COMPUTATIONAL TOOLS

Methods and models for unravelling human evolutionary history

Joshua G. Schraiber and Joshua M. Akey

Abstract | The genomes of contemporary humans contain considerable information about the history of our species. Although the general contours of human evolutionary history have been defined with increasing resolution throughout the past several decades, the continuing deluge of massively large sequencing data sets presents new opportunities and challenges for understanding human evolutionary history. Here, we review the signatures that demographic history imparts on patterns of DNA sequence variation, statistical methods that have been developed to leverage information contained in genome-scale data sets and insights gleaned from these studies. We also discuss the importance of using exploratory analyses to assess data quality, the strengths and limitations of commonly used population genomics methods, and factors that confound population genomics inferences.

Exploratory data analyses (EDA). The initial stages of 'digging into' a data set, usually by plotting low-dimensional summaries of the data.

Records of population history are embedded in the patterns of DNA sequence variation that exist among present day individuals. Indeed, genetics has become integral in delineating the evolutionary history of populations and species. However, reading the stories of population history written in DNA sequence variation is challenging and requires both comprehensive and accurate data and methods to interpret the complicated interplay of forces (including mutation, changes in population size, nonrandom mating, admixture and selection) that shape extant patterns of genetic diversity.

It seems almost difficult to recall that, until fairly recently, inferences of human history were hampered by small sample sizes and limited amounts of genetic data. Despite these challenges, a coherent narrative of human evolutionary history has been forged over the past several decades, providing striking insights into when and where anatomically modern humans arose (~200,000 years ago, probably in East Africa), the 'out-of-Africa' dispersal and consequent population bottleneck ~70,000 years ago, and the routes and timing of human migrations that led to colonization of all habitable parts of the world[1]. However, the development of massively parallel sequencing technology[2] has enabled genome-scale data sets to be generated at a frenetic pace[3,4]. Thus, the challenges are no longer related to a lack of data (although more comprehensive sampling of geographically diverse individuals is needed[5]), but rather to how to interpret the deluge of whole-genome and whole-exome sequences that have emerged and continue to emerge, which hold considerable promise in revealing more nuanced, realistic and complicated models of human history[4,6–8].

Here, we focus on the methodological and interpretive issues that arise in the inference of human evolutionary history from large-scale sequencing data. We concentrate largely on methods and approaches for inferring increasingly sophisticated models of human demographic history, and not adaptive evolution, which has been reviewed extensively[9–13]. We also emphasize the critically important role that quality control (QC) and exploratory data analyses (EDA) have in facilitating robust population genomics inferences, comprehensively discuss existing methodologies and their limitations (TABLE 1), and highlight potential pitfalls that can arise when attempting to extract parameter-rich (that is, complex) models of human history from genome-scale sequencing data. Although our emphasis is on human evolutionary history, the methodological tools and issues discussed are broadly applicable. Indeed, many population genomics methods have been pioneered in the context of human data owing to the availability of large data sets, but as sequencing technology makes it increasingly possible to obtain high-quality genomic data from other species, the issues and approaches described here extend well beyond the inference of human evolutionary history.

QC and EDA

The path from raw sequencing reads obtained from individuals to catalogues of single nucleotide variants (SNVs) involves many technical and analytical choices

Department of Genome Sciences, University of Washington, 3720 15th Avenue NE, box 355065, Seattle, Washington 98195–5065, USA.
e-mails: schraib@uw.edu; akeyj@uw.edu

doi:10.1038/nrg4005
Published online
10 November 2015

Table 1 | **Software for demographic inferences**

Name	Data type	Inference	Notes	Refs
STRUCTURE	Unlinked multi-allelic genotypes	Population structure, admixture	User-friendly GUI; can be computationally demanding	32
FRAPPE	Unlinked bi-allelic SNVs	Population structure, admixture	Alexander *et al.*[41] argue that convergence is not guaranteed	40
ADMIXTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Estimates the number of populations via cross-validation error	41
fastSTRUCTURE	Unlinked bi-allelic SNVs	Population structure, admixture	Obtains variational Bayesian estimates of posterior probability distribution	42
Structurama	Unlinked multi-allelic genotypes	Population structure, admixture	Uses a Dirichlet process to estimate the number of populations	43
HAPMIX	Phased haplotypes; reference panel	Chromosome painting	Requires populations to be specified a priori	48
fineSTRUCTURE	Phased haplotypes	Population structure, admixture, chromosome painting	Can be used to identify the number and identity of populations	49
GLOBETROTTER	Phased haplotypes	Population structure, admixture, chromosome painting	Extends the fineSTRUCTURE approach to estimate unsampled ancestral populations and admixture times	7
LAMP	Phased haplotypes; reference panel	Chromosome painting	Identifies local ancestry in windows, rather than using an HMM, so is more discrete than other approaches	52
PCAdmix	Phased haplotypes	Chromosome painting, population structure	Uses PCA in small chunks followed by an HMM to estimate local ancestry	53
dadi	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Requires some Python-coding skills; applicable to up to three populations	60
Fastsimcoal2	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Can also be used to simulate data under the SMC	62,63
Treemix	Frequencies of unlinked bi-allelic SNVs	Admixture graph	Highly multimodal likelihood surface and heuristic search; redo inference from many starting points	64
fastNeutrino	Frequency spectrum of unlinked bi-allelic SNVs	Demographic history	Applicable only to a single population; designed specifically for extremely large sample sizes	65
DoRIS	Lengths of IBD blocks between pairs of individuals	Demographic history	IBD must be inferred (for example, using Beagle or GERMLINE); specification of lower cut-off minimizes false-negative IBD tracts	71,72
IBS tract inference	Lengths of IBS blocks between pairs of individuals	Demographic	IBS can easily be confounded by missing data and/or sequencing errors	76
PSMC	Diploid genotypes from one individual	Demographic history	Best used in MSMC's PSMC mode, which uses the SMC to more accurately model recombination than the original PSMC; applicable to a single population	78
MSMC	Whole genome, phased haplotypes	Demographic history	Requires large amounts of RAM; cross-coalescence rate should not be interpreted as migration rate	82
CoalHMM	Whole genome, phased haplotypes	Demographic history	Multiple applications, including inference of population sizes, migration rates and incomplete lineage sorting	83–87
diCal	Medium-length, phased haplotypes	Demographic history	Uses shorter sequences than MSMC, but can be applied to multiple individuals in complex demographic models; infers explicit population genetic parameters for migration rates	89,92
LAMARC	Short, phased haplotypes	Demographic history	Requires Monte Carlo sampling of coalescent genealogies; very flexible	93
BEAST	Short, phased haplotypes	Species trees, effective population sizes	Used mainly as a method of phylogenetic inference. Can also infer population size history	94
MCMCcoal	Short, phased haplotypes	Divergence times between populations	Now incorporated into the software BPP[131]	95
G-PhoCS	Short, (un)phased haplotypes	Demographic history	Incorporates migration into the MCMCcoal framework. Averages over unphased haplotypes	96
Exact likelihoods using generating functions	Short, phased haplotypes	Demographic history	Implemented in Mathematica; applicable only to specific classes of multi-population models	97,98

BEAST, Bayesian evolutionary analysis by sampling trees; BPP, Bayesian phylogenetics and phylogeography; CoalHMM, coalescent HMM; *dadi*, diffusion approximations for demographic inference; diCal, demographic inference using composite approximate likelihood; DoRIS, demographic reconstruction via IBD sharing; G-PhoCS, generalized phylogenetic coalescent sampler; GERMLINE, genetic error-tolerant regional matching with linear-time extension; GUI, graphical user interface; HMM, hidden Markov model; IBD, identity by descent; IBS, identity by state; LAMARC, likelihood analysis with metropolis algorithm using random coalescence; LAMP, local ancestry in admixed populations; MCMC, Markov chain Monte Carlo; MSMC, multiple SMC; PCA, principal components analysis; PSMC, pairwise SMC; RAM, random access memory; SMC, sequentially Markov coalescent; SNVs, single nucleotide variants.

that may influence downstream analyses. Thus, initial QC and EDA are of considerable importance. Here, we outline some salient aspects of genotype calling, QC and EDA that are important for population genomics inferences.

There are two key issues involved in calling genetic variation from high-throughput sequencing data. First, short reads must be mapped to a reference genome; second, those reads must be used to determine genotypes. Read mapping is influenced by the fact that sequencing reads containing sites that differ from the reference genome will, in general, be harder to map, because they do not exactly match anywhere in the genome (or they may even by chance match a non-homologous sequence elsewhere in the genome). Although modern mapping software can deal with a small number of mismatches per read, regions with clustered mutations are still likely to seem less variable than they truly are owing to the difficulty of mapping highly divergent reads.

Given a set of reads that map with high quality, calling genotypes in haploid organisms is fairly straightforward: every site is either reference or alternative. However, in diploid organisms, such as humans, heterozygous sites introduce additional complexity into genotype calling. Accurate calling of heterozygous sites requires high-coverage data to mitigate the effects of sequencing errors and the stochasticity inherent in sampling each allele. For instance, at a site that is sequenced to a depth of 2×, observing 1 read supporting the reference allele and 1 read supporting the alternative allele may simply reflect a chance sequencing error. Moreover, for low-coverage sites, it is possible that one or the other allele is simply not sampled at all by any reads. However, for a site covered at 20×, observing 10 reads with a reference allele and 10 reads with an alternative allele would make a strong case for the site being truly heterozygous.

Several tools exist to call diploid genotypes from resequencing data. Early methods for SNV calling used simple, counting-based rules (often favouring a homozygous reference genotype), whereas most modern methods operate in a probabilistic framework by computing genotype likelihoods[14]. Genotype likelihoods quantify the probability of the observed data (that is, the reads covering a site), given every possible diploid genotype. Perhaps the most-commonly used software for making hard genotype calls in a probabilistic framework is the Genome Analysis Toolkit (GATK)[14–16]. GATK requires several tuning parameters that can influence SNV calling in unpredictable ways; thus, it is important to keep up to date on best practices in using GATK. Other methods have been developed to perform downstream population genomics analyses using genotype likelihoods, which obviates the need to call a particular genotype at every site and more-properly models the uncertainty inherent to inferences from short-read data (for example, REF. 17).

When genotypes have been called, they are typically subjected to several filtering criteria to ensure that only the most-accurate data are used in downstream analyses. Parameters that are most-commonly considered include sequencing depth, genotype and mapping quality[18], and measures of allelic bias (that is, the proportion of reads from each allele). Filtering is usually done on an individual basis, and then sites that are missing in a substantial fraction of individuals (for example, >10%) are removed from further analyses. Note that many population genetics statistics, such as nucleotide diversity, require knowledge of the total number of sites sequenced. Thus, it is critical to apply the same filtering criteria to all sites for each individual, and not just to those where a SNV was called. For example, there is a high probability that a heterozygous site with 2× coverage will be called as a homozygote. If a higher depth of coverage is used to filter SNV calls, then the same criteria should also be applied to all sites, so that an accurate estimate of the total number of sites used in the analysis can be obtained.

Another QC step that is particularly useful in large data sets is to filter sites that strongly deviate from Hardy–Weinberg equilibrium (HWE). Specifically, having an excess of heterozygous individuals in a data set (in the extreme case, sites can appear as heterozygous in all individuals) is diagnostic of paralogous sequence variants that can result from mismapping of recently duplicated regions, or simply sites that exist in a genomic context that is hard to sequence. Although there are genome-scale deviations from HWE[19], filtering out sites that show excessively strong departures from HWE will often result in removing only a small amount of data and avoiding serious errors in data analysis. Furthermore, maps of segmental duplications[20] and other potentially problematic sequences can also be used as filtering criteria to focus on regions that are most amenable to accurate read mapping and genotype calling.

When a set of robust genotypes is obtained, it is useful to perform some simple EDA to assess overall data quality, understand characteristics of the data and identify potential unwanted sources of variability. For example, principal components analysis (PCA) forms a central pillar of EDA for population genomics data. In brief, PCA finds the eigenvectors of the covariance matrix derived from genotypes among individuals. These eigenvectors provide the coefficients of the linear combinations of genotypes that most-effectively differentiate the various samples, without requiring a priori information on the classification of samples. Then, samples are typically plotted according to their loadings on the first few PCs. Ideally, samples will be separated according to their recent ancestry; individuals with more-similar genotypes on average will cluster closer in PC space, whereas more-distantly related individuals will lie further apart (as discussed below). However, PCA can also be used to identify technical sources of variation, which might arise if samples are sequenced by different instruments or at multiple facilities, or are processed with different batches of reagents[3]. It is particularly important to assess these factors when integrating data sets from distinct sources. Overall data summaries, such as the ratio of transitions to transversions, and individual-level metrics, such as the number of heterozygous sites, are also helpful in

Likelihoods
The probabilities of the data given various models and their parameters, thought of as functions of those parameters. The parameter values that maximize the probability of the data in each model are called maximum likelihood estimates.

Eigenvectors
Vectors that, when multiplied by a given matrix, still point in the same direction.

Covariance matrix
An $n \times n$ matrix describing the covariance between each pair in a sample of size n.

identifying systematic biases and poor-quality data from particular samples, respectively. In short, massively parallel sequencing data sets are complex and imperfect. The goals of QC and EDA are to minimize the effect of genotyping errors on subsequent analyses and to better understand the non-biological factors that contribute to patterns of variation.

Intelligently leveraging genomic-sequence data

A powerful feature of genome-scale data is that it allows careful choice of the genomic regions that are most appropriate to answer particular questions of interest. For example, in analyses of demographic history, it is important to mitigate the confounding effects of natural selection, which can lead to biased inferences. Although an elegant solution to this issue would be to explicitly account for the joint effects of selection and demography, the methods to robustly do so are not well developed (but see REFS 21,22 for strategies that can simultaneously model selection and demography). A more-straightforward approach is to restrict analyses to genomic regions that are least likely to be influenced by selection (FIG. 1). Indeed, carefully sampling putatively neutral regions has been shown to significantly influence inferences such as sex-biased gene flow[23,24] and recent growth rates[25]. These studies have largely focused on genomic regions that are not in close proximity to protein-coding regions. However, the development of comprehensive functional genomics data sets, such as the Encyclopedia of DNA Elements (ENCODE) Project[26] and the Roadmap Epigenomics Project[27], has created powerful new opportunities for population genomics inferences by facilitating the identification of more-carefully defined neutral regions. For instance, high-resolution maps of regulatory DNA defined by DNase I-hypersensitive sites (DHS), chromatin immunoprecipitation followed by sequencing (ChIP–seq) and histone marks now exist[26,27]. Such data, when integrated with information about evolutionary conservation[28], background selection[29], regions that have experienced recent positive selection[12] and sequences that show accelerated rates of evolution in the human lineage[30] (FIG. 1), allow putatively neutral sequences to be defined in ways that were previously not possible.

In addition, maps of recombination can also be integrated into a comprehensive sampling strategy to guide decisions on how best to minimize the effects of background selection (that is, choosing neutral regions that are not closely linked to functionally important sequences). Tools to automate the selection of putatively neutral regions have been developed[31], although at present they do not incorporate all of the potential features that could be leveraged and there is no consensus on what specific combination of functional and comparative genomics features is best for delineating putatively neutral sequences. When using exome data, for example, the wealth of functional genomics phenotypes and evolutionary conservation can also be leveraged to focus on the variants that are least affected by selection (for instance, unconstrained fourfold synonymous sites that do not overlap DHS), but background selection is still likely to be an issue when making inferences from protein-coding sequences. In short, a considerable amount of information now exists about the landscape of putatively functional sequences across the human genome, and this knowledge should be leveraged when making inferences of population demographic history.

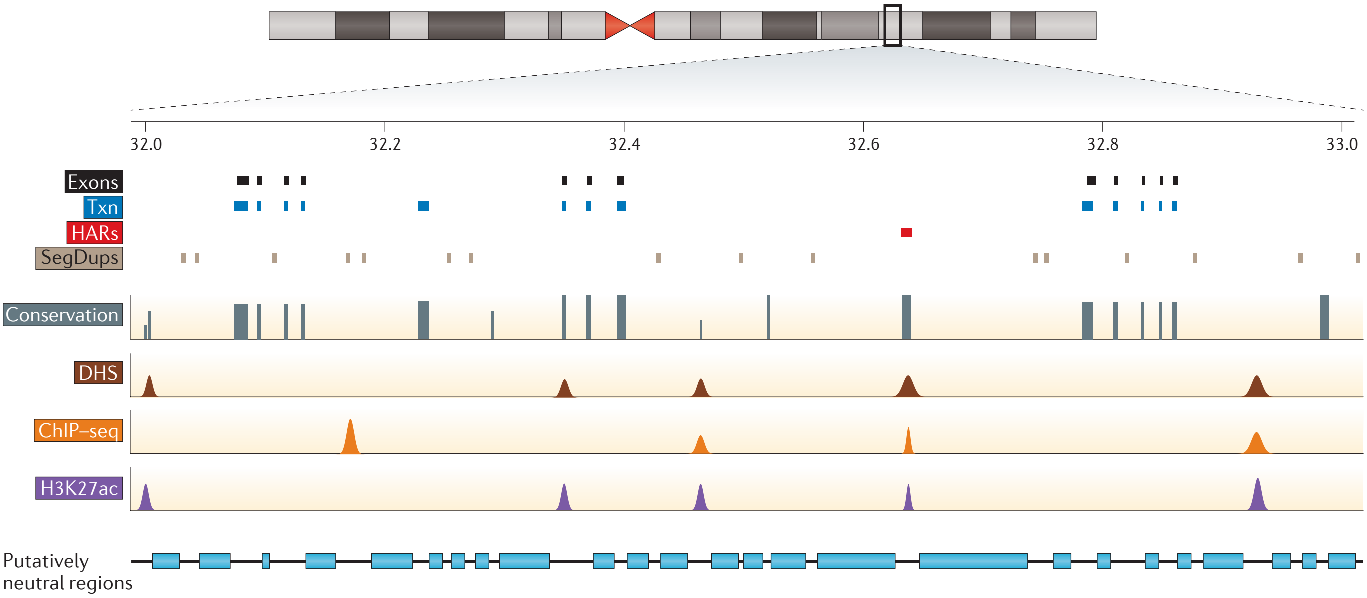

Figure 1 | **Identifying demographically informative genomic regions.** Functional and comparative genomics data can be leveraged to identify putatively neutral regions in a principled way. This schematic shows various functional and comparative genomics data, as well as sequences that are structurally complex (segmental duplications (SegDups)) or subject to adaptive evolution (human accelerated regions (HARs)). ChIP–seq, chromatin immunoprecipitation followed by sequencing; DHS, DNase I-hypersensitive sites; H3K27ac, histone H3 acetylated at lysine 27; Txn, transcription.

Inferences of population structure and admixture

After a set of robust genotypes from appropriate genomic regions is obtained, inferences of population structure are often performed. Understanding the genetic structure of sampled individuals serves two purposes. First, patterns of population structure are of direct interest in understanding human evolutionary history, times of population splitting, migration rates and patterns of mating among individuals. Second, many approaches for inferring parametric models of demographic history assume either a single, randomly mating population (a panmictic population) or knowledge of population structure. The development of software to effectively infer population structure using genotype data revolutionized population genomics, both as a tool for EDA and as a method for explicit hypothesis testing[32]. Here, we discuss different methods and approaches for identifying population structure and inferring population genetics parameters governing such patterns. FIGURE 2 illustrates the types of inferences that can be made from the methods described below, assuming a fairly simple three-population model with differential migration.

Identifying populations and testing for admixture. PCA has long been an important tool for inferring and visualizing population genetic structure[33]. PCA does not require a priori knowledge of population structure because it acts to project an individual's multilocus genotype onto a small number of dimensions (often two dimensions) that maximally separate the data (FIG. 2). With sufficient amounts of data, even very fine-scale patterns of population structure can be detected[34], which are sometimes more readily observed in lower-ranked PCs than in higher-ranked PCs[35]. When applying PCA to exome or genome sequencing data, it is important to prune SNVs based on patterns of linkage disequilibrium (LD), as PCA assumes that markers are independent. Furthermore, care must be taken to not over-interpret PCA results. Although it is possible to interpret PCA in terms of mean coalescence times between pairs of individuals[36], several caveats must accompany inferences made from PCA. In populations with recent, large-scale shifts in demography (such as humans), the directions of highest variability may be counterintuitive. For example, François *et al.*[37] found that, in models of range expansion, the directions of highest variability were perpendicular to the axis of expansion, which is contrary to the expectation that the direction of the highest variability should be parallel to the direction of expansion, as assumed by Cavalli-Sforza *et al.*[33]. Similarly, although it is possible that projections onto PC space may match geographical distributions of individuals under models of isolation by distance[34,38], this depends crucially on the sampling strategy and should not be expected in most cases in which PCA is applied to genetic data (but see REF. 39 for a PCA-like approach that explicitly incorporates spatial information).

In addition to PCA, many approaches for population assignment have been developed, but they are all largely similar to the popular software STRUCTURE[32], in that they look for groups of individuals that share common underlying allele frequencies and that are mutually in HWE. However, the specific technical details vary among methods. The original implementation of STRUCTURE is Bayesian and uses Markov chain Monte Carlo (MCMC) to average over underlying allele frequencies and assignments, and hence can be unreasonably slow for modern genomic data sets. By contrast, likelihood methods such as FRAPPE[40] and ADMIXTURE[41] are substantially faster and are recommended for large data sets. fastSTRUCTURE[42] allows for fast Bayesian inference of population structure, running on comparable timescales to ADMIXTURE but potentially providing more information about uncertainty in a given runtime, because ADMIXTURE requires bootstrap resampling to compute standard errors, whereas fastSTRUCTURE provides estimates of credible intervals as a natural part of the inference.

A key issue when running a STRUCTURE-like analysis is that the number of populations expected must be defined a priori (but see Structurama[43] for a method that uses a Dirichlet process to jointly infer the number of populations along with the population assignments). This is because the likelihood of the data will always be improved by adding more parameters (that is, more populations); however, those extra populations are merely overfitting the noise in the data. In the original publication, the researchers who developed STRUCTURE[32] used the marginal likelihood of the data with different numbers of populations to estimate the optimal number of populations. Although marginal likelihood comparison is theoretically well justified, it is difficult in practice to obtain stable estimates of the marginal likelihood from an MCMC run[44]. Therefore, care must be taken when using marginal likelihoods to infer the number of populations and, ideally, results of this analysis should be corroborated with alternative approaches. For example, ADMIXTURE recommends finding the number of populations that minimizes the cross-validation error; that is, the error in predicting the genotypes of held-out markers across individuals. Raj *et al.*[42] found that this approach does not work well with fastSTRUCTURE, and instead suggested using a marginal likelihood approximation combined with an estimate of the effective number of populations to obtain bounds on the possible number of populations.

In addition to identifying populations, STRUCTURE-like analyses provide an estimate of the fraction of each individual's genome that comes from each population (FIG. 2). In the case of STRUCTURE, specifically, the analysis is also able to assess the probability that an individual belongs to one or the other population; note that this is distinct from the more-common activity of estimating admixture proportions.

Local ancestry deconvolution: chromosome painting. When an individual is identified as admixed, it can be helpful to identify the regions of their genome that come from the different source populations. For instance, identifying segments of a specific ancestry can be useful in admixture mapping[45]. Moreover, by explicitly

Panmictic population
A group of individuals among whom random mating occurs.

Linkage disequilibrium
(LD). Nonrandom association between alleles at physically distinct genomic loci. Over time, this will be broken down by recombination.

Coalescence times
The times in the past when genomic regions shared a common genetic ancestor.

Isolation by distance
Genetic differentiation between individuals induced by geographic separation. Individuals that are closer geographically will be closer genetically.

Overfitting
By adding more parameters to a model, it will begin to model the noise in the observed data, rather than the true underlying mechanism of data generation. Overfit models will generalize poorly to new data sets.

Cross-validation error
The error in predicting the structure of a held-out portion of the data, when a model is trained on a subset of the whole data set. Minimizing cross-validation error is an effective way to choose parameters and hyperparameters.

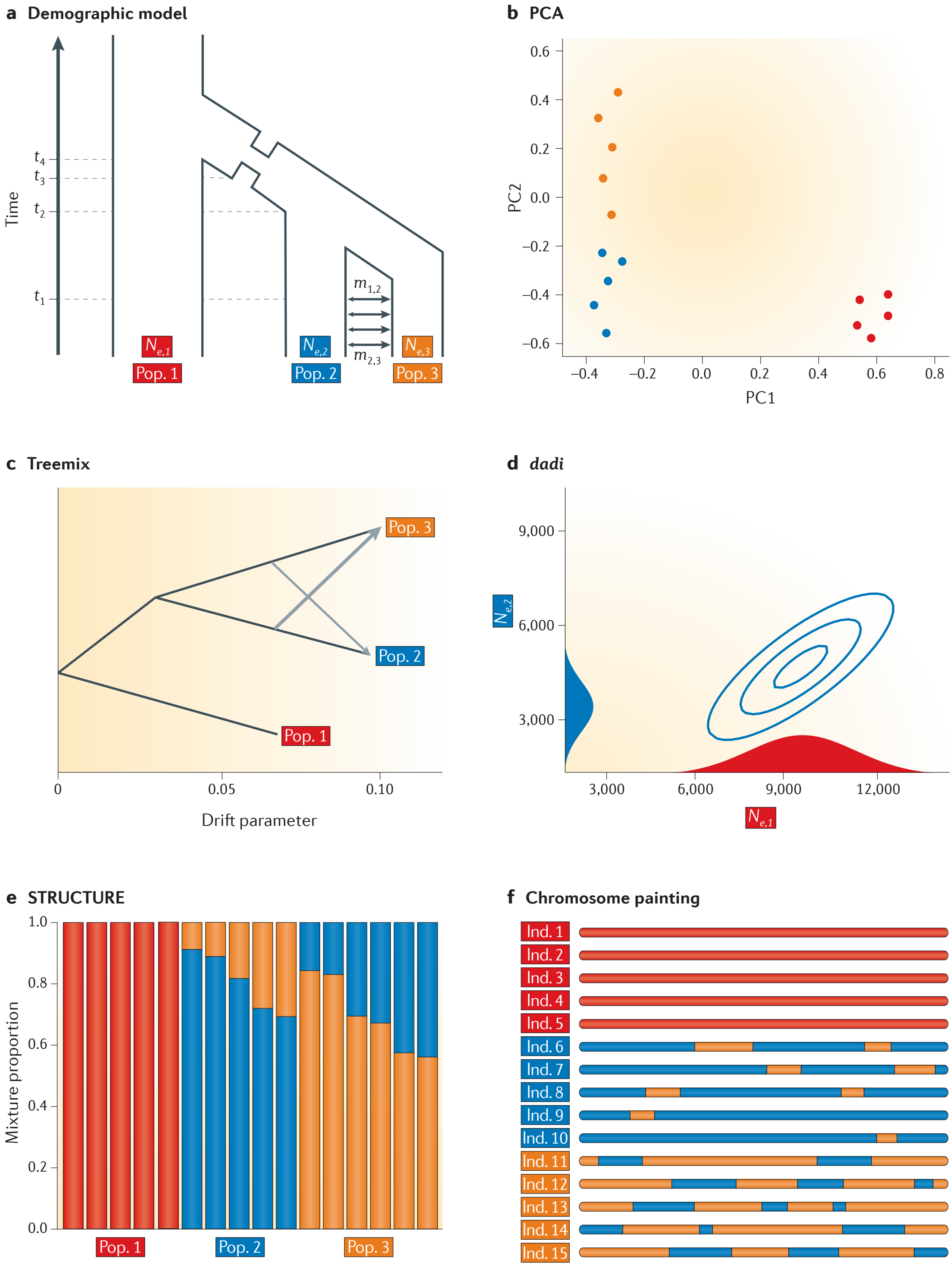

Figure 2 | **Inferring population demographic history.** A simple three-population model with changes in population size and asymmetric gene flow is shown. **a** | Demographic model. **b–f** | Schematics for the output of various methodological tools discussed in the text are illustrated. Principal components analysis (PCA) qualitatively illustrates population structure and admixture of populations 2 and 3 by the spread of individuals along PC2 (part **b**). Treemix illustrates different migration rates between populations 2 and 3 (part **c**). The difference in arrowhead size indicates asymmetrical migration rates. *dadi* (diffusion approximations for demographic inference) shows a contour plot of the likelihood surface of the effective population size of populations 2 and 3, as well as profile likelihoods of effective size for each population (part **d**). STRUCTURE provides estimates of the proportion of each individual genomes from populations 1, 2 and 3 (part **e**). Chromosomal painting shows the specific tracts of sequences inherited from ancestors in each population (part **f**). Ind., individual; $m_{i,j}$, migration rate from population i to population j; $N_{e,i}$, effective population size of population i; Pop., population.

modelling correlation in ancestry along the genome and the length distribution of ancestry segments, these approaches can obtain a more fine-grained picture of the history of admixture[46].

Production of such maps, known as local ancestry deconvolutions (or, more colloquially, as chromosome paintings; FIG. 2), requires a model of how ancestry changes along a genome. In principle, modelling genetic ancestry using the coalescent with recombination (which results in genealogical representations known as ancestral recombination graphs) provides a generative model of how individuals are related across different parts of their genome. Unfortunately, computation under the coalescent with recombination is extremely difficult, because very few calculations can be done analytically and the number of ancestral recombination graphs compatible with a given data set is very large. One of the most common approaches to inference of local ancestry is to use an approximation to the coalescent with recombination, which was proposed by Li and Stephens[47], and approximate the genome of the individual in question as a noisy mixture of the genomes of a reference set of individuals using a hidden Markov model (HMM). When this reference set of individuals is assigned to a priori populations, the population labels can be transferred to the chromosome, resulting in an assignment of each segment of chromosome to a particular ancestral population. Examples of algorithms using this approach include HAPMIX[48] and fineSTRUCTURE[49]. Note that, unlike HAPMIX, fineSTRUCTURE identifies populations that are not known a priori, which is similar to what STRUCTURE-like approaches do for population assignment. The method GLOBETROTTER modifies and extends the fineSTRUCTURE algorithm to account for ancestry from unsampled populations and has been used to reconstruct fine-scale population structure worldwide[7]. A key limitation of these methods is that they assume that admixture tract lengths are exponentially distributed and independent. For low rates of admixture, this is a reasonable assumption[6,50]; however, recent, strong admixture results in correlated admixture tracts that are stochastically larger than expected under an exponential distribution[51].

Several other methods for local ancestry deconvolution exist that do not work within a generalized Li and Stephens framework. However, all require some form of reference panel. Some, such as LAMP (local ancestry in admixed populations)[52] and PCAdmix[53], are fundamentally based on breaking the genome into windows and clustering relative to the reference panel within each window. The advantage of these methods is that they do not require assumption of a parametric population genetic model, which is necessarily an approximation to the complex dynamics of the underlying ancestral recombination graph. This makes such approaches applicable in cases in which the underlying demographic history is extremely complex or unknown, but can result in a substantial loss of power and interpretability. Moreover, choosing hyperparameters of such models, such as an optimal window size, is an important consideration. Windows that are too small will not have enough SNVs (and thus no information regarding ancestry), whereas windows that are too large may be disrupted by recombination and therefore may contain ancestry from different sources. LAMP provides an algorithm for choosing an optimal window size based on the assumed admixture parameters (for example, its intensity and how long ago it occurred), whereas the authors of PCAdmix state that their method performs well assuming windows contain at least ten SNVs.

Inferring complex demographic models

Armed with a better understanding of the genetic structure of a sample, it is now possible to explore more-complex and parameter-rich demographic models, including events such as population divergence, migration and changes in effective population size. In contrast to methods for determining population structure, there is no 'black-box software' to take genetic data and return an essentially unsupervised inference of demographic history. Therefore, these methods require the user to specify a model within the scope of inference of a given method and to provide parameter estimates within the context of that model. Most of the following methods are likelihood-based, and we make several best-practice recommendations about how to use them in BOX 1. For models in which the likelihood cannot be calculated, we examine the simulation-based strategy of approximate Bayesian computation (ABC) in BOX 2.

Methods based on the SFS. One of the most-useful summaries of population genomic data is the site frequency spectrum (SFS; FIG. 3). The SFS is a count of how many derived alleles in a sample of size n show up in $1/n$, $2/n$, … $(n\text{-}1)/n$ individuals (note that, if ancestral and derived statuses cannot be assigned, it is possible to use the folded SFS, which tracks minor allele frequency). Many common statistics in population genetics, including the number of segregating sites (S) and the average nucleotide diversity (π) are themselves summaries of the SFS.

The SFS contains information about both demography and natural selection in a population. For a panmictic population of a constant size subjected to no natural selection, the proportion of alleles found in i out of n chromosomes is proportional to $1/i$ (REF. 54), and deviations from this expectation can be used to make inferences about population history. There are several important qualitative features of a SFS that provide a clue towards population history (FIG. 3). For instance, recent population growth is indicated by an excess of low-frequency alleles compared with the expectation under neutrality; this is caused by the influx of new mutations in newly born individuals. Similarly, a recent population bottleneck results in an excess of low-frequency alleles, which is due to the reduction in population size causing low-frequency alleles to be lost. Population subdivision can have various effects on the SFS, including an increase in both medium-frequency and high-frequency alleles, depending on migration rates and divergence times.

Ancestral recombination graphs
Graph structures representing the genealogical history of a sample with a recombining genome. In addition to coalescence events (which bring two lineages together and therefore reduce the number of lineages in the graph), recombination events cause splits to occur, which increases the number of lineages in the graph.

Hidden Markov model
(HMM). A statistical model in which a set of underlying hidden states are assumed to follow Markov chain dynamics and induce a set of observed states.

Reference panel
A large number of individuals, related to samples of interest, for which some quality is known (for example, allelic phase).

Effective population size
The size that a theoretical population evolving under a Wright–Fisher model would need to be in order to match aspects of the observed genetic data.

Box 1 | General guidelines for likelihood inference from genomic data

Population genomic data enable inference of detailed, parameter-rich models. However, complicated models require complicated inferential steps; here, we outline several important steps that should be taken to ensure that parameter estimates are optimal.

Several methods described in this Review assume that genotyped sites are unlinked (for example, STRUCTURE or *dadi* (diffusion approximations for demographic inference)). Although application of the theory of composite likelihood generally ensures that correlated allele frequencies across loci do not result in biased parameter estimates, it will result in estimates of error that indicate a higher degree of confidence than is truly warranted by the data[107]. Hence, we recommend pruning linkage disequilibrium (LD) before analysing data in a model that assumes unlinked sites. This can be done by using software such as PLINK[108] to prune markers that are in high LD with each other. Another, less-principled approach, is to simply break the genome into windows on the scale of LD decay and take one site per window.

In contrast to the methods above, several methods assume that there is no recombination between sites (for example, LAMARC[93]). In this case, there is a trade-off between choosing sufficiently large genomic segments such that there are enough variable sites to be informative and selecting segments that are small enough so that they are not very affected by recombination. At a minimum, analysed data ought to pass the four-gamete test (that is, ensuring that only three of the four possible haplotypes are present for any pair of sites) to rule out regions that have been too strongly disrupted by recombination.

More generally, because the models of demographic history can be incredibly complex and the likelihood surfaces are potentially multimodal, it is important to perform inference in a careful manner. First, we recommend building up from simpler to more-complicated models. For instance, fitting a model to two populations before attempting to fit a three-population model, and using the two-population model fit as a guide for the three-population model fit. Although this 'greedy' approach, in which one sequentially selects more complicated models, may result in a suboptimal model fit, it simplifies the model-fitting procedure considerably by reducing the parameter space. Another critical step is to ensure that that the maximum likelihood estimate of the parameters is obtained from multiple, random initial parameter guesses. This will help to ensure that the maximum likelihood estimate obtained is a global maximum, rather than a local maximum.

To gain a more-quantitative understanding of population history, the Poisson random field (PRF) framework is used[55]. The PRF assumes that new mutations enter the population as a Poisson process, and that each new mutation is completely unlinked to any currently existing mutation. Under these assumptions, the counts of derived alleles at each frequency follow a Poisson distribution with a mean given by averaging a binomial sample over the underlying population frequency spectrum. The form of the population frequency spectrum is determined by the demographic and selective history under consideration. It is also possible to simply consider relative proportions of sites at each frequency, which follows a multinomial distribution. Therefore, it is possible to calculate the likelihood of the observed data and to use the principle of maximum likelihood to find the parameters that best explain the data.

An important caveat to inference based on the SFS is that, under the PRF, some types of demographic history are not statistically identifiable. Statistical identifiability refers to the ability to distinguish data generated under one set of parameters from data generated under a different set of parameters. Thus, if inference is made without a guarantee of identifiability, the underlying true demographic history will not be recovered, even with an infinite amount of data. Nonetheless, for many families of 'reasonable' demographic histories, the demographic model is identifiable from the SFS[56]; unfortunately, the uncertainty in parameter estimates from the SFS is substantially larger than the amount of uncertainty in many problems in classical statistics at a given sample size[57].

Early work looked at the SFS as a way to estimate parameters of natural selection (for example, REF. 58). However, in the past few years the SFS has mainly been used as a way to estimate demographic parameters. Although some earlier methods were able to model simple demographic histories in a single population[21,59], it was not until the release of *dadi* (diffusion approximations for demographic inference)[60] that complicated histories, including up to three populations, could be analysed. These models are specified using user-written Python scripts, and can be relatively complicated, featuring divergence times, admixture events, migration rates and arbitrary population size histories; *dadi* can be used to estimate all relevant parameters.

Despite its power, *dadi* has several limitations. Most critically, because it numerically solves a partial differential equation (PDE) to obtain the population frequency spectrum, it can be computationally intensive to analyse complicated population histories or large sample sizes. Thus, analysis in *dadi* is limited to three populations, and it is recommended to analyse only subsets of individuals when a data set contains a large sample size of individuals[6]. Moreover, the numerical solution of the PDE can be unstable, causing jagged likelihood surfaces, which results in gradient-based optimization methods (such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method included in the *dadi* package) performing poorly. Instead, we recommend using the Nelder–Mead simplex method, which is more robust to small, jagged areas in the likelihood surface. Another approach, still operating in the framework of PDEs but using orthogonal polynomials as opposed to

Poisson process
A stochastic process in which new events occur at a constant rate per unit of time. Often used to model mutation.

Identity by descent

(IBD). Whether a genomic region has descended from an ancestor unchanged. A genomic region in two (or more) individuals is identical by descent if it is inherited from a common ancestor without being broken up by recombination. Some authors require IBD segments to also be identical by state, that is, to also have no mutations in the region.

Identity by state

(IBS). Whether a genomic region has the same sequence as the corresponding region in another individual. A genomic region in two (or more) individuals is identical by state if it contains no mutations that distinguish the two individuals. Note that a region of IBS is not necessarily also identical by descent.

a finite element method, is able to handle histories of up to four populations[61], but is still computationally demanding.

Several other methods exist to infer demography from the SFS that are complementary to *dadi*. To overcome the limitation of only being able to analyse three populations, Excoffier *et al.*[62] developed a method using coalescent simulation to estimate the expected SFS under a given demographic history. This method, which is integrated into the software package fastsimcoal2 (REF. 63), can handle arbitrary numbers of populations and arbitrarily complicated demographic histories. However, it can be computationally intensive because it requires a large number of simulations to obtain a stable estimate of the SFS. Another method that can accommodate histories of more than three populations is Treemix[64]. Treemix does not model the SFS per se, but instead uses the covariance in allele frequency among populations to both learn an underlying population tree *ab initio* and infer admixture events between pairs of populations (FIG. 2). Bhaskar *et al.*[65] proposed a different method in fastNeutrino, which instead computes the theoretical SFS exactly and analytically using coalescent theory. Therefore, this method is amenable to more-rapidly converging gradient-based optimization algorithms. Importantly, fastNeutrino is readily able to scale to large data sets; for example, it was used to analyse a sample of 14,000 individuals of European ancestry[65]. However, fastNeutrino is only capable of estimating parameters for one population at a time. Therefore, it may be confounded by migration and admixture, which tend to inflate estimates of population size.

Box 2 | **Approximate Bayesian computation**

For many problems in population genomic inference, the likelihood function is intractable. Although both Monte Carlo integration and analytical approximations have enabled the computation of nearly exact likelihoods in many cases (see main text), there are still a substantial number of models in which it seems unlikely that an analytical likelihood can be calculated. In these cases, an attractive alternative method is approximate Bayesian computation (ABC). ABC is based on the intuitive idea that models that have a high posterior probability will produce summary statistics that are close to those calculated from the observed data. Several thorough reviews of ABC are available[109–112], and we briefly cover salient features of this method here.

ABC works by replacing the likelihood with approximate rejection sampling. Parameters from simulations that produce summaries that are close to the observed data are retained as approximate draws from the posterior probability distribution. Thus, it is applicable to any data set in which efficient simulation is possible, even if likelihood calculation is not possible. Several methods exist to obtain these approximate posterior probability distributions, including simple rejection sampling, ABC-MCMC (ABC Markov chain Monte Carlo)[113] and particle filters[114]. Many of these approaches are facilitated by the ABCtoolbox package[115].

The primary challenge of ABC is to choose appropriate summary statistics that are informative with regard to the parameters. It is theoretically optimal to use only sufficient statistics that are informative about the parameters of interest. However, it is often not easy to identify sufficient statistics. Thus, many analyses simply opt to use as many summary statistics as possible. However, one is then faced with the problem of dimensionality: the probability that any particular simulation matches all the summary statistics will be very low. A common solution to this problem is to use partial least squares regression to weight summary statistics by their relevance[113], although there are novel approaches that may provide better performance[116].

Methods based on haplotype data. Although allele frequencies can be used to make powerful inferences, it is ultimately desirable to make full use of the information contained in patterns of linkage between sites. However, modelling a recombining genome is substantially more difficult than modelling unlinked SNVs, as done with the PRF model. Using coalescent theory, it is possible to associate a genealogy with each position in the genome, with recombination causing the genealogy to change along the genome. Demography influences coalescence times, which in turn influence the patterns of mutations seen in the data (FIG. 3). For instance, during periods of small effective population sizes, more coalescences will occur, whereas at times of large effective size, there will be fewer coalescences. Haplotypes that coalesce more anciently will contain more mutations; therefore, the pattern of mutations encodes information about the underlying genealogy and also the demographic history. However, because the full coalescent with recombination includes long-range correlations between sites, the state space of genealogies compatible with the sample becomes extremely large[66,67]. Several methods attempt to circumvent this issue by considering summaries of haplotype diversity; specifically, a number of methods attempt to fit the distribution of regions of identity by descent (IBD) or identity by state (IBS).

A region is identical by descent between two individuals if it was inherited from a common ancestor without being disrupted by recombination. Because this is not directly observable from genetic data, methods that use patterns of IBD sharing require the data to be preprocessed by IBD-detection software, such as Beagle[68] or GERMLINE (genetic error-tolerant regional matching with linear-time extension)[69]. Ralph and Coop[70] used patterns of IBD sharing to obtain a qualitative understanding of relatedness in a large cohort of individuals of European ancestry. In a more-explicitly model-based framework, DoRIS (demographic reconstruction via IBD sharing)[71] uses coalescent theory to predict the distribution of lengths of IBD tracts within a single population (see also REF. 72 for an extension to multiple populations). One strength of this approach is that there is little need to model complicated recombination structures, because IBD segments are assumed to not be broken up by recombination. However, ascertaining IBD segments can be quite challenging, especially for small segments that correspond to more-ancient historical events[70]. Thus, great care needs to be taken during the preprocessing step of IBD detection before these methods are applied. A specific and important recommendation is to use only IBD blocks that are longer than a certain map length. At a minimum, IBD blocks used for parametric inference should be no shorter than 1 cM, and ideally no shorter than 2 cM, as recommended by the authors of DoRIS, and supported by power and false-positive analyses[73]. Most approaches that perform inference based on IBD blocks explicitly condition IBD tract size to be larger than a specific cut-off when calculating the likelihood; thus, using a minimum length cut-off will not result in biased inferences.

SMC'
A modification to the sequential Markov coalescent (SMC) that allows for hidden recombination events that do not change the local genealogy.

When the sequences under consideration may have experienced one or more recombination events, it becomes necessary to average over the possible recombination histories that may have shaped the observed haplotypes. Inspired by the work of Wiuf and Hein[67], who showed how to model the coalescent with recombination as a stochastic process along a DNA sequence, McVean and Cardin[74] introduced the sequentially Markov coalescent (SMC) to make calculations with recombination simpler. In brief, the SMC approximates the full coalescent with recombination by assuming that, when a recombination event occurs, the genealogy at the site to the right of the recombination event depends only on the genealogy at the site to the left of the recombination event. This assumption eliminates long-range correlations in genealogies that generally have very little effect on the data. A modified SMC, called the SMC' (REF. 75), increases the accuracy of the approximation substantially; most current inference strategies make use of the SMC'.

Using the SMC' approximation, Harris and Nielsen[76] developed a method similar to DoRIS that fits the distribution of IBS lengths to infer demographic history. In contrast to inference based on IBD segments, IBS tracts are, in principle, directly observable in the data. However, sequencing errors and missing data can make calling IBS tracts more difficult than naively expected. Nevertheless, this method has been used to analyse data from a diverse range of species, including humans and polar bears[77].

Li and Durbin[78] introduced one of the most popular methods that leverage the SMC to perform demographic inference, which is called pairwise SMC (PSMC). PSMC is directly applicable to whole-genome data from a single diploid individual without the need for phasing. Moreover, this method is capable of averaging over missing data, which helps it deal with the fact that many regions of the genome present difficulties for read mapping owing to repetitive elements and structural variation. PSMC is a HMM that moves along the sequence,

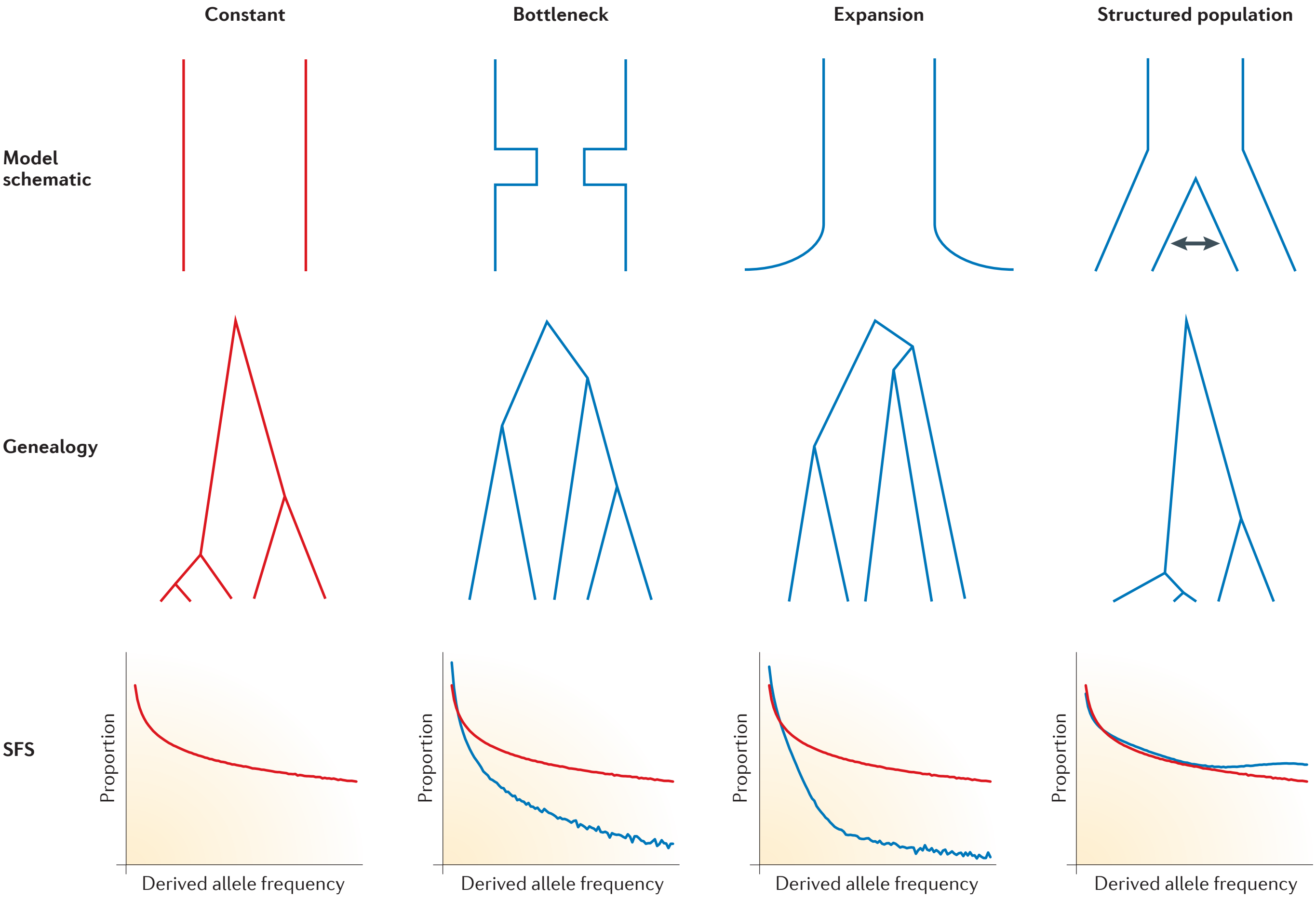

Figure 3 | **The effect of demographic perturbations on gene genealogies and the SFS.** Four simple population demographic models are shown: constant model, bottleneck model, expansion model and structured population model. Below each model schematic, we show average gene genealogies from five sampled lineages obtained by coalescent simulations and stylized site frequency spectrum (SFS; plotted on a logarithmic scale) generated from each model. The SFS from the constant-sized population model is shown in red on each subsequent plot to facilitate comparison among models. Demographic events influence the shape and structure of the genealogies, which in turn influence patterns of genetic variation, such as the SFS. Many popular methods leverage the SFS for inferring population demographic history. The double-ended arrow indicates bidirectional migration.

Box 3 | **Leveraging ancient DNA**

Ancient DNA (aDNA) gives us a window into the past that is normally unavailable when using data derived from contemporary individuals. Substantially different models of demography and selection can be consistent with a given set of modern DNA observations, and aDNA can be used to differentiate between the alternatives[117]. Because obtaining high-quality aDNA sequences is technically challenging and expensive, aDNA is best collected with the goal of testing specific hypotheses. However, aDNA studies frequently reveal unanticipated aspects of history.

Perhaps the quintessential example of aDNA providing insights into human demographic history is the revelation of Neanderthal admixture in modern humans[118]. Although some believed there was evidence of archaic admixture before the publication of the Neanderthal genome[119], direct comparison of human and Neanderthal DNA delivered conclusive proof of admixture (but see REF. 120 for a possible alternative hypothesis).

In addition to revealing demographic history, aDNA has been instrumental in understanding the effect of selection in humans. Motivated by the evidence of strong selection at the lactase locus in modern Europeans, several groups obtained targeted aDNA from the lactase locus in ancient Europeans[121–125]. Supporting the hypothesis of strong positive selection, the lactase persistence allele is found to be mostly absent in ancient Europeans. These and other observations have motivated the development of several methods to analyse aDNA time series to estimate selection coefficients[126–129].

A promising new development for using aDNA in population genomic analyses is the development of sequence capture. Using a combination of sequence capture and other enrichment techniques, numerous ancient genomes have been analysed across Eurasia, promising increasingly fine-scale information about history[130].

inferring the time of the most-recent common ancestor between the two haplotypes that make up the diploid sequence. Hence, PSMC can be used to infer the history of effective population size for that sample. Importantly, PSMC infers demographic history in a relatively non-parametric way compared to more-explicitly parametric models (such as exponential growth), by assuming a piecewise constant demographic history. This approach dates back to the introduction of Bayesian skyline plots, which were initially developed in the context of non-recombining haplotype data[79–81].

PSMC was extended to be able to handle data from multiple individuals, in an approach called multiple SMC (MSMC)[82]. In contrast to PSMC, MSMC uses the SMC' approximation, which increases its accuracy. When analysing multiple individuals, MSMC requires phased data; however, it is able to operate on a single, unphased diploid genome in the same way as PSMC. Because MSMC can be used to model haplotypes from individuals of different populations, it is able to estimate a proxy for the migration rate between populations. However, this parameter, called the relative cross coalescence rate, should not be interpreted as a direct estimate of the migration rate, because it simply measures the fraction of between-population coalescences compared with within-population coalescences.

Several other HMM-based approaches to inferring demographic history from full-genome data exist. In parallel to the development of the SMC and PSMC, Hobolth *et al.* (REF. 83) developed the coalescent HMM (CoalHMM) framework (see also REFS 84–87). This software is primarily focused on making comparisons between species and was essential in understanding the effect of incomplete lineage sorting between humans, chimpanzees and gorillas[88]. More recently, software called demographic inference using composite approximate likelihood (diCal)[89] was released, building on the theory of conditionally sampled alleles[90] using a modification of the approach developed by Paul *et al.*[91]. Early versions of diCal could be used to model multiple sampled haplotypes within a given population for estimating population size histories (in a similar way to PSMC). Subsequent developments (for example, REF. 92) have incorporated migration and allowed for inference of models with more than one population.

A substantial drawback of these haplotype-based methods is that they almost always require some form of data preprocessing. As mentioned, any method that requires detection of IBD segments can only be as good as the quality of the detected IBD segments allows. Similarly, all methods except PSMC (or MSMC in its PSMC mode) require data to be phased. Phasing requires large reference populations that are closely related to the samples of interest. Although the 1,000 Genomes Project provides these data for many world populations, as interest in the demography of more isolated populations grows, it may become difficult to phase individuals accurately.

Conclusions and future perspectives

The ability to obtain genome-scale data from multiple individuals in a population has created the opportunity to infer human demographic history with unprecedented resolution and accuracy. To make use of all of these data, numerous innovative and sophisticated methods have been developed to infer population structure and demographic history. Nonetheless, the strengths and limitations of methods should be carefully considered, and care needs to be taken to ensure that they are used properly and with the maximum power.

Models used for inference are becoming increasingly complicated and realistic, although they have not yet met the 'gold standard' of obtaining the full likelihood of genetic data given a demographic history. Although some models, such as LAMARC (likelihood analysis with metropolis algorithm using random coalescence)[93], BEAST (Bayesian evolutionary analysis by sampling trees)[94], MCMCcoal[95], and G-PhoCS (generalized phylogenetic coalescent sampler)[96] are capable of using Monte Carlo methods to average over the underlying genealogies, it is likely to be impossible to derive a closed-form expression for the likelihood with

Conditionally sampled alleles
Alleles that are sampled from a population given that a set of reference alleles is already in hand.

fully general demographic models. Future work in this area may be profitably aimed towards deriving exact likelihoods under simple models[97,98] or trying to make 'demography-free' inferences about the underlying genealogy, which would subsequently be used to make inferences about demography[99]. Furthermore, continued work is necessary to obtain accurate estimates of mutation rates, which are necessary to turn population genetic estimates of dates into units of generations or years (see Supplementary information S1 (box)), and we anticipate that, as sequencing technology improves, estimates of the *de novo* mutation rate derived from sequencing pedigrees[100] will have profound implications for inferences of human history.

Furthermore, as sample sizes become larger, specific details of the breeding structure in a population may begin to influence inferences, requiring the development of highly detailed and complicated models[101,102]. As a specific example, although the coalescent is extremely robust[103], certain kinds of breeding structure result in genealogies that are not compatible with the standard coalescent[104,105]. In Supplementary information S2 (box), we briefly discuss lambda coalescents, which allow for multiple merges in genealogical trees. Multiple merges may occur when single individuals contribute disproportionately to future generations, such as may be the case with Genghis Khan[106].

Finally, it is important to acknowledge that there are limits to the resolution afforded by genetic data. For example, because coalescent events are fundamentally stochastic and can never be directly observed, even if we were able to obtain genetic data from the entire human population we would not be able to obtain a perfect picture of human history. Although ancient DNA may be helpful in this regard (BOX 3), we believe that it is only through the continued synthesis of data and knowledge across disciplines that a more-complete story of human history will emerge.

1. Veeramah, K. R. & Hammer, M. F. The impact of whole-genome sequencing on the reconstruction of human population history. *Nat. Rev. Genet.* **15**, 149–162 (2014).
2. Metzker, M. L. Sequencing technologies — the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
3. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
This study describes an international project that created one of the most-comprehensive catalogues of sequence variation in geographically diverse populations.
4. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
This article represents one of the earliest large-scale, high-coverage exome data sets to be produced; it has been extensively used in evolutionary and medical genomics.
5. Bustamante, C. D., De La Vega, F. M. & Burchard, E. G. Genomics for the world. *Nature* **475**, 163–165 (2011).
6. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl Acad. Sci. USA* **108**, 11983–11988 (2011).
7. Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
8. Leslie, S. *et al.* The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
9. Nielsen, R. Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218 (2005).
10. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312**, 1614–1620 (2006).
11. Bamshad, M. & Wooding, S. P. Signatures of natural selection in the human genome. *Nat. Rev. Genet.* **4**, 99–111 (2003).
12. Akey, J. M. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* **19**, 711–722 (2009).
13. Fu, W. & Akey, J. M. Selection and adaptation in the human genome. *Annu. Rev. Genom. Hum. Genet.* **14**, 467–489 (2013).
14. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
15. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
16. Auwera, G. A. *et al.* From fastQ data to high-confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **15**, 1110 (2013).
17. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).
18. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
19. Schraiber, J. G., Shih, S. & Slatkin, M. Genomic tests of variation in inbreeding among individuals and among chromosomes. *Genetics* **192**, 1477–1482 (2012).
20. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067 (2009).
21. Williamson, S. H. *et al.* Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl Acad. Sci. USA* **102**, 7882–7887 (2005).
This study reports a clever approach to account for the effects of selection when making demographic inferences.
22. Živkovic, D., Steinrücken, M., Song, Y. S. & Stephan, W. Transition densities and sample frequency spectra of diffusion processes with selection and variable population size. *Genetics* **200**, 601–617 (2015).
23. Hammer, M. F. *et al.* The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. *Nat. Genet.* **42**, 830–831 (2010).
24. Gottipati, S., Arbiza, L., Siepel, A., Clark, A. G. & Keinan, A. Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. *Nat. Genet.* **43**, 741–743 (2011).
25. Gazave, E. *et al.* Neutral genomic regions refine models of recent rapid human population growth. *Proc. Natl Acad. Sci. USA* **111**, 757–762 (2014).
This study illustrates well how choosing neutral genomic regions carefully can lead to more-refined estimates of demographic parameters.
26. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
27. Romanoski, C. E., Glass, C. K., Stunnenberg, H. G., Wilson, L. & Almouzni, G. Epigenomics: roadmap for regulation. *Nature* **518**, 314–316 (2015).
28. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
29. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
30. Pollard, K. S. *et al.* Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* **2**, e168 (2006).
31. Arbiza, L., Zhong, E. & Keinan, A. NRE: a tool for exploring neutral loci in the human genome. *BMC Bioinformatics* **13**, 301 (2012).
32. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
This classic paper describes a nonparametric approach for inferring population structure.
33. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History And Geography Of Human Genes* (Princeton Univ. Press, 1994).
34. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
35. Biswas, S., Scheinfeldt, L. B. & Akey, J. M. Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. *Am. J. Hum. Genet.* **84**, 641–650 (2009).
36. McVean, G. A. Genealogical interpretation of principal components analysis. *PLoS Genet.* **5**, e1000686 (2009).
37. François, O. *et al.* Principal component analysis under population genetic models of range expansion and admixture. *Mol. Biol. Evol.* **27**, 1257–1268 (2010).
38. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40**, 646–649 (2008).
39. Yang, W.-Y., Novembre, J., Eskin, E. & Halperin, E. A model-based approach for analysis of spatial structure in genetic data. *Nat. Genet.* **44**, 725–731 (2012).
40. Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* **28**, 289–301 (2005).
41. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
42. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).
43. Huelsenbeck, J. P. & Andolfatto, P. Inference of population structure under a Dirichlet process model. *Genetics* **175**, 1787–1802 (2007).
44. Xie, W., Lewis, P. O., Fan, Y., Kuo, L. & Chen, M.-H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60**, 150–160 (2010).
45. Patterson, N. *et al.* Methods for high-density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74**, 979–1000 (2004).
46. Gravel, S. Population genetics models of local ancestry. *Genetics* **191**, 607–619 (2012).
47. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
48. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5**, e1000519 (2009).
49. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
50. Pool, J. E. & Nielsen, R. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711–719 (2009).

51. Liang, M. & Nielsen, R. The lengths of admixture tracts. *Genetics* **197**, 953–967 (2014).
52. Sankararaman, S., Sridhar, S., Kimmel, G. & Halperin, E. Estimating local ancestry in admixed populations. *Am. J. Hum. Genet.* **82**, 290–303 (2008).
53. Brisbin, A. *et al.* PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Hum. Biol.* **84**, 343–364 (2012).
54. Wakeley, J. *Coalescent Theory: An Introduction* (Robert & Co., 2009).
55. Sawyer, S. A. & Hartl, D. L. Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176 (1992).
56. Bhaskar, A. & Song, Y. S. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Ann. Statist.* **42**, 2469–2493 (2014).
57. Terhorst, J. & Song, Y. S. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proc. Natl Acad. Sci. USA* **112**, 7677–7682 (2015).
58. Bustamante, C. D., Wakeley, J., Sawyer, S. & Hartl, D. L. Directional selection and the site-frequency spectrum. *Genetics* **159**, 1779–1788 (2001).
59. Evans, S. N., Shvets, Y. & Slatkin, M. Non-equilibrium theory of the allele frequency spectrum. *Theor. Popul. Biol.* **71**, 109–119 (2007).
60. 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).
61. Lukic, 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).
62. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
63. Excoffier, L. & Foll, M. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**, 1332–1334 (2011).
64. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* **8**, e1002967 (2012).
65. Bhaskar, A., Wang, Y. & Song, Y. S. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Res.* **25**, 268–279 (2014).
66. Griffiths, R. C. & Marjoram, P. An ancestral recombination graph. *University of Canterbury* [online], http://www.math.canterbury.ac.nz/~r.sainudiin/recomb/ima.pdf (1997).
67. Wiuf, C. & Hein, J. Recombination as a point process along sequences. *Theor. Popul. Biol.* **55**, 248–259 (1999).
68. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
69. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19**, 318–326 (2009).
70. Ralph, P. & Coop, G. The geography of recent genetic ancestry across Europe. *PLoS Biol.* **11**, e1001555 (2013).
71. Palamara, P. F., 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).
72. Palamara, P. F. & Pe'er, I. Inference of historical migration rates via haplotype sharing. *Bioinformatics* **29**, i180–i188 (2013).
73. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
74. McVean, G. A. T. & Cardin, N. J. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 1387–1393 (2005).
This article introduces the SMC, which enabled important developments in population genomic inferencing from recombining sequences.
75. Marjoram, P. & Wall, J. D. Fast 'coalescent' simulation. *BMC Genet.* **7**, 16 (2006).
76. Harris, K. & Nielsen, R. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet.* **9**, e1003521 (2013).
77. Liu, S. *et al.* Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell* **157**, 785–794 (2014).
78. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
This study describes PSMC, which enables quasi-non-parametric inferencing of effective population size through time from a single diploid genome sequence.
79. Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192 (2005).
80. Heled, J. & Drummond, A. J. Bayesian inference of population size history from multiple loci. *BMC Evol. Biol.* **8**, 289 (2008).
This study details one of the first, and underappreciated, methods to infer population size history in a relatively non-parametric way from haplotype data.
81. Minin, V. N., Bloomquist, E. W. & Suchard, M. A. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25**, 1459–1471 (2008).
82. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
83. 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).
84. Dutheil, J. Y. *et al.* Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics* **183**, 259–274 (2009).
85. 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).
86. Hobolth, A., Dutheil, J. Y., Hawks, J., Schierup, M. H. & Mailund, T. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. *Genome Res.* **21**, 349–356 (2011).
87. 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).
88. Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169–175 (2012).
89. Sheehan, S., Harris, K. & Song, Y. S. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* **194**, 647–662 (2013).
90. Stephens, M. & Donnelly, P. Inference in molecular population genetics. *J. R. Stat. Soc. B* **62**, 605–655 (2000).
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).
92. Steinrücken, M., Paul, J. S. & Song, Y. S. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theor. Popul. Biol.* **87**, 51–61 (2013).
93. Kuhner, M. K. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **22**, 768–770 (2006).
94. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
95. Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645–1656 (2003).
96. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034 (2011).
97. Lohse, K., Harrison, R. J. & Barton, N. H. A general method for calculating likelihoods under the coalescent process. *Genetics* **189**, 977–987 (2011).
98. Lohse, K. & Frantz, L. A. F. Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. *Genetics* **196**, 1241–1251 (2014).
99. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* **10**, e1004342 (2014).
100. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* **15**, 47–70 (2014).
This review covers in great detail the recent controversy about the human genomic mutation rate and summarizes the different kinds of mutations in the human genome.
101. Bhaskar, A., Clark, A. G. & Song, Y. S. Distortion of genealogical properties when the sample is very large. *Proc. Natl Acad. Sci. USA* **111**, 2385–2390 (2014).
102. Wakeley, J., King, L., Low, B. S. & Ramachandran, S. Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics* **190**, 1433–1445 (2012).
103. Möhle, M. Robustness results for the coalescent. *J. Appl. Probab.* **35**, 438–447 (1998).
This important theory paper outlines the broad generality of the Kingman coalescent.
104. Pitman, J. Coalescents with multiple collisions. *Ann. Appl. Probab.* **27**, 1870–1902 (1999).
105. Sagitov, S. The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* **36**, 1116–1125 (1999).
106. Zerjal, T. *et al.* The genetic legacy of the Mongols. *Am. J. Hum. Genet.* **72**, 717–721 (2003).
107. Varin, C., Reid, N. & Firth, D. An overview of composite likelihood methods. *Statist. Sin.* **21**, 5–42 (2011).
108. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
109. Beaumont, M. A. Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* **41**, 379–406 (2010).
110. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
111. Sunnåker, M. *et al.* Approximate Bayesian computation. *PLoS Comput. Biol.* **9**, e1002803 (2013).
112. Csilléry, K., Blum, M. G. B., Gaggiotti, O. E. & François, O. Approximate Bayesian Computation (ABC) in practice. *Trends Ecol. Evol.* **25**, 410–418 (2010).
113. Wegmann, D., Leuenberger, C. & Excoffier, L. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182**, 1207–1218 (2009).
114. Sisson, S. A., Fan, Y. & Tanaka, M. M. Sequential Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. USA* **104**, 1760–1765 (2007).
115. Wegmann, D., Leuenberger, C., Neuenschwander, S. & Excoffier, L. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* **11**, 116 (2010).
116. Fearnhead, P. & Prangle, D. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc.* **74**, 419–474 (2012).
117. Pickrell, J. K. & Reich, D. Toward a new history and geography of human genes informed by ancient DNA. *Trends Genet.* **30**, 377–389 (2014).
118. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
119. Plagnol, V. & Wall, J. D. Possible ancestral structure in human populations. *PLoS Genet.* **2**, e105 (2006).
120. 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. USA* **109**, 13956–13960 (2012).
121. Burger, J., Kirchner, M., Bramanti, B., Haak, W. & Thomas, M. G. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proc. Natl Acad. Sci. USA* **104**, 3736–3741 (2007).
122. Malmström, H. *et al.* in *Migration in Prehistory: DNA and Stable Isotope Analysis of Swedish Skeletal Material* (ed. Linderholm, A.) (Stockholm University, 2008).
123. Malmström, H. *et al.* High frequency of lactose intolerance in a prehistoric hunter-gatherer population in northern Europe. *BMC Evol. Biol.* **10**, 89 (2010).

124. Lacan, M. *et al.* Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. *Proc. Natl Acad. Sci. USA* **108**, 9788–9791 (2011).
125. Plantinga, T. S. *et al.* Low prevalence of lactase persistence in Neolithic South-West Europe. *Eur. J. Hum. Genet.* **20**, 778–782 (2012).
126. Bollback, J. P., York, T. L. & Nielsen, R. Estimation of $2N_e s$ from temporal allele frequency data. *Genetics* **179**, 497–502 (2008).
127. Malaspinas, A.-S., Malaspinas, O., Evans, S. N. & Slatkin, M. Estimating allele age and selection coefficient from time-serial data. *Genetics* **192**, 599–607 (2012).
128. Mathieson, I. & McVean, G. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics* **193**, 973–984 (2013).
129. Steinrücken, M., Bhaskar, A. & Song, Y. S. A novel spectral method for inferring general diploid selection from time series genetic data. *Ann. Appl. Statist.* **8**, 2203–2222 (2014).
130. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
131. Yang, Z. & Rannala, B. Bayesian species delimitation using multilocus sequence data. *Proc. Natl Acad. Sci. USA* **107**, 9264–9269 (2010).

Acknowledgements
The author would like to acknowledge Kelley Harris for helpful discussions regarding SMC-based approaches.

Competing interests statement
The authors declare competing interests: see Web version for details.

SUPPLEMENTARY INFORMATION
See online article: S1 (box) | S2 (box)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF