Going down the rabbit hole: a review on how to link genome-wide data with ecology and evolution in natural populations

Yann X.C. Bourgeois

Abstract

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

Introduction

Genetic makeup of populations is shaped by multiple historical and selective factors. The advent of Next-Generation Sequencing (NGS) in the last 20 years has enhanced our understanding on how intermingled these factors are, and how they can impact genomic variation. Important results have been gathered on model species, or species with an economical interest. Such results include, among other examples, an improved perspective on human history of migrations, admixture and adaptation (e.g. Sabeti *et al.*, 2002; Abi-Rached *et al.*, 2011; Li and Durbin, 2011), elucidating the origin of domesticated species (e.g. Axelsson *et al.*, 2013; Schubert *et al.*, 2014), or characterizing the genetic bases of local adaptation in model or near-model species (e.g. Legrand *et al.*, 2009; Kolaczkowski *et al.*, 2011; Roux *et al.*, 2013; Kubota *et al.*, 2015). These studies have brought insights at an unprecedented scale on the links between genotype, phenotype and environment. Most of these studies relied on a precise knowledge of both population history and patterns of selection, together with functional validation of variants associated to selected phenotypes.

Translation of these methods into non-model species is part of a shift in evolutionary sciences that aims at better understanding biological diversity at various scales (Mandoli and Olmstead, 2000; Jenner and Wills, 2007; Abzhanov *et al.*, 2008). Recent breakthroughs brought by the study of initially non-model species (e.g. White *et al.*, 2010; Ellegren *et al.*, 2012; Weber *et al.*, 2013; Poelstra *et al.*, 2014) have confirmed the value of population genomics from this perspective. These advances are needed to broaden our view about the evolutionary process and improve sampling of distant clades. Ultimately, this process should provide a more balanced picture than the one brought by the study of a few model species (Abzhanov *et al.*, 2008). Genomic approaches also have the potential to improve conservation genetic inference by scaling up the amount of data available (Shafer *et al.*, 2015).

However, the widespread use of sophisticated analytical tools remains challenged by the lack of communication between fields (Shafer *et al.*, 2015), little user-friendliness of software and the ever-increasing amount of tools made available. Much effort has been put recently in addressing these issues, but a lack of clarity subsists and many uncertainties remain. The application of sometimes complex methods to species with little background has nonetheless become more accessible, and has the potential to bring valuable information.

In this paper, I propose various methods and suggestions to deal with usual questions in population genomics and genetics of adaptation in natural populations. I begin with a succinct review of methods available to obtain genome-wide polymorphism data before focusing on i) methods devoted to the study of population demographic history (Figure 1) and ii) methods aiming at detecting signatures of selection (Figure 2).

Glossary

SNP: single nucleotide polymorphism.

Variant calling: identifying confidently genomic variants from alignment data (in SAM/BAM format, see Li *et al.*, 2009). Classical SNP callers include the Genome Analysis Toolkit or GATK (McKenna *et al.*, 2010), freebayes (Garrison and Marth, 2012), samtools (Li *et al.*, 2009) or Platypus (Rimmer *et al.*, 2014). Other tools call large-scale variants such as inversions, translocations or copy-number variation (see main text).

Phasing: a process which identifies the alleles that are co-located on the same chromosome copy.

Pooled sequencing: a protocol where tens or hundreds samples are pooled in a single library prior sequencing (Futschik and Schlötterer, 2010). This prevents any individual identification of each sample.

Obtaining genetic markers and linking them to a genome

Common sequencing methods

I consider here two main ways of dealing with genomics in non-model species: reduced representation (Davey *et al.*, 2011) and whole-genome resequencing. Reduced representation allows sampling homogeneously variants across the genome by sequencing DNA fragments flanking restriction sites. Some of the best-known reduced representation techniques include RAD-sequencing (Baird *et al.*, 2008) and Genotyping by Sequencing or GBS (Elshire *et al.*, 2011). Their main interest is their relatively low cost and that they do not require any reference genome (see Davey *et al.*, 2011 for details). The amount of SNPs ranges from thousands to millions, which is most of the time enough to retrieve substantial information about demography and sometimes selection (see Puritz *et al.*, 2014 for a detailed summary of reduced-representation techniques).

Whole-genome resequencing requires a reference (at least at a draft stage) and is much more expensive, especially for species with long and complex genomes. However, this approach gives a complete overview of structural and coding variation, and allows some of the most powerful methods currently available to track signatures of selection (see below). Pooled sequencing (Futschik and Schlötterer, 2010) can be an option to reduce the costs, but restricts the analysis to methods focusing on allele frequencies, losing most of the information provided by variation in Linkage Disequilibrium (LD).

Shallow sequencing (1-5X per individual) may be a way to partly overpass this last issue for a similar cost (Buerkle and Gompert, 2013), but should not be used for methods requiring phasing and unbiased individual genotypes. Shallow shotgun sequencing also allows retrieving complete plastomes, due to the representation bias of mitochondrial or chloroplast sequences. Plastome sequences can provide insightful information about the evolutionary history of populations or species. Recent work has successfully used shallow sequencing to reconstruct mitochondrial or chloroplast sequences in plants (Malé *et al.*, 2014), animals (Hahn *et al.*, 2013) or old and altered museum samples (Besnard *et al.*, 2016). Methods such as MITObim (Hahn *et al.*, 2013) provide an automated and relatively user-friendly way to reconstitute plastome sequences, which can then be analyzed as a single non-recombining marker for phylogeny or population genetics.

Obtain positional information for markers

Whole-genome resequencing requires at least a draft genome, and reduced representations methods can also benefit from a reference, either to order markers or retrieve information about the nearest gene of a focal SNP. Methods inferring selection from haplotype extension and patterns of LD (described further below) require that the relative order of markers on genome sequence is known. A reference also allows analyzing separately sex chromosomes (that can be haploid) and autosomes to correct for variation in ploidy between males and females in gonochoric organisms. Obtaining a draft reference from deep Illumina sequencing is now relatively common, but requires a good knowledge of assembly methods to choose the tool adapted to the focal species. Initiatives such as Assemblathon (Bradnam *et al.*, 2013) have provided valuable insights and advices from this regard. Once a draft is produced, annotation of features is recommended since it allows linking variation at a locus to its putative function. This requires either RNA-seq data to be mapped back on the reference or at least that an annotation from a relatively close species is available.

It is possible to avoid these steps for species having a close relative already sequenced. Short-reads alignment algorithms like BWA (Li and Durbin, 2009) generally assume relatively low divergence between reads and reference. For species having less than 3% divergence, reads may be directly mapped back onto the nearest genome. For more distantly related species, a possible strategy would be using RAD-seq or GBS, build contigs for each locus with methods like Stacks (Catchen *et al.*, 2011) or PyRAD (Eaton, 2014), and map those loci on the reference with BLAT (Kent, 2002) or LASTZ (Schwartz *et al.*, 2003). Using a related reference requires that synteny is conserved between species. While this assumption is reasonable in, e.g., birds (Derjusheva *et al.*, 2004), it becomes more doubtful in other clades, like in plants (Molinari *et al.*, 2008; Soltis *et al.*, 2015). Before conducting a NGS study, it is therefore important to know how genomes vary in their structure across related species. Some methods do not even require any reference sequence to call SNPs from raw reads, like kSNP2 (Gardner and Hall, 2013). It is however advised to cautiously filter reads prior calling, since the method does not distinguish between sequencing errors and actual variants.

Checking for the presence of large structural variants can be informative when performing whole-genome resequencing. Structural variants include duplications and copy number variation (CNV), deletions, inversions or translocations. Neglecting this variation can lead to call spurious SNPs, for example in regions which are single copy in the reference but display CNV in some individual. This can distort estimates of nucleotide diversity or homozygosity, biasing analyses based on LD or allele frequencies. These variations can be partly masked by filtering SNPs on the basis of Hardy-Weinberg equilibrium or sequencing depth. However, more quantitative methods are available that allow to precisely characterize the nature and the position of this type of variation, like Delly (Rausch *et al.*, 2012) or Lumpy (Layer *et al.*, 2014). Regions that display changes in genomic structure can then be excluded for analyses requiring accurate estimates of diversity (e.g. Rasmussen *et al.*, 2014). On the other hand, these variations can be used for studying association with traits of interest.

Assessing population history

Exploring population structure

Checking for population structure is an essential step when performing analyses on genome-level datasets. Neglecting it can bias demographic inferences (Chikhi *et al.*, 2010; Heller *et* *al.*, 2013) or the detection of loci under selection (e.g. Nielsen *et al.*, 2007); thus, checking for outlier individuals and assessing the global structure is required prior any more sophisticated analysis. A simple approach that does not assume any *a priori* grouping is the Principal Component Analysis (PCA), based on analyzing variance-covariance structure among genotypes, which can be performed on both individual and pooled data. Methods such as SMARTPCA (Patterson *et al.*, 2006) or EIGENSTRAT (Patterson *et al.*, 2006) emerged from this framework. There are many software solutions and packages allowing to perform this type of analysis, such as SNPRelate (Zheng *et al.*, 2012), implemented in Bioconductor (Huber *et al.*, 2015), PLINK (Purcell *et al.*, 2007) or GenAbel (Aulchenko *et al.*, 2007). For large whole-genome data or high-density RAD-seq, reducing SNP redundancy by subsampling unlinked markers (having low LD or large physical distance between them) is a way to reduce computation time while keeping the relevant information.

Taking into account the relatedness of individuals is recommended, for example to evaluate the amount of inbreeding within a population. 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 them, in relation with geographical or ecological distance (e.g. Fields *et al.*, 2015). VCFTools (Danecek *et al.*, 2011) provides two ways calculating relatedness; unadjusted Ajk (Yang *et al.*, 2010) and a kinship coefficient also implemented in KING (Manichaikul *et al.*, 2010). It also allows calculating Hardy-Weinberg equilibrium. Population stratification and relatedness can also be explored in PLINK based on pairwise identity-by-state (IBS) distance or identity by descent (IBD).

Other approaches such as Structure (Pritchard *et al.*, 2000) and fastSTRUCTURE (Raj *et al.*, 2014) allow determining hierarchical population structure by grouping individuals in clusters without any *a priori*. FastSTRUCTURE is computationally faster and more efficient with large SNP datasets. These methods are also more efficient at detecting signatures of admixture. Geneland (Guillot *et al.*, 2012), available as a R package, allows determining the optimal number of population in a dataset by optimizing linkage and Hardy-Weinberg equilibrium within clusters, and is also able to incorporate geographic coordinates in the model to delineate their spatial organization. It can be useful to characterize the location and shape of hybrid zones.

In order to properly test for the existence of hierarchical population structure, methods based on differentiation measures (like F_{st}) can be used to build phylogenetic trees. POPTREE (Takezaki *et al.*, 2010) allows to use various differentiation metrics to infer relationships between populations. TreeMix (Pickrell and Pritchard, 2012) is a method building a population tree based on the covariance matrix of population allele frequencies. It allows tracking admixture events but requires the populations to be defined *a priori* (e.g. by a Structure analysis). Other methods can use individual SNP data to reconstruct phylogenies, like PhyML (Guindon *et al.*, 2010) or RAxML (Stamatakis, 2014). Splitstree (Huson and Bryant, 2006) is a user-friendly software to compute phylogenies and networks on SNP datasets and incorporate various methods for phylogeny reconstruction. Other pipelines, like SNPhylo (Lee *et al.*, 2014), propose a complete framework from SNP filtering to tree reconstruction that might help obtaining reliable topologies.

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 (Leache *et al.*, 2015). For this purpose, it might therefore be easier to extract genes or RAD contigs from the data and analyze them as DNA sequences in a software like BEAST2 (Drummond and Rambaut, 2007). In RAxML, a recent correction for bias on branch length has been implemented that requires the number of monomorphic sites to be known (Leache *et al.*, 2015) when providing only SNP alignment. Dating species or population divergence and changes in population sizes using SNP data is also possible in SNAPP (Bryant *et al.*, 2012), although the method requires long computing times when many markers are included. For dating purpose and resolution of individual and population/species trees, BEAST2 and BEAST* can also be used on sequence data for moderate-sized datasets (Drummond and Rambaut, 2007).

As a general word of caution, it is important to remind that RAD-sequencing and related methods display specific properties that can bias genome-wide estimates of diversity, like allelic dropout (Arnold *et al.*, 2013). However, this type of markers remains valuable for phylogenetic estimation, even for distantly related species (Cariou *et al.*, 2013).

To assess how diversity is partitioned across the different groups inferred by the methods described previously, it is advisable to perform an Analysis of Molecular Variance (AMOVA). Arlequin (Excoffier and Lischer, 2010) is particularly suited for this task. More generally, investigating patterns of nucleotide diversity, inbreeding, F_{st} or variation in LD between populations and across the genome is useful to have a preliminary idea of the amount of gene flow, admixture and variation in population sizes. These statistics can be easily retrieved with VCFTools or PopGenome (Pfeifer *et al.*, 2014).

Investigating population history with coalescent methods

The coalescent has first emerged to provide population geneticists a way of modeling alleles genealogy from a sample taken from a large population. Going backward in time, alleles merge (coalesce) in a stochastic way until reaching their most recent common ancestor (Kingman, 1982). A variety of methods used and enriched this theoretical framework to resolve complex population histories and their associated demographic parameters, such as divergence times, effective population sizes or gene flow. These parameters are usually scaled by mutation rate per generation. Converting those parameters into demographic estimates (e.g. time in years) requires that mutation rate and generation time be known or at least reasonably well estimated, for example from other close species with similar life history. Most well-known coalescent-based tools dedicated to population genetics include IMa (Hey and Nielsen, 2007), Migrate-n (Beerli and Palczewski, 2010) or Lamarc (Kuhner, 2009). Lamarc is the only one taking into account recombination in the model, the other ones requiring non-recombining blocks of sequence or markers to be used. Although they are powerful, these methods tend to be computationally slow (Excoffier *et al.*, 2013), since they require a full evaluation of the likelihood function associated to the model, a procedure that can be complex with hundreds or thousands of markers.

A way to bypass this issue has been the use of Approximate Bayesian Computation (ABC) methods, which compare to the actual data a set of simulated data produced by coalescent simulations under predefined scenarios. By measuring the distance between carefully chosen summary statistics describing each simulation with those from the observed dataset, it is possible to infer which scenario explains the data the best. DIYABC (Cornuet *et al.*, 2008) is a popular and user-friendly software allowing to perform a full ABC analysis (from simulations to model comparison), although it does not allow yet to model continuous gene flow between populations. Another approach, which provides more control to the user, consists in using coalescent simulators such as ms (Hudson, 2002) or fastsimcoal2 (Excoffier and Foll, 2011). A pipeline allowing to perform all these steps is also available in ABCToolbox (Wegmann *et al.*, 2010). Fastsimcoal is a bit slower to simulate data, but is more user-friendly than ms, and more effective when simulating recombination for sequence data. Once simulations are done, one can compute summary statistics for the simulated datasets (e.g. with Arlequin when using fastsimcoal2), then use packages like abc in R (Csilléry *et al.*, 2012) to perform model choice, cross-validation, estimate model misclassification and demographic parameters. More information on how to perform a proper

ABC analysis can be found in the work by Csilléry *et al.* (2010). The main advantage of ABC is that it allows handling arbitrarily complex models, unlike methods like IMa where the model is predefined. However, using summary statistics leads to the loss of potentially useful information.

More recently, new methods based on the allele frequency spectrum (AFS) emerged to facilitate and speed up the analysis of large SNP datasets. Different patterns of gene flow and demographic events all shape the AFS in specific ways (e.g. more alleles are likely to be found at similar frequencies in two recently diverged or highly connected populations). ∂a∂i (Gutenkunst *et al.*, 2009) does not rely on computationally intensive coalescent simulations but rather on a diffusion approximation of alleles, and computes likelihoods for the alternative models provided by the user. However, its current implementation does not handle more than three populations. More recently, another likelihood-based approach has been implemented in fastsimcoal2 (Excoffier *et al.*, 2013), that uses coalescent simulations and handles arbitrarily complex scenarios while not being limited by the number of populations included. These two methods assume that SNPs are under linkage equilibrium. Including SNPs in strong LD should not particularly bias model comparison, but can be an issue when estimating parameters (see fastsimcoal manual for more details). Note that the AFS can also be used as a set of summary statistics for ABC inference. Using allele frequencies estimated from pooled datasets should be feasible, although no study explored this possibility to my knowledge.

One drawback when using SNP data without considering monomorphic sites is that the mutation rate per generation is not directly taken into account. For example, in DIYABC, it does not matter when a mutation appears in the simulated genealogy, as long as it happens only once before coalescence, a reasonable assumption for SNP markers. However, this prevents any conversion of parameters into demographic estimates by using mutation rate. Again, it is also possible to extract the complete DNA sequence for a set of randomly selected markers and perform analyses on this dataset including monomorphic sites. Another possibility consists in a calibration of parameter estimates by including in the analysis a fixed parameter, such as population size or divergence time. This approach is also feasible when estimating parameters from the allele frequency spectrum, like in ∂a∂i or fastsimcoal2.

When whole genome data are available, it is then possible to use methods such as those based on Pairwise Sequentially Markovian Coalescent (PSMC), that require only a single diploid genome (Li and Durbin, 2011). This method allows tracking changes in population size across discrete time intervals. While powerful, PSMC is sensitive to confounding factors such as population structure (Orozco-terWengel, 2016) that leads to false signatures of expansion or bottleneck. It also does not allow studying recent demographic events. This is due to the fact that coalescence events for only two alleles from a single individual in the recent past are infrequent. However, extensions of the model allowing for several genomes have been developed to precise population history in the recent past, like MSMC (Schiffels and Durbin, 2014) or diCal (Sheehan *et al.*, 2013). As these methods require that heterozygous positions be properly called, it is required to correct for low depth of coverage (less than 8-10X) if needed. Recently an ABC framework, implemented in PopSizeABC, has been proposed to infer demographic variation from single genomes (Boistard *et al.*, 2016). The summary statistics used describe variation in LD and the AFS, while being robust to sequencing errors. This last method does not require phasing, which should limit the impact of phasing errors.

A recent extension of these methods takes into account population structure and aims at identifying the number of islands contributing to a single genome, assuming it is sampled from a Wright n-island meta-population (Mazet *et al.*, 2015). Such developments should help increasing the amount of information retrieved from only a few genomes. However, it is essential to keep in mind that natural populations are structured and connected in complex ways, which can bias demographic inferences, even for popular markers such as mitochondrial sequences (Heller *et al.*, 2013).

Reaching a high level of precision in demographic parameters estimation can be challenging when perspective is lacking about the evolutionary history of the species considered. At larger time-scales, the lack of fossil record can make difficult the calibration of molecular clocks. Thus, for some species, only qualitative interpretation will be possible.

Screening for selection and association

Selection and its impact on sequence variation

The impact of selection on genetic variation has been extensively studied, but still remains a central topic in evolutionary biology. Here I describe some features that are associated to different types of selection.

Selection acts both on correlations i) between alleles and environment at selected loci and ii) between alleles from different loci, either directly under selection or not. This is reflected respectively by i) variation in polymorphism within and between populations and ii) linkage disequilibrium between loci (Figure 2). A new mutation will see its frequency increase in a population where it provides a selective advantage (hard sweep). When such an allele arose recently, a large region around it can remain uniform, especially if selection is strong. As the allele rises quickly in frequency, it has too little time to recombine with other ancestral variants. This leads to an increase of linkage disequilibrium between variants associated to the advantageous mutation, as well as a decrease in nucleotide diversity around the selected locus. If selection occurs in one population but not others, it may be possible to observe a local increase in differentiation, like higher F_{st} values. If selection acts on standing variation or recurrent mutation, signature of selection can be less clear as several haplotypes surround the mutation under positive selection (see however Messer and Petrov, 2013; Jensen, 2014 for a discussion about the relative importance of soft selective sweeps).

Another type of selection is balancing selection, an umbrella term grouping all selective processes that lead to the maintenance of genetic polymorphism at a locus and to an excess of common alleles. Such processes include divergent selection (the same allele is under positive selection in one population and selected against in another one), negative frequency-dependent selection (a rare allele is preferably selected) or heterozygote advantage. In the case of recent balancing selection, the signature of selection is similar to a partial selective sweep, with the recently selected allele displaying reduced diversity and higher LD than the ancestral one. In the case of long-term balancing selection, there is an accumulation of genetic polymorphism around the selected loci, leading to the maintenance of haplotypes older and more diverse than in the rest of the genome. This increase in diversity can be associated to higher local estimates of effective population sizes and effective recombination rates. As alleles are older, coalescence times tend to be higher and can sometimes predate speciation, leading to trans-species polymorphism (see Charlesworth, 2006 for a detailed review). In some cases, an allele under balancing selection is stabilized at a single equilibrium frequency across populations, which can lead to a signature of lower differentiation compared to genomic background. There is still a lack of methods aiming at detecting specifically balancing selection compared to positive selection and recent hard sweeps (but see Fijarczyk and Babik, 2015).

In the following parts I present tools that can be used to detect signatures of selection. The methods that these tools implement fall into three main categories (partly reviewed in Vitti *et al.*, 2013), corresponding to the signature they try to target: i) study of variation in allele frequencies and polymorphism, ii) study of variation in linkage disequilibrium and iii) reconstruction of allele genealogies using the coalescent. Most of these methods assume that markers are ordered along a genome; although they can also be used to extract individual markers under selection that can be then be aligned (except for most LD-based methods).

Methods focusing on polymorphism

While demographic forces such as drift and migration will affect the whole genome in a similar way, local effects of selection should produce discrepancies with genome-wide polymorphism (Lewontin and Krakauer, 1973). Selection affects allele frequencies and polymorphism in predictable ways at the scale of single populations. Several statistics summarize them, like π, the nucleotide diversity (Nei and Li, 1979), Tajima's D (Tajima, 1989) or Fay and Wu's H (Fay and Wu, 2000). They are sensitive to population demographic history, that they allow characterizing as summary statistics in, e.g., ABC analyses. They have nonetheless the potential to highlight genomic regions displaying clear signatures of selection, or to confirm selection at candidate genes. For example, balancing selection should lead to an excess of common polymorphisms, similar to a recent bottleneck, leading to high Tajima's D and π values. Purifying selection leads to an opposite pattern, similar to a recent population expansion, with an excess of rare variants and low diversity. More sophisticated methods using allele frequency spectrum have been developed to detect positive selection, such as the recent improvement of the composite likelihood ratio (CLR) test (Nielsen *et al.*, 2005) performed in SweepFinder2 (Degiorgio *et al.*, 2016).

PopGenome (Pfeifer *et al.*, 2014) is a powerful R package that allows calculating AFS statistics (including the CLR test) across many genomes, as well as a variety of statistics on linkage disequilibrium and diversity. It also allows performing coalescent simulations to contrast observed polymorphism to neutral expectations. It is probably one of the most comprehensive tools to perform genome-wide analyses. Other possibilities include VCFTools, POPBAM (Garrigan, 2013) or Biopython libraries. For pooled data, Popoolation (Kofler, Orozco-terWengel, *et al.*, 2011; Kofler, Pandey, *et al.*, 2011) provides ways to calculate Tajima's D and nucleotide diversity, as well as measures of differentiation between populations.

Understanding the origin of genomic regions under selection highlights the evolutionary history of adaptive alleles (e.g. Abi-Rached *et al.*, 2011). Advantageous alleles can migrate from one population to another, or resist introgression from other populations (genomic islands of speciation/adaptation). The relative importance of these islands resisting gene flow after secondary contact has been recently discussed (Cruickshank and Hahn, 2014). Methods aiming at characterizing heterogeneity in introgression rates are in this context useful and can also refine the demographic history. A recent ABC framework has been developed to characterize this heterogeneity (Roux *et al.*, 2014). Methods such as PCAdmix (Brisbin *et al.*, 2012) can also be used to estimate the relative contributions of putative sources to a given sink population across the genome. A common test for introgression, available in PopGenome, is the ABBA-BABA test, summarized by Patterson's D (Durand *et al.*, 2011). Another possibility lies in the comparison of absolute and relative measures of divergence (Cruickshank and Hahn, 2014), such as d_{xy} and F_{st}, which can be calculated in PopGenome. Absolute measures of divergence are correlated to the time since coalescence. In the case of local introgression, both statistics should be reduced. For balancing selection, the decline in F_{st} is due to an excess of shared ancestral alleles, which should not impact d_{xy}, or should even make it higher than genomic background. However, these methods do not prevent false positives and results should (as usual) be interpreted with caution (Martin *et al.*, 2015).

When an allele is under positive selection in a population, its frequency tends to rise until fixation, unless gene flow from other populations or strong drift prevents it. It is therefore possible to contrast patterns of differentiation between populations adapted to their local environment to detect loci under divergent selection (e.g. displaying a high F_{st}). However, it is essential to control for population structure, as it may strongly affect the distribution of differentiation measures and produce high rates of false positive. First attempts to take into account population structure and variation in gene flow included FDIST2 (Beaumont and Nichols, 1996), which modeled populations as islands and aimed at detecting loci under selection by contrasting heterozygosity to F_{st} between populations. An extension of this model, able to take into account predefined hierarchical population structure, is implemented in Arlequin. More sophisticated methods are now available, dedicated to the detection of outliers in large genomic datasets. Most of them correct for relatedness across samples, and are reviewed extensively in the work by Francois *et al.*(2015). Some methods, like LFMM (Frichot *et al.*, 2013), aim at detecting variants correlated to environmental factors. Association methods may help targeting variants undergoing soft sweeps, weak selection or involved in polygenic control of traits (Pritchard *et al.*, 2010), for which signatures of selection are subtle and sometimes difficult to retrieve from allele frequencies data.

Other methods perform a "naïve scan" for outliers on the basis of differentiation, like BAYESCAN (Foll and Gaggiotti, 2008) which considers all populations to drift at different rates from a single ancestral pool. Most recent methods, like BAYENV (Günther and Coop, 2013) and its recent improvement, BAYPASS (Gautier, 2015), model demographic history by computing a kinship matrix between populations. Contrasting allele frequencies for each locus to the ones expected given this matrix allows testing deviation from neutrality. Those two last methods also include Bayesian tools to test for association with environmental features, facilitating further interpretation. BAYENV and BAYPASS also allow using pooled-sequencing data, making these methods polyvalent and possibly useful to many research teams. However, 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 formation of tension zones that may shift until they reach an environmental barrier where they can be trapped (Bierne *et al.*, 2011). Again, characterizing population history is required to conclude about the possible involvement of a genomic region in adaptation to environment. Sampling strategy must take into account the particular historical and demographic features of the species investigated to gain power (Nielsen *et al.*, 2007). The sequencing strategy has also to be carefully picked. Reduced representation methods do not cover all mutations in the genome and are thus more likely to miss those actually under selection. Special care in the choice of the restriction enzyme and determining the expected density of markers is needed to retrieve enough mutations close to genes under selection.

The methods described above focus on allele frequencies at the population scale, but do not allow characterizing properly association with a trait varying between individuals within populations (e.g. resistance to a pathogen, symbiotic association, individual size or flowering time). For this task, methods performing Genome-wide association analysis (GWAS) are better suited. Methods such as GenAbel in R (Aulchenko *et al.*, 2007) or PLINK (Purcell *et al.*, 2007) are powerful tool. Taking into account relatedness between samples and population history (e.g. using EIGENSTRAT or PC-adjustment corrections in GenAbel or stratified analyses in PLINK) is required to correct for false positives. This is especially recommended for species that undergo episodes of selfing or strong bottlenecks, for which sampling unrelated individuals may be unfeasible.

It is important to keep in mind that uncovering the genetic bases of complex, polygenic traits remains challenging, even in model species (Pritchard and Di Rienzo, 2010; Rockman, 2012).

It may be unavoidable in a first step to focus only on traits that are under a relatively simple genetic determinism. This can however lead to an overrepresentation of loci of major phenotypic effect, a fact that should be acknowledged when discussing the impact of selection on genome variation. The fact that loci of major effect are easier to target does not imply that they are the main substrate of selection (Rockman, 2012).

Detecting selection with methods focusing on LD

LD is increased and diversity is decreased in the vicinity of a selected allele, especially after recent selection. A class of methods aims at targeting those regions that display an excess of long homozygous haplotypes, such as the extended haplotype homozygosity (EHH) test (Sabeti *et al.*, 2002). It is also possible to compare haplotype extension across populations, with the XP-EHH test (McCarroll *et al.*, 2007) or Rsb (Tang *et al.*, 2007). Individuals included in the analysis should be as distantly related as possible to improve precision and avoid an excess of false positives. These approaches are more powerful with a relatively high density of markers, such as the ones obtained from whole-genome sequencing or high-density RAD-seq. They also require data to be phased in order to reconstruct haplotypes. This procedure can be performed with fastPhase (Scheet and Stephens, 2006), BEAGLE (Browning and Browning, 2011) or SHAPEIT2 (O'Connell *et al.*, 2014). The R package rehh (Gautier and Vitalis, 2012) allows calculating these statistics, as well as Sweep (http://www.broadinstitute.org/mpg/sweep/index.html). Statistics dedicated to the detection of soft sweeps are also available, like the H2/H1 statistics (Garud *et al.*, 2015), although further studies are still needed to understand to what extent hard and soft sweeps can actually be distinguished (Schrider *et al.*, 2015). This last statistics does not require data to be phased.

When the relative order of markers is not known, as it can be the case in RAD-seq studies without a reference genome, LDna (Kemppainen *et al.*, 2015) can be used to target sets of markers displaying strong linkage disequilibrium. This approach can be useful not only to detect selection but also structural variation such as large inversions.

Even hard selective sweeps can be challenging to detect with LD-based statistics (Jensen, 2014). It is advisable to combine several approaches to reach a better 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).

Detecting and characterizing selection with the coalescent

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

An advantage of full coalescent methods is that they provide a relatively complete picture of individual loci history, by modeling coalescence, recombination and taking into account variation in mutation rate. They are however computationally intensive, and thus difficult to apply to whole genomes. However, recent computational improvements make this procedure feasible, as illustrated by ARGWeaver (Rasmussen *et al.*, 2014). This method uses ancestral recombination graphs to model the genealogy of each non-recombining block in the genome. It allows extracting genealogies for these blocks and provides estimates for local recombination rate, coalescence time and local effective population size for each block. This approach is promising to characterize positive, purifying or balancing selection while taking into account variation in recombination and mutation rate. However, the high stochasticity in parameters estimation can limit resolution when targeting single genes.

Other methods use the theoretical framework of the coalescent to target sites under positive selection. A recent method (SCCT) using conditional coalescent trees (Wang *et al.*, 2014) claims to be faster and more precise in targeting selective sweeps. BALLET (DeGiorgio *et al.*, 2014) is a promising method to characterize ancient balancing selection. Most of those methods are designed for medium-to-high depth whole-genome resequencing, and require that individual genotypes be phased and well characterized.

Variants annotation

Characterizing the amount of synonymous or non-synonymous mutations is another way to detect whether a specific gene undergoes purifying or positive selection. An excess of non-synonymous mutations can signal positive or balancing selection, or a relaxation of selective constraints on a given gene. This requires that an annotated genome is available. Annotation of mutations can be done with SNPdat (Doran and Creevey, 2013), or directly in PopGenome, which can also perform at the genome scale tests of selection such as the MK test (McDonald and Kreitman, 1991). The MK test compares the amount of fixed and polymorphic mutations relative to an outgroup, according to their synonymous/non-synonymous state. 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. BLAST2GO (Conesa *et al.*, 2005) allows annotating genes by using a database of related species. It also allows performing GO enrichment analysis. These analyses must be carefully interpreted, depending on the level of divergence from the closest annotated species. It is important not to jump to the conclusion that orthologous genes must share the same function. When interpreting the link between selection and genetic variation, a careful review of literature can fruitfully complete the conclusions made using GO enrichment analyses.

Concluding remarks

In this contribution I highlighted different methods currently available to investigate how history and selection shape diversity in natural populations. It is important to understand that this dichotomy between selection and demography, while practical, remains artificial, and that the study of one benefits from studying the other. With the decreasing cost of sequencing it has been suggested that NGS should broaden quickly our perspective on complex evolutionary processes, from biogeography (Lexer *et al.*, 2013) to genetic bases of traits (Hohenlohe, 2014) or the maintenance of polymorphism (Hedrick, 2006). The study of DNA sequence variation, while already challenging by itself, needs to be combined with other disciplines such as ecology to be informative (Habel *et al.*, 2015). Although genome-scale analyses can be insightful to this regard, it is necessary to be conscious of their limits and to keep a biological perspective when interpreting their results. To do so, every analysis should always begin with a proper understanding of the methods used, to avoid using them as black boxes.

Before launching a project targeting thousands of markers in a species of interest, possibilities and limits of the chosen protocol must be evaluated. Focusing on species history will not necessarily require the same sampling strategy than focusing on local adaptation. While a small number of markers and populations may be enough to recover global structure and infer robustly demographic parameters, it will not provide enough resolution to target genes involved in local adaptation. In many cases, a preliminary study focusing on a few markers may already inform about the global species history and help to define an adapted design for NGS.

There is a need for a more collaborative and open culture in biology, allowing the free access to data and favoring good practices to allow repeatability of analyses (Nekrutenko and Taylor, 2012), although this cultural shift remains challenging (e.g. Mills *et al.*, 2015; Whitlock *et al.*, 2015). However, current challenges are not limited to data sharing, but also include dealing with the inflation of bioinformatics tools that sometimes overlap. Instead of working independently, researchers designing those tools could collaborate to propose free, robust and unified pipelines (Prins *et al.*, 2015). Such initiatives, like Galaxy (Goecks *et al.*, 2010) or Bioconductor (Huber *et al.*, 2015) are nonetheless emerging ; this should facilitate the emergence of a unified framework to limit the time dedicated to data analysis and focus on biological questions.

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 (80-)* **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.

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

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

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

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

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

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

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

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

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.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, *et al.* (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**: 10.

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.

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.

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

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.

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

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

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

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

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

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

Davey JW, Hohenlohe P a, 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.

Derjusheva S, Kurganova A, Habermann F, Gaginskaya E (2004). High chromosome conservation detected by comparative chromosome painting in chicken, pigeon and passerine birds. *Chromosome Res* **12**: 715–23.

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

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

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

Eaton DAR (2014). PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* **30**: 1844–1849.

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.

Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genet* **9**.

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

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

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

Fields PD, Reisser C, Dukic M, Haag CR, Ebert D (2015). Genes mirror geography in Daphnia magna. *Mol Ecol* **24**: 4521–4536.

Fijarczyk A, Babik W (2015). Detecting balancing selection in genomes: limits and prospects. *Mol Ecol* **24**: 3529–3545.

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.

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

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**.

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

Garrison E, Marth G (2012). Haplotype-based variant detection from short-read sequencing. *arXiv Prepr arXiv12073907*: 9.

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* **XXX**: XXXX–XXXX.

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.

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

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

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

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

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

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

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

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

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

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

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, *et al.* (2015).

Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**: 115–121.

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

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

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.

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.

Kent WJ (2002). BLAT — The BLAST -Like Alignment Tool. *Genome Res* **12**: 656–664.

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

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

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

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

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.

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

Layer RM, Chiang C, Quinlan AR, Hall IM (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**: R84.

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

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

Legrand D, Tenaillon MI, Matyot P, Gerlach J, Lachaise D, Cariou M-L (2009). Species-wide genetic variation and demographic history of Drosophila sechellia, a species lacking population structure. *Genetics* **182**: 1197–206.

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

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

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–60.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–9.

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.

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

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.

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

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

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

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

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, *et al.* (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303.

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

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

Molinari NA, Petrov DA, Price HJ, Smith JD, Gold JR, Vassiliadis C, *et al.* (2008). Synteny and Collinearity in Plant Genomes. *Science (80-)*: 486–489.

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

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

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

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

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

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

Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *PLoS Genet* **2**: 2074–2093.

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.

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 (80-)* **344**: 1410–1414.

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

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

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

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

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

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

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**.

Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: 333–339.

Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, *et al.* (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **46**: 912–918.

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.

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, Pauwels M, Ruggiero M-V, Charlesworth D, Castric V, Vekemans X (2013). Recent and ancient signature of balancing selection around the S-locus in Arabidopsis halleri and A. lyrata. *Mol Biol Evol* **30**: 435–47.

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

Scheet P, Stephens M (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**: 629–44.

Schiffels S, Durbin R (2014). Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**: 919–25.

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

200: 267–284.

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

Schwartz S, Kent W, Smit A (2003). Human–mouse alignments with BLASTZ. *Genome Res*: 103–107.

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. *Genetics* **194**: 647–662.

Soltis PS, Marchant DB, Van de Peer Y, Soltis DE (2015). Polyploidy and genome evolution in plants. *Curr Opin Genet Dev* **35**: 119–125.

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

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.

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

Wang M, Huang X, Li R, Xu H, Jin L, He Y (2014). Detecting recent positive selection with high accuracy and reliability by conditional coalescent tree. *Mol Biol Evol* **31**: 3068–3080.

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

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

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

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

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

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

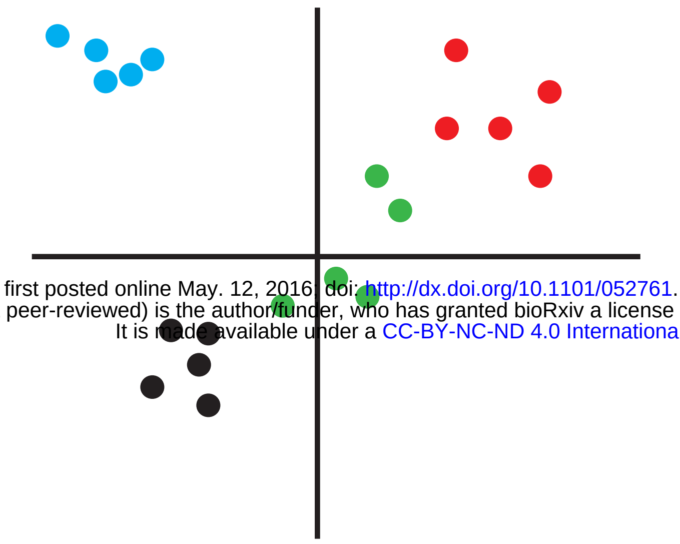
Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**: 3326–3328.

Figures

Figure 1: Summary of methods assessing the demographic history of populations (middle panel). References are provided in the main text.

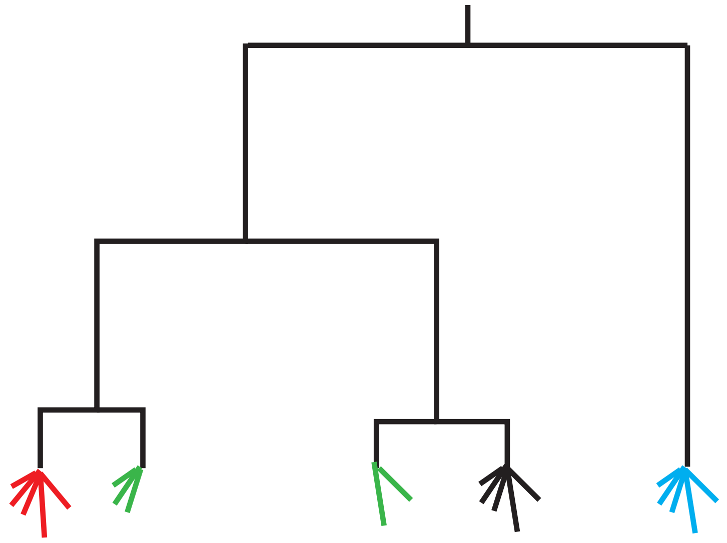
Figure 2: Summary of methods dedicated to the detection of various signatures of selection in the genome. In this simple example, the mutation on the left is under positive selection in one population (red) but not the other (black). The mutation on the right is under ancient balancing selection in the two populations.

Assessing population structure and family relationships

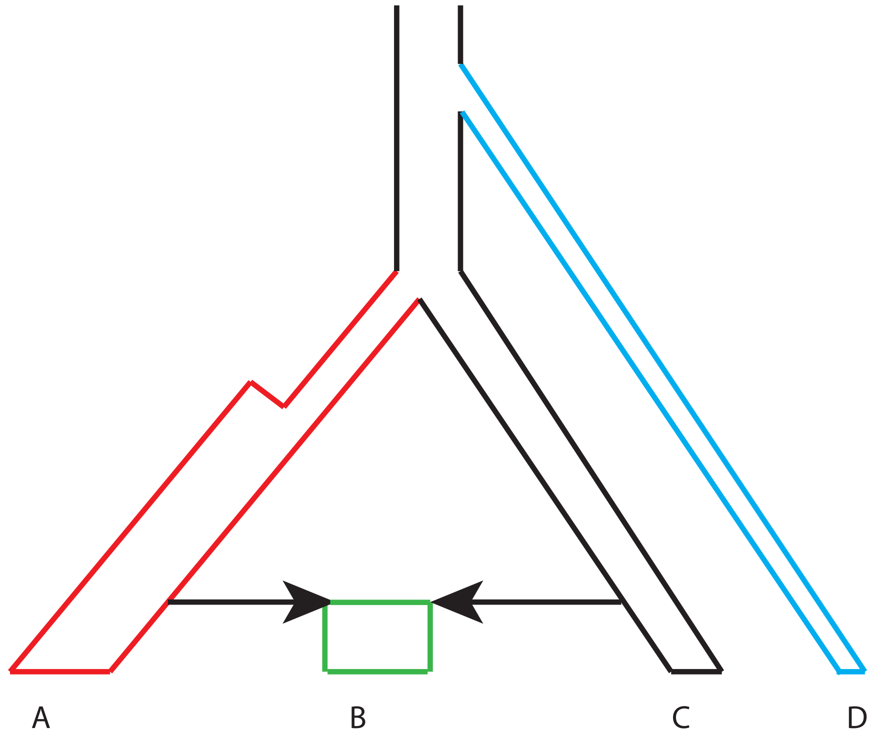


PCA (SNPRElate, GenAbel in R)
STRUCTURE and fastSTRUCTURE
GENELAND
PLINK, VCFTOOLS, KING (relatedness)
Arlequin (AMOVA)

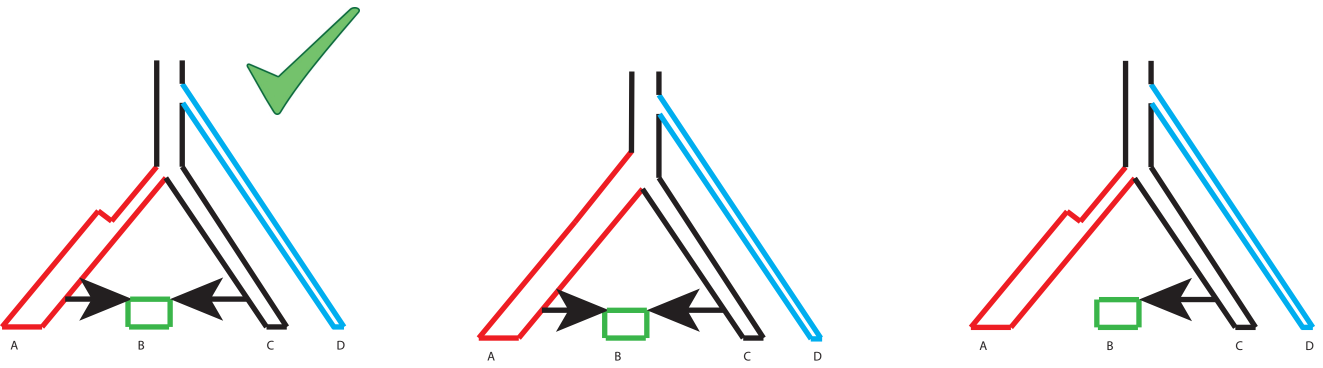
Phylogenetic relationships between individuals/populations



TREEMIX, fastSTRUCTURE (admixture)
POPTREE, PopGenome, VCFTOOLS (differentiation)
Splitstree, SNPhylo, RAxML, PhyML, BEAST2 (individual phylogeny)
POPTREE, BEAST*, SNAPP (Species/Population trees)

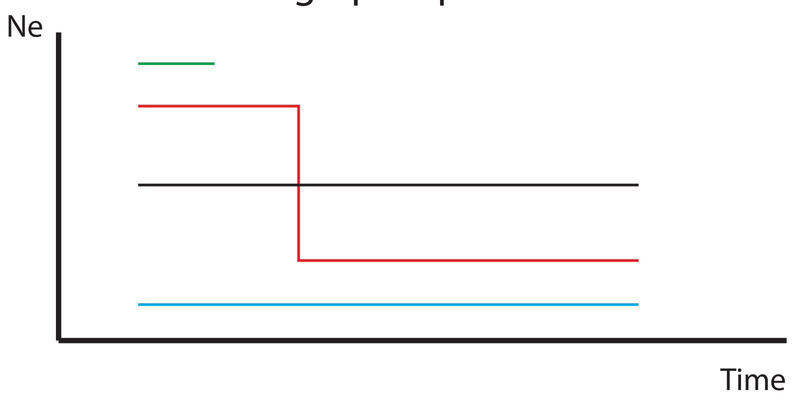


Model testing



Likelihood comparison: Migrate-n, IMa2 (for small datasets)
ABC methods: DIYABC, ms, fastsimcoal2, ABCToolbox, package abc in R.
Likelihood comparison using AFS (fastsimcoal2, dadi)

Demographic parameters inference



SNAPP, BEAST*, IMa, Migrate-n, LAMARC (for small datasets)
ABC methods: DIYABC, ms, fastsimcoal2
Inference from allele frequency spectrum (AFS): fastsimcoal2, dadi
diCal, PSMC, MSMC (for whole genome resequencing)

Positive selection
Balancing selection
Methods and tools available
π
Fst
Diversity
Differentiation/Association
Haplotype length and LD patterns
Coalescence and alleles history

PopGenome, VCFTOOLS, POPBAM

SweepFinder 2

Biopython

Popoolation (pooled data)

PopGenome, VCFTOOLS, Popoolation

Fst outlier methods: BAYPASS, BAYENV

Association with environment: BAYPASS, LFMM, BAYENV

Association with phenotype: GENABEL, PLINK

Genome-wide LD: PLINK, VCFTOOLS

EHH and Rsb tests: rehh package, Sweep

LD clusters: LDna

MSMS, PopGenome

ARGWeaver *

BALLET *

SCCT *

* Better suited for whole-genome resequencing