Understanding the origin of species with genome-scale data: modelling gene flow

Vitor Sousa and Jody Hey

Abstract | As it becomes easier to sequence multiple genomes from closely related species, evolutionary biologists working on speciation are struggling to get the most out of very large population genomic data sets. Such data hold the potential to resolve long-standing questions in evolutionary biology about the role of gene exchange in species formation. In principle, the new population genomic data can be used to disentangle the conflicting roles of natural selection and gene flow during the divergence process. However, there are great challenges in taking full advantage of such data, especially with regard to including recombination in genetic models of the divergence process. Current data, models, methods and the potential pitfalls in using them will be considered here.

Single-nucleotide polymorphisms
(SNPs). Sites in the DNA in which there is variation across the genomes in a population, usually comprising two alleles that correspond to two different nucleotides.

Ascertainment bias
Systematic bias introduced by the sampling design (for example, criteria used to select individuals and/or genetic markers) that induces a nonrandom sample of observations.

Department of Genetics, Rutgers, the State University of New Jersey, Piscataway, New Jersey 08854, USA. Correspondence to J.H.
e-mail: hey@biology.rutgers.edu
doi:10.1038/nrg3446
Published online 9 May 2013

One of the many doors that open when a species' genome is first sequenced is to the world of population genomics and to the unparalleled study of the evolutionary divergence of closely related populations and species. Ever since Darwin developed his model of how one species splits into two (his principle of divergence)[1], the field of evolutionary biology has been divided on the question of whether Darwin's model is correct. With growing access to very large amounts of genome data from multiple individuals of a species, it becomes increasingly likely that we will resolve this long-standing question.

Recent next-generation sequencing (NGS) technologies and assembly tools, such as restriction-site-associated DNA (RAD-tag)[2] sequencing and genotyping by sequencing (GBS)[3], now make it possible to obtain genome-scale data affordably from multiple individuals (reviewed in REFS 4,5). When individuals are sampled from multiple populations of a species, as has been done for humans[6–8], stickleback fish[2], fruitflies (*Drosophila* Population Genomics Project), *Arabidopsis thaliana* (1001 Genome Project)[9], dogs[10] and different species of great apes[11–13], among others[14,15], we gain an exceptional view not only on the variation within populations and species but also on the variation that lies between them. Data sets such as these can include millions of variable single-nucleotide polymorphisms (SNPs) and other kinds of polymorphisms and hold the promise of finally revealing the genetic side of how species and populations diverge.

However, a flood of new data may not lead directly to a commensurate gain in knowledge, and today, as new population genomic data sets are emerging, our skills of analysis and interpretation are partly overwhelmed. With the rise of large NGS data sets that reveal complex patterns of variation across species' genomes, we find that our best models and tools for explaining patterns of variation were designed for a simpler time and smaller data sets. In the first place, NGS data sets present unique challenges, apart from their size, that result from the way in which they are generated. For example, it is common to use a reference genome to aid the assembly of additional NGS data, and yet this introduces a form of ascertainment bias, or reference bias, that can affect one's findings[16,17] (BOX 1). Second, most models and methods available to analyse NGS data have limitations that prevent using all of the information in the data that bears on the processes of interest.

In this Review, we survey the state of the art of population divergence models and inference methods, with regard to population genomics data sets. We do not examine in detail the technical challenges that are related to NGS for correcting sequencing, assembly, SNP and genotype-calling errors, as these have recently been reviewed elsewhere[4,17,18]. Rather, we focus on models of population divergence and on methods to detect and to quantify gene flow, as well as methods to distinguish alternative modes of speciation. We discuss the limitations of these methods and provide examples of their application to recently available genome-wide data sets.

Models of species formation

Speciation in the absence of gene flow. When two populations become allopatric (that is, completely geographically separated), they can diverge without mixing genes and eventually become reproductively isolated[19]. Compared with divergence in the presence of gene exchange, this is a comparatively simple process, and this simplicity, together with clear biogeographic evidence (such as the finding that in island archipelagos it is common to find different species on different islands), convinced many that this was how nearly all new species formed[19–24].

Speciation in the presence of gene flow. Darwin, however, envisioned that natural selection can act in disparate ways over a species range to pull a species in different directions and eventually to split it, first into different varieties and finally into separate species. This model, which has come to be called 'sympatric speciation', was considered by many to be unlikely as gene exchange across a species range was considered to be a strong homogenizing force that counteracts divergence by natural selection. However, more recently, genetic data (for example, mitochondrial DNA sequences and microsatellites) together with biogeographic circumstances have provided compelling evidence that sympatric speciation has occurred in numerous contexts[25–27].

At the genetic level, Darwin's model raises complexities, as it predicts that divergence in the presence of gene flow can cause different genes to experience very different histories. Diversifying selection favours different alleles in different parts of a species range (and at one or more loci). However, the movement of all genes across the range of the species as the normal result of organisms reproducing and dispersing will regularly move alleles that are affected by the diversifying selection into the 'wrong' part of the species range. It will also cause the species to appear to be homogeneous when examined for patterns of variation at neutral genes.

An additional major player in the divergence process, particularly when gene flow is occurring, is recombination, which is the breaking and rejoining of chromosomes that happens during meiosis every generation and that allows different parts of the genome to have different histories. Because of recombination, an allele that is favoured by selection and that is increasing in frequency will carry with it in its trajectory towards fixation only those flanking regions to which it is most tightly linked[28,29]. Recombination also makes it possible for alleles at neutral loci to move by gene flow across the species range and to co-occur in the same population of genomes in which there are loci diverging by the action of diversifying selection[30,31]. Recombination thus allows a species to have a population of genomes with a split personality: to resemble two diverging gene pools at loci affected by diversifying selection and to resemble a single gene pool at loci that are not under selection in this way. Evidence of this kind of genomic schism has come from a diversity of systems in recent years, on the basis of DNA Sanger sequence and microsatellite data[32] and more recently from NGS data in stickleback fish[2], *Heliconius* butterflies[14] and flycatchers[15].

Modelling population divergence. A widely used theoretical framework for studying speciation using genetic data is the 'isolation with migration' model, so named because it includes both the separation of two populations (a process called isolation) following a splitting

Box 1 | **Challenges of NGS data for population genomics**

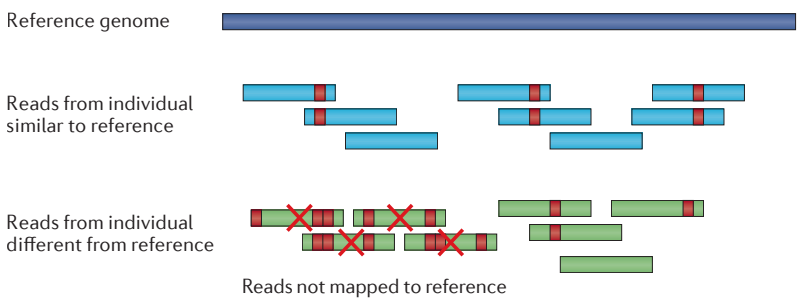
An ideal sequencing technology will output high-quality reads of lengths greater than the size of duplicated regions in the target genome. The current crop of technologies, however, generates sequences of short lengths (generally less than 500 bp and often less than 100 bp)[4,5]. Even with many reads and with paired-end libraries, genome assembly generally requires the aid of a reference genome. One of the main difficulties with this protocol, which is especially relevant for demographic inference, is genotype and single-nucleotide polymorphism (SNP) call uncertainty (reviewed in REFS 17,18), reference bias and phase uncertainty. Other challenges with next-generation sequencing (NGS) have recently been reviewed elsewhere[4,5,17].

Reference genome bias

Sequence assembly from short sequences using a reference genome results in ascertainment bias[116,117]. That is, NGS reads that are different from their homologous location in the reference genome at polymorphic sites will tend to be misassembled, and there will be a tendency to underestimate differences between the new data and the reference genome, as shown in the figure, in which reads from two individuals are differentially mapped to a reference genome. Reads from the individual shown in blue are preferentially mapped owing to their similarity with the reference genome, whereas reads from the other individual (shown in green) are not mapped as they contain more differences[17]. Although this bias can be mitigated by using high-coverage NGS data[118] or by doing *de novo* genome assemblies if possible, the bias is not easily dismissed in low-coverage contexts. For example, relaxation of criteria for matching NGS reads to a reference genome will help to avoid missing heterozygous positions but entails a concomitant increase in 'false positives'. Recently, investigators have begun to address this bias: for example, by including models of the process generating the variation[16] or by adding steps to the assembly pipeline that help to compensate for the bias in local regions of the genome[119].

Phase uncertainty

When sequencing a diploid individual, two genomes are sequenced, and it is difficult to know whether any pair of reads came from the same or from different genomes. This issue does not affect the identification of heterozygous positions, but this phase uncertainty greatly hinders the assembly of two genomes from one diploid sample, and it considerably complicates the assessment of linkage disequilibrium (LD) over longer distances. This can have an impact on demographic estimates, especially for methods based on haplotypes and that use LD information (as seen in simulation studies for hidden Markov model (HMM)-based[112] and identity by descent (IBD)-based[115] methods). One possible approach to deal with phase uncertainty consists of integrating over all possible phases, as implemented in a genealogy sampler approach[7]. The use of pair-ended reads can extend the lengths of regions over which two separate sequences can be resolved, beyond the length of the actual reads[120]. If data are available from an individual as well as from both of their parents (a so-called 'trio'), then it is possible to infer both chromosomes of the individual[6]. Alternatively, a population genetic statistical approach can be used to estimate phase when there are data from multiple individuals[121]. In the future, technology developments that allow for long reads from each chromosome are likely to reduce the problem of phase uncertainty[122].



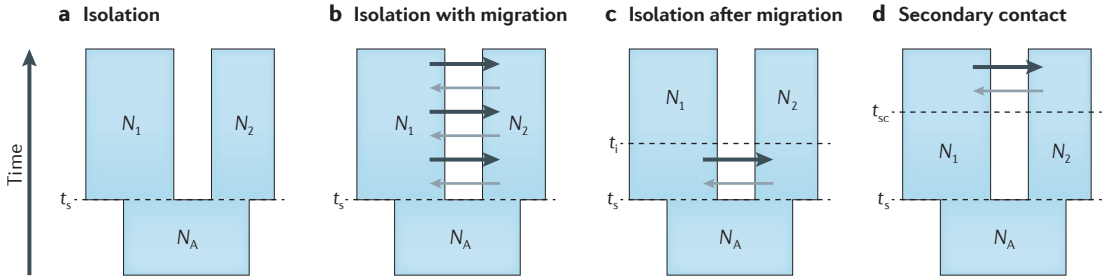


Figure 1 | **Alternative modes of divergence.** All models assume that an ancestral population of size N_A splits into two populations at time of split (t_s). The two present-day populations have effective sizes N_1 and N_2, respectively. Panel **a** shows the model in which migration rate is zero in both directions, which corresponds to an allopatric divergence scenario. Panels **b**–**d** represent alternative models in which populations have been exchanging migrants. Gene flow occurs at constant rates since the split from the ancestral population (**b**). Migration rates are assumed to be constant through time, but gene flow can be asymmetric: that is, one migration rate for each direction. Panel **c** shows a scenario in which populations begin diverging in the presence of gene flow but experience a cessation of gene flow after time since isolation (t_i). If the lack of current gene flow in this model is due to reproductive isolation then this represents a history in which divergence occurred to the point of speciation in the presence of gene flow. In panel **d**, we consider the alternative migration history in which populations were isolated and diverged for a period of time in the absence of gene flow, followed by secondary contact at time of secondary contact (t_{sc}) and the introgression of alleles from the other population by gene flow.

Paired-end libraries
Sequencing from each end of the fragments in a library. The two sequenced ends are typically separated by a gap.

Sympatric speciation
The process of divergence between populations or species occupying the same geographical area and in presence of gene flow.

Diversifying selection
Natural selection acting towards different alleles (or phenotypes) being favoured in different regions within a single population or among multiple connected populations.

Neutral genes
Genes for which genetic patterns are mostly affected by mutation and demographic factors, such as genetic drift and migration.

Allopatric divergence
The process of divergence between populations or species that are geographically separated, in the absence of gene flow.

Linkage disequilibrium
(LD). The nonrandom association of alleles at different sites or loci.

Islands of differentiation
Genomic regions of elevated differentiation owing to the action of natural selection.

F_{ST}
The proportion of the total genetic variability occurring among populations, typically used as a measure of the level of population genetic differentiation.

Island model
A model introduced by Sewall Wright to study population structure comprising multiple populations connected to each other through migration.

Metapopulation model
In the context of F_{ST}-based statistics, this is an idealized model in which several populations diverge without migration from a common ancestral gene pool (or metapopulation).

event from their common ancestral population as well as migration between populations[33–35]. At one extreme, we can consider a simple isolation model in which the migration rate is zero in both directions; this corresponds to an allopatric divergence scenario (FIG. 1a). Other models include isolation with migration (FIG. 1b), isolation after migration (FIG. 1c) and secondary contact (FIG. 1d). It has been shown that patterns of genetic variation in samples from two closely related populations or species can be used to distinguish a pure isolation model (FIG. 1a) from a model with migration[33,34] (FIG. 1b–d). Furthermore, the growing evidence of persistent gene exchange between closely related species means that divergence often arises in the midst of conflicting evolutionary processes[32].

Inferring the history of divergence

NGS data from multiple individuals offer the promise of disentangling the complex interplay between selection, gene flow and recombination that occurs during speciation with gene flow. First, by having information for essentially all parts of the genome, we can gain a more detailed and accurate picture of the demography of populations[36,37]. Second, it becomes possible to ask whether some parts of the genome have been exchanging genes more than others. Substantial variation in gene flow levels across the genome constitutes clear, albeit indirect, evidence that selection is acting against gene flow to a greater degree in some genome regions than in others[38,39]. Third, NGS data allow us to get better estimates of recombination rates and linkage disequilibrium (LD) patterns along the genome[40,41], and this can in principle be used to infer the timing and magnitude of gene flow. Finally, polymorphism and LD along the genome also bear information about selective sweeps and genes that are the targets of diversifying selection (reviewed in REFS 42,43). However, all of these inferences depend on having a theoretical framework that connects patterns of variation to an explicit model.

Genome scans using indicators of divergence. Depending on an investigator's question, it can sometimes be useful to take a fairly simple approach that does not use models with lots of parameters to study the levels of divergence between populations. This can be achieved by tailoring analyses to a specific component of the divergence process and scanning across the genome while calculating statistics that are expected to be sensitive to that feature. For example, there has been lots of interest in detecting 'islands of differentiation' by looking at the distribution of summary statistics that measure genetic differentiation, such as F_{ST}[44]. In the first study using RAD-tag sequencing, the differentiation of 45,789 SNPs along the genome between oceanic and freshwater populations of threespine sticklebacks (*Gasterosteus aculeatus*) showed, overall, reduced levels of differentiation (F_{ST} values close to zero)[2]. However, when a sliding window was run along the aligned genomes of freshwater and oceanic populations, the authors found evidence for genomic regions characterized by very high F_{ST} values (>0.35), potentially harbouring genes under divergent selection. Interestingly, the same genomic regions were highlighted in contrast to different freshwater populations, suggesting parallel adaptation to the freshwater environment. These results are in agreement with a larger study comprising seven pairs of closely related marine and freshwater populations comprising 5,897,368 SNPs[45].

Divergence summaries can also be used within demographic models of divergence, such as an island model or a metapopulation model (reviewed in REFS 46,47). One approach is to scan the genome using a hierarchical F_{ST} model that assumes a nested island model underlying the divergence process[48]. A related approach explicitly accounts for variation in read depth in NGS data within a Bayesian framework[49] and hence should be preferred for analysing such data.

Another type of genome scan that is targeted to identify recent admixture relies on comparing the

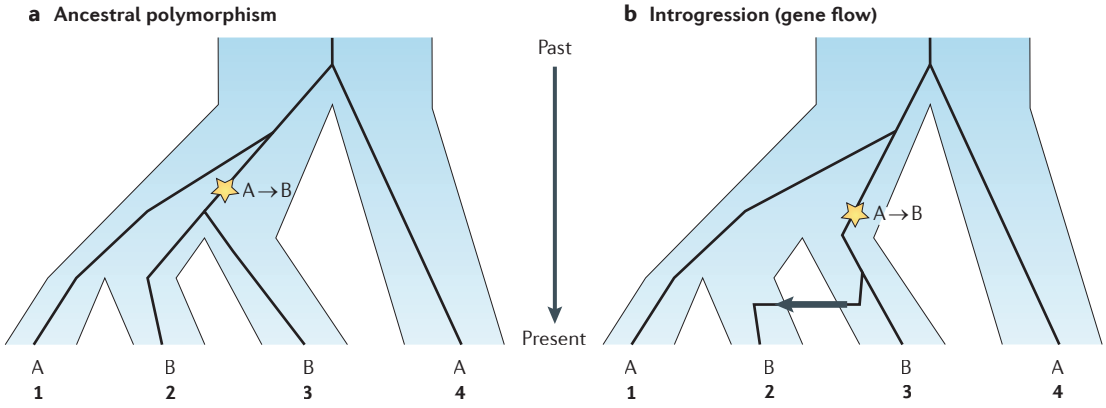


Figure 2 | **Disentangling ancestral polymorphism from gene flow (ABBA and BABA test).** The diagram shows the divergence of two sister populations (1 and 2), a third population (potential source of introgressed genes; 3) and an outgroup population (4) over time. The black line represents the gene tree of a given site, and the star represents a mutation from the ancestral state (allele A) to the derived state (allele B). The pattern ABBA can occur owing to an ancestral polymorphism (**a**): that is, coalescent of lineage from population 2 with lineage from population 3 in the ancestral population (population ancestral to populations 1, 2 and 3), or gene flow from population 3 to population 2 (**b**). Under a model with no gene flow, we expect that the pattern ABBA is as frequent as BABA owing to the fact that there is 50% chance that either the lineage from population 1 or from population 2 coalesces with lineage from population 3 in the population ancestral to populations 1, 2 and 3.

population tree (assumed to be known) with the gene trees inferred at a specific site. Incongruences between the population tree and the gene tree can be due to incomplete lineage sorting (shared ancestral polymorphism) or to gene flow. One statistic, called '*D*', was specifically designed to detect introgression from one population to another[50] (FIG. 2). Computing *D* requires a genome from each of two sister populations, a genome from a third population (a potential source of introgressed genes) and a fourth outgroup genome to identify the ancestral state (identified as the A allele). Focusing on SNPs in which the candidate source population has the derived allele (B) and in which the two sister genomes have different alleles, there are two possible configurations: either ABBA or BABA. Under the hypothesis of shared ancestral polymorphism, the number of tree topologies of ABBA and BABA are expected to be equal, and the expected *D* will be zero. Deviations from that expectation are interpreted as evidence of introgression. As with F_{ST} genome scans, investigators can look at the distribution of *D* along the genome, but when using *D*, the aim is to find genomic regions that specifically experienced introgression, whereas in the case of F_{ST}, the goal is to identify regions of high differentiation, regardless of the cause.

Genome scans using *D* were used, for instance, to detect admixture between archaic and modern humans[51,52] and to study the patterns of introgression in *Heliconius* butterflies[14]. In the case of modern and archaic humans, unidirectional introgression from Neanderthals to non-African humans was estimated to have occurred for 1–4% of the genome[51]. Similarly, data from 642,690 SNPs point to 4–6% of the present day Melanesian genomes being derived from admixture with Denisovans[52]. For *Heliconius* butterflies, RAD-tag sequencing of 4% of the genome (~12 Mb) indicated introgression from *Heliconius timareta* to *Heliconius melpomene amaryllis* (2–5% admixture), which are sympatric species that exhibit the same wing colour patterns. Interestingly, only a few regions exhibited significant *D* values, including genes known to contain the mimicry loci B/D and N/Yb. Despite the lack of an explicit test of positive selection, the fact that these regions harbour genes involved in mimicry is in agreement with an active role of selection promoting introgression at these regions. In these species, the patterns of differentiation along the genome suggest a case in which most of the genome is differentiated — consistent with a model of allopatric divergence (FIG. 1a) or divergence with limited gene flow (FIG. 1b) — whereas a few regions show evidence of secondary contact and uni- or bidirectional introgression of genes from one population (species) to the other (FIG. 1c). In both cases of humans and *Heliconius* spp., there was evidence of regions exchanged between populations that were already differentiated, pointing to the importance of secondary contact.

Although genome scan approaches are flexible and applicable to large genomic data sets, the focus on amenable summary statistics typically entails setting aside much of the information in a data set. A related limitation is that the same numerical value of a particular statistic can result from very distinct scenarios. For instance, a low F_{ST} can be due to shared ancestral polymorphism or due to gene flow[44]. Similarly, the *D* statistic can be significantly different from zero owing to other events rather than admixture. The evidence of admixture between modern human non-African populations and Neanderthals has been questioned by a simulation study showing that spatial expansions and population substructure without admixture could lead to *D* values that are similar to the observed ones[53].

Nested island model
A hierarchical island model with groups of populations in which migration among populations within the same group is higher than among populations in different groups.

Gene trees
Bifurcating trees that represent the ancestral relationships of homologous haplotypes sampled from a single or multiple populations. A gene tree includes coalescent events and, in models with gene flow, migration events. A gene tree is characterized by a topology, branch lengths, coalescence times and migration times.

Box 2 | **Contrasting the allele frequency spectrum with genealogy-sampling approaches**

Allele frequency spectrum

In two populations, the allele frequency spectrum (AFS) corresponds to a multidimensional matrix *X*, where each $x_{i,j}$ entry gives the number of single-nucleotide polymorphisms (SNPs) with an observed derived allele count of *i* in population 1 and *j* in population 2. The likelihood is easily computed, given the expected AFS under a given evolutionary model. Each entry in the expected AFS reflects the probability of a given SNP falling into that cell. Assuming that all SNPs are independent (that is, assuming free recombination between SNPs), these probabilities can be derived from the distribution of allele frequencies across populations, which in turn can be found with diffusion approximations to the evolutionary processes or with the coalescent. After the expected AFS is obtained under a given model, it is easy to compute the likelihood for an arbitrarily large number of SNPs, making this a method applicable to the analysis of genomic data.

Genealogy sampling

Coalescence-based models aim at extracting information about relevant selective and demographic events from gene trees relating homologous DNA sequences (haplotypes) sampled from multiple populations. Each locus may contain several SNPs, and hence haplotype data contain an extra layer of information when compared with AFS approaches. Most methods assume no recombination within each locus and free recombination among loci. Coalescence-based methods are usually based on samplers that collect genealogies from the posterior distribution. However, exploring the genealogical space can be extremely complex and relies on highly computationally intensive Monte Carlo algorithms, such as Markov chain Monte Carlo (MCMC), that do not easily extend to large genomic multi-locus data sets.

	AFS	Coalescence-based
Type of data	SNPs (biallelic markers)	Phased DNA segment (haplotype)
Assumptions about recombination	Free recombination among SNPs (all SNPs independent)	Free recombination among loci and complete linkage within loci
Assumptions about mutation	Mutation rates equal for all SNPs	Mutation rates vary across loci
Likelihood	Diffusion-based or coalescence-based	Coalescence-based
Methods	Composite-likelihoods; fairly fast and able to deal with millions of SNPs	Monte Carlo methods based on genealogy samplers (such as MCMC or importance sampling) or based on approximate methods (such as ABC or PAC); usually slow and computationally intensive, compromising their application to large genomic data sets

Likelihood and model-based methods. As useful as genome scans with indicator variables can be to identify components of the divergence process, they fall short of providing a full portrait of divergence unless they are combined with other analyses. In this light, the goal for many investigators is to be able to calculate the likelihood under a rich divergence model. For some model of divergence M, with a parameter set Θ, the likelihood is the probability (P) of the data given the parameters: that is, $P_M(Data \mid \Theta)$. Having a likelihood function at hand allows estimating the most likely parameters of a given model either with frequentist or Bayesian statistics[54]. Also, comparing the likelihood of alternative models opens the door to model choice approaches to infer the most probable divergence model. Currently, there are two main families of likelihood-based approaches for studying divergence: one based on the allele frequency spectrum (AFS) and a second based on sampling genealogies for short portions of the genome (BOX 2).

Likelihoods using the allele frequency spectrum. For a single SNP sampled in each of two populations, considered together with the base that is present in an outgroup genome, the data can be summarized as the number of copies of the derived allele in each of the two populations. For a large number of SNPs, these counts fill a discrete distribution — the allele frequency spectrum (AFS) — in two dimensions (one for each sampled population), which can be represented in graphical form (FIG. 3). This approach has seen renewed interest as large SNP data sets have become more common[55–57]. FIGURE 3 shows how the AFS can vary considerably for the different isolation with migration models shown in FIG. 1, particularly how simple isolation differs from models with gene flow. In the absence of gene flow (FIG. 3a), the frequencies of SNPs found in only one population are different from the SNPs in the other populations because genetic drift drives different alleles to fixation in each population. By contrast, in models with gene flow, the cells along the diagonal exhibit a higher density (FIG. 3b) because there are many SNPs with similar frequencies in the two populations. However, as exemplified in these AFSs, it can be difficult to separate alternative scenarios with gene flow, as these tend to be similar (FIG. 3b–d).

Although the expected AFS can be generated by simulations[55,58,59], it is also the focus of a population genetic theory in which differential equations describe the diffusion of allele frequencies in populations[60,61]. In recent years, the diffusion equation approach has been reawakened for the study of the AFS under isolation with migration models, such as the ones shown in FIG. 1 (REFS 57,62,63). If it is assumed that the SNPs segregate

Bayesian statistics
Statistical framework in which the parameters of the models are treated as random variables, allowing expression of the probability of parameters, given the data; this is called the posterior. The posterior probability is obtained by Bayes' rule, and it is proportional to the likelihood times the prior.

Allele frequency spectrum
(AFS). A distribution of the counts of single-nucleotide polymorphisms with a given observed frequency in a single or multiple populations.

Genetic drift
Stochastic changes in gene frequency owing to finite size of populations, resulting from the random sampling of gametes from the parents at each generation.

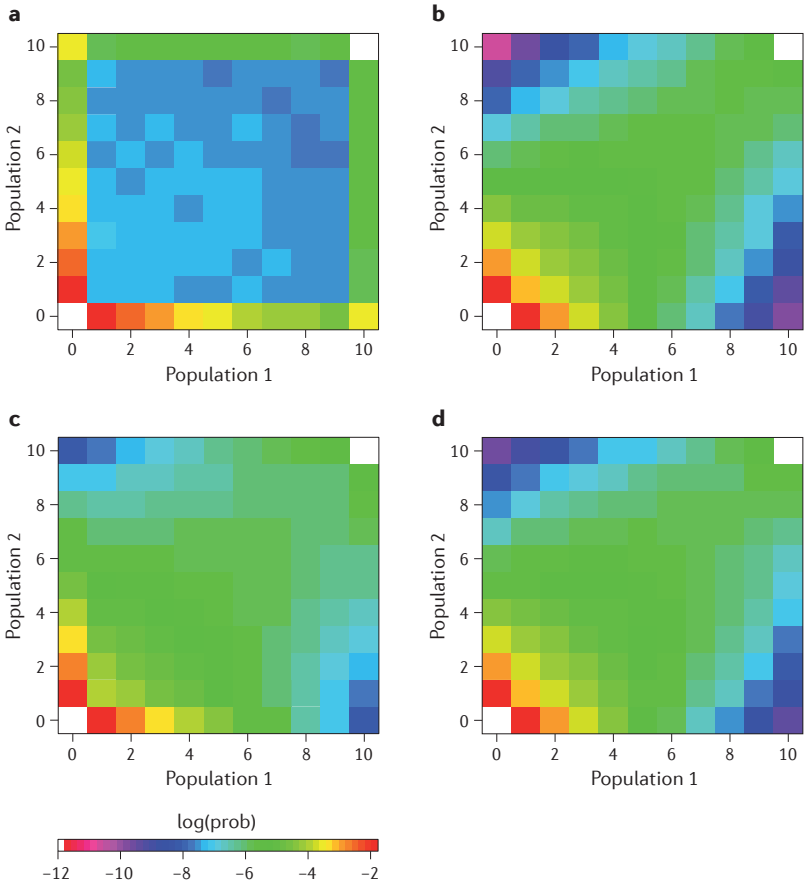


Figure 3 | **Allele frequency spectrum under alternative divergence models.** Each entry in the matrix (*x*,*y*) corresponds to the probability of observing a single-nucleotide polymorphism (SNP) with frequency of derived allele *x* in population 1 and *y* in population 2. The colours represent the log of the expected probability for each cell of the allele frequency spectrum (AFS). The white colour corresponds to –Inf: that is, to cells with an expected probability of zero. These AFSs are conditional on polymorphic SNPs, hence the cells (0,0) and (10,10) have zero probability. The likelihood for an observed AFS can be computed by comparing it with these expected AFSs. **a** | Isolation model. **b** | Isolation with migration. **c** | Isolation after migration. **d** | Secondary contact. The joint allele frequency spectrums for the different scenarios were obtained with coalescent simulations carried out with ms[125]. All scenarios were simulated, assuming all populations share the same effective sizes ($N = 10{,}000$), a time of split $t_s = 20{,}000$ generations ago ($t / 4N = 0.5$), symmetrical migration rate ($2N_1m_{12} = 5$, $2N_2m_{21} = 5$, for scenarios **b**, **c** and **d**, for scenario **c**, a time of isolation of $t_i = 2{,}000$ generations ago ($t_i / 4N = 0.05$) and, for scenario **d**, a time of secondary contact of $t_{sc} = 6{,}000$ generations ago ($t_{sc} / 4N = 0.15$).

Coalescent theory
A theory that describes the distribution of gene trees (and ancestral recombination graphs) under a given demographic model that can be used to compute the probability of a given gene tree.

independently, then given both an observed and an expected AFS for a model of interest, the likelihood can be directly calculated using a multinomial distribution. One difficulty is that in reality most data sets include many SNPs that are sufficiently close to one another that the assumption of independence does not apply. Still, the same likelihood calculation can be applied (now identified as a 'composite likelihood'[57]) without introducing bias to the parameter estimates, albeit with limited access to confidence intervals and other analyses for which a likelihood is often used[64]. By reducing the data to counts of SNP frequencies, AFS methods are also guilty of discarding all linkage information in the data. This means that these methods are not expected to be very sensitive to processes that can affect local LD patterns, such as gene flow or admixture. AFS-based analyses on population genomic data sets have so far mostly been conducted on human data[57,63,65], but the same approach can be used to study the divergence of closely related species. A nice example is the study of the divergence of Sumatran orangutans (*Pongo abelii*) and Bornean orangutans (*Pongo pygmaeus*)[13]. Low-coverage (8×) Illumina sequencing of 5 individuals from each species yielded a total of 12.74 million SNPs, and an AFS analysis led to an estimated speciation time of 400,000 years with a low level of gene exchange between the species[13].

The AFS approach has also been applied to more complex models with more than two populations or species. One example comes from the analysis of human data from the 1000 Genomes Project under a three-population isolation with migration model with gene flow and population expansions[65]. By considering only SNPs at synonymous sites and by explicitly modelling genotype calling errors, these authors estimated a time for expansion out of Africa around 51,000 years ago, a split between Europeans and East Asians around 23,000 years ago, recent population expansion in both Europeans and East Asians and statistically significant but reduced gene flow among all populations.

However, AFS-based methods become computationally challenging and expensive for models with more than three populations. There is thus considerable interest in finding suitable approximations to the diffusion process that do not rely on a full multidimensional AFS[66–68]. Recently, some new methods have appeared that implement simplified diffusion processes that do not include mutation models but that do account for divergence from common ancestry by genetic drift[66–68]. The lack of a mutational component means that these methods are intended for cases of recent divergence among populations. By modelling the branch lengths of the population and species tree as proportional to drift and by treating drift in different branches as independent (with no gene flow), it is possible to write down a likelihood function, opening the door to infer the population and species trees. It is also possible to include admixture within this framework by allowing for one population to have ancestry in multiple populations[66]. For instance, 60,000 SNPs from 82 dog breeds and wild canids (obtained with SNP arrays) supported a population tree with admixture events (as in FIG. 1c) rather than a pure isolation model[66] (as in FIG. 1a).

Likelihoods by sampling genealogies. If the recombination rate is low, such that it is unlikely to have occurred in the time since the common ancestor of a sample of sequences from one or various populations, as can be the case over a short region of the genome, the history of a sample of sequences can be described by a gene tree or genealogy (BOX 3). The depth and structure of such genealogies have been described by coalescent theory for a diversity of models, including the models shown in FIG. 1 (REFS 69–71), and this coalescent modelling has made it possible to calculate the likelihood for data

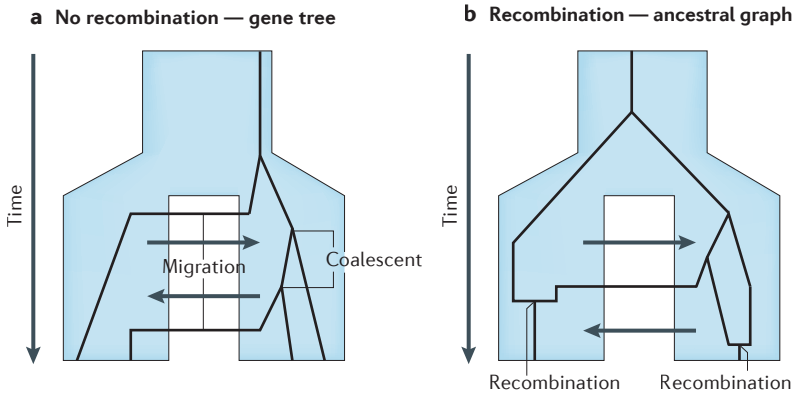
sets with multiple sequences of a short genomic region sampled from one or more populations. Rather than focusing on the best gene tree (as is often the case in phylogenetics), the likelihood is obtained by integrating over all possible genealogies[72]. Because this integration cannot be solved analytically except for small sample sizes, these approaches rely on computationally intensive methods[73,74]. The general principle of these methods is to sample a set of genealogies that is consistent with the data[74,75], which may in turn be used to obtain a posterior probability in a Bayesian approach[54]. These methods have seen tremendous advances in recent decades, making it possible to estimate effective population sizes, migration rates, admixture contributions and dates of population declines, among other parameters[37,73]. Moreover, through likelihood ratio tests[75] or through marginal likelihoods[76], it has become possible to assess the fit of alternative models of divergence.

A genealogy-sampling approach to the likelihood can easily be extended to multiple loci if each has not undergone recombination and if free recombination is assumed between loci. Under these assumptions, the overall likelihood is the product of that for each locus (BOX 3). Thus, in principle, a genealogy-sampling approach can be extended to a genome scale if the computational power is available to handle many thousands of genome segments. For small sample sizes, it is possible to obtain analytical solutions, and this has been found for a two-population isolation with migration model with constant gene flow for the special case of two sampled genomes[77]. When applied to the divergence between *Drosophila melanogaster* and *Drosophila simulans*, for a data set with 30,323 genomic segments (average length of 405 bp), a divergence time of 3.04 million years ago, and a non-zero migration rate from *D. simulans* to *D. melanogaster* was inferred[77]. An alternative approach to computing the likelihood for larger sample sizes consists of using generating functions, which so far has been shown to be possible for up to samples of three gene copies[78,79].

The largest data set analysed so far using a genealogy-sampling approach consists of six genomes (each divided into 37,574 segments of 1 kb in size): one from each of six human populations. The data were examined assuming an isolation model with five populations and migration between one single pair of populations[7]. Statistically significant migration was estimated between populations in Africa: namely, between San and Bantu, and San and Yoruba. Surprisingly, the estimates for the times of split pointed to a very old divergence between these African populations (108–157,000 years ago), suggesting an ancient and complex population structure in that continent.

Related approximate likelihood methods. One general family of methods (so-called 'likelihood-free' methods) sidesteps the actual calculation of likelihoods by using direct simulations under the model of interest. These include approximate Bayesian computation (ABC) methods, which have recently been reviewed[80–83]. One advantage of ABC methods is that it can be fairly straightforward to include recombination in the models[84]. However, the application of ABC to genome-wide data sets is still in its infancy mostly owing to the prohibitively high computational cost of simulating population genomic data, but in practice it has been shown to handle data sets with hundreds[85] to a few thousand loci[86].

Box 3 | **Challenges of computing likelihoods with recombination**



If there is free recombination among sites, the likelihood is simply the product of the likelihoods for each site. This is the assumption underlying the allele frequency spectrum (AFS)-based methods. At the other extreme, when all sites are fully linked, the ancestry of a sample is fully captured by a gene tree shared by all sites. This is the realm of coalescence-based methods and of most genealogy-sampling approaches. However, the reality lies in between these two extremes, and it is exactly for intermediate levels of recombination, when two portions of the genome are neither completely linked nor completely unlinked, that the calculation of the likelihood becomes very difficult.

In a genealogy-sampling method, likelihoods are computed by integrating over the genealogy space. Under a model characterized by a set of parameters, given data from L loci, $X=(X_1,\ldots,X_L)$ and its underlying gene trees, $G=(G_1,\ldots,G_L)$, where X_i and G_i represent the data and gene trees of the i^{th} locus ($i=1,\ldots,L$), respectively, the likelihood is found as a product over loci:

$$f(X|\Theta)=\prod_{i=1}^{L}\int f(X_i|G_i)f(G_i|\Theta)\mathrm{d}G_i$$

where $f(G_i|\Theta)$ is the probability of the genealogy given the parameters Θ, and $f(X_i|G_i)$ is the probability of the data at the i^{th} locus, given its genealogy. The ancestral relationships between sequences are described by a gene tree with coalescent and migration events (see part **a** of the figure). Recombination causes different parts of the genome to have different genealogical histories, and so the ancestry of a set of sequences is best pictured as a graph known as the ancestral recombination graph (ARG) with joining events (coalescent events) and the splitting of gene copies into two parental copies (recombination events; see part **b** of the figure)[70,101,123,124]. Each recombination event corresponds to a split of the sequence into two subsequences that carry different ancestral segments. Interestingly, given the ARG, denoted A, we can look at the marginal gene trees for each site along the sequence and compute the likelihood as a product over those marginal genealogies G_i as

$$f(X|\Theta)=\int f(A|\Theta)\prod_{i=1}^{S}f(X|G_i(A))\mathrm{d}A$$

where S is the number of sites, and $f(X|G_i(A))$ is the probability of the data at the i^{th} site, given its marginal genealogy G_i implied by the ARG. Note that in this case, the parameters (Θ) include the coalescent, migration and recombination processes. The marginal distribution of genealogies can be obtained given the ARG, but the ARG cannot be obtained given the marginal genealogies. This is at the core of the difficulties of dealing with recombination. First, in comparison with gene trees, the ARG is dramatically more complex, making the search through the ARG space intractable for population divergence models. Second, data typically contain diffuse information about which ARGs are more likely.

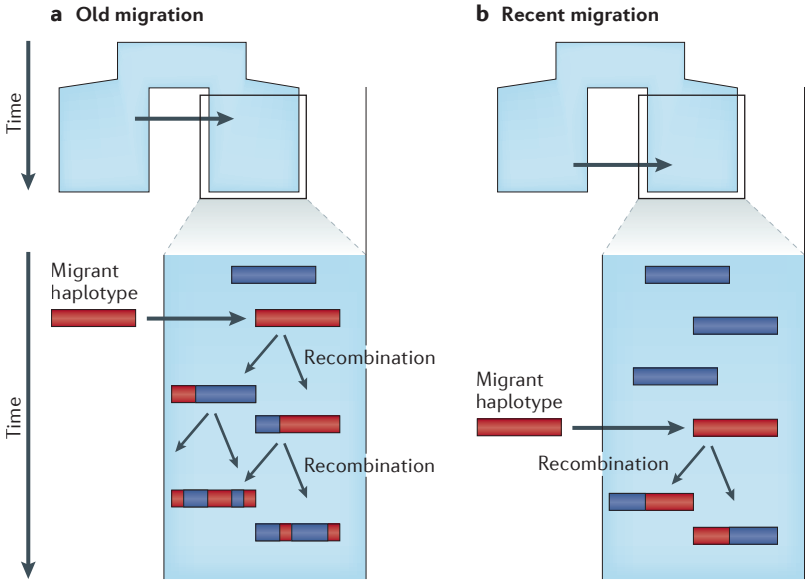


Figure 4 | **Distinguishing migration events based on linkage disequilibrium block structure.** Schematic representation of the expected distribution of the haplotype block lengths for an old migration event (**a**) and a recent migration event (**b**). The diagram shows two diverging populations that experience migration at some time in the past after the split and a zoom-in of what happens at the population that receives immigrant haplotypes. For simplicity, we assumed that all individuals share the same haplotype in the destination population (blue haplotype in the figure): that is, this haplotype has reached fixation. When a migrant haplotype (shown in red in the figure) enters a population, as times goes by, recombination breaks it into smaller fragments. Thus, blocks are expected to be shorter following an old migration event (**a**) than directly after a recent migration event (**b**), for which blocks are expected to be larger.

Another family of model-based methods approximates the likelihood as a sequence of conditional probabilities, with an additional term for each sequence that is added to a data set[87]. Such 'product of approximate conditionals' (PAC) or so-called 'copying model' methods have been extended to numerous demographic models[88,89], and finding better approximations has been an area of active research[87,90,91] that is likely to continue.

Historical gene flow and LD patterns

Population geneticists have long known that the movement of genes into a population can create strong patterns of LD in the regions of the genome experiencing that gene flow[41,92,93]. However, it remains a challenge to take advantage of this phenomenon to infer the history of gene flow[59,93,94]. One approach to disentangle alternative divergence models, such as the ones shown in FIG. 1, is based on the distribution of haplotype block lengths[95,96]. The principle is that when a migrant enters a population, it carries a set of chromosomes that, as time goes by, are broken into smaller fragments owing to recombination (FIG. 4). The distribution of block lengths depends not only on the recombination rate but also on the frequency at which a given population receives immigrants, and the older the migration event, the shorter the blocks are expected to be. Thus, the distribution of block lengths should allow disentangling alternative scenarios. A similar idea has recently been used to separate a scenario of admixture from ancestral population structure in the case of Neanderthals and modern humans[97]. In this study, by focusing on a subset of the data, the decay of an LD statistic as a function of the genetic distance among SNPs in present day European genomes supported a model with gene flow from Neanderthals, which is estimated to have occurred between 37,000 and 86,000 years ago. Other statistics have been proposed to detect more recent admixture events, which have mostly been applied to modern human populations[98,99]. Ideally, these statistics should be affected only by a specific factor, such as admixture. However, other demographic events (for example, bottlenecks) as well as selection can generate haplotype blocks[41,100], and it is still unclear how sensitive these LD statistics are to such events[97]. In principle, rather than looking at statistics of subsets of the data, a better description would be achieved with full-likelihood methods that express the probability of the entire data set under demographic models explicitly accounting for recombination and gene flow.

Likelihoods for models with recombination. Our ability to extract the information contained in LD patterns about migration and admixture relies on an explicit model of the process of recombination. However, obtaining likelihoods under such models implies complex expressions that are difficult to solve or to approximate (BOX 3). Full-likelihood methods that jointly estimate demography and recombination rates that have been developed so far use a model with just a single population[101–103]. Because of the difficulties of explicitly including intermediate levels of recombination (that is, neither zero recombination nor effectively free recombination), most likelihood methods are limited to small segments of the data, as is typical with genealogy samplers. Alternatively, the likelihood under simple single-population models can be obtained for pairs of loci, as a function of the recombination rates, and inference proceeds assuming independence of the pairs of loci as in composite likelihood approaches[104,105].

Among approaches that are being developed to include recombination in likelihood calculations are those based on the approximations of conditional likelihoods[87,90,91]. These methods seem promising, as these conditional distributions can be used to generate genealogies and ancestral recombination graphs (ARGs) that are consistent with the data, which in turn can be used to compute likelihoods by importance sampling[106]. Another promising approach for models with recombination for data from a small number of individuals (that is, three) but large numbers of loci is based on using generating functions for the underlying gene trees[78].

A promising family of approaches treats recombination as a spatial process along the genome[107,108]. In this framework, the ancestry of each site is modelled by a gene tree that changes at points of recombination as one moves along the genome, as a function of the recombination rates and of some underlying demographic model. This has been implemented in hidden Markov models (HMMs) to estimate divergence times and ancestral effective sizes[109,110] and to estimate population

Generating functions
Statistical technique used to obtain the distribution of sums of random variables, as required in computation of the probability of genealogies given the parameters of an underlying model.

Haplotype
A DNA sequence that is inherited as a single unit in the absence of recombination.

Bottlenecks
Reductions in the size of populations owing to stochastic events or owing to colonization of new areas (founder events).

Ancestral recombination graphs
(ARGs). Graphs that represent the ancestral relationship of homologous DNA sequences sampled from a single or multiple populations. In models with gene flow, an ARG includes coalescent, migration and recombination events.

Identity by descent (IBD). Two haplotypes are identical by descent if they are identical copies of a haplotype that are shared between individuals within families and hence are assumed to be identical by descent.

size changes[111] for data sets comprising a pair or trio of haploid genomes that are sampled from the same[111] or different populations[109,110]. The HMM framework is appealing as it allows obtaining likelihoods under complex models accounting for recombination and the correlation of genealogies of neighbouring sites. One main approximation is that coalescent times are treated in discrete time intervals rather than as a continuous variable. Recently, this was extended to models with gene flow followed by isolation[112]. For instance, looking at 10 Mb segments from each chromosome[12], the divergence between eastern gorillas (*Gorilla beringei*) and western gorillas (*Gorilla gorilla*) was estimated to involve a long period of continuous gene flow, since the split of the common ancestor (0.9–1.6 million years ago) until recently (80,000–200,000 years ago), after which gene flow ceased, fitting a model as depicted in FIG. 1c.

Finally, another promising avenue for further research is based on the distribution of haplotype block lengths as a function of immigration timing and rates[96]. This was implemented in a composite likelihood method based on the distribution of immigrant haplotype blocks ('migrant tracts'; FIG. 4), which was shown to have power to infer changes in migration rates up to 1,000 generations ago in a simulation study[96]. One of the limitations of this approach is that it assumes that the migrant haplotype blocks can be correctly identified without error, which is difficult to achieve for species with reduced differentiation. This approach has recently been extended to infer changes in migration rates through time[95] and was applied to humans to infer changes in historical gene flow rates from Europe using admixed African–American HapMap data. Other variations on this idea include methods that use summary statistics sensitive to LD[97,113] and methods to detect tracts of identity by descent (IBD) for informing on rates of migration[114]. One example is the recent derivation of the expected length of IBD tracts under different demographic models using coalescent arguments[115], showing that patterns of IBD in a sample of multiple individuals can be used to infer very recent demographic events (up to a few hundred generations ago). It is noteworthy that these models usually assume phased data, which are still difficult to obtain in practice (BOX 1).

Conclusions

Notwithstanding the difficulties of reference genome bias and phase uncertainty, population genomic data sets generated using NGS technologies offer tremendous potential for discerning the speciation process. However, in the quest to understand population divergence and speciation better, we wish to have theory and statistical methods that accommodate very large data sets and that connect the observed genomic patterns with relevant historical events for complex models of divergence. Currently, the available tools do not take full advantage of population genomic data sets, although there are sophisticated methods for taking a genome scan approach for particular aspects of the divergence process.

Going forwards, the greatest challenges on the theoretical and statistical side are to develop ways to include recombination fully in the analyses. Currently, AFS and genealogy-sampling approaches assume that different SNPs or loci are independently segregating, and other methods that take fuller account of recombination are restricted to smaller portions of the genome. NGS data have not yet changed our main paradigm of how populations diverge, but they have confirmed that natural selection is sometimes in conflict with gene exchange during the divergence process and that gene flow is a widespread process. We envision that great advances in population genomic inference will be achieved as comprehensive methods emerge for fully including recombination in our divergence models, as these will allow investigators to use all of the relevant information in their NGS data.

1. Darwin, C. *On the Origins of Species by Means of Natural Selection* (Murray, 1859).
2. Hohenlohe, P. A. *et al.* Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* **6**, e1000862 (2010).
This was the first study in which RAD-tag sequencing was used to scan genome-wide patterns of differentiation in the quest to find genes involved in adaptation.
3. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379 (2011).
4. Metzker, M. L. Sequencing technologies — the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).
This is an excellent Review of the NGS technologies, their applications, potential and limitations.
5. Davey, J. W. *et al.* Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Rev. Genet.* **12**, 499–510 (2011).
6. Altshuler, D. L. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
7. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nature Genet.* **43**, 1031–1034 (2011).
This study exemplifies the application of coalescence-based genealogy sampler methods to analyse NGS data, representing the largest data set analysed so far with such methods.
8. Lachance, J. *et al.* Evolutionary history and adaptation inferred from whole-genome sequences of diverse African hunter-gatherers *Cell* **150**, 457–469 (2012).
9. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genet.* **43**, 956–963 (2011).
10. von Holdt, B. M. *et al.* Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. *Nature* **464**, 898–902 (2010).
11. Prüfer, K. *et al.* The bonobo genome compared with the chimpanzee and human genomes. *Nature* **486**, 527–531 (2012).
12. Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012).
13. Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529–533 (2011).
14. The *Heliconius* Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).
15. Ellegren, H. *et al.* The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**, 756–760 (2012).
16. Kern, A. D. Correcting the site frequency spectrum for divergence-based ascertainment. *PLoS ONE* **4**, e5152 (2009).
17. Pool, J. E., Hellmann, I., Jensen, J. D. & Nielsen, R. Population genetic inference from genomic sequence variation. *Genome Res.* **20**, 291–300 (2010).
18. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nature Rev. Genet.* **12**, 443–451 (2011).
This provides a detailed Review on the challenges and recent developments on genotype and SNP calling for NGS data.
19. Dobzhansky, T. G. & Dobzhansky, T. *Genetics and the Origin of Species* (Columbia Univ. Press, 1937).
20. Coyne, J. A. & Orr, H. A. The evolutionary genetics of speciation. *Phil. Trans. R. Soc. B* **353**, 287 (1998).
21. Turelli, M., Barton, N. H. & Coyne, J. A. Theory and speciation. *Trends Ecol. Evol.* **16**, 330–343 (2001).
22. Futuyma, D. J. & Mayer, G. C. Non-allopatric speciation in animals. *Systemat. Biol.* **29**, 254–271 (1980).
23. Mayr, E. *Systematics and the Origin of Species: from the Viewpoint of a Zoologist* (Harvard Univ. Press, 1942).
24. Mayr, E. *Animal Species and Evolution* (Harvard Univ. Press, 1963).
25. Bolnick, D. I. & Fitzpatrick, B. M. Sympatric speciation: models and empirical evidence. *Annu. Rev. Ecol. Evol. Systemat.* **38**, 459–487 (2007).
26. Via, S. Sympatric speciation in animals: the ugly duckling grows up. *Trends Ecol. Evol.* **16**, 381–390 (2001).
27. Reznick, D. N. & Ricklefs, R. E. Darwin's bridge between microevolution and macroevolution. *Nature* **457**, 837–842 (2009).
28. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).

29. Barton, N. H. Genetic hitchhiking. *Phil. Trans. R. Soc. Lond. B* **355**, 1553–1562 (2000).
30. Wu, C. I. The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865 (2001).
31. Butlin, R. K. Recombination and speciation. *Mol. Ecol.* **14**, 2621–2635 (2005).
32. Pinho, C. & Hey, J. Divergence with gene flow: models and data. *Annu. Rev. Ecol. Evol. Systemat.* **41**, 215–230 (2010).
33. Nielsen, R. & Wakeley, J. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885–896 (2001).
This is one of the first papers in which a full likelihood approach based on genealogy samplers was applied to an isolation with migration model.
34. Hey, J. & Nielsen, R. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167**, 747–760 (2004).
35. Wakeley, J. & Hey, J. in *Molecular Approaches to Ecology and Evolution* 157–175 (Springer, 1998).
36. Luikart, G., England, P. R., Tallmon, D., Jordan, S. & Taberlet, P. The power and promise of population genomics: from genotyping to genome typing. *Nature Rev. Genet.* **4**, 981–994 (2003).
37. Nielsen, R. & Beaumont, M. A. Statistical inferences in phylogeography. *Mol. Ecol.* **18**, 1034–1047 (2009).
38. Levin, D. A. Interspecific hybridization, heterozygosity and gene exchange in *Phlox*. *Evolution* **29**, 37–51 (1975).
39. Wang, R. L., Wakeley, J. & Hey, J. Gene flow and natural selection in the origin of *Drosophila pseudoobscura* and close relatives. *Genetics* **147**, 1091–1106 (1997).
40. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310**, 321–324 (2005).
41. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Rev. Genet.* **9**, 477–485 (2008).
42. Nielsen, R. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218 (2005).
43. Stapley, J. *et al.* Adaptation genomics: the next generation. *Trends Ecol. Evol.* **25**, 705–712 (2010).
44. Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting FST. *Nature Rev. Genet.* **10**, 639–650 (2009).
45. Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
46. Beaumont, M. A. Adaptation and speciation: what can F_{ST} tell us? *Trends Ecol. Evol.* **20**, 435–440 (2005).
47. Gaggiotti, O. E. & Foll, M. Quantifying population structure using the F-model. *Mol. Ecol. Resources* **10**, 821–830 (2010).
48. Excoffier, L., Hofer, T. & Foll, M. Detecting loci under selection in a hierarchically structured population. *Heredity* **103**, 285–298 (2009).
49. Gompert, Z. & Buerkle, C. A. A. Hierarchical Bayesian model for next-generation population genomics. *Genetics* **187**, 903–917 (2011).
50. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
This provides a detailed description of the principles and properties of the *D* statistic (also known as the ABBA and BABA test), now widely used to detect and estimate rates of admixture and introgression.
51. Green, R. E. *et al.* A draft sequence of the neandertal genome. *Science* **328**, 710–722 (2010).
52. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
53. Eriksson, A. & Manica, A. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc. Natl Acad. Sci.* **109**, 13956–13960 (2012).
54. Beaumont, M. A. & Rannala, B. The Bayesian revolution in genetics. *Nature Rev. Genet.* **5**, 251–261 (2004).
55. Nielsen, R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942 (2000).
56. Williamson, S. H. *et al.* Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl Acad. Sci.* **102**, 7882–7887 (2005).
57. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
This was the first study solving the expected AFS for an isolation with migration model using the diffusion approximation, opening the door for computing likelihoods for genomic SNP data.
58. Excoffier, L. & Foll, M. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**, 1332–1334 (2011).
59. Adams, A. M. & Hudson, R. R. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* **168**, 1699–1712 (2004).
60. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97 (1931).
61. Kimura, M. Solution of a process of random genetic drift with a continuous model. *Proc. Natl Acad. Sci. USA* **41**, 144 (1955).
62. Lukić, S., Hey, J. & Chen, K. Non-equilibrium allele frequency spectra via spectral methods. *Theor. Popul. Biol.* **79**, 203–219 (2011).
63. Lukić, S. & Hey, J. Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics* **192**, 619–639 (2012).
64. Stephens, M. in *Handbook of Statistical Genetics* 3rd edn (eds Balding, D. J., Bishop, M. & Cannings, C.) 878–908 (Wiley, 2007).
65. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci.* **108**, 11983–11988 (2011).
66. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
67. Sirén, J., Marttinen, P. & Corander, J. Reconstructing population histories from single nucleotide polymorphism data. *Mol. Biol. Evol.* **28**, 673–683 (2011).
68. Gautier, M. & Vitalis, R. Inferring population histories using genome-wide allele frequency data. *Mol. Biol. Evol.* **30**, 654–668 (2013).
69. Kingman, J. F. C. On the genealogy of large populations. *J. Appl. Probab.* **19**, 27–43 (1982).
70. Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
71. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
72. Felsenstein, J. Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* **22**, 521–565 (1988).
73. Marjoram, P. & Tavaré, S. Modern computational approaches for analysing molecular genetic variation data. *Nature Rev. Genet.* **7**, 759–770 (2006).
74. Kuhner, M. K. Coalescent genealogy samplers: windows into population history. *Trends Ecol. Evol.* **24**, 86–93 (2009).
75. Hey, J. & Nielsen, R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl Acad. Sci. USA* **104**, 2785–2790 (2007).
76. Beerli, P. & Palczewski, M. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics* **185**, 313–326 (2010).
77. Wang, Y. & Hey, J. Estimating divergence parameters with small samples from a large number of loci. *Genetics* **184**, 363–379 (2010).
78. Lohse, K., Harrison, R. & Barton, N. H. A general method for calculating likelihoods under the coalescent process. *Genetics* **189**, 977–987 (2011).
This paper describes an interesting approach to obtain likelihoods for a large number of loci using generating functions that can be applied to isolation with migration models and can, in principle, deal with recombination.
79. Lohse, K., Barton, N. H., Melika, G. & Stone, G. N. A likelihood-based comparison of population histories in a parasitoid guild. *Mol. Ecol.* **21**, 4605–4617 (2012).
80. Beaumont, M. A. Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Systemat.* **41**, 379–406 (2010).
81. Sunnåker, M. *et al.* Approximate Bayesian computation. *PLoS Computat. Biol.* **9**, e1002803 (2013).
82. Hoban, S., Bertorelle, G. & Gaggiotti, O. E. Computer simulations: tools for population and evolutionary genetics. *Nature Rev. Genet.* **10**, 110–122 (2012).
83. Csilléry, K., Blum, M. G., Gaggiotti, O. E. & François, O. Approximate Bayesian computation (ABC) in practice. *Trends Ecol. Evol.* **25**, 410–418 (2010).
84. Becquet, C. & Przeworski, M. A new approach to estimate parameters of speciation models with application to apes. *Genome Res.* **17**, 1505–1519 (2007).
85. Nice, C. C. *et al.* Hybrid speciation and independent evolution in lineages of alpine butterflies. *Evolution* **67**, 1055–1068 (2013).
86. Li, S. & Jakobsson, M. Estimating demographic parameters from large-scale population genomic data using approximate Bayesian computation. *BMC Genet.* **13**, 22 (2012).
87. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
88. Davison, D., Pritchard, J. & Coop, G. An approximate likelihood for genetic data under a model with recombination and population splitting. *Theor. Popul. Biol.* **75**, 331–345 (2009).
89. Hellenthal, G., Auton, A. & Falush, D. Inferring human colonization history using a copying model. *PLoS Genet.* **4**, e1000078 (2008).
90. Steinrücken, M., Paul, J. S. & Song, Y. S. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theor. Popul. Biol.* 7 Sep 2012 (doi:org/10.1016/j.tpb.2012.08.004).
91. Paul, J. S., Steinrücken, M. & Song, Y. S. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics* **187**, 1115–1128 (2011).
This study describes a promising approximation for obtaining ARGs consistent with the data. This can in principle be applied to calculate likelihoods under isolation with migration models explicitly accounting for recombination.
92. Tachida, H. & Cockerham, C. C. Analysis of linkage disequilibrium in an island model. *Theor. Popul. Biol.* **29**, 161–197 (1986).
93. Nordborg, M. & Tavare, S. Linkage disequilibrium: what history has to tell us. *Trends Genet.* **18**, 83–90 (2002).
94. Myers, S., Fefferman, C. & Patterson, N. Can one learn history from the allelic spectrum? *Theor. Popul. Biol.* **73**, 342–348 (2008).
95. Gravel, S. Population genetics models of local ancestry. *Genetics* **191**, 607–619 (2012).
96. Pool, J. E. & Nielsen, R. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711–719 (2009).
This study proposes a solid theoretical framework to describe the haplotype block lengths in a population receiving immigrants.
97. Sankararaman, S., Patterson, N., Li, H., Pääbo, S. & Reich, D. The date of interbreeding between Neandertals and modern humans. *PLoS Genet.* **8**, e1002947 (2012).
98. Patterson, N. J. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
99. Loh, P.-R. *et al.* Inference of admixture parameters in human populations using weighted linkage disequilibrium. Preprint at arXiv [online], http://uk.arxiv.org/abs/1211.0251 (2012).
100. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Rev. Genet.* **4**, 587–597 (2003).
101. Griffiths, R. C. & Marjoram, P. Ancestral inference from samples of DNA sequences with recombination. *J. Computat. Biol.* **3**, 479–502 (1996).
102. Kuhner, M. K., Yamato, J. & Felsenstein, J. Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393–1401 (2000).
103. Wang, Y. & Rannala, B. Bayesian inference of fine-scale recombination rates using population genomic data. *Phil. Trans. R. Soc. B* **363**, 3921–3930 (2008).
104. Hudson, R. R. Two-locus sampling distributions and their application. *Genetics* **159**, 1805–1817 (2001).
105. McVean, G., Awadalla, P. & Fearnhead, P. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–1241 (2002).
106. De Iorio, M., Griffiths, R. C., Leblois, R. & Rousset, F. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theor. Popul. Biol.* **68**, 41–53 (2005).
107. Wiuf, C. & Hein, J. Recombination as a point process along sequences. *Theor. Popul. Biol.* **55**, 248–259 (1999).

108. Wiuf, C. & Hein, J. The ancestry of a sample of sequences subject to recombination. *Genetics* **151**, 1217–1228 (1999).
109. Hobolth, A., Christensen, O. F., Mailund, T. & Schierup, M. H. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.* **3**, e7 (2007).
110. Mailund, T., Dutheil, J. Y., Hobolth, A., Lunter, G. & Schierup, M. H. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS Genet.* **7**, e1001319 (2011).
111. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
112. Mailund, T. *et al.* A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genet.* **8**, e1003125 (2012).
This is the first application of HMM-based methods for isolation with migration models, explicitly accounting for recombination.
113. Pugach, I., Matveyev, R., Wollstein, A., Kayser, M. & Stoneking, M. Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol.* **12**, R19 (2011).
114. Browning, S. & Browning, B. Identity by descent between distant relatives: detection and applications. *Annu. Rev. Genet.* **46**, 617–633 (2012).
115. Francesco Palamara, P., Lencz, T., Darvasi, A. & Pe'er, I. Length distributions of identity by descent reveal fine-scale demographic history. *Am. J. Hum. Genet.* **91**, 809–822 (2012).
116. Rogers, A. R. & Jorde, L. B. Ascertainment bias in estimates of average heterozygosity. *Am. J. Hum. Genet.* **58**, 1033–1041 (1996).
117. Nielsen, R. Population genetic analysis of ascertained SNP data. *Hum. Genom.* **1**, 218–224 (2004).
118. Pool, J. E. *et al.* Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS Genet.* **8**, e1003080 (2012).
119. Corbett-Detig, R. B. & Hartl, D. L. Population genomics of inversion polymorphisms in *Drosophila melanogaster*. *PLoS Genet.* **8**, e1003056 (2012).
120. Li, R. Q. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
121. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nature Rev. Genet.* **12**, 703–714 (2011).
122. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nature Biotech.* **26**, 1146–1153 (2008).
123. Hudson, R. R. Gene genealogies and the coalescent process. *Oxford Surveys Evol. Biol.* **7**, 44 (1990).
124. Nordborg, M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923–929 (2000).
125. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).

Acknowledgements
This work was supported by grants from the US National Science Foundation and the US National Institutes of Health to J.H.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Jody Hey's homepage: http://genfaculty.rutgers.edu/hey
1000 Genomes Project: http://www.1000genomes.org
1001 Genomes Project: http://www.1001genomes.org
***Drosophila* Population Genomics Project:** http://www.dpgp.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF