An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes

Yann X C Bourgeois, Ben H Warren

Received: 30 December 2020 | Revised: 26 April 2021 | Accepted: 11 May 2021

DOI: 10.1111/mec.15989

SPECIAL ISSUE

MOLECULAR ECOLOGY WILEY

An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes

Yann X. C. Bourgeois[1] | Ben H. Warren[2]

[1]School of Biological Sciences, University of Portsmouth, Portsmouth, UK

[2]Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, CNRS, Sorbonne Université, EPHE, UA, CP 51, Paris, France

Correspondence
Yann X. C. Bourgeois, School of Biological Sciences, University of Portsmouth, King Henry Building, Portsmouth, UK.
Email: yann.bourgeois@port.ac.uk

Abstract

Characterizing the population history of a species and identifying loci underlying local adaptation is crucial in functional ecology, evolutionary biology, conservation and agronomy. The constant improvement of high-throughput sequencing techniques has facilitated the production of whole genome data in a wide range of species. Population genomics now provides tools to better integrate selection into a historical framework, and take into account selection when reconstructing demographic history. However, this improvement has come with a profusion of analytical tools that can confuse and discourage users. Such confusion limits the amount of information effectively retrieved from complex genomic data sets, and impairs the diffusion of the most recent analytical tools into fields such as conservation biology. It may also lead to redundancy among methods. To address these isssues, we propose an overview of more than 100 state-of-the-art methods that can deal with whole genome data. We summarize the strategies they use to infer demographic history and selection, and discuss some of their limitations. A website listing these methods is available at www.methodspopgen.com.

KEYWORDS
bioinformatics, demography, population genomics, selection, whole-genome sequencing

1 | INTRODUCTION

Comprehensive analyses of species history and selection contribute to our understanding of causation in biology, an effort that has included genetics, developmental science and ecology (Laland et al., 2011). The number of population genomic studies aimed at elucidating the history of natural populations has increased enormously in the last 10 years. A few examples include an improved understanding of the history of human migrations, admixture and adaptation (e.g., Abi-Rached et al., 2011; Li & Durbin, 2011; Sabeti et al., 2002), the origin of domesticated species (e.g., Axelsson et al., 2013; Cubry et al., 2018; Schubert et al., 2014) and the genetic basis of local adaptation (e.g., Kolaczkowski et al., 2011; Kubota et al., 2015; Legrand et al., 2009; Roux et al., 2013). Developments in whole-genome resequencing have continually improved the throughput of genetic data, while reducing the time and cost of their production. Increased data production has been accompanied by a drive to develop efficient computational methods to interpret patterns of genetic variation at the genomic scale. These interconnected developments have allowed species histories to be inferred even when little preliminary knowledge is available. Investigating variation across multiple genomes sampled across populations or closely related species is now a common task for teams studying evolutionary processes, who can rely on a diverse array of methods to infer demography and selection. Such progress has confirmed the value of population genomics in understanding biological diversity, beyond the initial

handful of model species upon which most of the field was built (Abzhanov et al., 2008; Ellegren et al., 2012; Jenner & Wills, 2007; Mandoli & Olmstead, 2000; Poelstra et al., 2014; Weber et al., 2013; White et al., 2010). Such advances are needed to broaden our view of evolutionary processes and improve sampling of distant clades. Ultimately, this should provide a more balanced picture than the one brought by the study of a few model species (Abzhanov et al., 2008). From an applied perspective, genomic approaches also have the potential to improve conservation genetic inference by scaling up the amount of data available (Shafer et al., 2015), understanding the past history of species (Leitwein et al., 2020), and identifying loci and alleles important for local adaptation, which can then be used to define relevant conservation units (Fraser & Bernatchez, 2001).

Much effort has recently been made in facilitating the dissemination of sometimes complex, state-of-the-art methods. Nevertheless, the last comprehensive review of methods for population genetics was performed more than 10 years ago (Excoffier & Heckel, 2006). Recent methodological advances have brought increased analytical complexity to the field, and an inflation in the number of methods covering any one topic. The widespread use of sophisticated analytical tools is made difficult by the lack of communication between fields (Shafer et al., 2015), little user-friendliness of software, inflation of data formats (Lischer & Excoffier, 2012) and the ever-increasing number of methods made available. As a consequence, it has become increasingly difficult for all potential users (and also developers) to follow developments and be sure of selecting the most appropriate method for the question and data at hand. Combining approaches is one of the current grand challenges in evolutionary biology (Cushman, 2014). While large-scale collaborations and sharing of skills between researchers allow for detailed analyses, a global summary of methods that can handle whole-genome resequencing data would be valuable for smaller research teams, so they can quickly start new projects and evaluate their experimental design. It would also facilitate communication between different subfields of evolutionary biology, by providing a common resource that can be used to identify methodological convergence and possible synergies. It may also avoid situations where similar methods are developed in parallel. Furthermore, the issue of anthropogenic environmental change and decline in biodiversity is pressing, and merits enhanced efforts to disseminate methods that can leverage genomic data, ultimately improving our understanding of the response of biodiversity to environmental change. Many conservation practitioners are receptive to using genetic tools, but do not always have access to the relevant expertise (Taylor et al., 2017). A freely accessible methodological summary may be useful in this context.

In this review, we assume that the reader is already familiar with the main concepts and current issues in population genomics, but needs an overview of the different methods associated with these concepts. We promote the idea that multiple approaches must be used and compared in any population genomics project. This has several benefits: it gives the investigator a better idea of the robustness of results and may reveal issues in raw data processing. In addition, different methods aim to detect slightly different signals, and their combination may provide a more comprehensive overview of the processes acting. We aim at providing a resource that, if not fully comprehensive, can act as an efficient starting point for researchers investigating whole-genome variation in the next few years. This article can be used in combination with other recent methodological reviews on selection (Haasl & Payseur, 2016; Koropoulis et al., 2020), demographic inference or simulation-based approaches (Schrider & Kern, 2018; Smith & Flaxman, 2020).

For the sake of simplicity, we divide our review into two sections (Figure 1): (i) methods devoted to the study of population structure and history (Tables 1 and 2), and (ii) detecting signatures of evolutionary processes along the genome (Tables 3 and 4). We end this review by outlining how different analyses can be combined, and present future directions that may be taken by the field of population genomics. We particularly insist on the interest—but also the challenges—of model-based approaches to test specific hypotheses, benchmark different methods and incorporate intrinsic properties of genomes (Table 4). The tables and a summary of the methods discussed in this paper will be kept updated to follow improvements, and are available at www.methodspopgen.com. Contributions are of course welcome, and can be sent to the following email address: methodspopgen@gmail.com.

2 | POPULATION STRUCTURE AND HISTORY

Genetic diversity and its genome-wide variance are directly impacted by variation in many factors including effective population sizes, population structure, inbreeding and migration. Moreover, the effects of selection on diversity at linked sites depends directly on local variation in the recombination rate. All these factors are important to characterize in any study of genome-wide variation. In this section, we describe methods aiming to quantify these aspects (see also Tables 1 and 2).

2.1 | Estimating familial relationships and reconstructing pedigrees

Understanding relatedness and structure both within and between populations is an important starting point for any study making inferences of selection or demographic history. Methods for estimating the relatedness of individuals are suited to studies relying on pedigree information (for example in quantitative genetics studies), or if there are reasons to suspect that familial relationships and inbreeding can play a major role in shaping the genetic structure of the population(s) considered. The most powerful methods in this category are likelihood-based and make use of heterozygous sites in each individual (e.g., COLONY in Wang, 2019, see also the detailed review in Huisman, 2017). Each pedigree configuration can be assigned a likelihood at each locus which depends on the probability of observing a given genotype conditional on the genotypes of assigned parents. Assuming independent loci, a composite likelihood can then be derived for a set of unlinked single nucleotide

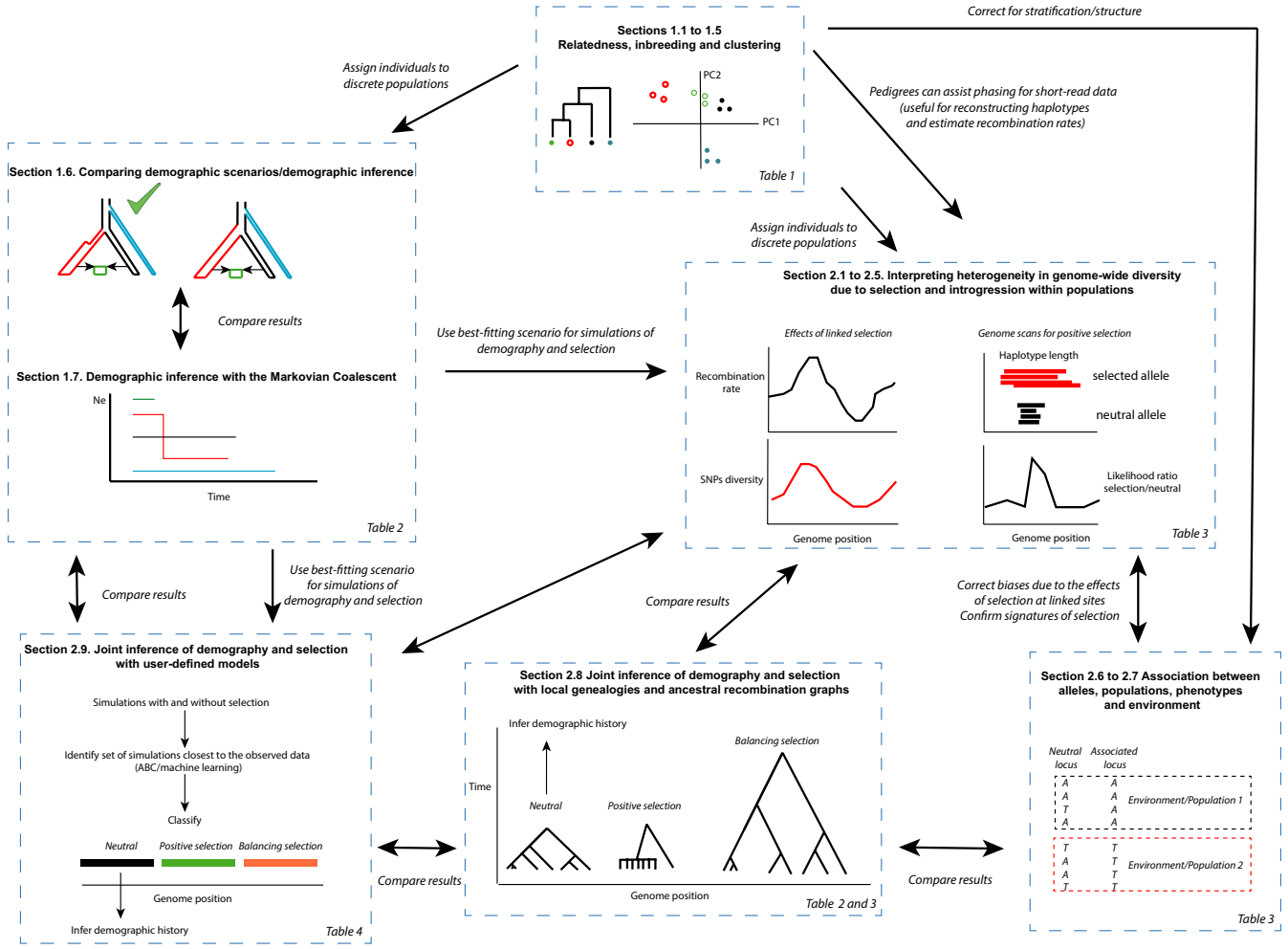

FIGURE 1 Graphical summary of this review. This work is divided into two main sections: the first section covers methods that generally assume neutrality and are generally used for demographic inference. The second section covers methods that aim at identifying loci under the direct and indirect effects of selection. The table listing the relevant methods is indicated at the bottom right of each box. The linear structure of this review does not necessarily reflect the network of possible comparisons between the results of different methods. These possible comparisons are indicated by double arrows. Results obtained from methods listed in different sections can be used to inform the next steps of an analysis (single arrows). We acknowledge that there is no one-size-fits-all pipeline, and elements of this general framework may be entirely omitted from an analysis depending on the research question

polymorphisms (SNPs). Information about pedigrees is important in order to filter out related individuals before carrying other population genetics analyses. Furthermore, mendelian constraints provide important information about haplotypes that can be used by phasing programs. Including related individuals can be useful when attempting to phase genotypes and generate a reference panel for further phasing in unrelated samples. The popular phasing algorithm Shapeit (Delaneau et al., 2019; Williams et al., 2012) can include familial information when reconstructing phased haplotypes.

2.2 | Using nonsupervised models to estimate relatedness and population structure

An elegant and efficient class of methods relies on using multivariate approaches such as principal component analysis (PCA) to infer relatedness between individuals and populations without a priori knowledge. These methods apply a dimension reduction procedure to matrices of individual genotypes, projecting genotypic variability along several axes of variation (Jombart et al., 2009). These approaches have been especially useful to study the consistency between geographical and genetic structure in human populations of Europe (Novembre et al., 2008). Procrustes rotation (Novembre et al., 2008) can be used to match geographical coordinates with PCA axes, showing how isolation by distance has shaped genetic structure. Since these methods do not have underlying assumptions based on (diploid) population genetics, they are suitable for analysing species displaying polyploidy or mixed-ploidy (Dufresne et al., 2014). They go beyond a mere description of data, since projections of individuals on PCA axes can be used to infer admixture proportions, and contain information about demographic processes shaping genetic diversity (McVean, 2009). PCA can be used as a summary of genetic

TABLE 1 Summary of methods dedicated to data description and assessing population structure. VCF: variant call format (see Danecek et al., 2011)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
SNMF	Clustering and characterizing admixture	Grouping individuals in clusters maximizing Hardy–Weinberg (HW) equilibrium and LD between loci	Fast (30× than ADMIXTURE)	Still slow computation time for very large data sets	http://membres-timc.imag.fr/Olivier.Francois/snmf/index.htm	Frichot et al. (2014)
STRUCTURE	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	User-friendly interface. Bayesian inference	Not suited for large whole genomes. Requires specific input format. Might be used on a small set of high-quality markers for small genomes	http://pritchardlab.stanford.edu/structure.html	Pritchard et al. (2000)
FASTSTRUCTURE	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	~100× faster than STRUCTURE	Approximate inference of the original STRUCTURE model	http://rajanil.github.io/fastStructure/	Raj et al. (2014)
ADMIXTURE	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	Maximum likelihood, faster than STRUCTURE. Can handle sex-linked markers	Often slower than its counterparts	https://www.genetics.ucla.edu/software/admixture/index.html	Alexander and Novembre (2009)
FINESTRUCTURE/GLOBETROTTER	Clustering and characterizing admixture	Chromosome painting, admixture and clustering	Estimates time since admixture, fast, set of scripts to facilitate analysis	Relies on STRUCTURE and FASTSTRUCTURE assumptions. Requires phased data	http://paintmychromosomes.com/	Hellenthal et al. (2014)
PCADMIX	Clustering and characterizing admixture	Chromosome painting	Fast, uses HMM to smooth out windows and limit noise due to low-confidence ancestry	Requires a priori definition of ancestral populations and phased haplotypes	https://sites.google.com/site/pcadmix/	Brisbin et al. (2012)
MOSAIC	Clustering and characterizing admixture	Chromosome painting, estimating admixture time and proportions	Can handle several source populations. These populations do not have to be good surrogates of populations that actually mixed	Requires phased data, but performs phasing error correction	https://maths.ucd.ie/~mst/MOSAIC/	Salter-Townshend and Myers (2019)
TFA	Clustering and characterizing admixture	Summarizing variance across loci and visualizing interindividual genetic distance	Uses latent factors to correct for drift and to position ancient samples in a PCA-like framework.	NA	https://bcm-uga.github.io/tfa/	François and Jay (2020)
CONSTRUCT	Clustering and characterizing admixture	Perform clustering while taking into account isolation by distance	Aims to extend STRUCTURE while avoiding the over-clustering that is produced by isolation by distance	Slow for large data sets	http://www.genescape.org/construct.html	Bradburd et al. (2017)

(Continues)

TABLE 1 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
DYSTRUCT	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	This method explicitly takes into account the age of samples. Useful when analysing mixtures of modern and ancient samples	Requires a genotype matrix in the eigenstrat format. Primarily tested on human data	https://github.com/tyjo/dystruct	Joseph and Pe'er (2019)Joseph and Pe'er (2019)
BEDASSLE	Differentiation and MCMC model testing	Identifies contribution of environment and geographical distance to population differentiation	Less biased than Mantel tests, provides tools for model testing	Uses population-level data.	https://cran.r-project.org/web/packages/BEDASSLE/index.html	Bradburd et al. (2013)
LOSTRUCT	Differentiation/ diversity	Chromosome painting	Performs local PCA along the genome. Identifies regions showing discrepancies with genome-wide structure, as often happens due to inversions	NA	https://github.com/petrelharp/local_pca	
NPSTAT	Differentiation/ diversity	Extracting summary statistics from pooled data	Explicitly corrects for sampling bias in pooled data. Allows computing tests using an outgroup (MK test, HKA test, Fay and Wu's *H*) and characterizing coding mutations	Mostly limited to summary statistics, but more complete than POPOOLATION	https://github.com/lucaferretti/npstat	Ferretti et al. (2013)
POPOOLATION/ POPOOLATION2/ POPOOLATION TE	Differentiation/ diversity/ recombination	Extracting summary statistics from pooled data	Explicitly corrects for sampling bias in pooled data. Can be used to detect TE polymorphisms.	Mostly limited to a few summary statistics. A pipeline dedicated to TE detection is also available	https://sourceforge.net/p/popoolation/wiki/Main/	Kofler, Orozco-terWengel, et al. (2011), Kofler, Pandey, et al. (2011)
POPGENOME	Differentiation/ diversity/ recombination	Computing summary statistics based on AFS and LD along genomes	Accepts VCF and GFF/GFT files, efficient and fast. Tests for admixture available (ABBA–BABA test). Includes basic coalescence simulations (MS and MSMS)	Mostly limited to summary statistics (but coalescent simulations are possible). No built-in SNP calling module	http://catchenlab.life.illinois.edu/stacks/	Pfeifer et al. (2014)
ANGSD	Differentiation/ diversity/ recombination	Computing summary statistics based on AFS and LD along genomes	Able to process BAM files, built-in procedures for data filtering, admixture analysis. Suited for low-depth data. Includes a suite of methods to estimate relatedness (NGSRELATE).	Mostly limited to summary statistics. Tutorials not always up-to-date.	https://github.com/ANGSD/angsd https://github.com/ANGSD/NgsRelate	Korneliussen et al. (2014), Hanghøj et al. (2019)
VCFTOOLS	Differentiation/ diversity/ recombination	Computing summary statistics based on AFS and LD along genomes	Fast. VCFTOOLS can also be used for SNP filtering	Less summary statistics than POPGENOME	https://vcftools.github.io/man_latest.html	Danecek et al. (2011)

(Continues)

TABLE 1 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
ATLAS	Differentiation/ diversity/ recombination	Low-depth sequencing/ancient samples analysis	Particularly suited for analysing ancient samples. Includes sets of tools to call variants, estimate post-mortem damage, inbreeding, genetic diversity. Produces the input file for PSMC (demography from a single diploid genome)	Better used in combination with GATK pipelines. Still in development	https://bitbucket.org/wegmannlab/atlas/wiki/Home	Link et al. (2017)
POPTREE2	Genetic differentiation	Visualizing a matrix of pairwise differentiation statistics as a tree	Can be used for pooled data sets, several statistics can be used	Differentiation measures alone do not necessarily retrieve the actual history of populations	http://www.med.kagawa-u.ac.jp/~genomelb/takezaki/poptree2/index.html	Takezaki et al. (2010)
EEMS	Landscape genomics	Estimating barriers to gene flow in a spatial context	Estimates pairwise relatedness between all samples, and compares it to isolation-by-distance expectations to identify barriers to gene flow and corridors of higher connectivity. Can handle both haploid and diploid data	Requires to convert VCF file into PLINK binary format. Estimates effective migration rates (does not disantangle migration rates and effective population sizes). Setting parameters for the MCMC chain requires some trial-and-error	https://github.com/dipetkov/eems	Petkova et al. (2015)
MAPS	Landscape genomics	Estimating barriers to gene flow in a spatiotemporal context	Expands on EEMS, but takes into account the phase to reconstruct past changes in connectivity. Can disentantle migration rates and effective population sizes (unlike EEMS)	Relies on identity-by-descent tracks, requiring phasing (e.g., using BEAGLE). A pipeline to obtain IBD tracks is available, with a few details here: https://github.com/halasadi/ibd_data_pipeline/issues/1	https://github.com/halasadi/MAPS	Al-Asadi et al. (2019)
TESS3R	Landscape genomics	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	Incorporates geographical information of samples. Can run genome scans of selection based on contrasting ancestral and modern allele frequencies.	Importing data requires using conversion tools found in the LEA suite	https://bcm-uga.github.io/TESS3_encho_sen/	Caye et al. (2016)

(Continues)

TABLE 1 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
SPACEMIX	Landscape genomics	Create maps based on genetic distance, and identify admixture	Creates a "geogenetic map" by embedding genetic distances in a map; anomalously high similarity in this map can be indicative of admixture	Can be slow for large data sets	http://www.genescape.org/spacemix.html	Bradburd et al. (2016)
SNPRELATE	Multivariate analysis	Summarizing variance across loci and visualizing interindividual genetic distance	Fast. Can use VCF files as an input	Requires careful interpretation (Jombard et al. 2009)	https://bioconductor.org/packages/release/bioc/html/SNPRelate.html	Zheng et al. (2012)
EIGENSTRAT/ SMARTPCA	Multivariate analysis	Summarizing variance across loci and visualizing interindividual genetic distance	Fast. Can use VCF files as an input	Requires careful interpretation (Jombard et al. 2009)	https://github.com/DReichLab/EIG/tree/master/EIGENSTRAT	Price et al. (2006)
DAPC (ADEGENET)	Multivariate analysis/ clustering	Maximizes divergence between groups identified by PCA	Fast. Less sensitive to HW eqilibrium assumptions. Claims to be more efficient than STRUCTURE	Requires careful interpretation (Jombard et al. 2009)	http://adegenet.r-forge.r-project.org/	Jombart et al. (2010)
SPCA (ADEGENET)	Multivariate analysis/ clustering	Spatially explicit model to assess population structure	Spatially explicit and able to detect cryptic structure. Fast	Does not take into account HW equilibrium or LD	http://adegenet.r-forge.r-project.org/	Jombart et al. (2008)
LAMP	Pedigree, identity by descent/ state	Chromosome painting, relatedness	LAMP also allows for association and pedigree analyses	Identifies local ancestry in windows (source of noise), requires phased data	http://lamp.icsi.berkeley.edu/lamp/	Baran et al. (2012)
PLINK	Pedigree, Identity by descent/ state	Estimating inbreeding and relatedness	Allows studying identity by descent and by state. PLINK is a multipurpose tool, facilitating data analysis within the same software	NA	http://pngu.mgh.harvard.edu/~purcell/plink/	Purcell et al. (2007)
VCFTOOLS	Pedigree, identity by descent/ state	Estimating inbreeding and relatedness	Computes unadjusted Ajk and kinship coefficient	NA	https://vcftools.github.io/man_latest.html	Danecek et al. (2011)
KING	Pedigree, identity by descent/ state	Estimating inbreeding and relatedness, multivariate analysis	Mendelian error checking, testing family structure, highly accurate kinship coefficient, association analysis, population structure inference	Kinship coefficient also computed in VCFTOOLS	http://people.virginia.edu/~wc9c/KING/Download.htm	Manichaikul et al. (2010)

(Continues)

TABLE 1 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
COLONY	Pedigrees	Pedigree inference from SNPs	Robust even with high error rates (e.g., low-depth sequencing). Can handle haplo-diploid systems (e.g., ants). Multithreaded	Can only simulate genotypes with the Windows version	https://www.zsl.org/science/software/colony	Wang (2019)
SEQUOIA	Pedigrees	Pedigree inference from SNPs	Can be applied to large pedigrees (>1,000 individuals). Accomodates unknown birth times	Handles hundreds of SNPs. For whole-genome data, preliminary filtering and LD-pruning may be recommended. Efficient with ~100 SNPs	https://cran.r-project.org/web/packages/sequoia/index.html	Huisman (2017)
LDHAT	Recombination	Estimating variation in recombination rates along a genome	Handles unphased and missing data, underlying model can be used for organisms such as viruses or bacteria	Limited to 300 sequences, specific format (not VCF), model for recombination hotspots based on human data	http://ldhat.sourceforge.net/	McVean et al. (2002)
LDHOT	Recombination	Identifying recombination hotspots	Specifically designed for detecting recombination hotspots	Requires data to be phased, working with LDHat	https://github.com/auton1/LDhot	Myers et al. (2005)
ISMC	Recombination	Recombination from a single diploid genome	No phasing needed. Accepts VCF files as input	Introgression and demographic misspecification may bias results. No detailed tutorial	https://github.com/gvbarroso/iSMC	Barroso et al. (2019)
LDHELMET	Recombination	Estimating variation in recombination rates along a genome	Higher accuracy than LDHat	Requires phased data. Does not handle VCF, only fasta and fastq formats. Requires dividing the genome into short segments to be analysed in parallel	https://sourceforge.net/projects/ldhelmet/	Chan et al. (2012)

TABLE 2 Summary of methods for demographic inference, detecting introgression and comparing demographic scenarios.

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
DSUITE	ABBA–BABA	Identifying past events of admixture between populations	Fast, handles VCF format. Suited for low-depth sequencing (handles uncertainties on genotypes). Provides a set of summary statistics that are useful to investigate complex admixture events	Requires an outgroup sequence. The methods cannot estimate the direction of gene flow.	https://github.com/millanek/Dsuite	Malinsky et al. (2021)
RENT+	Ancestral recombination graphs/ coalescence	Retracing the whole process of recombination and coalescence along a genome	Faster than first version of ARGWEAVER	Requires phased haplotypes. Specific input format. No built-in functions to extract information from genealogies	https://github.com/SajadMirzaei/RentPlus	Mirzaei and Wu (2017)
TREEMIX	Clustering and characterizing admixture	Admixture graph, infers most likely admixture events in a tree	Based on allele frequencies and can be used for pooled data	Requires multiple runs to properly assess the likelihood of each model	https://bitbucket.org/nygcresearch/treemix/src	Pickrell and Pritchard (2012)
G-PHOCS	Coalescence/ Bayesian	Estimating population divergence and migration parameters using a coalescent framework	Bayesian + MCMC, handles ancient samples	Parameters scaled by mutation rate, no admixture	http://compgen.cshl.edu/GPhoCS/	Gronau et al. (2011)
ABLE	Coalescence/ composite likelihood	Model comparison and parameter estimation	Uses both allele frequency spectrum and linkage disequilibrium within blocks of a prespecified size	Relies on MS syntax. Determining the most informative size for blocks requires performing pilot runs	https://github.com/champost/ABLE	Beeravolu et al. (2018)
STAIRWAY2	Coalescence/ composite likelihood	Inferring change in N_e with time	User-friendly. Fast. Suitable for pools or low-depth sequencing	Cannot handle migration or population splits	https://github.com/xiaoming-liu/stairway-plot-v2	Liu and Fu (2020)
FASTSIMCOAL2	Coalescence/ likelihood	Model comparison and parameter estimation	Performs coalescent simulations, parameter estimation and model testing using a fast likelihood method. Can handle arbitrarily complex scenarios for any type of marker	The maximum-likelihood method only uses the allele frequency spectrum. Several runs (20–100) are needed to explore the likelihood space	http://cmpg.unibe.ch/software/fastsimcoal2/	Excoffier et al. (2013)

(Continues)

TABLE 2 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
∂A∂I	Diffusion approximation of the AFS	Model comparison and parameter estimation	Run time does not depend on the number of SNPs included, does not require coalescent simulations, handles arbitrarily complex scenarios. Fast estimation of confidence intervals around parameter estimates (Godambe method). Suitable for pools/low-depth sequencing	Requires some knowledge of Python. Limited to three populations. Several runs (20–100) are needed to explore the likelihood space.	https://bitbucket.org/gutenkunstlab/dadi	Gutenkunst et al. (2009)
MOMENTS	Diffusion approximation of the AFS	Model comparison and parameter estimation	Based on Python, syntax similar to ∂A∂I. Can handle selection. Can use VCF files as input	Requires some knowledge of Python. Limited to five populations. Several runs (20–100) are needed to explore the likelihood space	https://bitbucket.org/simongravel/moments/src/master/	Jouganous et al. (2017)
MOMI2	Diffusion approximation of the AFS	Model comparison and parameter estimation	Can scale to 10 populations. Can simulate and read data in the VCF format. Detailed tutorials available	Does not handle continuous gene flow	https://github.com/popgenmethods/momi2	Kamm et al. (2020)
KIMTREE	Diffusion approximation/ Bayesian	Estimating divergence time between populations and testing for topologies. Estimate divergence times and past effective sex-ratio along branches of a populations tree	Fast and user-friendly. R scripts to obtain plots are available. Suitable for pools/low-depth sequencing. The method is conditional on a prior topology provided by the user. It computes DIC for a given topology, allowing to test for the best one	Strong selection on the sex chromosome can produce male-biased sex-ratios. Times are given in diffusion timescale, and can be converted in demographic times using independent estimates of N_e	http://www1.montpellier.inra.fr/CBGP/software/kimtree/download.html	Clemente et al. (2018)
GADMA	Genetic algorithm	Model comparison and parameter estimation	Based on moments and ∂A∂I. Automates the search for the best set of models explaining a given frequency spectrum	Limited to three populations at the moment	https://github.com/ctlab/GADMA	Noskova et al. (2020)
DORIS	Identity by descent (IBD) tract	Testing various demographic scenario	Uses variation in IBD tracts length to test for various demographic models	IBD must be inferred first with (e.g., BEAGLE). Handles a limited set of demographic scenarios. Modification in the code is required for more complex scenarios	https://github.com/pierpal/DoRIS	Palamara and Pe'er (2013)
UNNAMED.	Identity by state (IBS) tract	Predict observed patterns of IBS along a genome by fitting an appropriate, arbitrary complex demographic model	Allows bootstrapping and estimating confidence over parameter estimates with MS	Specific input format (similar to MSMC or ARGWEAVER)	https://github.com/kelleyharris/Inferring-demography-from-IBS	Harris and Nielsen (2013)

(Continues)

TABLE 2 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
ASTRAL-2	Phylogeny	Builds species trees using short nonrecombining sequences	Coalescence-based. Suitable for short loci (e.g., RAD-seq and GBS)	More reliable under high incomplete lineage sorting than SVDQUARTETS and NJST (Chou et al. 2015)	https://github.com/smirarab/ASTRAL	Mirarab and Warnow (2015)
BEAST2	Phylogeny	Network reconstruction and phylogenetic relationships	User-friendly. Can be used to track changes in effective population sizes (Bayesian Skyline Plots). Possible to estimate divergence times	Slow for large data sets. Requires sequence data that can be produced by, for example, STACKS for RAD-seq data	http://beast2.org/	Drummond and Rambaut (2007), Bouckaert et al. (2014)
IQ-TREE 2	Phylogeny	Divergence time estimation and phylogenetic relationships	User-friendly, can be run locally or on a web server, very detailed tutorials. Fast and accurate	Still no tutorial for analysing big data (last checked December 2020)	http://www.iqtree.org/	Minh et al. (2020)
MCMCTREE AND MCMCTREER	Phylogeny	Divergence time estimation and phylogenetic relationships	Included in PAML. An R program is designed to help choose relevant priors and interpret results https://github.com/PuttickMacroevolution/MCMCtreeR	Bayesian, sensitive to priors. Requires a resolved phylogeny and an alignment. Slow for large data sets. Not suited for recent divergence and high gene flow	http://abacus.gene.ucl.ac.uk/software/paml.html	Yang (2007), Puttick (2019)
NJST	Phylogeny	Builds species trees using short nonrecombining sequences	Available in the R package PHYBASE. Estimates populations/species tree from gene trees	Requires splitting part of the genome into nonrecombining "loci"	https://github.com/bomeara/phybase/	Liu and Yu (2010), Liu and Yu (2011)
PHRAPL	Phylogeny	Admixture graph, reticulated evolution	Uses trees in NEWICK format as an input to infer topology, migration rates, divergence times. Similar to ABC in spirit, using tree topology as a summary statistics	Cannot handle more than 16 taxa at a time, and requires subsetting larger data sets	http://www.phrapl.org/	Jackson et al. (2017)
PHYML	Phylogeny	Phylogenetic relationships	Maximum likelihood inference of phylogenetic relationships. An online version is available	Should be used on complex of species or divergent populations with little migration. Can be run on genomic windows to detect introgression (with, e.g., TWISST, DSUITE)	http://www.atgc-montpellier.fr/phyml/binaries.php	Guindon et al. (2010)
RAXML	Phylogeny	Network reconstruction and phylogenetic relationships	Maximum likelihood inference of phylogenetic relationships	Should be used on complex of species or divergent populations with little migration	http://sco.h-its.org/exelixis/web/software/raxml/index.html	Stamatakis (2014)

(Continues)

TABLE 2 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
SNAPP	Phylogeny	Phylogenetic relationships	Handles SNP data	Remains slow for medium to large data sets (>1,000 SNPs)	http://beast2.org/snapp/	Bryant et al. (2012)
SNPHYLO	Phylogeny	Network reconstruction and phylogenetic relationships	Complete pipeline from SNP filtering to tree reconstruction	Should be used on complex of species or divergent populations with little migration	http://chibba.pgml.uga.edu/snphylo/	Lee et al. (2014)
SVDQUARTETS	Phylogeny	Phylogenetic relationships	Estimates populations/species tree from gene trees	Remains slow for large data sets. Requires PAUP*	https://www.asc.ohio-state.edu/kubatko.2/software/SVDquartets/	Chifman and Kubatko (2014)
SVDQUEST	Phylogeny	Phylogenetic relationships	Estimates populations/species tree from gene trees	Faster than SVDQUARTETS	https://github.com/pranjalv123/SVDquest	Vachaspati and Warnow (2018)
*BEAST	Phylogeny and species tree inference	Divergence time estimation and phylogenetic relationships	Outputs a species tree instead of concatenated gene tree. Allows for testing consistency between phylogenetic signals at different loci	Slow for large data sets. Requires sequence data. Not suited for situations where gene flow/admixture is important	http://beast2.org/	Heled and Drummond (2010)
SPLITSTREE	Phylogeny/network	Network reconstruction and phylogenetic relationships	User-friendly interface, proposes a variety of methods for network reconstruction	Mostly descriptive	http://www.splitstree.org/	Huson and Bryant (2006)
DICAL2	Sequentially Markovian coalescent	Testing any arbitrary demographic scenario	Works with smaller, more fragmented data sets than PSMC. Handles more complex demographic models than MSMC (including admixture)	Requires phased whole genome data and a model to be defined	https://sourceforge.net/projects/dical2/	Sheehan et al. (2013)
MSMC AND MSMC-IM	Sequentially Markovian coalescent	Inferring change in N_e and migration rates with time between two populations	Allows tracking of population size changes in time without a priori. Allows estimating variation in cross-coalescence rate between two populations	Limited to the study of eight diploid individuals from two populations at once. Requires whole genome phased data and masking regions with insufficient sequencing depth	https://github.com/stschiff/msmc and https://github.com/wangke16/MSMC-IM	Schiffels and Durbin (2014)
SMC++	Sequentially Markovian coalescent	Inferring change in N_e with time and splitting time between two populations	Can analyse hundreds of individuals at a time and does not require phasing	Masking regions as in MSMC. The ancestral allele is assumed to be the reference allele by default. Assumes a clean split for population divergence. Future versions should allow gene flow inference	https://github.com/popgenmethods/smcpp	Terhorst et al., (2016)

(Continues)

TABLE 2 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
TWISST	Topology weighting	Chromosome painting, clustering and branching between populations	Retrieves the most likely coalescence pattern between several taxa along the genome. Can be seen as an extension of the ABBA/BABA test	Needs a priori grouping of individuals into taxa. Requires at least four taxa. Impractical for more than six taxa. Windows size must include enough SNPs to retrieve the correct topology but at the risk that regions with different histories are included	https://github.com/simonhmartin/twisst	Martin and Van Belleghem (2017)
BAYPASS/BAYENV	Variance/covariance matrix	Building a population covariance matrix across population allele frequencies, similar to TREEMIX	Can handle pooled data	Matrices are mostly designed to provide a neutral model for assessing selection, but can be used to infer population structure	http://www1.montpellier.inra.fr/CBGP/software/baypass/; https://bitbucket.org/tguenther/bayenv2_public/src	Günther and Coop (2013), Gautier (2015)

variation in a discriminant analysis, allowing clusters of individuals with highest genetic differentiation to be identified (e.g., using discriminant analysis of principal components [DAPC]; Jombart et al., 2010), and with potential to incorporate temporal sampling (e.g., using DYSTRUCT or TFA; Joseph & Pe'er, 2019; François & Jay, 2020), which is relevant for museum and ancient DNA studies.

2.3 | Model-based inference of population structure

Unlike the previous set of "algorithmic" approaches (see taxonomy proposed in Alexander & Novembre, 2009), model-based approaches model the probability of observing a set of genotypes given a predefined number of clusters (*K*). Some of these methods use a Bayesian (e.g., STRUCTURE; Pritchard et al., 2000, FASTSTRUCTURE; Raj et al., 2014) or a maximum-likelihood framework (e.g., ADMIXTURE; Alexander & Novembre, 2009) and are usually run for a range of *K* values. The optimal number of clusters can then be determined based on likelihood, although examining population structure for a range of *K* can allow substructure to be better identified. The main interest of these approaches is that they provide an estimate of coancestry coefficients, which are the proportions of an individual genome originating from multiple ancestral gene pools. Such information is more difficult to retrieve with approaches such as PCA (though not impossible, see McVean, 2009). There have been criticisms, however, regarding whether ambiguous assignment should actually be interpreted as a signal of admixture, and detailed inference requires thorough model testing and estimating the goodness of fit of a model with admixture (see Lawson et al., 2018).

2.4 | Heterogeneous structure in space: Landscape genomics

Some methods can explicitly use spatial information to inform clustering, allowing improved consideration of the effect of landscape heterogeneity on selection against migrants and drift (e.g., SPACEMIX, TESS3, Table 1). This spatial perspective can be useful to visualize the location and shape of hybrid zones (Guedj & Guillot, 2011). Simple Mantel tests have been popular to routinely investigate relationships between ecological variables and genetic differentiation while accounting for geographical distances. However, these tests are biased by spatial autocorrelation, assume linear dependence between variables, and do not allow testing the relative contribution of each variable (Guillot & Rousset, 2013; Legendre & Fortin, 2010). More recent methods such as EEMS (Petkova et al., 2015) divide the landscape with a dense geographical grid, and identify edges between samples where the effective migration rates are higher or lower than expectations based on isolation-by-distance. The method, however, cannot differentiate between scenarios that lead to the same amount of divergence between samples (e.g., divergence in isolation followed by secondary contact or a geographical barrier with constant, low gene flow). However, a recent expansion of the model, MAPS (Al-Asadi et al., 2019), makes the most of the information provided

TABLE 3 Summary of common methods for identifying loci under positive and balancing selection. The table also lists methods targeting loci associated with environmental features and phenotypes of interest. Note that GENABEL is no longer maintained

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
ARGWEAVER/ ARGWEAVER-D	Ancestral recombination graphs/ coalescence	Retracing the whole process of recombination and coalescence along a genome	Provides quantitative estimates for time to the most recent common ancestor (TMRCA) and topologies at each locus. ARGWEAVER-D can estimate introgression. Estimates effective population size. Provides tools to extract summary statistics for the topologies retrieved. Does not require phasing (but slower)	High computing cost. Slower on unphased or low depth data. ARGWEAVER-D is not part of the Anaconda (Python) distribution (http://compgen.cshl.edu/ARGweaver/doc/argweaver-d-manual.html)	Can be installed via conda: conda install -c genomedk argweaver and https://github.com/mjhubisz/argweaver and http://compgen.cshl.edu/ARGweaver/doc/argweaver-d-manual.html	Rasmussen et al. (2014); Hubisz et al. (2020)
GAPIT3	Association	Detecting association with environmental/ phenotypic features	Includes most methods for GWAS studies, including procedures for fast computation, mixed linear models, efficient mixed model association, Bayesian methods such as BLINK, diagnostics such as QQ plots and genotype filtering	May be slow for very large data sets	https://github.com/jiabowang/GAPIT3	Wang and Zhang (2020)
GEMMA	Association	Detecting association with environmental/ phenotypic features	Computationnally efficient for large-scale data sets	Imports data from PLINK format	http://www.xzlab.org/software.html	Zhou and Stephens (2012)
PLINK	Association	Detecting association with environmental/ phenotypic features	Handles a variety of tests for population structure and relatedness	Population structure/kinship need to be assessed in prior association analysis	http://pngu.mgh.harvard.edu/~purcell/plink/	Purcell et al. (2007)
TRINCULO	Association	Detecting association with environmental/ phenotypic features	Specifically designed to handle categorical variables with more than two categories. Performs multinomial logistic regression and provides frequentist and Bayesian frameworks	Requires lapack library in Unix. Allows fine-mapping by testing for corrrelations between adjacent markers	https://sourceforge.net/projects/trinculo/	Jostins and McVean (2016)
SAMBADA	Association/ environmental association	Detecting association with environmental/ phenotypic features	Designed to be fast, underlying models have been kept simple. Allows conversion from PLINK format. Takes into account spatial autocorrelation of individual genotypes. Allows correction for population structure	Does not work with pooled data. Possibly high levels of false positives. Relatedness between samples should be assessed independently. Should be used in combination with LFMM or BAYPASS	http://lasig.epfl.ch/sambada	Stucki et al. (2017)

(Continues)

TABLE 3 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
RELATE	Coalescence with recombination	Reconstruct genome-wide genealogies for hundreds of samples	Provides quantitative estimates for TMRCA and topologies at each locus. Infers past demography (similar to PSMC methods). Infers changes in mutation rates. Performs scans for positive selection over discrete time periods	Requires an outgroup to polarize alleles as ancestral/derived. Requires a recombination map. Does not reconstruct ARG *sensu stricto*, and does not estimate uncertainty of the local genealogies	https://myersgroup.github.io/relate/index.html	Speidel et al. (2019)
DICAL-IBD	Coalescent with recombination/IBD	Predicting IBD tracts from demographic models	High IBD sharing suggests recent positive selection.	Uses DICAL output to obtain expectations based on demographic scenarios	https://sourceforge.net/projects/dical-ibd/	Tataru et al. (2014)
VOLCANOFINDER	Composite likelihood test	Adaptive introgression	Detects a specific signature of increase then drop in diversity near a selected locus brought in a population through introgression	Private input format. Computationally intensive, needs to be run in parallel.	http://degiorgiogroup.fau.edu/vf.html	Setter et al. (2020)
SCCT	Conditional coalescent tree	Detecting positive selection	Designed for detecting recent positive selection. Claims to be more precise at identifying selected sites	The ancestral state of alleles must be obtained through an outgroup	https://github.com/wavefancy/scct	Wang et al. (2014)
LFMM	Environmental association	Detecting adaptation to environmental features	Corrects for population structure using latent factors, faster than BAYENV for large data sets	Only performs association with environment	http://membres-timc.imag.fr/Olivier.Francois/lfmm/software.htm	Frichot et al. (2013)
CLUES	Genealogies at selected loci	Estimate the time at which a beneficial allele rises in frequency	Previous version used ARGWEAVER output, current version uses RELATE. Provides scripts to plot the trajectory of selected alleles	Assumes a panmictic population, neglects the effects of selection at linked sites	https://github.com/35ajstern/clues	Stern et al. (2019)
PALM	Genealogies at selected loci	Estimate the strength and timing of selection on polygenic traits	Uses genealogies estimated from RELATE and results from GWAS to estimate timing and strength of selection for polygenic traits. Should be robust to pleiotropy and residual structure in GWAS	May overestimate selection for older events. Only tested in humans	https://github.com/35ajstern/palm	Stern et al. (2021)
STARTMRCA	Genealogies at selected loci	Estimate the time at which a beneficial allele rises in frequency	Compares genealogies between carriers and noncarriers of an advantageous mutation, assuming a star-genealogy at selected loci. Can handle VCF files	Requires a reference panel of noncarrier haplotypes. Sensitive to local diversity before the sweep, and to migration events during a sweep. More indicated for recent sweeps	https://github.com/jhavsmith/startmrca	Smith, Coop, Stephens, and Novembre (2018)
ANCESTRY _ HMM-S	Identity-by-state tracts	Adaptive introgression	Estimates the selective coefficient of the introgressed loci through a hidden-Markov chain approach	Requires the time and extent of introgression to be defined by the user	https://github.com/jesvedberg/Ancestry_HMM-S/	Svedberg et al. (2021)

(Continues)

TABLE 3 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
H12 TEST	LD	Detecting selection using signatures of high LD	Does not require phased data. Designed for detecting soft sweeps	Coalescent simulations are recommended to evaluate the likelihood of selection	https://github.com/ngarud/SelectionHapStats/	Garud et al. (2015)
LDNA	LD	Detecting selection using signatures of high LD	Can be used to address population structure or detect large inversions or indel polymorphism through LD	The user needs to play with parameters to ensure robustness of SNPs significantly linked	https://github.com/petrikemppainen/LDna	Kemppainen et al. (2015)
REHH	LD	Detecting selection using signatures of high LD	Can compute both XP-EHH and *Rsb*. Handles several input formats	Requires phased data and high density of markers	https://cran.r-project.org/web/packages/rehh/index.html	Gautier and Vitalis (2012)
SCAN FOR EPISTATIC INTERACTION (BASED ON LD)	LD	Polygenic selection/epistatic interactions	Uses genome-wide LD between a candidate locus and the rest of the genomes to identify epistatic interactions. Can test SNP–SNP interaction, or between genomic windows (summarizes genotypes through PCA)	Lack of a detailed tutorial	https://github.com/leaboyrie/LD_corpc1	Boyrie et al. (2020)
SELSCAN	LD	Detecting selection using signatures of high LD	Includes the *nSL* statistics dedicated to soft sweep detection	Does not include utilities to specify the ancestral state of alleles. Requires phased data and high density of markers	https://github.com/szpiech/selscan	Szpiech and Hernandez (2014)
BALLET	Likelihood test for balancing selection	Detecting balancing selection	Designed for detecting ancient balancing selection. Does not require phasing	Requires whole-genome data and recombination map. The ancestral state of alleles must be obtained through an outgroup	http://www.personal.psu.edu/mxd60/ballet.html	DeGiorgio et al. (2014)
BETASCAN2	Local associations of allele frequencies	Detecting balancing selection	Uses correlations in frequencies between genomically proximate SNPs to compute a score. Can incorporate information about ancestral/derived alleles, fixed derived variants and normalizes the statistics depending on the number of sites in a given genomic window. Very detailed tutorial and utilities	Requires estimating the length distribution of ancestral fragments on each side of the selected site. The 95% percentile can be estimated with the formula L = −log(0.05)/(T*rho), where T is the time since selection in generations and rho is the effective recombination rate/generation	https://github.com/ksiewert/BetaScan	Siewert and Voight (2017), Siewert and Voight (2020)
NCD STATISTICS	Local associations of allele frequencies	Detecting balancing selection	Examines the observed and expected frequency spectra of polymorphisms in genomic windows to test for selection. Can incorporate fixed differences with an outgroup (NCD2), but not mandatory (NCD1)	Private input format, requires simulations to calibrate the statistics. Requires to define the expected equilibrium frequency of alleles (usually between 0.3 and 0.5). Low sensitivity below these frequencies	https://github.com/bbitarello/NCD-Statistics	Bitarello et al. (2018)
PCADAPT	Population differentiation	Detecting positive selection and local adaptation	Does not require to define populations. Handles admixed populations and pooled data sets	False positive rate can be high	http://membres-timc.imag.fr/Michael.Blum/PCAdapt.html	Duforet-Frebourg et al. (2016)

(Continues)

TABLE 3 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
SELESTIM	Population differentiation	Detecting positive selection and local adaptation	Can estimate the coefficients of selection. Calibration using a simulated data set (can be used in combination with the R function simulate.baypass() in BAYPASS)	Assumes a Wrigth–Fisher island model.	http://www1.montpellier.inra.fr/CBGP/software/selestim/	Vitalis et al. (2014)
BAYENV, BAYPASS	Population differentiation/association	Detecting positive selection and adaptation to environmental features	Less sensitive to population demographic history than previous methods. Handles pooled data sets	Significance thresholds need to be determined from simulated data sets. Calibration with neutral SNPs is recommended. BAYPASS better estimates the kinship matrix	http://www1.montpellier.inra.fr/CBGP/software/baypass/; https://bitbucket.org/tguenther/bayenv2_public/src	Günther and Coop (2013), Gautier (2015)
FLK	Population differentiation/association	Detecting positive selection and local adaptation	Less sensitive to population demographic history than previous methods	Requires an outgroup population	https://qgsp.jouy.inra.fr/index.php?option=com_content&view=article&id=50&Itemid=55	Bonhomme et al. (2010)
LSD	Population differentiation/population-branch test	Detecting positive selection and local adaptation	Compares the level of exclusively shared differences between internal and external branches of a population tree. Allows testing selection occurring on the ancestral branch leading to two populations	Requires several populations to perform the test. May be less sensitive to selection on standing variation	https://bitbucket.org/plibrado/LSD	Librado and Orlando (2018)
POPBAM	Summary statistics	Detecting selection using AFS, differentiation	Extracts summary statistics directly from BAM files	Does not allow for sophisticated filtering and SNP calling	http://popbam.sourceforge.net/	Garrigan (2013)
VCFTOOLS	Summary statistics	Detecting selection using AFS, differentiation	Extracts summary statistics from VCF files. Also allows VCF filtering and conversion	Set of summary statistics not as extensive as POPGENOME	http://vcftools.sourceforge.net/	Danecek et al. (2011)
RAISD	Summary statistics/allele frequency spectrum +LD	Detecting positive selection and local adaptation	Scans the genome for composite signals of selective sweeps summarized by the μ statistics. Corrects for the effects of background selection by estimating a threshold value for the statistics based on simulations with background selection	Uses a single population of interest	https://github.com/alachins/raisd	Alachiotis and Pavlidis (2018)
TASSEL	Summary statistics/association	Detecting association with phenotype	User friendly (Java interface), corrects for relatedness, allows computing summary statistics (LD, diversity)	Requires relatedness to be assessed externally (with, e.g., STRUCTURE)	http://www.maizegenetics.net/tassel	Bradbury et al. (2007)

(Continues)

TABLE 3 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
ANGSD	Summary statistics/ association/ population branch test	Detecting selection using AFS, differentiation, association with functional traits	Allows for association using generalized linear models	Descriptive statistics. *p*-values need to be evaluated through coalescent simulations	http://www.popgen.dk/angsd/index.php/ANGSD	Korneliussen et al. (2014)
SWEED	Summary statistics/ composite likelihood test	Designed for whole genome data (or large continuous regions)	Supports Fasta and VCF formats. Estimates selection coefficients	NA	http://pop-gen.eu/wordpress/software/sweed	Degiorgio et al. (2016)
SELECTIONTOOLS	Summary statistics/ LD	Detecting selection using AFS, differentiation and LD statistics	Allows combining several tools in a single pipeline. Includes phasing tools	Set of available summary statistics remains limited (same as VCFTOOLS +Fay and Wu's *H*)	https://github.com/MerrimanLab/selectionTools	Cadzow et al. (2014)
GRROSS	Summary statistics/ allele frequency spectrum	Detecting selection in populations with complex admixture history	Computes the S_B statistics, which detects loci/regions deviating from neutral expectations for each branch leading to current populations. Supports VCF format (converter tools available). Runs with R	Requires that the history of admixture is known, and described with an admixture graph	https://github.com/FerRacimo/GRoSS	Refoyo-Martínez et al. (2019)
PAML/CODEML	Summary statistics/ phylogeny	Distribution of fitness effects/ selection on coding variation	Estimates selection along branches in a phylogeny for genes of interest, contrasting patterns of synonymous and nonsynonymous substitutions. A detailed tutorial is available here: https://link.springer.com/protocol/10.1007%2F978-1-4939-1438-8_4#Sec29	Slow for large data sets. Needs to be parallelized	http://abacus.gene.ucl.ac.uk/software/paml.html	Yang (2007)
POLYDFE2.0	Summary statistics/ phylogeny	Distribution of fitness effects/ selection on coding variation	Can test for invariance of DFEs across data sets (genomic regions within species, or different species). No need for divergence estimates (does not assume that the same DFE is shared between species and outgroup). Very detailed tutorial available here: https://link.springer.com/protocol/10.1007/978-1-0716-0199-0_6	Comparisons require a large number of SNPs for each data set for comparisons to be meaningful	https://github.com/paula-tataru/polyDFE	Tataru and Bataillon (2019)
POPGENOME	Summary statistics/ population branch test	Detecting selection using AFS, differentiation	Fast, embedded in R, allows using annotation files (GFF/GTF format)	Does not perform association, but can be used in combination with GENABEL within R	https://cran.r-project.org/web/packages/PopGenome/index.html	Pfeifer et al. (2014)

TABLE 4 Summary of common methods for simulating genome-wide data and performing simulation-based parameter inference and model comparison (supervised machine learning and Approximate Bayesian Computation)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
ABC/ABCRF	ABC	Performs all steps for model-checking and parameter estimation for ABC analyses. ABCRF includes random forest methods (a type of supervised machine-learning)	Informative vignette, allows graphical representation, complete and robust	Does not perform coalescent simulations (but can be used in combination with coala)	https://cran.r-project.org/web/packages/abc/index.html https://cran.r-project.org/web/packages/abcrf/index.html	Csilléry et al. (2012), Raynal et al. (2019)
ABCTOOLBOX	ABC	Complete ABC analysis, from simulations to model checking and parameter estimation	Modular, facilitates the computation of summary statistics	NA	https://bitbucket.org/wegmannlab/abctoolbox/wiki/Home	Wegmann et al. (2010)
DIYABC	ABC	Complete ABC analysis, from simulations to model checking and parameters estimation	User-friendly. Many ways to check goodness-of-fit. Good introduction to ABC models	Does not model continuous gene flow	http://www1.montpellier.inra.fr/CBGP/diyabc/	Cornuet et al. (2008)
POPSIZEABC	ABC	Inferring change in N_e using whole-genome data	Supposed to better assess recent events. Uses a set of summary statistics for the AFS and LD between markers. Handles multiple individuals	Approximate Bayesian approaches do not retrieve the whole information	https://forge-dga.jouy.inra.fr/projects/popsizeabc/	Boistard et al. (2016)
COALA	ABC/coalescent simulations	Combining coalescent simulators within a single framework	Facilitates the building of scenarios and computes summary statistics for simulations. Can be easily combined with the ABC or ABC-RF packages in R	Includes so far MS, MSMS and SCRM	https://cran.r-project.org/web/packages/coala/index.html	Staab and Metzler (2016)
FACSEXCOALESCENT	Coalescent simulations	Simulate demographic scenarios for asexual/facultatively sexual species	Can handle varying levels of sexual reproduction, inbreeding, selfing and cloning	Does not handle population size changes nor selection yet	https://github.com/MattHartfield/FacSexCoalescent https://github.com/MattHartfield/FacSexCoalescentR	Hartfield et al. (2016)
FASTSIMCOAL2	Coalescent simulations	Building any arbitrary scenario using a coalescent framework	Any arbitrary scenario can be implemented. Handles SNP, microsatellites and sequence data	Does not handle selection. Slower than MS with no recombination, much faster with recombination (see manual)	http://cmpg.unibe.ch/software/fastsimcoal2/	Excoffier and Foll (2011)
MS, MSMS, MSABC	Coalescent simulations	Building any arbitrary scenario using a coalescent framework	Any arbitrary scenario can be implemented. Handles SNP, microsatellites and sequence data. MSMS can include selection in the model	Syntax can be difficult to handle for new users compared to, e.g., FASTSIMCOAL2 (but see COALA)	http://www.bio.lmu.de/~pavlidis/home/?Software:msABC	Hudson (2002), Ewing and Hermisson (2010), Pavlidis et al. (2010)

(Continues)

TABLE 4 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
MSMS	Coalescent simulations	Simulate demographic scenarios including selection	Flexible, syntax similar to MS, handles arbitratily complex models. Can be used in an ABC framework to include selection as a parameter to be estimated	Syntax can be difficult to handle for the naive user (but see coala)	http://www.mabs.at/ewing/msms/index.shtml	Ewing and Hermisson (2010)
MSPRIME	Coalescent simulations	Building any arbitrary scenario using a coalescent framework	Faster than MS, Python interface. Syntax is more explicit than MS	Requires some knowledge of Python	https://github.com/tskit-dev/msprime	Kelleher et al. (2016)
SCRM	Coalescent simulations	Fast simulation of chromosome-scale sequences	Syntax similar to MS, handles any arbitrary scenario	Does not handle gene conversion and fixed number of segregating sites (unlike MS)	https://scrm.github.io/	Staab et al. (2015)
SPLATCHE3	Coalescent simulations	Simulating demographic scenarios in their spatial context	Coalescent simulator for genetic data, forward-in-time for demography in space. Spatially explicit.	Simulations can be slow (>1 hr) for large data sets (>100,000 SNPs) over more than 1,000 generations. Does not incorporate selection	http://www.splatche.com/splatche3	Currat et al. (2019)
COALESCENCE	Discoal	Simulate selective sweeps under arbitrary demographic scenarios	Relatively fast for short genomic fragments. Designed to simulate "hard" and "soft" sweeps	Mostly used with DIPLOS/HIC. Other simulators such as MSMS may be more suited for some scenarios	https://github.com/kr-colab/discoal	Kern and Schrider (2016)
QUANTINEMO2	Forward-in-time simulations	Simulating demographic and selection scenarios in their spatial context	Comprehensive simulator. Designed for the study of selection in a spatially explicit context. Simulates quantitative traits, fitness landscapes and underlying genetic variation with migration. Includes both population and individual-based simulations	Scan be slow for large/complex models	https://www2.unil.ch/popgen/softwares/quantinemo/	Currat et al. (2019), Neuenschwander et al. (2019)
SLIM3	Forward-in-time simulations	Simulating genomic sequences with intrinsic and extrinsic factors	One of the most comprehensive simulators. Can simulate genetic data in their spatiotemporal context, the effects of selection at linked sites, coding and noncoding variation, inbreeding and selfing. Supports tree-sequence recording for faster simulations. Large community	Slow for large genomic regions/large populations	https://messerlab.org/slim/	(Haller & Messer, 2019)

(Continues)

TABLE 4 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
DIPLOS/HIC	Supervised machine learning	Detecting selective sweeps	Classifies genomic windows as neutral, selected, or impacted by selection at linked sites. Also distinguishes between selection on standing and *de novo* variation. Uses a set of summary statistics describing frequency spectrum and LD, does not require phasing. Good tutorial explaining the pipeline	Good performance depends on the parameters used to simulate sweeps (window size, selective coefficient, demography). Requires some trial and error for new model species. Interpretation of "soft" and "hard" sweeps remains discussed	https://github.com/kr-colab/diploSHIC	Schrider and Kern (2016), Kern and Schrider (2018)
EVONET	Supervised machine learning	Detecting selective sweeps, balancing selection, and estimate demographic history	Uses deep-learning algorithms to classify genomic regions as selected or neutral, and estimate effective population sizes. Flexible (any number of summary statistics can be provided by the investigator)	Requires summary statistics as an input. Difficult for a naïve user	https://sourceforge.net/projects/evonet/?source=typ_redirect	Sheehan and Song (2016)
FASTEPRR	Supervised machine learning	Estimating effective recombination rates	Uses regression to estimate effective recombination rates from SNP alignments. Can use the VCF format. No clear bias due to phasing errors observed. Can incorporate demographic history (using MS command line)	Requires phased data	https://www.picb.ac.cn/evolgen/softwares/index.html	Gao et al. (2016)
FILET	Supervised machine learning	Detecting introgression	Uses Extra Trees classifiers and dedicated summary statistics to classify genomic windows as being introgressed or not. Identifies the direction of introgression	Targets pulse of introgression rather than continuous gene flow, but can detect the latter. Requires phased data in a fasta format	https://github.com/kr-colab/FILET	Gao et al. (2016)
GENOMATNN	Supervised machine learning	Desstecting adaptive introgression	Uses convolutional neural networks to identify adaptive introgression. Trained using the tree-sequence records obtained from SLIM3. Can handle VCF files and unphased data	Strong computational bottleneck with SLIM simulations	https://github.com/grahamgower/genomatnn	Gower et al. (2020)

(Continues)

TABLE 4 (Continued)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
IMAGENE	Supervised machine learning	Detecting selective sweeps	Uses convolutional neural networks to classify genomic windows in bins of distinct selection coefficients. Directly uses the image of the alignment, avoiding compression (i.e., using summary statistics)	Can be slow for large data sets	https://github.com/mfumagalli/ImaGene	Torada et al. (2019)
RELERNN	Supervised machine learning	Estimating recombination rates	Uses recurrent neural networks to estimate recombination rates from SNP alignments. Handles unphased and pooled data. Uses MSPRIME (Python implementation of MS) to generate simulations upon which the algorithm is trained. Can incorporate known demographic history provided by the user	Can be computationnally intensive for large effective population sizes. Accuracy on pooled data is modest for low depth of coverage. Absolute estimates of recombination rates depend on the accuracy of the mutation rate used for simulations	https://github.com/kr-colab/ReLERNN	Adrion, Galloway, et al. (2020)
SWIFR	Supervised machine learning	Detecting selective sweeps	Uses averaged one-dependence estimator to classify genomic regions as selected or neutral. Flexible in terms of which summary statistics are used. Can incorporate demographic history	Requires summary statistics as an input. Only distinguishes between selective sweeps and neutral regions	https://github.com/ramachandran-lab/SWIFr/blob/master/README.md	Sugden et al. (2018)

by whole genomes. By using phased haplotypes instead of genotypes, it can jointly estimate migration rates and local effective population sizes across the landscape. Because the length of shared haplotypes is associated with the time since they coalesced, analyses can focus on specific classes of haplotype lengths to reconstruct the past migration landscape at different time periods.

There exist methods that complement the approaches described above, by identifying which combination of geographical and ecological distance limits dispersal. A good example is BEDASSLE, which uses the deviation of allele frequencies at unlinked sites in local populations from the global average, and estimates genetic covariance between all pairs of populations. It then uses a spatial model to estimate the strength of association of covariance with environmental features, assuming a negative relationship between genetic and environmental distances. However, disentangling these effects has proved to be complex. A deeper analysis of genes more strongly impacted by either geography or ecology may be more informative when it comes to the proximate causes of reduced dispersion and differentiation, such as biased dispersal (Bolnick & Otto, 2013; Edelaar & Bolnick, 2012) or selection against migrants (Hendry, 2004). Landscape genomics now extends its focus to adaptive genetic variation, and benefits from new methods targeting signatures of selection (see below).

2.5 | Inferring phylogenetic relationships

Recent advances in molecular phylogenetic methods, and the employment of different types of next-generation sequencing (NGS) data is well beyond the scope of this review (see, e.g., Moriarty Lemmon & Lemmon 2013; Cruaud et al. 2014; Wen et al. 2015). In this respect, both maximum likelihood and Bayesian approaches have become popular to investigate evolutionary relationships between individuals from different populations, even when divergence is very recent (e.g., Wagner et al., 2013). These methods are implemented in software such as RAXML (Stamatakis, 2014) and BEAST2 (Drummond & Rambaut, 2007). Ultimately, all molecular phylogenies reconstruct the genealogy of the genes with which they have been constructed. A problem when applying phylogenetic methods, especially in the context of recent divergence, is the assumption that gene trees are representative of lineage history. This assumption is likely to be violated at the population level, since the influences of gene flow and incomplete lineage sorting are strong at this scale, and may cause gene trees to deviate from population history. Fortunately, many recent coalescent-based methods in phylogenomics explicitly model gene trees to fit inside a speciation framework.

When using genome-wide data at the population level, methods specifically dedicated to reconstructing multiple species coalescent (MSCs) models such as *BEAST (STAR-BEAST) may be preferred over concatenation (Edwards et al., 2016), since they accommodate fluctuations in genealogical history across the genome, allowing discordance between species trees and individual gene trees to be identified. However, in the presence of strong gene flow, MSC models can underestimate divergence times and overestimate effective population sizes (Leaché et al., 2014), because they attempt to explained the observed diversity with a strict isolation model. This issue is partially tackled by methods such as PHRAPL (Jackson et al., 2017), which estimates the likelihood of complex histories by examining genealogies at multiple genes and comparing them with coalescent simulations. Such integration is particularly needed for species and populations that are in the "grey zone of speciation" (Roux et al., 2016).

While useful to infer topologies, caution is advised when using branches lengths obtained from SNP-only data sets, foe example to calculate divergence times between different groups or species (Leaché et al., 2015). For this purpose, it might be more straightforward and reliable to extract from the data both variant and invariant sites at several genes (e.g., coding or conserved sequences), and analyse the whole sequences in software like BEAST2. Such analyses can also be performed in two steps: first, estimate the phylogenetic relationships between samples, then apply a molecular clock model to obtain times since divergence. This style of approach is implemented in the Bayesian method MCMCTree in the PAML package (Yang, 2007).

2.6 | Inferring demographic history with likelihood methods based on the allele frequency spectrum

The allele frequency spectrum (AFS) is the distribution of allele frequencies at polymorphic loci in one or several populations (called in that case joint or multipopulation spectrum). Different patterns of gene flow and demographic events all shape the AFS in specific ways (e.g., alleles are likely to occur at more similar frequencies if divergence is recent or if populations are highly connected). Several methods use the AFS to infer the demographic events explaining current genomic diversity. Two of the most popular methods (∂A∂I and FASTSIMCOAL2; Gutenkunst et al., 2009; Excoffier et al., 2013) fit population genetics model specified by the user to the observed spectrum using a maximum-likelihood approach. The AFS expected under a given scenario is obtained through simulation, either using a diffusion approach (∂A∂I), or coalescent simulations (FASTSIMCOAL2). These approaches quickly estimate parameters using composite likelihoods, but do not explicitly take into account correlations induced by linkage disequilibrium (LD) between physically linked markers (but see ABLE; Beeravolu et al., 2018). This might limit power to detect recent demographic events (e.g., migration, Jenkins et al., 2012). Including SNPs that are physically close together should not strongly bias parameter estimation. However, such an approach prevents direct comparisons of likelihoods from different models. Therefore, physically independent SNPs should be used to consider composite likelihoods as quasi likelihoods for model comparison (Excoffier et al., 2013). Using allele frequencies estimated from pooled data sets is also feasible, as illustrated by a recent study on hybridization in *Populus* species where AFS was estimated from pooled whole genome resequencing data (Christe et al., 2016). The same applies for low-depth-sequencing data, with software such as ANGSD or ATLAS that are able to extract the most likely AFS and other relevant summary statistics. Such approaches are particularly promising to analyse whole-genome data from species with large genomes, ancient

DNA samples, or when sequencing costs would otherwise be too prohibitive (Box 1).

2.7 | Inferring past demography with hidden-Markov model and sequentially Markovian coalescent methods

Methods have been developed to infer variation in population sizes with time using the whole genome of one or several diploid individuals. Briefly, these methods model successive genealogies along the genome sequence as a Markov process: the genealogy at one locus only depends on the genealogy at the previous locus. Changes in the topology

BOX 1 Analyzing pooled sequences, ancient DNA samples and low-depth data

Despite decreasing costs, whole-genome sequencing remains quite expensive, especially for species with large genomes. Classical experimental designs usually target a sequencing depth of about 20–40×. However, several options exist in situations in which this depth is not achievable. Pooled sequencing (Futschik & Schlötterer, 2010), in which individuals from the same sampling site/population are sequenced as a single library, can be an option to reduce costs. Summary statistics along the genome and allele frequency spectra can then be extracted for each population (e.g., using methods such as POPOOLATION; Kofler, Orozco-terWengel, et al., 2011; Kofler et al., 2011). Since individual information is not available, variation in LD across individuals cannot be fully exploited, but methods such as ∂A∂I can still be used to test complex demographic scenarios (Gutenkunst et al., 2009). Shallow shotgun sequencing (1–5× per individual) is another approach that gives access to individual information for a similar cost (Buerkle & Gompert, 2013), but might prevent using methods requiring accurate phasing and unbiased individual genotypes. Nevertheless, recent methods such as those implemented in the packages ANGSD (Korneliussen et al., 2014) or ATLAS (Link et al., 2017) are promising. For example, ATLAS includes an approach to reconstruct past demographic histories by applying the pairwise sequentially Markovian coalescent (PSMC) to low-depth ancient DNA samples. ANGSD comes with several methods that estimate relatedness in low-depth samples (NGSRELATEV2; Hanghøj et al., 2019), and can estimate allele frequency spectra that can be used for demographic and selection inference. Among the most powerful methods available, recent versions of ARG-WEAVER are promising since they can take into account genotype quality when reconstructing genealogies along the genome, and can therefore be applied to "low-quality" samples. One of the main drawbacks is that such analyses take time, making ARG-WEAVER more suited to investigating genealogies in a limited set of genomic regions of interest.

BOX 2 Efficiently simulate whole-genome data

Simulations of whole-genome data are poised to become a standard tool for researchers, and recent initiatives such as STDPOPSIM, an open library of population genetics simulation models for multiple species, might help design reproducible simulations (Adrion, Cole, et al., 2020). More than 145 genetic simulators are currently available, but not all can handle genome-sized data (see https://surveillance.cancer.gov/genetic-simulation-resources/). Simulated data can be used to define significance thresholds for summary statistics when trying to scan the genome for regions under selection. Simulations are also at the core of simulation-based algorithms such as ABC or supervised machine-learning. By comparing simulations with observed data, these methods can identify the processes that underlie diversity in any given genomic region.

There are two main categories of simulators, those based on coalescent ("backward in time" simulators), and forward-in-time simulators. The MS software, with its extensions (such as MSMS, Table 4), is one of the most versatile available. Coalescent simulators are generally fast, and can simulate large genomic regions of hundreds of kilobases efficiently. An important limitation of these simulators is that most only simulate SNP data, and were not intended to simulate other categories such as transposable elements. Moreover, despite the abundance of species that practice self-fertilization and asexual reproduction, only FACSEXCOALESCENT is able to model coalescence in facultatively sexual species.

Forward-in-time simulators such as SLIM3 (Haller & Messer, 2019) bypass the aforementioned limitations. They can accomodate an impressive diversity of scenarios and model genomic data in their spatiotemporal context, incorporate purifying and positive selection, and even go beyond Wright–Fisher approximations, for example by allowing overlapping generations. This comes at the cost of speed: long genome sequences can take days or weeks to be simulated. Simulation time can be reduced by scaling mutation rates, selective coefficients, times of demographic events and population sizes, but can still remain relatively long, requiring massive parallelization. However, SLIM3 now supports tree-sequence recording, which greatly reduces simulation time. Instead of explicitly simulating neutral mutations, the method outputs genealogies upon which mutations can be added at a later stage using the coalescent simulator msprime, implemented in Python (Kelleher et al., 2016).

are due to recombination events reconnecting branches in the tree. The whole genealogy is usually not estimated, however, which results in drastic gains in speed. Such methods have the advantage of requiring only a small number of individuals (1–10), no a priori knowledge of population history, and permitting time-varying gene flow to be incorporated (see MSMC-IM). One general drawback, however, is that they are limited to rather simple scenarios, and do not handle more than two populations as yet (but see DICAL2, Table 2). While powerful, they are sensitive to confounding factors such as population structure (OrozcoterWengel, 2016) that lead to false signatures of expansion or bottlenecks. These methods also do not allow extremely recent demographic events to be investigated, since the coalescence of two alleles from a single individual in the recent past (a few tens to hundreds of generations) is infrequent. Moreover, most of these methods require the data to be phased (but see SMC++; Terhorst et al., 2016), for example with FASTPHASE (Scheet & Stephens, 2006) or BEAGLE (Browning & Browning, 2011). In addition, phasing errors can lead to strong biases in parameter estimates for recent times (Terhorst et al., 2016). An extension of these methods takes into account population structure and aims to identify the number of islands contributing to a single genome, assuming it is sampled from a Wright *n*-island metapopulation (Mazet et al., 2015; Rodríguez et al., 2018). Such developments should improve the amount of information retrieved from only a few genomes.

Methods based on tracts of identity-by-descent (IBD, Palamara & Pe'er, 2013) constitute an interesting alternative for more complex model testing when whole genomes are available in large number. Such methods allow recent demographic events to be inferred with relative precision. They are used to predict the length of haplotypes shared by two individuals that are inherited from a common ancestor without recombination. However, IBD detection requires large cohorts and accurate phasing, and therefore application of these methods has been largely restricted to human populations so far (Browning & Browning, 2011; Palamara & Pe'er, 2013). Another approach has used tracts of identity-by-state (IBS) to perform demographic inference over a range of timescales (Harris & Nielsen, 2013). IBS tracts are directly observable since they are simply the intervals between pairwise differences in an alignment of sequences and do not require any assumption about coancestry to be defined. The method predicts the length distribution of IBS tracts for pairs of haplotypes under a range of demographic parameters. These predicted spectra are then compared to empirical data under a likelihood framework, as with methods based on the AFS.

3 | DETECTING LOCAL SIGNATURES OF EVOLUTIONARY PROCESSES ALONG THE GENOME

3.1 | Selection, introgression and their impact on sequence variation

While demographic forces such as drift and migration will affect the whole genome, selection in the presence of recombination is expected to be specific to particular portions of the genome, and therefore yield discrepancies with genome-wide polymorphism (Lewontin & Krakauer, 1973; but see section 3.9). Both positive and negative selection have long-distance effects on sites that are adjacent to those under selection, an effect often put under the umbrella term of "linked selection" (Cruickshank & Hahn, 2014; Ravinet et al., 2017). These effects are stronger in regions of low recombination, and may explain the correlations observed between nucleotide diversity, divergence metrics and recombination rates that are observed across many clades (Charlesworth et al., 1997; Cruickshank & Hahn, 2014). Using whole-genome resequencing data, it is possible to estimate the effective recombination rate along the genome (see the Recombination class of methods in Table 1). Such estimates are particularly useful in the absence of pre-existing genetic maps to assess how recombination and linked selection may bias estimates of diversity statistics or scans for selection. It also provides a way to determine a suitable window size to compute "independent" statistics along genomic windows.

In the sections that follow, we describe different methods aiming at identifying regions under selection by contrasting local patterns of diversity and divergence with genome-wide patterns. We begin with approaches focusing on single populations, and then summarize those focused on multiple populations.

3.2 | Quantifying positive and purifying selection on coding regions

The ratio between the number of nonsynonymous and synonymous mutations (also called dn/ds, K_A/K_S or ω) is often used to detect whether a specific gene is undergoing negative ($\omega < 1$) or positive ($\omega > 1$) selection. It is also useful to estimate the effects of demography on mutational load and ultimately extinction risk. An excess of nonsynonymous mutations can signal positive or balancing selection, or a relaxation of selective constraints on a given gene. More sophisticated tests, such as the MK test (McDonald & Kreitman, 1991), can use population data and compare the proportion of nonsynonymous and synonymous variation segregating within and between species. However, these approaches require an annotated genome and an outgroup to detect synonymous and nonsynonymous variants. Annotation of mutations can be performed with a dedicated software (e.g., SNPDAT; Doran & Creevey, 2013). The main issue with estimating ω from a single pair of species is that its value rarely exceeds 1, even in the case of positive selection, due to long-term effects of purifying selection. A more powerful approach lies in the comparison of nonsynonymous and synonymous mutations between orthologues from different species, and can be performed in packages such as PAML and CODEML (Yang, 2007). These methods are model-based and estimate the likelihood of different models of sequence evolution that can include selection at a specific codon, gene or branches along the phylogeny while accommodating variation in substitution rates, base composition or transition/transversion ratios.

The comparison of the AFS of synonymous (assumed neutral) and nonsynonymous polymorphisms is also useful to infer the distribution

of fitness effects (DFE), an informative measure in quantitative genetics regarding the adaptive potential of populations (Eyre-Walker & Keightley, 2007). This allows estimation of a fundamental parameter for coding sequences, α, the proportion of variants fixed by adaptive evolution. Several probability distributions have been proposed to fit the DFE (usually deriving from the Γ distribution, see Eyre-Walker & Keightley, 2007). Methods aiming at estimating the DFE derive the expected AFS for synonymous and nonsynonymous mutations under different probability distributions, and treat the effects of unknown demography and polarization errors as nuisance parameters shared by both categories of polymorphisms. The DFE is then obtained through comparison of the maximum-likelihood of different models. A well-developed set of models and distributions can be compared and tested in POLYDFE (Tataru & Bataillon, 2019). Note that a very detailed tutorial with scripts is available in Tataru and Bataillon (2020) for the latter method.

3.3 | Detecting selective sweeps (recent positive selection)

Selective sweeps reduce diversity in genomic regions flanking the selected site(s). This leads to local deviations in the shape of the AFS that can be captured by several summary statistics computed over genomic windows, such as π, the nucleotide diversity (Nei & Li, 1979), Tajima's *D* (Tajima, 1989), and Fay and Wu's *H* (Fay & Wu, 2000). Using a combination of these statistics allows targets of selection to be identified with greater precision, and minimizes the confounding effects of demography. However, defining a threshold beyond which the values of a set of statistics supports selection is nontrivial. Recent developments in machine learning and Approximate Bayesian Computation (ABC) may assist in this regard (see section 3.9 below). Directly contrasting genome-wide with local AFS is another option that does not require combining results from multiple summary statistics. This approach has been used to develop composite tests, such as the composite likelihood ratio (CLR) test (Degiorgio et al., 2016; Stamatakis et al., 2013) that aims to detect recent selective sweeps by maximizing the likelihood of a model with selection in a genomic window, and comparing it to a model built on SNPs sampled from the genomic background.

In regions near to a selected allele, it is expected that LD is increased and diversity is decreased, especially after recent positive selection. A class of methods are aimed at targeting those regions that display an excess of long homozygous haplotypes, such as the extended haplotype homozygosity (EHH) test (Sabeti et al., 2002). It is also possible to compare haplotype extension across populations, with the Cross Population Extended Haplotype Homozygosity test (*XP-EHH*; Sabeti et al., 2007) or *Rsb* (the standardized ratio of EHH at a given SNP site; Tang et al., 2007). These methods require data to be phased in order to reconstruct haplotypes, which can make them susceptible to switch-errors. Nevertheless, methods based on LD may be more sensitive to selection on standing variation or on multiple alleles that leave a more subtle signature (so called soft sweeps). Statistics dedicated to the detection of soft sweeps include the *nSL* statistics (Ferrer-Admetlla et al., 2014) in SELSCAN or the *H2*/*H1* statistics (Garud et al., 2015). These statistics usually examine the distribution of the length of homozygous haplotypes (in number of SNPs), comparing ancestral and derived haplotypes (*nSL*), or the second most frequent derived haplotype with the most frequent one (H2/H1). Further studies are still needed to understand to what extent hard and soft sweeps can actually be distinguished (Schrider et al., 2015), as well as their relative importance (Jensen, 2014; Messer & Petrov, 2013). Even hard selective sweeps can be challenging to detect with LD-based statistics especially under unstable demography and weak selection (Jensen, 2014). It is advisable to combine several approaches to improve confidence when pinpointing candidate genes for selection. Methods based on LD alone can sometimes miss the actual variants under selection due to the impact of recombination on local polymorphism that can mimic soft or ongoing hard sweeps (Schrider et al., 2015).

3.4 | Detecting long-term balancing selection

Unlike directional selection, balancing selection can lead to the maintenance of polymorphism at selected loci over long periods of time. This type of selection is extremely relevant for evolutionary biologists (Sellis et al., 2011), since it is at the core of strong coevolutionary dynamics such as host–parasite interactions (Ebert & Fields, 2020). Despite its importance, balancing selection has often been overlooked. This is mostly due to its narrow effects, particularly in the case of long-term balancing selection where recombination erodes association between loci under selection and neutral neighbouring loci. Nevertheless, the emergence of whole-genome resequencing data has facilitated the investigation of these narrow signals. Several recent methods and summary statistics (see Table 3, "Detecting balancing selection") have been specifically developed to detect this type of selection (Bitarello et al., 2018; DeGiorgio et al., 2014; Rasmussen et al., 2014; Siewert & Voight, 2017). These methods are all based on the AFS to some extent. Some methods examine the strength of correlations between allele frequencies at adjacent SNPs (Siewert & Voight, 2020), while others use a CLR approach, contrasting the likelihood of a model with selection in candidate windows with the likelihood computed for all sites in the genome (DeGiorgio et al., 2014). On the other hand, recent balancing selection may look similar to an incomplete selective sweep (Charlesworth, 2006), and be detected by methods aimed at detecting long haplotypes and low diversity.

3.5 | Detecting introgressed genomic regions

Understanding the origin of genomic regions under selection highlights the evolutionary history of adaptive alleles (e.g., Abi-Rached et al., 2011) and contributes to our understanding of the origin and maintenance of reproductive isolation. Studies focusing on hybrid

zones and introgression have provided inspiring examples of adaptive introgression (Hedrick, 2013), as demonstrated by recent work on localized introgression and inversions at a colour locus in *Heliconius* butterflies (The Heliconius Genome Consortium et al., 2012) or adaptive introgression of anticoagulant resistance alleles in mice (Song et al., 2011).

Summary statistics can be useful to obtain a first set of candidates for introgression and selection. One may, for example, plot the distribution of a differentiation measure such as F_{ST} (Weir & Cockerham, 1984) between populations, estimates of effective recombination rates and nucleotide diversity along the genome. Such an approach has been used in Darwin's finches, which uncovered genomic islands of divergence with low recombination rates resisting gene flow (Han et al., 2017). Other approaches, such as chromosome painting (Table 1), extend PCA and ADMIXTURE-like methods by incorporating information about the relative order of markers in the genome, allowing identification of regions for which ancestry differs from the rest of the genome. Recent developments also provide a fast and efficient way to test complex patterns of heterogeneous introgression along the genome (see, for example, Dsuite in Malinsky et al., 2021). These methods build upon the well-known ABBA–BABA statistics (Durand et al., 2011) and provide a variety of estimators that can be estimated for the whole genome or along genomic windows. They require that phylogenetic relationships between populations and species are known (see Section 2.5 above). Other methods allow the user to test the relative contribution of different topologies expected with and without gene flow (e.g., the topology weighting method implemented in TWISST; Martin & Van Belleghem, 2017).

3.6 | Identifying highly differentiated loci and associations between allele frequencies and environmental features

When an allele is under positive selection in a population, its frequency tends to rise to fixation, unless gene flow from other populations or strong drift prevents this from happening (Charlesworth et al., 1997). It is therefore possible to contrast patterns of differentiation between populations adapted to their local environment to detect loci under divergent selection (e.g., displaying a high F_{ST}). However, it is essential to control for population structure, as it may strongly affect the distribution of differentiation measures and produce high rates of false positives. Modern methods based on this principle (Table 3) correct for relatedness across populations, and can test association between allele frequencies and environmental features (see the extensive review by François et al., 2015). Methods such as BAYPASS (Gautier, 2015) are convenient in both describing population structure and providing preliminary insights into the proportion of loci that do not follow neutral expectations. When this proportion is not too high, outliers can be removed to avoid bias (Schrider et al., 2016) and the remaining loci can be used for demographic inference and model-testing. These estimated parameters can then be used to simulate sequences or independent SNPs and generate a neutral expectation. Loci that are more likely to be neutral can be used to further calibrate tests for selection (Lotterhos & Whitlock, 2014).

Detecting an association between environment and allele frequencies does not necessarily imply a role for local adaptation. For example, in the case of secondary contact, intrinsic genetic incompatibilities can lead to the emergence of tension zones that may shift until they reach an environmental barrier where they can be trapped (Bierne et al., 2011). In addition, the effects of selection at linked sites might generate false positives. The sampling strategy must take into account the particular historical and demographic features of the species investigated to gain power (Nielsen et al., 2007). The sequencing strategy must also be carefully considered to control for spatial autocorrelation of genotypes due to IBD and shared demographic history. For example, localized range expansion may produce a spurious association between environmental features and allele frequencies due to repeated founder effects and allele surfing (Excoffier & Ray, 2008). Including samples from populations not affected by such an expansion may avoid reaching biased conclusions by examining signatures of association at a broader scale.

3.7 | Identifying significant genotype–phenotype associations and and epistatic interactions between variants

The methods described above focus on allele frequencies at the population scale, but do not test association with traits that vary between individuals within populations (e.g., resistance to a pathogen, symbiotic association, individual size or flowering time). These traits can be under directional selection, but also under stabilizing selection across multiple populations (e.g., height). For this task, methods performing genome-wide association analysis (GWAS) are better suited. Detailed reviews on these methods and their biases are available (Liu & Yan, 2019; Tam et al., 2019; Wang et al., 2019). Initiatives such as GAPIT3 (Wang & Zhang, 2020) provide most of the currently available tools for GWAS in a single framework. The recent development of multivariate methods also allows loci putatively under selection to be identified in admixed or continuous populations without requiring information about individual phenotype (Duforet-Frebourg et al., 2016).

Uncovering the genetic basis of complex, polygenic traits remains challenging, even in model species (Pritchard & Di Rienzo, 2010; Rockman, 2012). It may be unavoidable as a first step to focus only on traits that are under relatively simple genetic determinism. This can, however, lead to the overrepresentation of loci of major phenotypic effect, a fact that should be acknowledged when discussing the impact of selection on genome variation. The fact that loci of major effect are the easiest to target does not imply that they are necessarily the main substrate of selection (Rockman, 2012). Association methods may help to target variants undergoing soft sweeps, weak selection or those involved in polygenic control of traits (Pritchard et al., 2010).

In such cases, signatures of selection may be subtle and sometimes difficult to retrieve from allele frequency data. Nevertheless, recent tools may have a higher sensitivity to polygenic selection (see section below on Ancestral recombination graphs), and a recent method uses genome-wide patterns of LD between a candidate gene (the "bait") and loci along the genome to detect candidate genes that may be involved in epistatic interactions (Boyrie et al., 2020). Such developments hold great promise in addressing the issue of nonadditive genetic effects.

3.8 | Inferring differences in history along the genome with ancestral recombination graphs

Ancestral recombination graphs (ARGs) are a generalization of the coalescent and describe the sequence of genealogies along a sample of recombining sequence. Genealogies are estimated for each nonrecombining block, and recombination between adjacent blocks is described by breaking the branch leading to the recombining haplotype and allowing it to recoalesce to the rest of the tree. This succession of local trees joined by recombination events provides a full description of the genealogical history of the data and is therefore a promising approach to characterize all modes of selection, introgression and demography while taking into account variation in recombination and mutation rate. Methods that are able to estimate or approximate these genealogies have long been computationally intensive and unapplicable to whole genomes. Fortunately, recent improvements make their application to whole genomes feasible (Table 4). A good example is ARGWEAVER (Rasmussen et al., 2014), which has allowed candidate genes for long-term balancing selection to be recovered from human data, and has recently been used in combination with machine learning methods to study speciation in capuchino seedeater birds (Hejase et al., 2020) and introgression in humans (Kuhlwilm et al., 2016). However, ARGWEAVER remains slow when analysing more that 50 diploid genomes, and is even slower for low-depth or unphased data. Another promising method is implemented in the RELATE software (Speidel et al., 2019). RELATE is able to estimate genome-wide genealogies, and uses this information to reconstruct past demographic trajectories, changes in mutation rate over time, identify loci under positive selection, and estimate when selection acted on candidate mutations. RELATE can handle thousands of genomes in a manageable amount of time, but requires an outgroup sequence to polarize derived alleles, and a recombination map. RELATE comes with two add-on methods, CLUES and PALM, which can use the RELATE output to estimate the strength of selection for single loci and a set of candidate loci identified by GWAS respectively. The latter method therefore provides a way to quantify polygenic selection and adaptive introgression at multiple loci, and constitutes a major advance in the field of population genomics.

3.9 | Jointly inferring demographic history and selection using ABC and supervised machine learning

It has recently become clear that the interactions between mutation and recombination rate, introgression, demography, selective sweeps and background selection have to be integrated into analyses of genetic variation (Andrew et al., 2013; Li et al., 2012; Ravinet et al., 2017). Simulation-based approaches hold great promise for incorporating this complexity. Two nonexclusive methodologies are promising: ABC and supervised machine learning approaches. Both rely on simulated data simulated under a range of parameters which are used to identify the combination of parameter values that are most likely to have generated the observed data (see Box 2 for a discussion on simulators). Both methodologies are flexible and powerful, and have become increasingly popular in population genomics (Csilléry et al., 2010; Schrider & Kern, 2018). These methods can accomodate any type of marker and arbitrarily complex models. By measuring the distance between carefully chosen summary statistics describing each simulation with those from the observed data set, it is possible to infer which combination of selective and demographic parameters best explains the data. These methods enable the rate of false positives to be estimated, for example by estimating how many times simulations of neutral sequences are classified as selected.

However, using summary statistics leads to the loss of potentially useful information (Robert et al., 2011). Machine learning presents three advantages with respect to this problem. First, whereas ABC is prone to show lower performance as the number of statistics summarizing simulations increases, machine learning tends to display better performance. Second, ABC usually excludes many simulations by retaining only those closest to the observed data, whereas machine learning makes use of all simulations to form a model. Third, machine learning algorithms can identify the set of summary statistics that is the most useful for inference, and some neural networks algorithms may also be trained directly on images of aligned sequences (e.g., IMAGENE, Table 4). An example of application of ABC and deep learning is provided in a study of African populations of *Drosophila melanogaster* (Sheehan & Song, 2016), in which the authors use these approaches to identify genomic regions under balancing and positive selection, and infer past demography.

The flexibility of ABC and supervised machine learning means that researchers can adapt the pipeline to their need, for example by using the simulator that generate simulations that are as close to the specifics of their study system as possible. However, this flexibility also means a steeper training curve for researchers learning these methods, due to the lack of a clearly unified analytical pipeline. Moreover, recent discussions on the power of machine learning pipelines to detect selective sweeps have highlighted that a careful consideration of demographic null models and methodological limitations is imperative (Harris et al., 2018).

4 | CONCLUSION

As illustrated by sections 3.8 and 3.9, the field of population genomics is now moving towards both better integrating the demographic framework in inferences of selection, and, conversely, taking into account selection when reconstructing demographic history. The joint

inference of loci under selection and quantification of demographic dynamics is of crucial importance in fields such as landscape genomics or the study of ongoing speciation. It might provide insights into the role of selection, recombination and gene flow in promoting or impairing local adaptation to new habitats. The growing availability of genome-wide data for nonmodel species is therefore promising, but requires caution and high stringency in our interpretation of observed patterns. With the decreasing cost of sequencing, it has been suggested that NGS will rapidly broaden our perspective on complex evolutionary processes, from biogeography (Lexer et al., 2013) to the genetic basis of traits (Hohenlohe, 2014) and the maintenance of polymorphisms (Hedrick, 2006). However, the study of DNA sequence variation is already challenging in its own right, and prone to storytelling (Pavlidis et al., 2012). In order to be informative about processes such as selection and demography, population genomics should ultimately be combined with other disciplines such as ecology and functional analyses (Habel et al., 2015). This can be achieved, for example, by assessing the function of selected genes, the consistency of demographic history with information retrieved from the fossil record or geological history, and the broader integration of population genomics with other fields and methods whenever possible, such as niche modelling, common garden experiments or the study of macro-evolutionary patterns of selection and diversification.

ACKNOWLEDGEMENTS

The University of Basel, New York University Abu Dhabi and the Univeristy of Portsmouth have supported Y.B.'s research in this area. We thank Stephane Boissinot, Joris Bertrand, Muriel Gros-Balthazard, Khaled Hazzouri, Anne Roulin and three anonymous reviewers for their insightful comments on previous versions of the manuscript. We also thank Gabriel Renaud and Peter Ralph for suggesting additional methods.

AUTHOR CONTRIBUTION

YB wrote the first draft of the manuscript and maintained the website. YB and BW agreed on the structure and revised the initial draft.

DATA AVAILABILITY STATEMENT

No data to provide. The tables shown in this article are available at www.methodspopgen.com

ORCID

Yann X. C. Bourgeois https://orcid.org/0000-0002-1809-387X
Ben H. Warren https://orcid.org/0000-0002-0758-7612

REFERENCES

Abi-Rached, L., Jobin, M. J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F. A., Kimani, J., Carrington, M., Middleton, D., Rajalingam, R., Beksac, M., Marsh, S. G. E., Maiers, M., Guethlein, L. A., Tavoularis, S., … Parham, P. (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science*, *334*, 89–95. https://doi.org/10.1126/science.1209202

Abzhanov, A., Extavour, C. G., Groover, A., Hodges, S. A., Hoekstra, H. E., Kramer, E. M., & Monteiro, A. (2008). Are we there yet? Tracking the development of new model systems. *Trends in Genetics*, *24*(7), 353–360. https://doi.org/10.1016/j.tig.2008.04.002

Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J. G., Gladstein, A. L., Gower, G., & Kern, A. D. (2020). A community-maintained standard library of population genetic models. *eLife*, *9*, e54967. https://doi.org/10.7554/eLife.54967

Adrion, J. R., Galloway, J. G., & Kern, A. D. (2020). Predicting the landscape of recombination using deep learning. *Molecular Biology and Evolution*, *37*(6), 1790–1808. https://doi.org/10.1093/molbev/msaa038

Alachiotis, N., & Pavlidis, P. (2018). RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology*, *1*(79), https://doi.org/10.1038/s42003-018-0085-8

Al-Asadi, H., Petkova, D., Stephens, M., & Novembre, J. (2019). Estimating recent migration and population-size surfaces. *PLoS Genetics*, *15*(1), 1–21. https://doi.org/10.1371/journal.pgen.1007908

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109.vidual

Andrew, R. L., Bernatchez, L., Bonin, A., Buerkle, C. A., Carstens, B. C., Emerson, B. C., Garant, D., Giraud, T., Kane, N. C., Rogers, S. M., Slate, J., Smith, H., Sork, V. L., Stone, G. N., Vines, T. H., Waits, L., Widmer, A., & Rieseberg, L. H. (2013). A road map for molecular ecology. *Molecular Ecology*, *22*(10), 2605–2626. https://doi.org/10.1111/mec.12319

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M. T., Perloski, M., & Lindblad-Toh, K. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature*, *495*(7441), 360–364. https://doi.org/10.1038/nature11837

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D. G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J. G., Avila, P. C., Rodriguez-Santana, J., Burchard, E. G., & Halperin, E. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, *28*(10), 1359–1367. https://doi.org/10.1093/bioinformatics/bts144

Barroso, G. V., Puzović, N., & Dutheil, J. Y. (2019). Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLOS Genetics*, *15*(11), e1008449. https://doi.org/10.1371/journal.pgen.1008449

Beeravolu, C. R., Hickerson, M. J., Frantz, L. A. F., & Lohse, K. (2018). ABLE: Blockwise site frequency spectra for inferring complex population histories and recombination. *Genome Biology*, *19*, 145. https://doi.org/10.1186/s13059-018-1517-y

Bierne, N., Welch, J., Loire, E., Bonhomme, F., & David, P. (2011). The coupling hypothesis: Why genome scans may fail to map local adaptation genes. *Molecular Ecology*, *20*(10), 2044–2072. https://doi.org/10.1111/j.1365-294X.2011.05080.x

Bitarello, B. D., De Filippo, C., Teixeira, J. C., Schmidt, J. M., Kleinert, P., Meyer, D., & Andres, A. M. (2018). Signatures of long-term balancing selection in human genomes. *Genome Biology and Evolution*, *10*(3), 939–955. https://doi.org/10.1093/gbe/evy054

Boistard, S., Rodriguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data – An approximate Bayesian computation approach. *PLoS Genetics*, 858–865, https://doi.org/10.1371/journal.pgen.1005877

Bolnick, D. I., & Otto, S. P. (2013). The magnitude of local adaptation under genotype-dependent dispersal. *Ecology and Evolution*, *3*(14), 4722–4735. https://doi.org/10.1002/ece3.850

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J. M., Blott, S., & San Cristobal, M. (2010). Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics*, *186*, 241–262. https://doi.org/10.1534/genetics.110.117275

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS Computational Biology*, *10*(4), 1–6. https://doi.org/10.1371/journal.pcbi.1003537

Boyrie, L., Moreau, C., Frugier, F., Jacquet, C., & Bonhomme, M. (2020). A linkage disequilibrium-based statistical test for Genome-Wide Epistatic Selection Scans in structured populations. *Heredity*, *126*(1), 77–91. https://doi.org/10.1038/s41437-020-0349-1

Bradburd, G. S., Coop, G. M., & Ralph, P. L. (2017). Inferring continuous and discrete population genetic structure across space. *Genetics*, *210*(September), 33–52. https://doi.org/10.1101/189688

Bradburd, G. S., Ralph, P. L., & Coop, G. M. (2013). Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution*, *67*(11), 3258–3273. https://doi.org/10.1111/evo.12193

Bradburd, G. S., Ralph, P. L., & Coop, G. M. (2016). A spatial framework for understanding population structure and admixture. *PLoS Genetics*, *12*(1), 1–38. https://doi.org/10.1371/journal.pgen.1005703

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics (Oxford, England)*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Brisbin, A., Bryc, K., Byrnes, J., Zakharia, F., Omberg, L., Degenhardt, J., Reynolds, A., Ostrer, H., Mezey, J. G., & Bustamante, C. D. (2012). PCAdmix: Principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human Biology*, *84*(4), 343–364. https://doi.org/10.3378/027.084.0401

Browning, B. L., & Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics*, *88*(2), 173–182. https://doi.org/10.1016/j.ajhg.2011.01.010

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & Roychoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, *29*(8), 1917–1932. https://doi.org/10.1093/molbev/mss086

Buerkle, C. A., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: How low should we go? *Molecular Ecology*, *22*(11), 3028–3035. https://doi.org/10.1111/mec.12105

Cadzow, M., Boocock, J., Nguyen, H. T., Wilcox, P., Merriman, T. R., & Black, M. A. (2014). A bioinformatics workflow for detecting signatures of selection in genomic data. *Frontiers in Genetics*, *5*(AUG), 1–8. https://doi.org/10.3389/fgene.2014.00293

Caye, K., Deist, T. M., Martins, H., Michel, O., & François, O. (2016). TESS3: Fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources*, *16*(2), 540–548. https://doi.org/10.1111/1755-0998.12471

Chan, A. H., Jenkins, P. A., & Song, Y. S. (2012). Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genetics*, *8*(12), e1003090. https://doi.org/10.1371/journal.pgen.1003090

Charlesworth, B., Nordborg, M., & Charlesworth, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetics Research*, *70*(02), 155–174. https://doi.org/10.1017/S0016672397002954

Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, *2*(4), e64. https://doi.org/10.1371/journal.pgen.0020064

Chifman, J., & Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics*, *30*(23), 3317–3324. https://doi.org/10.1093/bioinformatics/btu530

Chou, J., Gupta, A., Yaduvanshi, S., Davidson, R., Nute, M., Mirarab, S., & Warnow, T. (2015). A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics*, *16*, S2.

Christe, C., Stolting, K. N., Paris, M., Fraisse, C., Bierne, N., & Lexer, C. (2016). Adaptive evolution and segregating load contribute to the genomic landscape of divergence in two tree species connected by episodic gene flow. *Molecular Ecology*, *26*(1), 59–76. https://doi.org/10.1111/mec.13765

Clemente, F., Gautier, M., & Vitalis, R. (2018). Inferring sex-specific demographic history from SNP data. *PLoS Genetics*, *14*(1), 1–32. https://doi.org/10.1371/journal.pgen.1007191

Cornuet, J.-M., Santos, F., Beaumont, M. A., Robert, C. P., Marin, J.-M., Balding, D. J., Guillemaud, T., & Estoup, A. (2008). Inferring population history with DIY ABC: A user-friendly approach to approximate Bayesian computation. *Bioinformatics*, *24*(23), 2713–2719. https://doi.org/10.1093/bioinformatics/btn514

Cruaud, A., Gautier, M., Galan, M., Foucaud, J., Sauné, L., Genson, G., Dubois, E., Nidelet, S., Deuve, T., & Rasplus, J.-Y. (2014). Empirical assessment of RAD sequencing for interspecific phylogeny. *Molecular Biology and Evolution*, *31*(5), 1272–1274. https://doi.org/10.1093/molbev/msu063

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*(13), 3133–3157. https://doi.org/10.1111/mec.12796

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*(7), 410–418. https://doi.org/10.1016/j.tree.2010.04.001

Csilléry, K., François, O., & Blum, M. G. B. (2012). abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*(3), 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

Cubry, P., Tranchant-Dubreuil, C., Thuillet, A.-C., Monat, C., Ndjiondjop, M.-N., Labadie, K., Cruaud, C., Engelen, S., Scarcelli, N., Rhoné, B., Burgarella, C., Dupuy, C., Larmande, P., Wincker, P., François, O., Sabot, F., & Vigouroux, Y. (2018). The rise and fall of african rice cultivation revealed by analysis of 246 new genomes. *Current Biology*, *28*(14), 2274–2282.e6. https://doi.org/10.1016/j.cub.2018.05.066

Currat, M., Arenas, M., Quilodràn, C. S., Excoffier, L., & Ray, N. (2019). SPLATCHE3: Simulation of serial genetic data under spatially explicit evolutionary scenarios including long-distance dispersal. *Bioinformatics*, *35*(21), 4480–4483. https://doi.org/10.1093/bioinformatics/btz311

Cushman, S. A. (2014). Grand challenges in evolutionary and population genetics: The importance of integrating epigenetics, genomics, modeling, and experimentation. *Frontiers in Genetics*, *5*(JUL), 1–5. https://doi.org/10.3389/fgene.2014.00197

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Degiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., & Nielsen, R. (2016). SWEEPFINDER 2: Increased sensitivity, robustness, and flexibility. *Bioinformatics*, 10.111/mec.13351.RR.*32*(12), 1895–1897. https://doi.org/10.1093/bioinformatics/btw051

DeGiorgio, M., Lohmueller, K. E., & Nielsen, R. (2014). A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genetics*, *10*(8), e1004561. https://doi.org/10.1371/journal.pgen.1004561

Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, *10*(1), 24–29. https://doi.org/10.1038/s41467-019-13225-y

Doran, A. G., & Creevey, C. J. (2013). Snpdat: Easy and rapid annotation of results from de novo snp discovery projects for model and

non-model organisms. *BMC Bioinformatics*, *14*(1), 45. https://doi.org/10.1186/1471-2105-14-45

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*, 214. https://doi.org/10.1186/1471-2148-7-214

Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E., & Blum, M. G. B. (2016). Detecting genomic signatures of natural selection with principal component analysis: Application to the 1000 genomes data. *Molecular Biology and Evolution*, *33*(4), 1082–1093. https://doi.org/10.1093/molbev/msv334

Dufresne, F., Stift, M., Vergilino, R., & Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology*, *23*(1), 40–69. https://doi.org/10.1111/mec.12581

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252. https://doi.org/10.1093/molbev/msr048

Ebert, D., & Fields, P. D. (2020). Host–parasite co-evolution and its genomic signature. *Nature Reviews Genetics*, *21*(12), 754–768. https://doi.org/10.1038/s41576-020-0269-1

Edelaar, P., & Bolnick, D. I. (2012). Non-random gene flow: An underappreciated force in evolution and ecology. *Trends in Ecology & Evolution*, *27*(12), 659–665. https://doi.org/10.1016/j.tree.2012.07.009

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., & Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, *94*, 447–462.

Ellegren, H., Smeds, L., Burri, R., Olason, P. I., Backström, N., Kawakami, T., & Wolf, J. B. W. (2012). The genomic landscape of species divergence in *Ficedula flycatchers*. *Nature*, *491*(7426), 756–760. https://doi.org/10.1038/nature11584

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

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*(10), e1003905. https://doi.org/10.1371/journal.pgen.1003905

Excoffier, L., & Foll, M. (2011). Fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, *27*(9), 1332–1334. https://doi.org/10.1093/bioinformatics/btr124

Excoffier, L., & Heckel, G. (2006). Computer programs for population genetics data analysis: A survival guide. *Nature Reviews. Genetics*, *7*(10), 745–758. https://doi.org/10.1038/nrg1904

Excoffier, L., & Ray, N. (2008). Surfing during population expansions promotes genetic revolutions and structuration. *Trends in Ecology and Evolution*, *23*(7), 347–351. https://doi.org/10.1016/j.tree.2008.04.004

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, *8*(8), 610–618. https://doi.org/10.1038/nrg2146

Fay, J. C., & Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, *155*(3), 1405–1413. https://doi.org/10.1093/genetics/155.3.1405

Ferrer-Admetlla, A., Liang, M., Korneliussen, T., & Nielsen, R. (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, *31*(5), 1275–1291. https://doi.org/10.1093/molbev/msu077

Ferretti, L., Ramos-Onsins, S. E., & Pérez-Enciso, M. (2013). Population genomics from pool sequencing. *Molecular Ecology*, *22*(22), 5561–5576. https://doi.org/10.1111/mec.12522

François, O., & Jay, F. (2020). Factor analysis of ancient population genomic samples. *Nature Communications*, *11*(4661). https://doi.org/10.1038/s41467-020-18335-6

François, O., Martins, H., Caye, K., & Schoville, S. (2015). Controlling false discoveries in genome scans for selection. *Molecular Ecology*, *25*, 454–469. https://doi.org/10.1111/mec.13513

Fraser, D. J., & Bernatchez, L. (2001). Adaptive evolutionary conservation: Towards a unified concept for defining conservation units. *Molecular Ecology*, *10*(12), 2741–2752. https://doi.org/10.1046/j.1365-294X.2001.t01-1-01411.x

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics*, *196*(4), 973–983. https://doi.org/10.1534/genetics.113.160572

Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, *30*(7), 1687–1699. https://doi.org/10.1093/molbev/mst063

Futschik, A., & Schlötterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics*, *186*(1), 207–218. https://doi.org/10.1534/genetics.110.114397

Gao, F., Ming, C., Hu, W., & Li, H. (2016). New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era. *G3 Genes|genomes|genetics*, *6*(6), 1563–1571. https://doi.org/10.1534/g3.116.028233

Garrigan, D. (2013). POPBAM: Tools for evolutionary analysis of short read sequence alignments. *Evolutionary Bioinformatics*, *2013*(9), 343–353. https://doi.org/10.4137/EBO.S12751

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genetics*, *11*(2), 1–32. https://doi.org/10.1371/journal.pgen.1005004

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, *201*(4), 1555–1579. https://doi.org/10.1534/genetics.115.181453

Gautier, M., & Vitalis, R. (2012). Rehh an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, *28*(8), 1176–1177. https://doi.org/10.1093/bioinformatics/bts115

Gower, G., Picazo, P. I., Fumagalli, M., & Racimo, F. (2020). Detecting adaptive introgression in human evolution using convolutional neural networks. *BioRxiv*. https://doi.org/10.1101/2020.09.18.301069

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genetics*, *43*(10), 1031–1034. https://doi.org/10.1038/ng.937

Guedj, B., & Guillot, G. (2011). Estimating the location and shape of hybrid zones. *Molecular Ecology Resources*, *11*(6), 1119–1123. https://doi.org/10.1111/j.1755-0998.2011.03045.x

Guillot, G., & Rousset, F. (2013). Dismantling the Mantel tests. *Methods in Ecology and Evolution*, *4*(4), 336–344. https://doi.org/10.1111/2041-210x.12018

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, *195*(1), 205–220. https://doi.org/10.1534/genetics.113.152462

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*(10), e1000695. https://doi.org/10.1371/journal.pgen.1000695

Haasl, R. J., & Payseur, B. A. (2016). Fifteen years of genomewide scans for selection: Trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*, *25*(1), 5–23. https://doi.org/10.1111/mec.13339

Habel, J., Zachos, F., Dapporto, L., Rödder, D., Radespiel, U., Tellier, A., & Schmitt, T. (2015). Population genetics revisited – Towards a multidisciplinary research field. *Biological Journal of the Linnean Society*, *115*, 1–12. https://doi.org/10.1111/bij.12481

Haller, B. C., & Messer, P. W. (2019). Slim 3: Forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution*, *36*(3), 632–637. https://doi.org/10.1093/molbev/msy228

Han, F., Lamichhaney, S., Rosemary Grant, B., Grant, P. R., Andersson, L., & Webster, M. T. (2017). Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Research*, *27*(6), 1004–1015. https://doi.org/10.1101/gr.212522.116

Hanghøj, K., Moltke, I., Andersen, P. A., Manica, A., & Korneliussen, T. S. (2019). Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *GigaScience*, *8*(5), 1–9. https://doi.org/10.1093/gigascience/giz034

Harris, K., & Nielsen, R. (2013). Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genetics*, *9*(6), https://doi.org/10.1371/journal.pgen.1003521

Harris, R. B., Sackman, A., & Jensen, J. D. (2018). On the unfounded enthusiasm for soft selective sweeps II: Examining recent evidence from humans, flies, and viruses. *PloS Genetics*, *14*(12), e1007859–, https://doi.org/10.1371/journal.pgen.1007859

Hartfield, M., Wright, S. I., & Agrawal, A. F. (2016). Coalescent times and patterns of genetic diversity in species with facultative sex: Effects of gene conversion, population structure, and heterogeneity. *Genetics*, *202*(1), 297–312. https://doi.org/10.1534/genetics.115.178004

Hedrick, P. W. (2006). Genetic polymorphism in heterogeneous environments: The age of genomics. *Annual Review of Ecology, Evolution, and Systematics*, *37*(1), 67–93. https://doi.org/10.1146/annurev.ecolsys.37.091305.110132

Hedrick, P. W. (2013). Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, *22*(18), 4606–4618. https://doi.org/10.1111/mec.12415

Hejase, H. A., Salman-Minkov, A., Campagna, L., Hubisz, M. J., Lovette, I. J., Gronau, I., & Siepel, A. (2020). Genomic islands of differentiation in a rapid avian radiation have been driven by recent selective sweeps. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(48), 30554–30565. https://doi.org/10.1073/pnas.2015987117

Heled, J., & Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, *27*(3), 570–580. https://doi.org/10.1093/molbev/msp274

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*(February), 747–751. https://doi.org/10.1126/science.1243518

Hendry, A. P. (2004). Selection against migrants contributes to the rapid evolution of ecologically dependent reproductive isolation. *Evolutionary Ecology Research*, *6*(8), 1219–1236.

Hohenlohe, P. A. (2014). Ecological genomics in full colour. *Molecular Ecology*, *23*(21), 5129–5131. https://doi.org/10.1111/mec.12945

Hubisz, M. J., Williams, A. L., & Siepel, A. (2020). Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS Genetics*, *16*(8), 1–24. https://doi.org/10.1371/JOURNAL.PGEN.1008895

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Huisman, J. (2017). Pedigree reconstruction from SNP data: Parentage assignment, sibship clustering and beyond. *Molecular Ecology Resources*, *17*(5), 1009–1024. https://doi.org/10.1111/1755-0998.12665

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*(2), 254–267. https://doi.org/10.1093/molbev/msj030

Jackson, N. D., Morales, A. E., Carstens, B. C., & O'Meara, B. C. (2017). PHRAPL: Phylogeographic inference using approximate likelihoods. *Systematic Biology*, *66*(6), 1045–1053. https://doi.org/10.1093/sysbio/syx001

Jenkins, P. A., Song, Y. S., & Brem, R. B. (2012). Genealogy-based methods for inference of historical recombination and gene flow and their application in *Saccharomyces cerevisiae*. *PLoS One*, *7*(11), e46947. https://doi.org/10.1371/journal.pone.0046947

Jenner, R. A., & Wills, M. A. (2007). The choice of model organisms in evo-devo. *Nature Reviews. Genetics*, *8*(4), 311–319. https://doi.org/10.1038/nrg2062

Jensen, J. D. (2014). On the unfounded enthusiasm for soft selective sweeps. *Nature Communications*, *5*, 5281. https://doi.org/10.1038/ncomms6281

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. https://doi.org/10.1186/1471-2156-11-94

Jombart, T., Devillard, S., Dufour, A.-B., & Pontier, D. (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, *101*, 92–103. https://doi.org/10.1038/hdy.2008.34

Jombart, T., Pontier, D., & Dufour, A.-B. (2009). Genetic markers in the playground of multivariate analysis. *Heredity*, *102*, 330–341. https://doi.org/10.1038/hdy.2008.130

Joseph, T. A., & Pe'er, I. (2019). Inference of population structure from time-series genotype data. *American Journal of Human Genetics*, *105*(2), 317–333. https://doi.org/10.1016/j.ajhg.2019.06.002

Jostins, L., & McVean, G. (2016). Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes. *Bioinformatics*, *32*(12), 1898–1900. https://doi.org/10.1093/bioinformatics/btw075

Jouganous, J., Long, W., Ragsdale, A. P., & Gravel, S. (2017). Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. *Genetics*, *206*(3), 1549–1567. https://doi.org/10.1534/genetics.117.200493

Kamm, J., Terhorst, J., Durbin, R., & Song, Y. S. (2020). Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, *115*(531), 1472–1487. https://doi.org/10.1080/01621459.2019.1635482

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, *12*(5), e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Kemppainen, P., Knight, C. G., Sarma, D. K., Hlaing, T., Prakash, A., Maung Maung, Y. N., Somboon, P., Mahanta, J., & Walton, C. (2015). Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Molecular Ecology Resources*, *15*(5), 1031–1045. https://doi.org/10.1111/1755-0998.12369

Kern, A. D., & Schrider, D. R. (2016). Discoal: Flexible coalescent simulations with selection. *Bioinformatics*, *32*(24), 3839–3841. https://doi.org/10.1093/bioinformatics/btw556

Kern, A. D., & Schrider, D. R. (2018). diploS/HIC: An updated approach to classifying selective sweeps. *G3 Genes|genomes|genetics*, *8*(6), 1959–1970. https://doi.org/10.1534/g3.118.200262

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., Kosiol, C., & Schlötterer, C. (2011). PoPoolation: A toolbox for population genetic analysis of next generation sequencing

data from pooled individuals. *PLoS One*, *6*(1), e15925. https://doi.org/10.1371/journal.pone.0015925

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*, *27*(24), 3435–3436. https://doi.org/10.1093/bioinformatics/btr589

Kolaczkowski, B., Kern, A. D., Holloway, A. K., & Begun, D. J. (2011). Genomic differentiation between temperate and tropical Australian populations of *Drosophila melanogaster*. *Genetics*, *187*(1), 245–260. https://doi.org/10.1534/genetics.110.123059

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Koropoulis, A., Alachiotis, N., & Pavlidis, P. (2020). Detecting positive selection in populations using genetic data. In J. Y. Dutheil (Ed.). *Statistical population genomics* (pp. 87–123). Springer. https://doi.org/10.1007/978-1-0716-0199-0_5

Kubota, S., Iwasaki, T., Hanada, K., Nagano, A. J., Fujiyama, A., Toyoda, A., Sugano, S., Suzuki, Y., Hikosaka, K., Ito, M., & Morinaga, S.-I. (2015). A genome scan for genes underlying microgeographic-scale local adaptation in a wild arabidopsis species. *PLoS Genetics*, *11*(7), 1–26. https://doi.org/10.1371/journal.pgen.1005361

Kuhlwilm, M., Gronau, I., Hubisz, M. J., de Filippo, C., Prado-Martinez, J., Kircher, M., & Castellano, S. (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*, *530*(7591), 429–433. https://doi.org/10.1038/nature16544

Laland, K. N., Sterelny, K., Odling-Smee, J., Hoppitt, W., & Uller, T. (2011). Cause and effect in biology revisited: Is Mayr's proximate-ultimate dichotomy still useful? *Science (New York, N.Y.)*, *334*(6062), 1512–1516. https://doi.org/10.1126/science.1210879

Lawson, D. J., van van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat Commun*, 066431, *9*, 3258https://doi.org/10.1038/s41467-018-05257-7

Leaché, A. D., Banbury, B. L., Felsenstein, J., De Oca, A. N. M., & Stamatakis, A. (2015). Short tree, long tree, right tree, wrong tree: New acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology*, *64*, 1032–1047.

Leaché, A. D., Harris, R. B., Rannala, B., & Yang, Z. (2014). The influence of gene flow on species tree estimation: A simulation study. *Systematic Biology*, *63*, 17–30.

Lee, T.-H., Guo, H., Wang, X., Kim, C., & Paterson, A. H. (2014). SNPhylo: A pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics*, *15*(1), 162. https://doi.org/10.1186/1471-2164-15-162

Legendre, P., & Fortin, M. J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources*, *10*(5), 831–844. https://doi.org/10.1111/j.1755-0998.2010.02866.x

Legrand, D., Tenaillon, M. I., Matyot, P., Gerlach, J., Lachaise, D., & Cariou, M.-L. (2009). Species-wide genetic variation and demographic history of *Drosophila sechellia*, a species lacking population structure. *Genetics*, *182*(4), 1197–1206. https://doi.org/10.1534/genetics.108.092080

Leitwein, M., Duranton, M., Rougemont, Q., Gagnaire, P. A., & Bernatchez, L. (2020). Using haplotype information for conservation genomics. *Trends in Ecology and Evolution*, *35*(3), 245–258. https://doi.org/10.1016/j.tree.2019.10.012

Lemmon, E. M., & Lemmon, A. R. (2013). High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 99–121.

Lewontin, R. C., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, *74*(1), 175–195. https://doi.org/10.1093/genetics/74.1.175

Lexer, C., Mangili, S., Bossolini, E., Forest, F., Stölting, K. N., Pearman, P. B., Zimmermann, N. E., & Salamin, N. (2013). 'Next generation' biogeography: Towards understanding the drivers of species diversification and persistence. *Journal of Biogeography*, *40*(6), 1013–1022. https://doi.org/10.1111/jbi.12076

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Li, J., Li, H., Jakobsson, M., Li, S., SjÖdin, P., & Lascoux, M. (2012). Joint analysis of demography and selection in population genetics: Where do we stand and where could we go? *Molecular Ecology*, *21*(1), 28–44. https://doi.org/10.1111/j.1365-294X.2011.05308.x

Librado, P., & Orlando, L. (2018). Detecting signatures of positive selection along defined branches of a population tree using LSD. *Molecular Biology and Evolution*, *35*, 1520–1535. https://doi.org/10.1093/molbev/msy053

Link, V., Kousathanas, A., Veeramah, K., Sell, C., Scheu, A., & Wegmann, D. (2017). ATLAS: Analysis tools for low-depth and ancient samples. *BioRxiv*. https://doi.org/10.1101/105346

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298–299. https://doi.org/10.1093/bioinformatics/btr642

Liu, H. J., & Yan, J. (2019). Crop genome-wide association study: A harvest of biological relevance. *Plant Journal*, *97*(1), 8–18. https://doi.org/10.1111/tpj.14139

Liu, L., & Yu, L. (2010). Phybase: An R package for species tree analysis. *Bioinformatics*, *26*(7), 962–963. https://doi.org/10.1093/bioinformatics/btq062

Liu, L., & Yu, L. (2011). Estimating species trees from unrooted gene trees. *Systematic Biology*, *60*(5), 661–667. https://doi.org/10.1093/sysbio/syr027

Liu, X., & Fu, Y. X. (2020). Stairway Plot 2: Demographic history inference with folded SNP frequency spectra. *Genome Biology*, *21*(1), 1–9. https://doi.org/10.1186/s13059-020-02196-9

Lotterhos, K. E., & Whitlock, M. C. (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, *23*(9), 2178–2192. https://doi.org/10.1111/mec.12725

Malinsky, M., Matschiner, M., & Svardal, H. (2021). Dsuite – Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, *21*, 584–595. https://doi.org/10.1111/1755-0998.13265

Mandoli, D. F., & Olmstead, R. (2000). The importance of emerging model systems in plant biology. *Journal of Plant Growth Regulation*, *19*(3), 249–252. https://doi.org/10.1007/s003440000038

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, *26*(22), 2867–2873. https://doi.org/10.1093/bioinformatics/btq559

Martin, S. H., & Van Belleghem, S. M. (2017). Exploring evolutionary relationships across the genome using topology weighting. *Genetics*, *206*(1), 429–438. https://doi.org/10.1534/genetics.116.194720

Mazet, O., Rodriguez, W., & Chikhi, L. (2015). Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theoretical Population Biology*, *104*, 46–58. https://doi.org/10.1016/j.tpb.2015.06.003

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature*, *351*(6328), 652–654. https://doi.org/10.1038/351652a0

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genetics*, *5*(10), e1000686. https://doi.org/10.1371/journal.pgen.1000686

McVean, G., Awadalla, P., & Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene

sequences. *Genetics*, *160*(3), 1231–1241. https://doi.org/10.1093/genetics/160.3.1231

Messer, P. W., & Petrov, D. A. (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends in Ecology and Evolution*, *28*(11), 659–669. https://doi.org/10.1016/j.tree.2013.08.003

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Teeling, E. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, *31*(12), i44–i52. https://doi.org/10.1093/bioinformatics/btv234

Mirzaei, S., & Wu, Y. (2017). RENT+: An improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics*, *33*(7), 1021–1030. https://doi.org/10.1093/bioinformatics/btw735

Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. *Science*, *310*(5746), 321–324. https://doi.org/10.1126/science.1117196

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(10), 5269–5273. https://doi.org/10.1073/pnas.76.10.5269

Neuenschwander, S., Michaud, F., & Goudet, J. (2019). QuantiNemo 2: A Swiss knife to simulate complex demographic and genetic scenarios, forward and backward in time. *Bioinformatics*, *35*(5), 886–888. https://doi.org/10.1093/bioinformatics/bty737

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., & Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, *8*(11), 857–868. https://doi.org/10.1038/nrg2187

Noskova, E., Ulyantsev, V., Koepfli, K.-P., O'Brien, S. J., & Dobrynin, P. (2020). GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. *GigaScience*, *9*(3), 1–18. https://doi.org/10.1093/gigascience/giaa005

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., & Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), 98–101. https://doi.org/10.1038/nature07331

Orozco-terWengel, P. (2016). The devil is in the details: The effect of population structure on demographic inference. *Heredity*, *116*(4), 349–350. https://doi.org/10.1038/hdy.2016.9

Palamara, P. F., & Pe'er, I. (2013). Inference of historical migration rates via haplotype sharing. *Bioinformatics*, *29*(13), 180–188. https://doi.org/10.1093/bioinformatics/btt239

Pavlidis, P., Jensen, J. D., Stephan, W., & Stamatakis, A. (2012). A critical assessment of storytelling: Gene ontology categories and the importance of validating genomic scans. *Molecular Biology and Evolution*, *29*(10), 3237–3248. https://doi.org/10.1093/molbev/mss136

Pavlidis, P., Laurent, S., & Stephan, W. (2010). MsABC: A modification of Hudson's ms to facilitate multi-locus ABC analysis. *Molecular Ecology Resources*, *10*(4), 723–727. https://doi.org/10.1111/j.1755-0998.2010.02832.x

Petkova, D., Novembre, J., & Stephens, M. (2015). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, *48*(1), 94–100. https://doi.org/10.1038/ng.3464

Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, *31*(7), 1929–1936. https://doi.org/10.1093/molbev/msu136

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, *8*(11), e1002967. https://doi.org/10.1371/journal.pgen.1002967

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Baglione, V., & Wolf, J. B. W. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, *344*(6190), 1410–1414.

Price, A., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Pritchard, J. K., & Di Rienzo, A. (2010). Adaptation – Not by sweeps alone. *Nature Reviews Genetics*, *11*(10), 665–667. https://doi.org/10.1038/nrg2880

Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, *20*(4), R208–R215. https://doi.org/10.1016/j.cub.2009.11.055

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Puttick, M. N. (2019). MCMCtreeR: Functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics*, *35*(24), 5321–5322. https://doi.org/10.1093/bioinformatics/btz554

Raj, A., Stephens, M., & Pritchard, J. K. (2014). FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Rasmussen, M. D., Hubisz, M. J., Gronau, I., & Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, *10*(5), https://doi.org/10.1371/journal.pgen.1004342

Ravinet, M., Faria, R., Butlin, R. K., Galindo, J., Bierne, N., Rafajlović, M., & Westram, A. M. (2017). Interpreting the genomic landscape of speciation: Finding barriers to gene flow. *Journal of Evolutionary Biology*, *30*, 1450–1477. https://doi.org/10.1111/jeb.13047

Raynal, L., Marin, J. M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2019). ABC random forests for Bayesian parameter inference. *Bioinformatics*, *35*(10), 1720–1728. https://doi.org/10.1093/bioinformatics/bty867

Refoyo-Martínez, A., Da Fonseca, R. R., Halldórsdóttir, K., Árnason, E., Mailund, T., & Racimo, F. (2019). Identifying loci under positive selection in complex population histories. *Genome Research*, *29*(9), 1506–1520. https://doi.org/10.1101/gr.246777.118

Robert, C. P., Cornuet, J.-M., Marin, J.-M., & Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(37), 15112–15117. https://doi.org/10.1073/pnas.1102900108

Rockman, M. V. (2012). The QTN program and the alleles that matter for evolution: All that's gold does not glitter. *Evolution*, *66*(1), 1–17. https://doi.org/10.1111/j.1558-5646.2011.01486.x

Rodríguez, W., Mazet, O., Grusea, S., Arredondo, A., Corujo, J. M., Boitard, S., & Chikhi, L. (2018). The IICR and the non-stationary structured coalescent: Towards demographic inference with arbitrary changes in population structure. *Heredity*, *121*(6), 663–678. https://doi.org/10.1038/s41437-018-0148-0

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., & Bierne, N. (2016). Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLoS Biology*, *14*(12), e2000234. https://doi.org/10.1371/JOURNAL.PBIO.2000234

Roux, C., Pauwels, M., Ruggiero, M.-V., Charlesworth, D., Castric, V., & Vekemans, X. (2013). Recent and ancient signature of balancing selection around the S-locus in *Arabidopsis halleri* and *A.*

lyrata. *Molecular Biology and Evolution*, *30*(2), 435–447. https://doi.org/10.1093/molbev/mss246

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., & Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, *419*(6909), 832–837. https://doi.org/10.1038/nature01027.1

Sabeti, P. C., Varilly, P., Fry, B., McCarroll, S. A., Frazer, K. A., Ballinger, D. G., & Stewart, J. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913–918. https://doi.org/10.1038/nature06250

Salter-Townshend, M., & Myers, S. (2019). Fine-scale inference of ancestry segments without prior knowledge of admixing groups. *Genetics*, *212*(July), 869–889. https://doi.org/10.1534/genetics.119.302139

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, *78*(4), 629–644. https://doi.org/10.1086/502802

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*(8), 919–925. https://doi.org/10.1038/ng.3015

Schrider, D. R., & Kern, A. D. (2016). S/HIC: Robust identification of soft and hard sweeps using machine learning. *PLoS Genetics*, *12*(3), 1–31. https://doi.org/10.1371/journal.pgen.1005928

Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics*, *34*(4), 301–312. https://doi.org/10.1016/j.tig.2017.12.005

Schrider, D. R., Mendes, F. K., Hahn, M. W., & Kern, A. D. (2015). Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics*, *200*(1), 267–284. https://doi.org/10.1534/genetics.115.174912

Schrider, D. R., Shanku, A. G., & Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, *204*(3), 1207–1223. https://doi.org/10.1534/genetics.116.190223

Schubert, M., Jónsson, H., Chang, D., Der Sarkissian, C., Ermini, L., Ginolhac, A., & Orlando, L. (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(52), 201416991. https://doi.org/10.1073/pnas.1416991111

Sellis, D., Callahan, B. J., Petrov, D. A., & Messer, P. W. (2011). Heterozygote advantage as a natural consequence of adaptation in diploids. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(51), 20666–20671. https://doi.org/10.1073/pnas.1114573108

Setter, D., Mousset, S., Cheng, X., Nielsen, R., DeGiorgio, M., & Hermisson, J. (2020). VolcanoFinder: Genomic scans for adaptive introgression. *PLoS Genetics*, *16*(6), 1–44. https://doi.org/10.1371/journal.pgen.1008867

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., Colling, G., Dalén, L., De Meester, L., Ekblom, R., Fawcett, K. D., Fior, S., Hajibabaei, M., Hill, J. A., Hoezel, A. R., Höglund, J., Jensen, E. L., Krause, J., Kristensen, T. N., … Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. *Trends in Ecology & Evolution*, *30*(2), 78–87. https://doi.org/10.1016/j.tree.2014.11.009

Sheehan, S., Harris, K., & Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. *Genetics*, *194*, 647–662. https://doi.org/10.1534/genetics.112.149096

Sheehan, S., & Song, Y. S. (2016). Deep learning for population genetic inference. *PLoS Computational Biology*, *12*(3), 1–28. https://doi.org/10.1371/journal.pcbi.1004845

Siewert, K. M., & Voight, B. F. (2017). Detecting long-term balancing selection using allele frequency correlation. *Molecular Biology and Evolution*, *34*(11), 2996–3005. https://doi.org/10.1093/molbev/msx209

Siewert, K. M., & Voight, B. F. (2020). BetaScan2: Standardized statistics to detect balancing selection utilizing substitution data. *Genome Biology and Evolution*, *12*(2), 3873–3877. https://doi.org/10.1093/gbe/evaa013

Smith, C. C. R., & Flaxman, S. M. (2020). Leveraging whole genome sequencing data for demographic inference with approximate Bayesian computation. *Molecular Ecology Resources*, *20*(1), 125–139. https://doi.org/10.1111/1755-0998.13092

Song, Y., Endepols, S., Klemann, N., Richter, D., Matuschka, F.-R., Shih, C.-H., Nachman, M. W., & Kohn, M. H. (2011). Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Current Biology*, *21*(15), 1296–1301. https://doi.org/10.1016/j.cub.2011.06.043

Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, *51*(9), 1321–1329. https://doi.org/10.1038/s41588-019-0484-x

Staab, P. R., & Metzler, D. (2016). Coala: An R framework for coalescent simulation. *Bioinformatics*, *32*(12), 1903–1904. https://doi.org/10.1093/bioinformatics/btw098

Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). Scrm: Efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, *31*(10), 1680–1682. https://doi.org/10.1093/bioinformatics/btu861

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stamatakis, A., Alachiotis, N., Pavlidis, P., & Daniel, Z. (2013). SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. *Molecular Biology and Evolution*, *30*(9), 2224–2234. https://doi.org/10.1093/molbev/mst112

Stern, A. J., Speidel, L., Zaitlen, N. A., & Nielsen, R. (2021). Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies. *The American Journal of Human Genetics*, *108*(2), 219–239, https://doi.org/10.1016/j.ajhg.2020.12.005

Stern, A. J., Wilton, P. R., & Nielsen, R. (2019). An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLoS Genetics*, *15*(9), e1008384. https://doi.org/10.1371/journal.pgen.1008384

Stucki, S., Orozco-terWengel, P. Forester, B. R. Duruz, S. Colli, L. Masembe, C., Negrini, R., Landguth, E., Jones, M. R., The NEXTGEN Consortium, Bruford, M. W., Taberlet, P., & Joost, S. (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources*, *17*(5), 1072–1089. https://doi.org/10.1111/1755-0998.12629

Sugden, L. A., Atkinson, E. G., Fischer, A. P., Rong, S., Henn, B. M., & Ramachandran, S. (2018). Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nature Communications*, *9*(703), https://doi.org/10.1038/s41467-018-03100-7

Svedberg, J., Shchur, V., Reinman, S., Nielsen, R., & Corbett-Detig, R. (2021). Inferring Adaptive Introgression Using Hidden Markov Models. *Molecular Biology and Evolution*, *38*(5), 2152–2165. https://doi.org/10.1093/molbev/msab014

Szpiech, Z. A., & Hernandez, R. D. (2014). selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, *31*(10), 2824–2827. https://doi.org/10.1093/molbev/msu211

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

Takezaki, N., Nei, M., & Tamura, K. (2010). POPTREE2: Software for constructing population trees from allele frequency data and

computing other population statistics with windows interface. *Molecular Biology and Evolution*, *27*(4), 747–752. https://doi.org/10.1093/molbev/msp312

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1

Tang, K., Thornton, K. R., & Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biology*, *5*(7), 1587–1602. https://doi.org/10.1371/journal.pbio.0050171

Tataru, P., & Bataillon, T. (2019). PolyDFEv2.0: Testing for invariance of the distribution of fitness effects within and across species. *Bioinformatics*, *35*(16), 2868–2869. https://doi.org/10.1093/bioinformatics/bty1060

Tataru, P., & Bataillon, T. (2020). polyDFE: Inferring the distribution of fitness effects and properties of beneficial mutations from polymorphism data. In J. Y. Dutheil (Ed.). *Statistical population genomics* (pp. 125–146). Springer. https://doi.org/10.1007/978-1-0716-0199-0_6

Tataru, P., Nirody, J. A., & Song, Y. S. (2014). DiCal-IBD: Demography-aware inference of identity-by-descent tracts in unrelated individuals. *Bioinformatics*, *30*(23), 3430–3431. https://doi.org/10.1093/bioinformatics/btu563

Taylor, H. R., Dussex, N., & van Heezik, Y. (2017). Bridging the conservation genetics gap by identifying barriers to implementation for conservation practitioners. *Global Ecology and Conservation*, *10*, 231–242. https://doi.org/10.1016/j.gecco.2017.04.001

Terhorst, J., Kamm, J. A., & Song, Y. S. (2016). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, *49*(2), 303–309. https://doi.org/10.1038/ng.3748

The Heliconius Genome Consortium, Dasmahapatra, K. K., Walters, J. R., Briscoe, A. D., Davey, J. W., Whibley, A., & Jiggins, C. D. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, *487*(7405), 94–98. https://doi.org/10.1038/nature11041

Torada, L., Lorenzon, L., Beddis, A., Isildak, U., Pattini, L., Mathieson, S., & Fumagalli, M. (2019). ImaGene: A convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics*, *20*(Suppl 9), 1–12. https://doi.org/10.1186/s12859-019-2927-x

Vachaspati, P., & Warnow, T. (2018). SVDquest: Improving SVDquartets species tree estimation using exact optimization within a constrained search space. *Molecular Phylogenetics and Evolution*, *124*, 122–136. https://doi.org/10.1016/j.ympev.2018.03.006

Vitalis, R., Gautier, M., Dawson, K. J., & Beaumont, M. A. (2014). Detecting and measuring selection from gene frequency data. *Genetics*, *196*(3), 799–817. https://doi.org/10.1534/genetics.113.152991

Wagner, C. E., Keller, I., Wittwer, S., Selz, O. M., Mwaiko, S., Greuter, L., & Seehausen, O. (2013). Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, *22*(3), 787–798. https://doi.org/10.1111/mec.12023

Wang, J. (2019). Pedigree reconstruction from poor quality genotype data. *Heredity*, *122*(6), 719–728. https://doi.org/10.1038/s41437-018-0178-7

Wang, J., & Zhang, Z. (2020). GAPIT version 3: Boosting power and accuracy for genomic association and prediction. *BioRxiv*, https://doi.org/10.1101/2020.11.29.403170

Wang, M. H., Cordell, H. J., & Van Steen, K. (2019). Statistical methods for genome-wide association studies. *Seminars in Cancer Biology*, *55*, 53–60. https://doi.org/10.1016/j.semcancer.2018.04.008

Wang, M., Huang, X., Li, R., Xu, H., Jin, L., & He, Y. (2014). Detecting recent positive selection with high accuracy and reliability by conditional coalescent tree. *Molecular Biology and Evolution*, *31*(11), 3068–3080. https://doi.org/10.1093/molbev/msu244

Weber, J. N., Peterson, B. K., & Hoekstra, H. E. (2013). Discrete genetic modules are responsible for complex burrow evolution in Peromyscus mice. *Nature*, *493*(7432), 402–405. https://doi.org/10.1038/nature11816

Wegmann, D., Leuenberger, C., Neuenschwander, S., & Excoffier, L. (2010). ABCtoolbox: A versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, *11*, 116. https://doi.org/10.1186/1471-2105-11-116

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, *38*(6), 1358–1370.

Wen, J., Liu, J., Ge, S., Xiang, Q.-Y., & Zimmer, E. A. (2015). Phylogenomic approaches to deciphering the tree of life. *Journal of Systematics Evolution*, *53*, 369–370. https://doi.org/10.1111/jse.12175

White, B. J., Cheng, C., Simard, F., Costantini, C., & Besansky, N. J. (2010). Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Molecular Ecology*, *19*(5), 925–939. https://doi.org/10.1111/j.1365-294X.2010.04531.x

Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H., & Reich, D. (2012). Phasing of many thousands of genotyped samples. *American Journal of Human Genetics*, *91*(2), 238–251. https://doi.org/10.1016/j.ajhg.2012.06.013

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. https://doi.org/10.1038/ng.2310

Jombart, T., Pontier, D., & Dufour, A. -B. (2009). Genetic markers in the playground of multivariate analysis. *Heredity (Edinb)*, *102*, 330–341.

How to cite this article: Bourgeois YX, Warren BH. An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes. *Mol Ecol*. 2021;00:1–36. https://doi.org/10.1111/mec.15989