VolcanoFinder

genomic scans for adaptive introgression

Derek Setter[1][2][◕][¤], Sylvain Mousset[1][◕], Xiaoheng Cheng[3], Rasmus Nielsen[4], Michael DeGiorgio[5][6][‡], Joachim Hermisson[1][7][‡],

1 Department of Mathematics, University of Vienna, Vienna, Austria
2 School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom
3 Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA, USA
4 Departments of Integrative Biology and Statistics, University of California, Berkeley, CA, USA
5 Departments of Biology and Statistics, Pennsylvania State University, University Park, PA, USA
6 Institute for CyberScience, Pennsylvania State University, University Park, PA, USA
7 Max F. Perutz Laboratories, University of Vienna, Vienna, Austria

◕These authors contributed equally to this work.
‡These authors also contributed equally to this work.
¤Current Address: School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom
*Correspondence: joachim.hermisson@univie.ac.at (JH), mxd60@psu.edu (MD)

Abstract

Recent research shows that introgression between closely-related species is an important source of adaptive alleles for a wide range of taxa. Typically, detection of adaptive introgression from genomic data relies on comparative analyses that require sequence data from both the recipient and the donor species. However, in many cases, the donor is unknown or the data is not currently available. Here, we introduce a genome-scan method—`VolcanoFinder`—to detect recent events of adaptive introgression using polymorphism data from the recipient species only.

`VolcanoFinder` detects adaptive introgression sweeps from the pattern of excess intermediate-frequency polymorphism they produce in the flanking region of the genome, a pattern which appears as a volcano-shape in pairwise genetic diversity.

Using coalescent theory, we derive analytical predictions for these patterns. Based on these results, we develop a composite-likelihood test to detect signatures of adaptive introgression relative to the genomic background. Simulation results show that `VolcanoFinder` has high statistical power to detect these signatures, even for older sweeps and for soft sweeps initiated by multiple migrant haplotypes. Finally, we implement `VolcanoFinder` to detect archaic introgression in European and sub-Saharan African human populations, and uncovered interesting candidates in both populations, such as *TSHR* in Europeans and *TCHH-RPTN* in Africans. We discuss their biological implications and provide guidelines for identifying and circumventing artifactual signals during empirical applications of `VolcanoFinder`.

Author summary

The process by which beneficial alleles are introduced into a species from a closely-related species is termed adaptive introgression. We present an analytically-tractable model for the effects of adaptive introgression on non-adaptive genetic variation in the genomic region surrounding the beneficial allele. The result we describe is a characteristic volcano-shaped pattern of increased variability that arises around the positively-selected site, and we introduce an open-source method `VolcanoFinder` to detect this signal in genomic data. Importantly, `VolcanoFinder` is a population-genetic likelihood-based approach, rather than a comparative-genomic approach, and can therefore probe genomic variation data from a single population for footprints of adaptive introgression, even from *a priori* unknown and possibly extinct donor species.

Introduction

While classic species concepts imply genetic isolation [1], research of the past 30 years shows that hybridization between closely related species (or diverged subspecies) is widespread [2]. For adaptation research, this offers the intriguing perspective of an exchange of key adaptations between related species, with potentially important implications for our view of the adaptive process. Indeed, recent studies have brought clear evidence of cross-species introgression of advantageous alleles [3–6]. Well-documented examples cover a wide range of taxa, including the transfer of wing-pattern mimicry genes in *Heliconius* butterflies [7], herbivore resistance and abiotic tolerance genes in wild sunflowers [8, 9], pesticide resistance in mice [10] and mosquitoes [11], and new mating and vegetative incompatibility types in an invasive fungus [12]. Such adaptive introgressions also occurred in modern humans [13–15]: local adaptation to hypoxia at high-altitude was shown to be associated with selection for a

Denisovan-related haplotype at the *EPAS1* hypoxia pathway gene in Tibetan populations [16]; positive selection has been characterized for three archaic haplotypes, independently introgressed from Denisovans or Neanderthals in a cluster of genes involved in the innate immune response [17], and immunity related genes show evidence of selection for Neanderthal and Denisovan haplotypes [18, 19].

In all examples above, evidence of adaptive introgression rests on a comparative analysis of DNA from both donor and recipient species. In particular, studies in humans often rely on maps of introgressed Neanderthal or Denisovan fragments in the modern human genome [20–22]. The tell-tale signature of adaptive introgression is a segment of mutations from the donor population that is present in strong LD and in high frequency in the recipient population [13, 16]. Unfortunately, good data from a potential donor species may not always be available, especially in the case of an extinct donor. In the absence of a donor, introgression can sometimes be inferred from haplotype statistics in the recipient species [23, 24], the most recent methods making use of machine learning algorithms based on several statistics [25]. However, as observed in [13], there is currently no framework for a joint inference of admixture and selection, such as adaptive introgression, and selection is usually inferred from the unexpectedly high frequency of introgressed haplotypes [13, 19–22, 24, 26]. A recent article [27] on adaptive introgression in plants identified four different types of studies in this field, focusing on (i) introgression, (ii) genomic signatures of selection, (iii) adaptively relevant phenotypic variation, and (iv) fitness. Our work aims to bridge the gap between classes (i) and (ii), and detect the specific genomic signature of an introgression sweep.

The genomic signature of adaptation from a *de novo* beneficial mutation has been extensively studied. When such an allele fixes in the population, the neutral alleles initially physically linked to it hitchhike to high frequency, whereas those that are initially not linked to it might be rescued from extinction by recombination, creating a valley in heterozygosity around the selected allele—the classical pattern of a hard selective sweep [28–30]. If selection acts on standing genetic variation [31, 32], or if beneficial alleles enter the population recurrently through mutation or migration [33, 34], then multiple haplotypes from the ancestral population may survive the sweep, leading to distinctive patterns of *soft sweeps* [35, 36] with a more shallow sweep valley and typically a much weaker footprint. Recurrent hybridization may also cause a soft sweep in this sense.

In structured populations, theoretical studies have mostly focused on local adaptation and the effect of hitchhiking on differentiation indices [37–40]. However, a particularly relevant result in the context of adaptive introgression involves a structured population model with two demes connected by low migration [38]. As observed there, the pattern of a classical selective sweep is only reproduced in the subpopulation where the selected allele first arises, whereas it is highly different in the second subpopulation where the adaptive allele is later introduced by a migration event. In the latter population, heterozygosity is also reduced around the focal site, but this valley is surrounded by regions of increased heterozygosity, in which allelic variants from both subpopulations persist at intermediate frequencies.

Statistical methods to detect selective sweeps make use of patterns in both diversity within populations and differentiation among populations [41, 42]. Several widely-used tests require comparative data from two or more populations. Tests like `XP-CLR` [43] and `hapFLK` [44] can detect even soft sweeps under simple population structure and low migration rates [45]. Another family of model-based genome-scan methods identifies the effects of selection from the site frequency spectrum (SFS) and requires data from only a single population (and potentially an outgroup sequence). Using the composite-likelihood scheme suggested in [46], the `SweepFinder` software [47] detects local effects of positive selection on the SFS relative to the genome-wide genetic background SFS. The method was later extended to detect long term balancing selection [48, BALLET] and improved to include fixed differences in addition to polymorphic sites [49, SweepFinder2]. These methods compare how well two models fit the local SFS: a null model that assumes a genome-wide homogeneous SFS, and an alternative model that assumes selection acts at the focal locus. High detection power relies on modelling the specific effect of selection on the SFS for the alternative model (test 2 vs. test 1 in [47] and [48]).

The footprint of adaptive introgression, like sweeps from migration [38], differs strongly from the classical pattern of both hard or soft sweeps. The signal of adaptive introgression may therefore remain undetected by classical methods. Moreover, we are interested in distinguishing cases of adaptive introgression from adaptation within a species. For these reasons, we developed VolcanoFinder, a specialized method capable of detecting adaptive introgression when data from only the recipient species is available. The software and user manual are available at
http://www.personal.psu.edu/mxd60/vf.html.

The article is organized as follows. As a first step, we use a coalescent approach to model a recent introgression sweep in the recipient population after secondary contact with a possibly-unknown donor species. We use these results to characterize the introgression footprint by two parameters, one measuring the selection strength and the other, divergence to the donor. In the second step, these parameters are included in an extended composite-likelihood scheme, built on SweepFinder2 [50]. We use simulated data to assess the power of our method and compare it to that of SweepFinder2 and BALLET. Finally, we apply VolcanoFinder to human data sets in order to detect introgression sweeps in both the ancestral African and Central European populations, and we identify and discuss several candidate regions for each.

Results

Model and analysis

Evolutionary History

We consider a model with three species named *recipient*, *donor*, and *outgroup*, and their common ancestor species (see Fig. 1). We assume a diploid population size N for the recipient and the common ancestor, and size N' for the donor. All species evolve according to a Wright-Fisher model. The recipient and the donor species diverged at time T_d before present, their ancestor and the outgroup diverged at time $T_{sp} \geqslant T_d$. All times are measured pastward from the time of sampling in units of $4N$ generations. We assume an infinite sites model and complete lineage sorting in the ancestor. Polymorphic sites in the recipient species are polarized, *e.g.*, with the help of the outgroup. If a second, more distant outgroup is available, then we also assume that fixed differences between the recipient species and the first outgroup are polarized. With a mutation rate (per nucleotide and generation) of μ, and $\theta = 4N\mu$, the expected divergence between the recipient and the donor species is $D = 2\left(T_d + \frac{1}{2}\right)\theta$, and the expected divergence between the recipient species and its most recent common ancestor (MRCA) with the outgroup is $D_o = \left(T_{sp} + \frac{1}{2}\right)\theta$. If polarization of the fixed differences is unknown, then the full divergence $D'_o = 2D_o$ between the recipient and the outgroup species may be used instead. At time $T_i \ll T_d$, the donor and recipient species came into secondary contact, allowing for a single bout of introgression from the donor into the recipient.

Selection acts on a single locus with two alleles B (derived) and b (ancestral). The B allele is beneficial with selection coefficient $s > 0$ for Bb heterozygotes and $2s$ for BB homozygotes. We assume that, prior to introgression, the B allele is fixed in the donor population, but the ancestral b allele is fixed in the recipient. After introgression, the B allele survives stochastic loss and rises to fixation in the recipient species, sweeping away local genetic variation and pulling in foreign genetic variation in its wake. A sample of n lineages from the recipient population and one lineage from the distant outgroup is sampled at the time of observation, after the fixation of the beneficial allele. We model the effect of this recent introgression sweep on the polymorphism and divergence pattern at a neighbouring neutral locus, at distance d from the selected allele.

Structured coalescent approximation

We implemented the full model using both individual-based and coalescent-based simulations (see *Materials and Methods*). In order to describe the key features of the selection footprint to be included into a likelihood ratio test, we used a simple analytical model based on a structured coalescent approach. The genealogy at the focal neutral locus of a sample taken from the recipient population is structured by both selection and demography. Backward in time, the coalescence process is first structured by the effects of positive selection, where we distinguish lineages that are associated with alleles b and B at the selected locus, like in a classical sweep model. At the time of introgression, all B lineages move to the donor population, while all b lineages stay in the recipient population. The further history then follows a demographic model of divergence without migration. This separation into a brief period of positive selection and a long demographic phase allows for an efficient approximation.

For simplicity, we assume in the analytical model that the sweep is initiated by a single donor haplotype. Equivalently, we can assume that all B lineages quickly coalesce in the donor population (due to a bottleneck or recent origin of the B allele). As a consequence, we only need to follow a single ancestral lineage in the donor population and the donor population size does not enter the results.

Star-like approximation. During the selective phase, the B allele sweeps through the population following a frequency trajectory $X[t]$. At a neutral locus at recombination distance $R = rd$ from the selected site, any pair of lineages linked to the B allele may coalesce at rate $\frac{1}{2NX[t]}$, while any lineage may recombine to the b background at rate $R(1 - X[t])$ per generation [51]. Generally, $X[t]$ is a stochastic trajectory, but in large populations and for strong selection it is well approximated by a deterministic curve following logistic growth, $\dot{x}(t) = 4Nsx(t)(1 - x(t))$, where $x(0) = 1/(2N)$. In this case, any lineage at distance d from the selected locus may escape the selective sweep by recombining to the b background with the probability [47, 51]

$$P_e = 1 - e^{-\alpha d}, \tag{1}$$

with $\alpha = \frac{r}{s}\ln(2N)$. For strong selection, lineages recombine independently to the b background, so that the probability that exactly k lineages among n escape the sweep is given by the binomial distribution [47]:

$$P_e(k|\alpha, d) = \binom{n}{k} P_e^k (1 - P_e)^{n-k}. \tag{2}$$

The $n - k$ lineages that do not escape the sweep coalesce instantaneously to the single ancestral lineage on which the beneficial B allele first appeared. This star-like assumption ignores coalescence in the B background followed by recombination into the b background, but it permits an analytical approximation for the genealogical effects of the sweep even for large sample sizes n.

Demographic phase. Prior to the introgression event, coalescence of the remaining $k + 1$ lineages is structured demographically. The k lineages coalesce neutrally in the recipient population, while coalescence with the single ancestral B lineage only occurs once the lineages have traced back to the common ancestral population. For our analytical analysis, we make the simplifying assumption that the neutral coalescence of the k escaped lineages occurs before finding a common ancestor with lineage tracing through the donor population. That is, we assume complete lineage sorting. Note that this assumption does not affect predictions of genetic diversity, which rely on a sample of $n = 2$ individuals.

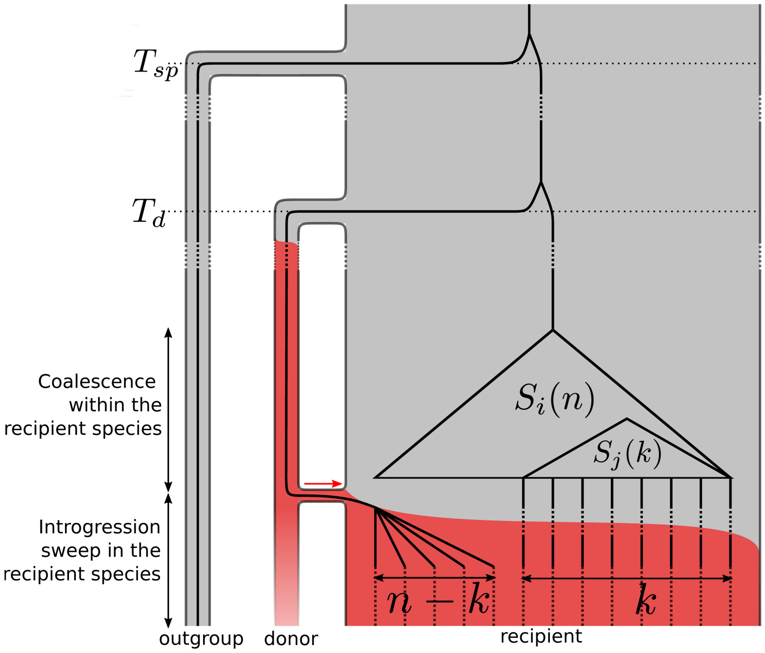
T_{sp}
T_d
$S_i(n)$
$S_j(k)$
Coalescence within the recipient species
Introgression sweep in the recipient species
$n-k$
k
outgroup
donor
recipient

Fig 1. Model of an introgression sweep after a secondary contact.
Species trees: phylogenetic relationships between the recipient, donor and outgroup species. Note that the time scale is not respected (T_d and T_{sp} are very large) and that all species are assumed to have the same size.
Coloured background: frequency of the selected allele in the different species. A single favourable haplotype is introduced into the recipient species through a rare hybridization event with the donor (red arrow) where it eventually reaches fixation.
Superimposed coalescent tree: coalescent tree of a sample of n lineages, taken from the recipient population at a neutral site located at a distance d from the focus of selection. k lineages escape the selective sweep (see eq. (2)) and their polymorphism is a subsample of the neutral site frequency spectrum (see eq. (6)). The other $n-k$ lineages trace back as a single lineage into the donor species.

Volcanoes of Diversity

The differences between introgression sweeps and classical sweeps can be seen in their respective footprint on the expected heterozygosity (pairwise nucleotide diversity) H at neighboring loci. As shown in Fig. 2A, introgression from a diverged donor population changes the typical valley shape of a classical sweep to a volcano shape, where diversity exceeds the genomic background in the flanking regions. We can understand this difference as follows.

Starting with a sample of size $n = 2$ taken from the recipient population directly after fixation of the B allele, there are four potential coalescent histories during the sweep phase. If both lineages do not coalesce during the sweep, then one or both must have escaped the sweep by recombination. We denote the probability of these events by P_{Bb} and P_{bb}. Alternatively, if the lineages coalesce, then their ancestral lineage can be associated with the B or the b allele, with respective probability P_B and P_b. Because the star-like approximation assumes that coalescence only occurs among B lineages at the start of the sweep, we have $P_b = 0$. The other probabilities are summarized in Table 1. The expected heterozygosity follows as $H = 2\theta\,\mathbb{E}[T_{coal,2}]$, with $\theta = 4N\mu$ and $\mathbb{E}[T_{coal,2}]$ the expected pairwise coalescence time, averaged over the four scenarios. Neglecting the time during the sweep, the coalescence times are entirely due to the demographic phase. For a classical sweep, this is just the neutral coalescence time in the study population (*i.e.*, $\mathbb{E}[T_{coal,2}] = 1/2$ in units of $4N$ generations, assuming standard neutrality). In the case of an introgression sweep, however, this time is increased if a single line has escaped the sweep (probability P_{Bb}). In this case, coalescence is only possible in the common ancestor of the donor and recipient species and $\mathbb{E}[T_{coal,2}] = T_d + 1/2$. The expected coalescence times for all cases are shown in Table 1.

Table 1. Summary of the effects of selection

Genealogical	Probability	Coalescence times	
history	star-like approx.	classic sweep	introgression sweep
$\{B, B\} \to \{B\}$	$P_B = e^{-2\alpha d}$	0	0
$\{B, B\} \to \{B, b\}$	$P_{Bb} = 2\left(1 - e^{-\alpha d}\right) e^{-\alpha d}$	1/2	$T_d + 1/2$
$\{B, B\} \to \{b, b\}$	$P_{bb} = \left(1 - e^{-\alpha d}\right)^2$	1/2	1/2
$\{B, B\} \to \{b\}$	$P_b = 0$	0	0

Possible (backward) coalescence or recombination events during the selective sweep with $n = 2$ lines linked to the beneficial allele at the time of sampling $\{B, B\}$, probabilities under the star-like approximation, and expected time to coalesence for both a classic sweep and an introgression sweep. Note that all times are measured in units of $4N$ generations.

Under the star-like approximation, we then obtain:

$$\begin{aligned} H_{\text{classic}} &= (1 - P_B)\theta = (P_{bb} + P_{Bb})\theta \\ H_{\text{intro}} &= (1 - P_B)\theta + 2T_d P_{Bb}\theta = P_{bb}\theta + P_{Bb}D, \end{aligned} \tag{3}$$

using $P_B + P_{Bb} + P_{bb} = 1$ and $D = (2T_d + 1)\theta$. For both introgression and classic sweeps, coalescence during the sweep (P_B) reduces genetic diversity, while *partial* escape through recombination (P_{Bb}) increases diversity only in the introgression case. Substituting the probabilities from Table 1, we obtain the expected heterozygosities as functions of $\alpha d = (R/s)\ln(2N)$ and D. In Fig. 2 (dashed lines) they are shown together with simulation data as function of the recombination distance R and D and s as parameters. Fig. 2A shows the effect of the divergence D of the donor population and Fig. 2B shows the effect of the strength of selection s acting on the the beneficial allele.

While divergence mostly affects the height of the volcano for introgression sweeps, the selection strength mostly scales the width of the footprint.

We can analyze the shape of the footprint in more detail using the star-like approximation. In this case, the width of the signal can be measured in terms of a single compound parameter αd. For a classical hard selective sweep, we find that the variation at a scaled distance $\alpha d = \frac{1}{2}\ln(1/X)$ from the selected site is reduced by a fraction of X (i.e., $H_{\text{classic}} = (1-X)\theta$). Due to the excess variation that is brought in from the diverged donor population, the central valley of an introgression sweep is narrower, with decreasing width as divergence D increases. At a distance

$$\alpha d = \ln\left(\frac{2D-\theta}{2D-2\theta}\right) \xrightarrow[D\to\infty]{} 0 \tag{4}$$

both effects compensate and we obtain an expected heterozygosity of $H_{\text{intro}} = \theta$. At larger distances, H_{intro} overshoots the background level and assumes a maximum value of

$$H^*_{\text{intro}} = \frac{D^2}{2D-\theta},$$

which is independent of the selection coefficient in the star-like approximation. Using $D = (2T_d+1)\theta$, we can express the relative height of the "volcano" above the background level as a function of the divergence time

$$\frac{H^*_{\text{intro}}-\theta}{\theta} = \frac{4T_d^2}{4T_d+1}.$$

This maximum is reached at a scaled distance $\alpha d = \ln\left(\frac{2D-\theta}{D-\theta}\right) \xrightarrow[D\to\infty]{} \ln(2) \approx 0.7$. The signal of the introgression sweep is therefore strongest at the distance where a classical sweep signal has already decayed by at least 75%. At a scaled distance of

$$\alpha d = \ln\left(\frac{2D-\theta}{D-\theta}\right) + \ln\left(\frac{10}{10-3\sqrt{10}}\right) \xrightarrow[D\to\infty]{} \ln\left(\frac{20}{10-3\sqrt{10}}\right) \approx 3.7 \tag{5}$$

the increased heterozygosity returns to 10% of the maximum value, and $H_{\text{intro}} = \theta + \frac{H^*_{\text{intro}}-\theta}{10}$. The footprint of an introgression sweep is therefore considerably wider than that of a classic sweep, in which, a 90% recovery of the decreased diversity is expected at distance $\alpha d = \frac{1}{2}\ln(10) \approx 1.2$.

Beyond the star-like approximation

While the star-like approximation (dashed lines in Fig. 2) provides qualitatively accurate results, it overestimates P_{Bb}, and consequently, the height of the volcano peaks. Simulations show that this height may also be slightly dependent on the selection coefficient (compare dashed lines and dots for simulated values in Fig. 2B). In the supplementary information, we provide a more accurate approximation for the probabilities in Table 1 using a stochastic approach based on Yule branching processes [51]. In particular, this approach allows for coalescence during the sweep as well as recombination of coalesced lineages to the b background. We thus obtain $P_b > 0$ and reduced values for P_{Bb} relative to the star-like approximation. As shown in Fig. 2 (solid lines, see also Text S1.1), this leads to an improved fit of the simulation data for pairwise diversity. However, an extension of this method to the site-frequency spectrum for larger samples is difficult. We therefore resort to the star-like approximation in what follows and in our parametric test.

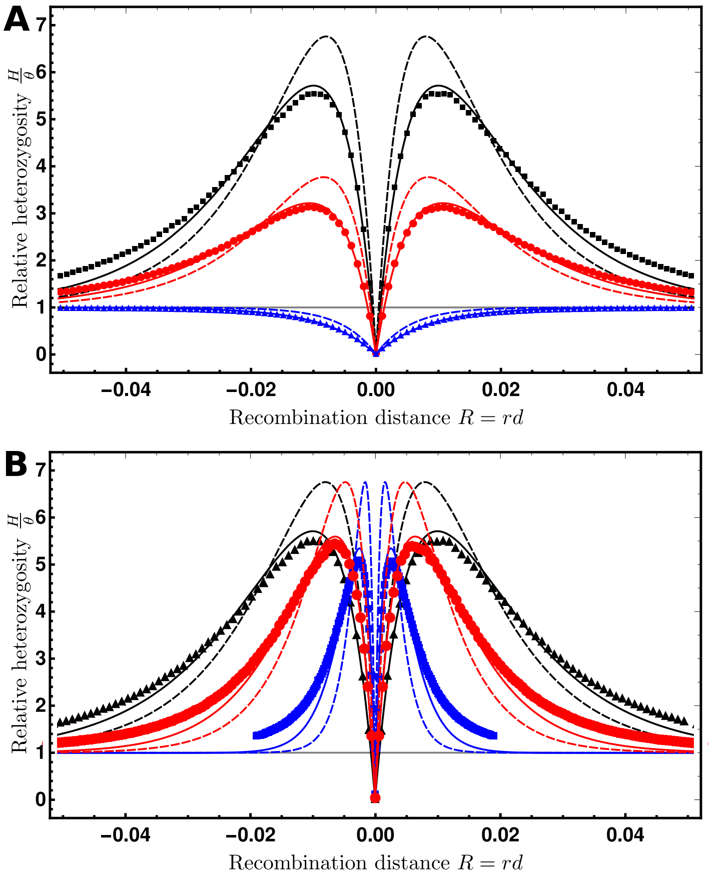

Fig 2. Volcanoes of Diversity.
Expected genetic diversity after the sweep relative to the initial heterozygosity, $\frac{H}{\theta}$ for a beneficial mutation centered at 0 as a function of the recombination distance $R = rd$ to a neutral locus on left ($-$) and right ($+$) sides. For both panels, the lines show the predictions under the star-like approximation (dashed) and the better approximation (solid, see Text S1.1). The dots show the average ± 3 standard errors about the mean (the error bars are smaller than the plot points). **A.** The effect of divergence of the donor population for an introgression sweep with $2Ns = 1\,000$. The divergence time (in units of $4N$ generations) is $T_d = 6$ (*i.e.*, $D = 13\theta$, black), 3 ($D = 7\theta$, red), and 0 ($D = \theta$, blue), where $T_d = 0$ is a classic sweep from a *de novo* mutation. **B.** The effect of the strength of selection for an introgression sweep with $T_d = 6$ ($D = 13\theta$). The strength of selection is $2Ns = 1\,000$ (black), 600 (red), or 200 (blue). For both panels, $\theta = 0.002$ ($N = 5\,000$, $\mu = 10^{-7}$), $r = 10^{-7}$, and the window size is 100 nt.

Single iterations

Footprints of introgression sweeps, like classical sweeps [46], are highly variable due to the stochastic events in the genealogical history of the sampled lineages. Single numerical replicates, as well as patterns in data, can deviate strongly from the "expected" volcano shape displayed in Fig. 2. In Fig. 3, we show a typical set of introgression footprints obtained from single replicate runs. We see that, under favorable conditions (large T_d and sampling directly after the fixation of the beneficial allele), volcano shapes are clearly discernible even in single iterations. However, we also see that the width and symmetry of the volcanoes varies greatly between replicates. The key reason for this variation is the early recombination events during the initial stochastic establishment phase of the beneficial allele. In the sample genealogy, the B allele can dissociate from the foreign haplotype if even a single recombination event occurs in the time between coalescence of all B lineages and the initial introgression of the B allele. As the volcano pattern is relatively broad, these recombination events occur with substantial probability. At distances beyond the recombination break point, only genetic variation from the recipient population hitchhikes, resulting in the classic sweep pattern from *de novo* mutation. Since independent recombination events are required to "cut the volcano" on both sides of the beneficial allele, strong asymmetries in the shape arise naturally.

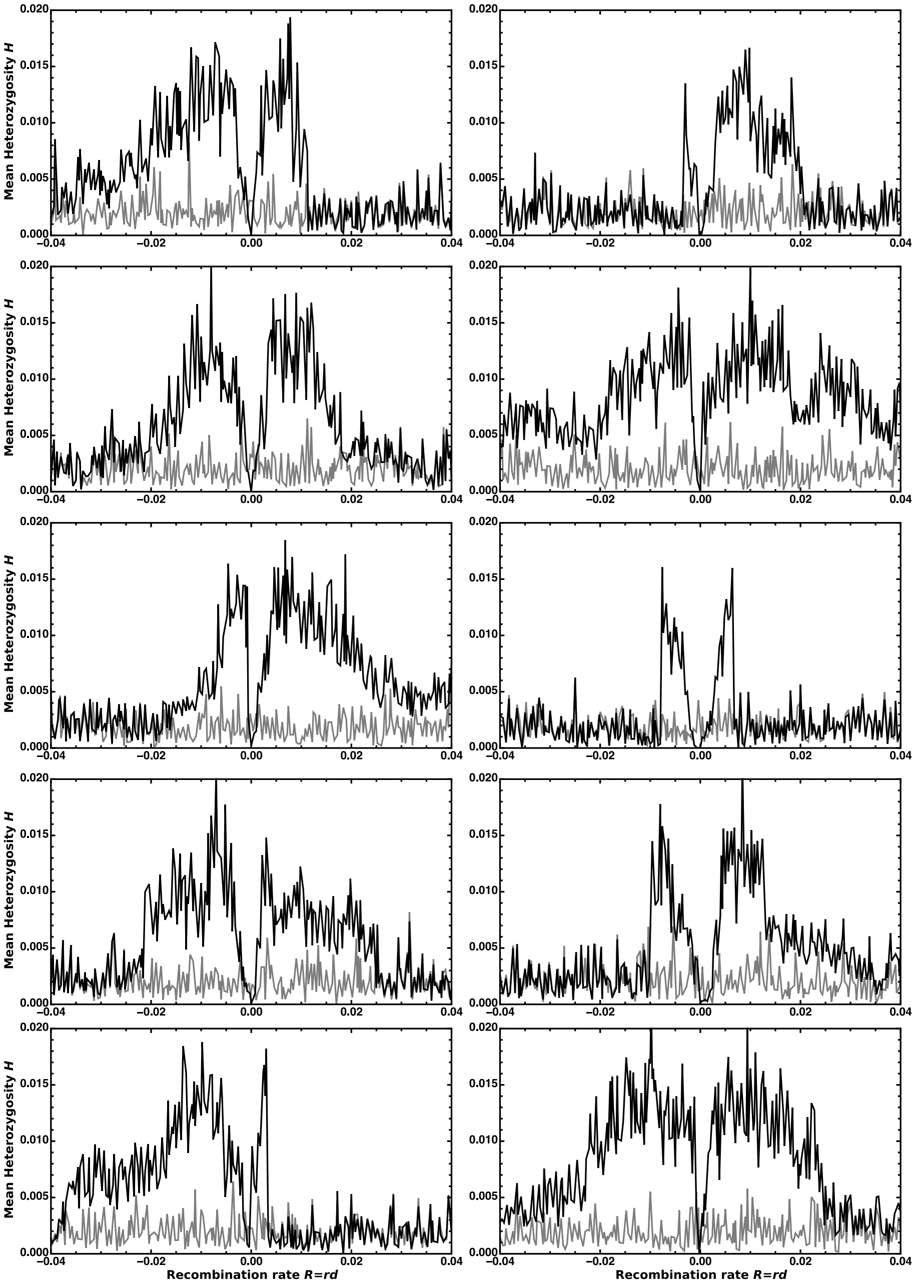

Fig 3. Single iterations of an adaptive introgression event.
Each panel shows an independent, randomly chosen simulation run. We calculated the whole-population mean genetic diversity in 401 non-overlapping non-adjacent one kb windows separated by one kb and centred on the selected locus. The initial heterozygosity and the genetic diversity at fixation of the beneficial B allele are shown in grey and black, respectively. Here, $\theta = 0.002$ ($N = 5\,000$, $\mu = 10^{-7}$), $r = 10^{-7}$, $T_d = 6$ ($D = 13\theta$), and $s = 0.06$ ($2Ns = 600$).

The footprint of adaptive introgression in the SFS

Following [47] we use a parametric approach to model the effect of a recent introgression sweep on the site frequency spectrum (SFS) at distance d from a recently-fixed beneficial allele. Our model includes the compound parameter α (sweep strength) of the classic hard sweep model in [47], as well as the additional parameter D (donor divergence) specific to an introgression sweep.

The background reference SFS

Consider an alignment of n sequences from the recipient species and one sequence from a distant outgroup. Assuming complete lineage sorting and the infinite sites model, all mutations in the alignment can be polarized with the help of a second more distant outgroup. We denote $S_i(n)$ the non-normalized SFS, *i.e.*, the probability that a mutation occurred at some nucleotide site and that exactly i $(0 \leqslant i \leqslant n)$ among the n sampled lineages in the recipient species harbor the inferred derived allele (see Fig. 1). The probability to observe a fixed difference between the recipient and the outgroup species is $S_0(n) + S_n(n)$. If a second outgroup is unavailable, then only polymorphic mutations in the recipient species can be polarized, but not the fixed differences. In this case, we arbitrarily label the state in the first outgroup as "ancestral" such that $S_0(n) = 0$. Following [47], the neutral reference SFS can be estimated from the observed genome-wide data. Given these estimates for the $S_i(n)$, the spectral probabilities $S_j(k)$ in subsamples of $k \leqslant n$ sequences follow as

$$S_j(k) = \sum_{i=j}^{n} S_i(n) \frac{\binom{i}{j}\binom{n-i}{k-j}}{\binom{n}{k}}. \qquad (6)$$

The conditional probability of observing i mutant alleles among n lineages given that the site is polymorphic is

$$p_{i,n} = \frac{S_i(n)}{\sum_{j=1}^{n-1} S_j(n)}. \qquad (7)$$

Similarly, the conditional probability of observing i mutant alleles given that the site is polymorphic or a fixed difference for which the recipient species has the inferred derived state is

$$q_{i,n} = \frac{S_i(n)}{\sum_{j=1}^{n} S_j(n)}. \qquad (8)$$

If the mutation rate μ varies along the genome, then the probabilities $S_i(n)$ will vary among sites, because $S_i(n)$ is proportional to $\theta = 4N\mu$. In contrast, the mutation rate cancels in the conditional probabilities $p_{i,n}$ and $q_{i,n}$, which are expected to be constant along the genome.

The expected SFS after the sweep

Following the star-like approximation, k lineages escape the introgression sweep with probability $P_e(k|\alpha, D)$ (eq. 2). We assume complete lineage sorting between these lineages and the single ancestral lineage of all lines that are caught in the sweep and transition to the donor species. An introgression sweep then transforms the SFS as follows. Let $S_i'(n|\alpha, d, D)$ denote the per-nucleotide probability of observing i mutant lineages in a sample of n lineages from the recipient species after an introgression sweep with strength parameter α and divergence parameter D at distance d. Below, we assume that the time for the coalescent process in the recipient species is negligible relative to the divergence time between the recipient and the donor species (see Fig. 1).

As shown in the supplement (see Text S1.2), this assumption can be relaxed. However, because the more complex model did not lead to a clear improvement of our statistical test, we focus on the simple approximation in the main text. In this case, the transformed SFS after the introgression sweep is given by ($1 \leqslant i \leqslant n-1$):

$$S_i'(n|\alpha, d, D) = \left(\sum_{k=i+1}^{n} P_e(k|\alpha, d) S_i(k) \right) + P_e(n-i|\alpha, d)\tfrac{D}{2} + P_e(i|\alpha, d)\tfrac{D}{2}. \tag{9}$$

The first term on the right hand side accounts for the contribution of mutations that occur during the coalescent process of the escaping lineages in the recipient species. The second and third terms, respectively, account for mutations on the long ancestral lineages in the donor and recipient population, which partition the $n-i$ lineages that are caught in the sweep from the i escape lineages. Because the expected coalescence time for these lineages is $T_d + 1/2$, the probability for a mutation to hit either lineage is $\theta(T_d + 1/2) = D/2$. The conditional probabilities given that the site is polymorphic follow as

$$p_{i,n}'(\alpha, d, D) = \frac{S_i'(n|\alpha, d, D)}{\sum_{j=1}^{n-1} S_j'(n|\alpha, d, D)}. \tag{10}$$

Because all terms in eq. (9) are proportional to θ, this normalization removes the dependence on the mutation rate, analogous to eq. (7).

If fixed differences are polarized, then a site will be a fixed difference for which all recipient lineages carry the mutant allele if a mutation occurred in the lineage that connects the MRCA of the sample to the MRCA of the recipient and the outgroup species, leading to the following probability:

$$S_n'(n|\alpha, d, D) = \left(\left(D_o - \tfrac{D}{2}\right) \sum_{k=1}^{n-1} P_e(k|\alpha, d) \right) + D_o P_e(0|\alpha, d) + D_o P_e(n|\alpha, d), \tag{11}$$

where D_o is the expected divergence between the recipient species and the MRCA of the recipient and the outgroup species. The first term in the right-hand side of eq. (11) accounts for the cases when some (but not all) lineages escape the sweep, whereas the second and third terms account for the cases when no lineages or all lineages escape the sweep. In our secondary contact model, D_o can be estimated from the data as

$$D_o = S_n(n) + \frac{1}{n} \sum_{i=1}^{n-1} i S_i(n), \tag{12}$$

which is equivalent to $D_o = S_1(1)$, as can be seen from eq. (6). The second term on the right hand side of eq. (12) is the mean number of mutations accumulated in each recipient lineage since their MRCA, related to the unbiased estimator of θ, $\hat{\theta}_L = \frac{1}{n-1} \sum_{i=1}^{n-1} i S_i(n)$ [52, eq. (6) and (8)]. If fixed differrences are not polarized, then eqs. (11) and (12) still hold when substituting D_o with the full divergence between the recipient species and the outgroup D_o'. Assuming constant mutation rates between the focal species and the outgroup, all three terms in eq. (11) are proportional to θ, making the conditional probabilities once again independent of he mutation rate,

$$q_{i,n}'(\alpha, d, D) = \frac{S_i'(n|\alpha, d, D)}{\sum_{j=1}^{n} S_j'(n|\alpha, d, D)}. \tag{13}$$

A composite likelihood ratio test

Our test builds on the composite-likelihood method first introduced in [46] and further developed in [47–49]. Sequence data are collected in an alignment of n chromosomes from the recipient species and possibly one chromosome from an outgroup species. We assume that mutations are polarized and consider only informative sites, *i.e.*, sites for which at least one chromosome in the recipient species harbors the inferred derived allele. Let L be the number of informative sites and X_ℓ the frequency of the derived allele at the ℓth informative site. We contrast the composite likelihoods of a reference and an alternative model for the empirical SFS. The reference model assumes that the distribution of the classes in the SFS is homogeneous along the chromosome. Accounting for fixed differences, the genome-wide SFS conditional probabilities are given by eq. (8), and the composite likelihood of the reference model is

$$CL_0 = \prod_{\ell=1}^{L} q_{X_\ell,n}. \tag{14}$$

The alternative model assumes that an introgression sweep event with unknown parameters α and D recently happened at some location on the chromosome, leading to an inhomogeneous altered SFS along the chromosome. Let d_ℓ be the distance of the locus of the introgression sweep to the ℓth informative site. The composite likelihood CL_1 of the alternative model including fixed differences uses the local SFS conditional probabilities from eq. (13),

$$CL_1(\alpha, D) = \prod_{\ell=1}^{L} q'_{X_\ell,n}(\alpha, d_\ell, D). \tag{15}$$

If fixed differences with a single outgroup are unavailable (for instance if different outgroup species were used to polarize polymorphic sites), then the test can also be set up without fixed differences, by using probabilities $p_{X_\ell,n}$ from eq. (7) in eq. (14) and $p'_{X_\ell,n}(\alpha, d_\ell, D)$ from eq. (10) in eq. (15).

For a given genomic position of the beneficial allele, maximum composite likelihood estimates $\hat{\alpha}$ and $\hat{D}$ are obtained such that $CL_1(\hat{\alpha}, \hat{D}) = \max_{\alpha,D}(CL_1(\alpha, D))$ with $\alpha > 0$ and $0 \leqslant D \leqslant 2D_o$ if fixed differences are polarized and $0 \leqslant D \leqslant D'_o$ otherwise. The test statistic for the composite likelihood ratio test is defined as

$$T_1 = 2\left(\ln CL_1(\hat{\alpha}, \hat{D}) - \ln CL_0\right). \tag{16}$$

The SFS after an introgression sweep

Fig. 4 displays the effect of adaptive introgression on the SFS of the recipient population in the simple model described above (Model 1, red columns) and the more complex model described in the supplement (Model 2, blue columns, see Text S1.2) relative to the neutral spectrum (black). Model 2 differs from Model 1 in that it does not ignore the coalescence time in the recipient species. The figure shows that near the sweep center (distance $\alpha d = 0.01$, top panel), hitchhiking reduces polymorphism and increases the proportion of fixed differences. For sites located at distances where single recombination events are likely ($\alpha d = 0.1$ and 1.0), partial hitchhiking of foreign variation increases polymorphism relative to the neutral expectation. This increase in diversity is accompanied by a decrease in the proportion of fixed differences relative to the third, outgroup species. Under the infinite sites mutation model, sites that diverge from the donor population must also diverge from the outgroup species. At these sites, introgression re-introduces the ancestral variant, sharply reducing the proportion of fixed derived sites in the sampled lineages. This is a key feature not seen in classic hard sweeps.

Fig. 4 also shows that the predicted SFS under the simple Model 1 (red) does not differ much from the SFS predicted under the slightly more accurate Model 2 (blue).

However, there is still a key difference. Model 2 is restricted to $D \geqslant 2\frac{n-1}{n}\hat{\theta}_L$, whereas Model 1 may take smaller values, including $D = \theta$ for a classic sweep from *de novo* mutation. For these reasons, we suggest to, in general, use Model 1. Unless otherwise noted, the VolcanoFinder results presented in this article are generated under Model 1 and fixed differences with the outgroup are polarized with the help of a second, distantly-related outgroup.

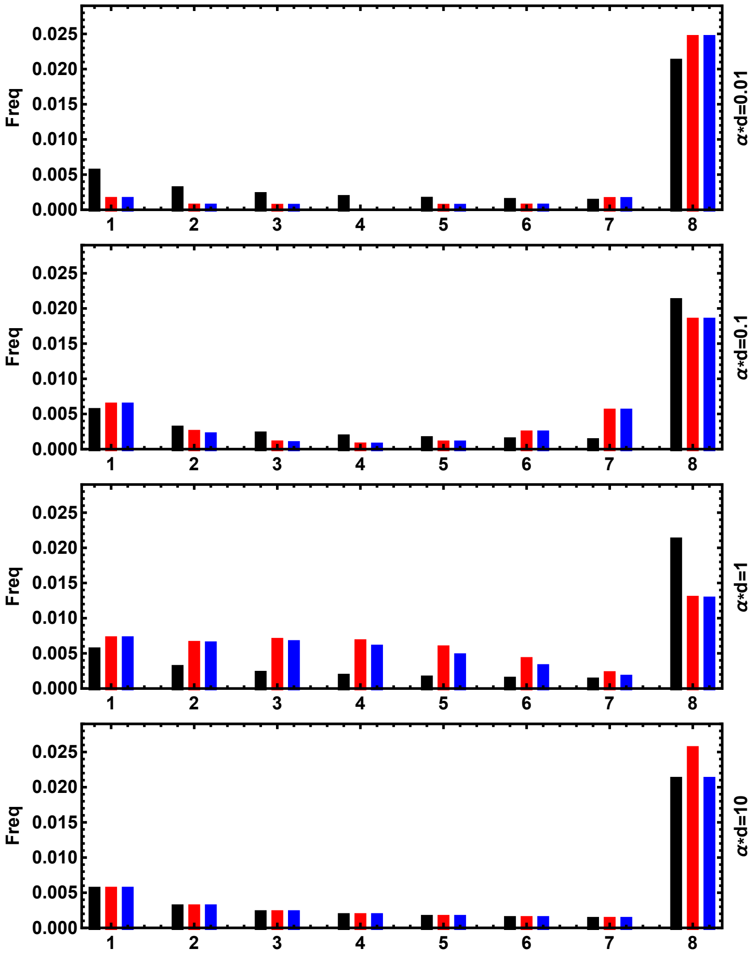

Fig 4. The site frequency spectrum after adaptive introgression. The SFS as a function of the relative distance αd from the sweep center. Model 1 (eq. 11) predictions are in red; Model 2 (eq. 22), blue. Here, $N = 5\,000$, $s = 0.1$, $\theta = 0.005$, $D = 0.026$ and $D_o = 0.05$

Although the star-like model is only a rough approximation, it allows for considerable flexibility to fit empirical patterns via optimization of the parameters α and D. While α modulates the width of the footprint, D mostly scales the height of the volcano. As shown in Fig. 3, the width of the pattern varies strongly between replicates. The model can partially compensate for this variation by adjusting α. Still, the average estimate of α (across replicates) closely matches the true value in simulated data (see Text S1.3). In contrast, the divergence is systematically underestimated. The downward bias of D compensates for the overestimation of the volcano height under the star-like approximation (*cf.* Fig. 2). This bias is not a problem as long as the method is only used to infer adaptive introgression, and no biological interpretation is attached to the fit parameters. Bias in D needs to be accounted for, however, if the method is used for biological parameter estimation.

Power analysis

We performed coalescent simulations involving an introgression sweep occurring with a 95% probability. This is achieved by adjusting the admixture level during the secondary contact episode (see Fig. S2.1 and Text S2.1). We studied the relative powers of `VolcanoFinder`, `SweepFinder2` [49], and `BALLET` [48] to detect non-neutrality in the simulated alignments. The effect of five parameters on the statistical power were assessed: the selection coefficient s of the beneficial allele, the age T_d of the speciation event that isolated the donor and the recipient species, the time elapsed since the end of the introgression sweep T_s, the presence of polymorphic genetic variation co-introgressing with the beneficial allele (hard or soft introgression sweeps), and the admixture level of the reference genomic background.

Hard and soft introgression sweeps

Hard and soft selective sweeps refer to sweeps that originate from a single or multiple copies of the beneficial allele, respectively. In the case of introgression, hard sweeps trace back to a single migrant from the donor population, while soft sweeps originate from multiple migrants. More generally, hard introgression sweeps represent all scenarios where the beneficial haplotype traces back to a very recent common ancestor in the donor population, such that no standing genetic variation from the donor population can co-introgress with the beneficial allele. Conversely, soft introgression sweeps allow for diversity among the introgression haplotypes. In our simulations, we maximize this diversity by assuming that the beneficial allele has fixed in the donor population a long time ago. As a consequence, all introgression haplotypes are related by a standard neutral coalescent in a donor population of size $N' = N$. While classical hard and soft sweeps in a single population can lead to strongly diverging footprints [36], hard and soft introgression sweeps both lead to very similar volcano patterns in the heterozygosity (compare Fig. S2.3 and Fig. S2.4). The central valley is slightly deeper for hard introgression sweeps, and the peaks are slightly higher for soft introgression sweeps.

Admixture in the genomic background

We present two series of power analyses, each with different assumptions about admixture in the genomic background. In the first scenario, we assume that the reference genomic background is free of admixture from the donor population. We thus test for signals of *local introgression* at the target locus against the alternative assumption of *no introgression*. In the second scenario, we assume that secondary contact leads to genome-wide admixture. I.e. we test for the power to detect *adaptive introgression* against a background of *neutral introgression*, with a uniform genome-wide admixture proportion. Both scenarios represent limiting cases of adaptive introgression events that may be observed in nature. If introgression is a very rare event and/or introgressed variation is usually deleterious and purged from the recipient population by selection, a non-admixed background is the appropriate reference. Conversely, genome-wide admixture can be expected with higher admixture rates and if genetic barriers to gene flow are weak.

Probability of detection of an introgression sweep in an outlier study

Analyses of test power typically display the true positive rate against the false positive rate in a so-called ROC curve. In the current study, ROC curves of this type are provided in the supporting information (Fig. S2.5 and Fig. S2.6). However, when a test is applied to actual data, the problem is slightly different. An introgression sweep event is identified in a genome-wide scan if the CLR values in the region involved in the introgression sweep rank among the highest genome-wide candidate peaks. We therefore define a detection probability given a number X of candidate peaks considered as the probability that the focal locus ranks among the top X CLR value peaks. The large number of neutral replicates we used for the power analysis is comparable to a full genome scan (for each parameter set, we produced 10^4 neutral replicates of 200 kb sequences leading to a 2 Gb alignment and 8×10^6 CLR values) and enabled us to estimate these detection probabilities (Fig. 5 and Fig. 6). Both ways to display test power are related: a high rejection rate for a very low false positive rate guarantees a high-ranking peak in a genome scan. However there are important differences: Rejection rates of ROC curves usually consider the maximum CLR statistics in windows around a focal site and are thus dependent on the (to some extent arbitrary) width of these windows. In contrast, the outlier-peak approach (like a scan of real data) uses the width of observed peaks to account for local linkage and therefore does not depend on a predefined window width.

Statistical power: non-admixed genomic background

In the limiting case of a non-admixed genomic background, `VolcanoFinder` clearly outperforms the other methods (Fig. 5 and Fig. S2.5). It detects both hard and soft introgression sweeps with strong or moderate selection strength with a probability close to 1 even in very small sets of outliers as long as the divergence from the donor species is large enough ($T_d \geqslant 2.5$, *i.e.*, $D \geqslant 6\theta$). In contrast to classical sweeps, even older introgression events are detected with high power (up to $T_s = 0.5$). The relative performances of `BALLET` and `SweepFinder2` depend on the age of the sweep. For highly diverged species ($T_d \geqslant 4$, $D \geqslant 9\theta$), `SweepFinder2` looses power faster than `BALLET` as the time since the selective sweep increases, because it is sensitive to the valley of expected heterozygosity induced by the selective sweep. This also explains the large reduction in power of `SweepFinder2` for soft sweeps. The better performance of `VolcanoFinder` in detecting introgression sweeps in smaller sets of outliers relies on its higher rejection rates for low false positive rate, see Fig. S2.5 (the lowest false positive rate on our ROC curves is 0.1%).

For some parameter sets, the power (or detection probability) of the tests exceeds the 95% probability that adaptive introgression occurs in the simulations. This is because the tests really detect local introgression in this setting, as described above. Even in the 5% of simulations where the adaptive allele is eventually lost, there may still be a significant excess of introgressed variation at the focal locus relative to the background. If these variants segregate at intermediate frequencies, then the signal is picked up by scans for adaptive introgression or long-term balancing selection. Note that, when rejection rates exceed 95%, higher rejection rates are observed for weak selection ($2Ns = 100$) than for strong selection ($2Ns = 1\,000$), consistent with the 10 fold higher admixture level needed in the weak selection case to achieve a 95% probability for an introgression sweep to occur.

Statistical power with an admixed reference genomic background

Fig. 6 and Fig. S2.6 show the power of all three tests assuming a constant genome-wide admixture proportion. Because this proportion is adjusted such that an introgression sweep occurs in 95% of all simulation, the maximal power (detection probability) that can be achieved by a "perfect" test in this case is 0.95 (as observed in the figures). It also means that the admixture proportion is larger for weak selection (3% for $2Ns = 100$) than for strong selection (0.3% for $2Ns = 1000$).

High levels of admixture in the genomic background leads to a strong reduction in power for all three methods (compare Fig. 5 and Fig. 6). Due to the higher admixture rate, this holds, in particular, for simulations with weak selection ($2Ns = 100$) whereas the reduction is moderate for $2Ns = 1\,000$. All methods need a relatively high false discovery rate to achieve rejection rates close to the expected maximum (Fig. S2.6), thus reducing the probability of an introgression sweep to be detected in small sets of outlying peaks. `VolcanoFinder` still performs better than other methods (Fig. 6), especially for recent introgression sweeps ($T_s = 0$) from donor species that are not too closely related ($T_d \geqslant 2.5$, *i.e.*, $D \geqslant 6\theta$). For instance, a recent introgression sweep ($T_s = 0$) from a moderately diverged donor species ($T_d = 2.5$, $D = 6\theta$) with a strongly selected allele ($2Ns = 1\,000$) will be associated with the genome-wide highest CLR with probability around 1/2 for `VolcanoFinder` 1/3 for `SweepFinder2` and close to 0 for `BALLET`. Notably, `VolcanoFinder` maintains some statistical power for much older selective events ($T_s \geqslant 0.25$) when the detection probability of other tests is close to 0.

As mentioned above, the rejection rates in ROC curves depend on the window size that is used to derive the maximum CLR statistics in the neutral reference. Narrower windows lead to smaller samples of CLR values for the null model, and thus to increased rejection rates. As the region showing the introgression sweep signal is ten times wider for strong selection ($2Ns = 1\,000$) than for weak selection ($2Ns = 100$), narrower windows can, in principle, be used for weaker selection. We therefore also computed the rejection rates based on the maximum CLR in regions of different width around the selected site (200 kb for $2Ns = 1\,000$ and 20 kb for $2Ns = 100$). As expected, the rejection rates for $2Ns = 100$ increase (Fig. S2.7): the gain of statistical power is especially noticeable for old introgression sweeps ($T_s \geqslant 0.1$) for which the rejection rates now clearly exceed the false positive rate. However, it does not reach the high values for the case $2Ns = 1\,000$ and a smaller admixture proportion. The effects of a smaller window size are similar for all three methods studied.

This approach with different window widths was also used when contrasting significant and non-significant tests in the distribution of the estimated selection parameters (position of the selected locus, selection strength, and divergence from the donor species) as inferred by `VolcanoFinder`. These results are described in Text S2.4 and Fig. S2.8 to Fig. S2.17.

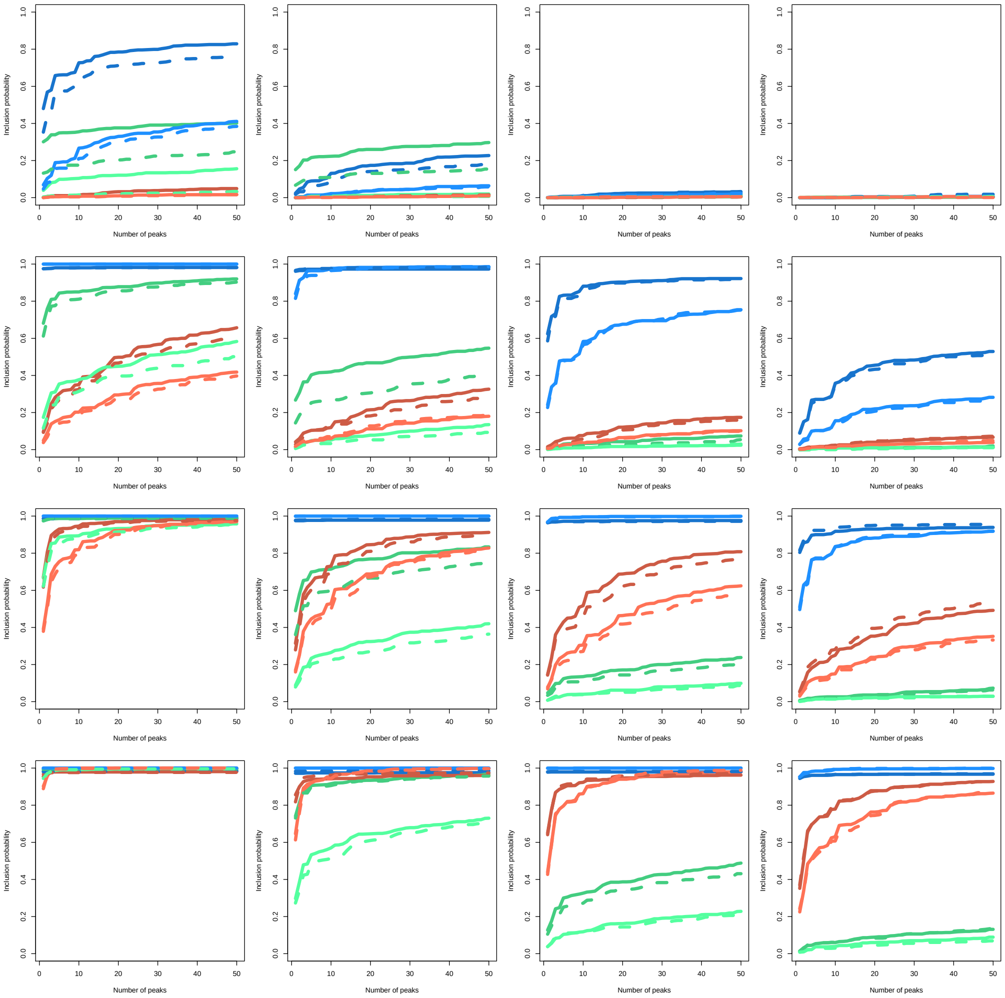

Fig 5. Detection probability of an introgression sweep (non admixed background)
Probability of an introgression sweep event to be detected in a genome-scan analysis using `VolcanoFinder` (blue), `BALLET` (brown) and `SweepFinder2` (green). The donor species diverged from the recipient species at (top to bottom) $T_d = 1, 2.5, 4, 5.5$ (*i.e.* $D = 3\theta, 6\theta, 9\theta, 12\theta$) and the selective sweep ended (from left to right) $T_s = 0, 0.1, 0.25, 0.5$ units of $4N$ generations in the past. Solid lines: no polymorphism in the donor species (hard introgression sweep). Dashed lines: polymorphism exists in the donor species (possible soft introgression sweep). Dark colour: $2Ns = 1\,000$; light colour: $2Ns = 100$. Analyses involved a non-admixed neutral genomic background as a reference.

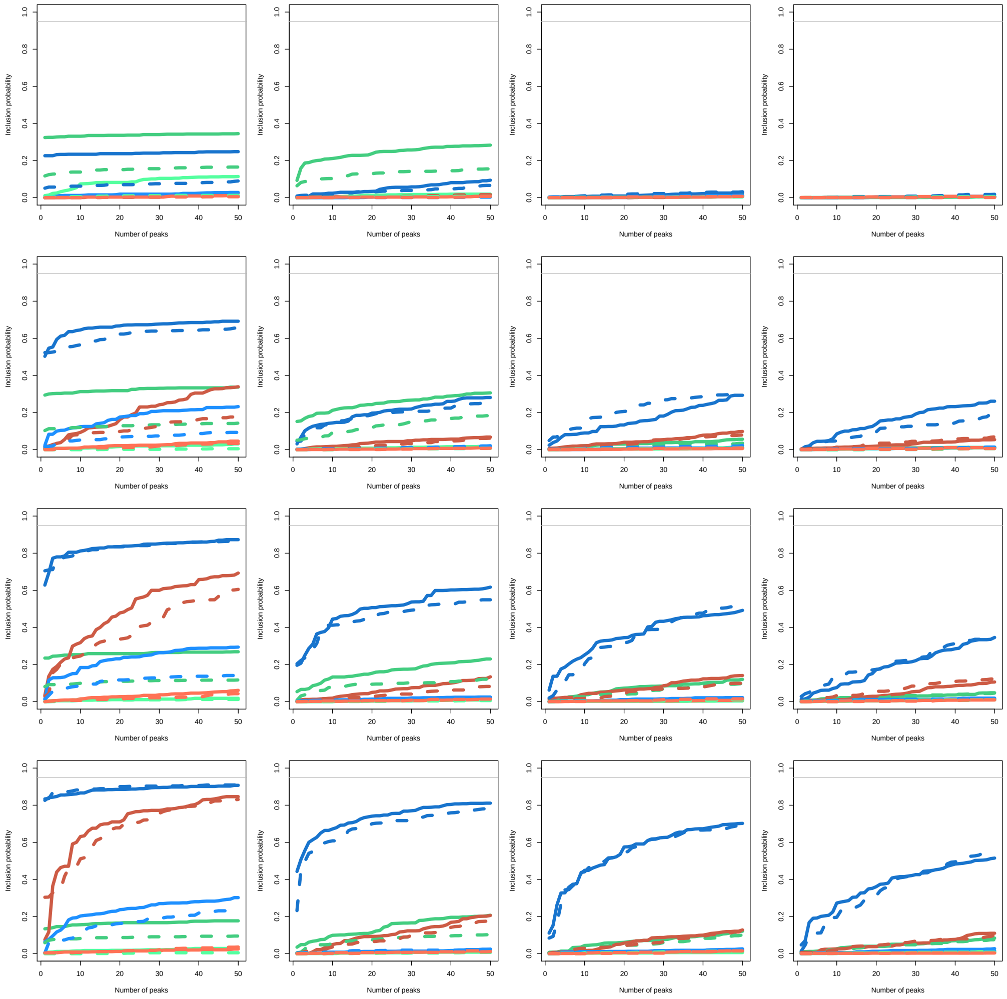

Fig 6. Detection probability of an introgression sweep (admixed background)
Probability of an introgression sweep event to be detected in a genome-scan analysis using `VolcanoFinder` (blue), `BALLET` (brown) and `SweepFinder2` (green). The donor species diverged from the recipient species at (top to bottom) $T_d = 1, 2.5, 4, 5.5$ (*i.e.* $D = 3\theta, 6\theta, 9\theta, 12\theta$) and the selective sweep ended (from left to right) $T_s = 0, 0.1, 0.25, 0.5$ units of $4N$ generations in the past. Solid lines: no polymorphism in the donor species (hard introgression sweep). Dashed lines: polymorphism exists in the donor species (possible soft introgression sweep). Dark colour: $2Ns = 1\,000$; light colour: $2Ns = 100$. Analyses involved a neutral admixed genomic background with the same level of admixture as a reference.

Robustness of VolcanoFinder to long-term balancing selection

Balancing selection increases the polymorphism-to-divergence ratio in regions surrounding the selected site [53]. Because this signal also occurs in the case of an introgression sweep, VolcanoFinder could falsely detect an introgression sweep in the case of long term balancing selection. To assess the robustness of VolcanoFinder, we compared the rejection rates of VolcanoFinder and BALLET under three demographic models inspired by [48] for increasingly old balancing selection (overdominance). The results are shown in Fig. S2.18. Unlike BALLET, the rejection rate of VolcanoFinder is close to the false positive rate (although a bit larger) for moderately old balanced polymorphisms ($T_s \leqslant 8.75$) and remains low (10% to 20% depending on the demographic model) even for very old balanced polymorphisms ($T_s = 20$). Interestingly, the effect of the demographic model on the power to detect the footprints of balancing selection acts in opposite directions for VolcanoFinder and BALLET, suggesting that these two methods are sensitive to opposite patterns in the SFS. Overall, VolcanoFinder was found to be relatively robust to long-term balancing selection.

Scans of human data

Despite the lack of contact with known archaic hominins such as Neanderthals or Denisovans, recent evidence suggests that the genomes of modern African human populations carry potentially-introgressed regions from unknown sources (*e.g.*, [54, 55]). In contrast, the genomes of non-Africans have been shown to harbor considerable levels of admixture with known archaic humans, such as Neanderthals [56, 57]. We therefore examined signals of adaptive introgression in African and non-African human populations by applying `VolcanoFinder` to the Yoruban (YRI) sub-Saharan African and a central European (CEU) human populations.

In particular, we employed bi-allelic single nucleotide variant calls from the human 1000 Genomes Project [58] and polarized alleles based on alignment with the chimpanzee reference sequence [59]. To circumvent potential technical artifacts, we filtered out regions of poor mappability and alignability, and also evaluated sequencing quality at outstanding candidate regions. Furthermore, we overlaid `VolcanoFinder` scan results with an independent scan using the T_2 statistic of `BALLET` [48] to investigate any co-localization with evidence for long-term balancing selection. We also examined the level of nucleotide diversity ($\hat{\theta}_\pi$) across the candidate regions, as well as the level of sequence uniqueness as a more stringent measure of mappability. In the scan on Europeans, we evaluated evidence for archaic introgression at candidate regions by examining non-synonymous differences with Neanderthals [60] as well as inferred Neanderthal or Denisovan introgression segments [20, 22]. See *Materials and Methods* for further details.

The top-scoring regions are reported in Table S3.1 (CEU population) and Table S3.2 (YRI population). Manhattan plots of the whole genome are shown in Fig. S3.1 (CEU) and Fig. S3.2 (YRI). In the CEU, we uncovered footprints of adaptive introgression on regions with putative Neanderthal ancestry, most notably the gene *TSHR* (Fig. 7) which encodes the receptor for thyroid stimulating hormone (TSH). Using eq. (4) and (5) with a recombination rate of $r = 10^{-8}$ recombination event per nucleotide per generation [61] and $N_e = 10^4$ [62], the inferred introgression parameters $\hat{\alpha}$ and $\hat{D}$ for the *TSHR* candidate region (Table S3.1) suggest a 41.7 kb volcano centered on a 2.4 kb valley. The ratio of polymorphic sites to fixed differences in the shoulders of this volcano $(175 : 372)$ is significantly higher than that of the genomic background (one-tailed binomial test, $p = 0.0137$) as well as that of the central region $(5 : 47)$ leading to a significant McDonald and Kreitman test [63] (one-tailed, $p = 2.6 \times 10^{-4}$). Since divergence between Neanderthals, Denisovans and modern humans is relatively recent (4.23–5.89% of the human-chimpanzee sequence divergence [57], leading to $D \approx 1.4\theta$–2θ according to our observations) and introgressed haplotypes typically do not reach high frequency in samples of modern human populations such as CEU, we do not expect `VolcanoFinder` to detect most of these signals.

On the other hand, we also found outstanding candidate regions devoid of known Neanderthal or Denisovan ancestries in the scan on Europeans. One such candidate is the *CHRNB3*-*CHRNA6* gene cluster (Fig. 8), which has been associated with substance dependence especially in Europeans (see *Discussion*). The inferred introgression parameters for this candidate region (Table S3.1) suggest a 36 kb volcano centered on a 2.1 kb valley. Once again, the ratio of polymorphic sites to fixed differences in the shoulders of the volcano $(178 : 259)$ is significantly higher than that of the genomic background (one-tailed binomial test, $p = 2.5 \times 10^{-9}$) as well as that of the central region $(5 : 21)$ leading to a significant McDonald and Kreitman test [63] (one-tailed, $p = 0.021$).

The most prominent signal across the genome in Europeans is also devoid of known archaic hominin ancestry. This region features the *APOL3* and *APOL4* (Fig. 9A) on chromosome 22, which encode apolipoportein L family proteins. The inferred introgression parameters for this candidate region (Table S3.1) suggest a 20 kb volcano centered on a 0.6 kb valley. Although this region is the most prominent candidate in our analysis, the polymorphic sites to fixed differences ratio is significantly higher than that of the genomic background in the right shoulder of the volcano only ($80 : 145$, one-tailed binomial test, $p = 0.006$). This indicates that the model-based method of `VolcanoFinder` relying on the whole SFS is more sensitive than the mere polymorphism:divergence ratio. The apolipoportein L family proteins are high density lipoproteins and take part in lipid transportation [64]. They are unique to the primate lineage, and have been hypothesized to be under positive selection in humans [65]. Intriguingly, we also estimated high likelihood ratio scores around this region in the African population scan (Fig. S3.3), although the peak locations in the two scans vary. Note that this candidate was not included in our final list of candidates for the YRI population (Table S3.2) due to the lack of data close to *APOL4* (Fig. S3.3A). However, the same genomic region in CEU (Fig. S3.4) does not exhibit high CLR scores despite the region devoid of data, lending support to the validity of the signals we observe in the scan on YRI. Instead of spanning across *APOL4* and *APOL3* like in CEU, the peak in YRI locates closer to *APOL2*, which closely neighbors *APOL1*.

In the African population scan, another interesting top-scoring region lies between the *TCHH* and *RPTN* genes on the epidermal differentiation complex (EDC) on chromosome 1 (Fig. 10). This gene complex features many genes essential for the late-stage differentiation of epidermal cells and is therefore important for the integrity and functionality of skin and skin appendages [66] such as hair and nails [67, 68]. The inferred introgression parameters for the *TCHH*-*RPTN* candidate region (Table S3.2) suggest a 23.3 kb volcano centered on a 1.3 kb valley. Although this region has the second-highest CLR in our candidate list, the ratio of polymorphic sites to fixed differences shows that the inferred shoulders are not enriched in polymorphic sites, although the one-tailed McDonald and Kreitman test between the shoulders and the valley is marginally significant ($64 : 159$ *vs.* $0 : 9$, $p = 0.0515$). In this case, `VolcanoFinder` may be sensitive to the skew in the SFS caused by the introgression sweep.

Lastly, we also applied `VolcanoFinder` on a dataset of 500 individuals drawn uniformly at random from the global set of samples from non-admixed populations in the 1000 Genome Project dataset. However we did not find strong support for any genomic region to have undergone adaptive introgression. This result agrees with our observations that the candidate regions in the scans on African and European populations barely overlap.

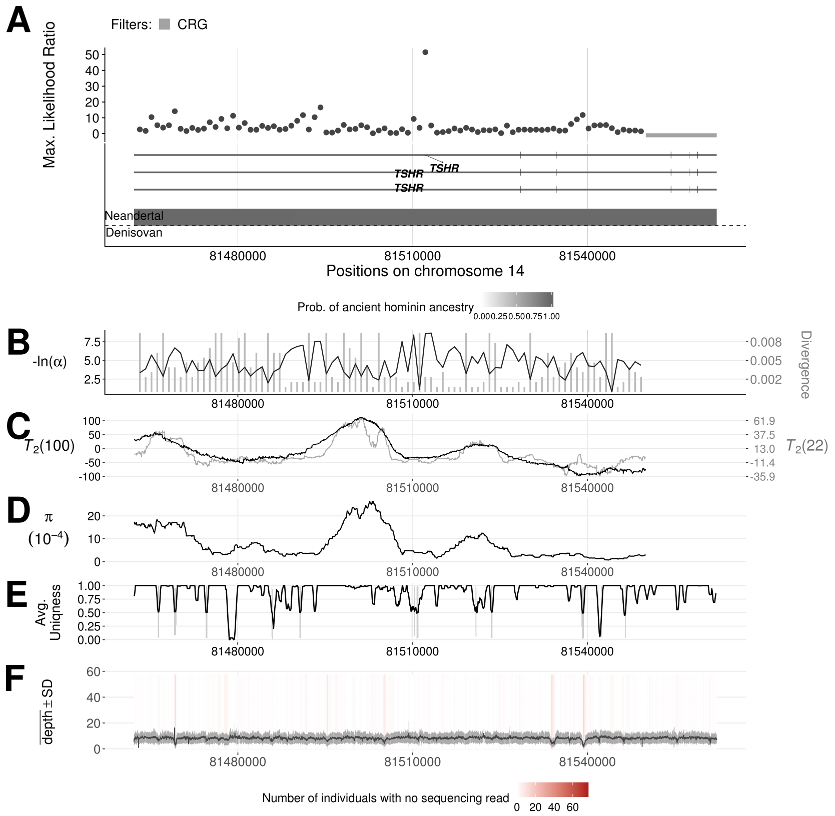

Fig 7. Introgression sweep signals, tracks of Neanderthal or Denisovan ancestry, parameter estimates, and sequencing properties across the 100 kb region on chromosome 14 covering the *TSHR* gene in CEU.

A. Likelihood ratio test statistic computed from Model 1 of `VolcanoFinder` on data on within-CEU polymorphism and substitutions with respect to chimpanzee. Horizontal light gray bars correspond to regions that were filtered based on mean CRG. Gene tracts and labels for key genes are depicted below the plot, with the wider bars representing exons. Tracks of putative regions with Neanderthal (above the horizontal line) or Denisovan (below the horizontal line) ancestry are located below gene diagrams. Higher probabilities of Neanderthal or Denisovan ancestry are depicted with darker colored bands (data from [22]). Non-synonymous mutations with Neanderthal are indicated in red. **B.** Values for α and divergence D corresponding to the maximum likelihood estimate of the data. Black line corresponds to $-\ln(\alpha)$ and vertical gray bars correspond to estimated D. **C.** Likelihood ratio test statistic computed from T_2 of `BALLET` on data on within-CEU polymorphism and substitutions with respect to chimpanzee using windows of 100 (black) or 22 (gray) informative sites on either side of the test site. **D.** Mean pairwise sequence difference ($\hat{\theta}_\pi$) computed in five kb windows centered on each polymorphic site. **E.** Mappability uniqueness scores for 35 nucleotide sequences across the region. **F.** Mean sequencing depth across the 99 CEU individuals as a function of genomic position, with the gray ribbon indicating standard deviation. The background heatmap displays the number of individuals devoid of sequencing reads as a function of genomic position, with darker shades of red indicating a greater number of individuals with no sequencing reads.

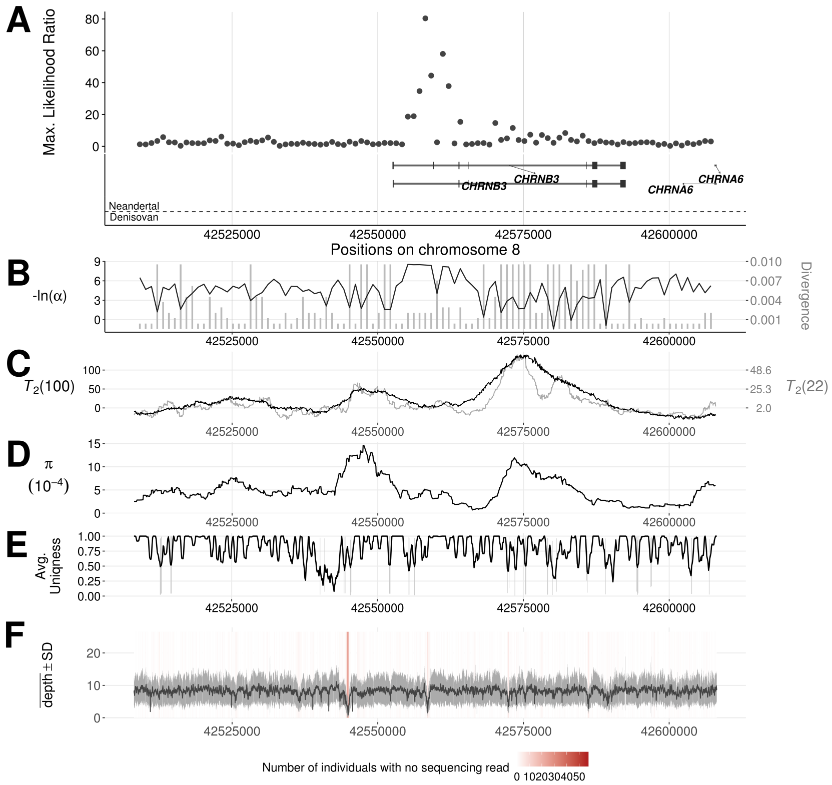

Fig 8. Introgression sweep signals, tracks of Neanderthal or Denisovan ancestry, parameter estimates, and sequencing properties across the 100 kb region on chromosome 8 covering the *CHRNB3* gene in CEU.
A. Likelihood ratio test statistic computed from Model 1 of `VolcanoFinder` on data on within-CEU polymorphism and substitutions with respect to chimpanzee. Horizontal light gray bars correspond to regions that were filtered based on mean CRG. Gene tracts and labels for key genes are depicted below the plot, with the wider bars representing exons. Tracks of putative regions with Neanderthal (above the horizontal line) or Denisovan (below the horizontal line) ancestry are located below gene diagrams. Higher probabilities of Neanderthal or Denisovan ancestry are depicted with darker colored bands (data from [22]). Non-synonymous mutations with Neanderthal are indicated in red. **B.** Values for α and divergence D corresponding to the maximum likelihood estimate of the data. Black line corresponds to $-\ln(\alpha)$ and vertical gray bars correspond to estimated D. **C.** Likelihood ratio test statistic computed from T_2 of `BALLET` on data on within-CEU polymorphism and substitutions with respect to chimpanzee using windows of 100 (black) or 22 (gray) informative sites on either side of the test site. **D.** Mean pairwise sequence difference ($\hat{\theta}_\pi$) computed in five kb windows centered on each polymorphic site. **E.** Mappability uniqueness scores for 35 nucleotide sequences across the region. **F.** Mean sequencing depth across the 99 CEU individuals as a function of genomic position, with the gray ribbon indicating standard deviation. The background heatmap displays the number of individuals devoid of sequencing reads as a function of genomic position, with darker shades of red indicating a greater number of individuals with no sequencing reads.

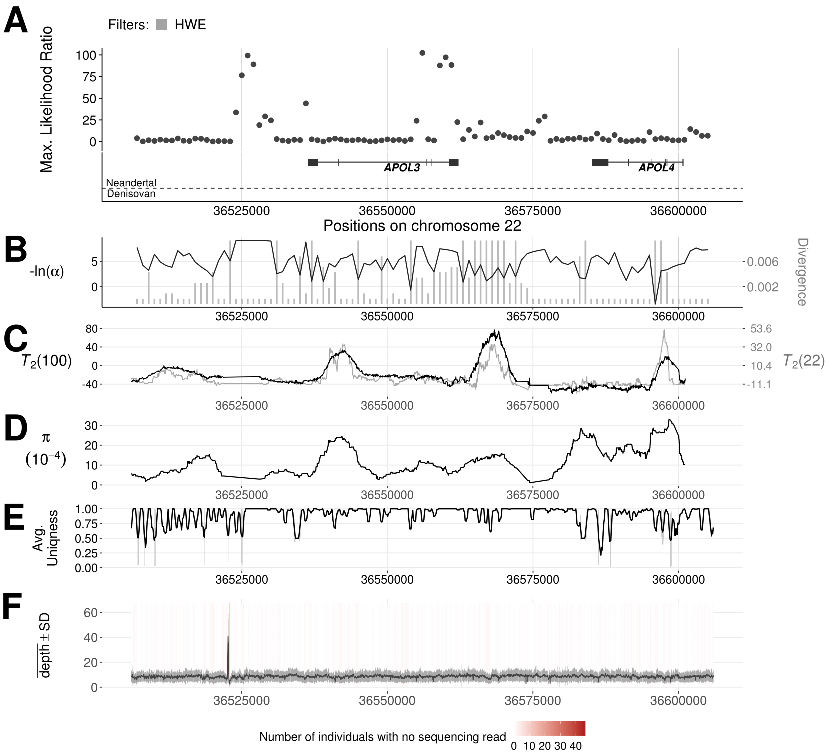

Fig 9. Introgression sweep signals, tracks of Neanderthal or Denisovan ancestry, parameter estimates, and sequencing properties across the 100 kb region on chromosome 22 covering *APOL* gene cluster in CEU.
A. Likelihood ratio test statistic computed from Model 1 of `VolcanoFinder` on data on within-CEU polymorphism and substitutions with respect to chimpanzee. Horizontal light gray bars correspond to regions that were filtered based on mean CRG. Gene tracts and labels for key genes are depicted below the plot, with the wider bars representing exons. Tracks of putative regions with Neanderthal (above the horizontal line) or Denisovan (below the horizontal line) ancestry are located below gene diagrams. Higher probabilities of Neanderthal or Denisovan ancestry are depicted with darker colored bands (data from [22]). Non-synonymous mutations with Neanderthal are indicated in red. **B.** Values for α and divergence D corresponding to the maximum likelihood estimate of the data. Black line corresponds to $-\ln(\alpha)$ and vertical gray bars correspond to estimated D. **C.** Likelihood ratio test statistic computed from T_2 of `BALLET` on data on within-CEU polymorphism and substitutions with respect to chimpanzee using windows of 100 (black) or 22 (gray) informative sites on either side of the test site. **D.** Mean pairwise sequence difference ($\hat{\theta}_\pi$) computed in five kb windows centered on each polymorphic site. **E.** Mappability uniqueness scores for 35 nucleotide sequences across the region. **F.** Mean sequencing depth across the 99 CEU individuals as a function of genomic position, with the gray ribbon indicating standard deviation. The background heatmap displays the number of individuals devoid of sequencing reads as a function of genomic position, with darker shades of red indicating a greater number of individuals with no sequencing reads.

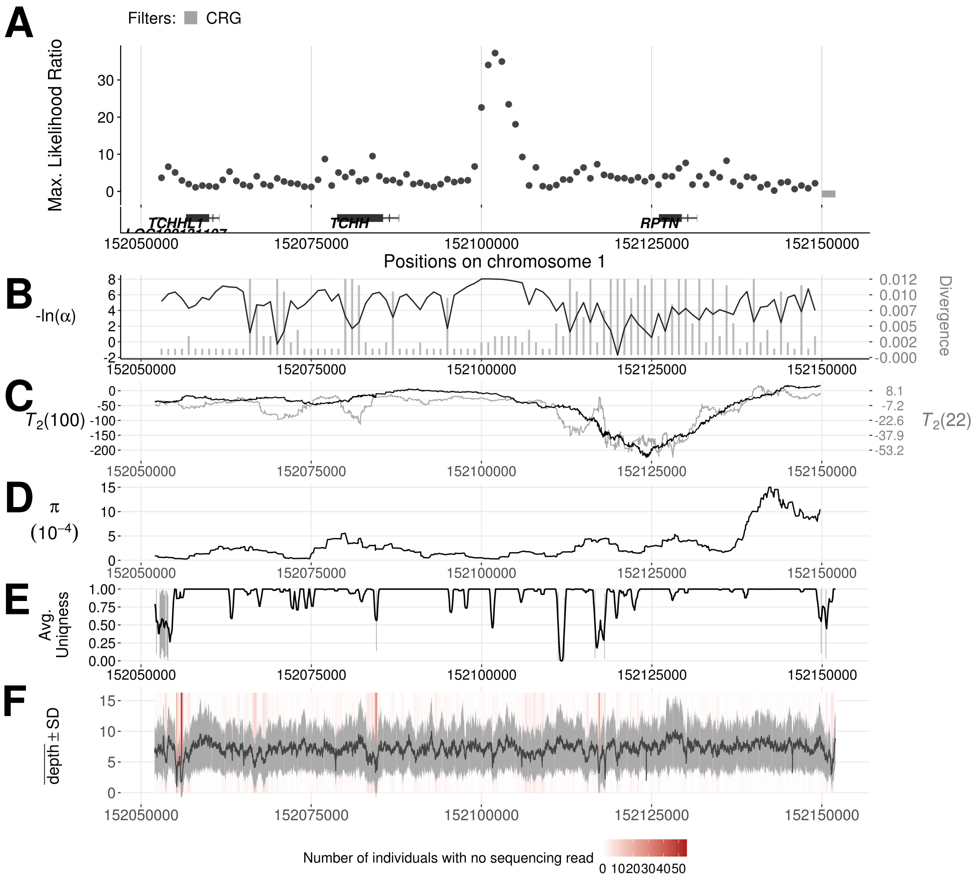

Fig 10. Introgression sweep signals, parameter estimates, and sequencing properties across the 100 kb region on chromosome 1 covering *TCHH* and *RPTN* genes in YRI.
A. Likelihood ratio test statistic computed from Model 1 of `VolcanoFinder` on data on within-YRI polymorphism and substitutions with respect to chimpanzee. Horizontal dark gray bars correspond to regions that were filtered based on mean CRG score. Gene tracts and labels for key genes are depicted below the plot, with the wider bars representing exons. **B.** Values for α and divergence D corresponding to the maximum likelihood estimate of the data. Black line corresponds to $-\ln(\alpha)$ and vertical gray bars correspond to estimated D. **C.** Likelihood ratio test statistic computed from T_2 of `BALLET` on data on within-YRI polymorphism and substitutions with respect to chimpanzee using windows of 100 (black) or 22 (gray) informative sites on either side of the test site. **D.** Mean pairwise sequence difference ($\hat{\theta}_\pi$) computed in five kb windows centered on each polymorphic site. **E.** Mappability uniqueness scores for 35 nucleotide sequences across the region. **F.** Mean sequencing depth across the 108 YRI individuals as a function of genomic position, with the gray ribbon indicating standard deviation. The background heatmap displays the number of individuals devoid of sequencing reads as a function of genomic position, with darker shades of red indicating a greater number of individuals with no sequencing reads.

Materials and Methods

Footprints of adaptive introgression: forward simulations

We used two distinct simulation approaches. The accuracy of the analytical predictions of the model was first studied using a mixed forward and backward method (described here) that fully simulates the stochastic trajectory of the selected allele, initially introduced in a single lineage. The power analysis was conducted in a second stage using a fully coalescent-based method (described in the section on power analysis below) that does not allow for direct control of the number of introgressed lineages, but enables to easily simulate hard and soft introgression sweeps, as well as to assess the effect of the genome-wide admixture resulting from secondary contact.

Due to the long divergence time, individual-based forward-time simulations of the full model are computationally expensive and time limiting. While the current coalescent-based method `msms` [69] can incorporate the effects of selection at a single locus, demography cannot be included when conditioning on the fixation of the foreign adaptive allele. This is because we cannot guarantee that, backward in time, the sweep will have completed before the allele returns to the common ancestral population.

To simulate the full model efficiently, we use a backward-time, forward-time approach. The coalescent simulator `msprime` [70] is capable of quickly simulating large genomic regions for even whole-population-sized samples. We use this to implement the model of divergence without gene flow among the donor and recipient populations, as well as a third distant outgroup. Sampling one lineage from the outgroup to polarize the data, we sample $2N-1$ lineages from the recipient population and one lineage from the donor population to form a diploid population containing a single hybrid individual. We then import the data to `simuPOP` [71]. In the foreign haplotype of the hybrid individual, we place at the center of the sweep region a beneficial allele with selective strength s. We repeatedly run the evolutionary model forward in time until an iteration with a successful sweep is found.

We simulate a genomic region that spans $R = rd = s$ left and right of the benefical mutation, as this covers the region where genetic diversity is increased. However, for computation speed, we do not simulate a continuous genome, but rather a set of 100-bp intervals centered at distance R. Here, the recombination rate per site r is low so that recombination within the windows is unlikely, but recombination between the windows occurs with appreciable chance. This ensures that the mean expected heterozygosity calculated for a given window is representative of the genealogical distribution specifically at that site. Furthermore, the mutation rate per site μ is chosen so that, even with high divergence, multiple mutation hits at a single site are unlikely.

Additional simulations that explore, for example, the genealogical distribution in Text S1.1, are also written using `simuPOP`. These are straightforward evolutionary models, and the additional simulation software is available upon request.

Software implementation

`VolcanoFinder` is implemented in the C programming language using much of the code base in `SweepFinder2` [50] as its foundation. The software takes in data on derived allele counts at biallelic sites ordered along a chromosome, employs information either on polymorphic sites or on both polymorphic sites and substitutions, and implements one of the four model combinations introduced here (Model 1 or 2, with or without fixed differences). The software also requires as input the empirical mutation frequency spectrum, which it uses as the null hypothesis in the composite likelihood calculation (as in [47]).

The user defines the number of test sites over which to compute the composite likelihood ratio test statistic, and these test sites are evenly spaced across the input genomic region or chromosome. Note that this implies that a test site does not need to be located on any particular data point. At a particular test site, VolcanoFinder searches a grid of divergence values D separating the donor and recipient populations and, for each, optimizes over the sweep strength α. By default, D is optimized over the grid $D \in \{\hat{\theta}_\pi, 2\hat{\theta}_\pi, \ldots, k\hat{\theta}_\pi\}$ under Model 1 and $D \in \{2\hat{\theta}_L, 3\hat{\theta}_L, \ldots, k\hat{\theta}_L\}$ under Model 2, where k is chosen as the maximum positive integer with $D \leqslant 2D_o$ if fixed differences are polarized or $D \leqslant D'_o$ if they are not. Here, D_o is the divergence between the recipient species and its MRCA with the outgroup species, D'_o is divergence between the recipient species and the outgroup, $\hat{\theta}_\pi$ is Tajima's estimator of the population-scaled mutation rate θ, and $\hat{\theta}_L$ is another unbiased estimator of θ. These values are computed internally in the software from the unnormalized site frequency spectrum, with D_o (or D'_o) computed as $S_1(1)$ (eq. (12) with polarized or non-polarized fixed differences), $\hat{\theta}_\pi = S_1(2)$, and $\hat{\theta}_L = \frac{1}{n-1}\sum_{i=1}^{n-1} iS_i(n)$. The user is also able to specify the set of D values that he wishes to cycle over instead.

Note that although VolcanoFinder can use all available data on an input chromosome to compute a composite likelihood ratio at a given test site, data points far from the test site will not alter this likelihood ratio, as the site frequency spectrum expected for such distant sites will be the same under neutrality as for adaptive introgression. For this reason, we follow the implementation used in SweepFinder [47] and cut the computation off when data points are distant enough from the test site. That is, we restrict the computation to data points in which $\alpha d \leqslant 12$, where d is the distance between the test site and a given data point. Furthermore, though the sweep strength parameter matches that of the original SweepFinder model [47], we found that the hard-coded limits on α in the SweepFinder implementations [47, 50] prevent the software from accurately detecting sweeps of that size. SweepFinder still has high power to observe the patterns of a classic hard sweep, identifying a model that underestimates the true strength of selection. Because VolcanoFinder relies on information further to the periphery of the sweep region, this generated a loss of power to detect sweeps, and strong sweeps are permitted in the VolcanoFinder software as a result. We therefore reduced the minimum α considered by VolcanoFinder by an order of magnitude compared to SweepFinder so that wide volcano patterns (*i.e.*, large d) can be observed by our method.

Because VolcanoFinder is computationally intensive, we provide several features in the software that allow introgression scans to run in parallel. First, for a given input dataset, the user can choose a number m such that the dataset is broken into m blocks of test sites with an equal number of contiguous test sites in each block. VolcanoFinder can then be applied to the same dataset m times, where each application it computes the values across the sites sites in one of the m blocks. These blocks of contiguous test sites can then be scanned separately on different compute cores, and an auxiliary script will merge the m scans into a single scan. In addition, for some users such a fine grid of D values may be unnecessary. To this end, the software also implements an option for specifying a single user-defined value for D—allowing to easily scan for adaptive introgression with many values of D simultaneously in parallel.

Power analysis

Model and simulation procedure

Coalescent simulations were performed with coala [72] as a frontend to msms [69]. We assume $n = 40$ lineages are sampled from a focal species and one lineage is sampled from an outgroup that diverged at time $T_{sp} = 10$ units of $4N$ generations in the past. Detailed descriptions are given in supp. Text S2.2.

For introgression sweeps, we model a secondary contact (Fig. S2.1) where the recipient (focal) and donor (unknown) species diverged at time $T_d < T_{sp}$ and a beneficial allele with selection coefficient s was introgressed from the donor into the recipient species during a short pulse of migration. The size of the donor species is adjusted either to enforce a hard introgression sweep or to allow the introgression of neutral polymorphism from the donor species, possibly leading to soft introgression sweeps. The migration parameters (migration rate, time and duration) are adjusted such that the fixation probability of the beneficial allele in the recipient species is high ($\pi_{\text{fix}} = 0.95$) and the introgression sweep ends at time T_s (see details in supp. Text S2.1). We assessed the effect of the divergence time ($T_d \in \{1, 2.5, 4, 5.5\}$, *i.e.* $D/\theta \in \{3, 6, 9, 12\}$), the ending time of the introgression sweep ($T_s \in \{0, 0.1, 0.25, 0.5\}$) and selection coefficient ($2Ns \in \{100, 1\,000\}$) for hard and soft introgression sweeps, leading to 64 parameter sets. Neutral coalescent simulations without admixture (one parameter set) or with the same level of admixture (64 parameter sets) were used as neutral references.

Coalescent simulations under three demographic models involving balancing selection (overdominance, Fig. S2.2) were also conducted to assess the robustness of `VolcanoFinder` to excess expected heterozygosity in the focal species caused by long term balancing selection starting at time T_s. Combining six values for $T_s \in \{1.25, 5, 8.75, 12.5, 16.25, 20\}$ and three demographic models leads to 18 parameter sets. Neutral coalescent simulations with the same demographic model (three parameter sets) were used as neutral references.

Statistical methods for power estimation

A detailed description is provided in Text S2.3. The genome-wide reference backgrounds used by all composite likelihood methods were obtained from neutral coalescent simulations.

For each simulated sequence, genome scan methods provide a list of locations for the selected locus and composite likelihood ratios. The maximum LR over a simulated sequence (or possibly in a smaller region) was used as a test statistics. For each parameter set, the null distributions of the test statistics were obtained from 10 000 neutral replicates and the rejection rates for increasing false positive rates (up to 5%) were estimated from 1 000 non neutral replicates.

In the case of introgression sweeps, two kinds of neutral references were used in separate analyses: either a non-admixed reference background (common to all parameter sets) or an admixed reference background (one per parameter set) with the same migration parameters as the associated non-neutral case. This enables to consider the two limiting cases where introgressed alleles are either quickly purged by natural selection (non-admixed background) or behave fully neutrally (admixed background).

The detection probability of an introgression sweep in a genome-wide study focussing on top candidates was estimated as the proportion of the 1 000 non-neutral replicates for which the highest LR would rank in the genome-wide top 50 peak values obtained under neutrality. Peak values were obtained from the 8×10^6 LR values generated by 10 000 neutral replicates. Neighbouring peaks (separated by less then 10 LR values) were merged.

Human data analysis: materials and methods to generate the CEU and YRI data

For each human population analyzed in this study, we used genotypes from variant calls of the 1000 Genomes Project Phase 3 dataset [58]. Alleles were polarized as derived or ancestral based on the allelic state in the aligned chimpanzee (panTro5) reference genome [59], and only mono- or bi-allelic single nucleotide sites that could be polarized were considered. As in [49], to ensure that we only used sites in regions of high mappability and alignability, we examined the mean CRG 100mer score for each 100 kilobase (kb) genomic region whose centers are spaced every 50 kb apart, and only considered sites in regions with a mean score no lower than 0.9.

Based on the filtered data, we summarized the non-normalized site frequency spectra for each population analyzed, and computed the per-site heterozygosity $\hat{\theta}_\pi$ across 108 Yoruban (YRI) and 99 European (CEU) individuals to be 0.001004392 and 0.0007584236, respectively, which is in line with previous estimates of the mutation rate [73]. Furthermore, from these frequency spectra, we also computed each population's divergence D_0' with chimpanzees as 0.01251347 and 0.01251496 for YRI and CEU, respectively, which is also in line with prior estimates [73]. The genome-wide proportion of polymorphic sites among informative sites (polymorphic sites and fixed differences) was 0.3905585 and 0.2763345 for YRI and CEU, respectively. We applied VolcanoFinder on their genomic data accordingly, placing a test site every one kb across each autosome. To mask the test sites falling in regions with missing or potentially problematic data, we removed from downstream analyses test sites in the aforementioned 100 kb windows with mean scores lower than 0.9, as well as test sites within 100 kb of a centromeric or telomeric region.

Candidate loci were defined as showing a peak of CLR values. We used a minimum ln(CLR) of 20 and a minimum distance of 15 kb between peaks. In order to remove artifactual candidates, we discarded candidates that stood in regions depleted of informative sites (the minimum distance to the nearest informative site had to be lower than the 0.9995 quantile of the distribution of the distances between consecutive informative sites on the same chromosome) and only retained candidates for which the inferred selection parameters were compatible with the typical volcano footprint of an introgression sweep ($\hat{D} > \hat{\theta}_\pi$ and a volcano half-width, inferred from $\hat{\alpha}$, $\hat{D}$ and eq. (5), larger than 5 kb). Eq. (5) suggests that this minimum half-width corresponds to a compound selection parameter $2N_e s \geqslant 2.7$ given realistic values for the recombination rate and the effective population size in humans, $r = 10^{-8}$ recombination events per nucleotide per generation [61] and $N_e = 10^4$ [62]. Such a low value enables us to take into account the variance of local recombination rates and the intrinsic trend of VolcanoFinder to underestimate the selection coefficient for old introgression events (Fig. S2.16 and Fig. S2.17).

To further curate empirical candidates, we generated the sequencing coverage based on the BAM files of each individual included in the dataset for a particular population (YRI or CEU). For each population, sample-wide mean sequencing depth and the corresponding standard deviation were computed and used as a reference for assessing candidate regions. As a complementary measure, we also considered the number of individuals devoid of sequencing reads at a particular genomic position to further examine data quality. Furthermore, we examined the mappability uniqueness of each 35 nucleotide sequence (data from [74]; accessed via UCSC Genome Browser) for all candidate regions. This criterion can further flag potential issues with sequence mapping. Moreover, to investigate potential sources of introgression, we also examined the non-synonymous differences between modern humans and Neanderthals [60], as well as the regions of mapped Neanderthal or Denisovan introgression segments that intersect candidate regions in the CEU population [20, 22]. Moreover, to investigate whether introgression sweep signals co-occur with signals of ancient balancing selection, we applied the T_2 from BALLET [48] statistic to the same polymorphism and substitution data on which VolcanoFinder was applied, and filtered the output with the same filters we applied to VolcanoFinder output.

Finally, in order to characterize a predicted increase of the polymorphism:divergence ratio in the shoulders of candidate volcanoes of introgression, the counts of polymorphic sites and fixed differences were obtained in the inferred valley and shoulder regions (distances from the LR-peak given by $\hat{\alpha}$, $\hat{D}$ and eq. (4) and (5)). The polymorphism:divergence ratios in the volcano shoulders were compared to that of the genomic background using a one-tailed binomial test and to those of their central valley in a one-tailed McDonald and Kreitman test [63].

Discussion and conclusions

The hitchhiking of foreign genetic variation during adaptive introgression from a diverged donor population generates a unique volcano-shaped signature in the genetic diversity of the recipient population. Such patterns have first been described for an island model in the limit of low migration rates [38]. Here, we characterize the pattern for a scenario of secondary contact and use it to construct a genome scan method to detect recent events of adaptive introgression from sequence variation in the recipient species, without the need to know the donor species.

In sharp contrast to a classical sweep, introgression sweeps have only a narrow (expected) valley of reduced diversity around the selected site, but broad flanking regions with an excess of intermediate-frequency polymorphism relative to fixed differences to an outgroup. This excess variation is the most prominent feature of the footprint and is observed for both hard and soft introgression sweeps (*i.e.* sweeps originating from one or several hybrids, see Fig. S2.3 and Fig. S2.4). It remains visible for extended periods of time after completion of the sweep (up to $\sim 2N$ generations, where N is the effective population size).

The construction of a mathematical model for the purpose of a parametric test requires a compromise between precision and tractability. Even for the simple measure of pairwise genetic diversity, accurate predictions require approximations with several parameters to account for the variance in coalescence time during the sweep [30, 51], see also our models in the electronic supplement (Text S1.1). However, our results show that an extended star-like approximation with only two parameters, α for the strength of the sweep and D for the divergence of the recipient population from the donor, offers a flexible scheme to match simulated volcano footprints for both hard and soft introgression sweeps.

The use of α and D as flexible fit parameters poses a challenge when interpreting them as estimators for the true strength of the sweep and true divergence of the donor population. In particular, comparison with accurate approximations and simulations shows that the star-like model overestimates the predicted genetic diversity. Hence, the optimal D found by `VolcanoFinder` is biased to underestimate the true divergence of the donor population.

There are further limits to the simple star-like model. Simulations show that volcano patterns are often strongly asymmetric and/or truncated due to early recombination events (Fig 3). The model also assumes that the population is sampled directly after completion of the sweep in the recipient population. Older sweep footprints may still show pronounced regions of excess variation, but could have recovered close to normal polymorphism level in the central sweep valley. More complex patterns are also expected if introgression haplotypes harbor more than a single selected allele in tight linkage. In particular, the beneficial allele can be linked to barrier genes that reduce the introgression probability and bias the footprints of successful introgression sweeps [75]. Inclusion of any such details into a statistical test would, however, require additional model parameters. For whole-genome scans, the higher-dimensional optimization that is required in this case can easily prove computationally prohibitive.

Power analysis

The footprint of adaptive introgression combines elements of a classic selective sweep (a sweep valley) with signals that are more typical of balancing selection (excess variation at intermediate frequencies). Accordingly, we tested the power of our new method `VolcanoFinder` to detect introgression sweeps relative to two standard methods that were designed to detect classic selective sweeps (`SweepFinder` 2 [47]) and long-term balancing selection (`BALLET`, [50]), respectively. In addition to ROC curves

(Fig. S2.5 to Fig. S2.7) that are typically presented in power analyses [45, 48, 49], we provide an alternative analysis that is closer to the use of a test in a real genome scan. To this end, we estimated the probability that an introgression locus ranks among the top 1 to 50 highest CLR peaks (Figs. 5 and 6) among peaks obtained from 8×10^6 CLR values from 10 000 neutral replicates, which represent a whole-genome background. This approach is particularly useful for composite-likelihood tests (all three tests considered here), where standard methods for multiple-testing correction [76] that rely on independent p-values do not apply.

Our model postulates that an introgression sweep occurred as a result of a rare hybridization event caused by a secondary contact between diverged species (see Fig 1). In our power analysis we adjusted the migration rate during this secondary contact to achieve a high fixation probability for the introgressed beneficial allele. This raises the question of the potential effect of admixture in the genomic background. We therefore explored two extreme cases: (i) a non-admixed genomic background and (ii) a neutrally admixed genomic background resulting from the same amount of admixture that allowed the introgression sweep to occur with a high probability. In natural populations, post-zygotic genetic barriers [77] will typically purge part of the introgressed variation, thus reducing the genome-wide admixture to some intermediate level between these limiting cases.

We find that VolcanoFinder has very high genome-wide power to detect introgression sweeps against a non-admixed background (test for local introgression, Fig. 5). It clearly outcompetes the methods that have been developed for other purposes. This power is strongly reduced if the genomic background harbors high levels of neutral admixture. However, the detection probability remains high if admixture in the background is moderate and if adaptation occurs from a strongly diverged donor population ($2Ns = 1\,000$ in Fig. 6).

Although our sweep model assumes that adaptation in the recipient population starts from a single hybrid individual, VolcanoFinder has virtually the same power to detect hard and soft introgression sweeps. This is in sharp contrast to the detection of classic sweeps in a single panmictic population by methods like SweepFinder 2. The small reduction in power for soft introgression sweeps is expected because the typical volcano patterns do not differ much between hard and soft sweeps, as explained above. We expect the same qualitative pattern also in the case of incomplete introgression sweeps, as long as the adaptive allele reaches sufficiently high frequencies $> 50\%$ in the recipient population. This suggests that VolcanoFinder may also detect these events with high power, but we did not test this case and quantitative predictions remain to be established.

A significant finding is the relatively high power of VolcanoFinder to detect old introgression sweeps. We tested this power for $T_s \leqslant 0.5$, or $2N$ generations, clearly beyond the detection limit of genome scanners for classic sweeps [49]. As an example, consider an introgression event with $2Ns = 1\,000$, $T_d = 4$ ($D = 9\theta$), and admixed background in Fig. 6. The average probability that the introgression locus ranks among the top 50 peaks is around 66% for recent events $0 \leqslant T_s \leqslant 0.1$, but still around 33% for old events $0.1 \leqslant T_s \leqslant 0.5$. Assuming a constant rate of introgression, we expect two times as many old events than recent events because of the four times larger time window for old events. This expected enrichment in old events is even stronger with a non-admixed genomic background (Fig 5).

For an estimation of selection strength and divergence, we studied the marginal distributions of the maximum composite likelihood parameter estimates. Our results confirm the expected underestimation of D (Fig. S2.14 and Fig. S2.15). In contrast, the estimation of the selection parameter $-\log_{10}(\alpha)$ is relatively accurate (Fig. S2.16 and Fig. S2.17). This also holds for the estimated location of the selected allele (Fig. S2.8

and Fig. S2.9). The variance of all estimates increases with older introgression sweep events. Notably, significant CLR values were only rarely associated with low estimates of $D = \theta$, typical of a classic selective sweep, even when we considered an introgression sweep from a very closely related donor species ($T_d = 1$, *i.e.* $D = 3\theta$). This suggests that CLR peaks in with low $D = \theta$, but very high $-\log_{10}(\alpha)$ should be considered with caution in genome scans.

Incomplete lineage sorting, which is ignored in our model, is likely in scenarios with closely-related donor and recipient species. Relaxing the assumption of complete lineage sorting may thus improve the model, especially for low divergence. However, any extension requires a more detailed knowledge of the past demography in the donor and recipient species or its estimation from genomic background variation [78, 79]. Also, introgression sweeps from a very recently diverged donor are only expected to leave weak volcano signals and may be more readily detectable with a classic genome scanner.

Several methods have been proposed to detect gene flow that could be used to identify introgressed regions (see [80] for a review). Some rely on the detection of outlier values for indicators of divergence such as F_{ST} [81], Patterson's D (also known as ABBA BABA, [56, 82]) or $G_{\min}$ [83]. Others are likelihood and model-based, relying on the site frequency spectrum [84, $\partial a \partial i$], hidden Markov models for the coalescent tree [85, `TreeMix`] or use approximate likelihood methods such as ABC [86]. Finally, simulated data can be used to train computer algorithms to detect footprints of introgression generated under a particular introgression model [20, 22, 25]. These methods are however not aimed at detecting the specific signature of genetic hitchhiking with an introgressed selected allele.

Like other SFS-based methods, `VolcanoFinder` assumes independence between neighbouring SNPs and is blind to strong LD patterns resulting from gene flow [87]. The rate of exponential decrease of linkage disequilibrium can be used to date admixture events [88], and sophisticated haplotype-based methods have been used to characterize admixture and selection in ancestral human populations [89, 90]. Positive selection also increases LD [91–93], and methods were proposed to employ haplotype structure to date the MRCA of a beneficial allele [94]. Haplotype-based methods are usually powerful at detecting even soft and partial classic selective sweeps [45]. In an introgression sweep, positive selection and gene flow synergistically create a pattern of long and very diverged haplotypes. Including haplotype information into `VolcanoFinder` would thus almost certainly improve its power, especially for recent introgression events, as haplotype structure is expected to be informative over shorter time scales than patterns in the site-frequency spectrum [95].

Assessing evidence for adaptive introgression at empirical candidates

We applied `VolcanoFinder` to variant calls to probe for footprints of adaptive introgression in contemporary sub-Saharan-African and European human populations. With careful filters and quality-checks both before and after scans, we identified several candidate regions that may lend insights to early human evolutionary history. For application of `VolcanoFinder`, we warrant caution during data preparation and scrutiny over result interpretation, and believe it is especially important to consider factors such as the sequencing and mapping quality as well as values of other key statistics.

When preparing input for `VolcanoFinder`, we considered only regions with high mapping quality, as erroneous mapping may produce mis-matched variant calls that artificially alter the diversity of a genomic region. Specifically, following [49], we filtered 100 kb genomic segments with mean CRG100 scores less than 0.9. Such extended segments were chosen due to sweeps often affecting large genomic regions. Because

VolcanoFinder places test sites evenly across a chromosome, for test locations within large masked regions (or in the middle of centromeres) devoid of data, the diversity levels at the edges of these regions may appear higher than expected under neutrality, coincidentally mirroring the “adaptive ridges” of increased diversity expected near an adaptive introgression allele. Consequently, test sites within masked regions may exhibit abnormally high likelihood ratio scores. Therefore, extended genomic regions of non-missing data are desired to circumvent this potential artifact. Furthermore, due to this characteristic, output test sites should also be filtered, and it is preferable that the mask applied on the output data be more stringent than that on the input data, such that abnormalities around the filtered regions can be removed. In particular, we computed the mean CRG100 filter in 100 kb windows that overlap by 50kb. Overlapping regions in which only one of the windows passed the CRG filter were retained in the input but were excluded in the output. Moreover, we removed regions flanking telomeres and centromeres which can be difficult to map [96, 97] and may harbor increased diversity due to their repetitive nature.

After removing regions based on CRG mappability and proximity to telomeric and centromeric regions *post hoc*, we still observed that many genomic regions which passed the filters (*e.g.*, the *PTPRN2* gene region in the scan on YRI; Fig. S3.5) exhibit extremely low D values and high $-\log_{10}(\alpha)$. Such parameter combinations are unlikely to result from true footprints of adaptive introgression and should not be considered as genuine signals. Moreover, we noticed that these test sites often appear within or near regions devoid of data. Because gaps in input data may also be introduced in regions without mappable outgroup sequences (*e.g.*, the *PCAT*/*CEACAM4* and *B4GALNT2* regions in the scan on YRI (Fig. S3.6 and Fig. S3.7), we further removed test sites falling in regions with outstandingly large between-informative-site distances compared with the empirical distribution of all distances. We advise users to adopt similar screening procedures on the output data from the scan in order to exclude artifacts.

To curate the candidate regions that passed all filters, we further consulted the 35-mer sequence uniqueness scores (a more stringent measure of mappability) and the sample-wise mean sequencing read depths in order to gauge how confident we can be in the accuracy of the input data. Specifically, regions with low uniqueness may be mapped to sequencing reads from other paralogous regions and exhibit artificially high levels of variation. Sequences with low read depth may harbor unreliable variant calls, whereas those with abnormally high depth may suggest either structural variation or that sequencing reads from other regions in the genome were erroneously mapped to the region. In this light, we flagged the candidate regions with low uniqueness or abnormal mean read depths, especially when these features manifest on the lips of the “volcano” where the sequence diversity $\hat{\theta}_\pi$ is high. Examples of such regions are the *MUC4* (Fig. S3.8) gene in the scan on CEU as well as the *CYP2B6*-*CYP2B7* gene region in the scan of YRI (Fig. S3.9)—both regions harbor areas of low sequencing depth close to the CLR peak. The *MUC4* candidate region was discarded from our final list because of the neighbouring CRG-filtered region (Fig. S3.8A) whereas the *CYP2B6*-*CYP2B7* gene region passed this filtering step and is actually the top candidate in the YRI list (Table S3.2). We advise users to consider such candidates with caution and possibly look for other evidence of an introgression sweep. As far as the *CYP2B6*-*CYP2B7* gene region is concerned, the polymorphism and divergence pattern in the CEU sample shows no support for an introgression sweep either, suggesting that this introgression signal might be an artifact.

Although the sequencing read depth and sequence uniqueness alone are insufficient to determine whether the observed high likelihood ratios are the result of artifacts, reasonable read depths and high sequence uniqueness nonetheless provide strong support that the footprints observed at candidate regions are genuine. To provide additional support for footprints of adaptive introgression, we also consulted the values of the `BALLET` T_2 statistic at putative adaptive regions. Because T_2 is sensitive to ancient balanced alleles, it may report slightly elevated scores for introgressed regions and low scores for sweep regions. Therefore, in putative adaptively-introgressed regions, we should not only see high likelihood ratios reported by `VolcanoFinder`, but also expect to see a concomitant dip in T_2 scores, consistent with the "volcano"-shaped footprint of nucleotide diversity. We are able to find these supporting features in *TSHR*, *CHRNB3*, and *APOL3* gene regions in the scan on CEU (Fig. 7, 8, and 9, respectively), as well as the *TCHH*-*RPTN* intergenic region in YRI (Fig. 10).

Implications of the `VolcanoFinder` scans in Europeans and Africans

After careful screening and curating of the candidate genes from our scans on contemporary Europeans (CEU) and sub-Saharan Africans (YRI), we reported 27 candidate regions in CEU and 7 candidate regions in YRI. With out-of-African populations having more contact with archaic hominins, it is sensible that we are identifying a greater number of candidate genes in Europeans than in Africans. Among the candidate regions reported, we found the *TSHR*, *CHRNB3*, and *APOL4* gene regions particularly interesting in CEU, and the *TCHH*-*RPTN* region highly interesting in YRI. Meanwhile, the lack of strong support for any genomic region in the scan on the pooled global population indicates that detectable adaptive introgression events with other hominins prior to the migration out of Africa may be unlikely.

In CEU, we found both strong evidence for adaptive introgression and Neanderthal ancestry in the *TSHR* gene (Fig. 7). This gene encodes the receptor for TSH, or thyrotropin, the pituitary hormone that drives the production of the thyroid hormones [98]. In addition to its pivotal role in thyroid functions and the thyroid-mediated energy metabolism in most tissues, the TSH receptor has also been shown to take part in skeletal remodeling [99, 100], epidermal functions and hair follicle biogenesis [101–103], gonad functionality [104], as well as immunity [105]. Moreover, accumulating evidence also show its expression in adipose tissues [106, 107], and that it can regulate lipolysis [108, 109] and thermogenesis [107, 110]. Considering the contrasting climates of Europe and Africa, we speculate that the selective pressure on the *TSHR* gene in Europeans may be explained by the need to update their thermo-regulation in response to the colder climate. As the Neanderthal would have been better adapted to the local environment by the time humans expanded out of Africa, it is also sensible that this genomic region carries considerable Neanderthal ancestry (Fig. 7A).

In contrast, the second highest candidate, the *CHRNB3* gene region, does not carry substantial Neanderthal ancestry (Fig. 8). This gene encodes a nicotinic cholinergic receptor, and modulates neuronal transmission on synapses. Multiple genetic variants on this locus have been repeatedly associated with substance dependence, including smoking behavior [111, 112], nicotine dependence [113, 114], alcohol consumption [115, 116], and cocaine dependence [116]. Furthermore, in cross-ethnicity studies, SNPs on this locus not only have higher allele frequencies in non-African populations [113], but also have a smaller effect on the nicotine dependence behavior in African Americans than European Americans [114]. The absence of Neanderthal or Denisovan ancestry around the footprints of adaptive introgression, the moderate inferred divergence value $\hat{D}$ (Fig. 8B, $\hat{D} = 0.023 = 3\hat{\theta}_\pi$), as well as the higher allele frequencies in non-Africans relative to African populations, may suggest a possible encounter with an unknown archaic hominin population during the out-of-Africa migration.

Also devoid of known archaic hominin ancestry, the top candidate *APOL* gene cluster in CEU (Fig. 9 and Fig. S3.1) not only exhibits substantially higher CLR scores than other candidates, but also shows evidence for adaptive introgression in YRI. Our closer inspection show that the peaks in the two scans do not co-localize, with the peak in CEU spanning *APOL3*, whereas the peak in YRI locating closer to *APOL2*, which closely neighbors *APOL1*. A potential interpretation for this observation is that an introgression event predated the split of African and non-Africans, with variants around *APOL2* and *APOL1* advantageous to the local environments of Africans, whereas other variants between *APOL4* and *APOL3* on a different haplotype were subject to a different source of selective pressure in non-Africans. In line with this interpretation, in addition to its influence on blood lipid levels [117], APOL1 can also form pores on lysosomes after being engulfed and kill *Trypanosoma* parasites [118]. The *Trypanosoma* are known for causing sleeping sickness (*i.e.*, trypanosomiasis) and have been rampant in Africa [119, 120]. Moreover, though some subspecies of *T. brucei* have evolved to be resistant to it, some genetic variants unique in African human populations have been shown to counteract their defense [121, 122]. In the absence of this pathogenic threat, however, enhancing APOL1's trypanosome lytic activity in turn elevates the risk of cardiovascular diseases and chronic kidney diseases [122–124]. These diverse features of the *APOL* gene cluster may provide a biological basis for distinct selective pressures in Africans and non-Africans.

Further echoing the recent evidence for archaic introgression in African humans, we found strong evidence for the *TCHH*-*RPTN* region in YRI to carry footprints of adaptive introgression. The gene *TCHH* encodes trichohyalin, a precursor protein crosslinked with keratin intermediate filaments in hair follicle root sheaths and hair medula [125–128], and is crucial for hair formation [127, 129]. In fact, SNPs in this gene have been associated with straighter hair in Europeans [127, 130], as well as Latin Americans [131]. The gene *RPTN*, on the other hand, encodes repetin, another keratin filament-associated protein expressed in skin [132]. Although its exact biological role awaits further elucidation, probably due to its relatively recent discovery, an increase of *RPTN* expression was observed in clinical cases of atopic dermatitis [133]. Further, variants in *RPTN* have also been recently reported to also associate with straight hair in both Europeans and East Asians [134]. In African populations, although it is suggested that variation in curly hair is a complex trait that involves many genes, *TCHH* is among the candidate genes [128]. The footprints of adaptive introgression on this locus therefore imply a potential setting in which the ancestors of contemporary African populations acquired the adaptive alleles from a possible admixture with an unknown archaic hominin, resulting in, at least, beneficial phenotypes of hair morphology and curvature.

Taken together, our scans for adaptive introgression on two human populations have not only recovered candidate regions in Europeans that align with previous observations of Neanderthal and Denisovan ancestry (*e.g.*, *TSHR*), but also revealed novel candidates in both Europeans and Africans that locate in regions without evidence for introgression from known archaic hominins. These results lend insights on the environmental selective pressure, such as lipid and energy metabolism and pathogen defense, that may have acted on early humans. Furthermore, together with the inferred divergence time as well as the reference of introgressed regions from known archaic hominins, we have assembled a set of clues related to the distribution of as-yet-unknown archaic humans and their interactions with our ancestors.

Acknowledgments

This research was funded by National Institutes of Health grant R35GM128590, National Science Foundation grant DEB-1753489, and the Alfred P. Sloan Foundation. DS was funded by the Austrian Science Fund (FWF): DK W-1225-B20, Vienna Graduate School of Population Genetics. https://www.popgen-vienna.at/ https://fwf.ac.at The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC). Portions of this research were conducted with Advanced CyberInfrastructure computational resources provided by the Institute for CyberScience at Pennsylvania State University.

References

1. Coyne JA, Orr HA, Orr HA. Speciation. Oxford University Press Inc; 2004.

2. Mallet J. Hybridization as an invasion of the genome. Trends in ecology & evolution. 2005;20(5):229–237.

3. Baack EJ, Rieseberg LH. A genomic view of introgression and hybrid speciation. Current opinion in genetics & development. 2007;17:513–518. doi:10.1016/j.gde.2007.09.001.

4. Arnold ML, Sapir Y, Martin NH. Review. Genetic exchange and the origin of adaptations: prokaryotes to primates. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2008;363:2813–2820. doi:10.1098/rstb.2008.0021.

5. Schwenk K, Brede N, Streit B. Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2008;363:2805–2811. doi:10.1098/rstb.2008.0055.

6. Hedrick PW. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Molecular ecology. 2013;22:4606–4618. doi:10.1111/mec.12415.

7. Consortium HG. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 2012;487:94–98. doi:10.1038/nature11041.

8. Whitney KD, Randell RA, Rieseberg LH. Adaptive introgression of herbivore resistance traits in the weedy sunflower *Helianthus annuus*. The American naturalist. 2006;167:794–807. doi:10.1086/504606.

9. Whitney KD, Randell RA, Rieseberg LH. Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus*. The New phytologist. 2010;187:230–239. doi:10.1111/j.1469-8137.2010.03234.x.

10. Song Y, Endepols S, Klemann N, Richter D, Matuschka FR, Shih CH, et al. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. Current biology : CB. 2011;21:1296–1301. doi:10.1016/j.cub.2011.06.043.

11. Norris LC, Main BJ, Lee Y, Collier TC, Fofana A, Cornel AJ, et al. Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. Proceedings of the National Academy of Sciences of the United States of America. 2015;112:815–820. doi:10.1073/pnas.1418892112.

12. Paoletti M, Buck KW, Brasier CM. Selective acquisition of novel mating type and vegetative incompatibility genes via interspecies gene transfer in the globally invading eukaryote *Ophiostoma novo-ulmi*. Molecular ecology. 2006;15:249–262. doi:10.1111/j.1365-294X.2005.02728.x.

13. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. Nature reviews Genetics. 2015;16:359–371. doi:10.1038/nrg3936.

14. Dannemann M, Racimo F. Something old, something borrowed: admixture and adaptation in human evolution. Current opinion in genetics & development. 2018;53:1–8. doi:10.1016/j.gde.2018.05.009.

15. Dolgova O, Lao O. Evolutionary and Medical Consequences of Archaic Introgression into Modern Human Genomes. Genes. 2018;9. doi:10.3390/genes9070358.

16. Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014;512:194–197. doi:10.1038/nature13408.

17. Dannemann M, Andrés AM, Kelso J. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. American journal of human genetics. 2016;98:22–33. doi:10.1016/j.ajhg.2015.11.015.

18. Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova JL, et al. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. American journal of human genetics. 2016;98:5–21. doi:10.1016/j.ajhg.2015.11.014.

19. Gittelman RM, Schraiber JG, Vernot B, Mikacenic C, Wurfel MM, Akey JM. Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. Current biology : CB. 2016;26:3375–3382. doi:10.1016/j.cub.2016.10.041.

20. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014;507:354–357. doi:10.1038/nature12961.

21. Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science (New York, NY). 2016;352:235–239. doi:10.1126/science.aad9416.

22. Sankararaman S, Mallick S, Patterson N, Reich D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Current biology : CB. 2016;26:1241–1247. doi:10.1016/j.cub.2016.03.037.

23. Plagnol V, Wall JD. Possible ancestral structure in human populations. PLoS genetics. 2006;2:e105. doi:10.1371/journal.pgen.0020105.

24. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell. 2018;173:53–61.e9. doi:10.1016/j.cell.2018.02.031.

25. Durvasula A, Sankararaman S. A statistical model for reference-free inference of archaic local ancestry. PLoS genetics. 2019;15:e1008175. doi:10.1371/journal.pgen.1008175.

26. Vernot B, Akey JM. Resurrecting surviving Neandertal lineages from modern human genomes. Science (New York, NY). 2014;343:1017–1021. doi:10.1126/science.1245938.

27. Suarez-Gonzalez A, Lexer C, Cronk QCB. Adaptive introgression: a plant perspective. Biology letters. 2018;14. doi:10.1098/rsbl.2017.0688.

28. Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974;23(1):23–35.

29. Kaplan NL, Hudson RR, Langley CH. The "hitchhiking effect" revisited. Genetics. 1989;123(4):887–899.

30. Barton NH. The effect of hitchhiking on neutral genealogies. Genet Res. 1998;72:123–133.

31. Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 2005;169(4):2335–2352. doi:10.1534/genetics.104.036947.

32. Przeworski M, Coop G, Wall JD. The signature of positive selection on standing genetic variation. Evolution; international journal of organic evolution. 2005;59:2312–2323.

33. Pennings PS, Hermisson J. Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration. Mol Biol Evol. 2006;23(5):1076–1084. doi:10.1093/molbev/msj117.

34. Pennings PS, Hermisson J. Soft sweeps III: the signature of positive selection from recurrent mutation. PLoS Genet. 2006;2(12):e186. doi:10.1371/journal.pgen.0020186.

35. Peter BM, Huerta-Sanchez E, Nielsen R. Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS genetics. 2012;8:e1003011. doi:10.1371/journal.pgen.1003011.

36. Hermisson J, Pennings PS. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods in Ecology and Evolution. 2017;8(6):700–716. doi:10.1111/2041-210X.12808.

37. Slatkin M, Wiehe T. Genetic hitchhiking in a subdivised population. Genet Res Camb. 1998;71:155–160.

38. Santiago E, Caballero A. Variation After a Selective Sweep in a Subdivided Population. Genetics. 2005;169(1):475–483.

39. Wiehe T, Schmid K, Stephan W. Selective sweeps in structured populations - empirical evidence and theoretical studies. In: Nurminsky D, editor. Selective sweeps. Georgetown, US: Landes Biosciences; 2005. p. 104–117.

40. Bierne N. The distinctive footprints of local hitchhiking in a varied environment and global hitchhiking in a subdivided population. Evolution: International Journal of Organic Evolution. 2010;64(11):3254–3272.

41. Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2010;365:185–205. doi:10.1098/rstb.2009.0219.

42. Hohenlohe PA, Phillips PC, Cresko WA. Using population genomics to detect selection in natural populations: key concepts and methodological considerations. International journal of plant sciences. 2010;171:1059–1071. doi:10.1086/656306.

43. Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. Genome Res. 2010;20(3):393–402. doi:10.1101/gr.100545.109.

44. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics. 2013;193:929–941. doi:10.1534/genetics.112.147231.

45. Vatsiou AI, Bazin E, Gaggiotti OE. Detection of selective sweeps in structured populations: a comparison of recent methods. Molecular ecology. 2016;25:89–103. doi:10.1111/mec.13360.

46. Kim Y, Stephan W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics. 2002;160(2):765–777.

47. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005;15(11):1566–1575. doi:10.1101/gr.4252305.

48. DeGiorgio M, Lohmueller KE, Nielsen R. A model-based approach for identifying signatures of ancient balancing selection in genetic data. PLoS Genet. 2014;10(8):e1004561. doi:10.1371/journal.pgen.1004561.

49. Huber CD, DeGiorgio M, Hellmann I, Nielsen R. Detecting recent selective sweeps while controlling for mutation rate and background selection. Mol Ecol. 2016;25(1):142–156. doi:10.1111/mec.13351.

50. DeGiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R. SweepFinder2: increased sensitivity, robustness and flexibility. Bioinformatics (Oxford, England). 2016;32:1895–1897. doi:10.1093/bioinformatics/btw051.

51. Durrett R, Schweinsberg J. Approximating selective sweeps. Theoretical Population Biology. 2004;66(2):129–138.

52. Zeng K, Fu YX, Shi S, Wu CI. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 2006;174:1431–1439. doi:10.1534/genetics.106.061432.

53. Charlesworth D. Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. PLoS Genet. 2006;2(4):1–6. doi:10.1371/journal.pgen.0020064.

54. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD. Genetic evidence for archaic admixture in Africa. Proceedings of the National Academy of Sciences. 2011;108(37):15123–15128.

55. Xu D, Pavlidis P, Taskent RO, Alachiotis N, Flanagan C, DeGiorgio M, et al. Archaic Hominin Introgression in Africa Contributes to Functional Salivary MUC7 Genetic Variation. Molecular Biology and Evolution. 2017;34(10):2704–2715. doi:10.1093/molbev/msx206.

56. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science (New York, NY). 2010;328:710–722. doi:10.1126/science.1188021.

57. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505(7481):43.

58. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.

59. Kuderna LF, Tomlinson C, Hillier LW, Tran A, Fiddes I, Armstrong J, et al. A 3-way hybrid approach to generate a new high quality chimpanzee reference genome (Pan_tro_3. 0). GigaScience. 2017;.

60. Burbano HA, Hodges E, Green RE, Briggs AW, Krause J, Meyer M, et al. Targeted investigation of the Neandertal genome by array-based sequence capture. Science. 2010;328(5979):723–725.

61. Dumont BL, Payseur BA. Evolution of the genomic rate of recombination in mammals. Evolution; international journal of organic evolution. 2008;62:276–294. doi:10.1111/j.1558-5646.2007.00278.x.

62. Charlesworth B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nature reviews Genetics. 2009;10:195–205. doi:10.1038/nrg2526.

63. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351(6328):652–654.

64. Duchateau PN, Pullinger CR, Orellana RE, Kunitake ST, Naya-Vigne J, O'Connor PM, et al. Apolipoprotein L, a new human high density lipoprotein apolipoprotein expressed by the pancreas Identification, cloning, characterization, and plasma distribution of apolipoprotein L. Journal of Biological Chemistry. 1997;272(41):25576–25582.

65. Smith EE, Malik HS. The apolipoprotein L family of programmed cell death and immunity genes rapidly evolved in primates at discrete sites of host–pathogen interactions. Genome research. 2009;.

66. Mlitz V, Strasser B, Jaeger K, Hermann M, Ghannadan M, Buchberger M, et al. Trichohyalin-like proteins have evolutionarily conserved roles in the morphogenesis of skin appendages. Journal of Investigative Dermatology. 2014;134(11):2685–2692.

67. Lee SC, Wang M, McBride OW, O'Keefe EJ, Kim IG, Steinert PM. Human trichohyalin gene is clustered with the genes for other epidermal structural proteins and calcium-binding proteins at chromosomal locus 1q21. Journal of investigative dermatology. 1993;100(1):65–68.

68. Kypriotou M, Huber M, Hohl D. The human epidermal differentiation complex: cornified envelope precursors, S100 proteins and the 'fused genes' family. Experimental Dermatology. 2012;21(9):643–649. doi:10.1111/j.1600-0625.2012.01472.x.

69. Ewing G, Hermisson J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics. 2010;26(16):2064–2065. doi:10.1093/bioinformatics/btq322.

70. Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS computational biology. 2016;12(5):e1004842.

71. Peng B, Kimmel M. simuPOP: a forward-time population genetics simulation environment. Bioinformatics (Oxford, England). 2005;21:3686–3687. doi:10.1093/bioinformatics/bti584.

72. Staab PR, Metzler D. `coala`: an R framework for coalescent simulation. Bioinformatics (Oxford, England). 2016;32:1903–1904. doi:10.1093/bioinformatics/btw098.

73. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061.

74. Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, et al. Fast computation and applications of genome mappability. PloS one. 2012;7(1):e30377.

75. Uecker H, Setter D, Hermisson J. Adaptive gene introgression after secondary contact. Journal of mathematical biology. 2015;70:1523–1580. doi:10.1007/s00285-014-0802-y.

76. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100(16):9440–5.

77. Orr HA. The population genetics of speciation: the evolution of hybrid incompatibilities. Genetics. 1995;139:1805–1813.

78. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–496. doi:10.1038/nature10231.

79. Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nature genetics. 2017;49:303–309. doi:10.1038/ng.3748.

80. Sousa V, Hey J. Understanding the origin of species with genome-scale data: modelling gene flow. Nature reviews Genetics. 2013;14:404–414. doi:10.1038/nrg3446.

81. Charlesworth B. Measures of divergence between populations and the effect of forces that reduce variability. Mol Biol Evol. 1998;15(5):538–543.

82. Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Molecular biology and evolution. 2011;28:2239–2252. doi:10.1093/molbev/msr048.

83. Geneva AJ, Muirhead CA, Kingan SB, Garrigan D. A New Method to Scan Genomes for Introgression in a Secondary Contact Model. PLOS ONE. 2015;10(4):e0118621. doi:10.1371/journal.pone.0118621.

84. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5(10):e1000695. doi:10.1371/journal.pgen.1000695.

85. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS genetics. 2012;8:e1002967. doi:10.1371/journal.pgen.1002967.

86. Csilléry K, Blum MG, Gaggiotti OE, François O. Approximate Bayesian computation (ABC) in practice. Trends in ecology & evolution. 2010;25(7):410–418.

87. Slatkin M. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nature reviews Genetics. 2008;9:477–485. doi:10.1038/nrg2361.

88. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. Science (New York, NY). 2014;343:747–751. doi:10.1126/science.1243518.

89. Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, et al. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. Science (New York, NY). 2017;356:543–546. doi:10.1126/science.aal1988.

90. Patin E, Quintana-Murci L. The demographic and adaptive history of central African hunter-gatherers and farmers. Current opinion in genetics & development. 2018;53:90–97. doi:10.1016/j.gde.2018.07.008.

91. Kelly JK. A test of neutrality based on interlocus associations. Genetics. 1997;146(3):1197–1206.

92. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the Human genome from haplotype structure. Nature. 2002;419(6909):832–837.

93. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449:913–918. doi:10.1038/nature06250.

94. Smith J, Coop G, Stephens M, Novembre J. Estimating Time to the Common Ancestor for a Beneficial Allele. Molecular biology and evolution. 2018;35:1003–1017. doi:10.1093/molbev/msy006.

95. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. Science (New York, NY). 2006;312:1614–1620. doi:10.1126/science.1124309.

96. Knight SJL, Lese CM, Precht KS, Kuc J, Ning Y, Lucas S, et al. An Optimized Set of Human Telomere Clones for Studying Telomere Integrity and Architecture. The American Journal of Human Genetics. 2000;67(2):320 – 332.

97. Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. Genomic and genetic definition of a functional human centromere. Science. 2001;294(5540):109–115.

98. Kopp P. Human Genome and Diseases: Review The TSH receptor and its role in thyroid disease. Cellular and Molecular Life Sciences CMLS. 2001;58(9):1301–1322.

99. Abe E, Marians RC, Yu W, Wu XB, Ando T, Li Y, et al. TSH is a negative regulator of skeletal remodeling. Cell. 2003;115(2):151–162.

100. Novack DV. TSH, the bone suppressing hormone. Cell. 2003;115(2):129–130.

101. Slominski A, Pisarchik A, Wortsman J, Kohn L, Ain KB, Venkataraman GM, et al. Expression of Hypothalamic–Pituitary–Thyroid Axis Related Genes in the Human Skin. Journal of Investigative Dermatology. 2002;119(6):1449–1455.

102. Bodó E, Kromminga A, Bíró T, Borbíró I, Gáspár E, Zmijewski MA, et al. Human female hair follicles are a direct, nonclassical target for thyroid-stimulating hormone. Journal of Investigative Dermatology. 2009;129(5):1126–1139.

103. Vidali S, Knuever J, Lerchner J, Giesen M, Tamás B, Klinger M, et al. Hypothalamic-pituitary-thyroid axis hormones stimulate mitochondrial function and biogenesis in human hair follicles. The Journal of investigative dermatology. 2014;134:33–42. doi:10.1038/jid.2013.286.

104. Sun SC, Hsu PJ, Wu FJ, Li SH, Lu CH, Luo CW. Thyrostimulin, but not thyroid-stimulating hormone (TSH), acts as a paracrine regulator to activate the TSH receptor in mammalian ovary. Journal of Biological Chemistry. 2010;285(6):3758–3765.

105. Coutelier JP, Kehrl JH, Bellur SS, Kohn LD, Notkins AL, Prabhakar BS. Binding and functional effects of thyroid stimulating hormone on human immune cells. Journal of clinical immunology. 1990;10(4):204–210.

106. Sorisky A, Bell A, Gagnon A. TSH receptor in adipose cells. Hormone and Metabolic Research. 2000;32(11/12):468–474.

107. Martinez-deMena R, Anedda A, Cadenas S, Obregon MJ. TSH effects on thermogenesis in rat brown adipocytes. Molecular and cellular endocrinology. 2015;404:151–158.

108. Elgadi A, Zemack H, Marcus C, Norgren S. Tissue-specific knockout of TSHr in white adipose tissue increases adipocyte size and decreases TSH-induced lipolysis. Biochemical and biophysical research communications. 2010;393(3):526–530.

109. Draman MS, Stechman M, Scott-Coombes D, Dayan CM, Rees DA, Ludgate M, et al. The role of thyrotropin receptor activation in adipogenesis and modulation of fat phenotype. Frontiers in endocrinology. 2017;8:83.

110. Endo T, Kobayashi T. Thyroid-stimulating hormone receptor in brown adipose tissue is involved in the regulation of thermogenesis. American Journal of Physiology-Endocrinology and Metabolism. 2008;295(2):E514–E518.

111. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, et al. Sequence variants at CHRNB3–CHRNA6 and CYP2A6 affect smoking behavior. Nature genetics. 2010;42(5):448.

112. Hoft NR, Corley RP, McQueen MB, Schlaepfer IR, Huizinga D, Ehringer MA. Genetic association of the CHRNA6 and CHRNB3 genes with tobacco dependence in a nationally representative sample. Neuropsychopharmacology. 2009;34(3):698.

113. Cui W, Wang S, Yang J, Yi S, Yoon D, Kim Y, et al. Significant association of CHRNB3 variants with nicotine dependence in multiple ethnic populations. Molecular psychiatry. 2013;18(11):1149.

114. Culverhouse RC, Johnson EO, Breslau N, Hatsukami DK, Sadler B, Brooks AI, et al. Multiple distinct CHRNB 3–CHRNA 6 variants are genetic risk factors for nicotine dependence in African Americans and European Americans. Addiction. 2014;109(5):814–822.

115. Hoft NR, Corley RP, McQueen MB, Huizinga D, Menard S, Ehringer MA. SNPs in CHRNA6 and CHRNB3 are associated with alcohol consumption in a nationally representative sample. Genes, Brain and Behavior. 2009;8(6):631–637.

116. Haller G, Kapoor M, Budde J, Xuei X, Edenberg H, Nurnberger J, et al. Rare missense variants in CHRNB3 and CHRNA3 are associated with risk of alcohol and cocaine dependence. Hum mol genet. 2013;23(3):810–819.

117. Page NM, Olano-Martin E, Lanaway C, Turner R, Minihane AM. Polymorphisms in the Apolipoprotein L1 gene and their effects on blood lipid and glucose levels in middle age males. Genes & nutrition. 2006;1(2):133–135.

118. Pérez-Morga D, Vanhollebeke B, Paturiaux-Hanocq F, Nolan DP, Lins L, Homblé F, et al. Apolipoprotein LI promotes trypanosome lysis by forming pores in lysosomal membranes. Science. 2005;309(5733):469–472.

119. Lambrecht FL. Aspects of evolution and ecology of tsetse flies and trypanosomiasis in prehistoric African environment. The Journal of African History. 1964;5(1):1–24.

120. Franco JR, Simarro PP, Diarra A, Jannin JG. Epidemiology of human African trypanosomiasis. Clinical epidemiology. 2014;6:257.

121. Lecordier L, Vanhollebeke B, Poelvoorde P, Tebabi P, Paturiaux-Hanocq F, Andris F, et al. C-terminal mutants of apolipoprotein LI efficiently kill both *Trypanosoma brucei brucei* and *Trypanosoma brucei rhodesiense*. PLoS pathogens. 2009;5(12):e1000685.

122. Farrall M. Cardiovascular twist to the rapidly evolving apolipoprotein L1 story. Circulation research. 2014;114(5):746.

123. Genovese G, Tonna SJ, Knob AU, Appel GB, Katz A, Bernhardy AJ, et al. A risk allele for focal segmental glomerulosclerosis in African Americans is located within a region containing APOL1 and MYH9. Kidney international. 2010;78(7):698–704.

124. Rosset S, Tzur S, Behar DM, Wasser WG, Skorecki K. The population genetics of chronic kidney disease: insights from the MYH9–APOL1 locus. Nature Reviews Nephrology. 2011;7(6):313.

125. Rogers GE, Harding HW, Llewellyn-Smith IJ. The origin of citrulline-containing proteins in the hair follicle and the chemical nature of trichohyalin, an intracellular precursor. Biochimica et Biophysica Acta (BBA)-Protein Structure. 1977;495(1):159–175.

126. Rothnagel JA, Rogers GE. Trichohyalin, an intermediate filament-associated protein of the hair follicle. J Cell Biol. 1986;102(4):1419–1429.

127. Steinert PM, Parry DA, Marekov LN. Trichohyalin mechanically strengthens the hair follicle multiple cross-bridging roles in the inner root sheath. Journal of Biological Chemistry. 2003;278(42):41409–41419.

128. Westgate GE, Ginger RS, Green MR. The biology and genetics of curly hair. Experimental Dermatology. 2017;26(6):483–490. doi:10.1111/exd.13347.

129. Steinert P, Marekov L. Multiple roles for trichohyalin in the inner root sheath. Experimental dermatology. 1999;8(4):331.

130. Pośpiech E, Karłowska-Pik J, Marcińska M, Abidi S, Andersen JD, van den Berge M, et al. Evaluation of the predictive capacity of DNA variants associated with straight hair in Europeans. Forensic Science International: Genetics. 2015;19:280–288.

131. Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacón-Duque JC, et al. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. Nature communications. 2016;7:10815.

132. Huber M, Siegenthaler G, Mirancea N, Marenholz I, Nizetic D, Breitkreutz D, et al. Isolation and characterization of human repetin, a member of the fused gene family of the epidermal differentiation complex. Journal of investigative dermatology. 2005;124(5):998–1007.

133. Trzeciak M, Sakowicz-Burkiewicz M, Wesserling M, Gleń J, Dobaczewska D, Bandurski T, et al. Altered expression of genes encoding cornulin and repetin in atopic dermatitis. International archives of allergy and immunology. 2017;172(1):11–19.

134. Pośpiech E, Lee SD, Kukla-Bartoszek M, Karłowska-Pik J, Woźniak A, Boroń M, et al. Variation in the RPTN gene may facilitate straight hair formation in Europeans and East Asians. Journal of dermatological science. 2018;.

135. Stephan W, Wiehe TH, Lenz MW. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theoretical Population Biology. 1992;41(2):237–254.

136. Otto SP, Barton NH. The evolution of recombination: removing the limits to natural selection. Genetics. 1997;147(2):879–906.

137. Uecker H, Hermisson J. On the fixation process of a beneficial mutation in a variable environment. Genetics. 2011;188:915–930.

138. Ewens WJ. Mathematical population genetics. vol. 9 of Biomathematics. Berlin, Heidelberg, New York: Springer-Verlag; 1979.

139. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123(3):585–595.