Going down the rabbit hole: a review on methods for population genomics in natural populations

Running title: Linking evolution and genome-wide data

Yann X.C. Bourgeois, Khaled M. Hazzouri

New York University Abu Dhabi, PO Box 129188, Saadiyat Island,Abu Dhabi, United Arab Emirates

Abstract

1. Characterizing species history and identifying loci underlying local adaptation is crucial in functional ecology, evolutionary biology, conservation and agronomy. The ongoing and constant improvement of next-generation sequencing (NGS) techniques has facilitated the production of an increasingly growing amount of genetic markers across genomes of non-model species.
2. The study of variation at these markers across natural populations has deepened the understanding of how population history and selection act on genomes. However, this improvement has come with a burst of analytical tools that can confuse users.
3. This confusion can limit the amount of information effectively retrieved from complex genomic datasets. In addition, the lack of a unified analytical pipeline impairs the diffusion of the most recent analytical tools into fields like conservation biology.
4. This requires efforts be made in providing introduction to these methods. In this paper we describe possible analytical protocols and list more than 70 methods dealing with genome-scale datasets, clarify the strategy they use to infer demographic history and selection, and discuss some of their limitations.

Keywords

Coalescent, Software, Molecular evolution

Introduction

Multiple historical and selective factors shape the genetic makeup of populations. The advent of Next-Generation Sequencing (NGS) in the last 20 years has enhanced our understanding on how intermingled these factors are, and how they can impact genomic variation. Important results have been gathered on model species, or species of economic interest. Such results include, among other examples, an improved perspective on human history of migrations, admixture and adaptation (e.g. Sabeti *et al.*, 2002; Abi-Rached *et al.*, 2011; Li and Durbin, 2011), elucidating the origin of domesticated species (e.g. Axelsson *et al.*, 2013; Schubert *et al.*, 2014), or characterizing the genetic bases of local adaptation in model or near-model species (e.g. Legrand *et al.*, 2009; Kolaczkowski *et al.*, 2011; Roux *et al.*, 2013; Kubota *et al.*, 2015). The amount of population genomic data that is aimed at elucidating the history of natural populations is now becoming very abundant and widespread, even for non-model ones. Studying genetic variation at the genome level allows to characterize how demographic factors shape species history. In return, the picture of demographic events is informative to robustly identify loci under selection; and even can help in conservation efforts by identifying locally adapted genes that can be used to define relevant conservation units (Fraser and Bernatchez, 2001).

Developments in NGS have so far constantly improved the throughput and quality of data, while reducing time and cost of their production. While they have become more affordable for teams studying evolutionary processes in biology, many methods to infer demography and selection have been developed. Consequently, this increase in data production has come at an analytical cost, with an inflation of methods each claiming to address specific issues, making difficult to follow the ongoing developments in the field. In addition, the widespread use of sophisticated analytical tools remains challenged by the lack of communication between fields (Shafer *et al.*, 2015), little user-friendliness of software and the ever-increasing amount of tools made available. Nevertheless, translation of these methods into non-model species is part of a shift towards genomics in evolutionary sciences that aims at better understanding biological diversity at various scales (Mandoli and Olmstead, 2000; Jenner and Wills, 2007; Abzhanov *et al.*, 2008). Recent breakthroughs brought by the study of initially non-model species (e.g. White *et al.*, 2010; Ellegren *et al.*, 2012; Weber *et al.*, 2013; Poelstra *et al.*, 2014) have confirmed the value of population genomics from this perspective. These advances are needed to broaden our view about the evolutionary process and improve sampling of distant clades. Ultimately, this process should provide a more balanced picture than the one brought

by the study of a few model species (Abzhanov *et al.*, 2008). Genomic approaches also have the potential to improve conservation genetic inference by scaling up the amount of data available (Shafer *et al.*, 2015). Much effort has been put recently in facilitating the diffusion of sometimes complex, state-of-the-art methods and their application to species with little background has nonetheless become more accessible, and has the potential to bring valuable information.

In this paper, we propose a decision-making pipeline (Figure 1) to help choosing appropriate methods dealing with questions in population genomics and genetics of adaptation in natural populations. We begin with a succinct review of methods available to obtain genome-wide polymorphism data (Box 1) before focusing on i) methods devoted to the study of population structure and identification of selected loci (Table 1 and 2) and ii) methods aiming at quantitatively characterize population structure (Table 3). We end this review by detailing how these analyses can be combined and provide perspective on the use of these methods for non-SNP datasets. Tables and summary of methods from this paper will be kept updated to follow improvements, and are available at www.methodspopgen.com.

Box 1. Common sequencing methods

Whole genome resequencing : Whole-genome resequencing requires a reference (at least at a draft stage) and is much more expensive, especially for species with long and complex genomes. However, this approach gives a complete overview of structural and coding variation, and allows some of the most powerful methods currently available to track signatures of selection (see below). Pooled sequencing (Futschik and Schlötterer, 2010) can be an option to reduce the costs, but restricts the analysis to methods focusing on allele frequencies. Since individual information is not available, variation in Linkage Disequilibrium across individuals (LD) cannot be exploited. Shallow sequencing (1-5X per individual) may be a way to partly overpass this last issue for a similar cost (Buerkle and Gompert, 2013), but should not be used for methods requiring phasing and unbiased individual genotypes. Shallow shotgun sequencing also allows retrieving complete plastomes, due to the representation bias of mitochondrial or chloroplast sequences. Plastome sequences can provide insightful information about the evolutionary history of populations or species. Recent work has successfully used shallow sequencing to reconstruct mitochondrial or chloroplast sequences in plants (Malé *et al.*, 2014), animals (Hahn *et al.*, 2013) or old and

altered museum samples (Besnard *et al.*, 2016). Methods such as MITObim (Hahn *et al.*, 2013) provide an automated and relatively user-friendly way to reconstitute plastome sequences, which can then be analyzed as a single non-recombining marker for phylogeny or population genetics.

RNAseq: RNAseq can be used with and without a reference genome. In the latter case, like any other reduced representation method, it does not provide information of linkage among genes. It has many applications along the evolutionary time scale. Since it is mostly coding regions, a deep phylogeny can be constructed with conserved orthologs. Depth of coverage is gene expression dependent, so calling genotypes varies across genes and it should be taking into consideration (Gayral *et al.*, 2013). RNAseq provide information about the regulation of biochemical process and pathways between different tissues of the same individual and between individuals and across different environments. Applying RNAseq in non-model organism to estimate differentially expressed genes can be challenging. One reason is because biological variance is much higher in the field studies than in controlled conditions, which requires more sampling of individuals to achieve statistical power significance (Todd *et al.*, 2016). If a reference genome is available, it is possible to call variants (Piskol *et al.*, 2013) and estimate differential gene expression with gene annotation (Love *et al.*, 2014). This can be done from alignment using RNAseq tools such Tophat2 (Kim *et al.*, 2013), HiSat (Kim *et al.*, 2015) or STAR (Dobin *et al.*, 2013) hat allow for splicing reads. Using a reference genome to bridge regions with low coverage, software like Cufflinks and more recent ones such as Bayesembler (Maretty *et al.*, 2014) and StringTie (Pertea *et al.*, 2015) are able to assemble more lowly expressed transcripts than a *de novo* approach (Maretty *et al.*, 2014). However, reference guided methods generally ignore variation from the reference since they focus on the structure of the exon overall in a transcript. Therefore, tools for de novo assembly such as Trinity, Oases, SOAPdenovo-Trans and Bridger are more suitable to retrieve information about variation for population genetic inferences when no reference genome is available (Grabherr *et al.*, 2011; Schulz *et al.*, 2012; Xie *et al.*, 2014; Chang *et al.*, 2015).

Targeted sequencing: This method facilitates the development of markers for a single species. Since the specificity of the probe does not have to be very high, the same probe can used with different related species (Nicholls *et al.*, 2015). Conservation of the target genomic region under study is important. High conservation may lead to higher efficiency of capture but can artificially reduce representation of polymorphic regions. Different technologies allow for

targeted sequence capture that can be classified by enrichment methods (hybridization-based; PCR-based; molecular inversion probe-based, see (Mamanova *et al.*, 2010)(Mamanova et al. 2010). Targeted sequencing reduces the genomic representation compared to whole genome sequencing and it allows for multiple individuals to be multiplexed, lowering the cost of sequencing per sample. In addition, the analysis complexity is reduced compared to WGS since only a subset of genomic regions are sequenced. By allowing an improvement to spatial and temporal sampling, targeted sequencing could reconstruct dispersal routes and migration between varieties and subspecies (Nadeau *et al.*, 2012; da Fonseca *et al.*, 2016).

RADseq: Reduced representation allows sampling homogeneously variants across the genome by sequencing DNA fragments flanking restriction sites. Some of the best-known reduced representation techniques include RAD-sequencing (Baird *et al.*, 2008) and Genotyping by Sequencing or GBS (Elshire *et al.*, 2011). Their main interest is their relatively low cost and that they do not require any reference genome (see (Davey *et al.*, 2011) for details). They therefore allow to sequence many individuals at a low cost, making them widely used for the study of population structure, demography and selection. As a general word of caution, note that RAD-sequencing and related methods display specific properties that can bias genome-wide estimates of diversity, like allelic dropout (Arnold *et al.*, 2013). However, this type of markers remains valuable for phylogenetic estimation, even for distantly related species (Cariou *et al.*, 2013), and allelic dropout can be compensated by focusing only on markers sequenced in all individuals. Many pipelines have been specifically designed to account for RAD-seq specificities, including Stacks (Catchen *et al.*, 2011) or TASSEL-UNEAK (Lu *et al.*, 2013), facilitating the reproducibility of analyses. Reduced representation methods do not cover all mutations in the genome and are thus more likely to miss those actually under selection. Special care in the choice of the restriction enzyme and determining the expected density of markers is needed to retrieve enough mutations close to genes under selection. The number of SNPs ranges from thousands to millions, which is most of the time enough to retrieve substantial information about demography and sometimes selection (see Puritz *et al.*, 2014 for a detailed summary of reduced-representation techniques).

Population structure and data description

Exploring population structure

Many tools currently exist to infer population structure (Table 1). An elegant and efficient class of methods relies on using multivariate approaches to infer relatedness between individuals and populations without *a priori*. Since these methods do not have underlying assumptions based on population genetics, they are suitable to analyze species displaying polyploidy or mixed-ploidy (Dufresne *et al.*, 2014). A detailed review about these methods has been already performed (Jombart *et al.*, 2009) and an exhaustive list of their applications is beyond the scope of this review. A simple approach that does not assume any *a priori* grouping is the Principal Component Analysis (PCA), based on analyzing variance-covariance structure among genotypes, which can be performed on both individual and pooled data. These approaches have been especially useful to study the consistency between geographical and genetic structure in European human populations (Novembre *et al.*, 2008). A recent application of this technique using Procrustes rotation to match geographical coordinates with PCA axes has been performed on populations of a freshwater crustacean (*Daphnia magna*) using RAD-sequencing, allowing to show how isolation by distance shaped genetic structure.

Methods allowing to estimate the relatedness of individuals are suited for studies relying on pedigree information, or if there are reasons to suspect that familial relationships can play a major role in shaping genetic structure of the population(s) considered. When each individual in a study is sampled from a different location or environment, estimating relatedness also provides a way to assess the genetic distance between individuals, in relation with geographical or ecological distance. For example, in a recent study using more than 1000 *Arabidopsis thaliana* genomes, estimates of relatedness have allowed identifying possibly relict populations that may have subsisted in Europe during last Ice Age (Alonso-Blanco *et al.*, 2016). VCFTools (Danecek *et al.*, 2011) provides two ways calculating relatedness; unadjusted Ajk (Yang *et al.*, 2010) and a kinship coefficient also implemented in KING (Manichaikul *et al.*, 2010). Population stratification and relatedness can also be explored in PLINK based on pairwise identity-by-state (IBS) distance or identity by descent (IBD). These last methods can further be used to identify genomic regions displaying a high sharing of alleles, which can suggest positive selection (see below).

Landscape (and seascape) genetics has widely contributed to our understanding of how ecological and geographical variation affects species history and adaptation (Manel and Holderegger, 2013). Of central importance in this field is the identification of how populations are connected and how organisms move in the landscape matrix. Complementary to these approaches, identifying how and where populations (or closely related species, see Roux et al. 2016) hybridize is crucial when it comes to characterize colonization trajectories, tension zones and secondary contacts (Gay *et al.*, 2008; Bierne *et al.*, 2011).

Approaches such as Structure (Pritchard *et al.*, 2000) and fastSTRUCTURE (Raj *et al.*, 2014) have been widely popular in this framework, allowing to determine hierarchical population structure and admixed populations by grouping individuals in clusters without any *a priori*. The optimal number of clusters (K) can then be determined based on likelihood, although examining population structure for a range of K can allow better identifying substructure. Since these methods can be slow for large whole-genome data or high-density RAD-seq, reducing SNP redundancy by subsampling unlinked markers (those known to have low LD, e.g. due to a large physical distance between them) is a way to reduce computation time while keeping the relevant information. More precise inference and testing for directional introgression can be then performed in a software such as TreeMix (Pickrell and Pritchard, 2012).

Some methods can explicitly use spatial information to inform clustering. A practical R package is Geneland (Guillot *et al.*, 2012), that allows determining the optimal number of population in a dataset by optimizing linkage and Hardy-Weinberg equilibrium within clusters, and is also able to incorporate geographic coordinates in the model to delineate their spatial organization. It can be useful to characterize the location and shape of hybrid zones. On the other hand, methods such as BEDASSLE (Bradburd *et al.*, 2013) can be used to complement these approaches, and identify which combination of geographical and ecological distance limits dispersion. However, disentangling these effects has been proved to be complex and 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 (Edelaar and Bolnick, 2012; Bolnick and Otto, 2013) or selection against migrants (Hendry, 2004).

Phylogenetic methods such as RAxML (Stamatakis, 2014) or BEAST2 (Drummond and Rambaut, 2007) have been popular to cluster individuals into populations at the species level. Their underlying assumptions (e.g. homoplasia occurs through mutation, not recombination)

should however be restricted to complexes of populations with low gene flow or ongoing recombination and with sufficient divergence, since even methods dedicated to reconstruct species tree such as BEAST* can be strongly biased when it comes to estimate divergence times and effective population sizes (Leaché *et al.*, 2014). Methods implemented in Splitstree (Huson and Bryant, 2006), make less assumptions and are therefore more suited for building networks linking individuals. While useful to infer topologies, caution is advised when using branches lengths obtained from SNP-only datasets, e.g. to calculate divergence times between different groups or species (Leaché *et al.*, 2015). For this purpose, it might therefore be easier to extract from the data both variant and invariant sites at several genes or RAD contigs, and analyze the whole sequences in a software like BEAST2.

To assess how diversity is partitioned across the different groups inferred by the methods described previously, it is advisable to perform an Analysis of Molecular Variance (AMOVA). Arlequin (Excoffier and Lischer, 2010) is particularly suited for this task. More generally, investigating patterns of nucleotide diversity, inbreeding, F_{st} or variation in LD between populations and across the genome is useful to have a preliminary idea of gene flow, admixture and variation in population sizes. These statistics can be easily retrieved with VCFTools or PopGenome (Pfeifer *et al.*, 2014). If a reference genome is available, these statistics can also allow to scan for regions under selection or those more likely to display introgression while controlling for recombination (e.g. with LDHat, Table 1).

Screening for selection and association

Selection and its impact on sequence variation

Checking for population structure is an essential step when performing analyses on genome-level datasets. Neglecting it can bias demographic inferences (Chikhi *et al.*, 2010; Heller *et al.*, 2013) or the detection of loci under selection (e.g. Nielsen *et al.*, 2007); thus, checking for outlier individuals and assessing the global structure is required prior to any more sophisticated analysis. On the other hand, selection acts both on correlations i) between alleles and environment at selected loci and ii) between alleles from different loci, either directly under selection or not. This is reflected respectively by i) variation in polymorphism within and between populations and ii) linkage disequilibrium (LD) between loci. If selection is

widespread in the genome, the study of population history can thererore be biased, making necessary the joint study of selection and population structure.

In the sections that follow we present tools that can be used to detect signatures of selection (Table 2), but are also informative to assess how heterogeneous variation at a genome scale can be, an information that can be used to retrieve, e.g., signatures of introgression or identifying loci involved in reproductive isolation. The methods that these tools implement fall into three main categories (partly reviewed in Vitti *et al.*, 2013), corresponding to the signature they try to target: i) study of variation in allele frequencies and polymorphism, ii) study of variation in linkage disequilibrium and iii) reconstruction of allele genealogies using the coalescent. Most of these methods assume that markers are ordered along a genome; although they can also be used to extract individual markers under selection that can be then be aligned (except for most LD-based methods).

It can happen that researchers report the results obtained from only a few methods when studying selection (François *et al.*, 2015). However, many methods (even popular ones, such as Bayescan) can suffer from high positive rates under some demographic scenarios (Lotterhos and Whitlock, 2014). Combining methods can therefore help prevent this.

Methods focusing on polymorphism

While demographic forces such as drift and migration will affect the whole genome, local effects of selection should produce discrepancies with genome-wide polymorphism (Lewontin and Krakauer, 1973). Selection affects allele frequencies and polymorphism in predictable ways at the scale of single populations. Several statistics summarize them, like π, the nucleotide diversity (Nei and Li, 1979), Tajima's D (Tajima, 1989) or Fay and Wu's H (Fay and Wu, 2000). They are sensitive to population demographic history, and allow characterizing of summary statistics (e.g., in ABC analyses). They have nonetheless the potential to highlight genomic regions displaying clear signatures of selection, or to confirm selection at candidate genes. For example, balancing selection should lead to an excess of common polymorphisms, similar to a recent bottleneck, leading to high Tajima's D and π values. Purifying selection leads to the opposite pattern, similar to a recent population expansion, with an excess of rare variants and low diversity. Combination of these statistics allow to identify more precisely targets of selection, and has been used to develop composite tests, like the composite likelihood ratio (CLR) test (Nielsen *et al.*, 2005).

When an allele is under positive selection in a population, its frequency tends to rise until fixation, unless gene flow from other populations or strong drift prevents it. 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. First attempts to take into account population structure and variation in gene flow included FDIST2 (Beaumont and Nichols, 1996), which modeled populations as islands and aimed at detecting loci under selection by contrasting heterozygosity to F_{st} between populations. More sophisticated methods are now available, dedicated to the detection of outliers in large genomic datasets. Most of them correct for relatedness across samples, and are reviewed extensively by Francois *et al.*(2015). Some methods, like LFMM (Frichot *et al.*, 2013), aim at detecting variants correlated to environmental factors.

Other methods perform a "naïve scan" for outliers on the basis of differentiation, like BAYESCAN (Foll and Gaggiotti, 2008) which considers all populations to drift at different rates from a single ancestral pool. Most recent methods, like BAYPASS (Gautier, 2015), model demographic history by computing a kinship matrix between populations. Those methods are particularly well suited for the study of RAD-sequencing data, for which allele frequencies are often the only information available in the absence of any reference genome.

Detecting 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). Characterizing population history is required to draw conclusions about the possible involvement of a genomic region in adaptation to environment. 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 has also to be carefully picked to control for spatial autocorrelation of genotypes due to isolation by distance and shared demographic history.

The methods described above focus on allele frequencies at the population scale, but do not allow characterizing association with a trait varying between individuals within populations (e.g. resistance to a pathogen, symbiotic association, individual size or flowering time). For this task, methods performing Genome-wide association analysis (GWAS) are better suited,

although the recent development of multivariate methods such as PCAdapt (Duforet-Frebourg *et al.*, 2016) also allows to identify outlier loci in admixed or continuous populations. Methods such as GenAbel in R (Aulchenko *et al.*, 2007) or PLINK (Purcell *et al.*, 2007) are powerful tools. Taking into account relatedness between samples and population history is required to correct for false positives. This is especially recommended for species that undergo episodes of selfing or strong bottlenecks, for which sampling unrelated individuals may be unfeasible.

It is important to keep in mind that uncovering the genetic bases of complex, polygenic traits remains challenging, even in model species (Pritchard and Di Rienzo, 2010; Rockman, 2012). It may be unavoidable in a first step to focus only on traits that are under a relatively simple genetic determinism. This can however lead to an 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 easier to target does not imply that they are the main substrate of selection (Rockman, 2012). Association methods may help targeting variants undergoing soft sweeps, weak selection or involved in polygenic control of traits (Pritchard *et al.*, 2010), for which signatures of selection are subtle and sometimes difficult to retrieve from allele frequency data.

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 understanding the origin and maintenance of reproductive isolation.To address the adapative contribution of introgressed segments, one may want to first identify these segments, estimate the relative contribution of each parental population (chromosome painting), then assess whether they display signatures of selection (Racimo *et al.*, 2015). Advantageous alleles can migrate from one population to another, or resist introgression from other populations, and the relative importance of these islands resisting gene flow after secondary contact has been recently discussed (Cruickshank and Hahn, 2014). Many studies having focused on hybrid zones and introgression provide inspiring examples (Hedrick, 2013), as demonstrated by recent work focusing on patterns of heterogenous gene flow in *Mytilus* mussels (Roux *et al.*, 2014), localized introgression and inversions at a color locus in *Heliconius* butterflies (The Heliconius Genome Consortium *et al.*, 2012) or adaptive introgression of anticoagulant resistance alleles in mice (Song *et al.*, 2011).

Methods aimed at characterizing heterogeneity in introgression rates are useful to detect adaptive introgression and refine demographic history. A common test for introgression,

available in PopGenome, is the ABBA-BABA test, summarized by Patterson's D (Durand *et al.*, 2011). Another possibility lies in the comparison of absolute and relative measures of divergence (Cruickshank and Hahn, 2014), such as d_{xy} and F_{st}, which can be calculated in PopGenome. Phylogenetic methods able to contrast gene trees to species trees, such as BEAST* can be used to infer whether a substantial proportion of loci display inconsistent information. A recent ABC framework has also been proposed to characterize genome-wide heterogeneity in migration rates (Roux *et al.*, 2014).

Absolute measures of divergence are correlated to the time since coalescence. In the case of local introgression, both statistics should be reduced. For balancing selection, the decline in F_{st} is due to an excess of shared ancestral alleles, which should not impact d_{xy}, or should even make it higher than genomic background. However, these methods do not prevent false positives and results should be interpreted with caution (Martin *et al.*, 2015).

Detecting selection with methods focusing on LD

LD is increased and diversity is decreased near a selected allele, especially after recent selection. A class of methods aims 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 XP-EHH test (McCarroll *et al.*, 2007) or Rsb (Tang *et al.*, 2007). Individuals included in the analysis should be as distantly related as possible to improve precision and avoid an excess of false positives. These methods require data to be phased in order to reconstruct haplotypes. This procedure can be performed with fastPhase (Scheet and Stephens, 2006), BEAGLE (Browning and Browning, 2011) or SHAPEIT2 (O'Connell *et al.*, 2014). The R package rehh (Gautier and Vitalis, 2012) allows calculating these statistics, as well as Sweep (http://www.broadinstitute.org/mpg/sweep/index.html) or selscan (Szpiech and Hernandez, 2014). Statistics dedicated to the detection of soft sweeps and selection on standing variation are also available, like the nSL statistics (Ferrer-Admetlla *et al.*, 2014) in selscan or the H2/H1 statistics (Garud *et al.*, 2015), although further studies are still needed to understand to what extent hard and soft sweeps can actually be distinguished (Schrider *et al.*, 2015). When the relative order of markers is not known, as it can be the case in RAD-seq studies without a reference genome, LDna (Kemppainen *et al.*, 2015) can be used to target sets of

markers displaying strong linkage disequilibrium. This approach can be useful not only to detect selection but also structural variation such as large inversions.

Even hard selective sweeps can be challenging to detect with LD-based statistics (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).
These approaches are more powerful with a relatively high density of markers, such as the ones obtained from whole-genome sequencing or high-density RAD-seq, and benefits from using statistics focusing on polymorphism and allele sharing. In a recent study of local adaptation in sticklebacks (Roesti *et al.*, 2015), these methods have been used on dense RAD-sequencing data to characterize the extent of selection at markers displaying high differentiation (F_{ST}), allowing to pinpoint new candidates and confirming previous ones (such as *Ectodysplasin* gene). In addition, the identification of large regions displaying high divergence and LD revealed the importance of large-scale structural variation in shaping genome structure.

Detecting and characterizing selection with the coalescent

When a candidate locus has been identified, it is possible to use coalescent simulations to evaluate the strength of selection and estimate the age of alleles. A software such as msms (Ewing and Hermisson, 2010), which is also available in PopGenome, can then be used. This requires that the neutral history of population be known to properly control for, e.g., population structure and gene flow.

An advantage of full coalescent methods is that they provide a relatively complete picture of individual loci history, by modeling coalescence, recombination and considering variation in mutation rate. They are however computationally intensive, and thus difficult to apply to whole genomes. However, recent computational improvements make this procedure feasible, as illustrated by ARGWeaver (Rasmussen *et al.*, 2014), which allowed recovering known candidate genes for balancing selection in human data. This method uses ancestral recombination graphs to model the genealogy of each non-recombining block in the genome. It permits extracting genealogies for these blocks and provides estimates for local recombination rate, coalescence time and local effective population size for each block. This

approach is promising to characterize positive, purifying or balancing selection while taking into account variation in recombination and mutation rate. However, the high stochasticity in parameter estimation can limit resolution when targeting single genes. Other methods use the theoretical framework of the coalescent to target sites under positive selection. A recent method (SCCT) using conditional coalescent trees (Wang *et al.*, 2014) claims to be faster and more precise in targeting selective sweeps. BALLET (DeGiorgio *et al.*, 2014) is a promising method to characterize ancient balancing selection. Most of those methods are designed for medium-to-high depth whole-genome resequencing, and require that individual genotypes be phased and well characterized.

Identifying variants of functional interest

Characterizing the amount of synonymous or non-synonymous mutations is another way to detect whether a specific gene undergoes purifying or positive selection. An excess of non-synonymous mutations can signal positive or balancing selection, or a relaxation of selective constraints on a given gene. This requires that an annotated genome is available. Annotation of mutations can be done with SNPdat (Doran and Creevey, 2013), or directly in PopGenome, which can also perform at the genome scale tests of selection such as the MK test (McDonald and Kreitman, 1991). Another popular test of selection is the comparison of non-synonymous and synonymous mutations between orthologs from different species and can be performed in packages such as PAML (Yang, 2007). To recover information about the putative function of a gene or a genomic region, it may be useful to perform a genome ontology (GO) enrichment analysis, using tools such as BLAST2GO (Conesa *et al.*, 2005). When interpreting the link between selection and genetic variation, a careful review of literature can fruitfully complete the conclusions made using GO enrichment analyses.

While suggestive, genome scans for selection and association in natural populations cannot be considered as conclusive evidence for the function of a given gene, and needs to be combined with functional evidence (Vitti *et al.*, 2013), which might be provided by variation in the expression of a candidate highlighted by RNA-sequencing data (but see Box 1), but more generaly implies that developmental studies be performed, a step that is not always possible for non-model organisms. Pinpointing the exact genetic mutation leading to a change in phenotype is challenging even when combining several tests for selection, and requires whole-genome sequencing data to obtain a near-exhaustive list of mutations. It has been proposed to combine QTL analyses with population genomics to facilitate identification of candidate loci (Stinchcombe and Hoekstra, 2008). Basically, controlled crosses allow for

identifying genomic regions associated with a selected phenotype, while the study of variation in natural populations facilitates the fine-mapping of variants actually selected in natural populations. However, this requires that the species of interest can be raised in laboratory, which is unpractical for many research teams. An alternative is the study of candidate genes, for which an extensive description of functional variation is available. For example, in a recent study on bananaquits, GBS data have been used to obtain a neutral distribution to which patterns of substitution and differentiation at candidate genes for color variation were compared (Uy *et al.*, 2016). In another study on color polymorphism in *Peromyscus* mice, a combination of field experiment, targeted sequencing of candidate genes and neutral regions, genome-scan for selection and association was able to show how selection at many mutations at the same locus drive adaptive phenotypic divergence (Linnen *et al.*, 2013).

The combination of tests aiming at different signatures of selection can allow reducing the size of candidate regions. For example, combining results from environmental association mapping and genomic scans for selection allows the identification of candidate genes for which a function can be proposed (François *et al.*, 2015). Another common approach relies on the combination of rests targeting different signatures of selection, typically those using the allele frequency spectrum and those using haplotype length. A test of this sort has been proposed in human genetics (Grossman *et al.*, 2013), called the composite of multiple signals (CMS) test. Specifically, CMS integrates F_{ST} with iHS and XP-EHH and other statistics describing the AFS. Nevertheless, signatures of selection can be elusive, and obtaining an exhaustive list of genes under positive selection is unlikely. Further advances will require that methods targeting selection be able to better take into account epistatic interaction and weak selection.

Population history

The coalescent has first emerged to provide population geneticists a way of modeling alleles genealogy from a sample taken from a large population. Going backward in time, alleles merge (coalesce) in a stochastic way until reaching their most recent common ancestor (Kingman, 1982). The most well-known coalescent-based tools dedicated to population genetics include IMa (Hey and Nielsen, 2007), Migrate-n (Beerli and Palczewski, 2010) or Lamarc (Kuhner, 2009) (Table 3). Obtaining demographic estimates (e.g. time in years) for parameters usually requires that mutation rate and generation time be known or at least

reasonably well estimated, for example from other close species with similar life history. Since selection impacts allele frequencies, it is common that loci that are candidate for selection be removed prior any analysis for methods using the AFS.

Recent methods have been developed to handle whole genome datasets that allow inferring variation in population sizes with time without *a priori*, such as those based on Pairwise Sequentially Markovian Coalescent (PSMC), that require only a single diploid genome (Li and Durbin, 2011). One general drawback of these methods is that they are limited to rather simple scenarios, not handling more than two populations yet (but see diCal2, Table 2). While powerful, PSMC is sensitive to confounding factors such as population structure (Orozco-terWengel, 2016) that leads to false signatures of expansion or bottleneck. It also does not allow studying recent demographic events since coalescence events for only two alleles from a single individual in the recent past are infrequent. However, extensions of the model allowing for several genomes have been developed for precise population history in the recent past, like MSMC (Schiffels and Durbin, 2014) or diCal (Sheehan *et al.*, 2013). Recently an ABC framework, implemented in PopSizeABC, has been proposed to infer demographic variation from single genomes (Boistard *et al.*, 2016). A recent 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 meta-population (Mazet *et al.*, 2015). Such developments should help increase the amount of information retrieved from only a few genomes. However, it is essential to keep in mind that natural populations are structured and connected in complex ways, which can bias demographic inferences, even for popular markers such as mitochondrial sequences (Heller *et al.*, 2013).

A computationally faster approach is to use Approximate Bayesian Computation (ABC) methods, which compare the empirical data with a set of simulated data produced by coalescent simulations under predefined scenarios. By measuring the distance between carefully chosen summary statistics describing each simulation with those from the observed dataset, it is possible to infer which scenario explains the data the best. More information on how to perform ABC analysis is described by Csilléry *et al.*(2010). The main advantage of ABC is that it allows handling any type of markers and arbitrarily complex models, contrarily to methods like IMa where the model is predefined. However, using summary statistics leads to the loss of potentially useful information.

Methods based on IBS (Harris and Nielsen, 2013) and IBD tracts (Palamara and Pe'er, 2013) constitute an interesting alternative for model testing when high density RAD-seq data or

whole genome datasets are available in large number (more than 100 individuals were required to infer recent demographic events with DoRIS). The large sample sizes required to infer recent events make these methods mostly relevant to researchers working on near-model species for which a substantial amount of data is available.

More recently, new methods such as ∂a∂i based on the allele frequency spectrum (AFS) emerged to facilitate and speed up the analysis of large SNP datasets. Different patterns of gene flow and demographic events all shape the AFS in specific ways (e.g. more alleles are likely to be found at similar frequencies in two recently diverged or highly connected populations). These methods generally assume that SNPs are under linkage equilibrium. Including SNPs in strong LD should not particularly bias model comparison, but can be an issue when estimating parameters (see fastsimcoal manual for more details). Note that the AFS can also be used as a set of summary statistics for ABC inference. Using allele frequencies estimated from pooled datasets 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).

One drawback when using SNP data without considering monomorphic sites is that the mutation rate per generation is not directly taken into account. For example, in DIYABC, it does not matter when a mutation appears in the simulated genealogy, as long as it happens only once before coalescence, a reasonable assumption for SNP markers. However, this prevents any conversion of parameters into demographic estimates by using mutation rate. Again, it is also possible to extract the complete DNA sequence for a set of randomly selected markers and perform analyses on this dataset including monomorphic sites. Another possibility consists of a calibration of parameter estimates by including in the analysis a fixed parameter, such as population size or divergence time. This approach is also feasible when estimating parameters from the allele frequency spectrum, like in ∂a∂i or fastsimcoal2. Reaching a high level of precision in demographic parameters estimation can be challenging when information is lacking about the evolutionary history of the species considered. At larger time-scales, the lack of fossil record can make difficult the calibration of molecular clocks. Thus, for some species, only qualitative interpretation will be possible.

There is currently a balance between methods allowing for arbitrarily complex models that are defined a priori by the user (e.g. ABC), and methods that allow to more "naively" track evolution in population size or migration (e.g. MSMC). While the first one allows to better model the actual complexity of most living systems, the second one is less prone to be

impacted by the user's bias. Using both methods can therefore help retrieving robustly the evolutionary history of a given model. For example, a recent study on maize demographic and selective history used both ∂a∂i and MSMC to characterize bottleneck and expansion associated to domestication (Beissinger *et al.*, 2016). Note that for this study, all scripts and methods have been made available online, enhancing its reproducibility.

Suggestions and perspectives

Estimating selection and demography jointly along a heterogeneous genome

As stated by Lewontin and Krakauer in 1973, "while natural selection will operate differently for each locus and each allele at a locus, the effect of breeding structure is uniform over all loci and all alleles"(Lewontin and Krakauer, 1973). Since then, traditional studies on selection have mostly considered that demographic processes act on all loci in the same way across a genome, and that positive selection is mostly rare. The traditional approach has thus tended to disconnect the study of selection from the study of demography (Li *et al.*, 2012).

However, this assumption may be incorrect, and a good understanding of demography is crucial to understand how efficient selection can be. On the other hand, removing loci under selection is needed to retrieve the actual demographic story of a set of populations. For example, the large effective population sizes of *Drosophila* have been hypothesized to facilitate a widespread effect of selection across the genome (Sattath *et al.*, 2011; discussion in Li *et al.*, 2012), making both demographic inference and detection of outliers difficult. Other counfounding factors include variation in recombination and mutation rates and background selection (Ewing and Jensen, 2016). Even for model species, these counfounding effects have been precisely characterized only recently and obtaining precise recombination and mutation maps is challenging for non-model species. It has been shown in the last few years that variation in introgression rates along the genome and coupling between loci involved in reproductive isolation and those involved in local adaptation can bias inference about selection and demography (Bierne *et al.*, 2011; Roux *et al.*, 2014). Locally low recombination rate can lead to reduced polymorphism, and be mistaken with a signature of purifying selection.

These issues can only be addressed by going beyond categorization between methods assigned to either the study of selection or demography, and use the results obtained by one

method to inform the other. The availability of a reference genome facilitating the positioning of markers is helpful in this regard. In their study of the two different lineages of the european Sea Bass, and using a RAD-sequencing approach, Tine et al. took into account variation in recombination rate along the genome to interpret signatures of reduced polymorphism as being the result of selection or low recombination (Tine *et al.*, 2014). Since differentiation along the genome seemed to reveal islands resisting gene flow, they could fit a model taking into account variation in introgression rates, providing a better fit to the data and highlighting that islands of high divergence were more likely to be due to locally reduced gene flow after secondary contact. This example illustrates how a combination of descriptive statistics and coalescent analyses can be used to retrieve information from genomic data about both selection and demography.

Most methods do not actually estimate demography and selection jointly, but rather rely on a process where neutral expectations are first drawn from a set of neutral SNPs (e.g. intergenic SNPs), followed by a step where the likelihood for a marker to be under selection is evaluated. Methods such as BAYPASS or PCAdapt are convenient to describe both population structure and give first insights about the proportion of loci that do not follow neutral expectations. When this proportion is not too high (which would suggest recent introgression or an excess of markers displaying high LD due to, e.g., large inversions), outliers can be removed and remaining loci be used to compare neutral models and estimate demographic parameters (e.g. using an ABC framework). 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 such as FLK or BAYPASS (Lotterhos and Whitlock, 2014).

Some recent methods are especially relevant to study both demography and selection at once, while taking into account variation in recombination and mutation rates. For whole-genome data, methods reconstructing ancestral recombination graphs (such as ARGWeaver) have a high potential, since they allow retrieving genealogies along the genome and inform about the timing of coalescence events and therefore selection or migration. A recent application of the method in human paleogenomics allowed to quantitatively characterize introgression between modern humans, Neandertal and Denisovans using only a few whole genomes (Kuhlwilm *et al.*, 2016). This method has however a high computing and sequencing cost, and is therefore not suited for the study of many individuals.

Caution must prevail when attempting to apply sophisticated methods to disentangle selection and demography. In a recent review, Cruickshank et al. suggested that IMa2, which is commonly used to estimate migration rates, did not allow to reliably distinguish between loci under selection and those resisting gene flow (Cruickshank and Hahn, 2014). In the specific case they highlight (*Oryctolagus cuniculus* rabbits, (Sousa *et al.*, 2013)), a descriptive statistics that should have captured introgression signatures (d_{xy}) did not reveal any evidence for differential gene flow between loci categorized by IMa2. This controversy illustrates that basic description of the data is needed in combination with more sophisticated methods, that sometimes make assumptions that can be violated, such as neutrality or no recombination within loci.

In general, and for all types of datasets, description of the data is essential to assess the proportion of loci displaying consistent pattern and characterize the genomic landscape of a species. One may for example plot the distribution of F_{ST} between populations, mean linkage disequilibrium, nucleotide diversity or p-values of association with a trait. Such an approach has been used in *Ficedula* flycatchers, allowing to clearly highlight genomic islands of divergence and the higher differentiation on sexual chromosomes due to ongoing reproductive isolation (Ellegren *et al.*, 2012).

To sum up, the field of population genomics is now moving towards a better integration of selection into a historical framework, while 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, as it should provide perspective about how proximate mechanisms and gene flow can promote or impair local adaptation to new habitats. The growing availability of genome-wide data for non-model 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 should broaden quickly our perspective on complex evolutionary processes, from biogeography (Lexer *et al.*, 2013) to genetic bases of traits (Hohenlohe, 2014) or the maintenance of polymorphism (Hedrick, 2006). While genome heterogeneity in migration, mutation or recombination rates do not necessarily make impossible any conclusion about evolutionary dynamics, they have the potential to blur inferences. The study of DNA sequence variation, while already challenging by itself, needs therefore to be combined with other disciplines such as ecology and functional analyses to be informative (Habel *et al.*, 2015), in accordance with the Dobzhanskyan dictum

stating that biological sense can only be derived from evolutionary context. This can be done 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 modeling, common garden experiments or the study of macro-evolutionary patterns of selection and diversification.

Data sharing, consistency and robustness

Most literature in population genomics has focused so far on single species at once, or on sets of closely related species and subspecies. However, many questions require a more global approach to provide general insights about key processes such as speciation or genetic bases of convergent evolution. While this approach becomes more feasible given the increase in the amount of NGS data (e.g. (Romiguier *et al.*, 2014; Roux *et al.*, 2016)), it requires i) that datasets are made available by researchers, ii) that methods used for analyzing data can be reproduced through unified pipelines. There is a need for a more collaborative and open culture in biology, allowing the free access to data and favoring good practices to allow repeatability of analyses (Nekrutenko and Taylor, 2012), although this cultural shift remains challenging (e.g. Mills *et al.*, 2015; Whitlock *et al.*, 2015).

However, current challenges are not limited to data sharing, but also include dealing with the inflation of bioinformatics tools that sometimes overlap. Overall, our present survey of methods reveals a lack of unified pipeline dealing at once with main aspects of population genomics. Instead of working independently, researchers designing those tools could collaborate to propose free, robust and unified pipelines (Prins *et al.*, 2015). Such initiatives, like Galaxy (Goecks *et al.*, 2010) or Bioconductor (Huber *et al.*, 2015) are nonetheless emerging, and propose clear tutorials facilitating their use. ANGSD (Korneliussen *et al.*, 2014) already provides useful utilities allowing to perform both extensive pre- and post-processing of data (Table 1, Table 3). R has been a popular software among biologists, and now proposes a set of packages able to handle genome-wide datasets and compatible with the VCF files produced by most SNP callers (Paradis *et al.*, 2016). For example, one may consider using only R to perform clustering analyses, PCA and identify outliers (using sNMF, SNPRelate, PCAdapt, and Bioconductor packages), study population structure (adegenet package, geneland), compute summary statistics, test for selection (PopGenome, rehh), study

association with environment (LFMM in LEA package) and performing coalescent simulations and ABC inference for studying demography (coala and abc package). Now that population genomics begins to reach maturity, it is to be hoped that more integrated pipelines will allow minimizing the time spent looking for appropriate software and focus on biological questions.

Beyond SNPs: including structural variation/transposable elements/epigenetics

Most studies of selection and demography have so far focused on SNP, since they are relatively easy to detect with current technlogy and their mutation mechanism produces mostly biallelic alleles, making them easier to use for statistical tests. However, many other heritable genetic alterations can affect genomes, including transposable elements insertions, epigenetics modifications such as methylation, duplications, inversions, deletions or translocations. One of the main issue with this type of variation is that their diversity and their impact on the genome can make them difficult to detect in a systematic way (Iskow *et al.*, 2012), especially for species having only a draft genome. It is however possible to use variation at these markers to study selection, for example by using differentiation statistics, association to environment or extension of haplotypes. Combining information about variant position and SNP variation in flanking regions is also a powerful way to detect variants under selectionas highlighted by a recent study of transposable elements insertions in *Drosophila* (Kofler *et al.*, 2012). Recent work also shows that classical summary statistics such as Tajima's D can be adapted to non-SNP datasets, such as methylations (Wang and Fan, 2014).

Sets of neutral SNPs can be used to control for demography and relatedness between samples when inferring selection. For example, this type of appproach has recently began to be exploredfor studying selection on methylation patterns. In a recent Molecular Ecology issue (Verhoeven *et al.*, 2016), a study using bisulfite precipitation was able to replace strongly associated methylated variants near to genes known to be involved in response to environment, while another could show a stronger pattern of Isolation by Distance for methylation-sensitive AFLPs than for regular AFLPs and microsatellites, suggesting a stronger impact of environment on methylation patterns than expected under neutrality (Herrera *et al.*, 2016).

Another potential issue with this type of variation is that there is currently a lack of tools able to simulate their mutation model, complicating any comparison driven from neutral models

built from SNPs. This is the case for transposable elements, for which the assumption of mutation/drift equilibrium is challenging, making comparisons of their allele frequency spectrum with neutral SNPs potentially difficult. For example, a recent burst of transposition can lead to an excess of low frequency elements and recent insertions compared to the expectation under equilibrium, even if transposable elements (TEs) are not under purifying selection (Bergman and Bensasson, 2007; Blumenstiel *et al.*, 2014). More generally, neutral models would benefit from new ways to model the appearance of genomic variation through time for non-SNP data, allowing to provide even more conservative assessments of either negative or positive selection.

Acknowledgements

The University of Basel and New York University Abu Dhabi have supported this research. I want to thank two anonymous reviewers, Stephane Boissinot, Joris Bertrand, Anne Roulin and Ben Warren for their insightful comments on previous versions of the manuscript.

References

Abi-Rached L, Jobin M, Kulkarni S, McWhinnie A, Dalva K, Gragert L, *et al.* (2011). The shaping of modern human immune systems by multiregional admixture with archaic humans. *Science* **334**: 89–95.

Abzhanov A, Extavour CG, Groover A, Hodges SA, Hoekstra HE, Kramer EM, *et al.* (2008). Are we there yet? Tracking the development of new model systems. *Trends Genet* **24**: 353–60.

Alexander DH, Novembre J (2009). Fast Model-Based Estimation of Ancestry in Unrelated Individuals. *Genome Res*: 1655–1664.

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KMM, *et al.* (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**: 481–491.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol Ecol* **22**: 3179–90.

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007). GenABEL: An R library for genome-wide association analysis. *Bioinformatics* **23**: 1294–1296.

Axelsson E, Ratnakumar A, Arendt M-L, Maqbool K, Webster MT, Perloski M, *et al.* (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**: 360–4.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, *et al.* (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**: e3376.

Baran Y, Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Eng C, *et al.* (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **28**: 1359–1367.

Beaumont MA, Balding DJ (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* **13**: 969–980.

Beaumont MA, Nichols RA (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc London Biol Sci*: 1619–1626.

Beerli P, Palczewski M (2010). Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics* **185**: 313–26.

Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J (2016). Recent demography drives changes in linked selection across the maize genome. *Nat Plants* **2**: 16084.

Bergman CM, Bensasson D (2007). Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **104**: 11340–11345.

Besnard G, Bertrand JAM, Delahaie B, Bourgeois YXC, Lhuillier E, Thébaud C (2016). Valuing museum specimens: high-throughput DNA sequencing on historical collections of New Guinea crowned pigeons (Goura). *Biol J Linn Soc* **117**: 71–82.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011). The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Mol Ecol* **20**: 2044–72.

Blumenstiel JP, Chen X, He M, Bergman CM (2014). An age-of-allele test of neutrality for transposable element insertions. *Genetics* **196**: 523–538.

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 Genet*: 858–865.

Bolnick DI, Otto SP (2013). The magnitude of local adaptation under genotype-dependent dispersal. *Ecol Evol* **3**: 4722–4735.

Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah JM, Blott S, *et al.* (2010). Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended. *Genetics*: 241–262.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, *et al.* (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput Biol* **10**: 1–6.

Bradburd GS, Ralph PL, Coop GM (2013). Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution (N Y)* **67**: 3258–3273.

Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, *et al.* (2012). PCAdmix: Principal Components-Based Assignment of Ancestry along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Hum Biol* **84**: 343–364.

Browning BL, Browning SR (2011). A fast, powerful method for detecting identity by descent. *Am J Hum Genet* **88**: 173–182.

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, Roychoudhury A (2012). Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Mol Biol Evol* **29**: 1917–1932.

Buerkle CA, Gompert Z (2013). Population genomics based on low coverage sequencing: how low should we go? *Mol Ecol* **22**: 3028–35.

Cadzow M, Boocock J, Nguyen HT, Wilcox P, Merriman TR, Black MA (2014). A bioinformatics workflow for detecting signatures of selection in genomic data. *Front Genet* **5**: 1–8.

Cariou M, Duret L, Charlat S (2013). Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecol Evol* **3**: 846–852.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011). Stacks: building

and genotyping Loci de novo from short-read sequences. *G3 (Bethesda)* **1**: 171–82.

Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, *et al.* (2015). Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biol* **16**: 30.

Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA (2010). The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* **186**: 983–995.

Christe C, Stolting KN, Paris M, Frayisse 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. *Mol Ecol*.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**: 3674–3676.

Cornuet J-M, Santos F, Beaumont MA, Robert CP, Marin J-M, Balding DJ, *et al.* (2008). Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* **24**: 2713–9.

Cruickshank TE, Hahn MW (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol* **23**: 3133–3157.

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010). Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol* **25**: 410–8.

Csilléry K, François O, Blum MGB (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol Evol* **3**: 475–479.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* (2011). The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* **12**: 499–510.

Degiorgio M, Huber CD, Hubisz MJ, Hellmann I, Nielsen R (2016). Genetics and population analysis SWEEPFINDER 2 : Increased sensitivity , robustness , and flexibility. *Bioinformatics*.

DeGiorgio M, Lohmueller KE, Nielsen R (2014). A model-based approach for identifying signatures of ancient balancing selection in genetic data. *PLoS Genet* **10**: e1004561.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al.* (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Doran AG, Creevey CJ (2013). Snpdat: easy and rapid annotation of results from de novo snp discovery projects for model and non-model organisms. *BMC Bioinformatics* **14**: 45.

Drummond AJ, Rambaut A (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**: 214.

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB (2016). Detecting genomic signatures of natural selection with principal component analysis: Application to the 1000 genomes data. *Mol Biol Evol* **33**: 1082–1093.

Dufresne F, Stift M, Vergilino R, Mable BK (2014). Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. *Mol Ecol* **23**: 40–69.

Durand EY, Patterson N, Reich D, Slatkin M (2011). Testing for ancient admixture between closely related populations. *Mol Biol Evol* **28**: 2239–2252.

Edelaar P, Bolnick DI (2012). Non-random gene flow: an underappreciated force in evolution and ecology. *Trends Ecol Evol* **27**: 659–65.

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, *et al.* (2012). The genomic landscape of species divergence in Ficedula flycatchers. *Nature* **491**: 756–60.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, *et al.* (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS One* **6**: e19379.

Ewing G, Hermisson J (2010). MSMS: A coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**: 2064–2065.

Ewing GB, Jensen JD (2016). The consequences of not accounting for background selection in demographic inference. *Mol Ecol* **25**: 135–141.

Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013). Robust Demographic

Inference from Genomic and SNP Data. *PLoS Genet* **9**.

Excoffier L, Foll M (2011). Fastsimcoal: a Continuous-Time Coalescent Simulator of Genomic Diversity Under Arbitrarily Complex Evolutionary Scenarios. *Bioinformatics* **27**: 1332–4.

Excoffier L, Lischer HEL (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **10**: 564–7.

Fay JC, Wu CI (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–13.

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014). On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol* **31**: 1275–1291.

Foll M, Gaggiotti O (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**: 977–93.

da Fonseca RR, Albrechtsen A, Themudo GE, Ramos-Madrigal J, Sibbesen JA, Maretty L, *et al.* (2016). Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Mar Genomics* **30**: 1–11.

François O, Martins H, Caye K, Schoville SD (2015). Controlling False Discoveries in Genome Scans for Selection. *Mol Ecol* **55**: in press.

Fraser DJ, Bernatchez L (2001). Adaptive evolutionary conservation: Towards a unified concept for defining conservation units. *Mol Ecol* **10**: 2741–2752.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196**: 973–983.

Frichot E, Schoville SD, Bouchard G, François O (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol* **30**: 1687–1699.

Futschik A, Schlötterer C (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* **186**: 207–18.

Garrigan D (2013). POPBAM: Tools for evolutionary analysis of short read sequence alignments. *Evol Bioinforma* **2013**: 343–353.

Garud NR, Messer PW, Buzbas EO, Petrov DA (2015). Recent Selective Sweeps in North

American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLoS Genet* **11**: 1–32.

Gautier M (2015). Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics* **201**: 1555–1579.

Gautier M, Vitalis R (2012). Rehh An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* **28**: 1176–1177.

Gay L, Crochet P-A, Bell D a, Lenormand T (2008). Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models. *Evolution* **62**: 2789–806.

Gayral P, Melo-Ferreira J, Glémin S, Bierne N, Carneiro M, Nabholz B, *et al.* (2013). Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate-Invertebrate Gap. *PLoS Genet* **9**.

Goecks J, Nekrutenko A, Taylor J (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11**: R86.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644–52.

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet* **43**: 1031–1034.

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, *et al.* (2013). Identifying recent adaptations in large-scale genomic data. *Cell* **152**: 703–713.

Guillot G, Renaud S, Ledevin R, Michaux J, Claude J (2012). A unifying model for the analysis of phenotypic, genetic, and geographic data. *Syst Biol* **61**: 897–911.

Guindon S, Dufayard JF, 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. *Syst Biol* **59**: 307–321.

Günther T, Coop G (2013). Robust identification of local adaptation from allele frequencies. *Genetics* **195**: 205–220.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5**.

Habel J, Zachos F, Dapporto L, Rödder D, Radespiel U, Tellier A, *et al.* (2015). Population genetics revisited – towards a multidisciplinary research field. *Biol J Linn Soc* **115**: 1–12.

Hahn C, Bachmann L, Chevreux B (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - A baiting and iterative mapping approach. *Nucleic Acids Res* **41**.

Harris K, Nielsen R (2013). Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLoS Genet* **9**.

Hedrick PW (2006). Genetic Polymorphism in Heterogeneous Environments: The Age of Genomics. *Annu Rev Ecol Evol Syst* **37**: 67–93.

Hedrick PW (2013). Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol Ecol* **22**: 4606–4618.

Heled J, Drummond AJ (2010). Bayesian Inference of Species Trees from Multilocus Data. *Mol Biol Evol* **27**: 570–580.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, *et al.* (2014). A Genetic Atlas of Human Admixture History. *Science* **343**: 747–751.

Heller R, Chikhi L, Siegismund HR (2013). The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History. *PLoS One* **8**.

Hendry AP (2004). Selection against migrants contributes to the rapid evolution of ecologically dependent reproductive isolation. *Evol Ecol Res* **6**: 1219–1236.

Herrera CM, Medrano M, Bazaga P (2016). Comparative spatial genetics and epigenetics of plant populations: Heuristic value and a proof of concept. *Mol Ecol* **25**: 1653–1664.

Hey J, Nielsen R (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci U S A* **104**: 2785–90.

Hohenlohe PA (2014). Ecological genomics in full colour. *Mol Ecol* **23**: 5129–31.

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, *et al.* (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**: 115–121.

Hudson RR (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337–338.

Huson DH, Bryant D (2006). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**: 254–267.

Iskow RC, Gokcumen O, Lee C (2012). Exploring the role of copy number variants in human adaptation. *Trends Genet* **28**: 245–257.

Jenner RA, Wills MA (2007). The choice of model organisms in evo-devo. *Nat Rev Genet* **8**: 311–319.

Jensen JD (2014). On the unfounded enthusiasm for soft selective sweeps. *Nat Commun* **5**: 5281.

Jombart T, Devillard S, Balloux F, Falush D, Stephens M, Pritchard J, *et al.* (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* **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 (Edinb)* **101**: 92–103.

Jombart T, Pontier D, Dufour A-B (2009). Genetic markers in the playground of multivariate analysis. *Heredity (Edinb)* **102**: 330–41.

Kemppainen P, Knight CG, Sarma DK, Hlaing T, Prakash A, Maung Maung YN, *et al.* (2015). Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Mol Ecol Resour*: 1031–1045.

Kern AD, Schrider DR (2016). Discoal: flexible coalescent simulations with selection. *Bioinformatics* **32**: 3839–3841.

Kim D, Langmead B, Salzberg SL (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357–60.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013). TopHat2: accurate

alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

Kingman JFC (1982). The coalescent. *Stoch Process their Appl* **13**: 235–248.

Kofler R, Betancourt AJ, Schlötterer C (2012). Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. *PLoS Genet* **8**.

Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, *et al.* (2011). PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* **6**: e15925.

Kofler R, Pandey RV, Schlötterer C (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27**: 3435–6.

Kolaczkowski B, Kern AD, Holloway AK, Begun DJ (2011). Genomic differentiation between temperate and tropical Australian populations of Drosophila melanogaster. *Genetics* **187**: 245–60.

Korneliussen TS, Albrechtsen A, Nielsen R (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**: 356.

Kubota S, Iwasaki T, Hanada K, Nagano AJ, Fujiyama A, Toyoda A, *et al.* (2015). A Genome Scan for Genes Underlying Microgeographic-Scale Local Adaptation in a Wild Arabidopsis Species. *PLoS Genet* **11**: 1–26.

Kuhlwilm M, Gronau I, Hubisz MJ, de Filippo C, Prado-Martinez J, Kircher M, *et al.* (2016). Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**: 429–433.

Kuhner MK (2009). Coalescent genealogy samplers: windows into population history. *Trends Ecol Evol* **24**: 86–93.

Leaché AD, Banbury BL, Felsenstein J, De Oca ANM, Stamatakis A (2015). Short tree, long tree, right tree, wrong tree: New acquisition bias corrections for inferring SNP phylogenies. *Syst Biol* **64**: 1032–1047.

Leaché AD, Harris RB, Rannala B, Yang Z (2014). The influence of gene flow on species tree

estimation: A simulation study. *Syst Biol* **63**: 17–30.

Lee T-H, Guo H, Wang X, Kim C, Paterson AH (2014). SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* **15**: 162.

Legrand D, Tenaillon MI, 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**: 1197–206.

Lewontin RC, Krakauer J (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**: 175–195.

Lexer C, Mangili S, Bossolini E, Forest F, Stölting KN, Pearman PB, *et al.* (2013). 'Next generation' biogeography: towards understanding the drivers of species diversification and persistence (M Carine, Ed.). *J Biogeogr* **40**: 1013–1022.

Li H, Durbin R (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493–496.

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? *Mol Ecol* **21**: 28–44.

Linnen CR, Poh Y-P, Peterson BK, Barrett RDH, Larson JG, Jensen JD, *et al.* (2013). Adaptive evolution of multiple traits through multiple mutations at a single gene. *Science* **339**: 1312–1316.

Lotterhos KE, Whitlock MC (2014). Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol Ecol* **23**: 2178–2192.

Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, *et al.* (2013). Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. *PLoS Genet* **9**.

Malé PJG, Bardon L, Besnard G, Coissac E, Delsuc F, Engel J, *et al.* (2014). Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Mol Ecol Resour* **14**: 966–975.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, *et al.* (2010). Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7**: 111–118.

Mandoli DF, Olmstead R (2000). The importance of emerging model systems in plant biology. *J Plant Growth Regul* **19**: 249–252.

Manel S, Holderegger R (2013). Ten years of landscape genetics. *Trends Ecol Evol* **28**: 614–621.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**: 2867–2873.

Maretty L, Sibbesen JA, Krogh A, Heber S, Alekseyev M, Sze S-H, *et al.* (2014). Bayesian transcriptome assembly. *Genome Biol* **15**: 501.

Martin SH, Van Belleghem SM (2016). Exploring evolutionary relationships across the genome using topology weighting. *bioRxiv*: 69112.

Martin SH, Davey JW, Jiggins CD (2015). Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol Biol Evol* **32**: 244–257.

Mazet O, Rodriguez W, Chikhi L (2015). Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theor Popul Biol* **104**: 46–58.

McCarroll SA, Sabeti PC, Frazer KA, Varilly P, Fry B, Ballinger DG, *et al.* (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**: 913–8.

McDonald JH, Kreitman M (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**: 652–4.

McVean G, Awadalla P, Fearnhead P (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**: 1231–1241.

Mills JA, Teplitsky C, Arroyo B, Charmantier A, Becker PH, Birkhead TR, *et al.* (2015). Archiving Primary Data: Solutions for Long-Term Studies. *Trends Ecol Evol* **30**: 581–589.

Myers S (2005). A Fine-Scale Map of Recombination Rates and Hotspots Across the Human

Genome. *Science* **310**: 321–324.

Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, *et al.* (2012). Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philos Trans R Soc Lond B Biol Sci* **367**: 343–53.

Nei M, Li WH (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A* **76**: 5269–73.

Nekrutenko A, Taylor J (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat Rev Genet* **13**: 667–72.

Nicholls JA, Pennington RT, Koenen EJM, Hughes CE, Hearn J, Bunnefeld L, *et al.* (2015). Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus Inga (Leguminosae: Mimosoideae). *Front Plant Sci* **6**: 710.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007). Recent and ongoing selection in the human genome. *Nat Rev Genet* **8**: 857–868.

Nielsen R, Williamson S, Kim Y, Nielsen R, Williamson S, Kim Y, *et al.* (2005). Genomic scans for selective sweeps using SNP data Genomic scans for selective sweeps using SNP data. *Genome Res*: 1566–1575.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, *et al.* (2008). Genes mirror geography within Europe. *Nature* **456**: 98–101.

O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, *et al.* (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genet* **10**.

Orozco-terWengel P (2016). The devil is in the details: the effect of population structure on demographic inference. *Heredity (Edinb)* **116**: 349–350.

Palamara PF, Pe'er I (2013). Inference of historical migration rates via haplotype sharing. *Bioinformatics* **29**: 180–188.

Paradis E, Gosselin T, Goudet J, Jombart T, Schliep K (2016). Linking genomics and population genetics with R. *Mol Ecol Resour*: 54–66.

Pavlidis P, Laurent S, Stephan W (2010). MsABC: A modification of Hudson's ms to facilitate multi-locus ABC analysis. *Mol Ecol Resour* **10**: 723–727.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290–5.

Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ (2014). PopGenome: An efficient swiss army knife for population genomic analyses in R. *Mol Biol Evol* **31**: 1929–1936.

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

Piskol R, Ramaswami G, Li JB (2013). Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet* **93**: 641–651.

Poelstra JW, Vijay N, Bossu CM, Lantz H, Ryll B, Baglione V, *et al.* (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* **344**: 1410–1414.

Price A, Patterson NJ, Plenge RM, Weinblatt ME, Shadick N a, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–9.

Prins P, de Ligt J, Tarasov A, Jansen RC, Cuppen E, Bourne PE (2015). Toward effective software solutions for big biology. *Nat Biotech* **33**: 686–687.

Pritchard JK, Pickrell JK, Coop G (2010). The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. *Curr Biol* **20**: R208–R215.

Pritchard JK, Di Rienzo A (2010). Adaptation – not by sweeps alone. *Nat Rev Genet* **11**: 665–667.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, *et al.* (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**: 559–575.

Puritz JB, Matz M V., Toonen RJ, Weber JN, Bolnick DI, Bird CE (2014). Demystifying the RAD fad. *Mol Ecol* **23**: 5937–5942.

Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E (2015). Evidence for archaic

adaptive introgression in humans. *Nat Rev Genet* **16**: 359–371.

Raj A, Stephens M, Pritchard JK (2014). FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**: 573–589.

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A (2014). Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genet* **10**.

Rockman M V (2012). The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution (N Y)* **66**: 1–17.

Roesti M, Kueng B, Moser D, Berner D (2015). The genomics of ecological vicariance in threespine stickleback fish. *Nat Commun* **6**: 8767.

Romiguier J, Gayral P, Ballenghien M, Bernard a., Cahais V, Chenuil a., *et al.* (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**: 261–3.

Roux C, Fraisse C, Castric V, Vekemans X, Pogson GH, Bierne N (2014). Can we continue to neglect genomic variation in introgression rates when inferring the history of speciation? A case study in a Mytilus hybrid zone. *J Evol Biol* **27**: 1662–1675.

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 Biol*.

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. *Mol Biol Evol* **30**: 435–47.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, *et al.* (2002). Detecting recent positive selection in the human genome from haplotype structure. **419**.

Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G (2011). Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in drosophila simulans. *PLoS Genet* **7**.

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. *Am J Hum Genet* **78**: 629–44.

Schiffels S, Durbin R (2014). Inferring human population size and separation history from

multiple genome sequences. *Nat Genet* **46**: 919–25.

Schrider DR, Mendes FK, Hahn MW, Kern AD (2015). Soft shoulders ahead: Spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics* **200**: 267–284.

Schubert M, Jónsson H, Chang D, Der Sarkissian C, Ermini L, Ginolhac A, *et al.* (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proc Natl Acad Sci* **111**: 201416991.

Schulz MH, Zerbino DR, Vingron M, Birney E (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**: 1086–1092.

Shafer AB a., Wolf JBW, Alves PC, Bergström L, Bruford MW, Brännström I, *et al.* (2015). Genomics and the challenging translation into conservation practice. *Trends Ecol Evol* **30**: 78–87.

Sheehan S, Harris K, Song YS (2013). Estimating Variable Effective Population Sizes from Multiple Genomes : A Sequentially Markov Conditional Sampling Distribution Approach. *Genetics* **194**: 647–662.

Song Y, Endepols S, Klemann N, Richter D, Matuschka FR, Shih CH, *et al.* (2011). Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Curr Biol* **21**: 1296–1301.

Sousa VC, Carneiro M, Ferrand N, Hey J (2013). Identifying loci under selection against gene flow in isolation-with-migration models. *Genetics* **194**: 211–233.

Staab PR, Metzler D (2016). Coala: An R framework for coalescent simulation. *Bioinformatics* **32**: 1903–1904.

Staab PR, Zhu S, Metzler D, Lunter G (2015). Scrm: Efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics* **31**: 1680–1682.

Stamatakis A (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Stinchcombe JR, Hoekstra HE (2008). Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity (Edinb)* **100**: 158–170.

Szpiech ZA, Hernandez RD (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol* **31**: 2824–2827.

Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–95.

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. *Mol Biol Evol* **27**: 747–752.

Tang K, Thornton KR, Stoneking M (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol* **5**: 1587–1602.

The Heliconius Genome Consortium, Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, *et al.* (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**: 94–98.

Tine M, Kuhl H, Gagnaire P-A, Louro B, Desmarais E, Martins RST, *et al.* (2014). European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nat Commun* **5**: 5770.

Todd E V., Black MA, Gemmell NJ (2016). The power and promise of RNA-seq in ecology and evolution. *Mol Ecol* **25**: 1224–1241.

Uy JAC, Cooper EA, Cutie S, Concannon MR, Poelstra JW, Moyle RG, *et al.* (2016). Mutations in different pigmentation genes are associated with parallel melanism in island flycatchers. *Proc R Soc B* **283**: 2115–2118.

Verhoeven KJF, VonHoldt BM, Sork VL (2016). Epigenetics in ecology and evolution: What we know and what we need to know. *Mol Ecol* **25**: 1631–1638.

Vitti JJ, Grossman SR, Sabeti PC (2013). Detecting Natural Selection in Genomic Data. *Annu Rev Genet* **47**: 97–120.

Wang J, Fan C (2014). A neutrality test for detecting selection on DNA methylation using single methylation polymorphism frequency spectrum. *Genome Biol Evol* **7**: 154–171.

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. *Mol Biol Evol* **31**: 3068–3080.

Weber JN, Peterson BK, Hoekstra HE (2013). Discrete genetic modules are responsible for complex burrow evolution in Peromyscus mice. *Nature* **493**: 402–5.

Wegmann D, Leuenberger C, Neuenschwander S, Excoffier L (2010). ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* **11**: 116.

White BJ, Cheng C, Simard F, Costantini C, Besansky NJ (2010). Genetic association of physically unlinked islands of genomic divergence in incipient species of Anopheles gambiae. *Mol Ecol* **19**: 925–939.

Whitlock MC, Bronstein JL, Bruna EM, Ellison AM, Fox CW, McPeek MA, *et al.* (2015). A Balanced Data Archiving Policy for Long-Term Studies. *Trends Ecol Evol* **xx**: 1–2.

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, *et al.* (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**: 1660–1666.

Yang Z (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Others (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**: 565–569.

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**: 3326–3328.

Zhou X, Stephens M (2012). Genome-wide efficient mixed model analysis for association studies. *Nat Genet* **44**: 821–824.

Tables

Table 1. Summary of methods dedicated to data description and assessing population structure. Methods highlighted in bold can be combined in a pipeline within the R software.

Software	Type of method	Purpose	Specifics	Issues and warnings	Link	Reference
SPRelate	**Multivariate analysis**	**Summarizing variance across loci and visualizing inter-individual 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 inter-individual 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 HWE 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)**
BEDASSLE	**Differentiation and MCMC model testing**	**Identifies contribution of environment and geographical distance to populations 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)**
GENELAND	**Clustering and characterizing admixture**	**Grouping individuals in spatially consistent clusters maximizing HW equilibrium**	**Takes into account spatial variation, supposed to detect weak structure, framed in R**	**Immigrant alleles are assumed to be found only in new immigrants**	**https://cran.r-project.org/web/packages/Geneland/**	**(Guillot *et al.*, 2012)**
sNMF	**Clustering and characterizing admixture**	**Grouping individuals in clusters maximizing HW equilibrium and LD between loci**	**Fast (30X than ADMIXTURE)**	**Still slow computation time for large datasets**	**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.	Slow for large datasets. Requires specific input format	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	~100X 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, claimed to be faster than Structure	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, specific tools for RAD-seq, 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)
Splitstree	Phylogeny/Network	Network reconstruction and phylogenetic relationships	User friendly interface, proposes a variety of methods for networks reconstruction	Mostly descriptive	http://www.splitstree.org/	(Huson and Bryant, 2006)
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)
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)
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 datasets. Requires sequence data that can be produced by , e.g., Stacks for RAD-seq data	http://beast2.org/	(Drummond and Rambaut, 2007; Bouckaert *et al.*, 2014)
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	http://www.atgc-montpellier.fr/phyml/binaries.php	(Guindon *et al.*, 2010)

SNAPP	Phylogeny	Phylogenetic relationships	Handles SNP data	Remains slow for medium to large datasets (>1,000SNPs)	http://beast2.org/snapp/	(Bryant *et al.*, 2012)
*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 datasets. Requires sequence data. Not suited for situations where gene flow/admixture occurrs	http://beast2.org/	(Heled and Drummond, 2010)
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)
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 4 taxa. Impractical for more than 6 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, 2016)
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 multi-purpose 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)
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)
Arlequin	AMOVA	Characterizing hierarchical population structure	Arlequin allows for a variety of other analyses of diversity	Requires a priori assignment of individuals to populations, data formatting is required prior analysis	http://cmpg.unibe.ch/software/arlequin35/Arl35Downloads.html	(Excoffier and Lischer, 2010)
POPTREE2	Genetic distance	Visualizing a matrix of pairwise differentiation statistics as a tree	Can be used for pooled datasets, 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)
Stacks	Differentiation/Diversity/Phylogeny	Processing RAD-seq data and facilitate their analysis	Designed for RAD-seq data, variety of output formats for downstream analyses. Allows to retrieve DNA sequences for each locus	NA	http://catchenlab.life.illinois.edu/stacks/	(Catchen *et al.*, 2011)
Popoolation/Popoolation2/Popoolation TE	Differentiation/Diversity/Recombination	Extracting summary statistics from pooled data	Explicitely corrects for sampling bias in pooled data	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	Mostly limited to summary statistics	https://github.com/ANGSD/angsd	(Korneliussen *et al.*, 2014)
Arlequin	Differentiation/Diversity/Recombination	Computing summary statistics based on AFS and LD along genomes	Can output AFS for further analysis in fastsimcoal2	Slower than PopGenome, requires a private format	http://cmpg.unibe.ch/software/arlequin35/Arl35Downloads.html	(Excoffier and Lischer, 2010)
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)
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, private format, 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, 2005)

Table 2. Summary of common methods for identifying loci under selection. Methods available in R are highlighted in bold.

Software	Class of methods	Purpose	Specifics	Issues and warnings	Link	Reference
Bayescan	Population differentiation	Detecting positive selection and local adaptation	Incorporates uncertainty on allele frequencies due to low sample sizes	Sensitive to priors on the ratio of selected:neutral sites. False positive rates can be high under scenarios of demographic expansion, admixture and isolation by distance	http://cmpg.unibe.ch/software/BayeScan/	(Foll and Gaggiotti, 2008)
FDIST2	**Population differentiation**	**Detecting positive selection and local adaptation**	**Allows to control for hierarchical population structure**	**False positive rate is high when an island model cannot be assumed**	**http://datadryad.org/resource/doi:10.5061/dryad.v8d05**	**(Beaumont and Balding, 2004)**
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)**
PCAdapt	**Population differentiation**	**Detecting positive selection and local adaptation**	**Does not require to define populations. Handles admixed populations and pooled datasets**	**False positive rate can be high**	**http://membres-timc.imag.fr/Michael.Blum/PCAdapt.html**	**(Duforet-Frebourg *et al.*, 2016)**
Bayenv, BayPass	Population differentiation/Association	Detecting positive selection and adaptation to environmental features	Less sensitive to population demographic history than previous methods. Handle pooled datasets	Significance thresholds need to be determined from pseudo-observed datasets. 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)
LFMM	**Environmental association**	**Detecting adaptation to environmental features**	**Corrects for population structure using latent factors, faster than BAYENV for large datasets**	**Only performs association with environment**	**http://membres-timc.imag.fr/Olivier.Francois/lfmm/software.htm**	**(Frichot *et al.*, 2013)**
PLINK	Association	Detecting association with environmental/phenotypical features	Handles a variety of tests for population structure and relatedness	Population structure/kinship need to be assessed prior association analysis	http://pngu.mgh.harvard.edu/~purcell/plink/	(Purcell *et al.*, 2007)

GENABEL	**Association**	**Detecting association with environmental/phenotypic features**	**Modularity, facilitates correction for population structure/relatedness.**	**Imports data from PLINK format**	**http://www.genabel.org/**	**(Aulchenko et al., 2007)**
GEMMA	Association	Detecting association with environmental/phenotypical features	Computationnally efficient for large scale datasets	Imports data from PLINK format	http://www.xzlab.org/software.html	(Zhou and Stephens, 2012)
ANGSD	Summary statistics/Association	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)
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	(Korneliussen *et al.*, 2014)
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)
POPGenome	**Summary statistics**	**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)**
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)
SweeD	Composite Likelihood test	Designed for whole genome data (or large continuous regions)	Supports Fasta and VCF formats. Estimates for selection coefficients.	Better suited for whole genome datasets	http://pop-gen.eu/wordpress/software/sweed	(Degiorgio *et al.*, 2016)
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)
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)**
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)**
ARGWeaver	Ancestral recombination graphs	Detecting selection by screening for variation in topology and age of alleles	Provides quantitative estimates for TMRCA and topologies at each locus. Can be used to infer demographic history. Especially useful to identify signature of long-term balancing selection (older coalescence times)	High computing cost. Requires phased whole-genome data.	https://github.com/mdrasmus/argweaver	(Rasmussen *et al.*, 2014)
msms	Coalescence	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)
discoal	Coalescence	Simulate selective sweeps under arbitrary demographic scenarios	More specifically designed for studying soft and hard sweeps	Redundant with msms	https://github.com/kern-lab/discoal	Publication embargoed (Kern and Schrider, 2016)
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/	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4296155/
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)
SCCT	Conditional coalescent tree	Detecting positive selection	Designed for detecting recent positive selection. Clains to be more precise at identifying selected sites	Requires whole-genome data. The ancestral state of alleles must be obtained through an outgroup	https://github.com/wavefancy/scct	(Wang *et al.*, 2014)

Table 3. Summary of methods for demographic inference, simulations and scenarios comparisons. Methods available in R are highlighted in bold.

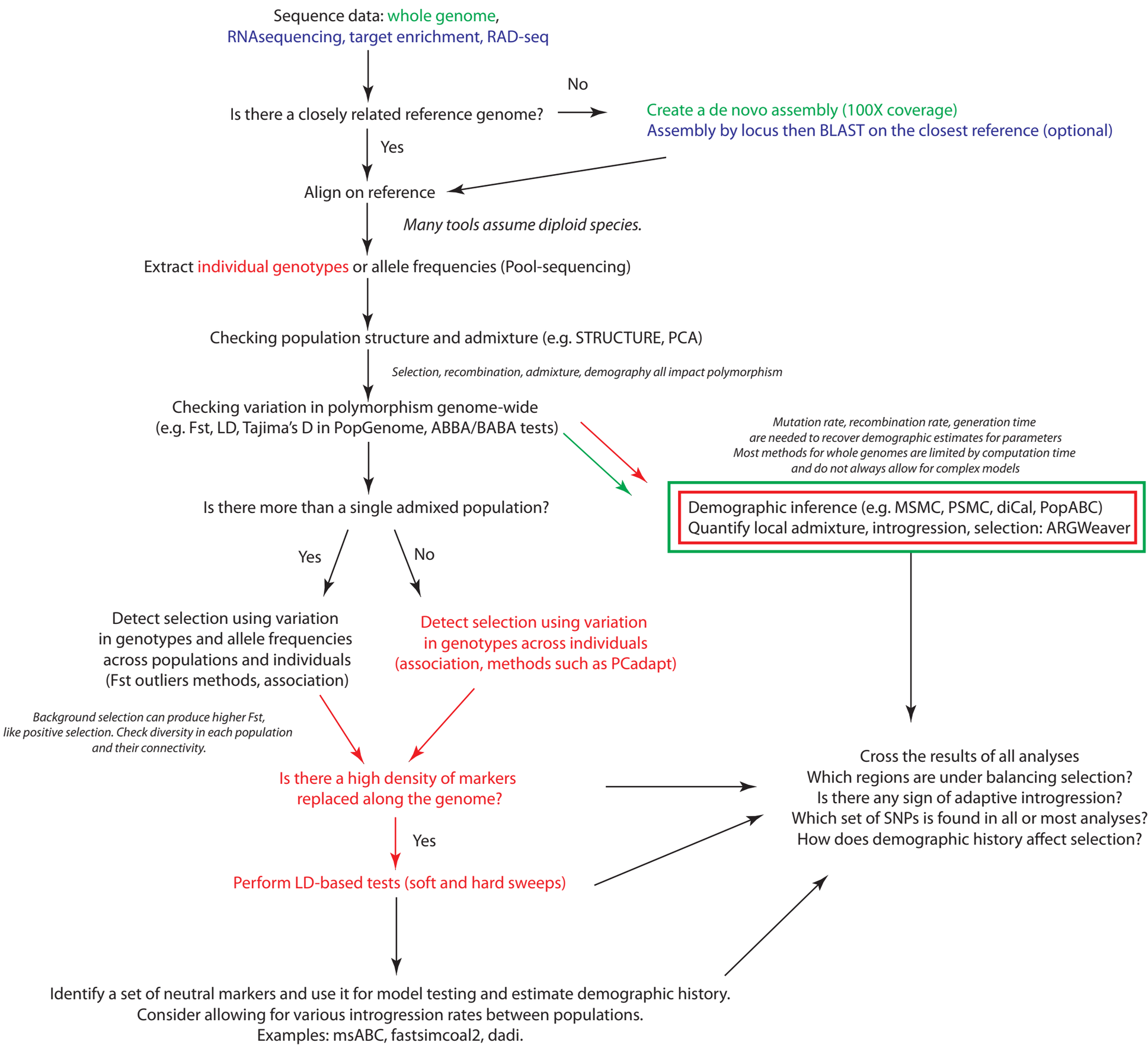
Methods	Software	Purpose	Specifics	Issues and warnings	Link	Reference
Identity by state tract	Unnamed	Predict observed patterns of Identity by state along a genome by fittingan appropriate, arbitrary complex demographic model	Allows bootstrapping and estimating confidence over parameter estimates with ms	Specific input format (similar to MSMC or ARGWeaver)	https://github.com/kelleyharris/Inferring-demography-from-IBS	(Harris and Nielsen, 2013)
IBD tract	DoRIS	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)
Sequentially Markovian coalescent	diCal2	Testing any arbitrary demographic scenario	Works with smaller, more fragmented datasets 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)
Sequentially Markovian coalescent	PSMC	Inferring change in Ne with time using a single diploid genome	Allows to track population size changes in time without a priori.	Limited to one population and one diploid individual. Better used within MSMC. Requires phased whole genome data and masking regions with insufficient sequencing depth	https://github.com/lh3/psmc	(Li and Durbin, 2011)
Sequentially Markovian coalescent	MSMC	Inferring change in Ne and migration rates with time between two populations	Allows to track population size changes in time without a priori. Allows estimating variation in cross-coalescence rate between two populations	Limited to the study of 8 diploid individuals from 2 populations at once. Requires whole genome phased data and masking regions with insufficient sequencing depth	https://github.com/stschiff/msmc	(Schiffels and Durbin, 2014)
Bayesian	Migrate-n	Inferring migration rates	Both ML and bayesian methods can be used to estimate parameters	Only estimates population sizes and migration rates. Not suited for large datasets. Private input format	http://popgen.sc.fsu.edu/Migrate/Migrate-n.html	(Beerli and Palczewski, 2010)

Bayesian	IMa2	Inferring parameters from an isolation with migration model	Fully bayesian approach, can perform joint estimates of parameters in L-mode and test for nested models	IM model is the only one available. Discrete admixture cannot be tested. Long computation times. Recent splits lead to overestimate migration rates	https://bio.cst.temple.edu/~hey/software/software.htm#IMa2	(Hey and Nielsen, 2007)
Bayesian	G-PhoCS	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)
Ancestral Recombination Graphs/coalescence	ARGWeaver	Retracing the whole process of recombination and coalescence along a genome	Provides quantitative estimates for TMRCA and topologies at each locus. Estimates effective population size. Provides tools to extract summary statistics for the topologies retrieved.	High computing cost. Requires phased whole-genome data.	https://github.com/mdrasmus/argweaver	(Rasmussen *et al.*, 2014)
coalescent simulations	ms, msms, msABC	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.	Can be difficult to handle for the naive user (but see coala)	http://www.bio.lmu.de/~pavlidis/home/?Software:msABC	(Hudson, 2002; Ewing and Hermisson, 2010; Pavlidis *et al.*, 2010)
coalescent simulations	scrm	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)
coalescent simulations	fastsimcoal2	Building any arbitrary scenario using a coalescent framework	Any arbitrary scenario can be implemented. Handles SNP, microsatellites and sequence data.	Does not handle selection	http://cmpg.unibe.ch/software/fastsimcoal2/	(Excoffier and Foll, 2011)
ABC/Likelihood	fastsimcoal2	Model comparison and parameters estimation	Performs coalescent simulations, parameter estimation and model testing using a fast likelihood method. Can handle arbitrarily complex scenarios for any type of marker	Summary statistics need to be calculated through Arlequin, slowing down their computation	http://cmpg.unibe.ch/software/fastsimcoal2/	(Excoffier *et al.*, 2013)

ABC/coalescent simulations	**coala**	**Combining coalescent simulators within a single framework**	**Facilitates the building of scenarios and computes summary statistics for simulations**	**Includes so far ms, msms and scrm**	**https://cran.r-project.org/web/packages/coala/index.html**	**(Staab and Metzler, 2016)**
ABC	**abc**	**Performs all steps for model-checking and parameters estimation for ABC analyses**	**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**	**(Csilléry *et al.*, 2012)**
ABC	DIYABC	Complete ABC analysis, from simulations to model checking and parameters estimation	User-friendly	Does not allow to model continuous gene flow	http://www1.montpellier.inra.fr/CBGP/diyabc/	(Cornuet *et al.*, 2008)
ABC	ABCToolbox	Complete ABC analysis, from simulations to model checking and parameters estimation	Modular, facilitates the computation of summary statistics	Current version is Beta (15/01/2016)	https://bitbucket.org/phaentu/abctoolbox-public/	(Wegmann *et al.*, 2010)
ABC	PopSizeABC	Inferring change in Ne 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)
Diffusion approximation of the AFS	∂a∂i	Model comparison and parameters estimation	Run time does not depend on the number of SNPs included, does not require coalescent simulations, handles arbitrarily complex scenarios	requires some knowledge of Python. Limited to 3 populations	https://bitbucket.org/gutenkunstlab/dadi	(Gutenkunst *et al.*, 2009)

Figures

Figure 1. A possible pipeline for analysing population genomics data using methods described in this paper. In red are indicated options that are generally not suited for pool-seq data. In green are indicated steps that require genome-wide datasets.


Sequence data: whole genome,
RNAsequencing, target enrichment, RAD-seq
Is there a closely related reference genome?
No
Create a de novo assembly (100X coverage)
Assembly by locus then BLAST on the closest reference (optional)
Yes
Align on reference
Many tools assume diploid species.
Extract individual genotypes or allele frequencies (Pool-sequencing)
Checking population structure and admixture (e.g. STRUCTURE, PCA)
Selection, recombination, admixture, demography all impact polymorphism
Checking variation in polymorphism genome-wide
(e.g. Fst, LD, Tajima's D in PopGenome, ABBA/BABA tests)
Mutation rate, recombination rate, generation time
are needed to recover demographic estimates for parameters
Most methods for whole genomes are limited by computation time
and do not always allow for complex models
Demographic inference (e.g. MSMC, PSMC, diCal, PopABC)
Quantify local admixture, introgression, selection: ARGWeaver
Is there more than a single admixed population?
Yes
No
Detect selection using variation
in genotypes and allele frequencies
across populations and individuals
(Fst outliers methods, association)
Detect selection using variation
in genotypes across individuals
(association, methods such as PCadapt)
Background selection can produce higher Fst,
like positive selection. Check diversity in each population
and their connectivity.
Is there a high density of markers
replaced along the genome?
Yes
Perform LD-based tests (soft and hard sweeps)
Cross the results of all analyses
Which regions are under balancing selection?
Is there any sign of adaptive introgression?
Which set of SNPs is found in all or most analyses?
How does demographic history affect selection?
Identify a set of neutral markers and use it for model testing and estimate demographic history.
Consider allowing for various introgression rates between populations.
Examples: msABC, fastsimcoal2, dadi.