Using Haplotype Information for Conservation Genomics
Maeva Leitwein, Maud Duranton, Quentin Rougemont, Pierre-Alexandre Gagnaire, Louis Bernatchez

To cite this version:

HAL Id: hal-02395119
https://hal.archives-ouvertes.fr/hal-02395119
Submitted on 5 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Using Haplotype Information for Conservation Genomics

Maeva Leitwein,1,3,* Maud Duranton,2,3 Quentin Rougemont,1,3 Pierre-Alexandre Gagnaire,2,4 and Louis Bernatchez1,4

The particular combinations of alleles that define haplotypes along individual chromosomes can be determined with increasing ease and accuracy by using current sequencing technologies. Beyond allele frequencies, haplotype data collected in population samples contain information about the history of allelic associations in gene genealogies, and this is of tremendous potential for conservation genomics. We provide an overview of how haplotype information can be used to assess historical demography, gene flow, selection, and the evolutionary outcomes of hybridization across different timescales relevant to conservation issues. We address technical aspects of applying such approaches to nonmodel species. We conclude that there is much to be gained by integrating haplotype-based analyses in future conservation genomics studies.

The Potential of Haplotypes for Conservation Biology

Societal recognition of global biodiversity and the dramatic erosion caused by human activity is relatively recent [1]. The emergence of conservation biology in the early 1980s [2] has given birth to a crisis discipline that aims to propose strategies to curb biodiversity loss [3]. Conservation genetics approaches contribute to these efforts by documenting levels of genetic variation within and among populations to estimate key evolutionary parameters [4]. By contributing to a better assessment of the demography and evolutionary potential of wild populations, this field now plays a major role in species conservation and management [5].

The development of next-generation sequencing (NGS) technologies and the ensuing availability of whole-genome polymorphism data has moved the field from conservation genetics to conservation genomics [6,7]. The increased number of neutral markers has enabled a more accurate estimation of effective population size \((N_e)\) (see Glossary) and migration rate \((m)\) [6], two fundamental parameters in conservation biology. For instance, populations with a small \(N_e\) have increased homozygosity for partially recessive deleterious mutations, and are therefore more susceptible to inbreeding depression [8,9]. In addition, small populations usually accumulate more deleterious mutations when drift prevails over selection [7]. This may synergistically interact with demography to cause extinction through a mutational meltdown process [10]. Similarly, genomic data now provide more robust estimates of migration rates and intergenerational dispersal distances to address genetic connectivity [11,12]. When immigrants effectively transmit their genes following dispersal into a recipient population, the resulting gene flow may either promote or counteract local adaptation [13], increase or mask genetic load [14], erode species boundaries [15], or have potential long-term effects through adaptive or maladaptive incorporation of foreign genetic material [16].

Natural and human-induced gene flow between divergent evolutionary lineages can result in genetic admixture or introgression [17,18]. Such exchanges of foreign genetic material raise several conservation and management questions [19], especially when they occur between wild and domesticated populations [20], or between endangered and nonendangered species [21]. Population genomic studies have been addressing these issues with increasing power over the past decade [5], but did not fully exploit the information contained in linkage disequilibrium (LD) among neighboring markers [22]. Recently, however, some studies started to use microhaplotypes to increase the accuracy of individual assignment, relatedness, and population structure inference [23,24].

Highlights

Access to genome-wide genotype data has recently catalyzed new avenues of conservation genomics and biodiversity research.

However, the rich information provided by linkage disequilibrium among genotypes at linked loci remains largely underexploited in the context of conservation genomics.

Retrieving haplotype information within populations substantially improves the estimation of numerous parameters of relevance for conservation that pertain to population demography, gene flow, and selection.

Haplotype data also contribute to understanding the consequences of genetic admixture by characterizing the genomic mosaic of local ancestry. This allows dissection of variation in introgression rates across the genome, thus casting light on the evolutionary processes that shape genome-wide ancestry.

1Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, Canada
2ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France
3Co-first authors
4Co-last authors
*Correspondence: maeva.leitwein.pro@gmail.com
On a broader genomic scale, analyzing the particular combinations of alleles that define haplotypes along individual chromosomes represents an important, but untapped, source of information to decipher the complex interplay of evolutionary forces shaping genetic variation across the genome.

The signal classically obtained from allele frequencies ignores LD information (we term this the ‘vertical signal’ to reference the way in which allele frequencies are read in sequence alignments). By contrast, the signal contained in haplotype data captures the information of LD among neighboring sites along the genome (termed the ‘horizontal signal’ to illustrate how haplotypes appear in sequence alignments), which provides a better understanding of the demographic and selective processes that influence genetic diversity and population structure [19]. For example, the use of the horizontal signal contained in admixture tracts allowed Duranton et al. [25] to estimate dispersal distance in a Mediterranean population of the European sea bass (Dicentrarchus labrax) [25] (Figure 1), which could be useful for delineating Marine Protected Areas (MPAs).

Haplotype data are inherently related to the rate of recombination and its variation across the genome, a major modulator of selection efficiency [26]. For instance, tight linkage is expected to amplify the genomic footprint of linked selection [27] owing to combined effects of background selection [8] and selective sweeps [28]. As a result, linked selection tends to prevail in low-recombining regions, which reduces nucleotide diversity below the genome-wide background level through local reduction in $N_e$ [4]. In the same vein, selection acting on linked mutations affects the outcome of genetic admixture [29,30]. As a consequence, haplotype data allow a better understanding of the evolutionary mechanisms that shape the genomic mosaic of local ancestry tracts following admixture.

With the continuing development of sequencing technologies (i.e., linked-read and long-read sequencing) and analytical methods (i.e., haplotype reconstruction approaches), haplotype information is becoming accessible for nonmodel organisms, thus opening new opportunities for conservation genomic studies (Box 1). We review here how this horizontal signal contained in genomic variation has the potential to promote future advances in different contexts relevant to conservation genomics, ranging from genetic variation within single populations to genetically interacting populations or species.

Spatiotemporal Inference in Metapopulations

Using haplotype information can improve the inference of population demographic parameters including $N_e$ and $m$. Currently, the most widely used methods to infer demographic history of a population from phased whole-genome haplotype data rely on the sequential Markovian coalescent (SMC) approximation [41]. SMC approaches, that are typically implemented using a hidden Markov model, have provided invaluable insights into changes in population size through time in various taxonomic groups. For instance, using phased genome sequences, Yang et al. [42] provided evidence for continuous decline in the critically endangered population of ironwood tree (Ostrya rehderiana), accompanied by an increased number of deleterious mutations. However, some SMC-based methods make several assumptions, including the absence of population structure, migration, or admixture, which may bias inferences [43,44]. To circumvent those limitations, the most recent extensions of SMC-based methods, including the multiple sequential Markovian coalescent (MSMC), have the potential to handle larger sample sizes [30,41,42] as well as more complex demographic models (e.g., more than one population, asymmetric migration rates, variable $N_e$ along the genome, etc.) [26], without necessarily needing phased data [45].

Between-Population Demographic Inferences Using IBD and IBS Tracts

Long identity-by-descent (IBD) tracts (Box 2) can be exploited to quantify effective population size and migration rates. IBD segments incorporate LD information stemming from recently shared ancestry (Box 2). Long IBD segments can inform us about the demographic history occurring after the time of the most recent common ancestor (TMRCA) [46–48]. Consequently, recent changes in $N_e$ or migration rates are expected to affect levels of shared long IBD segments [49]. A high number of long IBD segments indicates many recent coalescent events (recent TMRCAs) and thus a small
recent \( N_e \), whereas the opposite indicates a large \( N_e \). These properties allow better estimations of recent changes in population size [50] while accounting simultaneously for group ancestry [51] (Figure 2 for an example). Furthermore, IBD methods can also fit complex demographic histories, and are thus more accurate for very recent histories (as recent as four to 10 generations ago), which makes them highly valuable for conservation genetics. For instance, recent migration among small-sized demes is expected to increase IBD sharing between demes but to decrease it within demes. In humans, rates of IBD sharing decay with increasing geographic distance between European populations, where Europeans from neighboring populations share from two to 12 common ancestors in the past 1500 years [52]. Finally, Palamara et al. [49] developed a model for inferring population size change up to 10 generations ago, and they extended their framework to accommodate multiple demes and infer recent fine-scale migration rates [53]. To date, these methods have been mostly used in human genetics studies and with a few other species such as flycatchers (Ficedula spp.) [54]. These methods can provide valuable information in a conservation genomics context via a better understanding of the recent history of population size and genetic connectivity (gene flow) among populations, with the potential to focus on particular time-periods. Ultimately, such information could help to understand which factors are the most threatening to endangered populations.

Another metric of interest is the intergenerational dispersal distance (i.e., the variance in parent–offspring distances) which can be inferred jointly with the effective population density using the slope of an isolation-by-distance model [63]. To overcome the issue of separating density and dispersal, Ringbauer et al. [64] recently developed an inference framework based on Barton et al. [65] that describes the expected number of IBD segments of a given length in a given pair of samples as a function of their distance. Using this method, the authors [64] were able to estimate a dispersal rate of approximately 50–100 km/\( \sqrt{\text{generations}} \)) in European human populations. The estimated dispersal parameter provided by this approach is of direct interest for conservation purposes [64] because it dissociates recent dispersal distances from past effective population density. Finally, unlike IBD segments, identity-by-state (IBS) tracts can be directly observed without the need to infer historical ancestry (Box 2). Therefore, IBS tracts can be easily used to infer demographic parameters. For instance, the composite likelihood framework developed by Harris and Nielsen [60] uses IBS tracts to infer temporal changes in \( N_e \), as well as divergence time and admixture.

Within-Population Demography and Inbreeding Using Runs Of Homozygosity

A run of homozygosity (ROH) corresponds to an IBD segment within a single individual that descends from shared parental ancestry (i.e., when parents carry identical sequences that coalesce to one recently shared ancestor). ROH analysis can inform us about levels of population size reduction, inbreeding, or natural selection acting on the genome. Knowledge of the distribution of both ROH number and length is informative with regards to \( N_e \), with expectations identical to those for IBD segments [50,66,67]. For instance, the abundance of ROH in different length classes was used to quantitatively compare \( N_e \) among four species of flycatcher (Ficedula spp.) in different historical time-periods [54] (Figure 2). ROH may also inform conservation geneticists about inbreeding, which can decrease fitness because of unmasking of partially recessive deleterious alleles [9]. For instance, inbreeding estimates were recently obtained from ROH in an endangered population of gray wolf (Canis lupus) [68]. In particular, the authors were able to finely characterize ROH on nearly completely homozygous chromosomes, and they showed that the majority of ROH stem from common ancestors that were shared less than 10 generations ago. A particularly important feature of IBD and ROH is that they can be inferred without haplotype phasing (although better estimates of IBD block will be obtained if accurate phasing is available) making these approaches particularly attractive for nonmodel species in a conservation context [31].

Interactions between Differentiated Genomes

Over the years, many different analytical approaches have been developed to estimate the timing and magnitude of gene flow [69,70] or admixture proportions in wild individuals (e.g., [71]). Nevertheless, new methods considering linkage information, in addition to allele frequencies at independent loci, have only recently started to emerge. Despite their potentially widespread benefits, these
methods have been mostly used to study human populations. Hybridization between species or divergent populations generally leads to the introgression of migrant chromosomes within a recipient genetic background. Such migrant tracts will subsequently be shortened at each generation of backcrossing by recombination, and long migrant tracts are therefore expected to have introgressed more recently than short tracts [72]. Admixed individual genomes can be represented as a mosaic of local ancestry tracts originating from two (or more) differentiated populations or species [73], and these can be dissected using linkage information [74]. Many different methods have been developed to infer the ancestry of local tracts along individual genomes using different types of data (Box 3).

Once revealed, this mosaic of introgressed tracts carries much information pertaining to the timing, magnitude, and variation of gene flow along the genome [74]. Based on this principle, Leitwein et al. [86] introduced a metric that captures the unevenness of ancestry proportions between chromosome homologs, the chromosomal ancestry imbalance (CAI) metric, which can be used to distinguish between early- and late-generation hybrids. This metric revealed a multiple-way admixture (i.e., admixture among more than two populations) between wild populations of brown trout (Salmo trutta) and two domesticated stocked strains, as well as the temporal dynamics of hybridization relative to each domestic strain [86].

Under simplifying neutral assumptions, introgressed tract lengths should follow an exponential distribution [87] which would allow the timing of admixture events to be inferred [62,74,86] and
Box 1. Haplotype Phasing

In diploid species, every individual carries one autosomal chromosome copy inherited from its mother and one from its father, each comprising particular combinations of genetic variants. Classically, when diploid individuals are genotyped, differences between the two parental haplotypes appear as heterozygous sites among which linkage information is lost because sequence reads are usually too short to span pairs of contiguous heterozygote positions (Figure I). Haplotype phasing allows this information to be retrieved by using two broad categories of phasing methods.

(i) Indirect approaches [31,32] can use short-read NGS data from related or unrelated individuals. If individuals are related and a pedigree is available, Mendelian transmission rules can be used to perform phasing-by-transmission. For example, in a mother–father–child trio, if the mother and child are heterozygous (A/a) and the father is homozygous (a/a), the derived variant (A) is necessarily on the maternal haplotype. Applying this rationale to all heterozygous sites in the genome of the child allows chromosome-sized haplotypes to be reconstructed. Different software can be used to perform phasing-by-transmission, such as Merlin [33], GATK [34], and Hapi [35]. If individuals are not related, observed frequencies and associations among alleles within a population can be used to perform statistical phasing to estimate the probability of every possible haplotype. The most commonly used software are Eagle [36], Beagle [37], and Shapelt [38].

(ii) Direct approaches [39] are based on whole-genome sequencing of a single individual using long contiguous DNA fragments. One option is to group long DNA fragments into pools within which genomic regions are uniquely represented. Each resulting pool is then converted to a uniquely identified shotgun-sequencing library. In each pool, short reads mapping to the same genomic region thus belong to the same haplotype, allowing phase reconstruction. Liked-reads technologies such as chromium genome sequencing (10X Genomics) have been developed based on this principle. Alternatively, third-generation sequencing technologies such as PacBio and Oxford Nanopore allow us to directly access the phase information by sequencing long reads of several tens of kilobases.

Both methods present advantages and disadvantages. Statistical phasing is the most straightforward and least expensive method but requires large sample size and is less accurate because low-frequency SNPs may not be phased. Phasing-by-transmission approaches are more accurate but are also more expensive because they require closely related individuals, which may be a major problem in wild populations. Direct approaches are the most accurate methods but are also the most expensive. Recently it has been shown that combining both approaches improves the accuracy of the inferred haplotype structure [40].

### Figure I. Schematic Representation of the Information Obtained through Haplotype Phasing.

One sequence is represented with invariable positions in gray and three SNPs in green. The different possible allelic associations between these three SNPs form four different possible haplotypes. Phasing-by-transmission allows the true parental allelic associations to be identified by determining whether each variant was paternally (blue) or maternally (yellow) inherited.
Box 2. Identity-by-Descent (IBD) and Identity-by-State (IBS)

Haplotype similarities between individuals (or between homologous sequences within a diploid individual) can result from sharing a common ancestor, where allelic combinations remain unbroken by recombination. Such segments show IBD (Figure I). The length distribution of IBD blocks reflects the age of shared ancestry because short blocks will have undergone, on average, more recombination events and will therefore represent longer time to the most recent common ancestor (MRCA). By contrast, longer IBD blocks will be indicative of a more recent MRCA.

A common difficulty with the analysis of pairwise IBD (other than ROHs) is that ancestry inferences are necessary to delineate them. Currently available IBD detection methods (e.g., [31,55–57]) are more accurate for identifying long tracts (length >2 centimorgan (cM)) because intermediate tracts (1–2 cM) can result from the confluations of shorter tracts [58]. Moreover, nearly all methods developed to identify IBD were optimized using human datasets or simulations mimicking human genome properties and demographic history. The appropriateness of these methods in species exhibiting highly different demographies has not yet been tested, and more simulation studies may be necessary before they can be applied more broadly.

Haplotypes defined as IBS are identical sequences delimited by two polymorphic sites. They do not require a shared ancestry and, consequently, IBS does not necessarily imply IBD. Some authors (e.g., [59]) also consider that IBD segments can bear new mutations, and therefore do not always imply IBS. The major difference between IBD and IBS is related to the MRCA: IBD is mostly used to infer ‘recent’ demography, whereas IBS often refers to both long and short segments, and therefore may provide information on longer timescales [60]. IBD tracts can be a good alternative to IBD tracts [60] because they are directly observed from the data. However, IBS are also influenced by sequencing and phasing errors. Although IBS tracts have not been widely used in nonmodel species (but see [61] and [62]), they can be analyzed with methods that incorporate linkage information and also accommodate complex demographic models of split, mixture, and population size change.

Figure I. Identity-by-Descent (IBD) (Left) and Identity-by-State (IBS) (Right) Segments.

IBD segments are displayed in the case of a half-sibling. IBS does not necessarily imply a shared common ancestor and can be inherited by any individuals. The yellow and green colors represent ancestry tracts broken by recombination over time. Abbreviation: MRCA, most recent common ancestor.
introgressed population, several methods focusing on LD patterns have been proposed [94]. A new LD statistic that weights SNPs according to their level of differentiation between two admixing populations was first used to study admixture between sub-Saharan African and West Eurasian human populations [95], and was subsequently improved in following studies [96–98]. This approach was recently modified to consider LD originating from the source population while modeling multiple waves of admixture events [99] and continuous gene flow [100]. This type of approach was used to

Figure 2. The use of Highly Homozygous Identity-by-Descent (IBD) Segments (Runs Of Homozygosity, ROH) To Study the Demography of a Nonmodel Species, the Collared Flycatcher (Ficedula spp.). (Upper panel) Distribution of the genome proportion in ROH for two classes of TMRCA (time to the most recent common ancestor). The greater abundance of ROH in a given class indicates a small effective population size ($N_e$) during the period considered. This information was used to quantitatively compare $N_e$ among six populations from four species of flycatcher (Ficedula spp.) in different historical time-periods. Each color corresponds to a different species: orange (collared flycatchers); green (pied flycatchers); gray (Atlas flycatchers); and light blue (semicollared flycatchers). (Bottom right) Change in recent $N_e$ (black line) and its 95% confidence interval (broken line) inferred from pairwise IBD segments in the Baltic collared flycatcher. The analysis of pairwise IBD revealed that the Baltic population of the collared flycatcher was founded <60 generations ago and displayed the smallest $N_e$ of all populations. Adapted, with permission, from Kardos et al. [54].
study admixture between the gray wolf and domestic dog (C. lupus familiaris), which allowed more efficient conservation practices to be proposed that do not solely rely on external phenotypes to identify hybrids [20]. Methods based on the length distribution of IBS segments can also be used to study admixture [60]. For example, these have been used to study how polar bears (Ursus maritimus) diverged from brown bears (Ursus arctos) and adapted to life in the high Arctic. They revealed that several ancient hybridization events have most likely occurred between the two species [61].

Novel methods have also been developed to perform local ancestry inference while estimating the timing of a single [84] or several admixture pulses [101], without prior knowledge on the genetic structure of admixture groups [97]. One main advantage of these methods is that phased data are not needed, and they are also appropriate for low-coverage or pool-sequencing data [84,101]. These methods were tested on simulated data and returned estimates consistent with previous studies on the admixture history of Drosophila melanogaster populations [84,101].

**Selective Outcomes of Hybridization**

Hybridization between differentiated populations or species often results in heterogeneous patterns of local ancestry where genomic regions show increased or decreased frequencies of introgressed ancestry [29,30,74,102]. Such patterns might be modulated by neutral, positive, or negative selective forces [102]. To understand which forces are involved, it is important to consider local variation in the recombination rate that modulates genome-wide ancestry profiles through different types of interactions between selection and recombination [103].

Furthermore, introgressed haplotypes are expected to be shortened faster in high compared with low-recombining regions [102]. Because the level of LD between introgressed variants modulates the efficiency of selection acting on them [87], the number of generations since hybridization is also an important factor to consider. Indeed, selective effects interfere at the scale of large tracts in first hybrid generations. By contrast, after hundreds of generations, introgressed haplotypes are sufficiently shortened by recombination that selective effects can start to operate at a local (i.e., locus) scale [60,86,87,102].

**Selective Effects at Large Tract Scales**

Relatively recent hybridization events (i.e., roughly up to 12 generations ago) will generally result in the occurrence of long foreign haplotypes. Consequently, both favorable and detrimental fitness effects will act at the scale of long ancestry tracts. In this situation, potential positive effects such as heterosis (i.e., hybrid vigor) [104,105] are expected to occur through local associative overdominance, masking the expression of partially recessive deleterious alleles (Figure 3) [106,107]. This is particularly expected to predominate when a small population exhibiting high genetic load is introgressed by a foreign nonloaded population [30,108]. Moreover, the accumulation of weakly deleterious alleles in small populations could translate into a strong genetic load particularly in isolated, inbred populations [6,109]. Negative effects on fitness because of outbreeding depression [110] are also expected in situations of genetic incompatibilities between alleles from foreign and recipient populations (e.g.,

---

**Box 3. Local Ancestry Inference**

Local ancestry inference is used to characterize mosaic ancestries resulting from admixture and the introgression of foreign alleles within recipient populations. Different local ancestry inference methods have been developed that rely on different types of data (phased or unphased) and techniques [75,76]. The wide majority of these are based on hidden Markov models (HMMs) where hidden states correspond to the different possible ancestries. The aim is to estimate, for every variable position along the genome, the probability that a variant originates from a particular ancestral population, thus allowing the reconstruction of a mosaic of continuous ancestry blocks along the genome. However, the number of populations to be considered, and preliminary knowledge of admixture parameters and linkage information, depends on the method used. Recently, new methods have been proposed that can simultaneously estimate local ancestry and infer admixture parameters [84,85]. A nonexhaustive list of the most commonly used methods and their main characteristics is presented in Table I.
<table>
<thead>
<tr>
<th>Software</th>
<th>Technique</th>
<th>Data for admixed individuals/ reference individuals</th>
<th>Type of data</th>
<th>Number of source populations</th>
<th>Accounting for background LD in ancestral population</th>
<th>Biological parameters needed</th>
<th>Inferred parameters</th>
<th>Ploidy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SABER (Tang et al. [73])</td>
<td>MHMM</td>
<td>Phased/ phased</td>
<td>High-density SNPs panel + genetic distances</td>
<td>≥ 2</td>
<td>Yes</td>
<td>None</td>
<td>None</td>
<td>Diploid</td>
</tr>
<tr>
<td>HAPMIX (Price et al. [77])</td>
<td>HMM</td>
<td>Unphased/ phased</td>
<td>High-density SNPs panel + genetic distances</td>
<td>2</td>
<td>Yes</td>
<td>Admixture time and genome-wide admixture proportions</td>
<td>None</td>
<td>Diploid</td>
</tr>
<tr>
<td>PCAdmix (Bryc et al. [78])</td>
<td>Principal component analysis + HMM</td>
<td>Unphased/ unphased</td>
<td>High-density SNPs panel + genetic distances</td>
<td>≥ 2</td>
<td>No</td>
<td>Admixture time</td>
<td>None</td>
<td>Diploid</td>
</tr>
<tr>
<td>ChromoPainter (Lawson et al. [79])</td>
<td>HMM</td>
<td>Phased/ phased</td>
<td>High-density SNPs panel + genetic distances</td>
<td>≥ 2</td>
<td>Yes</td>
<td>None</td>
<td>None</td>
<td>Diploid</td>
</tr>
<tr>
<td>LAMP-LD/ LAMP-HAP (Baran et al. [80])</td>
<td>HMM (window-based framework)</td>
<td>Unphased/ unphased</td>
<td>High-density SNPs panel + physical positions</td>
<td>≥ 2</td>
<td>Yes (and Mendelian segregation in family trios)</td>
<td>None</td>
<td>None</td>
<td>Diploid</td>
</tr>
<tr>
<td>RFMix (Maples et al. [81])</td>
<td>Conditional random field (CRF)</td>
<td>Phased/phased (phasing error correction)</td>
<td>High-density SNPs panel + genetic distances</td>
<td>≥ 2</td>
<td>No</td>
<td>Admixture time</td>
<td>None</td>
<td>Diploid</td>
</tr>
<tr>
<td>ELAI (Guan [82])</td>
<td>Two-layer HMM</td>
<td>Unphased/ unphased (also works with phased reference)</td>
<td>High-density SNPs panel + genetic distances</td>
<td>≥ 2</td>
<td>Yes</td>
<td>Admixture time</td>
<td>None</td>
<td>diploid</td>
</tr>
<tr>
<td>Ancestry_HMM (Corbett-Detig and Nielsen [84])</td>
<td>HMM</td>
<td>Unphased/ unphased</td>
<td>Read pileup data</td>
<td>2</td>
<td>No</td>
<td>Global ancestry proportion and chromosome number</td>
<td>Admixture time</td>
<td>Arbitrary ploidy</td>
</tr>
<tr>
<td>Loter (Dias-Alves et al. [83])</td>
<td>Analytical resolution</td>
<td>Phased/phased (phasing error correction for two source populations)</td>
<td>High-density SNPs panel + physical positions</td>
<td>≥ 2</td>
<td>No</td>
<td>None</td>
<td>None</td>
<td>Diploid</td>
</tr>
<tr>
<td>MOSAIC (Salter-Townshend and Myers [85])</td>
<td>HMM</td>
<td>Phased/ phased (phasing error correction)</td>
<td>High-density SNPs panel + genetic distances</td>
<td>≥ 2</td>
<td>Yes</td>
<td>None</td>
<td>Admixture time and proportion, and (F_{ST})*</td>
<td>Diploid</td>
</tr>
</tbody>
</table>

\(*F_{ST}\) (the fixation index that varies between 0 and 1 and measures the extent of genetic differentiation among subpopulations)
Bateson–Dobzhansky–Muller incompatibilities). These might also be revealed by admixture between diverged populations (Figure 3) or species that differ in their genomic architectures (e.g., the presence/absence of large inversions) [29,111].

Localized Selective Effects

In older hybrid generations, selective effects are more likely to act at the locus scale [87,101]. Maladaptive fitness effects of introgressed alleles could thus emerge only after a long time following hybridization events, when deleterious alleles become dissociated (through recombination events) from each other and from potentially beneficial alleles at other loci. In particular, this is expected when admixture occurs among populations of small \( N_e \) [109]. This process is expected to be accompanied by a progressive decrease in associative overdominance effects through time. For instance, it was proposed that the occurrence of several diseases in modern humans was a result of ancient introgression events with Neanderthals [112,113]. Conversely, adaptive introgression has also been documented in modern human populations.
The introgression of Neanderthal and Denisovan DNA has apparently conferred selective advantages to modern humans, for example skin pigmentation, immune response to pathogens, and adaptation to altitude [87]. However, the occurrence of adaptive or maladaptive introgression has almost never been investigated in nonmodel species undergoing natural or anthropogenic hybridization (i.e., genetic rescue [114]) which would be of interest for conservation (but see [113]). In brook char (Salvelinus fontinalis), for example, short-term positive effects of introgression in stocked populations have been documented [116]. This result, however, does not necessarily indicate long-term positive effects because hybridization with the domestic strain used for supplementation is recent, and thus the effects of potentially maladaptive alleles of domestic origin could be revealed later after the dissipation of associative overdominance. This highlights the importance of considering the temporal dynamics of introgression in a conservation context. In summary, selective pressures, either negative or positive, can modulate variation in introgression rates across the genome, and this, as a function of many interacting parameters, including recombination rate or variation in effective population size.

Concluding Remarks

Although NGS methods allow the generation of huge amount of genomic data relatively quickly and cheaply, genomic variation in species other than humans is still largely analyzed on the basis of independent SNPs (except for micro-haplotype studies), without tapping into the substantial source of information contained in patterns of LD variation across the genome (see Outstanding Questions). Nevertheless, the recent studies and new analytical developments reviewed above clearly show that this represents a missed opportunity toward improving the use of genomics to guide our conservation decisions and management strategies. At a within-population level, much can be learned from haplotype information, which provides a powerful means to perform demographic inference. At the between-population level, haplotype information provides an in-depth picture of the magnitude of both contemporary and historical gene flow between populations by retrieving the mosaic of ancestry tracts [29,62].

Clearly, the outcomes of both empirical and simulation studies performed at the haplotype level highlight the importance of considering time since the onset of hybridization events in a conservation context. In particular, a frequently unappreciated outcome of hybridization events is that the directionality of selection acting on a given gene may vary over time as a function of the decreasing size of linkage blocks. As a consequence, positive effects following introgression can occur during the first hybrid generations, driven by the masking of partially recessive deleterious mutations (i.e., through associative overdominance), increasing the fraction of introgressed local foreign ancestry. Later, these potentially deleterious alleles might reveal their individual effects with the diminishing local fraction of foreign ancestry and the shortening of linkage blocks [109]. Therefore, to establish appropriate conservation strategies that would take the ‘hybridization problem’ into account, it appears crucial to document the temporal dynamics of introgressive hybridization [6,86,109,117]. It is also important to take into consideration variations of introgression and recombination rate along the genome because differential selective forces might also operate along the genome (i.e., favorable or unfavorable to the introduced alleles [116]). To conclude, our review is an attempt to encourage consideration of the great potential of LD information to improve our knowledge of the demographic history (both past and recent) of populations and to understand why admixture and/or introgression rate fluctuate along the genome [29,30,102,118]. Clearly, there is much to be gained by integrating haplotype-based analyses in future studies pertaining to conservation genomics (see Outstanding questions).

Acknowledgments

We thank Andrea Stephans, Marty Kardos, and an anonymous reviewer for their constructive comments that greatly improved this manuscript.

References


Outstanding Questions

Will the horizontal signal contained in genome-wide genotype data (i.e., haplotype structure) broaden the amount of information useful for conservation compared with the vertical information of allele frequencies?

Will haplotype studies help to better address conservation and management issues such as population structure, inbreeding, genetic connectivity, and the consequences of anthropogenic hybridization?

Can conservation and management strategies benefit from improved estimates of contemporary population sizes and dispersal distances through the use of haplotype information?

What additional understanding of the temporal dynamics and evolutionary consequences of introgressive hybridization can we gain from local ancestry inference versus conventional admixture analyses?

To what extent does the length of admixture tracts interplay with the different selective mechanisms occurring, and how does this affect the efficiency of genetic rescue?


