bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Going down the rabbit hole: a review on methods characterizing selection and demography in natural populations

Running title: Linking evolution and genome-wide data

Yann X.C. Bourgeois[1], Khaled M. Hazzouri[1], Ben H. Warren[2]

[1]**New York University Abu Dhabi**, PO Box 129188, Saadiyat Island, Abu Dhabi, United Arab Emirates

[2]Department of Systematic and Evolutionary Botany, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland

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 ever-increasing number of genetic markers across genomes of non-model species.
2. The study of variation in these markers across natural populations has deepened the understanding of how population history and selection act on genomes. Population genomics now provides tools to better integrate selection into a historical framework, and take into account selection when reconstructing demographic history. However, this improvement has come with a burst of analytical tools that can confuse users.
3. Such 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. To address this need, we describe possible analytical protocols and link these with more than 70 methods dealing with genome-scale datasets. We summarise the

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

strategies they use to infer demographic history and selection, and discuss some of their limitations. A website listing these methods is available at www.methodspopgen.com.

Keywords

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 10 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 understanding of the history of human migrations, admixture and adaptation (e.g. Sabeti *et al.*, 2002; Abi-Rached *et al.*, 2011; Li and Durbin, 2011), the origin of domesticated species (e.g. Axelsson *et al.*, 2013; Schubert *et al.*, 2014), and the genetic basis of local adaptation in both model and non-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 has increased enormously in the last five years, even for non-model species. Studying genetic variation at the genome level allows the demographic factors shaping species history to be characterised. Further, understanding demographic history is important in correctly identifying loci under selection. Such data can even help in conservation efforts by identifying locally adapted genes that can be used to define relevant conservation units (Fraser and Bernatchez, 2001).

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In the last 10 years, developments in NGS have continually improved the throughput of data, while reducing time and cost of their production. These methods have become more affordable for teams studying evolutionary processes in biology, and many new methods to infer demography and selection have been developed. However, these methodological advances have brought increased analytical complexity to the field, and an inflation in the number of methods covering any one topic. As a consequence, it has become increasingly difficult for all potential users to follow developments and be sure of selecting the most appropriate method for the question and data in hand.

An overarching theme that concerns new users in a wide range of contexts is understanding patterns of heterogenous diversity along the genome. Patterns of nucleotide variation in genomes are shaped by both intrinsinc and extrinsic factors. Even within a single isolated panmictic population, interaction between recombination, selection and historical variation in population size will lead to heterogeneous diversity along the genome. At the scale of several connected populations or even between emerging species, these processes will affect the rate at which migration homogenizes the genome (Wolf and Ellegren, 2016).

A prime example is the situation of a researcher primarily interested in identifying signatures of recent positive selection in a species of interest. Since a new mutation will see its frequency increase in a population where it provides a selective advantage (*i.e.* hard selective sweep), a large region around it can remain uniform, especially if selection is strong (Sabeti *et al.*, 2002; McVean, 2007; Vitti *et al.*, 2013). This can lead to an increase in linkage disequilibrium (LD) between variants associated to the advantageous mutation, as well as a decrease in the age of the positively selected alleles and their nucleotide diversity. If positive selection occurs only

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in some populations, it may be possible to observe an increase in differentiation at this locus (Charlesworth *et al.*, 1997). To detect this signature of selection, some methods can track particularly long haplotypes and linkage disequilibrium along the genome. Others will rather focus on allele frequency spectrum and nucleotide diversity. Association methods will take advantage of preliminary knowledge of a phenotype or environment to identify loci displaying correlated allele frequencies. A few methods aim at inferring the whole history of coalescence and recombination along genomes, but still make simplifying assumptions and often require whole-genome resequencing data, which remain unaffordable for many teams.

Therefore, the choice of methods of any such researcher will depend on the available data and specifics of the question being addressed. One key aspect is that all these methods and questions do not have the same requirements in terms of reference genomes and marker density. For example, recent discussion of RAD-markers has been interesting from this perspective (Lowry *et al.*, 2016; Catchen *et al.*, 2017). The density of markers obtained along a genome depends on the choice of the restriction enzyme, and this choice must take into account the average extent of LD. Genome scans of selection will lose power if this density is not enough to cover mutations in strong linkage with variants under selection.

In the absence of any unified framework, combining several tools is necessary to interpret results. It must be borne in mind that recombination rates vary along the genome, which can possibly bias tests based on LD. It can therefore be important to characterize the recombination landscape in natural populations, requiring the use of another method (e.g. LDHat, Table 1). Background selection can lead to signatures of high differentiation that mimick disruptive selection (Charlesworth *et al.*, 1997). An assessment of genetic diversity

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

within populations, haplotype frequencies and possibly association with phenotype in each population would therefore be needed to explore this possibility (Charlesworth *et al.*, 1997). Demographic history impacts patterns of LD, allele age and frequencies at the genome scale, and affects the efficiency of selection at specific genes. This calls for at least basic checking of demographic structure and history and ideally building neutral demographic models to estimate the expected frequency of outliers without involving selection. In addition, most methods estimating selection coefficients require estimating effective population sizes. Finally, including markers under selection can bias demographic inference by skewing allele frequency spectra and LD, which requires careful data filtering and removal of outliers.

In this simplified example, we see that a reciprocal feedback between different aspects of evolutionary genomics is needed (Figure 1). Combining approaches is one of the current grand challenges in evolutionary biology (Cushman, 2014). While large-scale collaborations and sharing of skills between researchers allow for detailed analyses, a regularly updated list of methods would be valuable for smaller research teams to quickly start new projects and evaluate their experimental design.

In addition to methodological and technical challenges, the widespread use of sophisticated analytical tools is made difficult by the lack of communication between fields (Shafer *et al.*, 2015), little user-friendliness of software, inflation of data formats (Lischer and Excoffier, 2012) and the ever-increasing number of methods made available. Fields like landscape genetics and phylogeography have largely focussed on identifying general patterns in populations history and species diversification. Other researchers are more interested in identifying specific genes that are involved in adaptation in natural populations. All these

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

views contribute to our understanding of causation in biology, an effort that has included genetics, developmental science and ecology (Laland *et al.*, 2011). A global summary of methods used in these different fields would therefore facilitate communication between disciplines.

The last extensive review of methods in population genetics was performed 10 years ago (Excoffier and Heckel, 2006). Since then there has been increasing drive to translate these methods into approaches applicable to genomic data and non-model species. This drive has confirmed the value of population genomics on non-model species in understanding biological diversity at various scales (Mandoli and Olmstead, 2000; Jenner and Wills, 2007; Abzhanov *et al.*, 2008; White *et al.*, 2010; Ellegren *et al.*, 2012; Weber *et al.*, 2013; Poelstra *et al.*, 2014). Such 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 recently been made in facilitating the diffusion of sometimes complex, state-of-the-art methods. Their application to species with little background data has become more accessible, bringing the potential to add much valuable information.

In this paper, we propose possible pipelines (Figures 1, 2 and 3) to help choose appropriate methods dealing with current 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

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

population structure and quantitative characterization of population history (Table 1 and 2) and ii) methods aimed at identifying selected loci (Table 3). We end this review by detailing how these analyses can be combined, and present future directions that may be taken by the field of population genomics. The tables and a summary of the methods discussed in this paper will be kept updated to follow improvements, and are available at www.methodspopgen.com.

Box 1. Common sequencing methods

RAD-seq: Reduced representation allows broad sampling of variants across the genome by sequencing DNA fragments flanking restriction sites. Such sampling is not specific to any particular kind of region (e.g. coding or non-coding). Some of the best-known reduced representation techniques include RAD-sequencing (Baird *et al.*, 2008) and Genotyping by Sequencing (GBS; Elshire *et al.*, 2011). Their main interest is their low cost and that they do not require any reference genome (see Davey *et al.*, 2011 for details), although a reference can be useful to identify outlier genomic regions and retrieve linkage disequilibrium information between markers. Use of a reference genome also limits the bias due to paralogy and mapping errors (Hand *et al.*, 2015). Reduced representation allows many individuals to be genotyped at once, and so is widely used for the study of population structure, demography and selection. It does not cover all mutations in the genome and the choice of the restriction enzyme is crucial to control for the density of markers. This choice further controls the mean sequencing depth, the number of mutations close to genes under selection, and the accurate calling of genotypes. The number of SNPs ranges from thousands to millions, which is usually enough to retrieve substantial information about demography and sometimes selection (see Puritz *et al.*, 2014 for a detailed summary of reduced-representation techniques). As a general word of caution, note that RAD-sequencing and related methods display specific properties that can bias genome-

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

wide estimates of diversity, e.g. allelic dropout (Arnold *et al.*, 2013, Puritz *et al.* 2014). However, this type of marker remains valuable for phylogenetic estimation, even for distantly related species (Cariou *et al.*, 2013), and allelic dropout can be compensated for by focusing only on markers sequenced in all individuals. Variations on the original RADseq protocol have been developed to overcome some of these caveats (ddRAD, Peterson *et al.*, 2012; ezRAD, Toonen *et al.*, 2013; 2b-RAD, Wang *et al.*, 2012). Many pipelines have been specifically designed to account for RAD-seq specificities, including Stacks (Catchen *et al.*, 2011), TASSEL-UNEAK (Lu *et al.*, 2013) or TASSEL-GBS for GBS data (Glaubitz *et al.*, 2014).

Targeted sequencing: This class of methods allows sequencing and genotyping the same set of genomic fragments or single nucleotide polymorphisms (SNP arrays) across individuals, and has been recently promoted to study non-model species (Jones and Good, 2016). Since the specificity of the probe does not have to be very high, the same probe can be used among closely 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). Commercial products, such as Agilent's SureSelect, MYcroarray's MYbaits or Roche NimbleGen's SeqCap offer these methods or a derivation (Grover *et al.*, 2012).

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 complexity of analysis is reduced compared to whole genome sequencing (WGS), since only a subset of genomic regions is sequenced. By allowing

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

an improvement in spatial and temporal sampling, targeted sequencing can reconstruct dispersal routes and migration between varieties and subspecies (Nadeau *et al.*, 2012; da Fonseca *et al.*, 2016). Another commonly used technique includes Single nucleotide polymorphism (SNP) genotyping arrays have frequently been used in studies aimed at detecting phenotype/genotype associations or to study population struc ture (Gautier *et al.*, 2010; Johnston *et al.*, 2011). However, regenotyping of ascertained SNPs in a new population can lead to bias which can be problematic for demographic inference (Albrechtsen *et al.*, 2010; Lachance and Tishkoff, 2013).

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 applications on many different evolutionary time scales. Since it mostly sequences 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 which must be taken into consideration (Gayral *et al.*, 2013). If a reference genome is available, it is possible to call variants (Piskol *et al.*, 2013). This method is cost-effective and an alternative to whole genome sequencing. However, common variant callers do not behave well with RNAseq due to reads encompassing intronic regions as well as bias introduced during the sequencing library preparation. One of the common variant calling pipelines available is GATK which suggests best practices for calling variants on RNAseq (https://software.broadinstitute.org/gatk/best-practices/). Another variant calling protocol specifically designed for RNAseq is Opossum (Oikkonen and Lise, 2017), which can be used with haplotype-based callers such as Platypus and GATK haplotypeCaller. This software maintains precision and improves the sensitivity of SNP calling compared to the GATK best practice pipeline. RVboost (Wang, Davila, *et al.*, 2014) was developed using the method of

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variant prioritization, using a so-called boosting method that uses a set of high-confidence variants to set a model of good quality variants. All RNA variants are then prioritized and called based on this model. It outperforms Variant Quality Score Recalibration (VQSR) from the Genome Analysis Tool Kit (GATK) and the RNA-Seq variant calling pipeline SNPiR (Piskol *et al.*, 2013). RVboost can indentify false variants introduced by random hexamer priming during library preparation.

Whole genome resequencing: Whole-genome resequencing requires a well assembled reference and is more expensive than RAD-seq or targeted sequencing, especially for species with long and complex genomes. Some methods do not actually require any reference sequence to call SNPs from raw reads, like kSNP2 (Gardner and Hall, 2013) or DiscoSNP (Uricaru *et al.*, 2015). However, this limits the main interest of this approach, since mapping back on a reference has the potential to provide a complete overview of structural and coding variation. It also allows the use of powerful methods to track signatures of selection (see below). Pooled sequencing (Futschik and Schlötterer, 2010) can be an option to reduce costs, but generally restricts analyses 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 into the evolutionary history of populations or species, and 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

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Population structure and data description

Population structure and diversity

Description of the data is essential to assess the proportion of loci displaying a consistent pattern, and characterize how genetic diversity is partitioned within species. Genetic diversity and its genome-wide variance are directly impacted by variation in many factors including effective population sizes, population structure, inbreeding, migration, and recombination rates. Their characterization must be performed prior to any analysis to get insights into the forces and constraints acting on populations.

A key aspect when describing a new dataset is the assessment of relatedness between individuals or localities. Neglecting population structure can dramatically bias demographic inference, especially when gene flow is not accounted for or panmixia is assumed (Chikhi *et al.*, 2010; Heller *et al.*, 2013). It also biases the detection of loci under selection (e.g. Nielsen *et al.*, 2007). Cryptic population structure is typically a confounding effect in studies of phenotype-genotype association studies, when a given feature or trait is disproportionally found in a population or a set of related individuals (Balding, 2006). Fortunately, the abundance of SNP data produced by typical genomic studies is often enough to thoroughly assess relatedness between individuals.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Many tools currently exist to infer population structure (Table 1, Figure 2). An elegant and efficient class of methods relies on using multivariate approaches such as principal component analysis (PCA) to infer relatedness between individuals and populations without *a priori* knowledge. Since these methods do not have underlying assumptions based on population genetics, they are suitable for analyzing species displaying polyploidy or mixed-ploidy (Dufresne *et al.*, 2014). A detailed review of these methods has been already performed (Jombart *et al.*, 2009) and an exhaustive list of their applications is beyond the scope of this review. These approaches have been especially useful to study the consistency between geographical and genetic structure in human populations of Europe (Novembre *et al.*, 2008). They were also recently applied to RAD-sequenced populations of a freshwater crustacean (*Daphnia magna*). Procrustes rotation (Novembre *et al.*, 2008) was used to match geographical coordinates with PCA axes, showing how isolation by distance has shaped genetic structure (Fields et al. 2015).

Methods for estimating the relatedness of individuals are suited to 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. Genetic distance can then be compared with geographical or ecological distance. For example, in a recent study using more than 1000 *Arabidopsis thaliana* genomes, estimates of relatedness have allowed the identification of putatively relictual populations that may have persisted in Europe since the last Ice Age (Alonso-Blanco *et al.*, 2016).

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Approaches such as Structure (Pritchard *et al.*, 2000) and fastSTRUCTURE (Raj *et al.*, 2014) have been widely used to determine hierarchical population structure and admixed populations by grouping individuals in clusters. The optimal number of clusters (K) can then be determined based on likelihood, although examining population structure for a range of K can allow substructure to be better identified. The main interest of these approaches is that they provide a measure of coancestry coefficients, which are the proportions of an individual genome originating from multiple ancestral gene pools. Such information is more difficult to retrieve with approaches such as PCA. There have been criticisms however about whether ambiguous assignment could be actually interpreted as a signal of admixture, and detailed inference requires thorough model testing and estimating the goodness of fit of a model with admixture (see Falush *et al.*, 2016).

Heterogenous patterns of divergence between species along their genomes

Advantageous alleles can migrate from one population to another, resist introgression from other populations, reach fixation and erase diversity around them. This is one scenario leading to heterogenous patterns of divergence along the genome, the so-called islands of divergence (Wolf and Ellegren, 2016). Alternative scenarios leading to similar patterns were recently highlighted (Cruickshank and Hahn, 2014). Understanding the origin of genomic regions under selection highlights the evolutionary history of adaptive alleles (e.g. Abi-Rached *et al.*, 2011) and contributes to our understanding of the origin and maintenance of reproductive isolation. Studies focusing on hybrid zones and introgression have provided inspiring examples (Hedrick, 2013), as demonstrated by recent work focusing on patterns of heterogenous gene flow in *Mytilus* mussels (Roux *et al.*, 2014), localized introgression and

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

inversions at a color locus in *Heliconius* butterflies (The Heliconius Genome Consortium *et al.*, 2012) and adaptive introgression of anticoagulant resistance alleles in mice (Song *et al.*, 2011). Descriptive statistics computed along genomes provide valuable information in this context. One may for example plot the distribution of a differentiation measure such as F_{ST} (Weir and Cockerham, 1984) between populations, mean linkage disequilibrium or nucleotide diversity. Such an approach has been used in *Ficedula* flycatchers, which uncovered clear genomic islands of divergence and the higher differentiation on sexual chromosomes due to ongoing reproductive isolation (Ellegren *et al.*, 2012). Other approaches, such as chromosome painting (Table 1), extend PCA and Structure-like methods by incorporating information about the relative order of markers in the genome, allowing identification of regions for which ancestry differs from the rest of the genome.

Heterogeneous structure in space: landscape genomics

Landscape (as well as seascape and lakescape) 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. Environmental heterogeneity has a strong impact on how genetic diversity is shaped by migration success between populations, for example after a range expansion (Wegmann *et al.*, 2006). A spatially explicit perspective provides context to understand the evolution of locally adapted genes. Moreover, identifying how and where populations (or closely related species, see Roux et al. 2016) hybridize is crucial when it comes to characterizing colonization trajectories, tension zones and secondary contacts (Gay *et al.*, 2008; Bierne *et al.*, 2011).

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Some methods can explicitly use spatial information to inform clustering, allowing improved consideration of the effect of landscape heterogeneity on selection against migrants and drift. This spatial perspective can be useful to visualize the location and shape of hybrid zones (Guedj and Guillot, 2011). Landscape genetics has valuable application in management and conservation, where it is useful to identify the relevant evolutionary significant units displaying spatial and ecological divergence. Furthermore, researchers are often interested in testing the impact of ecological variation on genetic structure. Mantel tests have been popular to investigate relationships between ecological variables and genetic differentiation while accounting for geographical distances. However, these tests are biased by spatial autocorrelation, assume linear dependence between variables, and do not allow testing the relative contribution of each variable (Legendre and Fortin, 2010; Guillot and Rousset, 2013). 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 dispersal. However, disentangling these effects has 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). Landscape genomics now extends its focus to adaptive genetic variation, and benefits from new methods targeting signatures of selection (Figure 2 and below).

Population history

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Phylogeny

Phylogenetics has a long history that is linked to the broader topic of systematics (Moritz & Hillis 1996; Baum & Smith 2013). Since their inception in the 1980s, molecular phylogenetic methods have been used to address a wide range of problems at different taxonomic scales, including intraspecific population history. Recent advances in molecular phylogenetic methods, and the employment of different types of NGS data is well beyond the scope of this review (see e.g. Moriarty Lemmon & Lemmon 2013; Cruaud et al. 2014; Wen et al. 2015). Rather we focus on the use of phylogeny within the context of studies of intra-specific population history and selection. In this respect, both Maximum Likelihood and Bayesian approaches have become popular to investigate evolutionary relationships between individuals from different populations, even when divergence is very recent (e.g. Wagner *et al.*, 2013). These methods are implemented in softwares such as RAxML (Stamatakis, 2014) and BEAST2 (Drummond and Rambaut, 2007). Ultimately, all molecular phylogenies reconstruct the geneaology of the genes with which they have been constructed. Therefore, a basic assumption when using them to infer lineage history at any taxonomic level (populations, species, and higher taxonomic units) is that the gene tree is representative of lineage history. This assumption is likely to be particularly weak at the population level, since the influences of gene flow, selection, and incomplete lineage sorting are strong at this scale, and may cause gene trees to deviate from population history. Nonetheless, such phylogenies can provide a useful starting point for inferences that are complemented with other methods.

When using genome-wide data at the population level, methods specifically dedicated to reconstructing multiple species coalescent models (MSC) such as *BEAST (STAR-BEAST) should be preferred over concatenation (Edwards *et al.* 2016), since they allow discordance between species trees and individual gene trees to be identified. Note that these methods can

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

be strongly biased when it comes to estimate divergence times and effective population sizes (Leaché *et al.*, 2014). The impact of gene flow and recombination on phylogenetic methods is however an alley of research that will allow better integration between phylogeny and population genetics (Edwards *et al.*, 2016). Such integration is particularly needed for species and populations that are in the “grey zone of speciation” (Roux *et al.*, 2016). Recent advances in MSC methods handling extremely short, non-recombining fragments (see Chou *et al.*, 2015 for a comparison) are promising, especially for datasets such as those produced by GBS.

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. Network methods implemented in Splitstree (Huson and Bryant, 2006), make less assumptions and account for potentially conflicting signals due to high gene flow. Unfortunately, such methods remain mostly descriptive.

Approximate Bayesian Computation

Phylogenetic methods tend to be slow for large datasets, and generally do not attempt to account for many effects that are crucial in population genetic interpretation, such as gene flow and recent demographic events within species. A more suitable framework for microevolutionary studies relies on coalescence theory. Population geneticists first developed coalescent theory as a way of modeling the genealogy of alleles from a sample of a large population. Going backward in time, alleles merge (coalesce) in a stochastic way until reaching their most recent common ancestor (Kingman, 1982). Obtaining demographic estimates (e.g. time in years) for parameters usually requires that mutation rate and generation

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

time be known or at least reasonably well estimated, for example from closely-related species with similar life history.

Computationally fast approaches include Approximate Bayesian Computation (ABC), which compares the empirical data with a set of simulated data produced by coalescent simulations under scenarios predefined by the user (Table 2). 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 analyses are described by Csilléry *et al.* (2010). The main advantage of ABC is that it allows handling any type of marker and arbitrarily complex models, contrary to methods like IMa where the model is predefined. However, using summary statistics leads to the loss of potentially useful information (Robert *et al.*, 2011).

Likelihood methods based on the allele frequency spectrum (AFS)

Recently, new likelihood methods based on the 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. alleles are likely to occur at more similar frequencies if divergence is recent or if populations are highly connected). These approaches quickly estimate parameters using composite likelihoods, and do not explicitly take into account correlations induced by LD between physically linked markers (but see ABLE, Table 2). This might limit power to detect recent demographic events (e.g. migration, Jenkins *et al.,* 2012). Including SNPs that are physically close together should not strongly bias parameter estimation. However, such an approach prevents direct comparisons of likelihoods from different models. Therefore, physically independent SNPs should be used to consider

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

composite likelihoods as quasi likelihoods for model comparison (Excoffier *et al.*, 2013). 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).

The number of mutations found in a given length of DNA sequence directly depends on the mutation rate. One drawback when using SNP data without considering monomorphic sites is that the mutation rate per generation can not be used to convert parameters into demographic estimates (Excoffier *et al.*, 2013). Another possibility consists of calibrating parameter estimates by including a fixed parameter in the analysis, such as population size or divergence time. An issue specific to SNP arrays is ascertainment bias, which is the systematic deviation of allele frequencies from theoretical expectations due to the choice of individuals used at the step of SNP discovery. For example, if SNPs found in one population are the only ones genotyped in another population, a whole set of markers polymorphic in the second population but not in the first will be missed, biasing the AFS (Lachance and Tishkoff, 2013).

Reaching a high level of precision when estimating demographic parameters can be challenging when information is lacking about the evolutionary history of the species considered. However, even when such information is lacking it is possible to compare the likelihoods of different demographic scenarios, a procedure that has been successfully applied to many species to shed light on the process of speciation (Roux *et al.*, 2016).

Methods using whole-genome resequencing

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Recently, methods have been developed to infer variation in population sizes with time using the whole genome of just one diploid individual. This began with the Pairwise Sequentially Markovian Coalescent (PSMC, Li and Durbin, 2011), and extensions have been made to this model to allow for several genomes. Such methods have the advantage of requiring only a few individuals, and no *a priori* knowledge of population history. One general drawback, however, is that they are limited to rather simple scenarios, and do not handle more than two populations as yet (but see diCal2, Table 2). While powerful, they are sensitive to confounding factors such as population structure (Orozco-terWengel, 2016) that lead to false signatures of expansion or bottleneck. They also do not allow extremely recent demographic events to be investigated, since the coalescence of two alleles from a single individual in the recent past (a few tens to hundreds generations) is infrequent. Moreover, most of these methods require the data to be phased (but see SMC++, Table 2), for example with fastPhase (Scheet and Stephens, 2006) or BEAGLE (Browning and Browning, 2011). In addition, phasing errors can lead to strong biases in parameters estimates for recent times (Terhorst *et al.*, 2016). An extension of these methods takes into account population structure and aims to identify the number of islands contributing to a single genome, assuming it is sampled from a Wright n-island meta-population (Mazet *et al.*, 2015). Such developments should improve the amount of information retrieved from only a few genomes. However, 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).

Methods based on tracts of identity-by-descent (IBD, Palamara and Pe’er, 2013) constitute an interesting alternative for more complex model testing when whole genome or densely genotyped datasets are available in large number. Such methods allow recent demographic events to be inferred with relative precision. They are used to predict the length of haplotypes

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

There is currently a tradeoff to be made between methods allowing for arbitrarily complex models that are defined *a priori* by the user (e.g. ABC), and methods that allow population history to be inferred agnostically (e.g. PSMC). While the first category of methods are typically the highest performers at inferring complex population history from a moderate number of markers, it is currently only the second category of methods that are able to make use of the full information provided by whole genome data. Using both methods can therefore help in accurately retrieving the evolutionary history of a given species. For example, a recent study on maize demographic and selective history used both ∂a∂i and and Markovian Coalescent methods to characterize the bottleneck and expansion associated with domestication (Beissinger *et al.*, 2016).

Screening for selection and association

Selection and its impact on sequence variation

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

While demographic forces such as drift and migration will affect the whole genome, selection is expected to be specific to particular portions of the genome, and therefore yield discrepancies with genome-wide polymorphism (Lewontin and Krakauer, 1973). Selection affects allele frequencies and polymorphism in predictable ways at the scale of single populations (Charlesworth, 2006; Charlesworth and Charlesworth, 2010). Several statistics summarize them, such as π, the nucleotide diversity (Nei and Li, 1979), Tajima's D (Tajima, 1989), and Fay and Wu's H (Fay and Wu, 2000). Using a combination of these statistics allows targets of selection to be identified with greater precision, and minimizes the confounding effects of demography (Nielsen *et al.*, 2005). This approach has been used to develop composite tests, such as the composite likelihood ratio (CLR) test (Nielsen *et al.*, 2005) that aim to detect recent selective sweeps.

Methods based on population subdivision

When an allele is under positive selection in a population, its frequency tends to rise to fixation, unless gene flow from other populations or strong drift prevents this from happening (Charlesworth *et al.*, 1997). It is therefore possible to contrast patterns of differentiation between populations adapted to their local environment to detect loci under divergent selection (e.g. displaying a high F_{st}). However, it is essential to control for population structure, as it may strongly affect the distribution of differentiation measures and produce high rates of false positives. First attempts to take into account population structure and variation in gene flow included FDIST2 (Beaumont and Nichols, 1996). This method models populations as islands and is aimed at detecting loci under selection by contrasting heterozygosity to F_{st} between populations. More sophisticated methods are now available (Table 3), dedicated to the detection of outliers in large genomic datasets. Most of them correct for relatedness across samples, and can test association between allele frequencies and

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

environmental features (see the extensive review by François *et al.*, 2015). These 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. The sampling strategy must take into account the particular historical and demographic features of the species investigated to gain power (Nielsen *et al.*, 2007). The sequencing strategy must also be carefully considered to control for spatial autocorrelation of genotypes due to isolation by distance and shared demographic history.

Genome-wide association

The methods described above focus on allele frequencies at the population scale, but do not test association with traits that vary between individuals within populations (e.g. resistance to a pathogen, symbiotic association, individual size or flowering time). For this task, methods performing Genome-wide association analysis (GWAS) are better suited. The recent development of multivariate methods such as PCAdapt (Duforet-Frebourg *et al.*, 2016) also allow loci putatively under selection to be identified in admixed or continuous populations without requiring information about individual phenotype.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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 are aimed at targeting those regions that display an excess of long homozygous haplotypes, such as the extended haplotype homozygosity (EHH) test (Sabeti *et al.*, 2002). It is also possible to compare haplotype extension across populations, with the Cross Population Extended Haplotype Homozygosity test (XP-EHH (McCarroll *et al.*, 2007)) or Rsb (the standardized ratio of EHH at a given SNP site (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. Statistics dedicated to the detection of selection on standing variation or on multiple alleles (so called soft sweeps) 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

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

actually be distinguished (Schrider *et al.*, 2015), as well as their relative importance (Messer and Petrov, 2013; Jensen, 2014).

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).

All LD-based approaches are more powerful with a relatively high density of markers, such as the ones obtained from whole-genome sequencing, SNP-arrays or high-density RAD-seq, and benefit from using statistics focusing on polymorphism and allele sharing. In a recent study of local adaptation in sticklebacks (Roesti *et al.*, 2015), these statistics have been used on dense RAD-sequencing data to look for recent selection at loci displaying high differentiation (F_{ST}). This approach has allowed new candidate loci to be pinpointed, and has confirmed the involvement of those implicated previously (e.g. the *Ectodysplasin* gene). In addition, the identification of large regions displaying high divergence and LD has revealed the importance of large-scale structural variation in shaping genome structure, such as inversions (Roesti *et al.*, 2015).

Detecting and characterizing selection with the coalescent

If a candidate locus or genomic region 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

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

be used. However, this requires that population history is known in order to control for other phenomena such as population structure and gene flow. An advantage of full coalescent methods is that they provide a relatively complete picture of the history of individual loci. This can be achieved by modeling coalescence and recombination, and considering variation in mutation rate. However, such methods have long been computationally intensive, and thus difficult to apply to whole genomes. Fortunately, recent computational improvements make their application to whole genomes feasible. A good example is ARGWeaver (Rasmussen *et al.*, 2014), which has allowed candidate genes for long-term balancing selection to be recovered from human data. This method uses ancestral recombination graphs to model the genealogy of each non-recombining block in the genome. Ancestral recombination graphs (ARG) are a generalization of the coalescent and describe the sequence of genealogies along a sample of recombining sequence. Genealogies are estimated for each non-recombining block, and recombination between adjacent blocks is described by breaking the branch leading to the recombining haplotype and allowing it to re-coalesce to the rest of the tree. This succession of local trees joined by recombination events provides a full description of the genealogical history of the data and is therefore a promising approach to characterize positive, purifying or balancing selection while taking into account variation in recombination and mutation rate.

Identifying variants of functional interest

Characterizing the number of synonymous versus non-synonymous mutations is another approach to detect whether a specific gene is undergoing purifying or positive selection. However, this approach requires an annotated genome. An excess of non-synonymous mutations can signal positive or balancing selection, or a relaxation of selective constraints on a given gene. Annotation of mutations can be done with SNPdat (Doran and Creevey, 2013), or directly in PopGenome, which can also perform tests of selection such as the MK test at the

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genome scale (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).

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 need to be combined with functional evidence (Vitti *et al.*, 2013). Such evidence can sometimes be provided by variation in the expression of a candidate gene highlighted by RNA-sequencing data. More often, developmental studies are required, 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). Essentially, controlled crosses allow genomic regions associated with a selected phenotype to be identified, while the study of variation in natural populations facilitates the fine-mapping of selected variants in natural populations. However, this requires that the species of interest can be raised in a laboratory or greenhouse, 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 passerines (bananaquits), GBS data have been used to obtain a neutral distribution to which patterns of substitution and differentiation were compared at candidate genes for color variation (Uy *et al.*, 2016). Another study on color polymorphism in *Peromyscus* mice used a combination of field experiments, targeted sequencing of candidate genes and neutral regions,

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and genome-scans for selection. Tests for association between these data were able to show how selection on many mutations at the same locus drive adaptive phenotypic divergence (Linnen *et al.*, 2013).

The combination of tests aimed at different signatures of selection can allow the size of candidate regions to be reduced. 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 different tests targeting signatures of selection, typically those using the allele frequency spectrum and those using haplotype length. A test of this type has been proposed in human genetics (Grossman *et al.*, 2013), and is called the composite of multiple signals (CMS) test. 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.

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". 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. This traditional approach has thus tended to disconnect the study of selection from the study of demography (Li *et al.*, 2012).

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

However, this assumption may be incorrect, and a joint understanding of demography and selection is crucial from this perspective (Figure 3). 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), which are difficult to assess with precision in non-model species. Moreover, it has been shown in the last few years that loci involved in reproductive isolation are often also involved in local adaptation. This, combined with variation in introgression rates along the genome, can bias inference about selection and demography (Bierne *et al.*, 2011; Roux *et al.*, 2014). Genomic regions with low recombination rates can lead to reduced polymorphism, and be mistaken for signatures 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 using the results obtained by one method to inform the other. Such an approach was taken by Tine et al. 2014 in investigating the two different lineages of the European Sea Bass, 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 possibly being the result of selection, low recombination, or a combination of the two (Tine *et al.*, 2014). Since differentiation along the genome seemed to reveal islands resisting gene flow, they could fit a model incorporating variation in introgression rates. This provided improved fit to the data and suggested that islands of differentiation are most likely to be due to locally reduced gene flow after secondary contact. This example illustrates how a combination of descriptive statistics and

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 SNPs presumed to be neutral (e.g. intergenic SNPs), followed by a step where the likelihood of a marker being under selection is evaluated. Methods such as BAYPASS or PCAdapt (Table 3) are convenient in both describing population structure and providing preliminary insights into the proportion of loci that do not follow neutral expectations. If this proportion is high, it would suggest recent introgression or an excess of markers displaying high LD (e.g. due to large inversions). However, when this proportion is not too high, outliers can be removed to avoid bias (Schrider *et al.*, 2016) and the remaining loci 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 high potential. They allow genealogies to be retrieved along the genome as well as the timing of coalescence events. Such information is ultimately useful for making inferences regarding selection and migration. Recently this method was used in human paleogenomics to quantitatively characterize introgression between modern humans, Neandertals and

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Denisovans using only a few whole genomes (Kuhlwilm *et al.*, 2016). However, the approach has a high computing and sequencing cost, and is therefore not suitable for studies requiring sampling of many individuals.

Caution must prevail when attempting to apply sophisticated methods to disentangle selection and demography. In a recent review, Cruickshank & Hahn suggest that IMa2, which is commonly used to estimate migration rates, is not able to reliably distinguish between loci resisting gene flow, and those under selection in the absence of gene flow (Cruickshank and Hahn, 2014). In the specific case they highlight (*Oryctolagus cuniculus* rabbits, Sousa *et al.*, 2013), a descriptive statistic 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 prior using more sophisticated methods. Note however that Cruickshank & Hahn did not address the case of secondary contact, and other methods such as ABC may better detect interruption in gene flow (Sup. Text in Roux *et al.*, 2016).

To sum up, the field of population genomics is now moving towards both better integrating the demographic framework in inferences of selection, and, conversely, taking into account selection when reconstructing demographic history. The joint inference of loci under selection and quantification of demographic dynamics is of crucial importance in fields such as landscape genomics or the study of ongoing speciation. It should provide insights into the role of selection, recombination and gene flow in promoting or impairing local adaptation to new habitats. The growing availability of genome-wide data for non-model species is therefore promising, but requires caution and high stringency in our interpretation of observed patterns.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

With the decreasing cost of sequencing, it has been suggested that NGS will rapidly broaden our perspective on complex evolutionary processes, from biogeography (Lexer *et al.*, 2013) to the genetic basis of traits (Hohenlohe, 2014) and the maintenance of polymorphisms (Hedrick, 2006). While genome heterogeneity in migration, mutation and 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 is already challenging in its own right. Nonetheless, in order to be informative about processes such as selection and demography it should ultimately be combined with other disciplines such as ecology and functional analyses (Habel *et al.*, 2015). 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.

Beyond SNPs: studying structural variation, transposable elements and epigenetic modifications

Most genome-scale studies of selection and demography have so far focused on SNPs, since they are relatively easy to detect with current technology 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 insertions of transposable elements (Villanueva-Cañas *et al.*, 2017), epigenetics modifications such as methylation (Danchin *et al.*, 2011), duplications, inversions, deletions and translocations (Iskow *et al.*, 2012). One of the main issues 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

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

et al., 2012), especially for species for which only a draft genome is available. It is however possible to use variation in such genetic alternations 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 selection (Villanueva-Cañas *et al.*, 2017) as highlighted by a recent study of transposable element 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 approach has recently been adopted in studies of selection on methylation patterns. In a recent Molecular Ecology issue (Verhoeven *et al.*, 2016), a study using bisulfite precipitation in Valley Oak trees (Gugger *et al.* 2016) was able to place methylated variants associated to climatic variables near to genes known to be involved in response to environment. Another study 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 models of mutation, complicating any comparison with neutral models built from SNPs. This is particularly true 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 are not under purifying selection

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(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. This would provide even more conservative assessments of negative and positive selection.

Acknowledgements

The University of Basel and New York University Abu Dhabi have supported YB's research in this area. We want to thank two anonymous reviewers, Stephane Boissinot, Joris Bertrand and Anne Roulin 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.

Albrechtsen A, Nielsen FC, Nielsen R (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol* **27**: 2534–2547.

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**:

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Balding DJ (2006). A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**: 781–91.

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.

Beeravolu CR, Hickerson MJ, Frantz LAF, Lohse K (2016). Approximate Likelihood Inference of Complex Population Histories and Recombination from Multiple Genomes. *bioarXiv*: 1–31.

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).

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Catchen JM, Hohenlohe PA, Bernatchez L, Funk WC, Andrews KR, Allendorf FW (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Mol Ecol Resour*: 362–365.

Charlesworth D (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet* **2**: e64.

Charlesworth B, Charlesworth D (2010). *Elements of evolutionary genetics*.

Charlesworth B, Nordborg M, Charlesworth D (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet Res, Camb* **70**: 155–174.

Chifman J, Kubatko L (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30**: 3317–3324.

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.

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

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.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Cushman SA (2014). Grand challenges in evolutionary and population genetics: The importance of integrating epigenetics, genomics, modeling, and experimentation. *Front Genet* **5**: 1–5.

Danchin É, Charmantier A, Champagne F a, Mesoudi A, Pujol B, Blanchet S (2011). Beyond DNA: integrating inclusive inheritance into an extended theory of evolution. *Nat Rev Genet* **12**: 475–86.

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.

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

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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

Edwards S V., Xi Z, Janke A, Faircloth BC, McCormack JE, Glenn TC, *et al.* (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Mol Phylogenet Evol* **94**: 447–462.

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, Heckel G (2006). Computer programs for population genetics data analysis: a survival guide. *Nat Rev Genet* **7**: 745–58.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Falush D, Dorp L van, Lawson D (2016). A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots. *bioRxiv*: 66431.

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.

Ferretti L, Ramos-Onsins SE, Pérez-Enciso M (2013). Population genomics from pool sequencing. *Mol Ecol* **22**: 5561–5576.

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.

Gardner SN, Hall BG (2013). When whole-genome alignments just won’t work: KSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. *PLoS One* **8**.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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, Laloë D, Moazami-Goudarzi K (2010). Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. *PLoS One* **5**: 1–11.

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.

Gautier M, Vitalis R (2013). Inferring population histories using genome-wide allele frequency data. *Mol Biol Evol* **30**: 654–668.

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**.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, *et al.* (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS One* **9**.

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.

Grover CE, Salmon A, Wendel JF (2012). Targeted sequence capture as a powerful tool for evolutionary analysis. *Am J Bot* **99**: 312–9.

Guedj B, Guillot G (2011). Estimating the location and shape of hybrid zones. *Mol Ecol Resour* **11**: 1119–1123.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Guillot G, Rousset F (2013). Dismantling the Mantel tests. *Methods Ecol Evol* **4**: 336–344.

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**.

Hand BK, Hether TD, Kovach RP, Muhlfeld CC, Amish SJ, Boyer MC, *et al.* (2015). Genomics and introgression: Discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Curr Zool* **61**: 146–154.

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

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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.

Jenkins PA, Song YS, Brem RB (2012). Genealogy-Based Methods for Inference of Historical Recombination and Gene Flow and Their Application in Saccharomyces cerevisiae. *PLoS One* **7**.

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.

Johnston SE, McEwan JC, Pickering NK, Kijas JW, Beraldi D, Pilkington JG, *et al.* (2011). Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. *Mol Ecol* **20**: 2555–2566.

Jombart T, Devillard S, Balloux F, Falush D, Stephens M, Pritchard J, *et al.* (2010).

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Jones MR, Good JM (2016). Targeted capture in evolutionary and ecological genomics. *Mol Ecol* **25**: 185–202.

Jostins L, McVean G (2016). Trinculo: Bayesian and frequentist multinomial logistic regression for genome-wide association studies of multi-category phenotypes. *Bioinformatics* **32**: 1898–1900.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Lachance J, Tishkoff SA (2013). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *Bioessays* **35**: 780–786.

Laland KN, Sterelny K, Odling-Smee J, Hoppitt W, Uller T (2011). Cause and effect in biology revisited: is Mayr's proximate-ultimate dichotomy still useful? *Science* **334**: 1512–6.

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.

Legendre P, Fortin MJ (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Mol Ecol Resour* **10**: 831–844.

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.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Lischer HEL, Excoffier L (2012). PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**: 298–299.

Liu L, Yu L (2011). Estimating species trees from unrooted gene trees. *Syst Biol* **60**: 661–667.

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

Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, *et al.* (2016). Breaking RAD: An evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. *Mol Ecol Resour*.

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.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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 (2007). The structure of linkage disequilibrium around a selective sweep. *Genetics* **175**: 1395–406.

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

Messer PW, Petrov DA (2013). Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol Evol* **28**: 659–669.

Mirarab S, Warnow T (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**: i44–i52.

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.

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

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Oikkonen L, Lise S (2017). Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. *Wellcome Open Res* **2**: 6.

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.

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.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* **7**: e37135.

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.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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**.

Robert CP, Cornuet J-M, Marin J-M, Pillai NS (2011). Lack of confidence in approximate Bayesian computation model choice. *Proc Natl Acad Sci U S A* **108**: 15112–7.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Schrider DR, Shanku AG, Kern AD (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics* **204**: 1207–1223.

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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Stucki S, Orozco-Terwengel P, Bruford MW, Colli L, Masembe C, Negrini R, *et al.* (2016). High performance computation of landscape genomic models integrating local indices of spatial association. *Mol Ecol Resour*: 1–15.

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.

Terhorst J, Kamm JA, Song YS (2016). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat Genet* **49**: 303–309.

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.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Toonen RJ, Puritz JB, Forsman ZH, Whitney JL, Fernandez-Silva I, Andrews KR, *et al.* (2013). ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ* **1**: e203.

Uricaru R, Rizk G, Lacroix V, Quillery E, Plantard O, Chikhi R, *et al.* (2015). Reference-free detection of isolated SNPs. *Nucleic Acids Res* **43**: e11.

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.

Villanueva-Cañas JL, Rech GE, de Cara MAR, González J (2017). Beyond SNPs: how to detect selection on transposable element insertions. *Methods Ecol Evol* **8**: 728–737.

Vitalis R, Gautier M, Dawson KJ, Beaumont MA (2014). Detecting and measuring selection from gene frequency data. *Genetics* **196**: 799–817.

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

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, *et al.* (2013). Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol Ecol* **22**: 787–98.

Wang C, Davila JI, Baheti S, Bhagwate A V., Wang X, Kocher JPA, *et al.* (2014). RVboost: RNA-seq variants prioritization using a boosting method. *Bioinformatics* **30**: 3414–3416.

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.

Wang S, Meyer E, McKay JK, Matz M V (2012). 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat Methods* **9**: 808–810.

Weber JN, Peterson BK, Hoekstra HE (2013). Discrete genetic modules are responsible for

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

complex burrow evolution in Peromyscus mice. *Nature* **493**: 402–5.

Wegmann D, Currat M, Excoffier L (2006). Molecular diversity after a range expansion in heterogeneous environments. *Genetics* **174**: 2009–20.

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

Weir BS, Cockerham CC (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution (N Y)* **38**: 1358–1370.

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.

Wolf JBW, Ellegren H (2016). Making sense of genomic islands of differentiation in light of speciation. *Nat Rev Genet* **18**: 87–100.

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

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.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
Arlequin	AMOVA (Analysis of Molecular Variance)	Characterizing hierarchical population structure	Arlequin allows for a variety of analyses of diversity (see below)	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)
ADMIXTURE	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and	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)

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.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

		LD between loci				
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)
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)
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)**
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)
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)
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)
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)**
npstat	Differentiation/Diversity	Extracting summary statistics from pooled data	Explicitely corrects for sampling bias in pooled data. Allows computing tests using an outgroup (MK	Mostly limited to summary statistics, but more complete than Popoolation.	https://github.com/lucaferretti/npstat	(Ferretti *et al.*, 2013)

			test, Fay and Wu's H) and characterizing coding mutations.			
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)
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)
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)**
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)
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)
POPTREE2	Genetic differentiation	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)
Kimtree	Genetic distance	Estimating divergence time between populations	The method is conditional on a prior topology provided by	Times are given in diffusion time scale, and can be converted	http://www1.montpellier.inra.fr/CBGP/software/kimtree/index.	(Gautier and Vitalis, 2013)

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

		and testing for topologies	the user. It computes DIC for a given topology, allowing to test for the best one.	in demographic times using independent estimates of Ne.	html	
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)
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)**
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)**
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)
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)
ASTRAL-2	Phylogeny	Builds species trees using short non-recombining	Coalescence-based. Suitable for short loci (e.g. RAD-seq and	More reliable under high incomplete lineage sorting that	https://github.com/smirarab/ASTRAL	(Mirarab and Warnow, 2015)

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

		sequences	GBS)	SVDQuartets and NJst (Chou *et al.* 2015)		
BEAST2	Phylogeny	Network reconstruction and phylogenetic relationships	User friendly. Can be used to track changes in effective population sizes (Bayesian Skyline Plots). Possible to estimate divergence times	Slow for large 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)
NJst (in phybase)	**Phylogeny**	**Builds species trees using short non-recombining sequences**	**Coalescence-based. Suitable for short loci (e.g. RAD-seq and GBS)**	**See ASTRAL-2 and Chou *et al.* 2015**	**https://code.google.com/archive/p/phybase/downloads**	(Liu and Yu, 2011)
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)
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)
SNAPP	Phylogeny	Phylogenetic relationships	Handles SNP data	Remains slow for medium to large datasets (>1,000SNPs)	http://beast2.org/snapp/	(Bryant *et al.*, 2012)
SNPhylo	Phylogeny	Network reconstruction and phylogenetic relationships	Complete pipeline from SNP filtering to tree reconstruction	Should be used on complex of species or divergent populations with little migration	http://chibba.pgml.uga.edu/snphylo/	(Lee *et al.*, 2014)
SVDQuartets	Phylogeny	Builds species trees using short non-recombining sequences	Coalescence-based. Suitable for short loci (e.g. RAD-seq and GBS)	See ASTRAL-2 and Chou *et al.* 2015	http://www.stat.osu.edu/~lkubatko/software/SVDquartets/	(Chifman and Kubatko, 2014)
*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)

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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)
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)
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)
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)

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
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)**

Table 2. Summary of methods for demographic inference, simulations and scenarios comparisons. Methods available in R are highlighted in bold.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABCToolbox	ABC	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)
DIYABC	ABC	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)
PopSizeABC	ABC	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)
coala	**ABC/coalescent simulations**	**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)**
fastsimcoal2	ABC/Likelihood	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)
ARGWeaver	Ancestral Recombination Graphs/coalescence	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)
G-PhoCS	Bayesian	Estimating population divergence and migration parameters using a coalescent framework	Bayesian + MCMC, handles ancient samples	Parameters scaled by mutation rate, no admixture	http://compgen.cshl.edu/GPhoCS/	(Gronau *et al.*, 2011)
IMa2	Bayesian	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)

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Migrate-n	Bayesian	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)
ABLE	Coalescence/Composite Likelihood	Model comparison and parameters estimation	Uses both allele frequency spectrum and linkage disequilibrium within blocks of a pre-specified size. Handles whole-genome data and RAD-seq.	Relies on ms syntax. Determining the most informative size for blocks requires performing pilot runs.	https://github.com/champost/ABLE	(Beeravolu *et al.*, 2016))
fastsimcoal2	coalescent simulations	Building any arbitrary scenario using a coalescent framework	Any arbitrary scenario can be implemented. Handles SNP, microsatellites and sequence data.	Does not handle selection. Slower than ms with no recombination, much faster with recombination (see manual)	http://cmpg.unibe.ch/software/fastsimcoal2/	(Excoffier and Foll, 2011)
ms, msms, msABC	coalescent simulations	Building any arbitrary scenario using a coalescent framework	Any arbitrary scenario can be implemented. Handles SNP, microsatellites and sequence data. msms can include selection in the model.	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)
scrm	coalescent simulations	Fast simulation of chromosome-scale sequences	Syntax similar to ms, handles any arbitrary scenario	Does not handle gene conversion and fixed number of segregating sites (unlike ms)	https://scrm.github.io/	(Staab *et al.*, 2015)
∂a∂i	Diffusion approximation of the AFS	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)
DoRIS	IBD tract	Testing various demographic scenario	Uses variation in IBD tracts length to test for various demographic models.	IBD must be inferred first with, e.g., BEAGLE. Handles a limited set of demographic scenarios. Modification in the code is required for more complex scenarios	https://github.com/pierpal/DoRIS	(Palamara and Pe'er, 2013)

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Unnamed	Identity by state tract	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)
diCal2	Sequentially Markovian coalescent	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)
MSMC	Sequentially Markovian coalescent	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)
PSMC	Sequentially Markovian coalescent	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)
SMC++	Sequentially Markovian coalescent	Inferring change in Ne with time and splitting time between two populations	Can analyze hundreds of individuals at a time and does not require phasing	The ancestral allele is assumed to be the reference allele by default. Assumes a clean split for populations divergence	https://github.com/popgenmethods/smcpp	(Terhorst *et al.*, 2016)

Table 3. Summary of common methods for identifying loci under selection. Methods available in R are highlighted in bold.

Software	**Class of method**	**Purpose**	**Specifics**	**Issues and warnings**	**Link**	**Reference**

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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)
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)
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)**
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)
Trinculo	Association	Detecting association with environmental/phenotypical features	Specifically designed to handle categorical variables with more than 2 categories. Performs multinomial logistic regression and provides frequentist and bayesian frameworks.	Requires lapack library in Unix. Allows fine-mapping by testing for corrrelations between adjacent markers.	https://sourceforge.net/projects/trinculo/	(Jostins and McVean, 2016)
SAMBADA	Association/Environmental association	Detecting association with environmental/phenotypical features	Designed to be fast, underlying models have been kept simple. Allows conversion from PLINK format. Takes into account spatial autocorrelation of individual genotypes. Allows correction for population structure	Does not work with pooled data. Possibly high levels of false positives. Relatedness between samples should be assessed independently. Should be used in combination with LFMM or BayPass.	http://lasig.epfl.ch/sambada	(Stucki *et al.*, 2016)
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)
msms	Coalescence	Simulate demographic scenarios including selection	Flexible, syntax similar to ms, handles arbitratily complex models. Can be used in an ABC	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)

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

			framework to include selection as a parameter to be estimated			
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/
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)
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, Huang, *et al.*, 2014)
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)**
H12 test	LD	Detecting selection using signatures of high LD	Does not require phased data. Designed for detecting soft sweeps	Coalescent simulations are recommended to evaluate the likelihood of selection	https://github.com/ngarud/SelectionHapStats/	(Garud *et al.*, 2015)
LDna	**LD**	**Detecting selection using signatures of high LD**	**Can be used to address population structure or detect large inversions or indel polymorphism through LD**	**The user needs to play with parameters to ensure robustness of SNPs significantly linked**	**https://github.com/petrikemppainen/LDna**	**(Kemppainen *et al.*, 2015)**
rehh	**LD**	**Detecting selection using signatures of high LD**	**Can compute both XP-EHH and Rsb. Handles several input formats**	**Requires phased data and high density of markers**	**https://cran.r-project.org/web/packages/rehh/index.html**	**(Gautier and Vitalis, 2012)**
Selscan	LD	Detecting selection using signatures of high LD	Includes the nSL statistics dedicated to soft sweep detection	Does not include utilities to specify the ancestral state of alleles. Requires phased data and high density of markers	https://github.com/szpiech/selscan	(Szpiech and Hernandez, 2014)
BALLET	Likelihood test for balancing selection	Detecting balancing selection	Designed for detecting ancient balancing selection. Does not require phasing	Requires whole-genome data and recombination map. The ancestral state of alleles must be obtained through an outgroup	http://www.personal.psu.edu/mxd60/ballet.html	(DeGiorgio *et al.*, 2014)

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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)**
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)**
SelEstim	Population differentiation	Detecting positive selection and local adaptation	Can estimate the coefficients of selection. Calibration using a pseudo-observed dataset to obtain (can be used in combination with the R function simulate.baypass() in BayPass).	Assumes an island model.	http://www1.montpellier.inra.fr/CBGP/software/selestim/	(Vitalis *et al.*, 2014)
Bayenv, BayPass	Population differentiation/Association	Detecting positive selection and adaptation to environmental features	Less sensitive to population demographic history than previous methods. 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)
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)**
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)
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)**

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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)
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)

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/052761; this version posted June 18, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

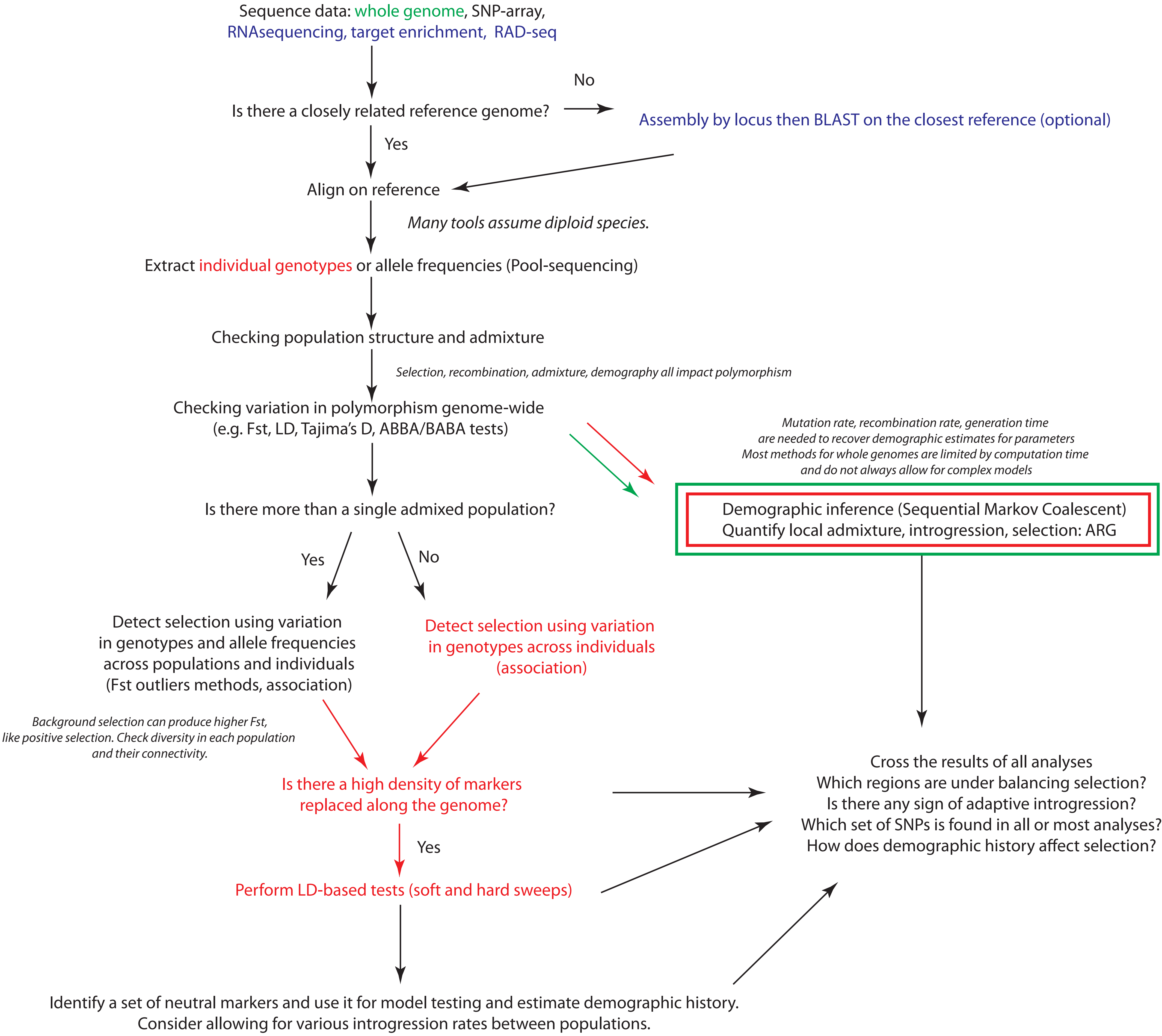
Figures

Figure 1. A possible general 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. ARG: Ancestral Recombination Graph (see main text).

Figure 2. Set of questions and relevant methods to characterize population structure and local adaptation. Proposed methods mostly use common data formats for input files, facilitating their integration in a single pipeline. PGDSpider (Lischer and Excoffier, 2012) can be used to automate file conversion for methods requiring private input format. The proposed methods are not exhaustive, see tables for a more detailed list.

Figure 3: Set of questions and relevant methods to characterize demography and selection. *: requires reference genome; ** requires reference genome and whole genome resequencing.

certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Sequence data: whole genome, SNP-array, RNAsequencing, target enrichment, RAD-seq

Is there a closely related reference genome?

No

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

Selection, recombination, admixture, demography all impact polymorphism

Checking variation in polymorphism genome-wide
(e.g. Fst, LD, Tajima's D, 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 (Sequential Markov Coalescent)
Quantify local admixture, introgression, selection: ARG

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)

Background selection can produce higher Fst,
like positive selection. Check diversity in each population
and their connectivity.

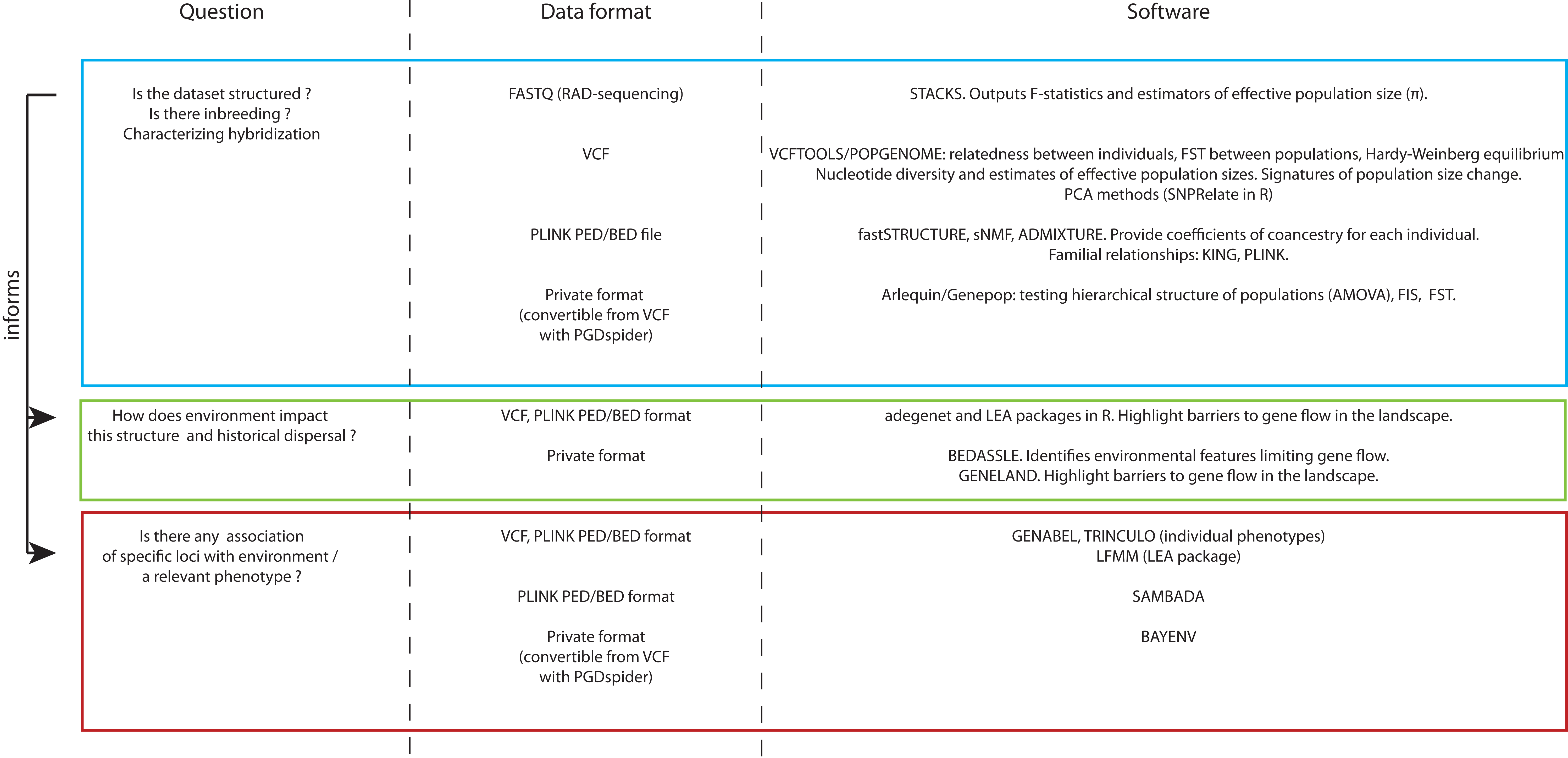
Is there a high density of markers replaced along the genome?

Yes

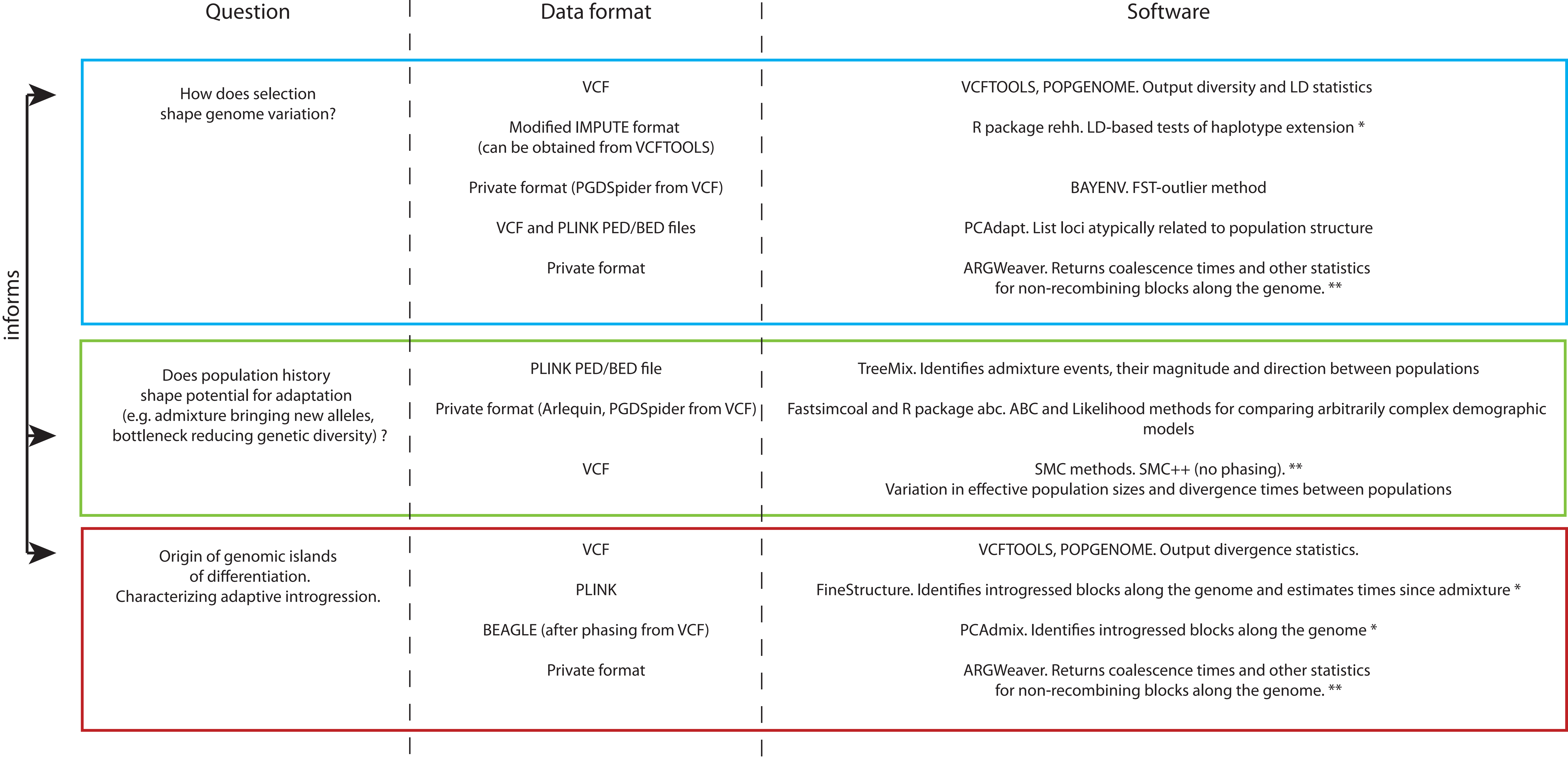
Perform LD-based tests (soft and hard sweeps)

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.



Question	Data format	Software
Is the dataset structured ? Is there inbreeding ? Characterizing hybridization	FASTQ (RAD-sequencing)	STACKS. Outputs F-statistics and estimators of effective population size (π).
	VCF	VCFTOOLS/POPGENOME: relatedness between individuals, FST between populations, Hardy-Weinberg equilibrium Nucleotide diversity and estimates of effective population sizes. Signatures of population size change. PCA methods (SNPRelate in R)
	PLINK PED/BED file	fastSTRUCTURE, sNMF, ADMIXTURE. Provide coefficients of coancestry for each individual. Familial relationships: KING, PLINK.
	Private format (convertible from VCF with PGDspider)	Arlequin/Genepop: testing hierarchical structure of populations (AMOVA), FIS, FST.
How does environment impact this structure and historical dispersal ?	VCF, PLINK PED/BED format	adegenet and LEA packages in R. Highlight barriers to gene flow in the landscape.
	Private format	BEDASSLE. Identifies environmental features limiting gene flow. GENELAND. Highlight barriers to gene flow in the landscape.
Is there any association of specific loci with environment / a relevant phenotype ?	VCF, PLINK PED/BED format	GENABEL, TRINCULO (individual phenotypes) LFMM (LEA package)
	PLINK PED/BED format	SAMBADA
	Private format (convertible from VCF with PGDspider)	BAYENV



Question	Data format	Software
How does selection shape genome variation?	VCF	VCFTOOLS, POPGENOME. Output diversity and LD statistics
	Modified IMPUTE format (can be obtained from VCFTOOLS)	R package rehh. LD-based tests of haplotype extension *
	Private format (PGDSpider from VCF)	BAYENV. FST-outlier method
	VCF and PLINK PED/BED files	PCAdapt. List loci atypically related to population structure
	Private format	ARGWeaver. Returns coalescence times and other statistics for non-recombining blocks along the genome. **
Does population history shape potential for adaptation (e.g. admixture bringing new alleles, bottleneck reducing genetic diversity) ?	PLINK PED/BED file	TreeMix. Identifies admixture events, their magnitude and direction between populations
	Private format (Arlequin, PGDSpider from VCF)	Fastsimcoal and R package abc. ABC and Likelihood methods for comparing arbitrarily complex demographic models
	VCF	SMC methods. SMC++ (no phasing). ** Variation in effective population sizes and divergence times between populations
Origin of genomic islands of differentiation. Characterizing adaptive introgression.	VCF	VCFTOOLS, POPGENOME. Output divergence statistics.
	PLINK	FineStructure. Identifies introgressed blocks along the genome and estimates times since admixture *
	BEAGLE (after phasing from VCF)	PCAdmix. Identifies introgressed blocks along the genome *
	Private format	ARGWeaver. Returns coalescence times and other statistics for non-recombining blocks along the genome. **