MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11 (Suppl. 1),** 123–136 doi: 10.1111/j.1755-0998.2010.02943.x

ANALYTICAL APPROACHES

Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges

S. J. HELYAR,* J. HEMMER-HANSEN,† D. BEKKEVOLD,† M. I. TAYLOR,* R. OGDEN,‡ M. T. LIMBORG,† A. CARIANI,§ G. E. MAES,¶ E. DIOPERE,¶ G. R. CARVALHO* and E. E. NIELSEN†
*Molecular Ecology and Fisheries Genetics Laboratory (MEFGL), School of Biological Sciences, University of Bangor, Environment Centre Wales, Bangor, Gwynedd LL57 2UW, UK, †National Institute of Aquatic Resources, Technical University of Denmark, Vejlsøvej 39, DK-8600 Silkeborg, Denmark, ‡TRACE Wildlife Forensics Network, Royal Zoological Society of Scotland, Edinburgh EH12 6TS, UK, §Molecular Genetics for Environmental & Fishery Resources Laboratory – GenMAP, Interdepartmental Centre for Research in Environmental Sciences, University of Bologna, Ravenna 163-48100, Italy, ¶Laboratory of Animal Diversity and Systematics, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium

Abstract

Recent improvements in the speed, cost and accuracy of next generation sequencing are revolutionizing the discovery of single nucleotide polymorphisms (SNPs). SNPs are increasingly being used as an addition to the molecular ecology toolkit in nonmodel organisms, but their efficient use remains challenging. Here, we discuss common issues when employing SNP markers, including the high numbers of markers typically employed, the effects of ascertainment bias and the inclusion of nonneutral loci in a marker panel. We provide a critique of considerations specifically associated with the application and population genetic analysis of SNPs in nonmodel taxa, focusing specifically on some of the most commonly applied methods.

Keywords: ascertainment bias, nonneutral loci, outlier detection, population genomics, population structure, software

Received 14 July 2010; revision received 23 September 2010; accepted 27 September 2010

Introduction

Recent improvements in the speed, cost and accuracy of next generation sequencing (NGS) and advances in the accompanying bioinformatic tools are revolutionizing the opportunities for generating genetic resources in nonmodel organisms. In turn, this is driving a shift from anonymous markers such as microsatellites to direct analyses of sequence variation including single nucleotide polymorphisms (SNPs). This shift has evolved from the initial uptake of such markers in humans and other commercially important species, to their application in a wide range of nonmodel species.

SNPs are attractive markers for many reasons (for reviews see Brumfield *et al.* 2003; Morin *et al.* 2004), including the availability of high numbers of annotated markers, low-scoring error rates, relative ease of calibration among laboratories compared to length-based markers and the associated ability to assemble combined temporal and spatial data sets from multiple laboratories. Additionally, the potential for high-throughput genotyping improved genotyping results for poor quality samples [such as historical, noninvasive or degraded samples (Morin & McCarthy 2007; Smith *et al.* 2011)], a simple mutation model, and the ability to examine both neutral variation and regions under selection offers unparalleled scope for expansive screening of genomes and large sample sizes from natural populations. Although several early studies questioned the advantage of SNPs over neutral markers such as microsatellites (e.g. Rosenberg *et al.* 2003), more recent studies have shown that SNPs are also showing promise as highly informative markers, as many studies with access to very large numbers of SNPs (mainly human) have shown that a small fraction of the SNPs have a very high information content for population structure analysis (e.g. Lao *et al.* 2006; Paschou *et al.* 2007), outperforming microsatellites (Liu *et al.* 2005). Despite microsatellites typically displaying far greater allelic diversity per locus, individual SNPs can segregate strongly among populations (Freamo *et al.* 2011; Karlsson *et al.* 2011).

Correspondence: S.J. Helyar, Fax: 01248 370731;
E-mail: s.helyar@bangor.ac.uk

Although SNPs are increasingly being used as an addition to the molecular ecology toolkit, their use as a standard tool in nonmodel organisms remains challenging, with debate over how to utilize them most efficiently. A recent study by Garvin *et al.* (2010) reviewed the technical aspects of SNP discovery and genotyping, but there are also challenges associated with the analysis of SNP data. These concerns vary depending on the questions being addressed: some specific issues have been covered in other papers (e.g. parentage assignment, Anderson & Garza 2006; Hauser *et al.* 2011; power assessment, Morin *et al.* 2009; development of linkage maps, Ball *et al.* 2010 and relatedness, Krawczak 1999). However, an overview of the considerations specifically associated with the application of SNPs and their appropriate analysis in population genetic studies of nonmodel organisms appears timely. We focus specifically on some of the most commonly applied methods and first discuss the challenges common to all analyses; problems arising from the dramatic increase in the number of markers that are available, the effects of ascertainment bias and the inclusion of nonneutral loci in a marker panel.

Number of loci

Using SNP data to analyse population structure is theoretically straightforward, but until recently a major obstacle was the identification of software that could handle large data sets. However, for many of the standard analyses, such as basic descriptive statistics, authors have modified their software to accept several thousand loci (see Table 1). Nevertheless, many packages are still limited by either the number of loci or the sum of individuals × loci that can be analysed. Additional problems may also arise when using some analytical methods that are computationally intensive, such as Bayesian MCMC methods. While such software may accept very large data sets, the time taken for a standard desktop computer to conduct the analysis may be prohibitive.

Ascertainment bias

Ascertainment bias is the systematic deviation from the expected allele frequency distribution that occurs because of the sampling processes used to find (ascertain) marker loci. In SNPs, this may occur as the markers are generally identified in a small panel of individuals from part of the species' range (ascertainment width). Likewise, only SNPs occurring more than a predefined number (k) of times in the ascertainment sample are included (ascertainment depth). When these SNPs are then genotyped on a larger sample of individuals, an 'ascertainment bias' is introduced (Nielsen 2000; Albrechtsen *et al.* 2010). Because of the small size of the ascertainment panel (compared to the population), the probability that a SNP is identified in this panel is a function of its minor allele frequency (MAF), i.e. SNPs with a very low MAF are less likely to be discovered than those with a higher MAF.

Ascertainment bias may compromise analyses based on diversity measures, for example, any statistical measure that relies on allele frequency may be affected. Because there is a bias towards not sampling rare SNPs, the average diversity of polymorphic sites is overestimated, while the average diversity across all sites is underestimated. This may lead to a bias in estimates of nucleotide diversity, population size, demographic changes, linkage disequilibrium, selective sweeps and inferences of population structure (Nielsen 2000; Schlötterer & Harr 2002; Akey *et al.* 2003; Nielsen & Signorovitch 2003; Marth *et al.* 2004; Rosenblum & Novembre 2007; Storz & Kelly 2008; Guillot & Foll 2009; Chen *et al.* 2010; Moragues *et al.* 2010). The size and direction of the bias depend on the sampling strategy used for the ascertainment panel; for example, studies on both humans and *Drosophila* suggest that genetic diversity will be underestimated if individuals from the ancestral population range are not included in the ascertainment panel (Schlötterer & Harr 2002; Romero *et al.* 2009). However, a panel based on purely ancestral (African) *Drosophila* did not underestimate the diversity in the European populations. Moreover, a study by Rosenblum & Novembre (2007) that examined a spatially structured population of lizards found that choosing individuals at random from across the geographical range minimized the resulting bias. However, some studies with small ascertainment panels are not addressing these issues (e.g. Kerstens *et al.* 2009; Li *et al.* 2010).

Three main approaches have been used to address and correct for ascertainment bias in studies of natural populations: (i) the application of more robust methods, such as those based on haplotype structure (e.g. Sabeti *et al.* 2007, however, this requires a full genome as reference), (ii) the simulation of data based on the ascertainment process to derive appropriate critical values and confidence intervals taking the ascertainment into account (e.g. Carlson *et al.* 2004; Voight *et al.* 2006) and (iii) the direct correction of the statistical estimators and statistics using specific models (e.g. Nielsen 2000; Wakeley *et al.* 2001; Nielsen & Signorovitch 2003; Polanski & Kimmel 2003; Marth *et al.* 2004; Nielsen *et al.* 2004;. Also see Table 1). However, a major restriction is that the correction of the allele frequency spectrum restricts downstream analyses to corrected summary statistic data (allele frequencies) with the loss of the observed individual genotypes that are needed for many applications (e.g. determining population structure, individual

Table 1 Computer software used for the most common aspects of population genetics

Programme	Functions	Maximum number of loci	Maximum number of individuals	Reference and web address
PEAS v1	Multiple data manipulation and summary statistics	None	None	Xu *et al*. (2010). http://www.picb.ac.cn/~xushua/index.files/Download_PEAS.htm Data manipulation includes file conversion for other population genetics programmes
SNPator	Multiple data manipulation and summary statistics	None	None	Morcillo-Suarez *et al*. (2008). http://www.snpator.org/public/downloads/aRamirez/tajimasDCorrector/
POPGENE	Multiple summary statistics	1000	1400 pops/150 groups	http://www.ualberta.ca/~fyeh/
Arlequin 3.5*	Multiple summary statistics	None	None	Excoffier & Lischer (2010). http://cmpg.unibe.ch/software/arlequin35/
Genepop v4	Multiple summary statistics	None	None	Rousset (2008). http://kimura.univ-montp2.fr/~rousset/Genepop.htm
popgen†	Multiple	None	None	http://mathgen.stats.ox.ac.uk/software.html
FSTAT2.9.4	Multiple summary statistics	10 000	200	Goudet (1995). http://www2.unil.ch/popgen/softwares/fstat2.9.4_10kloc_9all_200pops.zip
HIERFSTAT†	F-statistics	None	None	Goudet (2005). http://www.unil.ch/popgen/softwares/hierfstat.htm.
GenAlEx6.4	Multiple summary statistics	127 or 8192‡	65 500	Peakall and Smouse (2006). http://www.anu.edu.au/BoZo/GenAlEx/index.php
Genetix4.05	Multiple	None	None	http://www.genetix.univ-montp2.fr/genetix/genetix.htm
AscB†	Correction for Ascertainment Bias	None	None	Guillot & Foll (2009) http://www2.imm.dtu.dk/~gigu/AscB/
trueFS	Correction for Ascertainment Bias	None	None	Nielsen *et al*. (2004). http://people.binf.ku.dk/rasmus/webpage/truefs.html
Plink1.07§	Multiple	None	None	Purcell *et al*. (2007). http://pngu.mgh.harvard.edu/purcell/plink/
DetSel	Outlier locus detection	None	None	Vitalis *et al*. (2003). http://www.genetix.univ-montp2.fr/detsel.html
FDIST2	Outlier locus detection	None	None	Detection of loci under selection from hierarchical F-statistics, implemented in Arlequin (see above)
BAYESFST¶	Outlier locus detection	None	None	Beaumont & Balding (2004). http://www.reading.ac.uk/Statistics/genetics/software.html
LOSITAN	Outlier locus detection	None	None	Antao *et al*. (2008) http://popgen.eu/soft/lositan/
BayeScan	Outlier locus detection	None	None	Foll & Gaggiotti (2008). http://www-leca.ujf-grenoble.fr/logiciels.htm
matSAM v2	Outlier locus detection	None	None	Joost *et al*. (2008). http://www.econogene.eu/software/sam/
Structure 2.3.3	(Spatial) Genetic Structure	The maximum data set size around 100 million genotypes (loci × ind.)**††		Pritchard *et al*. (2000). http://pritch.bsd.uchicago.edu/structure.html
PCAGEN	Genetic Structure	50	5000 ind. 500 pops	http://www2.unil.ch/popgen/softwares/pcagen.htm

Table 1 Continued

Programme	Functions	Maximum number of loci	Maximum number of individuals	Reference and web address
adegenet†	Genetic Structure	None	None	Jombart (2008). http://adegenet.r-forge.r-project.org/
Geneland†	Spatial Genetic Structure	None**	None**	Guillot & Santos (2009). http://www2.imm.dtu.dk/~gigu/Geneland/
TESS 2.3	Spatial Genetic Structure	None**	None**	Chen *et al.* (2007). http://membres-timc.imag.fr/Olivier.Francois/tess.html
BAPS	Genetic Structure	None**	None**	Corander *et al.* (2008). http://web.abo.fi/fak/mnf/mate/jc/smack_software_eng.html
GESTE	Genetic Structure	None**	None**	Foll & Gaggiotti (2006). http://www-leca.ujf-grenoble.fr/logiciels.htm
GeneClass2	Assignment	None**	None**	Piry *et al.* (2004). http://www.ensam.inra.fr/URLB/GeneClass2/Setup.htm
WHICHLOCI	Locus selection	None**	None	Banks *et al.* (2003). http://www.bml.ucdavis.edu/whichloci.htm
GAFS 1.1	Locus selection	None**	None	Topchy *et al.* (2004). http://www.fw.msu.edu/~scribne3/molecularecology/programs.htm
BELS	Locus selection	None**	None	Bromaghin (2008). http://alaska.fws.gov/fisheries/biometrics/programs.htm

*Although Arlequin is not an R package, the latest version interfaces with R to produce the graphs.
†An R package. Additional packages may be found at http://cran.r-project.org/web/views/Genetics.html
‡The number of loci is dependant of the version of excel that you use, for pre-2007 as the number of columns in Excel was 256, but this has increased in Excel 2007 to 16 384 columns. For versions of GenAlEx 6.3 onwards, users are given the choice of installing either GenAlEx6.3.xla or GenAlEx 6.3 for 2007.xla. Both versions will run in Excel 2007, but to take advantage of full compatibility with Excel 2007 you should instal the Excel 2007 specific option.
§Extensible with via R function plug-ins.
¶R scripts also available on the website.
**While the are no physical constraints on the numbers of loci or individuals that can be submitted to this programme, the number of permutations that may needed for the computation of some options may make the calculation prohibitive on a standard desktop computer.
††The authors suggest reducing the data set for the exploratory analysis. Additionally, for large data sets, the default settings for BURNIN and NUMREPS can be reduced, without affecting the accuracy.

assignment, multilocus heterozygosity estimates, mixed stock analysis).

The generation of more and longer reads will eventually lead to the next step in SNP genotyping, where individuals are directly (single track) sequenced, followed by a high-confidence assembly and phasing of sequence reads [using for example; Phase (Stephens *et al.* 2001), FastPhase (Scheet & Stephens 2006), Shape-IT (Delaneau *et al.* 2008)]. Alternatively, the genotyping-by-sequencing approach, used for instance in RAD sequencing, combines the power of high throughput sequencing and large-scale polymorphism genotyping in one step (for a limited number of individuals) significantly reducing the problem of ascertainment bias (Baird *et al.* 2008; Hohenlohe *et al.* 2010).

However, as NGS data is likely to remain the basis of SNP development for the foreseeable future (see conclusions), and considering the inherent properties of newly developed SNPs, ascertainment bias is likely to remain a problem in the near future and may lead to incorrect population genetic inferences. Consequently, attempts must be made both to minimize the effects by careful design of the ascertainment panel. This can be achieved by the geographical sampling of multiple individuals, the tagging of individuals used in the sequencing for later genotype/haplotype reconstructions and a sufficient sequencing depth for *in silico* frequency spectra to be assessed before final SNP genotyping (for instance, by combining long (454 Roche) and short (ABI, Illumina) read sequencing runs for reference assembly and SNP discovery,

respectively). Accounting for the bias in the resulting data with the use of up to date statistical and simulation/modelling tools will allow the robustness of results to be assessed, despite assumption violations (Balzer *et al.* 2010). However, as explored in more detail in the following sections, ascertainment bias need not always pose a problem.

Nonneutral loci

The availability of thousands of genetic markers reinforces the need for careful evaluation of the markers used for a specific population genetic study, as markers in genic and nongenic regions may generally differ with respect to basic properties such as levels of variation and population differentiation, which will affect the outcome of downstream analyses. In genome-wide association (GWA) studies in humans, it has been found that SNPs in genic regions are more likely to display signatures of both positive and negative selection than those in nongenic regions (Barreiro *et al.* 2008; Coop *et al.* 2009) and that genetic variation is generally lower in gene-rich regions (Cai *et al.* 2009). While the degree that these findings apply to nonmodel organisms remains unknown, they do indicate that markers situated in or close to genes may not provide a representative picture of genome-wide effects of neutral evolutionary forces. Additionally, genomes contain gene regulatory networks (GRNs) that are highly conserved regions within the noncoding DNA (Davidson *et al.* 2002; Woolfe *et al.* 2005); this implies both that these regions will not be identified by transcriptome sequencing and also that there are sections of noncoding DNA that are under selection.

SNPs represent the most widespread type of sequence variation in genomes, and the combination of the continuing decrease in costs for NGS and new efficient methodologies, such as RAD-tag sequencing (e.g. Miller *et al.* 2007; Baird *et al.* 2008) and RRS (reduced representation sequencing—e.g. Castano-Sanchez *et al.* 2009), is showing great promise for fast, efficient SNP detection in nonmodel species. While there are methods that can preferentially target noncoding regions (e.g. EPIC markers, Palumbi & Baker 1994), there are also increasing expressed sequence tag (EST) resources available for many taxa, increasing the likelihood that many SNP loci that are being developed will be located either within or very close to coding regions. However, it is now thought that animal genomes are pervasively transcribed (Ponting *et al.* 2009) with a large number of noncoding transcripts being polyadenylated, which will therefore be included in EST collections. Consequently, the representation of the genome might be larger and have fewer constraints on sequence variability than previously thought.

For some applications, this potential bias in genome coverage has been highlighted as an advantage, if for example, the aim is to identify candidate genes under selection (Bonin 2008; Brieuc & Naish 2011; Hemmer-Hansen *et al.* 2011 and also see the discussion in the section 'Detection of Outliers' below). However, issues could arise if the purpose of a study is to make general inferences about neutral evolutionary processes, such as genetic drift and gene flow. In such cases, markers under selection should be removed prior to analyses (Beaumont & Nichols 1996), as they may bias results significantly (see also discussion in Laval *et al.* 2010 and below). On the other hand, markers under selection could be exploited for specific purposes, such as investigating population structure on ecological rather than evolutionary timescales (Waples & Gaggiotti 2006), and for increasing the power for assigning individuals to populations of origin (Nielsen *et al.* 2009b).

With these caveats in mind, we now review the application of the most common analytical methods in population genetics to SNP data, paying special attention to the significant issues described, particularly how ascertainment bias and nonneutral loci affect analyses and how such effects can be addressed. Finally, we highlight salient priorities for further research in the integration of SNPs into molecular ecology.

Population genetic data analyses

Measures of genetic differentiation and population structure

With the ever increasing opportunities for SNP mining in nonmodel species, it is becoming increasingly evident that the apparent shortcomings of individual SNPs to detect population structure compared to microsatellites (Rosenberg *et al.* 2003) can be overcome by the relative ease with which large numbers of SNP markers can be developed and screened. The statistical power to detect population structure is related to the total number of alleles examined, and the discriminatory power of ~100 (neutral) SNPs is very roughly equivalent to 10–20 microsatellites (Kalinowski 2002). Moreover, the most informative SNP markers (i.e. those that show the greatest allele frequency variation among populations) in a panel may rival (or even exceed) the average information content of microsatellite markers (e.g. Liu *et al.* 2005; Smith *et al.* 2007). Using SNP markers to investigate population structure is theoretically straightforward, and most standard population genetic software packages allow for inclusion of large numbers of loci. However, there are also practical considerations, as some (especially Bayesian) methods are computationally intensive and may have problems handling very large data sets.

Wright's *F*-statistics are arguably the most commonly used descriptive statistics in population and evolutionary genetics (Wright 1931). As their original development, many related statistics have been described either as improvements or for specific applications, for example, for microsatellite data (G_{ST}, θ, and R_{ST}), sequence data (Φ_{ST}) and for quantitative traits (Q_{ST}) (see Holsinger & Weir 2009 for a review). One issue that has caused much debate is how to compare diversity estimates among markers, with much focus on the effect of differing mutation rates and levels of heterozygosity between highly polymorphic markers, such as microsatellites, and less variable markers, such as allozymes and SNPs (Waples & Gaggiotti 2006; Allendorf & Luikart 2007). In 2005, Hedrick proposed the new statistic G'_{ST} to provide a measure of differentiation that allows comparison among loci with different levels of genetic variation, such as among microsatellites, or between different marker types, such as allozymes/SNPs and microsatellites; measures such as this and the more recent D_{EST} (Jost 2008) are increasingly being used (e.g. De Carvalho *et al.* 2010; White *et al.* 2010). However, G'_{ST} has also been criticized as uninformative when migration is not expected to be negligible (Ryman & Leimar 2008). Mutation rates are in general considerably lower for SNPs than for microsatellites (Foll & Gaggiotti 2008; Excoffier *et al.* 2009), and more importantly while the expected locus-specific heterozygosity may reach more than 0.95 for a microsatellite marker, the maximum expected heterozygosity that can be reached by a bi-allelic SNP is 0.5. Such constraints mean that single locus F_{ST} estimates derived from SNP markers are likely to be more comparable than those derived from microsatellite loci. Many of the most frequently used programmes for calculating F_{ST} and related statistics have recently extended their capacity for numbers of loci and samples (details shown in Table 1).

Within human genetics, large-scale GWA studies are increasingly focusing on the population genetics of the samples, as unidentified structure may lead to spurious associations between traits and markers/genes. While such factors have enhanced the development of some SNP-specific software, such as Plink (Purcell *et al.* 2007), it has yet to be seen how applicable these are to more traditional population genetic approaches in nonmodel organisms.

In nonmodel species, global and pairwise F_{ST} values are typically estimated over all loci; as all markers are assumed to be effectively neutral, there should not be any major inconsistencies between loci. However, when loci are potentially under different selective pressures the estimates may be different for each locus, requiring per locus estimates. Xu *et al.* (2009) proposed a new measure of population structure specifically for SNPs. It is based on the c parameter (Nicholson *et al.* 2002), which is population-specific and measures the differentiation of the population from the common ancestral population. In contrast, the new measure C is an index of the overall levels of population structure across populations. Extensive simulations in Xu *et al.* (2009) show that C takes into account ascertainment bias and correlates well with Wright's F_{ST}. The correlation increases with increasing information (more SNPs and/or more subpopulations in the samples).

Clustering algorithms such as Bayesian MCMC clustering approaches are frequently utilized in genetic analyses. These methods define populations by minimizing departures from Hardy–Weinberg and maximizing linkage equilibrium. Clustering analyses can be performed independently of spatial information or be linked analytically to spatial and/or environmental parameters; the latter commonly termed 'landscape genetics' (Guillot *et al.* 2009). Genetic clustering and analyses of spatial structure can be based on neutral marker variation, on markers under selection or on a combination, with the last of these commonly being of particular interest in many EST-derived SNP approaches. User-friendly software for conducting such analyses includes Structure (Pritchard *et al.* 2000), BAPS (Corander *et al.* 2008), GESTE (Foll & Gaggiotti 2008) and Geneland (Guillot *et al.* 2005), the current versions of which all allow for the inclusion of larger numbers of loci (see Table 1). However, the assumptions of no linkage disequilibrium between markers common to many of these applications are likely to be violated with denser SNP coverage/representation across chromosomal regions, although some applications do allow the inclusion of linkage information (Falush *et al.* 2003). Including a relatively low number of markers in linkage disequilibrium is not likely to bias estimates of population differentiation, but may lead to overestimates of clusters (Kaeuffer *et al.* 2007). However, the effects of including markers with different levels of linkage disequilibrium on estimates of cluster numbers and divergence are not well described. As an alternative to Bayesian clustering, principal component analysis (PCA) and related approaches have been applied in several SNP studies of human population structure (Patterson *et al.* 2006). An advantage of PCA-based approaches, compared to Bayesian methods, is that PCA can be performed quickly on desktop computers. PCA approaches also facilitate the identification of subsets of markers that effectively describe differences among populations (Paschou *et al.* 2007), and it has even been argued that PCA outperforms Bayesian methods for inferring population structure when many loci are available and the structure is subtle (Reeves &

Richards 2009). However, PCA methods are sensitive to missing data and sampling effects, especially for species and populations with continuous distributions (Novembre & Stephens 2008), which can limit inference about underlying historical and demographic processes [although ways of circumventing these problems have been proposed for SNP data (Paschou *et al.* 2007)].

SNP-based estimates of population structure are potentially affected by ascertainment bias if the SNP panel used was developed for populations (or species) other than those analysed (Nielsen 2000). Nonetheless, few statistical assessments of the effect of ascertainment bias on fundamental measures such as F_{ST} estimates have been reported (although see Schlötterer & Harr 2002; Albrechtsen *et al.* 2010; and Moragues *et al.* 2010). Including information for loci either under directional or balancing selection themselves or loci tightly linked to regions under selection leads to violation of assumptions for most neutral population genetic models and may cause erroneous inference about population demographic parameters, such as rates of genetic drift and migration between individual demes. Several reports of population structure based on presumably neutral marker information are likely to (unknowingly) have incorporated nonneutral markers (Nielsen *et al.* 2006). In weakly structured species, the effect of just a few loci on overall patterns could be significant, but provided selected loci make up only a small proportion of the total marker number, biological inference is not generally expected to be severely biased (Luikart *et al.* 2003). Nonetheless, with SNP markers often developed from transcriptomic sequencing, the dramatic increase in genome coverage implies that some proportion of the markers are likely to be linked to genes/regions under selection, making it of paramount importance to test for marker 'neutrality' prior to exploring population structure (for example, by using outlier tests as outlined in the section below). Studies that combine information from neutral and nonneutral markers in analyses of population structure and estimation of demographic parameters are still scarce for nonmodel organisms (for examples see Gaggiotti *et al.* 2009; Nielsen *et al.* 2009a), and there is a need for development of analytical tools that allow integration across marker classes (Guillot *et al.* 2009).

Detection of outliers

The search for signatures of selection in molecular data has a long tradition in evolutionary biology. Most methods rely on the concept of genetic hitch-hiking (Maynard Smith & Haigh 1974), where a marker is linked to a site under selection, and although not the target of selection, the 'hitch-hiking' marker fails to display patterns of neutrality. For molecular markers, the methods to detect outlier loci can be divided into two broad categories, the first based on linkage disequilibrium between markers, and the second based on differences in levels of genetic variation and levels of genetic divergence between samples (see also Vasemägi & Primmer 2005).

Genome scan approaches (see Luikart *et al.* 2003 and Storz 2005 for reviews) have now been applied to an increasing number of nonmodel organisms (e.g. Anderson *et al.* 2005; Bonin *et al.* 2006; Hayes *et al.* 2007; Eveno *et al.* 2008; Moen *et al.* 2008; Namroud *et al.* 2008; Nielsen *et al.* 2009a), and this has generated insight into the pros and cons to the various approaches for detecting markers under selection in the wild.

Many nonmodel species still have little or no genomic resources, and the location of SNPs within the genome is therefore often unknown, rendering methods relying on detailed analyses of linkage disequilibrium unfeasible. Methods based on comparisons of genetic variation in random sets of markers have been developed both for microsatellites (Schlötterer 2002; Kauer *et al.* 2003; Marshall & Weiss 2006) and SNP-based haplotypes (Voight *et al.* 2006; Sabeti *et al.* 2007); however, these do not seem to be relevant for a relatively limited number of SNPs without genomic information. In contrast, many methods based on comparisons of levels of genetic divergence between samples can be applied to markers where information about genomic location is missing. Hence, these methods appear better suited for studies in nonmodel species.

Most methods based on comparisons of divergence among samples are based on the original Lewontin–Krakauer test, which compares single locus estimates of F_{ST} to an expected neutral distribution of F_{ST} (Lewontin & Krakauer 1973). The original Lewontin–Krakauer test is now rarely used, mainly because of concerns over its performance when allele frequencies are correlated between samples leading to an increased number of false positives (Robertson 1975; Beaumont 2005). However, several closely related methods have been proposed to overcome the shortcomings of the original approach. With a very large number of markers, it may be possible simply to estimate the expected distribution of F_{ST} from the markers themselves (e.g. Akey *et al.* 2002), but for most nonmodel organisms, the available number of markers is too limited and simulations must be used to generate the neutral distribution. In these cases, the model used for the simulations is crucially important, as it will effect the identification of outlier loci. For instance, Vitalis *et al.* (2001, 2003) developed a method (implemented in DetSel) based on pairwise population comparisons of individual locus F_{ST} to a simulated distribution of F_{ST} generated under a model of two fully isolated populations descended from a common ancestral population. Beaumont & Nichols (1996) developed FDIST2, which

uses a classical island model to generate the expected neutral distribution of F_{ST} estimates. While these methods remove the need to directly use genotyped markers as the baseline, they do so indirectly by using the estimated overall F_{ST} as a starting point for simulations. Thus, including loci under selection in the initial F_{ST} estimate may generate a bias in the simulated distribution. Additionally, the models used for the simulation of data in the two methods are unlikely to match most natural situations, because many populations are significantly connected through asymmetrical patterns of gene flow. The two limitations above have been addressed in later Bayesian methods based on logistic regression models of locus and population effects on F_{ST}. Both BAYESFST (Beaumont & Balding 2004) and BayeScan (Foll & Gaggiotti 2008) allow F_{ST} to vary between populations and identify loci potentially under selection through estimates of locus effects on F_{ST}. The two methods are based on the same basic regression model, but differ in the way that the effect of selection is inferred. While BAYESFST does not conduct a formal statistical test, BayeScan uses a likelihood ratio test to assess the most likely of the two alternative models (no effect of selection vs. effect of selection). Both programmes have been widely applied, but they have also recently been found to be vulnerable to complex population structure scenarios, such as when populations are hierarchically structured, leading to correlated allele frequencies among samples (Excoffier *et al.* 2009). A modified, hierarchical, version of FDIST2 implemented in the Arlequin 3.5 software may be more appropriate for such situations (Excoffier *et al.* 2009). The implementation of a hierarchical island model results in higher variance between simulated neutral loci and thus leads to a more conservative estimate of the number of outlier loci (Excoffier *et al.* 2009). It seems inevitable that the lower false-positive rate comes at the expense of a higher false-negative rate; however, the method has so far only been evaluated with simulated neutral loci, focusing on the discovery of false positives, rather than the power for discovering true positives.

While the Bayesian methods may be relatively powerful for detecting directional selection, they have low power for detecting loci under balancing selection, particularly for SNP applications (Beaumont & Balding 2004; Foll & Gaggiotti 2008). This may be problematic in situations with low levels of population structure, when the power for detecting directional selection could be substantially higher than the power for discriminating between loci under balancing selection and loci evolving under neutrality.

In general, a low number of samples also substantially reduces the statistical power of these methods (Foll & Gaggiotti 2008), meaning that pairwise comparisons (e.g. between populations under different environmental forcing) will detect only extreme outlier loci, and many potential candidate loci may be missed. In contrast, too many samples could also bias results, particularly if allele frequencies are correlated among samples, resulting in increased false-positive rates (Excoffier *et al.* 2009). This bias could be reduced through analysing balanced subsets of samples, i.e. using a similar number of samples from each of a number of populations or groups of populations identified through other approaches, such as clustering methods. Thus, a balanced design could minimize effects from complex population structure not easily handled by many current methods. Furthermore, it is possible to evaluate the effect of study design by running several tests on different subsets of samples.

The genetic resource originally used for developing the genetic markers can impact results of outlier detection approaches in several ways. For instance, it must be remembered that in current studies of nonmodel organisms, markers will often mainly be linked to the variation in coding (and expressed) parts of the genome (see section on nonneutral loci). Although the effects of such an ascertainment strategy on genome scans have yet to be assessed, in some approaches, these markers will be used to generate the expected 'neutral' distribution of F_{ST} values. However, if this baseline is biased, then results may not truly reflect the proportion of loci under selection. Further biases may be introduced through ascertainment bias (see introduction and discussion in Nielsen *et al.* 2009b). In addition, loci in linkage disequilibrium could bias results by introducing biased genome coverage among the markers, for instance biassing F_{ST} through physical linkage of loci displaying elevated or lowered levels of structuring.

Although the aforementioned methods have their limitations, they have all been developed to handle relatively large data sets and they are very useful for providing a general overview of the data at hand. Again, the important thing is to have clarity in the question that is being addressed. If the goal is to identify sets of markers with high discriminatory power between different populations/groups of populations, then in principle it does not matter if a detected outlier is truly subject to selection, or if it is a false positive, provided that the signal is temporally stable. In this case, the outlier detection can be viewed as an explorative and preliminary exercise supporting downstream analyses. However, if evolutionary or demographic processes are being investigated, the inclusion of loci under selection may influence results significantly and careful attention should be paid to the design of the scan for outlier loci.

Power analysis

Several population genetic applications, such as conservation management, product traceability and forensic

genetic analysis, involve the assignment of individuals, or collections of individuals, to population of origin based on their (multilocus) genotypes (Manel *et al.* 2005). Here, the inclusion of markers exhibiting evidence for diversifying selection need not violate assumptions and can dramatically increase assignment success, at least if all (or most) reference populations are represented in the baselines against which samples are compared. Analyses combining marker types should, however, be accompanied by simulations of how potential sampling effects could influence assignment (see Anderson 2010). Likewise, inclusion of nonneutral markers may be advantageous when attempting to estimate genetic admixture of individuals or populations.

For applications such as individual assignment (IA), there are many advantages (for example, the reduction in costs, time and computational demands) in using a reduced panel of markers that have been identified as maximizing the power available. For example, selection of breed-informative SNP markers for IA in cattle enabled a reduction in panel size from 54 000 to 200 SNPs with negligible loss of assignment power in twelve European cattle breeds (Wilkinson *et al.*, pers.com). However, a marker panel that has been reduced for this purpose is not suitable for many standard population genetic analyses because of the bias introduced through the high grading of markers that segregate among target populations (Waples 2010).

Identifying loci with maximum power. Not all genotyped loci are necessary for increasing assignment power. Loci may have high-genotyping error rates, be noninformative with little discriminatory power or be strongly correlated (linked) with other markers, thereby yielding redundant information. For some purposes, it may be desirable to create 'minimal panels with maximum power', for example; panels for assigning individuals to major groups, or very specific panels for discriminating between two alternative hypotheses in relation to individual assignment. The selection of loci to form SNP panels for assignment will be driven by the complexity of the assignment question involved. A biallelic marker will only ever be able to segregate two populations; therefore, multiple SNPs will be needed for IA when there are multiple candidate source populations. By assessing assignment power at the level of the individual SNP, there will always be a risk that the SNPs selected with most power (e.g. highest F_{ST} values), will be biased towards the most differentiated populations and will not allow for assignment to more finely differentiated groups. When dealing with large numbers of SNP markers, automated methods for selecting loci with the most power across a range of application scenarios are required; simply ranking SNPs by F_{ST} values is unlikely to lead to an optimum, minimal panel of markers for complex assignment problems, as it is particular combinations of loci that are likely to contain the highest discrimination power.

Three different approaches for locus selection have been developed together with accompanying software. WHICHLOCI (Banks *et al.* 2003) initially estimates the assignment power of individual loci from empirical data and ranks them according to individual assignment (and/or misassignments). In a second round of assignment, loci are added to an assignment trial from the top of the individual power list until the user specified level of accuracy is achieved. The programme and approach is relatively simple and straightforward. However, an important caveat is that the programme does not explore the potential power of certain combinations of loci, which may maximize IA, but may not include loci from the top of the list. An alternative approach is genetic algorithm-based feature selection (GAFS, Topchy *et al.* 2004). This programme uses a 'genetic algorithm' optimization technique, by exploring different locus combinations where the highest classification accuracy is the parameter of interest that is being searched for. The programme works on many solutions simultaneously in contrast to other optimization algorithms using incremental improvement (see above). Although the programme allows for an exhaustive search of all potential combinations, it may not be computationally feasible to explore all combinations, thereby leaving potentially highly discriminatory combinations unexplored. The third and most recently described option is 'backward elimination locus selection' using the programme BELS (Bromaghin 2008). The programme excludes each locus in the baseline data temporarily, and the baseline accuracy for assignment (or Mixed Stock Analysis) of remaining loci is evaluated iteratively. After all loci have been evaluated, the locus causing the least power reduction is permanently excluded. The procedure is repeated until only one locus is left or the level of accuracy reaches a user-defined minimum. The advantage of the programme is that (like GAFS) it exploits possible synergistic effects among loci. The downside is that with many loci and populations, it takes a long time to run on a standard desktop computer. Another shortcoming of the BELS procedure is in cases of forensic assignment where selection for the smallest subset of loci, providing 100% correct assignment is the goal. In this case, the programme is unable to rank loci as elimination of any locus from the full data set will not lead to a drop in overall assignment power (100%). Instead, a reverse procedure where loci are added according to their individual assignment power and subsequently eliminated using subsets of loci where assignment power is below 100% could be applied (J. Bromaghin, personal communication). Overall, it appears that the two latter

programmes represent the most optimal approaches for SNP loci under selection, as they search for 'synergistic' combinations of loci providing the highest overall level of assignment power regardless of their individual power.

A final note of caution for the selection of particular loci with elevated assignment power was pointed out in a recent paper by Anderson (2010). The programmes described in this section all use the same data for ranking loci and assessing their power, leading to biased and over-optimistic estimates of assignment power. Instead, Anderson suggested a procedure called THL (training, holdout, leave-one-out), where a subset of samples (training samples) is used for selection of highly informative loci to be included in the final panel of loci. These samples are combined with another subset of data (the holdout samples) to form the baseline for assignment using the final panel. By assigning the holdout samples using the full baseline sample employing a leave-one-out procedure, it is possible to separate the process of locus selection or 'high grading' from the evaluation of assignment power, while at the same time making use of the whole data set. This approach should be encouraged and implemented as a standard for evaluation of assignment power of loci under selection.

Power for detecting population differentiation. A recent paper by Morin *et al.* (2009) addresses the issue of the number of SNPs and sample size that should be used to maximize statistical power to identify evolutionary significant units (ESUs) and demographic independents units (DIPs) using the programme POWSIM (Ryman & Palm 2006). The 'effect sizes', i.e. the magnitude of differentiation required to detect two scenarios was F_{ST} = 0.2 and F_{ST} = 0.0025. The study assessed sample sizes within 10–100, number of loci 10–75 and MAFs 0.01–0.5. Overarching results showed that approximately 30 neutral loci were required to detect ESUs (N_em = 0.1), while identification of DIPs may require >75 loci. Different MAFs had little effect on power; haplotypes (linked loci) from different SNPs within the same locus could improve power, though sample size had a strong effect on power. For example, with 75 SNPs and F_{ST} = 0.0025, an increase in sample size from 50 to 100 provided a twofold increase in power (proportion of significant tests) from 0.4 to 0.8. Accordingly, if the aim is specifically to address the issue of microgeographical population structure, it may be advisable to use relatively large sample sizes. Also, including loci suspected to be under selection may increase power to detect differentiation; however, the stability of the pattern has to be investigated because contemporary selection may alter allele frequencies even within a cohort (see Nielsen *et al.* 2009a).

Glover *et al.* (2010) compared the IA resolution between analyses with 309 mapped SNPs (global F_{ST} −0.002 to 0.316; only one 'outlier locus') and 14 microsatellite markers (global F_{ST} 0.033–0.115) in wild and domesticated strains of Atlantic salmon (*Salmo salar*). They found that proportions of correctly assigned individuals was 0.65, 0.73 and 0.73 when assigned with 14 microsatellites, 300 SNPs and 195 'mapped' (>1 cM) SNPs, respectively. Overall, assignment was best (80% correct) when ~100 unlinked SNP loci were used. Above 100 loci, assignment success decreased. Comparing marker types, the most informative 15 salmon SNPs matched the level of assignment achieved by the most informative four microsatellite loci (ranked by maximizing allelic variation). If linkage information is available, Structure (Pritchard *et al.* 2000; Falush *et al.* 2003) may outperform Geneclass (Cornuet *et al.* 1999), as Structure enables the use of a linkage model, taking marker distance into consideration in computations, whereas Geneclass treats loci as independent. In the study by Glover *et al.* (2010) using Structure, the use of a linkage model led to 88% correct self-assignment when using 300 SNPs, whereas correct assignment was 80% with Geneclass. This study suggests that the identification of a highly informative set of SNPs from a larger panel is likely to give significantly more accurate individual genetic self-assignment compared to any combination of microsatellite loci. However, there is a risk of an upwards bias of the estimates of assignment success when 'high-grading' loci, as described by Anderson (2010). The study by Glover *et al.* (2010) also underlines the importance of using an appropriate method for modelling the statistical power and assignment resolution when choosing subsets of markers for targeted assignment analyses.

Conclusions

In several of the aforementioned sections, attention has been drawn to some of the concerns associated with the discovery of SNPs from NGS data. Some of these issues, such as the bias in genome coverage achieved, or the complications of not having a reference genome, are being dealt with by advances in technology (e.g. reducing the bias in terminal end sequencing (Korbel *et al.* 2007), paired-end reads for sequence assembly without a reference sequence (Li *et al.* 2010), also see Harismendy *et al.* 2009 for an evaluation of the different issues between platforms and Everett *et al.* 2011 for an assessment of the potential to assemble sequences to publicly available EST databases). Other major drawbacks such as the conversion rate from NGS data to validated SNPs, and the inherent ascertainment bias in the data still need practical solutions (for reviews see Hudson 2008; Shendure & Ji

2009; Garvin *et al.* 2010). NGS is one of the most powerful tools currently available, but its use must be undertaken with its limitations in mind. Meanwhile major advances in sequencing—such as the third generation technologies—are promising to resolve many of the difficulties with the current systems with less expensive, longer read, more accurate systems promised in the near future (Eid *et al.* 2009; Metzker 2009; Rusk 2009). However, although it has been suggested that ecologists may soon be able to perform population genetics at a genome, rather than a gene level (Hudson 2008), these technologies are likely to remain out of reach for the majority of studies on nonmodel organisms for the foreseeable future. Additionally, the replacement of SNP genotyping by the analysis of the full genome sequence data is also currently out of reach for the majority of nonmodel species.

The continued increase in speed and decrease in cost for SNP genotyping nonmodel organisms is undoubtedly going to lead to further major changes in relation to the availability of data on a genomic scale for population genetic analysis in the near future. Currently, we are in a transition period where population structure is typically inferred from relatively few genetic markers for some wild organisms, while thousands of markers and even whole genomes (Hohenlohe *et al.* 2010) are being analysed in others. Accordingly, we expect to see an increased movement towards genome wide analyses to gain a general understanding of the relative importance of neutral and adaptive processes in wild populations. Such a development will result in a conceptual change as it will no longer be feasible to manually edit or check data quality. In turn, further developments will be required in relation to statistical tools and associated software for analysing data orders of magnitude larger than is currently standard, some of which have been highlighted above. However, the fundamental principles of population genetics remain the same and specific research questions will continue to require appropriate analysis dependant on the nature of the markers used.

Although the data sets that we have access to are increasing in size, there will continue to be a need for small panels of 'genetic tags' for ecological, management and forensic purposes where the assignment of individuals and groups of individuals to the population of origin is desired. We expect these applications to grow tremendously and become commonplace as the costs of genotyping decline progressively. To generate added momentum, there is an enhanced need for genomic data for nonmodel taxa, from where the high grading of the most informative loci for individual assignment can take place to create cost-effective panels of minimum size with maximum power.

Acknowledgements

The discussions that formed the basis of this manuscript began at a FishPopTrace Consortium workshop held in February 2010. We thank all the members of the consortium present at that meeting, especially L. Bargelloni, for their contributions to the discussions; we also thank the two anonymous reviewers for their helpful comments.

Conflict of interest

The authors have no conflict of interest to declare and note that the sponsors of the issue had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805–1814.

Akey JM, Zhang K, Xiong M, Jin L (2003) The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Molecular Biology and Evolution*, **20**, 232–242.

Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution*, **24**, 1–20.

Allendorf FW, Luikart G (2007) *Conservation and the Genetics of Populations*. Blackwell Publishing, Oxford, UK.

Anderson EC (2010) Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources*, **10**, 701–710.

Anderson EC, Garza JC (2006) The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, **172**, 2567–2582.

Anderson TJC, Nair S, Sudimack D *et al.* (2005) Geographical distribution of selected and putatively neutral SNPs in Southeast Asian malaria parasites. *Molecular Biology and Evolution*, **22**, 2362–2374.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics*, **9**, 323.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Ball AD, Stapley J, Dawson DA, Birkhead TR, Terry Burke T, Slate J (2010) A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (Taeniopygia guttata). *BMC Genomics*, **11**, 218.

Balzer S, Malde K, Lanzén A, Sharma A, Jonassen I (2010) Characteristics of 454 pyrosequencing data – enabling realistic simulation with flowsim. *Bioinformatics*, **26**, i420–i425.

Banks MA, Eichert W, Olsen JB (2003) Which genetic loci have greater population assignment power? *Bioinformatics*, **19**, 1436–1438.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. *Nature Genetics*, **40**, 340–345.

Beaumont MA (2005) Adaptation and speciation: what can F_{ST} tell us? *Trends in Ecology and Evolution*, **20**, 435–440.

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

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London – Series B: Biological Sciences*, **263**, 1619–1626.

Bonin A (2008) Population genomics: a new generation of genome scans to bridge the gap with functional genomics. *Molecular Ecology*, **17**, 3583–3584.

Bonin A, Taberlet P, Miaud C, Pompanon F (2006) Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Molecular Biology and Evolution*, **23**, 773–783.

Brieuc M, Naish K (2011) Detecting signatures of positive selection in partial sequences generated on a large scale: pitfalls, procedures and resources. *Molecular Ecology Resources*, **11** (Suppl. 1), 172–183.

Bromaghin J (2008) BELS: backward elimination locus selection for studies of mixture composition or individual assignment. *Molecular Ecology Resources*, **8**, 568–571.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) Single nucleotide polymorphisms (SNPs) as markers in phylogeography. *Trends in Ecology & Evolution*, **18**, 249–256.

Cai JJ, Macpherson JM, Sella G, Petrov DA (2009) Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genetics*, **5**, e1000336.

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. *Nature*, **429**, 446–452.

Castaño-Sánchez C, Smith TPL, Wiedmann RT *et al.* (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, **10**, 559.

Chen C, Durand E, Forbes F, Francois O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, **7**, 747–756.

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. *Genome Research*, **20**, 393–402.

Coop G, Pickrell JK, Novembre J *et al.* (2009) The role of geography in human adaptation. *PLoS Genetics*, **5**, e1000500.

Corander J, Siren J, Arjas E (2008) Bayesian spatial modeling of genetic population structure. *Computational Statistics*, **23**, 111–129.

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, **153**, 1989–2000.

Davidson EH, Rast JP, Oliveri P *et al.* (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.

De Carvalho D, Ingvarsson PK, Joseph J *et al.* (2010) Admixture facilitates adaptation from standing variation in the European aspen (*Populus tremula* L.), a widespread forest tree. *Molecular Ecology*, **19**, 1638–1650.

Delaneau O, Coulonges C, Zagury JF (2008) Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, **9**, 540.

Eid J, Fehr A, Gray J *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Eveno E, Collada C, Guevara MA *et al.* (2008) Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. *Molecular Biology and Evolution*, **25**, 417–437.

Everett M, Grau E, Seeb J (2011) Short reads and non-model species: exploring the complexities of next generation sequence assembly and SNP discovery in the absence of a reference genome. *Molecular Ecology Resources*, **11** (Suppl. 1), 93–108.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Foll M, Gaggiotti O (2006) Identifying the environmental factors that determine the genetic structure of Populations. *Genetics*, **174**, 875–891.

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**, 977–993.

Freamo H, O'Reilly P, Berg P, Lien S, Boulding E (2011) Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. *Molecular Ecology Resources*, **11** (Suppl. 1), 243–256.

Gaggiotti OE, Bekkevold D, Jorgensen HBH *et al.* (2009) Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: atlantic herring as a case study. *Evolution*, **63**, 2939–2951.

Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms to non-model species: a technical review. *Molecular Ecology Resources*, **10**, 915–934. (doi:10.1111/j.1755-0998.2010.02891.x).

Glover KA, Hansen MM, Lien S, Als TD, Høyheim B, Skaala Ø (2010) A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genetics*, **11**, 2–12.

Goudet J (1995) fstat version 1.2: a computer program to calculate *F*-statistics. *Journal of Heredity*, **86**, 485–486.

Goudet J (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, **5**, 184–186.

Guillot G, Foll M (2009) Correcting for ascertainment bias in the inference of population structure. *Bioinformatics*, **25**, 552–554.

Guillot G, Santos F (2009) A computer program to simulate multilocus genotype data with spatially auto-correlated allele frequencies. *Molecular Ecology Resources*, **9**, 1112–1120.

Guillot G, Mortier F, Estoup A (2005) GENELAND: a computer package for landscape genetics. *Molecular Ecology Notes*, **5**, 712–715.

Guillot G, Leblois R, Coulon A, Frantz AC (2009) Statistical methods in spatial genetics. *Molecular Ecology*, **18**, 4734–4756.

Harismendy O, Ng PC, Strausberg RL *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology*, **10**, R32.

Hauser L, Baird M, Hilborn R, Seeb L, Seeb J (2011) An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (*Oncorhynchus nerka*) population. *Molecular Ecology Resources*, **11** (Suppl. 1), 150–161.

Hayes B, Laerdahl J, Lien S *et al.* (2007) An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences. *Aquaculture*, **265**, 82–90.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution*, **59**, 1633–1638.

Hemmer-Hansen J, Nielsen E, Meldrup D, Mittelholzer C (2011) Identification of single nucleotide polymorphisms in candidate genes for growth and reproduction in a nonmodel organism; the Atlantic cod, Gadus morhua. *Molecular Ecology Resources*, **11** (Suppl. 1), 71–80.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresk WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLOS Genetics*, **6**, e1000862.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Reviews Genetics*, **10**, 639–650.

Hudson M (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Joost S, Kalbermatten M, Bonin A (2008) Spatial analysis method (SAM): a software tool combining molecular and environmental data to identify candidate loci for selection. *Molecular Ecology Resources*, **8**, 957–960.

Jost L (2008) G_{ST} and its relatives do not measure differentiation. *Molecular Ecology*, **17**, 4015–4026.

Kaeuffer R, Reale D, Coltman DW, Pontier D (2007) Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity*, **99**, 374–380.

Kalinowski ST (2002) How many alleles per locus should be used to estimate genetic distances? *Heredity*, **88**, 62–65.

Karlsson S, Moen T, Lien S, Glover K, Hindar K (2011) Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP-chip. *Molecular Ecology Resources*, **11** (Suppl. 1), 236–242.

Kauer M, Dieringer D, Schlotterer C (2003) Nonneutral admixture of immigrant genotypes in African *Drosophila melanogaster* populations from Zimbabwe. *Molecular Biology and Evolution*, **20**, 1329–1337.

Kerstens HH, Crooijmans RP, Veenendaal A *et al.* (2009) Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics*, **10**, 479–489.

Korbel JO, Urban AE, Affourtit JP *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.

Krawczak M (1999) Informativity assessment for biallelic single nucleotide polymorphisms. *Electrophoresis*, **20**, 1676–1681.

Lao O, van Duijn K, Kersbergen P, de Knijff P, Kayser M (2006) Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *American Journal of Human Genetics*, **78**, 680–690.

Laval G, Patin E, Barreiro LB, Quintana-Murci L (2010) Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS ONE*, **5**, e10284.

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.

Li R, Fan W, Tian G *et al.* (2010) The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317.

Liu N, Chen L, Wang S, Oh C, Zhao H (2005) Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics*, **6**(Suppl. 1), S26.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, **4**, 981–994.

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology and Evolution*, **20**, 136–142.

Marshall JM, Weiss RE (2006) A Bayesian heterogeneous analysis of variance approach to inferring recent selective sweeps. *Genetics*, **173**, 2357–2370.

Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations GENETICS, **166**(1): 351–372.

Maynard Smith J, Haigh J (1974) Hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.

Metzker M (2009) Sequencing in real time. *Nature Biotechnology*, **27**, 150–151.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, **17**, 240–248.

Moen T, Hayes B, Nilsen F *et al.* (2008) Identification and characterisation of novel SNP markers in Atlantic cod: evidence for directional selection. *BMC Genetics*, **9**, 18.

Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theoretical and Applied Genetics*, **120**, 1525–1534.

Morcillo-Suarez C, Alegre J, Sangros R *et al.* (2008) SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics analyses upon SNP data. *Bioinformatics*, **24**, 1643–1644.

Morin PA, McCarthy M (2007) Highly accurate SNP genotyping from historical and low-quality samples. *Molecular Ecology Notes*, **7**, 937–946.

Morin PA, Luikart G, Wayne RK, SNP_Workshop_Group (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, **19**, 208–216.

Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, **9**, 66–73.

Namroud MC, Beaulieu J, Juge N, Laroche J, Bousquet J (2008) Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce. *Molecular Ecology*, **17**, 3599–3613.

Nicholson G, Smith AV, Jónsson F *et al.* (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 695–715.

Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**, 931–942.

Nielsen R, Signorovitch J (2003) Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theoretical Population Biology*, **63**, 245–255.

Nielsen R, Hubisz MJ, Clark AG (2004) Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics*, **168**, 2373–2382.

Nielsen EE, Hansen MM, Meldrup D (2006) Evidence of microsatellite hitch-hiking selection in Atlantic cod (*Gadus morhua* L.): implications for inferring population structure in non-model organisms. *Molecular Ecology*, **15**, 3219–3229.

Nielsen EE, Hemmer-Hansen J, Poulsen NA *et al.* (2009a) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology*, **9**, 276.

Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009b) Population genomics of marine fishes: identifying adaptive variation in space and time. *Molecular Ecology*, **18**, 3128–3150.

Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, **40**, 646–649.

Palumbi SR, Baker CS (1994) Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. *Molecular Biology and Evolution*, **11**, 426–435.

Paschou P, Ziv E, Burchard EG *et al.* (2007) PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics*, **3**, 1672–1686.

Patterson N, Price A, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, e190.

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A (2004) GeneClass2: a Software for Genetic Assignment and First-Generation Migrant Detection. *Journal of Heredity*, **95**, 536–539.

Polanski A, Kimmel M (2003) New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, **165**(1): 427–436.

Ponting CP, Oliver PL, Reik W (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.

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 *et al.* (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, **81**, 559–575.

Reeves PA, Richards CM (2009) Accurate Inference of Subtle Population Structure (and Other Genetic Discontinuities) Using Principal Coordinates. *PLoS ONE*, **4**, e4269.

Robertson A (1975) Remarks on the Lewontin-Krakauer test. *Genetics*, **80**, 396–396.

Romero IG, Manica A, Goudet J, Handley LL, Balloux F (2009) How accurate is the current picture of human genetic variation? *Heredity*, **102**, 120–126.

Rosenberg N, Li L, Ward R, Pritchard J (2003) Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics*, **73**, 1402–1422.

Rosenblum EB, Novembre J (2007) Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. *Journal of Heredity*, **98**, 331–336.

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Molecular Ecology Resources. **8**, 103–106.

Rusk N (2009) Cheap third-generation sequencing. *Nature Methods*, **6**, 244–245.

Ryman N, Leimar O (2008) Effect of mutation on genetic differentiation among nonequilibrium populations. *Evolution*, **62**, 2250–2259.

Ryman N, Palm S (2006) POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Molecular Ecology Notes*, **6**, 600–602.

Sabeti PC, Varilly P, Fry B *et al.* (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–919.

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, **78**, 629–644.

Schlötterer C (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics*, **160**, 753–763.

Schlötterer C, Harr B (2002) Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in *Drosophila melanogaster*. *Molecular Ecology*, **11**, 947–950.

Shendure J, Ji H (2009) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.

Smith CT, Antonovich A, Templin WD, Elfstrom CM, Narum SR, Seeb LW (2007) Impacts of marker class bias relative to locus-specific variability on population inferences in Chinook salmon: a comparison of single-nucleotide polymorphisms with short tandem repeats and allozymes. *Transactions of the American Fisheries Society*, **136**, 1674–1687.

Smith M, Pascal C, Grauvogel Z, Habicht C, Seeb J, Seeb L (2011) Multiplex preamplification PCR and microsatellite validation allows accurate single nucleotide polymorphism (SNP) genotyping of historical fish scales. *Molecular Ecology Resources*, **11** (Suppl. 1), 257–266.

Stephens M, Smith N, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

Storz JF, Kelly JK (2008) Effects of spatially varying selection on nucleotide diversity and linkage disequilibrium: insights from deer mouse globin genes. *Genetics*, **180**, 367–379.

Topchy A, Scibner K, Punch W (2004) Accuracy-driven loci selection and assignment of individuals. *Molecular Ecology Notes*, **4**, 798–800.

Vasemägi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology*, **14**, 3623–3642.

Vitalis R, Dawson K, Boursot P (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics*, **158**, 1811–1823.

Vitalis R, Dawson K, Boursot P, Belkhir K (2003) DetSel 1.0: a computer program to detect markers responding to selection. *Journal of Heredity*, **94**, 429–431.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A Map of Recent Positive Selection in the Human Genome. *PLoS Biology*, **4**, e72.

Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K (2001) The discovery of single-nucleotide polymorphisms – and inferences about human demographic history. *American Journal of Human Genetics*, **69**, 1332–1347.

Waples R (2010). Perspective. High grading bias: subtle problems with assessing power of selected subsets of loci for population assignment. *Molecular Ecology*, **19**, 2599–2601.

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

White C, Selkoe KA, Watson J, Siegel DA, Zacherl DC, Toonen RJ (2010) Ocean currents help explain population genetic structure. *Proceedings of the Royal Society B-Biological Sciences*, **277**, 1685–1694.

Woolfe A, Goodson M, Goode DK *et al.* (2005) Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biology*, **3**, e7. doi:10.1371/journal.pbio.0030007.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Xu H, Sarkar B, George V (2009) A new measure of population structure using multiple single nucleotide polymorphisms and its relationship with FST. *BMC Research Notes*, **2**, 21.

Xu S, Gupta S, Jin L (2010) PEAS V1.0: a package for elementary analysis of SNP data. *Molecular Ecology Resources*, **10**, 1085–1088. Online Early.