those based on d_{min} and d_{xy}) as well as 5 new summary statistics that we describe below. Our reasoning for this approach was that by combining many statistics for detecting introgression we should achieve sensitivity to introgression across a larger range of scenarios than accessible to any individual statistic. Buoyed by our recent work showing the power and flexibility of Extra Trees classifiers [53] for population genomic inference [54, 55], we leveraged this machine learning paradigm for identification of introgression. Using simulations we show that FILET is far more powerful and versatile than competing methods for identifying introgressed loci.

Further we apply FILET to examine patterns of introgression between *Drosophila simulans* and its island endemic sister taxon *Drosophila sechellia*.

The speciation event that gave rise to the island endemic *Drosophila sechellia* from a *Drosophila simulans*-like ancestor is a textbook example of a specialist species that evolved from a presumably generalist ancestor [56, 57]. Indeed, *D. sechellia* has quite remarkably specialized to breed on the toxic fruit of *Morinda citrifolia* [58], while *D. simulans* (and *D. mauritiana*) do not tolerate the organic volatile compounds in the ripe fruit [59-61]. The genetic and neurological underpinnings of this key ecological difference have been identified, at least in part [62-67] making the *D. simulans*/*D. sechellia* pair one of the most successful cases of genetic dissection of the causes of an ecologically relevant trait. While this is so, the population genetics of divergence between these species has only been examined in the context of either population samples from a handful of loci [68-71] or in the absence of population data [72]. These studies estimated population divergence time between *D. simulans* and *D. sechellia* to be as early as ~250,000 years ago [72] or as old as ~413,000 years ago [70]. All population genomic surveys demonstrate that *D. sechellia* harbors little genetic variation in comparison to *D. simulans*, perhaps as a result of a population size crash/founder event from which the population has not recovered [68, 71]. Moreover it has been suggested that what little variation there is in *D. sechellia* shows little population genetic structure among separate island populations in the Seychelles archipelago [71]. Lastly there is some evidence of introgression between each pair of species within the *D. simulans* complex [72], and *D. simulans* and *D. sechellia* have been found to hybridize in the field [73]. Here we characterize the population genetics of divergence between *D. sechellia* and *D. simulans,* combining existing whole-genome sequences from a mainland population of *D. simulans* [74] with newly generated genome sequences from *D. sechellia.* Applying FILET to these data confirms previous reports of introgression between these species and reveals that this gene flow is primarily in the direction of *D. simulans* to *D. sechellia.* Finally, the success of our approach underscores the potential power of supervised machine learning for evolutionary and population genetic inference.

MATERIALS AND METHODS

Statistics capturing the population genetic signature of introgression

We set out to assemble a set of statistics that could be used in concert to reliably determine whether a given genomic window had experienced recent gene flow. Several statistics that have been designed to this end ask whether there is a pair of samples exhibiting a lower than expected degree of sequence divergence within the window of interest. The most basic of these is d_{min}, the minimum pairwise divergence across all cross-population comparisons (S1 Fig; [50]). The reasoning behind d_{min} is that even if only a single sampled individual contains an introgressed haplotype, d_{min} should be lower than expected and the introgression event may be detectable. A related statistic is G_{min}, which is equal to d_{min}/d_{xy} [51]; the presence of this term in the

denominator is meant to control for variation in the neutral mutation rate across the genome. RND_{min} accomplishes this by dividing d_{min} by the average divergence of all sequences from either species to an outgroup sequence [52]. The name of this statistic is derived from its constituent parts, d_{min}, and *RND* [75].

As described in the following section, we incorporated a number of previously devised statistics into our classification approach, including some of those based on d_{min}. We also included some novel statistics that we designed to have improved sensitivity to particularly recent introgression. The first of these is defined as:

$$d_{d1} = d_{min}/\pi_1$$

where π_1 is nucleotide diversity [49] in population 1. Similarly, $d_{d2} = d_{min}/\pi_2$. d_{d1} and d_{d2} statistics are so named because they compare d_{min} to diversity within populations 1 and 2, respectively. The rationale behind these statistics is that, if there has been recent introgression from population 1 into population 2, and at least one sampled chromosome from population 2 contains the introgressed haplotype, then the cross-population pair of individuals yielding the value of d_{min} should both trace their ancestry to population 1. Thus, the sequence divergence between these two individuals should on average be equal to π_1. Similarly, if there was introgression in the reverse direction d_{min} would be on the order of π_2. Following similar rationale, we devised two related statistics: $d_{d\text{-}Rank1}$ and $d_{d\text{-}Rank2}$. $d_{d\text{-}Rank1}$ is the percentile ranking of d_{min} among all pairwise divergences *within* population 1; the value of this statistic should be lower when there has been introgression from population 1 into population 2. $d_{d\text{-}Rank2}$ is the analogous statistic for introgression from population 2 into population 1. We also included a statistic comparing average linkage disequilibrium within populations to average LD within the global population (i.e. lumping all individuals from both species together), as follows:

$$Z_X = (Z_{nS1} + Z_{nS2}) / (2 \times Z_{nSG})$$

where Z_{nS1}, and Z_{nS2} measure average LD [76] between all pairs of variants within the window in population 1 and population 2, respectively, and Z_{nSG} which measures LD within the global population. The reasoning behind this statistic is based on the assumption that, in the presence of gene flow, LD may be elevated within the recipient population(s) but not in the global population. S2 Fig shows that the distributions of these statistics do indeed differ substantially between genealogies with and without introgression (simulation scenarios described below), especially when this introgression occurred recently. In addition to these and other statistics summarizing diversity across the two population samples, we also incorporated several single-population statistics into our classifier (see below), as these may also contain information about recent introgression. For example, separate measures of nucleotide diversity in our two population samples would contain useful information because introgression is expected to increase diversity in the recipient population, especially if the source population was large or if the two populations split long ago.

Description of FILET classifier

We used a supervised machine learning approach to assign a genomic window to one of three distinct classes on the basis of a "feature vector" consisting of a number of statistics summarizing patterns of variation within the window from two closely related populations. These three classes are: introgression from population 1 into population 2, introgression from population 2 into population 1, and the absence of introgression. Specifically, we used an Extra-Trees classifier [53], which is an extension of random forests [77], an ensemble learning technique that creates a large ensemble of semi-randomly generated binary decision trees [78], before taking a vote among these decision trees in order to decide which class label should be assigned to a given data instance (i.e. genomic window in our case). In an Extra-Trees classifier, the tree building process is even more randomized than in typical random forests: in addition to selecting a random subset of features when generating a tree, the separating threshold for each feature is randomly chosen, rather than selected the threshold that optimally separates the data classes. We require example regions for each class in order to train the Extra-Trees classifier, so we used coalescent simulations to generate these training examples (described below). Our ultimate goal was to detect introgression within 10 kb windows in *Drosophila*, so to train our classifier properly we simulated chromosomal regions approximating this length (simulation details are given below). The target window size could easily be altered by changing the length of the regions simulated for training (i.e. by adjusting the recombination and mutation rates, θ and ρ).

FILET's feature vector contains a number of single-population summaries of per-base pair genetic variation: π, the variance in pairwise distances among individuals, the density of segregating sites, the density of polymorphisms private to the population, Fay and Wu's H and θ_H statistics [79], and Tajima's D [80]. The feature vector also includes two single-population summary statistics that are not normalized per base pair: Z_{nS} (which is averaged across all pairs of SNPs), and the number of distinct haplotypes observed in the window. Each feature vector included values of these 9 statistics for each population, yielding 18 single-population statistics in total. In addition, the two-population statistics included in FILET's feature vector were as follows: F_{ST} (following ref. [81]), Hudson's S_{nn} [82], per-bp d_{xy}, per-bp d_{min}, G_{min}, d_{d1}, d_{d2}, $d_{d\text{-}Rank1}$, $d_{d\text{-}Rank2}$, Z_X, IBS_{MaxB} (the length of the maximum stretch of identity by state, or IBS, among all pairwise between-population comparisons), and IBS_{Mean1} and IBS_{Mean2} which capture the average IBS tract length when comparing all pairs of sequences within populations 1 and 2, respectively. These IBS statistics are calculated by examining all pairs of individual sequences within a population (or across populations in the case of IBS_{MaxB}), noting the positions of differences, and examining the distribution of lengths between these positions (as well as between the first position and the beginning of the window and between the last position and the end of the window). Note that we did not include RND_{min} or other measures such as Patterson's D and F_4 statistics [83] that require alignment to one or more additional species along with the focal pair, because we designed FILET so that it would not require outgroup information. We note however that through its use of supervised machine learning, FILET could easily be extended to incorporate such data. Instead, in order to improve robustness to mutational

variation, we adopted the approach of drawing the mutation rate from a wide range of values when generating training examples to train FILET to classify data from our *Drosophila* samples (see below). All code necessary to run the FILET classifier (including calculating summary statistics on both simulated and real data sets, training, and classification) along with the full results of our application to *D. simulans* and *D. sechellia* (described below) are available at https://github.com/kern-lab/FILET/.

Simulated test scenarios

Following Rosenzweig et al. [52], we used the coalescent simulator msmove (https://github.com/geneva/msmove) to simulate data for testing FILET's power to detect introgression in populations with four different values of T_D (the time since divergence): $0.25\times4N$, $1\times4N$, $4\times4N$, and $16\times4N$ generations ago, where N is the population size. For each of these simulations the population size was held constant (i.e. the ancestral population size equals that of both daughter populations). We developed a classifier for each of these scenarios of population divergence. Supervised machine learning techniques such as the Extra-Trees classifier require training data consisting of examples from each of the three classes, but in practice a large number of example loci known to have experienced introgression may not be available. We therefore simulated training data sets for each of the four values of T_D. Again following Rosenzweig et al. [52], the relevant parameters for each of these simulations include: T_M, the time since the introgression event, which we drew from $\{0.01\times T_D, 0.05\times T_D, 0.1\times T_D, 0.15\times T_D, \ldots, 0.9\times T_D\}$ (i.e. multiples of $0.05\times T_D$ up to 0.9, and also including $0.01\times T_D$); and P_M, the probability that a given lineage would migrate from the source population to the sink population during the introgression event, which we drew from $\{0.05, 0.1, 0.15, \ldots, 0.95\}$. We simulated an equal number of training examples for each combination of these two parameter values for both directions of gene flow, yielding 10^4 simulations in total for both of these classes, conditioning that each of these instances must have contained at least one migrant lineage. Finally, we simulated an equivalent number of samples without introgression, yielding a balanced training set (10^4 examples for each class).

Next, we computed feature vectors as described above for each of these training examples, and proceeded with training our Extra-Trees classifiers by conducting a grid search of all training parameters precisely as described in Schrider and Kern [54], and setting the number of trees in the resulting ensemble to 100. All training and classification with the Extra-Trees classifier was performed using the scikit-learn Python library (http://scikit-learn.org; [84]). We also calculated feature importance and rankings thereof by training an Extra-Trees classifier of 500 decision trees on the same training data (using scikit-learn's defaults for all other learning parameters), and then using this classifier's "feature_importances_" attribute. Briefly, this feature importance score is the average reduction in Gini impurity contributed by a feature across all trees in the forest, always weighted by the probability of any given data instance reaching the feature's node as estimated on the training data [85]; this measure thus captures both how well a

feature separates data into different classes and how often the feature is given the opportunity to split (i.e. how often it is visited in the forest). The values of these scores are then normalized across all features such that they sum to one.

For each T_D, we evaluated the appropriate classifier against a larger set of 10^4 simulations generated for each parameter combination along a grid of values of T_M and P_M. The values of P_M were drawn from the same set as those in training as described above, while one additional possible value of T_M was included: $0.001 \times T_D$. Also note that for these simulations we did not require at least one migrant lineage as we had done for training (information that is recorded by msmove). For test simulations with bidirectional migration, we did not require each replicate sample to contain at least one migrant lineage, though we modified msmove to record which samples did in fact experience migration. In addition to test examples for each direction of gene flow, we simulated 10^4 examples where no migration occurred in order to assess false positive rates. In order to examine the performance of FILET when an unsampled ghost lineage was the source of introgression, we generated test simulations with the same values of T_D, T_M, and P_M as above, but where the source of the introgression was a third, unsampled population that separated from the two sampled populations at time T_D. In all of our simulations, both for training and testing, we set locus-wide population mutation and recombination rates θ and ρ to 50 and 250, respectively, similar to autosomal values in 10 kb windows in *D. melanogaster* [86] and sampled 15 individuals from each population. We also experimented with different window sizes, simulating training and test data (1,000 replicates for each class for each set) with window sizes corresponding to 1 kb, 10 kb, 5 kb, and 50 kb by multiplying θ and ρ by the appropriate scalar. When testing the sensitivity of our method on these data, we considered a window to be introgressed if FILET's posterior probability of the no-introgression class was <0.05, except for the scenario with T_D equal to $16 \times 4N$ generations ago in which case we used a posterior probability cutoff of 0.01, as we found that this step mitigated the elevated false positive rate under this scenario (reducing the rate from >10% to the estimate of 6% shown in S3 Fig). In windows labeled as introgressed, the direction of gene flow was determined by asking which of the two introgression classes had a higher posterior probability. Note that we used the same demographic scenario for both the training and test data for each T_D, and discuss the implications of demographic model misspecification in the Results and Discussion.

In order to compute receiver operator characteristic (ROC) curves we constructed balanced binary training sets composed of 10^4 examples with no introgression, and 10^4 examples allowing for introgression (with equal representation to each combination of T_M, P_M, and direction of introgression. The score that we obtained for each test example in order to compute the ROC curve was one minus the posterior probability of no introgression as generated by the Extra-Trees classifier (i.e. the classifier's estimated probability of introgression, regardless of directionality).

Comparison to ChromoPainter

335 We compared FILET's accuracy to that of ChromoPainter [46], a software program that utilizes
336 a variant of the copying model first proposed by Li and Stephens [87]. In this model each sample
337 haplotype is a mosaic composed of chromosomal segments chosen from a set of possible
338 ancestral haplotypes, allowing for differences caused by mutation and the potential for changes
339 in ancestry at recombination breakpoints. Such an approach can thus be used to predict the
340 ancestry of each individual at each polymorphism—these predictions are referred to as
341 "paintings" by ChromoPainter. To this end we repeated our simulations above but increased the
342 size of the chromosomal segments to 1 Mb by increasing θ and ρ to 5000 and 25000. In these
343 simulations only gene flow from population 2 to population 1 was allowed, and we modified
344 msmove to record the coordinates of introgressed segments, and to restrict introgression events
345 to those involving segments falling entirely within the central 100 kb of the chromosome. For
346 each combination of T_M and P_M we generated 10 replicate simulations, including 10 replicates
347 without introgression.

348 We ran ChromoPainter with the following parameters: the "-a 0 0" switch to model each
349 individual haplotype as a mosaic of each other individual rather than using a set of predefined
350 reference haplotypes, "-i 10" and "-ip" options to estimate copying proportions over 10
351 Expectation-Maximization (EM) iterations, and the default "-s 10" switch to stochastically draw
352 10 chromosome paintings for each individual from the HMM following EM. We then used the
353 output from ChromoPainter to predict introgressed chromosomal segments as follows:

354 For each polymorphism, we examine each haplotype i among our n haplotypes, and
355 record which of the other n-1 haplotypes serves as the best ancestor for i in each of our 10
356 paintings. We then examine each individual in population 2 (the recipient population), and count
357 the number of paintings for which the ancestral haplotype is from population 1. If this number is
358 > 5 (i.e. a majority) for any of our individuals in population 2, then we consider this focal
359 polymorphism to be introgressed. If two adjacent polymorphisms are predicted to be
360 introgressed, all sites between them are also considered to be introgressed. If only 1
361 polymorphism is predicted, then just that one site is considered introgressed. We also produced a
362 more stringent version of these predictions by only retaining introgressed segments consisting of
363 at least 25 consecutive introgressed polymorphisms. Note that ChromoPainter requires base pair
364 positions, and msmove uses an infinite sites model where polymorphisms are located in a
365 continuous space between zero and one. Thus in order to perform this analysis we had to map
366 msmove's continuous locations to physical locations, which we accomplished by multiplying by
367 10^6 and rounding to the nearest available position.

368 We compared ChromoPainter to a sliding-window application of FILET's classifier for
369 10 kb windows (with 1 kb step sizes). We also produced finer-scale FILET predictions using a 1
370 kb classifier (with 100 bp step sizes) to refine predictions made by the 10 kb classifier: only
371 sliding windows predicted as introgressed by the 1 kb classifier and lying within introgressed
372 segments predicted by the 10 kb classifier were retained as candidates by this version. For the
373 refinement step, FILET's posterior probability cutoff for introgression was relaxed to 0.5 (i.e.
374 introgression more probable than not); a more lenient cutoff is appropriate here because this

classifier was only applied within regions already predicted to be introgressed by the 10 kb classifier.

Drosophila sechellia collection

Drosophila sechellia flies were collected in the islands of Praslin, La Digue, Marianne and Mahé with nets over fresh *Morinda* fruit on the ground. All flies were collected in January of 2012. Flies were aspirated from the nets by mouth (1135A Aspirator – BioQuip; Rancho Domingo, CA) and transferred to empty glass vials with wet paper balls (to provide humidity) where they remained for a period of up to three hours. Flies were then lightly anesthetized using FlyNap (Carolina Biological Supply Company, Burlington, NC) and sorted by sex. Females from the *melanogaster* species subgroup were individualized in plastic vials with instant potato food (Carolina Biologicals, Burlington, NC) supplemented with banana. Propionic acid and a pupation substrate (Kimwipes Delicate Tasks, Irving TX) were added to each vial. Females were allowed to produce progeny and imported using USDA permit P526P-15-02964. The identity of the species was established by looking at the taxonomical traits of the males produced from isofemale lines (genital arches, number of sex combs) and the female mating choice (i.e., whether they chose *D. simulans or D. sechellia* in two-male mating trials).

Sequence data and variant calling and phasing

We obtained sequence data from 20 *D. simulans* inbred lines [74] from NCBI's Short Read Archive (BioProject number PRJNA215932). We also sequenced wild-caught outbred *D. sechellia* females (see above) from Praslin (n=7 diploid genomes), La Digue (n=7), Marianne (n=2), and Mahé (n=7). These new *D. sechellia* genomes are available on the Short Read Archive (BioProject number PRJNA395473). For each line we then mapped all reads with bwa 0.7.15 using the BWA-MEM algorithm [88] to the March 2012 release of the *D. simulans* assembly produced by Hu et al. [89] and also used the accompanying annotation based on mapped FlyBase release 5.33 gene models [90]. Next, we removed duplicate fragments using Picard (https://github.com/broadinstitute/picard), before using GATK's HaplotypeCaller (version 3.7; [91-93]) in discovery mode with a minimum Phred-scaled variant call quality threshold (-stand_call_conf) of 30. We then used this set of high-quality variants to perform base quality recalibration (with GATK's BaseRecalibrator program), before again using the HaplotypeCaller in discovery mode on the recalibrated alignments. For this second iteration of variant calling we used the --emitRefConfidence GVCF option in order to obtain confidence scores for each site in the genome, whether polymorphic or invariant. Finally, we used GATK's GenotypeGVCFs program to synthesize variant calls and confidences across all individuals and produce genotype calls for each site by setting the --includeNonVariantSites flag, before inferring the most probable haplotypic phase using SHAPEIT v2.r837 [94]. The genotyping and phasing steps were performed separately for our *D. simulans* and *D. sechellia* data, and for each of step in the

pipeline outlined above we used default parameters unless otherwise noted. In order to remove potentially erroneous genotypes (at either polymorphic or invariant sites), we considered genotypes as missing data if they had a quality score lower than 20, or were heterozygous in *D. simulans*. After throwing out low-confidence genotypes, we masked all sites in the genome missing genotypes for more than 10% of individuals in either species' population sample, as well as those lying within repetitive elements as predicted by RepeatMasker (http://www.repeatmasker.org). Only SNP calls were included in our downstream analyses (i.e. indels of any size were ignored). These phased and masked data are available at https://zenodo.org/record/1166021.

Demographic inference

Having obtained genotype data for our two population samples, we used ∂a∂i [95] to model their shared demographic history on the basis of the folded joint site frequency spectrum (downsampled to n=18 and n=12 in *D. simulans* and *D. sechellia*, respectively); using the folded spectrum allowed us to circumvent the step of producing whole genome alignments to outgroup species in *D. simulans* coordinate space in order to attempt to infer ancestral states. We used an isolation-with-migration (IM) model that allowed for continual exponential population size change in each daughter population following the split. This model includes parameters for the ancestral population size (N_{anc}), the initial and final population sizes for *D. simulans* (N_{sim_0} and N_{sim}, respectively), the initial and final sizes for *D. sechellia* (N_{sech_0} and N_{sech}, respectively), the time of the population split (T_D), the rate of migration from *D. simulans* to *D. sechellia* ($m_{sim \rightarrow sech}$), and the rate of migration from *D. sechellia* to *D. simulans* ($m_{sech \rightarrow sim}$). We also fit our data to a pure isolation model that was identical to our IM model but with $m_{sim \rightarrow sech}$ and $m_{sech \rightarrow sim}$ fixed at zero. Finally, we fit our data to an admixture model identical to the isolation model but with the addition of two parameters: the time of a pulse admixture event from *D. simulans* into *D. sechellia* (T_{AD}) and the proportion of individuals in *D. sechellia* migrating from *D. simulans* during this event (P_{AD}). Our optimization procedure consisted of an initial optimization step using the Augmented Lagrangian Particle Swarm Optimizer [96], followed by a second step of optimization refining the initial model using the Sequential Least Squares Programming algorithm [97], both of which are included in the pyOpt package for optimization in Python (version 1.2.0; [98]) as in Schrider et al. [99]. We performed ten optimization runs fitting both of these models to our data, each starting from a random initial parameterization, and assessed the fit of each optimization run using the AIC score. Code for performing these optimizations can be obtained from https://github.com/kern-lab/miscDadiScripts, wherein 2popIM.py, 2popIsolation.py, 2popIsolation_admixture.py fit the IM, isolation, and admixture models described above, respectively. For scaling times by years rather than numbers of generations, we assumed a generation time of 15 gen/year as has been estimated in *D. melanogaster* [100].

Training FILET to detect introgression between *D. simulans* and *D. sechellia*

Having obtained a demographic model that provided an adequate fit to our data, we set out to simulate training examples under this demographic history for each of our three classes (i.e. no migration, migration from *D. simulans* to *D. sechellia*, and from *D. sechellia* to *D. simulans*). For training examples including introgression, T_M was drawn uniformly from the range between zero generations ago and $T_D/4$, while P_M raged uniformly from (0, 1.0]. In addition, in order to make our classifier robust to uncertainty in other parameters in our model, for each training example we drew values of each of the remaining parameters from $[x-(x/2), x+(x/2)]$, where x is our point estimate of the parameter from ∂a∂i. In addition to the parameters from our demographic model (T_D, ρ, N_{anc}, N_{sim}, and N_{sech}), these include the population mutation rate $\theta=4N\mu$ (where μ was set to 3.5×10^{-9}), and the ratio of θ to the population recombination rate ρ (which we set to 0.2). Continuous migration rates were set to zero (i.e. the only migration events that occurred were those governed by the T_M and P_M parameters, and the no-migration examples were truly free of migrants). In total, this training set comprised of 10^4 examples from each of our three classes.

As described above, we masked genomic positions having too many low confidence genotypes or lying within repetitive elements (described above) before proceeding with our classification pipeline. While doing so, we recorded which sites were masked within each 10 kb window in the genome that we would later attempt to classify. In order to ensure that our masking procedure affected our simulated training data in the same manner as our real data, for each simulated 10 kb window we randomly selected a corresponding window from our real dataset and masked the same sites in the simulated window that had been masked in the real one. We then trained our classifier in the same manner as described above.

In order to ensure that this classifier would indeed be able to reliably uncover loci experiencing recent gene flow between our two populations, we assessed its performance on simulated test data. First, we applied the classifier to test examples simulated under this same model (again, 10^4 for each class). Next, to address the effect of demographic model misspecification, we applied our classifier to an isolation model with a different parameterization and no continuous size change in the daughter populations. For this model we simply set N_{sim} and N_{sech} to $\pi_{sim}/4\mu$ and $\pi_{sech}/4\mu$, respectively, where π for a species is the average nucleotide diversity among all windows included in our analysis after filtering, and μ was again set to 3.5×10^{-9}. We then set N_{anc} to be equal to N_{sim}, and set T to $d_{xy}/(2\mu) - 2N_{anc}$ generations where d_{xy} is the average divergence between *D. simulans* and *D. sechellia* sequences across all windows. This latter value is simply the expected TMRCA for cross-species pairs of genomes, minus the expected waiting time until coalescence during the one-population (i.e. ancestral) phase of the model. This simple model thus produces samples with similar levels of nucleotide diversity for the two daughter populations as those produced under our IM model, but that would differ in other respects (e.g. the joint site frequency spectrum and linkage disequilibrium, which would be affected by continuous population size change after the split). For both test sets we masked sites in the same manner as for our training data before running FILET.

495 **Classifying genomic windows with FILET**
496
497 We examined 10 kb windows in the *D. simulans* and *D. sechellia* genomes, summarizing
498 diversity in the joint sample with the same feature vector as used for classification (see above),
499 ignoring sites that were masked as described above. We omitted from this analysis any window
500 for which >25% of sites were masked, and then applied our classifier to each remaining window,
501 calculating posterior class membership probabilities for each class. We then used a simple
502 clustering algorithm to combine adjacent windows showing evidence of introgression into
503 contiguous introgressed elements: we obtained all stretches of consecutive windows with >90%
504 probability of introgression as predicted by FILET (i.e. the probability of no-introgression class
505 <10%), and retained as putatively introgressed regions those that contained at least one window
506 with >95% probability of introgression. In order to test for enrichment of these introgressed
507 regions for genic/intergenic sequence or particular Gene Ontology (GO) terms from the FlyBase
508 5.33 annotation release [90], we performed a permutation test in which we randomly assigned a
509 new location for each cluster or introgressed windows (ensuring the entire permuted cluster
510 landed within accessible windows of the genome according to our data filtering criteria). We
511 generated 10,000 of these permutations.
512
513 **RESULTS AND DISCUSSION**
514
515 **FILET detects introgressed loci with high sensitivity and specificity**
516
517 We sought to devise a bioinformatic approach capable of leveraging population genomic data
518 from two related population samples to uncover introgressed loci with high sensitivity and
519 specificity. In the Materials and Methods, we describe several previous and novel statistics
520 designed to this end. However, rather than preoccupying ourselves with the search for the ideal
521 statistic for this task, we took the alternative approach of assembling a classifier leveraging many
522 statistics that would in concert have greater power to discriminate between introgressed and non-
523 introgressed loci. Supervised machine learning methods have proved highly effective at making
524 inferences in high-dimensional datasets and are beginning to make inroads in population genetics
525 [101]. In this vein, we designed FILET, which uses an extension of random forests called an
526 Extra-Trees classifier [53]. We previously succeeded in applying Extra-Trees classifiers for a
527 separate population genetic task—finding recent positive selection and discriminating between
528 hard and soft sweeps [54, 55]—though other methods such as support vector machines [102] or
529 deep learning [103] could also be applied to this task.
530 Briefly, FILET assigns a given genomic window to one of three distinct classes—recent
531 introgression from population 1 into population 2, introgression from population 2 into 1, or the
532 absence of introgression—on the basis of a vector of summary statistics that contain information
533 about the two-population sample's history. This feature vector contains a variety of statistics
534 summarizing patterns of diversity within each population sample, as well as a number of

535 statistics examining cross-population variation (see Materials and Methods for a full description).
536 FILET must be trained to distinguish among these three classes, which we accomplish by
537 supplying 10,000 simulated example genomic windows of each class, with each example
538 represented by its feature vector. Because we expect that the majority of introgression events to
539 be non-adaptive, these simulations did not include natural selection. Once this training is
540 complete, FILET can then be used to infer the class membership of additional genomic windows,
541 whether from simulated or real data.

542 We began by assessing FILET's power on a number of simulated datasets, examining
543 windows roughly equivalent to 10 kb in length in *Drosophila* (Materials and Methods). In
544 particular, because the power to detect introgression depends on the time since their divergence,
545 T_D, we measured FILET's performance under four different values of T_D, training a separate
546 classifier for each. In Figure 1 (T_D=0.25×4N) and S3 Fig (T_D values of 1, 4, and 16×4N), we
547 compare FILET's power to that of two related statistics that have been devised to detect
548 introgressed windows, d_{min} and G_{min} (Materials and Methods). These figures show that FILET
549 has high sensitivity to introgression across a much wider range of introgression timings (T_M) and
550 intensities (P_M) than either of these statistics under each value of T_D, and that this disparity is
551 amplified dramatically for smaller values of T_D. Furthermore, these figures demonstrate that
552 FILET infers the correct directionality of recent introgression with high accuracy, whereas d_{min}
553 and G_{min} contain no information about the direction of gene flow. Finally, FILET does not appear
554 especially sensitive when the source of gene flow is an unsampled ghost population rather than
555 one of the two sequenced populations (S4 Fig), though it could potentially be trained to detect
556 such cases if desired.

557 We also note that for d_{min} and G_{min} we established 95% significance thresholds from our
558 simulated training data without introgression, thereby achieving a false positive rate of 5%. In
559 order to assess FILET's false positive rate, we classified a set of test simulations without
560 introgression and found that FILET's false positive rate was considerably lower (Figure 1 and S3
561 Fig) except for our largest value of T_D, where it was initially higher (0.4% for T_D=0.25×4N but
562 >10% for T_D=16×4N), despite our selection of a posterior probability cutoff of 95% (Methods).
563 This illustrates an important issue with posterior probability estimates produced by supervised
564 machine learning methods: they may occasionally be miscalibrated. We therefore adjusted the
565 cutoff for the T_D=16×4N scenario (to 99% probability of introgression) which lowered our false
566 positive rate to 6% as shown in S3 Fig. Thus, when an appropriate posterior probability cutoff is
567 chosen—a task that can be performed in a straightforward manner by testing on simulated data—
568 FILET achieves much greater sensitivity to introgression than d_{min} and G_{min} often at a much
569 lower false positive rate. We also demonstrate the FILET's greater power than these statistics via
570 ROC curves (S5 Fig), where it outperforms each statistic under each scenario. Specifically, the
571 difference in power between FILET and d_{min} is dramatic for smaller values of T_D (area under
572 curve, or AUC, of 0.85 versus 0.73 when T_D=0.25×4N for FILET and d_{min}, respectively) but
573 comparatively miniscule for our largest T_D (AUC of 0.94 versus 0.93 when T_D=16×4N). It
574 therefore appears that FILET's performance gain relative to single statistics is highest for the

575 more difficult task of finding introgression between very recently diverged populations, while for
576 the easier case of detecting introgression between highly diverged populations some single
577 statistics may perform nearly as well.
578 We also experimented with smaller training sets, finding similar classification power
579 (measured by AUC) as above when we trained FILET using only 1000 or even 100 simulated
580 examples per class (S6 Fig), though in the latter case estimated class posterior probabilities were
581 far less accurate. In addition, we examined the effect of altering the target window size used
582 when training and testing FILET (S7 Fig).
583
584 **A comparison of the power and resolution of FILET and ChromoPainter**
585
586 Methods designed to uncover changes in ancestry along a recombining chromosome within
587 admixed populations can also be used to recover introgressed regions. To this end we used
588 ChromoPainter [46] which has the advantage of not requiring a set of "reference haplotypes"
589 known to be free of introgression from each population, and can instead predict for each
590 haplotype, which of all the other haplotypes in the sample (from either population) is most
591 closely related. We simulated two-population samples for 1 Mb chromosomes where
592 introgression from population 2 to population 1 was allowed in the central 100 kb window, and
593 used ChromoPainter to identify introgressed loci (see Methods). We then ran FILET on these
594 simulations, this time using a sliding-window approach to detect introgressed segments
595 (Methods).
596 Figure 2 shows that FILET has substantially higher sensitivity than ChromoPainter—
597 summing across the entire parameter space (including many scenarios where introgression is
598 quite difficult to detect) FILET recovered 27.7% of introgressed base pairs compared to 19.4%
599 for ChromoPainter—while having a roughly 20-fold lower false positive rate (0.42% for FILET
600 versus 9.31% for ChromoPainter). For scenarios with more ancient and less intense
601 introgression, we did observe somewhat higher sensitivity in ChromoPainter's predictions.
602 However, this seems to be driven largely by ChromoPainter's propensity to identify a larger
603 fraction of base pairs as introgressed regardless of their true ancestry, as evidenced by its higher
604 false positive rate. To demonstrate this further we show for the positive predictive value (the
605 number of base pairs correctly predicted to be introgressed divided by the total number of base
606 pairs predicted to be introgressed) for each method in S8 Fig. This figure shows that FILET's
607 positive predictive is consistently far higher than ChromoPainter's. We sought to improve this by
608 adopting a more stringent threshold for ChromoPainter's predictions, requiring at least 25
609 adjacent polymorphisms to be called introgressed in order to retain the candidate region. This
610 approach did succeed at reducing the false positive rate to 1.15%, though this is still substantially
611 higher than FILET's, but this improvement came at the cost of ChromoPainter's sensitivity,
612 which was reduced to 8.6%, roughly one-third that of FILET (Figure 2 and S8 Fig). We also
613 tried an intermediate threshold (5 polymorphisms), but did not observe a substantial increase in
614 specificity compared to our initial more lenient approach (8.49% false positive rate). Thus, while

615 we cannot rule out that it may be possible to devise a method to leverage ChromoPainter's model
616 to predict introgression that exceeds the performance of our application of ChromoPainter here,
617 our results suggest that it is unlikely that such an approach would eclipse the performance of
618 FILET. We note that ChromoPainter does have the advantage of not requiring simulated training
619 data. ChromoPainter also has the potential to identify donor and recipient haplotypes, which
620 FILET currently does not, but the far greater power of FILET demonstrated above will make it
621 preferable to many researchers who are interested in identifying introgressed regions. Moreover
622 our above results imply that predictions of the span and origin of introgressed haplotypes made
623 directly from ChromoPainter's output may not always be particularly reliable.
624 It is important to note that in the above simulations many introgressed regions will be
625 considerably smaller than our 10 kb window size. This fact, combined with our use of accuracy
626 measurements counting the number of individual base pairs correctly classified, makes the
627 results presented above useful for evaluating FILET's resolution and the impact of window size
628 on our predictions. By these measures FILET outperforms ChromoPainter, which does not use
629 windows and is only limited in scale by the density of polymorphisms. This suggests that when
630 using sliding windows FILET is able to achieve adequate resolution regardless of its use of a
631 predefined window size. Nonetheless we sought to improve our resolution further by using a
632 finer-scale FILET classifier trained on 1 kb windows as described above to refine the location of
633 putatively introgressed regions identified by the 10 kb classifier (see Methods). While this did
634 marginally reduce our false positive rates and increase our positive predictive values (see Figure
635 2 and S8 Fig), sensitivity was also somewhat reduced (to 25.7%; Figure 2). The relatively minor
636 effect of adding this refinement step reinforces the notion that a predefined window size is not a
637 major hindrance to our methods' effectiveness. Thus for most applications a window size that
638 yields enough polymorphisms to reliably calculate the statistics included in our feature vectors
639 may suffice.
640 Overall, FILET detects introgressed regions with greater power and resolution than
641 ChromoPainter, a method designed to detect ancestry tracks along recombining chromosomes.
642 However we note that many methods of this class exist, and it is possible that some may achieve
643 greater accuracy in some circumstances (e.g. if reference haplotype panels are used).
644
645 **Sensitivity to continuous gene flow**
646
647 While FILET is designed for identifying particular genomic windows that experienced
648 introgression as a result of a pulse migration event, genomic windows with genealogies that
649 include introgression may of course also be produced by continuous migration, with the timing
650 of geneflow varying from window to window. We therefore simulated genomic windows
651 experiencing a variety of bidirectional migration rates under each of our values of T_D and
652 recorded the fraction of windows for which our sampled individuals contained at least one
653 migrant lineage. Next, for each simulated window we applied the FILET classifier trained under
654 the appropriate divergence time as described above, recording the fraction of windows with at

655 least one migrant lineage that were classified as introgressed by FILET. The results of these tests
656 (S1 Table) show that for each value of T_D, the lowest bidirectional migration rates that we tested
657 do not produce migrant lineages, while higher rates will produce a small to modest fraction
658 migrants, most of which are undetected (e.g. when m=0.01, 23% and 59% of windows contain at
659 least one migrant when T_D=0.25 and T_D=1, respectively, but <5% are detected by FILET). Thus,
660 FILET, as currently trained, may not be sensitive to gene flow produced by low levels of
661 continuous migration. However, as the migration rate increases further, more and more of these
662 migrant windows will be detected (e.g. when m=1, 70% and 100% of windows are detected as
663 migrants by FILET when T_D=0.25 and T_D=1, respectively).
664
665 **Ranking the importance of statistics for detecting introgression**
666
667 Although our goal was to use a set of statistics to perform more accurate inference than possible
668 using individual ones, another benefit of our Extra-Trees approach is that it also allows for a
669 natural way to evaluate the extent to which different statistics are informative under different
670 scenarios of introgression. To this end, we used the Extra-Trees classifier to calculate feature
671 importance, which captures the ability of each statistic to separate the data into its respective
672 classes (Materials and Methods). We find that for our lowest T_D (a split N generations ago) the
673 top four features, all with similar importance, are the density of private alleles in population 1,
674 the density of private alleles in population 2, $d_{d\text{-}Rank1}$, and $d_{d\text{-}Rank2}$. For our next-lowest T_D (4N
675 generations ago), the top four, again with similar importance score estimates, are F_{ST}, Z_X, d_{d1}, and
676 d_{d2}. Thus our newly devised d_d and Z_X statistics seem to be particularly informative in the case of
677 recent introgression between closely related populations. For the larger values of T_D, d_{xy} and d_{min}
678 rise to prominence. The complete lists of feature importance for each T_D are shown in S2 Table.
679 The exceptional accuracy with which FILET uncovers introgressed loci underscores the
680 potential for machine learning methods to yield more powerful population genetic inferences
681 than can be achieved via more conventional tools which are often based on a single statistic.
682 Indeed, machine learning tools have been successfully leveraged in efforts to detect recent
683 positive selection [54, 104-107], to infer demographic histories [108], or even to perform both of
684 these tasks concurrently [109].
685
686 **Joint demographic history of *D. simulans* and *D. sechellia***
687
688 As described in the Materials and Methods, we used publically available *D. simulans* sequence
689 data [74], and collected and sequenced a set of *D. sechellia* genomes. We mapped reads from
690 these genomes to the *D. simulans* assembly [89], obtaining high coverage >28× for each
691 sequence (see sampling locations, mapping statistics, and Short Read Archive identifier
692 information listed in S3 Table). We do not expect that our reliance on the *D. simulans* assembly
693 resulted in any appreciable bias, as reads from our *D. sechellia* genomes were successfully
694 mapped to the reference genome at nearly the same rate as reads from *D. simulans* (S3 Table).

695 After completing variant calling and phasing (Materials and Methods), we performed
696 principal components analysis on intergenic SNPs at least 5 kb away from the nearest gene in
697 order to mitigate the bias introduced by linked selection [99, 110, 111]. While this is unlikely to
698 completely eliminate the confounding effect of linked selection in *Drosophila*, the fraction of
699 mutations that are deleterious is far greater in coding regions than in intergenic regions (~90%
700 versus <50% according to [112]); thus it is reasonable to presume that the impact of linked
701 selection is reduced several kilobases away from genes [113]. We observed evidence of
702 population structure within *D. sechellia*. In particular, the samples obtained from Praslin
703 clustered together, while all remaining samples formed a separate cluster (S9 FigA). Running
704 fastStructure [114] on this same set of SNPs yielded similar results: when the number of
705 subpopulations, *K*, was set to 2 (the optimal value for *K* selected by fastStructure's chooseK.py
706 script), our data were again subdivided into Praslin and non-Praslin clusters. Indeed, across all
707 values of *K* between 2 and 8, fastStructure's results were suggestive of marked subdivision
708 between Praslin and non-Praslin samples, and comparatively little subdivision within the non-
709 Praslin data (S9 FigB). This surprising result differs qualitatively from previous observations
710 from smaller numbers of loci [71, 115], and underscores the importance of using data from many
711 loci—preferably intergenic and genome-wide—in order to infer the presence or absence of
712 population structure.
713 Next, we examined the site frequency spectra of the Praslin and non-Praslin clusters,
714 noting that both had an excess of intermediate frequency alleles in comparison to that of the *D.*
715 *simulans* dataset (S10 Fig), in line with our expectations of *D. sechellia's* demographic history.
716 We also note that the Praslin samples contained far more variation (50,243 sites were
717 polymorphic within Praslin) than non-Praslin samples (4,108 SNPs within these samples). This
718 difference in levels of variation may reflect a much lesser degree of population structure and/or
719 inbreeding on the island of Praslin than across the other islands, or may result from other
720 demographic processes. Additional samples from across the Seychelles would be required to
721 address this question. In any case, in light of this observation we limited our downstream
722 analyses of *D. sechellia* sequences to those from Praslin.
723 Because we required a model from which to simulate training data for FILET, we next
724 inferred a joint demographic history of our population samples using ∂a∂i [95]. In particular, we
725 fit three demographic models to our dataset: an isolation-with-migration (IM) model allowing for
726 continuous population size change and migration following the population divergence, an
727 isolation model with the same parameters but fixing migration rates at zero, and an isolation
728 model with one burst of pulse migration from *D. simulans* into *D. sechellia* (Materials and
729 Methods). In S4 Table we show our model optimization results, which show clear support for the
730 IM model over the other models. The IM model that provided the best fit to our data (Figure 3A)
731 includes a much larger population size in *D simulans* than *D. sechellia* (a final size of 9.3×10^6
732 for *D. simulans* versus 2.6×10^4 for *D. sechellia*), consistent with the much greater diversity
733 levels in *D. simulans* [10, 71]. This model also exhibits a modest rate of migration, with a
734 substantially higher rate of gene flow from *D. simulans* to *D. sechellia* ($2\times N_{anc}m$=0.086) than

735 vice-versa ($2\times N_{anc}m$=0.013). Thus, the results of our demographic modeling are consistent with
736 the observation of hybrid males in the Seychelles [73], and the possibility of recent introgression
737 between these two species across a substantial fraction of the genome (see refs. [72, 116]).
738 An interesting characteristic of the model shown in Figure 3A is that, assuming 15
739 generations per year, the estimated time of the *D. simulans*- *D. sechellia* population split is ~86
740 kya, or 1.3×10^6 generations ago. This contrasts with a recent estimate of 2.5×10^6 generations ago
741 from Garrigan et al. [72] which was based on single genomes rather than population genomic
742 data, but did account for ancestral polymorphism, as did estimates from Obbard et al. [117]
743 which yielded even older split times. Supporting our inference, we note that our average
744 intergenic cross-species divergence of 0.017 yields an average TMRCA of ~2.5×10^6 generations
745 ago, assuming a mutation rate of 3.5×10^{-9} mutations per generation as observed in *D.*
746 *melanogaster* [112, 118], and this estimate would include the time before coalescence in the
747 ancestral population. Unless the mutation rate the *D. simulans* species complex is substantially
748 lower than in *D. melanogaster*, a population split time of 2.5×10^6 generations ago therefore
749 seems unlikely given that the ancestral population size (and therefore the period of time between
750 the population divergence and average TMRCA) was probably large (>500,000 by our estimate).
751 Thus, we conclude that the *D. simulans* and *D. sechellia* populations may have diverged more
752 recently than previously appreciated, perhaps within the last 100,000 years.
753 Although the specific parameterization of our model should be regarded as a preliminary
754 view of these species' demographic history that is adequate for the purposes of training FILET,
755 future efforts with larger sample sizes will be required to refine this model. That being said, the
756 basic features of this model—a much larger *D. simulans* population size than *sechellia*, and a
757 fairly large ancestral population size—are unlikely to change qualitatively.
758
759 **Widespread introgression from *D. simulans* to *D. sechellia***
760
761 *Accuracy and robustness of FILET under estimated model:* Having obtained a suitable model of
762 the *D. simulans*- *D. sechellia* joint demographic history, we proceeded to simulate training data
763 and train FILET for application to our dataset (Materials and Methods). After training FILET
764 and applying it to simulated data under the estimated demographic model, we find that we have
765 good sensitivity to introgression (56% of windows with introgression are detected, on average),
766 and a false positive rate of only 0.2% (Figure 3B). Thus, while we may miss some introgressed
767 loci, we can have a great deal of confidence in the events that we do recover. Our feature
768 rankings for this classifier are included in S2 Table—under this scenario the most informative
769 feature is our newly devised $d_{d\text{-}sim}$. Note that we achieve high accuracy despite some of the
770 difficulties presented by the demographic model in Figure 3A, most notably the asymmetry in
771 effective population sizes between our two species. Indeed, because our method is trained under
772 this demographic history, such characteristics of genealogies produced under the assumed
773 demographic history (such as asymmetry in π) with and without introgression become the signals
774 used by FILET to make its classifications.

775 As shown in Figure 3B we find that this classifier has greater sensitivity to introgression
776 from *D. sechellia* to *D. simulans* than vice-versa. The cause of a stronger signal of *D. sechellia*→
777 *D. simulans* introgression can be understood from a consideration of the d_{min} statistic under each
778 of our three classes. When there is no introgression, d_{min} will be similar to the expected
779 divergence between *D. simulans* and *D. sechellia*; when there is introgression from *D. simulans*
780 to *D. sechellia*, we may expect d_{min} to be proportional to π_{sim}, which may only be a moderate
781 reduction relative to the no-introgression case given the large population size in *D. simulans*;
782 when there is introgression from *D. sechellia* to *D. simulans* then d_{min} is proportional to π_{sech}
783 which is dramatically lower than the expectation without introgression. While many of our
784 statistics do not rely on d_{min}, this example illustrates an important property of the genealogies
785 which include introgression from *D. sechellia* to *D. simulans* that would make them easier to
786 detect than gene flow in the reverse direction.
787 We also tested this classifier's performance on a different demographic scenario (S4
788 Table) in order to examine the effect of model misspecification during training. In particular, we
789 devised a simple island model with two population sizes: a larger size for *D. simulans* and the
790 ancestral population (7.6×10^5), and a smaller size for *D. sechellia* (5.7×10^4) with a split time of
791 ~59 kya. Our simple procedure for estimating these values is described in the Materials and
792 Methods. Again, we find that we have good power to detect introgression with a very low false
793 positive rate (0.28%; S11 Fig). Although there are myriad incorrect models that we could test
794 FILET against, this example suggests that FILET's performance is robust to at least some
795 scenarios of demographic misspecification.
796
797 *Application to population genomic data:* We applied FILET to 10,185 non-overlapping 10 kb
798 windows that passed our data quality filters (101.85 Mb in total, or 86.7% of the five major
799 chromosome arms; Materials and Methods). FILET classified 267 windows as introgressed with
800 high-confidence, which we clustered into 94 contiguous regions accounting for 2.93% of the
801 accessible portion of the genome (2.99 Mb in total; Materials and Methods). This finding is
802 qualitatively similar to a previous estimate (4.6%) by Garrigan et al. [72] based on comparisons
803 of single genomes from each species in the *D. simulans* complex. Unlike this previous effort,
804 FILET is able to infer the directionality of introgression with high confidence (Figure 3B), and
805 we find evidence that the majority of this introgression has been in the direction of *D. simulans*
806 to *D. sechellia*: only 21 of the 267 (7.9%) putatively introgressed windows were classified as
807 introgressed from *sechellia* to *D. simulans*. This finding is not a result of a detection bias, as we
808 have greater power to detect gene flow from *D. sechellia* to *D. simulans* than in the reverse
809 direction. Given that our *D. simulans* sequences are from the mainland, one interpretation of this
810 result is that although there has been recent gene flow from *D. simulans* into the Seychelles,
811 where *D. simulans* and *D. sechellia* occasionally hybridize, there does not appear to be an
812 appreciable rate of back-migration to the mainland of *D. simulans* individuals harboring
813 haplotypes donated from *D. sechellia*. On the other hand, *D. sechellia* alleles may often be
814 purged from *D. simulans* by natural selection. This may be in part due to the reduced ecological

815 niche size of *D. sechellia*, such that any alleles which may introgress into *D. simulans* and lead to
816 preference for or resistance to *Morinda* fruit may prove deleterious in other environments. More
817 generally, *D. sechellia* haplotypes introgressing into *D. simulans* may harbor more deleterious
818 alleles due to their smaller population size, which will be more effectively purged in the larger *D.*
819 *simulans* population if mutations are not fully recessive [28]. Tests of these hypotheses will have
820 to wait for a population sample of genomes from *D. simulans* collected in the Seychelles.
821 We asked whether our candidate introgressed loci were enriched for particular GO terms
822 using a permutation test (Materials and Methods), finding no such enrichment. We did observe a
823 deficit in the number of genes either partially overlapping or contained entirely within
824 introgressed regions in our true set versus the permuted set; although a paucity of introgressed
825 genes would be consistent with introgressed functional sequence often being deleterious, this
826 difference was not significant (297 vs. 373.2, respectively; P=0.083; one-sided permutation test).
827 One notable introgressed region on 3R that FILET identified had been previously found
828 by Garrigan et al. as containing a 15 kb region of introgression. We show that gene flow in this
829 region actually extends for over 200 kb (Figure 4). When Brand et al. [119] sequenced the 15 kb
830 region originally flagged by Garrigan et al. in a number of *D. simulans* and *D. sechellia*
831 individuals, they also uncovered evidence of a selective sweep in *D. sechellia* originating from
832 an adaptive introgression from *D. simulans*. Our data set also supports the presence of an
833 adaptive introgression event at this locus: a 10 kb window lying within the putative sweep region
834 (highlighted in Figure 4) is in the lower 5% tail of both d_{min} (consistent with introgression) and
835 π_{sech} (consistent with a sweep in *sechellia*); this is the only window in the genome that is in the
836 lower 5% tail for both of these statistics. This region contains two ionotropic glutamate
837 receptors, *CG3822* and *Ir93a*, which may be involved in chemosensing among other functions
838 [120], and the latter of which appears to play a role in resistance to entomopathogenic fungi
839 [121]. Also near the trough of variation within *D. sechellia* are several members of the *Turandot*
840 gene family, which are involved in humoral stress responses to various stressors including heat,
841 UV light, and bacterial infection [122, 123], and perhaps parasitoid attack as well [124]. On the
842 other hand, Brand et al. [119] hypothesize that the target of selection may be a transcription
843 factor binding hotspot between *RpS30* and *CG15696*, and the phenotypic target of this sweep
844 remains unclear.
845 Interestingly, this particular window is the only one in this region that is classified by
846 FILET as having recent gene flow from *D. sechellia* to *D. simulans*. However this classification
847 may be erroneous as one might expect FILET, which was not trained on any examples of
848 adaptive introgression, to make an error in such a scenario because rather than gene flow
849 increasing polymorphism in the recipient population, diversity is greatly diminished if the
850 introgressed alleles rapidly sweep toward fixation. We note that this window is immediately
851 flanked by a large number of windows classified as introgressed from *D. simulans* to *D. sechellia*
852 and which show a large increase in diversity in the recipient population as expected. Moreover,
853 Brand et al.'s phylogenetic analysis of introgression in this region also supported gene flow in
854 this direction. Brand et al. also found evidence suggesting that the introgressed haplotype began

855 sweeping to higher frequency in *D. simulans* (though it has not reached fixation in this species)
856 prior to the timing of the introgression and subsequent sweep in *D. sechellia*. Thus we conclude
857 that the adaptive allele probably did indeed originate in *D. simulans* before migrating to *D.*
858 *sechellia*, and FILET's apparent error in this case underscores the genealogical differences
859 between adaptive gene flow and introgression events involving only neutral alleles.
860
861 **Concluding remarks**
862
863 Here we present a novel machine learning approach, FILET, that leverages population genomic
864 data from two related populations in order to determine whether a given genomic window has
865 experienced gene flow between these populations, and if so in which direction. We applied
866 FILET to a set of *D. simulans* genomes as well as a new set of whole genome sequences from the
867 closely related island endemic *D. sechellia*, confirming widespread introgression and also
868 inferring that this introgression was largely in the direction of *D. simulans* to *D. sechellia*. Future
869 work leveraging *D. simulans* data sampled from the Seychelles will be required to determine
870 whether this asymmetry is a consequence of low rate of migration of *D. simulans* back to
871 mainland Africa (where our *D. simulans* data were obtained), or whether the directionality of
872 gene flow is biased on the islands themselves. In addition to creating FILET, we devised several
873 new statistics, including the d_d statistics and Z_X which our feature rankings show to be quite
874 useful for uncovering gene flow.
875 Despite the success of FILET on both simulated data sets and real data from *Drosophila*,
876 there are several improvements that could be made. First, by framing the problem as one of
877 parameter estimation (i.e. regression) rather than classification, we may be able to precisely infer
878 the values of relevant parameters of introgression events (i.e. the time of the event and the
879 number of migrant lineages). Deep learning methods, which naturally allow for both
880 classification and regression, may prove particularly useful for this task [103]. Indeed, Sheehan
881 and Song [109] used deep learning to infer demographic parameters (regression) while
882 simultaneously identifying selective sweeps (classification). Another step we have not taken is to
883 explicitly handle adaptive introgression, which could potentially greatly improve our approach's
884 power to detect such events.
885 While population genetic inference has traditionally relied on the design of a summary
886 statistic sensitive to the evolutionary force of interest, a number of highly successful supervised
887 machine learning methods have been put forth within the last few years [54, 104-109]. These
888 methods are often thought of as black boxes, a characterization that may not always be fair [125].
889 Indeed in the context of evolutionary genetics such machine learning approaches are easily
890 interpreted as we have strong generative models that guide our intuition. Nonetheless, classical
891 statistical estimation from parametric models may often be more interpretable. Hybrid
892 approaches combining machine learning techniques with Bayesian approaches to estimate
893 posterior distributions of evolutionary parameters (e.g. [127]) thus represent an attractive
894 alternative to either approach in their "pure" form. As genomic data sets continue to grow, we

895 argue that machine learning approaches—in whatever shape they eventually take—leveraging
896 high dimensional feature spaces have the potential to revolutionize evolutionary genomic
897 inference.
898
899 **ACKNOWLEDGMENTS**
900
901 We thank Michael Lan for his work on an early iteration of this project. We also thank Joshua
902 Schraiber and two anonymous reviewers for comments on the manuscript.
903
904 **REFERENCES**
905
906 1. Mallet J. Hybridization as an invasion of the genome. Trends in ecology & evolution.
907 2005;20(5):229-37.
908 2. Whitney KD, Ahern JR, Campbell LG, Albert LP, King MS. Patterns of hybridization in
909 plants. Perspectives in Plant Ecology, Evolution and Systematics. 2010;12(3):175-82.
910 3. Barton NH. The role of hybridization in evolution. Mol Ecol. 2001;10(3):551-68.
911 4. Tung J, Barreiro LB. The contribution of admixture to primate evolution. Current opinion
912 in genetics & development. 2017;47:61-8.
913 5. Baack EJ, Rieseberg LH. A genomic view of introgression and hybrid speciation. Current
914 opinion in genetics & development. 2007;17(6):513-8.
915 6. Goulet BE, Roda F, Hopkins R. Hybridization in plants: old ideas, new techniques. Plant
916 Physiol. 2017;173(1):65-78.
917 7. Gladieux P, Ropars J, Badouin H, Branca A, Aguileta G, Vienne DM, et al. Fungal
918 evolutionary genomics provides insight into the mechanisms of adaptive divergence in
919 eukaryotes. Mol Ecol. 2014;23(4):753-73.
920 8. Schardl C, Craven K. Interspecific hybridization in plant-associated fungi and oomycetes:
921 a review. Mol Ecol. 2003;12(11):2861-73.
922 9. Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. Speciation and introgression
923 between Mimulus nasutus and Mimulus guttatus. PLoS Genet. 2014;10(6):e1004410.
924 10. Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, et al. Population
925 genomics: whole-genome analysis of polymorphism and divergence in Drosophila
926 simulans. PLoS Biol. 2007;5(11):e310.
927 11. Kulathinal RJ, Stevison LS, Noor MA. The genomics of speciation in Drosophila:
928 diversity, divergence, and introgression estimated using low-coverage genome
929 sequencing. PLoS Genet. 2009;5(7):e1000550.
930 12. Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, et al.
931 Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome
932 Res. 2013;23(11):1817-28.

933 13. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, et al.
934 Extensive introgression in a malaria vector species complex revealed by phylogenomics.
935 Science. 2015;347(6217):1258524.

936 14. Nürnberger B, Lohse K, Fijarczyk A, Szymura JM, Blaxter ML. Para-allopatry in
937 hybridizing fire-bellied toads (Bombina bombina and B. variegata): Inference from
938 transcriptome-wide coalescence analyses. Evolution. 2016;70(8):1803-18.

939 15. Rothfels CJ, Johnson AK, Hovenkamp PH, Swofford DL, Roskam HC, Fraser-Jenkins
940 CR, et al. Natural hybridization between genera that diverged from each other
941 approximately 60 million years ago. The American Naturalist. 2015;185(3):433-42.

942 16. Nadeau NJ, Ruiz M, Salazar P, Counterman B, Medina JA, Ortiz-Zuazaga H, et al.
943 Population genomics of parallel hybrid zones in the mimetic butterflies, H. melpomene
944 and H. erato. Genome Res. 2014;24(8):1316-33.

945 17. Turissini DA, Matute DR. Fine scale mapping of genomic introgressions within the
946 Drosophila yakuba clade. bioRxiv. 2017:152421.

947 18. Bachtrog D, Thornton K, Clark A, Andolfatto P. Extensive introgression of
948 mitochondrial DNA relative to nuclear genes in the Drosophila yakuba species group.
949 Evolution. 2006;60(2):292-302.

950 19. Leavitt DH, Marion AB, Hollingsworth BD, Reeder TW. Multilocus phylogeny of
951 alligator lizards (Elgaria, Anguidae): Testing mtDNA introgression as the source of
952 discordant molecular phylogenetic hypotheses. Mol Phylogenet Evol. 2017;110:104-21.

953 20. Sarver BA, Demboski JR, Good JM, Forshee N, Hunter SS, Sullivan J. Comparative
954 phylogenomic assessment of mitochondrial introgression among several species of
955 chipmunks (Tamias). Genome Biol Evol. 2016;9(1):7-19.

956 21. Carneiro M, Albert FW, Afonso S, Pereira RJ, Burbano H, Campos R, et al. The genomic
957 architecture of population divergence between subspecies of the European rabbit. PLoS
958 Genet. 2014;10(8):e1003519.

959 22. Maroja LS, Larson EL, Bogdanowicz SM, Harrison RG. Genes with restricted
960 introgression in a field cricket (Gryllus firmus/Gryllus pennsylvanicus) hybrid zone are
961 concentrated on the X chromosome and a single autosome. G3: Genes, Genomes,
962 Genetics. 2015;5(11):2219-27.

963 23. Muirhead CA, Presgraves DC. Hybrid incompatibilities, local adaptation, and the
964 genomic distribution of natural introgression between species. The American Naturalist.
965 2016;187(2):249-61.

966 24. Phifer-Rixey M, Bomhoff M, Nachman MW. Genome-wide patterns of differentiation
967 among house mouse subspecies. Genetics. 2014;198(1):283-97.

968 25. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft
969 sequence of the Neandertal genome. Science. 2010;328(5979):710-22.

970 26. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The
971 genomic landscape of Neanderthal ancestry in present-day humans. Nature.
972 2014;507(7492):354-7.

973 27. Turner TL, Hahn MW, Nuzhdin SV. Genomic islands of speciation in Anopheles
974 gambiae. PLoS Biol. 2005;3(9):e285.

975 28. Harris K, Nielsen R. The genetic cost of Neanderthal introgression. Genetics.
976 2016;203(2):881-91.

977 29. Juric I, Aeschbacher S, Coop G. The strength of selection against Neanderthal
978 introgression. PLoS Genet. 2016;12(11):e1006340.

979 30. Hedrick PW. Adaptive introgression in animals: examples and comparison to new
980 mutation and standing variation as sources of adaptive variation. Mol Ecol.
981 2013;22(18):4606-18.

982 31. Norris LC, Main BJ, Lee Y, Collier TC, Fofana A, Cornel AJ, et al. Adaptive
983 introgression in an African malaria mosquito coincident with the increased usage of
984 insecticide-treated bed nets. Proceedings of the National Academy of Sciences.
985 2015;112(3):815-20.

986 32. Pardo-Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo-Ready W, Joron M, et al.
987 Adaptive introgression across species boundaries in Heliconius butterflies. PLoS Genet.
988 2012;8(6):e1002752.

989 33. Song Y, Endepols S, Klemann N, Richter D, Matuschka F-R, Shih C-H, et al. Adaptive
990 introgression of anticoagulant rodent poison resistance by hybridization between old
991 world mice. Curr Biol. 2011;21(15):1296-301.

992 34. Bechsgaard J, Jorgensen TH, Schierup MH. Evidence for Adaptive Introgression of
993 Disease Resistance Genes Among Closely Related Arabidopsis Species. G3: Genes,
994 Genomes, Genetics. 2017;7(8):2677-83.

995 35. Cheeseman K, Ropars J, Renault P, Dupont J, Gouzy J, Branca A, et al. Multiple recent
996 horizontal transfers of a large genomic region in cheese making fungi. Nature
997 Communications. 2014;5:2876.

998 36. Huerta-Sánchez E, Jin X, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, et al. Altitude
999 adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature.
1000 2014;512(7513):194-7.

1001 37. Melo MC, Salazar C, Jiggins CD, Linares M. Assortative mating preferences among
1002 hybrids offers a route to hybrid speciation. Evolution. 2009;63(6):1660-5.

1003 38. Salazar C, Baxter SW, Pardo-Diaz C, Wu G, Surridge A, Linares M, et al. Genetic
1004 evidence for hybrid trait speciation in Heliconius butterflies. PLoS Genet.
1005 2010;6(4):e1000930.

1006 39. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in
1007 unrelated individuals. Genome Res. 2009;19(9):1655-64.

1008 40. Anderson E, Thompson E. A model-based method for identifying species hybrids using
1009 multilocus genetic data. Genetics. 2002;160(3):1217-29.

1010 41. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus
1011 genotype data. Genetics. 2000;155(2):945-59.

1012 42. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide
1013 allele frequency data. PLoS Genet. 2012;8(11):e1002967.

1014 43. Guan Y. Detecting structure of haplotypes and local ancestry. Genetics. 2014;196(3):625-
1015 42.

1016 44. Tang H, Coram M, Wang P, Zhu X, Risch N. Reconstructing genetic ancestry blocks in
1017 admixed individuals. The American Journal of Human Genetics. 2006;79(1):1-12.

1018 45. Sohn K-A, Ghahramani Z, Xing EP. Robust estimation of local genetic ancestry in
1019 admixed populations using a nonparametric Bayesian approach. Genetics.
1020 2012;191(4):1295-308.

1021 46. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using
1022 dense haplotype data. PLoS Genet. 2012;8(1):e1002453.

1023 47. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, et al. Sensitive
1024 detection of chromosomal segments of distinct ancestry in admixed populations. PLoS
1025 Genet. 2009;5(6):e1000519.

1026 48. Wright S. The genetical structure of populations. Ann Hum Genet. 1949;15(1):323-54.

1027 49. Nei M, Li W-H. Mathematical model for studying genetic variation in terms of restriction
1028 endonucleases. Proceedings of the National Academy of Sciences. 1979;76(10):5269-73.

1029 50. Joly S, McLenachan PA, Lockhart PJ. A statistical approach for distinguishing
1030 hybridization and incomplete lineage sorting. The American Naturalist. 2009;174(2):E54-
1031 E70.

1032 51. Geneva AJ, Muirhead CA, Kingan SB, Garrigan D. A new method to scan genomes for
1033 introgression in a secondary contact model. PLoS ONE. 2015;10(4):e0118621.

1034 52. Rosenzweig BK, Pease JB, Besansky NJ, Hahn MW. Powerful methods for detecting
1035 introgressed regions from population genomic data. Mol Ecol. 2016;25(11):2387-97. doi:
1036 10.1111/mec.13610.

1037 53. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning.
1038 2006;63(1):3-42.

1039 54. Schrider DR, Kern AD. S/HIC: Robust Identification of Soft and Hard Sweeps Using
1040 Machine Learning. PLoS Genet. 2016;12(3): e1005928.

1041 55. Schrider DR, Kern AD. Soft sweeps are the dominant mode of adaptation in the human
1042 genome. Mol Biol Evol. 2017;34(8):1863–77. Epub 2017/05/10. doi:
1043 10.1093/molbev/msx154. PubMed PMID: 28482049.

1044 56. Jones CD. The genetic basis of Drosophila sechellia's resistance to a host plant toxin.
1045 Genetics. 1998;149(4):1899-908.

1046 57. Jones CD. The genetics of adaptation in Drosophila sechellia. Genetica. 2005;123(1-
1047 2):137.

1048 58. Louis J, David J. Ecological specialization in the Drosophila melanogaster species
1049 subgroup: a case study of D. sechellia. Acta oecologica Oecologia generalis.
1050 1986;7(3):215-29.

1051 59. Farine J-P, Legal L, Moreteau B, Le Quere J-L. Volatile components of ripe fruits of
1052 Morinda citrifolia and their effects on Drosophila. Phytochemistry. 1996;41(2):433-8.

1053 60. Legal L, Chappe B, Jallon JM. Molecular basis ofMorinda citrifolia (L.): Toxicity on
1054 drosophila. J Chem Ecol. 1994;20(8):1931-43.

1055 61. Legal L, Moulin B, Jallon JM. The relation between structures and toxicity of oxygenated
1056 aliphatic compounds homologous to the insecticide octanoic acid and the chemotaxis of
1057 two species of Drosophila. Pestic Biochem Physiol. 1999;65(2):90-101.

1058 62. Andrade López J, Lanno S, Auerbach J, Moskowitz E, Sligar L, Wittkopp P, et al.
1059 Genetic basis of octanoic acid resistance in Drosophila sechellia: functional analysis of a
1060 fine-mapped region. Mol Ecol. 2017;26(4):1148-60.

1061 63. Dekker T, Ibba I, Siju K, Stensmyr MC, Hansson BS. Olfactory shifts parallel
1062 superspecialism for toxic fruit in Drosophila melanogaster sibling, D. sechellia. Curr
1063 Biol. 2006;16(1):101-9.

1064 64. Huang Y, Erezyilmaz D. The genetics of resistance to Morinda fruit toxin during the
1065 postembryonic stages in Drosophila sechellia. G3: Genes, Genomes, Genetics.
1066 2015;5(10):1973-81.

1067 65. Hungate EA, Earley EJ, Boussy IA, Turissini DA, Ting C-T, Moran JR, et al. A locus in
1068 Drosophila sechellia affecting tolerance of a host plant toxin. Genetics.
1069 2013;195(3):1063-75.

1070 66. Matsuo T, Sugaya S, Yasukawa J, Aigaki T, Fuyama Y. Odorant-binding proteins
1071 OBP57d and OBP57e affect taste perception and host-plant preference in Drosophila
1072 sechellia. PLoS Biol. 2007;5(5):e118.

1073 67. Shiao M-S, Chang J-M, Fan W-L, Lu M-YJ, Notredame C, Fang S, et al. Expression
1074 divergence of chemosensory genes between Drosophila sechellia and its sibling species
1075 and its implications for host shift. Genome Biol Evol. 2015;7(10):2843-58.

1076 68. Hey J, Kliman RM. Population genetics and phylogenetics of DNA sequence variation at
1077 multiple loci within the Drosophila melanogaster species complex. Mol Biol Evol.
1078 1993;10(4):804-22.

1079 69. Kern AD, Jones CD, Begun DJ. Molecular population genetics of male accessory gland
1080 proteins in the Drosophila simulans complex. Genetics. 2004;167(2):725-35.

1081 70. Kliman RM, Andolfatto P, Coyne JA, Depaulis F, Kreitman M, Berry AJ, et al. The
1082 population genetics of the origin and divergence of the Drosophila simulans complex
1083 species. Genetics. 2000;156(4):1913-31.

1084 71. Legrand D, Tenaillon MI, Matyot P, Gerlach J, Lachaise D, Cariou M-L. Species-wide
1085 genetic variation and demographic history of Drosophila sechellia, a species lacking
1086 population structure. Genetics. 2009;182(4):1197-206.

1087 72. Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, et al.
1088 Genome sequencing reveals complex speciation in the Drosophila simulans clade.
1089 Genome Res. 2012;22(8):1499-511.

1090 73. Matute D, Ayroles J. Hybridization occurs between Drosophila simulans and D. sechellia
1091 in the Seychelles archipelago. J Evol Biol. 2014;27(6):1057-68.

1092 74. Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR. Landscape of
1093 standing variation for tandem duplications in Drosophila yakuba and Drosophila
1094 simulans. Mol Biol Evol. 2014;31(7):1750-66.

1095 75. Feder JL, Xie X, Rull J, Velez S, Forbes A, Leung B, et al. Mayr, Dobzhansky, and Bush
1096 and the complexities of sympatric speciation in Rhagoletis. Proceedings of the National
1097 Academy of Sciences. 2005;102(suppl 1):6573-80.

1098 76. Kelly JK. A test of neutrality based on interlocus associations. Genetics.
1099 1997;146(3):1197-206.

1100 77. Breiman L. Random forests. Machine Learning. 2001;45(1):5-32.

1101 78. Quinlan JR. Induction of decision trees. Machine Learning. 1986;1(1):81-106.

1102 79. Fay JC, Wu C-I. Hitchhiking under positive Darwinian selection. Genetics.
1103 2000;155(3):1405-13.

1104 80. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA
1105 polymorphism. Genetics. 1989;123(3):585-95.

1106 81. Hudson RR, Slatkin M, Maddison W. Estimation of levels of gene flow from DNA
1107 sequence data. Genetics. 1992;132(2):583-9.

1108 82. Hudson RR. A new statistic for detecting genetic differentiation. Genetics.
1109 2000;155(4):2011-4.

1110 83. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture
1111 in human history. Genetics. 2012;192(3):1065-93.

1112 84. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-
1113 learn: Machine learning in Python. Journal of Machine Learning Research.
1114 2011;12(Oct):2825-30.

1115 85. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees: CRC
1116 press; 1984.

1117 86. Chan AH, Jenkins PA, Song YS. Genome-wide fine-scale recombination rate variation in
1118 Drosophila melanogaster. PLoS Genet. 2012;8(12):e1003090.

1119 87. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination
1120 hotspots using single-nucleotide polymorphism data. Genetics. 2003;165(4):2213-33.

1121 88. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
1122 arXiv. 2013. doi: 1303.3997.

1123 89. Hu TT, Eisen MB, Thornton KR, Andolfatto P. A second-generation assembly of the
1124 *Drosophila simulans* genome provides new insights into patterns of lineage-specific
1125 divergence. Genome Res. 2013;23(1):89-98.

1126 90. Gramates LS, Marygold SJ, Santos Gd, Urbano J-M, Antonazzo G, Matthews BB, et al.
1127 FlyBase at 25: looking to the future. Nucleic Acids Res. 2017;45(D1):D663-D71.

1128 91. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The
1129 Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA
1130 sequencing data. Genome Res. 2010;20(9):1297-303.

1131 92. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A
1132 framework for variation discovery and genotyping using next-generation DNA
1133 sequencing data. Nat Genet. 2011;43(5):491-8.

1134 93. Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al.
1135 From FastQ data to high-confidence variant calls: the genome analysis toolkit best
1136 practices pipeline. Current protocols in bioinformatics. 2013;43:11.0. 1-.0. 33.

1137 94. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease
1138 and population genetic studies. Nat Methods. 2013;10(1):5-6.

1139 95. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint
1140 demographic history of multiple populations from multidimensional SNP frequency data.
1141 PLoS Genet. 2009;5(10):e1000695.

1142 96. Jansen PW, Perez RE. Constrained structural design optimization via a parallel
1143 augmented Lagrangian particle swarm optimization approach. Computers & Structures.
1144 2011;89(13):1352-66.

1145 97. Kraft D. A software package for sequential quadratic programming: DFVLR
1146 Obersfaffeuhofen, Germany; 1988.

1147 98. Perez RE, Jansen PW, Martins JR. pyOpt: a Python-based object-oriented framework for
1148 nonlinear constrained optimization. Structural and Multidisciplinary Optimization.
1149 2012;45(1):101-18.

1150 99. Schrider DR, Shanku AG, Kern AD. Effects of Linked Selective Sweeps on
1151 Demographic Inference and Model Selection. Genetics. 2016;204(3):1207-23. doi:
1152 10.1534/genetics.116.190223.

1153 100. Pool JE. The mosaic ancestry of the Drosophila genetic reference panel and the D.
1154 melanogaster reference genome reveals a network of epistatic fitness interactions. Mol
1155 Biol Evol. 2015;32(12):3236-51.

1156 101. Schrider DR, Kern AD. Supervised Machine Learning for Population Genetics: A New
1157 Paradigm. Trends Genet. 2018:10.1016/j.tig.2017.12.005.

1158 102. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20(3):273-97.

1159 103. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.

1160 104. Lin K, Li H, Schlötterer C, Futschik A. Distinguishing positive selection from neutral
1161 evolution: boosting the performance of summary statistics. Genetics. 2011;187(1):229-
1162 44.

1163 105. Pavlidis P, Jensen JD, Stephan W. Searching for footprints of positive selection in whole-
1164 genome SNP data from nonequilibrium populations. Genetics. 2010;185(3):907-22.

1165 106. Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, et al.
1166 Hierarchical boosting: a machine-learning framework to detect and classify hard selective
1167 sweeps in human populations. Bioinformatics. 2015;31(24):3946-52.

1168 107. Ronen R, Udpa N, Halperin E, Bafna V. Learning natural selection from the site
1169 frequency spectrum. Genetics. 2013;195(1):181-93.

1170 108. Pudlo P, Marin J-M, Estoup A, Cornuet J-M, Gautier M, Robert CP. Reliable ABC model
1171 choice via random forests. Bioinformatics. 2016;32(6):859-66.

1172 109. Sheehan S, Song YS. Deep learning for population genetic inference. PLoS Comput Biol.
1173 2016;12(5):e1004845.

1174 110. Gazave E, Ma L, Chang D, Coventry A, Gao F, Muzny D, et al. Neutral genomic regions
1175 refine models of recent rapid human population growth. Proceedings of the National
1176 Academy of Sciences. 2014;111(2):757-62.

1177 111. Ewing GB, Jensen JD. The consequences of not accounting for background selection in
1178 demographic inference. Mol Ecol. 2016;25(1):135-41.

1179 112. Schrider DR, Houle D, Lynch M, Hahn MW. Rates and genomic consequences of
1180 spontaneous mutational events in Drosophila melanogaster. Genetics. 2013;194(4):937-
1181 54.

1182 113. Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, Pool JE, et al. Genomic
1183 variation in natural populations of *Drosophila melanogaster*. Genetics. 2012;192(2):533-
1184 98.

1185 114. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population
1186 structure in large SNP data sets. Genetics. 2014;197(2):573-89.

1187 115. Legrand D, Vautrin D, Lachaise D, Cariou M-L. Microsatellite variation suggests a
1188 recent fine-scale population structure of Drosophila sechellia, a species endemic of the
1189 Seychelles archipelago. Genetica. 2011;139(7):909.

1190 116. Navascués M, Legrand D, Campagne C, Cariou M-L, Depaulis F. Distinguishing
1191 migration from isolation using genes with intragenic recombination: detecting
1192 introgression in the Drosophila simulans species complex. BMC Evol Biol.
1193 2014;14(1):89.

1194 117. Obbard DJ, Maclennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM. Estimating
1195 divergence dates and substitution rates in the Drosophila phylogeny. Mol Biol Evol.
1196 2012;29(11):3459-73.

1197 118. Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter M. Analysis of the
1198 genome sequences of three Drosophila melanogaster spontaneous mutation accumulation
1199 lines. Genome Res. 2009;19:1195-201.

1200 119. Brand CL, Kingan SB, Wu L, Garrigan D. A selective sweep across species boundaries in
1201 Drosophila. Mol Biol Evol. 2013;30(9):2177-86.

1202 120. Benton R, Vannice KS, Gomez-Diaz C, Vosshall LB. Variant ionotropic glutamate
1203 receptors as chemosensory receptors in Drosophila. Cell. 2009;136(1):149-62.

1204 121. Lu H-L, Wang JB, Brown MA, Euerle C, Leger RJS. Identification of Drosophila
1205 mutants affecting defense to an entomopathogenic fungus. Scientific reports. 2015;5.

1206 122. Ekengren S, Hultmark D. A family of Turandot-related genes in the humoral stress
1207 response of Drosophila. Biochem Biophys Res Commun. 2001;284(4):998-1003.

123. Ekengren S, Tryselius Y, Dushay MS, Liu G, Steiner H, Hultmark D. A humoral stress response in Drosophila. Curr Biol. 2001;11(9):714-8.

124. Salazar-Jaramillo L, Jalvingh KM, de Haan A, Kraaijeveld K, Buermans H, Wertheim B. Inter-and intra-species variation in genome-wide gene expression of Drosophila in response to parasitoid wasp attack. BMC Genomics. 2017;18(1):331.

125. Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science. 2001;16(3):199-231.

126. Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, MÃžller K-R. How to explain individual classification decisions. Journal of Machine Learning Research. 2010;11(Jun):1803-31.

127. Blum MG, François O. Non-linear regression models for Approximate Bayesian Computation. Statistics and Computing. 2010;20(1):63-73.

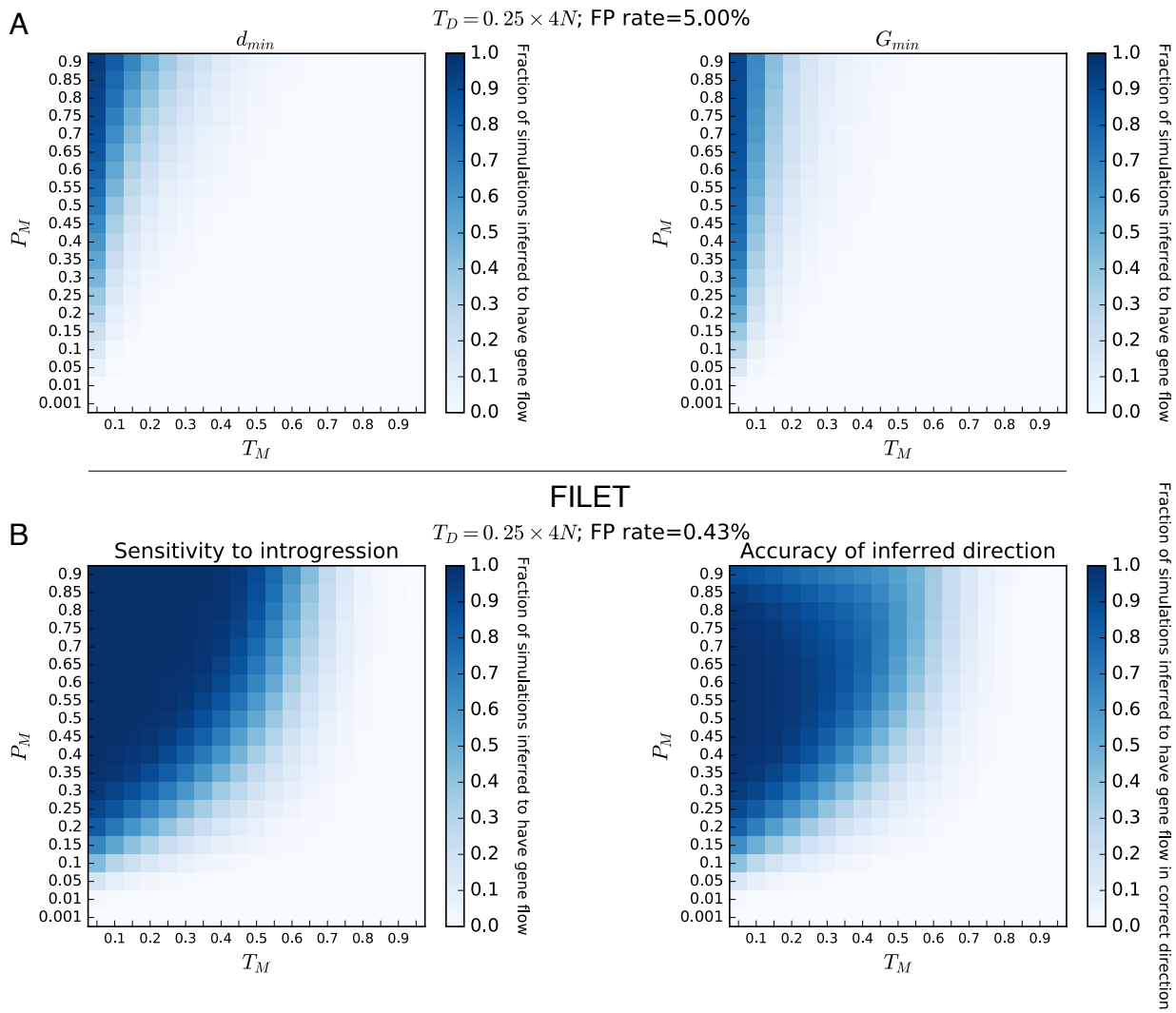

Fig. 1. Heatmaps showing several methods' sensitivity to detect introgression. We show the fraction of simulated genomic regions with introgression occurring under various combinations of migration times (T_M, shown as a fraction of the population divergence time T_D) and intensities (P_M, the probability that a given lineage will be included in the introgression event) that are detected successfully by each method. (*A*) Accuracy of d_{min} and G_{min} statistics, where a simulated region is classified as introgressed if the values of these statistics are found in the lower 5% tail of the distribution under complete isolation (from simulations). Thus, the false positive rate is fixed at 5%. (*B*) The accuracy of FILET on these same simulations. On the left we show the fraction of regions correctly classified as introgressed (compare to panel *A*). On the right, we show the fraction of all simulated regions that are not only classified as introgressed, but also for which the direction of gene flow was correctly inferred (i.e. if the direction is inferred with 100% accuracy for a given cell in the heatmap, the color shade of that cell will be identical to that in the heatmap on the left). The false positive rate, as determined from applying FILET to a simulated test set with no migration, is also shown.

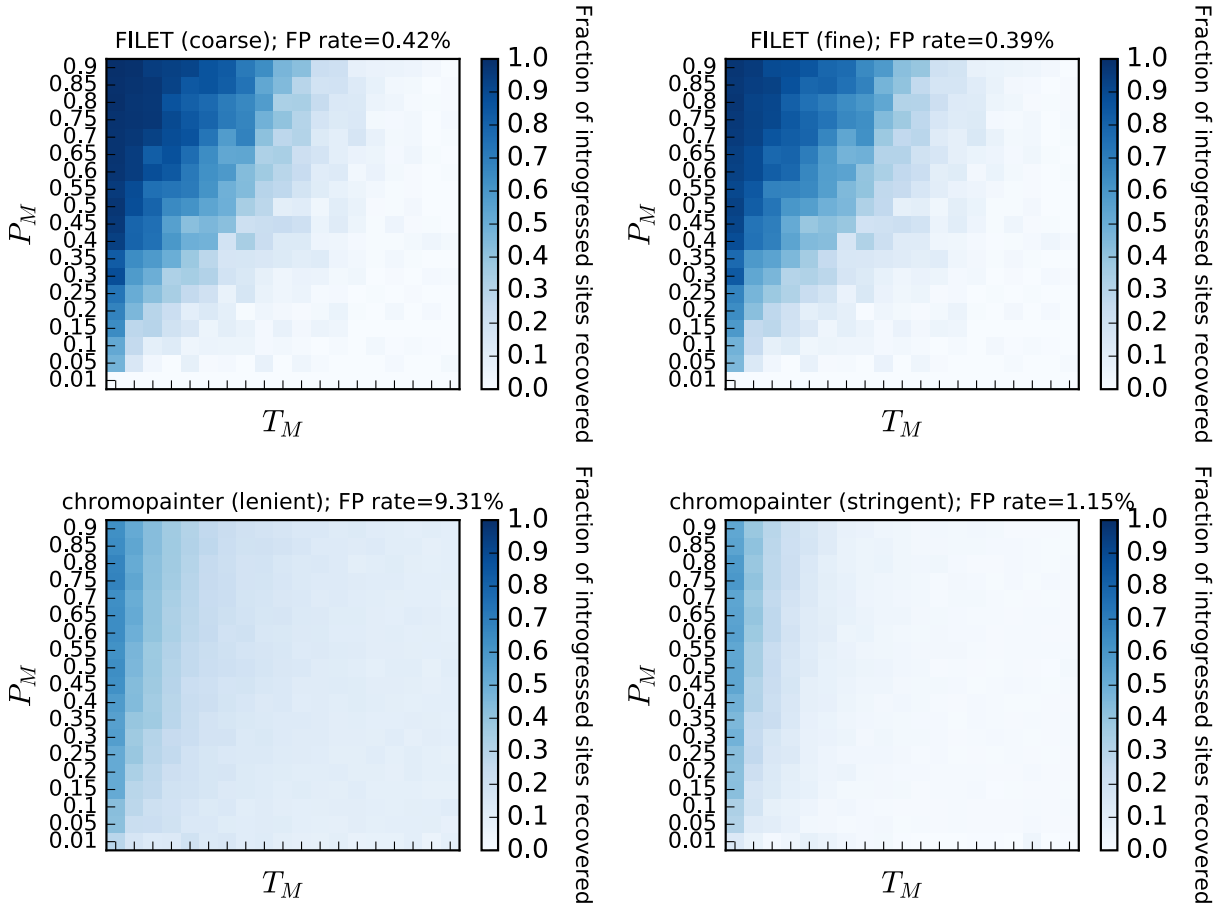

Fig. 2. A comparison of the power and resolution of FILET and ChromoPainter using simulations of a 1 Mb chromosome where introgression was allowed within the central 100 kb region. As in Figure 1, the population split time was set to *N* generations ago, and the darkness of the heatmap shows sensitivity to introgression. Unlike Figure 1, here we are measuring sensitivity at the level of the individual base pair rather than evaluating the question of whether a window at large was recovered as containing introgressed alleles. The "coarse" version of FILET refers to a FILET classifier trained to detect introgression in 10 kb windows, which was applied to sliding windows (1 kb step size) across the chromosome. The "fine" version of FILET applied a classifier trained on 1 kb windows to sliding windows (100 bp step size) within those regions classified as introgressed by the FILET classifier. The lenient version of ChromoPainter required evidence of introgression at a single SNP to identify introgression, while the stringent version required candidate regions to contain at least 25 consecutive SNPs supporting introgression.

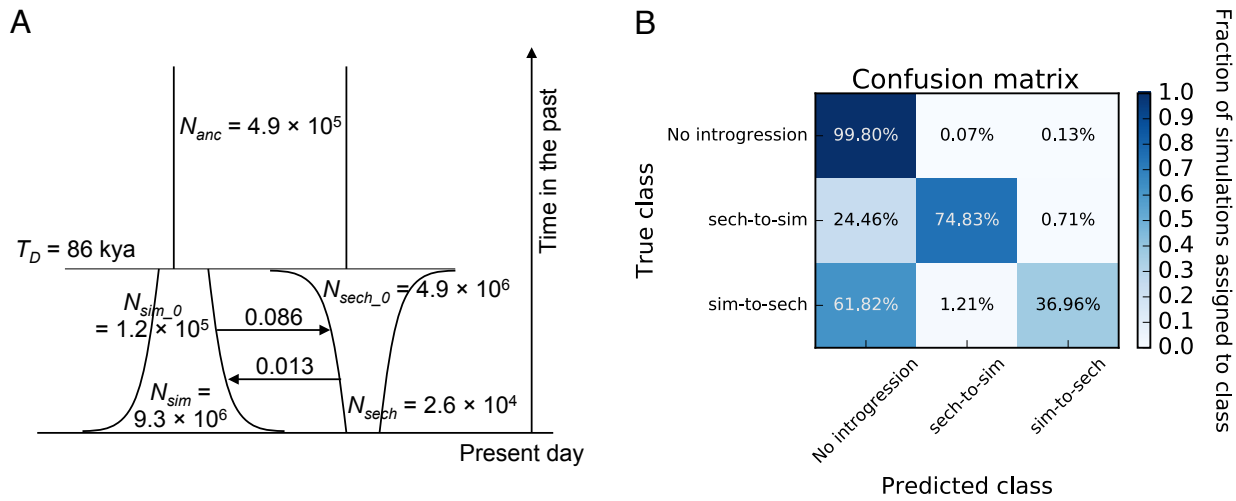

Fig. 3. Inferred joint population history of *D. simulans* and *D. sechellia*, and power to detect introgression under this model. (*A*) The parameterization of our best-fitting demographic model. Migration rates are shown by arrows, and are in units of $2 \times N_{anc} m$, where m is the probability of migration per individual in the source population per generation. (*B*) Confusion matrix showing FILET's classification accuracy under this model as assessed on an independent simulated test set. Perfect accuracy would be 100% along the entire diagonal from top-left to bottom-right, and the false positive rate is the sum of top-middle and top-right cells.

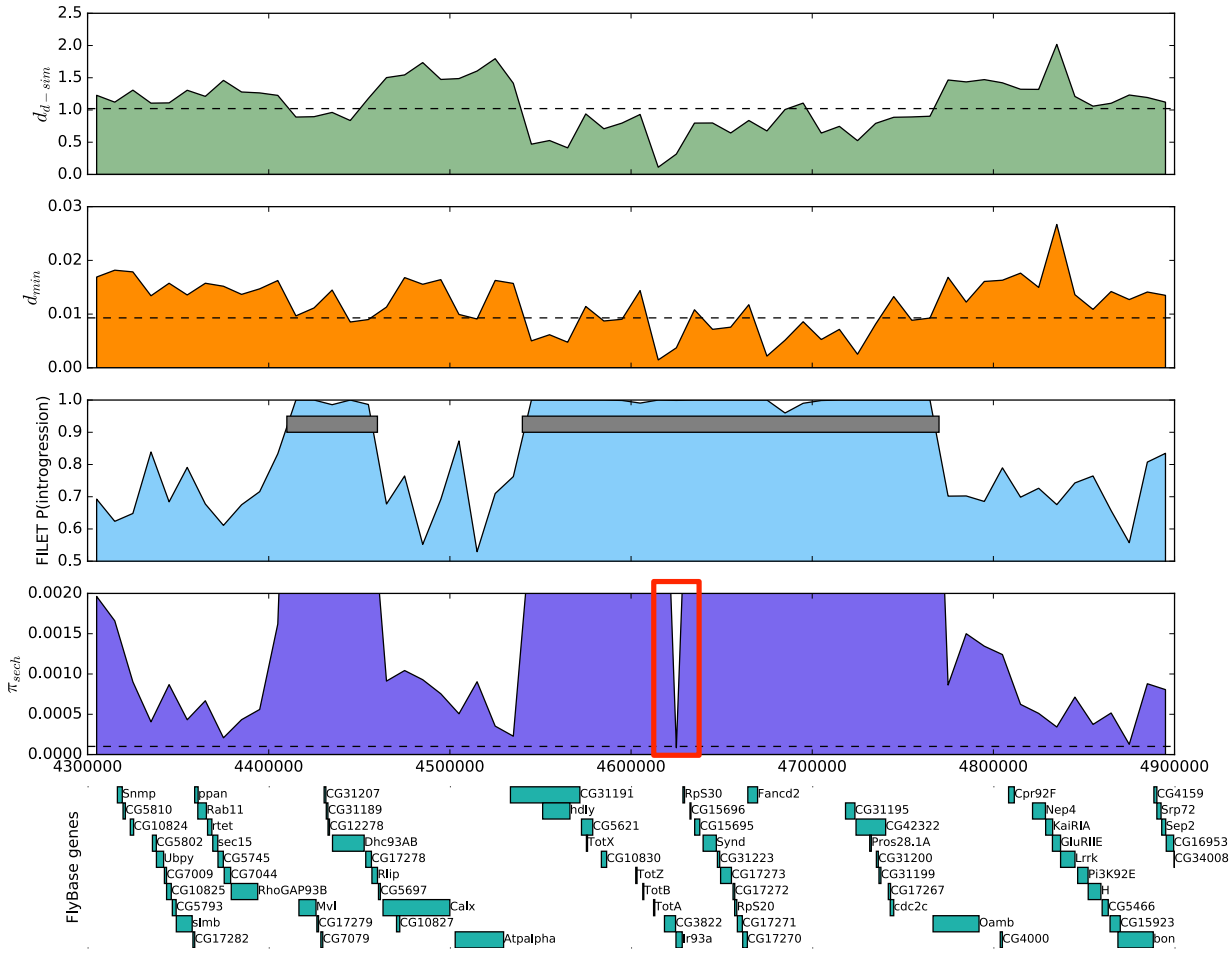

Fig. 4. A large genomic region on 3R classified by FILET as introgressed from *D. simulans* to *D. sechellia*. Values of the $d_{d\text{-}sim}$ and d_{min} (upper two panels) within each 10 kb window in the region are shown, along with the posterior probability of introgression from FILET (i.e. 1 – *P*(no introgression)). Clustered regions classified as introgressed are shown as gray rectangles superimposed over these probabilities. Also shown are windowed values of π in *D. sechellia*, with the sweep region highlighted in red, and the locations of annotated genes with associated FlyBase identifiers [90].

SUPPLEMENTAL FIGURE AND TABLE LEGENDS

S1 Fig. Illustration of the difference in values of the d_{min} statistic calculated from joint population samples with and without introgression.

S2 Fig. Violin plots showing the values of d_{min}, all four d_d statistics, and Z_X under simulated scenarios including introgression or lacking it for each values of T_D. The values of these statistics were obtained from the training data sets described in the Materials and Methods.

S3 Fig. Heatmaps showing several methods' sensitivity to detect introgression. Same as Figure 1, but for other values of T_D. (*A*) Accuracy for d_{min} and G_{min} when $T_D = 1\times4N$ generations. (*B*) Accuracy of FILET when $T_D = 1\times4N$. (*C*) and (*D*) show the same when $T_D = 4\times4N$. (E) and (F) show the same when $16\times4N$.

S4 Fig. Heatmaps showing FILET's sensitivity to introgression from an unsampled ghost population. (*A*) Sensitivity when $T_D = 0.25\times4N$ generations. (*B*) Sensitivity when $T_D = 1\times4N$ generations. (*C*) $T_D = 4\times4N$ generations. (*D*) $T_D = 16\times4N$.

S5 Fig. ROC curves showing power of FILET, d_{min} and G_{min} under each value of T_D. In order to generate these curves we transformed the classification task into a binary one: discriminating between isolation and introgression in either direction. (*A*) $T_D = 0.25\times4N$ generations. (*B*) $T_D = 1\times4N$ generations. (*C*) $T_D = 4\times4N$. (*D*) $T_D = 16\times4N$. Training and test sets for these problems contained equal numbers of examples of introgression from population 1 into 2 and introgression from population 2 into 1.

S6 Fig. ROC curves showing power of versions of FILET trained with decreasing numbers of training instances (ranging from 100 to 10000 for each class). (*A*) $T_D = 0.25\times4N$ generations. (*B*) $T_D = 1\times4N$ generations. (*C*) $T_D = 4\times4N$. (*D*) $T_D = 16\times4N$.

S7 Fig. ROC curve showing FILET's power when trained and tested on simulated examples with different window sizes with $T_D = 0.25\times4N$ generations.

S8 Fig. A comparison of the positive predictive value of FILET and ChromoPainter on the same simulated data used for Fig 2. The "coarse" and "fine" versions of FILET, and lenient and stringent versions of ChromoPainter's predictions, are as defined for Fig 2. In cases where the positive predictive value is undefined (i.e. no base pairs were predicted to be introgressed), it is displayed as zero (i.e. a white cell in the heatmap).

S9 Fig. Population structure within *D. sechellia.* (*A*) The top three principal components of all *D. sechellia* diploid genomes. The cluster on the left shows the individuals from Praslin, while the

cluster on the right shows all other individuals. Note that the cluster on the right is far less dispersed due to the very small amount of polymorphism among these individuals. The numbers in parentheses on each axis show the fraction of the variance explained by each principal component. (*B*) Results of running fastStructure on our *D. sechellia* samples with the number of subpopulations (*K*) ranging from 2 to 8.

S10 Fig. Site frequency spectra of *D. sechellia* samples from Praslin, *D. sechellia* samples from all other locations, and *D. simulans* samples. The *D. sechellia* samples were both downsampled to n=12 as described in the text, while *D. simulans* was downsampled to n=18 (i.e. the same sample sizes used for our demographic inference). These SFS show the fraction of all polymorphisms found in each bin rather than the raw number of polymorphisms, and thus do not contain information about the total number of SNPs. As described in the text, there is >12-fold more polymorphism in the Praslin samples than in the non-Praslin samples.

S11 Fig. Confusion matrix showing FILET's classification accuracy when trained under out inferred model of the *simulans-sechellia* joint demographic history, but applied to test data generated under a different model (described in Materials and Methods and shown in S4 Table). under this model as assessed on an independent simulated test set. Perfect accuracy would be 100% along the entire diagonal from top-left to bottom-right, and the false positive rate is the sum of top-middle and top-right cells.

S1 Table. Results when applying FILET to simulations with constant bidirectional migration. 10000 simulated replicates were tested for each parameter combination.

S2 Table. Feature importance and rankings for each classifier used in this study.

S3 Table. Sampling location, sequencing/mapping statistics, and SRA identifiers for each genome included in this study.

S4 Table. Demographic parameter estimates inferred by ∂a∂i, along with a simple naïve model.