INVITED REVIEWS AND SYNTHESES

Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools

FRANCE DUFRESNE,* MARC STIFT,† ROLAND VERGILINO‡ and BARBARA K. MABLE§

*Département de Biologie, Université du Québec à Rimouski, Québec, QC, Canada, G5L 3A1, †Department of Biology, University of Konstanz, Konstanz, D 78457, Germany, ‡Department of Integrative Biology, University of Guelph, Guelph, ON, Canada, N1G 2W1, §Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK

Abstract

Despite the importance of polyploidy and the increasing availability of new genomic data, there remain important gaps in our knowledge of polyploid population genetics. These gaps arise from the complex nature of polyploid data (e.g. multiple alleles and loci, mixed inheritance patterns, association between ploidy and mating system variation). Furthermore, many of the standard tools for population genetics that have been developed for diploids are often not feasible for polyploids. This review aims to provide an overview of the state-of-the-art in polyploid population genetics and to identify the main areas where further development of molecular techniques and statistical theory is required. We review commonly used molecular tools (amplified fragment length polymorphism, microsatellites, Sanger sequencing, next-generation sequencing and derived technologies) and their challenges associated with their use in polyploid populations: that is, allele dosage determination, null alleles, difficulty of distinguishing orthologues from paralogues and copy number variation. In addition, we review the approaches that have been used for population genetic analysis in polyploids and their specific problems. These problems are in most cases directly associated with dosage uncertainty and the problem of inferring allele frequencies and assumptions regarding inheritance. This leads us to conclude that for advancing the field of polyploid population genetics, most priority should be given to development of new molecular approaches that allow efficient dosage determination, and to further development of analytical approaches to circumvent dosage uncertainty and to accommodate 'flexible' modes of inheritance. In addition, there is a need for more simulation-based studies that test what kinds of biases could result from both existing and novel approaches.

Keywords: hybridization, mixed modes of reproduction, next-generation sequencing, polyploidy

Introduction

Polyploidy is a prominent feature of plant genomes (Tate *et al.* 2005). Although polyploidy is much rarer in the animal kingdom than in plants, there are numerous examples of polyploid invertebrates, fish and amphibians (Gregory & Mable 2005; Mable *et al.* 2011). Even organisms that are now genetically diploid often have a paleopolyploid history. In plants and yeast, early genome-sequencing projects revealed that numerous diploid species show signs of ancient genome duplications (Arabidopsis, Blanc *et al.* 2000; rice, Bowers *et al.* 2003; yeast, Kellis *et al.* 2004; poplar, Tuskan *et al.* 2006; grapevine, Jaillon *et al.* 2007). In animals, whole-genome duplication events have coincided with the origin of vertebrates,

Correspondence: France Dufresne, Fax: 418 724 1849;
E mail: france dufresne@uqar.qc.ca

gnathostomes and teleosts (Holland *et al.* 1994; Postlethwait *et al.* 2000; Crow *et al.* 2006). A whole-genome duplication event is thought to have facilitated the survival of flowering plant lineages during the mass extinction events during the Cretaceous-Tertiary transition (Fawcett *et al.* 2009). This has led to the generally accepted view that polyploidization plays an important role in evolution, in both plants and animals.

Despite the important role of polyploidization in evolution, our basic understanding of polyploids is still poor compared with diploids. This is largely due to the more complex nature of their genome evolution. Polyploids are typically classified as either autopolyploids or allopolyploids (Stebbins 1947). Autopolyploids originate after genome doubling within a single species, so that each chromosome is represented by more than two homologous copies. These homologous copies theoretically can at least initially pair in all possible combinations, leading to polysomic inheritance. However, even in autopolyploids divergence, neo-functionalization, or loss of duplicate copies over time (Lynch & Conery 2000) inevitably leads to disomic inheritance for at least some loci (Ohno 1970). Allopolyploids originate after hybridization of different species and subsequent genome doubling so that each chromosome is represented by two (or more) sets of divergent chromosomes, in which chromosomes within a set are termed homologues, and chromosomes from different sets (i.e. derived from different ancestral species) homoeologues (see Box 1). With sufficient divergence between homoeologues, meiotic pairing only takes place between chromosomes from the same parental origin, leading to disomic inheritance. In cases for which the homoeologous chromosomes can pair in meiosis and produce viable gametes, allopolyploids also may show a mixture of disomic and polysomic inheritance patterns. Moreover, inheritance patterns can vary across the genome within individuals, leading to disomic inheritance at some loci and polysomic at others. Due to the time and expense of assessing segregation within progeny arrays for every locus and every individual or species compared, it has not been quantified how frequently deviations from strictly disomic or strictly polysomic inheritance occur. However, where segregation has been tested, it is rare to find either extreme across all loci. For example, the family Salmonidae originated through polyploidization, but allozyme data originally suggested that inheritance patterns can vary between species, within species or even among tissue types within individuals (Danzmann & Bogart 1982; Allendorf & Danzmann 1997). Similar conclusions about deviations from strictly disomic or strictly polysomic inheritance have been described for plants (Jannoo *et al.* 2004; Stift *et al.* 2008; Kamiri *et al.* 2011; Koning-Boucoiran *et al.* 2012).

The existence of complex inheritance patterns complicates the genetic analysis of polyploids, because analytical frameworks normally assume a specific mode of inheritance. Assumptions about inheritance patterns are important because expected dosage of alleles (i.e. copy number of each allele) at individual loci will differ depending on the mode of segregation and models predicting the rate of loss or change of duplicate genes depend on the degree of redundancy of duplicate copies (Ohno 1970; Ferris & Whitt 1977; Allendorf 1978). This introduces both conceptual (e.g. how many alleles and gene copies are to be expected) and methodological (e.g., resolving allele and gene copy numbers) issues with obtaining markers for population genetic analyses. A major challenge for most existing markers used for population genetic analyses is reliably resolving dosage of alleles in polyploids and so enabling calculation of observed and expected allele frequencies, which is fundamental to many population genetic based inferences (Cockerham 1973; Kreitman 1987). Continuing advances in sequencing technology mean that it should soon be possible to consider genome-wide variation in segregation patterns, but most population genetics models must currently be applied in the absence of knowledge about segregation, expected dosage, and allele or gene copy number.

In addition, variation in mode of segregation patterns can make it difficult to disentangle the effects of genome duplication from hybridization. For allopolyploids, analyses would be most robust if copies from each parent could be identified and treated separately during analysis of genetic variation. However, past genome duplication events make it difficult to distinguish true single nucleotide polymorphisms (SNPs) or orthologous allelic copies from fixed differences between homoeologous duplicate chromosomal regions and from tandemly-duplicated paralogous regions (Everett *et al.* 2011; Seeb *et al.* 2011b). This is confounded by the difficulty of resolving whether polyploid lineages have arisen through allo- or autopolyploidization.

Although many polyploid fish, amphibians (Bogart 1980; Otto & Whitton 2000) and plants (Suomalainen *et al.* 1987) reproduce sexually, an additional complexity arises due to the frequent association of specific reproductive systems with polyploidy. In the animal kingdom, the majority of polyploid invertebrates and reptiles reproduce asexually, and it has been estimated that 99% of apomictic plant species are polyploids (Suomalainen *et al.* 1987). In some cases, such as found in the planarian flatworm, *Schmidtea polychroa*, polyploid individuals can produce viable sperm that may lead to rare sexual processes (Sánchez-Navarro *et al.* 2013). As asexually reproducing plants and animals often have uneven ploidy levels (e.g. triploid) but coexist with even ploidy (e.g. diploid or tetraploid) individuals that reproduce sexually (Neiman *et al.* 2011), a substantial challenge is to include multiple ploidy levels with different expected heterozygosities (due to differences in both allelic dosage and mating system) into the same population genetic analyses, particularly for inferences that rely on accurate estimation of allele frequencies.

The main aim of this review is to provide an overview of the molecular and statistical tools that are currently available for polyploid population genetics, to provide examples of their application, and to identify the main areas where further development of molecular techniques and statistical theory is required to advance the field. Our review is organized into two sections. The first section deals with the issue of obtaining informative markers for polyploids. We first discuss the application of traditional markers [amplified fragment length polymorphism (AFLP), microsatellites, Sanger sequencing] in polyploids, and their pros and cons. We then show that new sequencing technologies still suffer from similar problems as traditional markers and introduce some of their own, but do hold promise for ultimately reducing these problems. The second section focuses on the analytical side and deals with the problem of extending standard methodologies for diploids to polyploid data. We discuss how classical approaches (allele frequency estimation, assignment and clustering methods, fixation indices, similarity/distance indices and multivariate analyses, custom models) can be used with polyploid data and identify priorities for further development of methodology and software. In particular, we conclude that there is a strong need for simulations to evaluate the appropriateness of the various creative solutions that have been proposed for analysing polyploid data.

Box 1
Glossary

Allelic dosage Number of copies of each allele at a particular locus in a polyploid genotype.

Allopolyploid Polyploid that has originated by genome doubling after hybridization, so that two homoeologous sets of the same chromosome exist. The dogma is that this generally leads to disomic inheritance, because there is preferential pairing between chromosomes from the same ancestral genome. However, polysomic inheritance is often still possible, at least at some loci or chromosomal regions.

Autopolyploid Polyploid that has been originated by genome doubling within a species, so that all variants of the same chromosome are homologous. The dogma is that this generally leads to polysomic inheritance, because there is no preferential pairing between certain chromosomes. However, as genome doubling inevitably leads to divergence among copies, specialization of function, or loss of copies, a return to disomic inheritance is predicted over time. Hybridization between closely related species or differentiated populations of the same species (sometimes referred to as segmental allopolyploidy) can be difficult to distinguish from autopolyploidy, but it is expected that there will be at least some disomic inheritance.

Disomic inheritance Type of inheritance typical for allopolyploids due to preferential pairing between the chromosomes derived from the same ancestral species. This means that alleles derived from the same ancestral species segregate as for diploids, so offspring receive only one copy from a given parent. There is thus not expected to be recombination between the copies derived from the different parents (i.e. homoeologues).

Double reduction Meiotic process in polyploids with polysomic inheritance in which recombination takes place between the locus and centromere and sister chromatids migrate to the same pole (i.e. segregate in the same gamete).

Homoeologues Divergent loci or chromosomes in allopolyploid genomes that usually do not pair together during meiosis because they are derived from different parental lineages.

Homologues Loci or chromosomes that usually pair together during meiosis because they are derived from the same parental lineage.

Orthologues Gene copies that diverged after a speciation event.

Paralogues Gene copies that diverged after a gene or genome duplication event.

Partial heterozygote In diploids, genotypes for a given locus can be homozygotes (e.g. AA, BB, CC) or heterozygotes (e.g. AB, AC, BC, CD). In polyploids, genotypes can be homozygotes (e.g. AAAA, BBBB, CCCC), full heterozygotes (e.g. ABCD, ABFG, CDEF) or partial heterozygotes where one or more alleles are present multiple times (e.g. ABBC, ABFF, ABBB). Resolving partial heterozygotes is one of the biggest challenges for applying population genetics approaches to polyploids, for the majority of existing methods.

Null allele An allele that fails to amplify using locus-specific primers or that is not observed due to incomplete sampling (e.g. not enough clones sequenced or not enough coverage during deep sequencing).

Polysomic (or multisomic) inheritance type of inheritance typical for autopolyploids, where all variants of the same chromosome can pair in meiosis. This means that parental alleles will be combined in the same gamete in all possible combinations. Depending on the position of the locus relative to the centromere, a maximum of one-sixth of the gametes can be the result of double reduction.

Stutter bands artefacts due to replication slippage during the PCR amplification of highly repetitive sequences (e.g. microsatellites), visible as one or more shadow bands, or one or multiple repeat lengths shorter or longer than the actual allele length.

Molecular genetic and genomic markers for polyploid population genetics

General caveats for genetic marker analysis in polyploids

Molecular markers that are standardly used for population genetics in diploids can in principle also be used in polyploids. However, one of the most important challenges when working with polyploid genomes is the difficulty of resolving the allelic constitution of individual loci (i.e. allelic dosage), which would be necessary to implement methods that rely on allele frequency-based inferences or those that require complete genotyping of individuals. Uncertainties in dosage can also compound problems associated with homoplasy due to null alleles or artefacts associated with either replication slippage (e.g. stutter bands) or unequal amplification of alleles of different lengths (e.g. allelic dominance) in markers requiring PCR amplification; as the number of alleles at a locus could vary from 1 to *k* in a *k*-ploid, detecting alleles that either do not amplify consistently or 'extra' alleles is not straightforward for ploidy levels higher than diploid (*k* 2). Most tests for detecting such artefacts are based on Hardy Weinberg (HW) equilibrium (MICROCHECKER, Van Oosterhout *et al.* 2004, 2006), but complete dosage information would be required to calculate expected allele and genotype frequencies. In addition, as many polyploids also show a shift to self-fertilization (Mable 2004a) or reproduce asexually (Stenberg & Saura 2013), tests that assume HW equilibrium also would not be useful for detecting homoplasy in these cases.

The presence of an uncertain number of allelic copies could also be problematic for sequence-based analyses; for example, in tests for selection where the relative frequency of particular alleles is informative or in calculation of inbreeding coefficients based on observed and expected heterozygosity (which would of course also apply to codominant markers). In addition, if there is sufficient divergence among duplicated copies that the different sets (homoeologues) segregate independently, then analyses that cannot distinguish between homoeologues could result in inaccurate inferences about population genetic structure and levels of genetic diversity.

In this section, we will discuss the implications of these general issues as well as specific problems or benefits associated with applying the most commonly used markers for population genetics to polyploid genomes. We have divided this into 'traditional markers' (AFLPs, microsatellites, Sanger sequencing) and 'new markers' (rapidly advancing deep sequencing approaches).

Traditional markers

AFLP. Amplified fragment length polymorphism fingerprinting has been popular in population genetics, but especially in plants (Bensch & Åkesson 2005), where the frequency of polyploidy is high (Masterson 1994). It is attractive because a single fingerprint includes information for a large number of anonymous nuclear markers that are assumed to be scattered over the entire genome (Meudt & Clarke 2007). A disadvantage compared with codominant markers such as microsatellites (see below 1.2) is that AFLP markers are dominant (i.e. they contain no direct information on heterozygosity), which could actually be an advantage when working with polyploids, to avoid problems with dosage uncertainty.

A further attractive feature of AFLPs is that fingerprints can in principle be simultaneously generated for diploids and polyploids, thus allowing interploidal comparisons. For this reason, AFLPs have frequently been used to reconstruct origins of allopolyploids (e.g. in Dactylorhiza, Hedrén *et al.* 2001; Achillea, Guo *et al.* 2005; and Ranunculus, Paun *et al.* 2006) and for the analysis of population structure and Analysis of Molecular Variance (e.g. in polyploid Knautia, Kolář *et al.* 2012; and alpine Ranunculus, Burnier *et al.* 2009). However, these applications have revealed a potential drawback that AFLPs in polyploids tend to produce higher numbers of AFLP fragments than diploids (reviewed by Fay *et al.* 2005; Meudt & Clarke 2007). AFLP markers are prone to homoplasy (comigration of nonhomologous fragments), which increases in proportion to the total number of AFLP bands (Caballero & Quesada 2010).

AFLPs in species with larger genomes (higher ploidy levels) also frequently result in a small number of high-intensity fragments and many low-intensity fragments that are difficult to score, which effectively results in a relatively high frequency of null alleles. These phenomena have been attributed to repetitive elements related to retrotransposon activity (Fay *et al.* 2005), but it remains to be tested if they could cause any bias. Nevertheless, the sheer abundance of informative markers that AFLPs can generate appears to outweigh potential scoring issues. Hence, we conclude that AFLPs provide a powerful source of information for addressing questions related to origins of allopolyploids and population genetic structure.

Microsatellites (simple sequence repeats). In population genetics, microsatellites are an attractive alternative to dominant AFLPs, because they are by nature codominant. This means that they allow (at least in diploids) directly distinguishing between heterozygotes and homozygotes, which is important for inferring levels of inbreeding and using allele frequency-based inferences. Typical applications of microsatellites involve the analysis of population structure, genetic diversity and population differentiation (Sunnucks 2000). Moreover, if one is willing to assume certain models of repeat evolution, microsatellite data can be used to calculate migration rates or to reconstruct geneaeologies, which can be used to test models of demographic history based on coalescent models (e.g. Beaumont 1999). Next-generation sequencing (NGS) technologies now allow the efficient identification of large numbers of microsatellites at a fraction of the cost and effort of traditional approaches, so these markers will probably remain popular for population genetics studies, despite continuing advances in technology.

In polyploids, inability to reliably utilize codominant scoring reduces the usefulness of microsatellites relative to diploids and to AFLPs. The nature of the problem is best illustrated with an example. A tetraploid genotyped with three different alleles scored at a microsatellite locus could have three possible genotypes: AABC, ABBC or ABCC. If there is a null allele that does not amplify, the true genotype could be ABCX. Homoplasy could also result if there are stutter bands caused by replication slippage during the PCR process, which could make it look like the genotype was ABCD, when in fact D is not a true allele. Which genotype is correct would affect the allele frequency distribution of the alleles and in turn inferences about population genetic structure. Theoretically, allelic configurations for microsatellites could be resolved based on the ratios between peak intensities to determine the relative number of copies of each allele (MAC-PR method: Esselink *et al.* 2004), but in practice, this has only proved feasible in cases where segregation analyses within families were used to confirm dosage patterns; for example, in *Rosa* × *hybrida* (Esselink *et al.* 2004), *Thymus praecox* (Landergott *et al.* 2006) and *Rorippa amphibia* (Luttikhuizen *et al.* 2007). Such segregation data are essential to reliably resolve the exact allelic configuration based on peak intensities but are rarely performed in practice due to the extra samples, time and effort required to perform the tests for families from each individual or even each population sampled. In addition, segregation data cannot be obtained in asexual polyploids. This effectively means that codominant microsatellite data have to be treated as dominant, which reduces the information content and precludes analyses that take into account observed heterozygosity of individuals or allele frequency distributions.

Null alleles are a further problem for use of microsatellites in polyploids. Null alleles of course form a general problem in population genetics for codominantly scored molecular markers (irrespective of ploidy), because they lead to an overestimation of homozygosity (e.g., see Dakin & Avise 2004). The risks could be magnified in polyploids (particularly allopolyploids) for several reasons. First, loci developed for one species may not amplify equally well in other species. This is a general problem regardless of ploidy level when distantly related taxa are compared with markers developed in only one of the taxa. However, allopolyploid taxa combine multiple diverged genomes in a single individual, so that even population genetic comparisons within a single species may be affected by null alleles. The severity of the problem depends on the degree of similarity between the homoeologues (Röder *et al.* 1995; McQuown *et al.* 2002). Second, polyploidization and hybridization often lead to increased transposon activity and sequence loss due to genomic rearrangements (Parisod *et al.* 2009), which could affect primer binding sites. Third, the presence of multiple alleles at each locus increases the chances of differential amplification of alleles (i.e. allelic dominance; Vergilino *et al.* 2009). This makes the problem of not being able to test for the presence of null alleles problematic for polyploids, particularly when combined with dosage uncertainty.

Despite the complications associated with genotyping, microsatellites have been used to analyse population structure and address phylogeographic questions in polyploids. For example, dominantly scored microsatellites have been used to identify a cryptic invasive European lineage of hexaploid reed *Phragmites australis* in North America (Saltonstall 2003), to infer that multiple genotypes of the red alga *Asparagoformis taxiformis* have invaded the Mediterranean Sea (Andreakis *et al.* 2009) and that clonal diversity has increased in refugial island populations of octoploid prune tree *Prunus lusitanica* (García-Verdugo *et al.* 2013). In the relatively few cases where dosage has been determined reliably, microsatellites have provided powerful markers for polyploid population genetics and have the ability to include diploids and polyploids in the same analysis. For example, in a phylogeographic study of hawthorn

(*Crataegus*), complete genotypes were resolved using peak ratios (Esselink *et al.* 2004) and used to show that diploid sexuals were more diverse than triploid apomicts (Lo *et al.* 2009). Codominantly scored microsatellites have also been used to show that *Rorippa amphibia* autotetraploid plants have higher genetic diversity than diploids, exactly matching predictions based on the larger effective population size of tetraploids (Luttikhuizen *et al.* 2007). In cases where resolving dosage is unrealistic (which is probably the case for ploidy levels higher than tetraploid), it is questionable if the increased information content per locus (i.e. multiple allelic states that can be identified) outweighs the loss of marker number compared with AFLPs and the increased risks of artefacts caused by null alleles and homoplasy. Although microsatellites are widely used, they cannot be used to their full potential in polyploids unless segregation is tested at each locus or until analytical solutions that can implement dosage uncertainty are adequately tested. With future developments in NGS technologies, the sequencing of microsatellite alleles may someday replace current genotyping methods and allow the characterization of hundreds of individuals at thousands of loci (Guichoux *et al.* 2011). This would reduce the influence of homoplasy, provided that sequencing errors are minimized by bioinformatics treatment.

Sanger sequencing. A major advantage of using DNA sequences for population genetics compared with fragment-based analyses is that complex substitution models can be fit to the data (e.g. Swofford *et al.* 1996), which allows application of more rigorous tests of demographic history, genealogical relationships, migration rates, recombination and selection (e.g. Rozas & Rozas 1999). Different regions of DNA evolve at different rates and so can be used to address questions from relatedness among individuals to deep species relationships. For example, introns and noncoding sequences tend to evolve at a faster rate than coding regions and so can be useful for examining close relationships; analysis of SNPs across a wide range of genes has the potential to increase fine-scale resolution compared with focusing on single genes. In theory, models of evolution based on sequences can be extended to polyploids, as long as complete information can be obtained about nucleotide substitution patterns, heterozygosity and allele frequencies.

A disadvantage of using nuclear DNA sequences for analyses that rely on resolving patterns of allele sharing and observed heterozygosity is that even in diploids it is often difficult to resolve the phase of substitutions, meaning that labour-intensive cloning is required to determine the exact allelic composition in heterozygotes (Zhang & Hewitt 2003). Cloning is also required if heterozygotes include sequences of different lengths. Even for diploids it can also be difficult to distinguish paralogues (i.e. alleles arising from gene duplications) from orthologues (i.e. alleles that have arisen through common descent at a single locus) in gene families. These problems are exacerbated in polyploids due to the increase in the number of possible alleles at a locus, unknown copy number of genes, and reticulate evolution in allopolyploids.

As most polyploids undergo some degree of diploidization following the initial genome duplication event, there can be random losses of gene copies in different taxa or even in different individuals from the same taxa, leading to widespread presence absence variation and copy number variation (CNV; Griffin *et al.* 2011). This makes resolution of phylogenetic trees and population genetic inferences difficult if orthologues cannot be reliably distinguished from paralogues. In allopolyploids, if there is high sequence conservation among parental copies, there is the added difficulty of identifying homoeologues, and origins through hybridization mean that assumptions of strictly bifurcating models of evolution are violated. One approach would be to focus on genes that do not remain duplicated in polyploids, but this in itself might be evidence that such genes are under selection, and so not strictly appropriate for population genetic tests that assume neutrality. Alternatively, network-based approaches that allow reticulation, such as SplitsTree (Huson & Bryant 2006), are frequently used to resolve origins and phylogenies of polyploids based on nuclear gene sequences (e.g. Schmickl *et al.* 2008; Brysting *et al.* 2011; Talavera *et al.* 2013). This approach can reduce problems associated with duplicate gene copies as well as hybridization if paralogues can be resolved based on phylogenetic clustering and then analysed separately by designing paralogue-specific primers (e.g. Evans *et al.* 2011).

Except for plastid DNA (mitochondria and chloroplasts) and ribosomal RNA repeats (which are both present in high copy number in each cell), traditional Sanger sequencing has required either a PCR or cloning step, with PCR the most popular since the early 1990s (Swofford *et al.* 1996). However, this means that DNA sequencing suffers from some of the same problems as PCR-based fragment analyses (e.g. microsatellites): lack of ability to determine allelic dosage; uneven amplification of alleles; and possibility of null alleles. In addition, increasing the number of alleles at a locus and/or the number of gene copies increases the risk of artefacts due to recombination during the PCR process, and cloning is nearly always required if there are more than two alleles at a locus. Although the proportion of clones of a particular allele could be used as an indication of its relative dosage, this would require even amplification of each allele; there is also a risk of missing alleles (i.e. null alleles) if some alleles amplify less strongly than others and if insufficient numbers of clones are sequenced. Particularly for polyploids arising through hybridization, a substantial challenge when using cloning is to distinguish real recombinants among parental copies from PCR-based artefacts (e.g. Jørgensen *et al.* 2012). However, with sufficient effort, even complex gene families can be resolved and interpreted in polyploids using segregation analyses and cloning (Mable *et al.* 2004). Thus, the problem is not as fundamentally insurmountable as for microsatellites.

Despite these caveats, DNA sequencing has revealed important insights into polyploid evolution and still holds the greatest potential for population genetic inferences. It was in allopolyploid cotton that it was first discovered that ribosomal gene arrays, which had been assumed to evolve under complete concerted evolution so that every copy in an individual is identical in sequence (Hillis & Davis 1988), could include multiple sequence types (Wendel *et al.* 1995). Furthermore, it was demonstrated that copies could be present from either parent or both and that this could vary by individual. There have been extensive studies investigating phylogeography in closely related diploids and polyploids using plastid sequences for both animals (e.g. Ptacek *et al.* 1994; Ludwig *et al.* 2001; Tsigenopoulos *et al.* 2002; Stenberg *et al.* 2003; Evans *et al.* 2004; Stöck *et al.* 2005; Culling *et al.* 2006; Lampert & Schartl 2008) and plants (e.g. Soltis *et al.* 1989; Brochmann *et al.* 1992; Van Dijk & Bakx-Schotman 1997; Segraves *et al.* 1999; Wu *et al.* 2010); because of their uniparental inheritance and lack of variation among copies within individuals, they can be treated as effectively equivalent in diploids and polyploids. Many studies have also combined nuclear and plastid sequence data to investigate complex evolutionary histories of polyploids in both plants (Soltis & Soltis 2000; Baumel *et al.* 2002; Huang *et al.* 2002; Schmickl *et al.* 2008; Ainouche *et al.* 2009; Krak *et al.* 2013) and animals (Evans *et al.* 2005; Holloway *et al.* 2006; Saitoh *et al.* 2010), and the combination of organelle and nuclear data can help to disentangle incomplete lineage sorting from past hybridization events (e.g. Vergilino *et al.* 2011). Some studies have combined plastid or nuclear genes with other types of markers such as AFLPs (e.g. Burnier *et al.* 2009; Ma *et al.* 2010) to resolve complex polyploid complexes. Given the rapid developments in sequencing technology, resolution of complete genotypes in polyploids should be achievable in the near future, but the fundamental issues related to interpreting sequence variation in duplicated genes (i.e. assigning alleles to loci, distinguishing phase, resolving copy number and allelic dosage, inferring recombination) remain a substantial challenge.

New markers

Rapid advances in technology enabling whole-genome perspectives on genetic variation hold great promise for increasing the range of inferences possible using polyploid genomes (reviewed by Aversano *et al.* 2012; Buggs *et al.* 2012; Egan *et al.* 2012; Madlung 2013) but cannot yet solve all of the issues with previous markers and introduce some of their own challenges. Researchers working on polyploid genomes have been at the forefront of advanced genomic approaches for understanding changes in gene expression, epigenetics and genome shock associated with hybridization and gene duplication (Ainouche & Jenczewski 2010; Stöck & Lamatsch 2013). Although this is at least partly due to the fact that many economically important crop plants (reviewed by Edwards *et al.* 2013) and fish (reviewed by Mable *et al.* 2011) are polyploid, important genome-scale insights have also been obtained from nonmodel organisms with intriguing evolutionary histories of recent polyploidy, such as *Spartina* (Ainouche *et al.* 2004; Salmon *et al.* 2005; Chelaifa *et al.* 2010; de Carvalho *et al.* 2013), *Senecio* (Hegarty *et al.* 2006, 2008, 2009) and *Tragopogon* (Soltis *et al.* 2004; Buggs *et al.* 2009, 2010, 2012).

While there has as yet been little focus on implications of polyploidy for population genomics, it will still be critical to resolve issues associated with gene duplication, allelic dosage, copy number variation, resolution of homoeologues, and recombination. In addition, reliable assembly of duplicated genes, repetitive sequences and highly divergent regions of polymorphism remains one of the largest challenges for whole-genome reconstruction and annotation; even genomes that are considered well-resolved (e.g. *Arabidopsis thaliana*) retain uncertainty in these types of regions. In addition, most NGS methods currently suffer from higher error rates than traditional Sanger sequencing, which can introduce additional biases; while this problem applies equally to diploids, dosage uncertainties again make the problem potentially more difficult to solve in polyploids. However, the major advantage is the overwhelming number of sequence-based characters available for population genetics analyses of nonmodel species and being able to take a genomewide perspective on consequences of introgression through hybridization, fate of duplicate genes, and patterns of selection and recombination.

Below we outline some of the main types of characters that have been used in population genomic approaches and discuss current strategies for dealing with polyploid genomes. A major difference with NGS approaches is that technology and analyses are advancing so quickly that there is not a ‘stable state’ of issues and solutions that can be applied as easily as for the older methods. We expect that it will soon be possible to apply the same types of population genetic analyses developed for traditional Sanger sequencing at a whole-genome scale, but it is the sheer volume of data that will be the biggest challenge for implementation. We thus concentrate the review on where we think the major challenges currently lie in generating the data, rather than making specific recommendations for application of population genetic models to NGS data obtained from polyploids.

Genome-wide SNP markers. Development of microarray technology was the first major advance in making genome-scale approaches accessible to ecological and evolutionary questions (e.g. Gibson 2002; Shiu & Borevitz 2006). Although microarrays have been applied to interesting questions related to gene expression in polyploids (Chen *et al.* 2004; Slotte *et al.* 2007; Buggs 2008; Hegarty *et al.* 2008, 2009; Mavarez *et al.* 2009; Chelaifa *et al.* 2010; Flagel & Wendel 2010; Pignatta *et al.* 2010; de Carvalho *et al.* 2013), a major issue is with unknown copy number changes between the individuals compared on the array, which could lead to spurious conclusions about expression differences. Although this could theoretically be corrected using DNA arrays to estimate copy number (Auer *et al.* 2007), inability to distinguish sequence divergence (i.e. preventing hybridization on the arrays) from loss of duplicated copies, could affect such interpretations (e.g. Parkin *et al.* 2010). Expression changes in allopolyploids can also be highly complex. For example, detailed studies using cDNA-AFLP approaches have clearly demonstrated that not only changes in gene expression but stochastic loss or over-representation of parental copies occur frequently in newly synthesized polyploids (e.g. Wang *et al.* 2006; Gaeta *et al.* 2007; Buggs *et al.* 2009, 2010; Jackson & Chen 2010). Thus, differences in hybridization of paralogues have represented an important challenge for microarray-based studies of changes in gene expression following polyploidization.

Transcriptome analyses using RNA-sequence hold more promise for distinguishing the evolutionary dynamics of duplicate genes, because they should not be as sensitive to bias in the representation of paralogous copies. As for all analyses of polyploids, emerging results are complex but intriguing (de Carvalho *et al.* 2013). Large genome size, large gene families and high repetitive sequence content remains problematic for genome and transcriptome assembly, particularly in nonmodel organisms (e.g. Vijay *et al.* 2013), but new approaches are constantly being developed that could improve resolution of polyploid genomes. For example, following up on microarray-based experiments (Flagel *et al.* 2008; Flagel & Wendel 2010; Salmon *et al.* 2010), Yoo *et al.* (2013) used Illumina technology to sequence the transcriptomes of wild and cultivated cotton to distinguish between expression changes due to biases in which parental genome is expressed in an allopolyploid and 'dominance' in the expression patterns from one parent (i.e. where hybrids show similar expression patterns to those in one parent, rather than preferentially expressing the allelic copy from one parent; reviewed by Buggs 2013). Such complications emphasize that even with advanced technology, phylogenetic and population genetic analyses of polyploids could remain problematic due to their biology, rather than just methodological issues.

Despite these issues, SNP arrays based on transcriptome analyses have led to useful insights into the population genetics of polyploid organisms (e.g. Atlantic salmon: Bourret *et al.* 2012). For example, based on 454 transcriptome sequencing of polyploid wheat genomes, Lai *et al.* (2012) modified a tool developed for SNP detection in diploid crop plants (AutoSNPdb) to enable integration of SNP and gene annotation information with a graphical viewer even for such highly complex genomes. In polyploid sturgeons, Hale *et al.* (2009) applied a rarefaction approach taken from theoretical ecology to assess the relationship between sequence coverage and gene discovery and discussed whether normalization is a useful approach to reduce coverage of repetitive sequences such as rRNA subunits. Normalization could be particularly problematic for polyploids because relative levels of gene expression among homoeologues are often of particular interest for understanding evolutionary and functional processes in polyploids and so important information might be lost through the standardization. In addition, if diploids and polyploids are included in the same analyses, it might not be possible to apply a single normalization strategy to all individuals, due to differences in relative coverage. Although some success has been achieved using distant diploid relatives as references (e.g. Everett *et al.* 2011), the current lack of sequenced polyploids also hinders assembly and resolution of SNPs for most polyploid genomes.

Continuing technological developments mean that genomic-based SNP generation is now also feasible, even in large polyploid genomes. However, problems with distinguishing between paralogous copies and the presence of high copy numbers of repetitive elements in many polyploids (Leitch & Leitch 2008; Koukalova *et al.* 2010; Buggs *et al.* 2012; Piednoël *et al.* 2012) mean that full-genome annotations remain challenging (e.g. Seeb *et al.* 2011a; Brenchley *et al.* 2012; Wang *et al.* 2012), reducing the potential to interpret population genomics patterns in the context of potential for selection. In some

Table 1 Software and statistical packages used in population genetics and population genomics studies on polyploid or mixed-ploidy level populations, including the type of polyploids for which they are applicable, whether or not they support large datasets (i e for analysis of next-generation sequence data, what types of markers they have been developed for, and the operating systems on which they can be run

Software	Type of polyploids	Supporting large datasets	Marker type	Operating system
Assembly, SNP discovery and genotyping				
CLC Bio genomic workbench http://www.clcbio.com/products/clc-genomics-workbench/	All	Yes	Sequences SNP	Mac OS X Windows Unix
Genome Analysis Tool Kit (gatk) http://www.broadinstitute.org/gatk/	All	Yes	Sequences SNP	Mac OS X Windows Unix
Stacks http://creskolab.uoregon.edu/stacks/	Diploids*	Yes	Sequences SNP	Mac OS X Linux
fitTetraR package http://www.wageningenur.nl/en/show/Software-fitTetra.html	Tetraploids	Yes	Bi-allelic	Mac OS X Windows Unix
superMassa http://statgen.esalq.usp.br/SuperMASSA/	All	Yes	SNP	Online
Distance-based methods				
popdist http://genetics.agrsci.dk/~bernt/popgen/	Asexuals (mixedploidies)	No	SSR	Mac OS X Windows Unix
Estimation of allele frequencies and *F*-statistics based methods				
polySegratio/polySegratioMM http://cran.r-project.org/web/packages/polySegratio/index.html http://cran.r-project.org/web/packages/polySegratioMM/index.html	Autopolyploids	Yes	SNP AFLP SSR	Mac OS X Windows Unix
atetra http://www.vub.ac.be/APNA/ATetra.html	Tetraploids	No	SSR	Windows
StAMPP r package http://cran.rproject.org/web/packages/StAMPP/index.html	Mixed ploidies	Yes	SNP	Mac OS X Windows Unix
Bayesian clustering methods				
Structure http://pritch.bsd.uchicago.edu/structure.html	Autopolyploids	Yes	SNP SSR	Mac OS X Windows Unix
InStruct http://cbsuapps.tc.cornell.edu/InStruct.aspx	Autopolyploids Allopolyploids	Yes	SNP SSR	MacOS X Windows Linux Online
Packages implementing multiple methods				
adegenet r package http://cran.r-project.org/web/packages/adegenet/	Various but no mixed ploidies	Yes	All	Mac OS X Windows Linux
PolySat r package http://openwetware.org/wiki/Polysat	Polysomic inheritance (mixed ploidies)	No	SSR	Mac OS X Windows Linux
SPAGeDi http://ebe.ulb.ac.be/ebe/Software.html	Autopolyploids	Yes	Dominant Codominant	Mac OS X Windows Linux/Unix
GenoType/GenoDive www.patrickmeirmans.com/software/GenoDive.html	Asexuals (mixed ploidies) Polysomic inheritance	Yes	Dominant Codominant	Mac OS X Windows

AFLP, amplified fragment length polymorphism; SNP, single nucleotide polymorphisms; SSR, simple sequence repeats

*Although this software is primarily adapted for diploids, some studies have used this software more or less successfully to analyse SNPs in polyploids (Ogden *et al* 2013; Wang *et al* 2013) instances, duplicated genes are intentionally excluded to simplify genomic assembly, with linkage maps based on only the nonduplicated portion of the genome (Everett *et al.* 2011, 2012). As distinguishing what types of genes are retained in duplicate is often a critical goal to understand selection pressures following gene duplication (e.g. Birchler & Veitia 2007), this could be an important omission. Nevertheless, whole-genome-based population genetic inferences on polyploid genomes are starting to emerge. Hollister *et al.* (2012) resequenced 12 individual plants from four populations of tetraploid *Arabidopsis arenosa* and aligned them to reference sequences from two diploid relatives (*Arabidopsis thaliana* and *Arabidopsis lyrata*) and used the three-way comparisons to interpret patterns of selection in the tetraploid genome. The novelty was that they also tested the mode of inheritance using a simulation approach compared to the observed SNP frequency distribution. Although only a portion of the sequence space that was found at a threshold read depth in *A. arenosa* and aligned to both other genomes could be used, the study demonstrated the utility of implicitly considering the different types of allele-frequency spectra expected in polyploids into analyses of selection at a genomewide scale.

There have already been some developments in strategies for incorporating gene duplication into models of genome assembly, and we anticipate that continuing improvements in both sequencing technology and bioinformatics pipelines will result in generation of well-annotated and complete polyploid genomes in the near future. Increasing the stringency (e.g. allowing differentiation of two divergent sequences as two different loci and not two alleles from the same locus) when assembling genomes may help to eliminate combining paralogues during SNP discovery analyses and could help to differentiate homoeologous sequences from each other in allopolyploids (Hohenlohe *et al.* 2011). For example, the Stacks software (Table 1; Catchen *et al.* 2013, 2011), which operates by ordering matching reads into different short-read 'stacks', could allow differentiation of paralogous (or homoeologous) from homologous sequences. By increasing the number of 'stacks' per locus in the module USTACKS (Catchen *et al.* 2013) and modulating the mismatch parameter used to produce these 'stacks', the user should be able to differentiate alleles from duplicated genes as well as alleles from homoeologous loci in allopolyploids (depending on the divergence between homoeologous loci). However, increasing the stringency of the assembly risks separating polymorphic loci that include highly divergent alleles at single loci (e.g. immune genes at the Major Histocompatibility Complex, MHC) into multiple loci (Seeb *et al.* 2011a). Comparison with a completely resolved and annotated reference genome is needed to distinguish divergent alleles from duplicated loci (Wang *et al.* 2013). Thus, there remains the circular problem of initially resolving duplicated or highly divergent genomes.

Another important issue related to all current NGS-sequencing approaches has to do with error rates. While the scale of the problem varies by method, for all current methods heterozygote genotypes can be falsely produced by the incorporation of spurious mutations during the sequencing (or amplifying) steps, and heterozygotes can be missed with insufficient sequence coverage. Taking into account the sequencing error rate and the depth of coverage is critical for properly characterizing homozygote and heterozygote genotypes and estimating allele frequencies, even in diploid populations (Lynch 2009; Hohenlohe *et al.* 2010). However, as the depth of coverage used to sequence and detect variants has to be sufficient to sample all variants present at a given locus, it should be increased proportionately to the ploidy level to account for the possibility of increased number of alleles. Again, dosage uncertainty in polyploids means that a simple calculation of read number in relation to expected heterozygosity at a given locus cannot be used to predict whether there has been sufficient coverage, as has been used for diploids (Catchen *et al.* 2013). There would also be difficulties with combining different ploidy levels in the same analysis, as it would be difficult to completely normalize read depths.

Various genomic assemblers (see Table 1) such as the CLCbio genomic workbench and the Genome Analysis Tool Kit (GATK; McKenna *et al.* 2010; DePristo *et al.* 2011) can incorporate the ploidy level as a parameter to discover or estimate the presence of variants in polyploids. The CLCbio genomic workbench uses a modified version of Neighbourhood Quality Standard (Altshuler *et al.* 2000; Brockman *et al.* 2008) to detect variants, taking into account the quality of the sequences. GATK, an open-source community platform, uses a Bayesian framework, taking into account *phred* quality score (Ewing *et al.* 1998) to disentangle spurious mutations from real variants (McKenna *et al.* 2010; DePristo *et al.* 2011). However, these approaches still often consider true variants to have a frequency of 0.5 in heterozygous genotypes and so might not be directly applicable to assessing reliability of SNP calls in polyploids. Simulation studies are required to assess how sensitive such approaches might be to assuming diploid inheritance in polyploid genomes or to individual loci showing polysomic inheritance, and to predict what types of biases might result.

For high-throughput SNP-genotyping platforms, there are some analytical approaches that can incorporate partial heterozygosity (i.e. heterozygotes with different dosage patterns), and we suggest that this is an area where further analytical solutions should continue to be developed, not only for these rapid genotyping methods but also for assessing reliability of SNPs obtained from whole-genome sequences. Using mixture models, the fitTetra R package allows genotyping and estimation of partial heterozygote tetraploid individuals using data obtained from high-throughput SNP genotyper platforms (Voorrips *et al.* 2011). Serang *et al.* (2012) have provided a Bayesian algorithm to genotype individuals and estimate SNP frequencies in populations with complex mixed-ploidy levels, which is currently compatible with Illumina GoldenGate assays and the Sequenom iPlex MassARRAY®. This algorithm is implemented in the software SuperMASSA (see Table 1). Once again, the problem of uncertainty in allele dosage remains a challenge: both software packages assume that the intensity of hybridization is directly proportional to the copy number (i.e. allelic dosage) at a given SNP site, which has not been systematically tested. Simulation studies to assess the sensitivity of these types of analyses to deviations from the expected dosage should be conducted to evaluate the utility of such approaches and identify where improvements should be made.

Multiplex amplicon sequencing. High-throughput targeted sequencing approaches hold great promise for understanding the evolutionary history of polyploid organisms and for identifying patterns of genetic diversity at adaptively important genes. This method has been used, for example, as a 'digital cloning' approach to resolving complex gene families in autotetraploid plants (Jørgensen *et al.* 2012). However, although the approach is more efficient than cloning in terms of coverage of amplicon products and confidence in resulting genotyping, potential biases associated with PCR-based techniques are not completely solved by a deep-sequencing approach. Uneven representation of allelic products can still be apparent within and between individuals or between PCR runs, and PCR recombinants can remain difficult to distinguish from genuine recombinant alleles. Differences in annealing of the tagged primers in allopolyploids due to divergence between the parental sequences could also complicate the interpretation of parental genome contributions (e.g. Bundock *et al.* 2009). Nevertheless, tagged amplicon sequencing has been applied to allopolyploids to simultaneously investigate linkage of multiple homologues of candidate genes coding for important traits (e.g. Gholami *et al.* 2012) and to investigate phylogeography of polyploids using a combination of nuclear and organellar genes (Griffin *et al.* 2011). Lessons learned from the analysis of complex gene families in diploids (e.g. MHC: Sommer *et al.* 2013) will be a useful source of solutions to increasing genotype reliability using tagged amplicons, which can be applied to both diploids and polyploids. There has been a recent switch to using Illumina-based sequencing technology, which produces shorter sequences but with lower rates of error than for 454; the rapid advances in both the technology (e.g. read length) and analyses (e.g. methods for detecting chimeric sequences, Quince *et al.* 2011) of these types of data should further increase the utility of this approach to applying population genetics models to sequences obtained from duplicated sequences.

Targeted sequence capture. Another type of approach that is increasingly being applied and that holds great promise for isolating multiple whole genes for use in population genetic studies of polyploids is the enrichment or targeting of particular parts of the genome (targeted sequence capture). Salmon *et al.* (2012) analysed heterozygosity of hundreds of homoeologues genes in wild and domesticated cotton *Gossypium hirsutum* with the aid of custom hybridization probes (targeting 500 pairs of homoeologues from the transcriptome). A similar approach was used to sequence 56.5 Mb of genomic DNA from allohexaploid bread wheat (Winfield *et al.* 2012) to assess variation at 500 000 SNPs, not only among gene copies but also among varieties. Bundock *et al.* (2012) used information from Sorghum (*Sorghum bicolor*) to capture the sequences of two closely related sugarcane genotypes (*Saccharum officinarum* and a hybrid cultivar) and were able to develop SNPs using Agilent Sure Select arrays and Illumina sequencing. The approach has also been applied to highly complex gene families (plant resistance genes) to identify not only already known genes but to identify hundreds more copies than had been identified from scans of complete genome sequences (Jupe *et al.* 2013) and to pull out orthologous sequences from distantly related plant species (potato and tomato). O'Neill *et al.* (2013) applied parallel tagged amplicon sequencing to better resolve species boundaries in *Ambystoma tigrinum*, a species with a large and complex genome. EST information from two related species was used, and 95 PCR-targeted unlinked nuclear loci in 93 individuals were used to assign individuals to different geographical regions using the STRUCTURE software (Pritchard *et al.* 2000). This combined sequencing and bioinformatics approaches resulted in a genomewide data set with relatively low levels of missing data and a wide range of nucleotide variation. The advantage of these types of methods for polyploids is that problems with unequal coverage across the genome due to large size and duplications would be reduced by focusing on a smaller number of target genes, for which read depth could be optimized to allow inference of number of alleles. Although it is not yet feasible to reliably infer copy number, given that this is also an area of concern for duplicated genes in diploids, we predict that creative solutions will appear in the near future.

Genotyping by sequencing. A currently expanding area of research is the use of complexity-reducing techniques to enable population-scale analyses of nonmodel organisms. 'Genotyping by sequencing' approaches are one such class of methods. Although there are a variety of approaches, restriction-associated DNA (RAD) sequencing (Baird *et al.* 2008) has been used the most frequently for population genetic applications (Hohenlohe *et al.* 2010, 2011; Rowe *et al.* 2011). RAD-Seq provides the ability to examine tens of thousands of genetic loci simultaneously in groups of individuals. The principle of this approach is similar to AFLPs in that genomic DNA is cut with restriction enzymes, but the digested fragments are then ligated to adapters and bar-coded to enable multiplex sequencing using NGS platforms. It yields two kinds of data: presence absence of markers resulting from polymorphism in the restriction enzyme cut site, and substitutional (SNP, indel) markers in tagged sequences. For polyploids, the advantage is that, with sufficient coverage, it should be possible to obtain all four copies (in a tetraploid) at a given polymorphic site and so theoretically determine allelic dosage. However, this assumes no bias in representation of allelic copies and equal read coverage across all loci, so that sequences can be normalized to a standard. Currently, this is not feasible even in diploids but if possible, would lead to a major breakthrough in sequence-based analyses of polyploid genomes. Although phase of substitutions is limited to a relatively short fragment of DNA flanking each cut site, the use of paired-end sequencing with a reference genome or using more than one restriction enzyme (double digest RAD: Peterson *et al.* 2012) has the potential to distinguish between paralogues by considering patterns of nucleotide substitutions over a larger sequence fragment and so to enable multilocus haplotype-based analyses (e.g. STRUCTURE analyses: Pritchard *et al.* 2000; Falush *et al.* 2007). One important drawback of RAD sequencing is the fact that mutations at restriction sites will make it impossible to observe the associated SNP allele, resulting in allele dropout. In addition, if restriction digest sites are present in transposons, large numbers of reads will not be informative; thus, stringent data filters are required (Twyford & Ennos 2012). Simulation studies have shown that including loci with missing data can lead to an over-estimation of F_{ST} values (Arnold *et al.* 2013; Gautier *et al.* 2013). The ascertainment of sites with missing data will be even more important in polyploids, given their duplicated loci. Simulation studies are required to better assess the effects of allele dropout in both auto- and allopolyploid organisms. The major advantage compared with AFLPs and microsatellites is being able to apply a testable model of evolution to the data and so increase the scale of inference possible about evolutionary and demographic processes.

So far, most studies that have used RAD sequencing for mapping have excluded potential paralogues in downstream analyses (e.g. sockey salmon: Everett *et al.* 2012), but testing segregation of variants within families could help to distinguish how many copies are present at a particular RAD 'locus' (i.e. the contiguous sequence next to a cut site). For allopolyploids, if it is possible to separate reads into the diploid contributions from each parent, then data can be analysed as if it were effectively diploid. For example, Hohenlohe *et al.* (2011) distinguished candidate SNPs for differentiation between *Oncorhynchus mykiss* and native westslope cutthroat trout (*Oncorhynchus clarkii lewisi*) by detecting excessively high observed heterozygosity and deviations from HW equilibrium. However, they appear to have assumed strict disomic inheritance; again, uncertainties in segregation patterns at each locus would affect the model for expected genotype distributions and so could bias these types of analyses.

Reduced representation NGS techniques suffer from the fact that mutations in the restriction enzyme restriction sites, along with the random sequencing of genomic fragments, may result in a large number of missing orthologues. This is of particular concern in large complex genomes because the larger sequence length means that there is a higher probability of stochastic differences in which SNPs are sequenced in different individuals (O'Neill *et al.* 2013). Uncertainties in allelic and gene copy number also means that errors remain more difficult to detect in polyploids than in diploids (as for the other NGS-based methods), but this is complicated by strategies for filtering data. The rediploidization process that occurs following genome duplication means that individuals could differ in which gene copies they retain. For genome-sampling approaches such as RAD sequencing, this means that filtering data to include only loci that are found in all individuals could omit important information on the fate of duplicate genes and could confound interpretation of paralogues. This would also be problematic when including multiple ploidy levels in the same analysis, as a uniform filtering strategy might lead to biases across ploidies.

Regardless of these cautions, complexity reduction approaches should in theory be easier to apply to polyploids than whole-genome approaches because of the reduced difficulties with ensuring sufficient coverage provided by sequencing only a targeted portion of the genome. There also should be no theoretical barrier to using assemblers and SNP genotypers developed for diploids. However, for very large and complex genomes, current methods might still be limited by uneven coverage across the genome. For example, in the complex case of sturgeon, where ploidy level can be as high as 2n 8x, but there has been varying degrees of rediploidization, Ogden *et al.* (2013) were able to discover SNPs using a RAD tag sequencing technique on a Illumina Hiseq2000 platform. However, they were unable to recover all of the polymorphisms expected from genotyping within a family (two parents and six offspring). A current but potentially transient benefit of complexity reduction approaches for polyploid genomes is that such approaches can be applied without assembly to a reference sequence, but inferences remain more powerful where this is possible. For example, in polyploid birch, paralogues were differentiated from homologues using the features of the Stacks assembler by comparing RAD sequences to a reference genome library, but not when comparing *de novo* RAD sequences to each other (Wang *et al.* 2013). While these approaches can reduce the cost of SNP discovery and genotyping by sequencing, the continued increase in data volumes at an ever-reducing cost may make whole-genome sequencing more efficient for SNP discovery in the future.

Combining methodologies. Even for diploids, there has been recognition that combining approaches has the greatest potential for resolving large and complex genomes. For example, long-read technologies that are prone to high error rates but can be used to generate scaffolds where a reference genome is not available, with higher accuracy short-read approaches used for detailed SNP identification. For example, You *et al.* (2011) used such a combined approach for SNP discovery in the diploid ancestor of the D genome of polyploid wheat (*Aegilops tauschii*), which itself has a genome size of over 4 Gb, with 90% repetitive sequences, making *de novo* assembly difficult. They combined Roche 454 shotgun reads with low-genome coverage of one genotype to distinguish single copy sequences and repeat junctions from repetitive sequences and sequences shared by paralogous genes and then mapped shotgun reads from other genotypes generated with SOLiD or Solexa to the annotated Roche 454 reads to identify putative SNPs. Mayer *et al.* (2011) combined chromosome sorting, NGS, array hybridization, and synteny comparisons with model grasses to construct an ordered scaffold of barley (*Hordeum vulgare*). Seeb *et al.* (2011a) included a high-resolution melt curve analysis (HRMA; Wu *et al.* 2008) and Sanger sequencing, as additional stringency steps, to validate transcriptome-based SNPs in tetraploid chum salmon. Such combined approaches hold the most promise for identifying individual markers that could be used for population genetic inference in polyploid genomes, to allow resolution of the full complexity of the evolutionary process when changes in copy number are critical for understanding relationships among populations.

Extending population genetic tools used for diploids to polyploids

General caveats for genetic marker analysis in polyploids

Analysis of allele and genotype frequencies and the quantification of deviations from the HW equilibrium are a central aspect of population genetics. Although the concepts of population genetics theory have predominantly been developed for diploids (Wright 1943, 1951), the same core principles apply to polyploids. The HW equilibrium principle can be applied to the diploid subgenomes of allopolyploids with strict disomic inheritance, if one can reliably identify the homoeologous copies. The principle has also been extended to autopolyploids, where polysomic inheritance and double reduction complicate matters (Haldane 1930; Geiringer 1949; Parsons 1959; see Bever & Felber 1992 for a review). For a polyploid with polysomic inheritance (without double reduction), expected genotype frequencies for a bi-allelic locus in HW equilibrium are predicted by the formula $(p + q)^{2m}$, in which p and q represent the frequencies of both allelic states and m is the 'haploid' ploidy level (Haldane 1930). The main effect of double reduction is that it causes the expected frequencies of homozygous genotypes to increase (Bever & Felber 1992 and references therein), resembling the effect of inbreeding (see Geiringer 1949; Parsons 1959; Bennett 1968 for some formulae for predicting genotype frequencies of polyploids with double reduction). This relates to a more general theoretical issue with the use of HW equilibrium in autopolyploids. Compared with diploids, the random mating equilibrium is not reached as fast in autopolyploids (Haldane 1930; Geiringer 1949; Bever & Felber 1992) and depends on the frequency of double reduction (Parsons 1959; Bennett 1968). This questions whether any method that is based on deviation from HW equilibrium is actually appropriate for autopolyploids. To the best of our knowledge, there are no theoretical studies that have addressed this issue.

In any case, the theoretical basis for population genetic analysis in polyploids is frequently not always possible to apply in practice. The reasons for this are mainly related to issues that have already been identified in the previous sections: (i) inheritance can deviate from strict disomic or polysomic and can vary from locus to locus and over time; (ii) dosage/copy number uncertainty and null alleles prevent reliable assessment of observed allele and genotype frequencies; and (iii) differences in ploidy level within a taxon or between closely related taxa included in the same analyses add an additional level of complexity to the population genetic analysis of polyploid species. It is of course possible to avoid difficulties with mixed ploidy (often referred to as mixed cytotypes) by analysing different ploidy levels separately, and to refrain from any interploidy comparison. This only seems reasonable in situations where different ploidy levels are indeed reproductively or spatially isolated. In *Aster amellus*, for example, diploids and hexaploids are completely reproductively isolated from each other, despite being morphologically indistinguishable and occurring in close vicinity (Münzbergová *et al.* 2013). However, although experimental crosses between ploidy levels tend to result in a much lower seed-set than crosses between plants with equal ploidy, reproductive isolation between ploidy levels can be incomplete (e.g. Hardy *et al.* 2001; Husband & Sabara 2003; Stift *et al.* 2010; Mraz *et al.* 2012). This means that gene flow between ploidy levels is possible and so population structure should be considered across cytotypes. Using molecular markers, gene flow across ploidy levels has, for example, been detected between diploids and tetraploids in *Arabidopsis arenosa* and *A. lyrata* (Jørgensen *et al.* 2011), and between diploids and apomictic triploids in *Taraxacum* (Menken *et al.* 1995). Given the frequent genetic exchange between ploidy levels, and the fact that polyploids are often recently derived from ancestors with a lower ploidy level, it is clearly undesirable to analyse different ploidy levels separately.

In this section, we will discuss the most commonly used approaches for population genetic analysis in polyploids, and how assumptions related to the inheritance mode and dosage uncertainty may affect these approaches. This will provide a thorough evaluation of the approach-specific pros and cons and allows us to make recommendation of work that is most critically needed to advance the field. We discuss some of the main statistical packages that implement these methods in the main text, Table 1, and boxes 2 and 3 and discuss creative solutions that are being suggested for extending analyses to polyploids. Some of the most exciting developments are being implemented in flexible programming environments that allow direct user additions, such as R (http://www.r-project.org/, R Development Core Team 2004). We anticipate that future advances will continue using these platforms.

Estimating allele frequencies

Estimation of allele frequencies is of great importance in the study of demographic factors influencing population structure such as migration, population growth or bottlenecks. Accurate allele frequencies are a prerequisite for the calculation of expected heterozygosities and estimates of population differentiation and fixation indices. Unlike in diploids, direct calculation of allele frequencies in polyploids can rarely be determined unless there is no uncertainty in allele copy number. A way around this problem is to incorporate dosage uncertainty into the inference of population genetic parameters. Unfortunately, there is not a single straightforward method for doing this. A first way is to estimate allele frequencies by considering that each allele in partial heterozygotes has an equal likelihood of being present in more than one copy (implemented in SPAGeDi assuming polysomic inheritance; Hardy & Vekemans 2002). This leads to an underestimation of common allele frequencies and an overestimation of rare allele frequencies (Clark & Jasieniuk 2011). A second method works by assigning the state of the unknown double dose allele based on the total sample or population allele frequencies (implemented in GenoDive assuming polysomic inheritance; Meirmans & Van Tienderen 2004). A problem arises here due to circularity caused by the very fact that the uncertainty in allelic dosage means that accurate population allele frequencies cannot be calculated and that assigning a particular allelic state changes the allele frequencies that the assignment was based on. A third method is to calculate allele frequencies and levels of heterozygosity in polyploid populations only based on unambiguous genotypes and ignoring genotypes with missing data (StAMPP, Pembleton *et al.* 2013; ATETRA, Van Puyvelde *et al.* 2010; and TETRASAT, Markwith *et al.* 2006; the latter two assuming disomic inheritance). This may cause biased allele frequencies because partial heterozygotes are ignored. In an extreme example of a tetraploid population of two individuals with genotypes ABCD and ABBB (which would be scored as ABX due to the dosage uncertainty), the true frequency of B is 1/2 but would be estimated as 1/4 if ABBB were excluded. Similarly, in a hexaploid population with ABBBBB and ABCDEF, the true frequency of B (1/2) would be estimated as 1/6. For ploidy levels above tetraploid, this method is probably obsolete anyway, as there will probably not be any unambiguous genotypes. Even in triploids and tetraploids, allele frequencies will be inaccurate for loci with limited variability and hence many ambiguous genotypes. This is because the frequency estimates for such loci can only be based on the limited number of individuals with unambiguous genotypes. A fourth way is to recalculate actual allele frequencies in the population from the 'allelic phenotype'

Box 2

Population differentiation indices

Analysis of population differentiation (*F*-statistics) is a key component of population genetic studies. Here, we list some of the F_{ST} analogues and interpopulation differentiation measures that have been developed for diploids but have been applied to polyploids and discuss what is expected under disomic and polysomic inheritance.

F_{ST} related measures

The first and the most widely used summary statistics in population genetics is Sewall Wright's F_{ST} (Wright 1943, 1965):

$$F_{ST} \cong \frac{\mathrm{Var}(p)}{p(1-p)}$$

where Var(*p*) is the variance of local allele frequencies among subpopulation and *p* is the mean allele frequency. The properties of this index have been well studied under island and isolation by distance models (Wright 1943; Wright 1946; Wright 1965; Slatkin & Barton 1989; Whitlock 2011). Under the finite island model and with a low mutation rate μ, F_{ST} is only dependent on the effective population size *N* and the migration rate *m*, such that:

$$F_{ST} = \frac{1}{4Nm+1}$$

if $m << 1$ and $\mu << m$ (Wright 1951).

Weir & Cockerham (1984) proposed θ as an estimate of F_{ST} using a simple Analysis of Variance (ANOVA) to calculate the variances within and among subpopulations. Similarly, different F_{ST} analogues have been proposed to analyse population structure by taking into account haplotype sequences, ϕ_{ST} (Excoffier *et al.* 1992), or a stepwise mutation model in microsatellites, R_{ST} (Slatkin 1995). Nei's G_{ST} (Nei 1973, 1987) is equivalent to Wright's F_{ST} but defined in terms of heterozygosity within subpopulations (H_S) and heterozygosity of the entire set of subpopulations under the assumption of HW equilibrium (H_T). It has been designed to account for the analysis of loci with multiple alleles:

$$G_{ST} = \frac{H_T - H_S}{H_S}$$

As many authors have shown than G_{ST} has the undesirable property of being constrained to a maximum value of <1 when the mutation rate is high, G'_{ST} was proposed by Hedrick (2005) to adjust for the number of alleles in a subpopulation as:

$$G'_{ST} = \frac{G_{ST}(d - 1 + H_S)}{(d-1)(1-H_S)}$$

with *d* being the number of subpopulation studied. Such standardization can be applied to other F_{ST} analogues (F'_{ST}, ϕ'_{ST} or θ') by weighting by their maximum values (Meirmans & Hedrick 2011). Moreover, as Hedrick's G'_{ST} (2005) may be biased when few subpopulations have been sampled, Meirmans & Hedrick (2011) have proposed a standardization to account for small population sample size, G''_{ST}.

Each of these measures can be adapted to autopolyploids but difficulties with inferring dosage again restrict their usage. In polyploid organisms with complete disomic inheritance in which heterozygosity can be fixed, even if complete genotypes can be resolved the expected heterozygosity (H_S) may be overestimated; hence, fixation indices such as F_{ST}, Nei's G_{ST} and G''_{ST} would be underestimated (Meirmans & Van Tienderen 2013). An alternative measure *Rho*, was proposed by Ronfort *et al.* (1998) after Tachida & Yoshimaru (1996) and Waller and Knight (1989). It has a theoretical background linked to Wright's $F_{ST,}$ with the following equation:

$$\frac{Rho}{1-Rho} = \frac{2F_{ST}}{(1+F_{IS})(1-F_{ST})}$$

where F_{IS} is the inbreeding coefficient of an individual within a subpopulation. Additionally, Ronfort *et al.* (1998) provided a method to estimate *Rho* using the ANOVA framework of Weir & Cockerham (1984). In their simulation studies (see Population differentiation), Meirmans & Van Tienderen (2013) found that this measure was least sensitive to the ploidy level, selfing rate and double reduction rate (and therefore mode of inheritance) and recommended it as the population differentiation measure of choice for polyploids.

Jost's *D*

Jost (2008) proposed a summary statistic that accounts for the number of alleles in the population:

$$D = \frac{d(H_T - H_S)}{(1-d)(1-H_S)}$$

This summary statistic measures the departure from total differentiation, which should not be confused with fixation indices (e.g. F_{ST}) that measure the departure from panmixia, at least in the finite island model (Whitlock 2011). Jost's *D* is not informative about migration between populations or other demographic processes and is dependent on neutral genetic diversity and the mutation rate. The behaviour of *D* according to the mode of inheritance is hard to interpret and *D* should therefore be avoided for the analysis of polyploid organisms (Meirmans & Van Tienderen 2013). However, it is implemented for extension to polyploid data in GENODIVE (Table 1).

frequencies (estimation of allele frequencies using phenotypes instead of genotypes) based on an iterative process. De Silva *et al.* (2005) developed a maximum-likelihood-based approach to do this, using the expectation maximization algorithm of Dempster *et al.* (1977), under the assumption of random mating and either disomic or polysomic inheritance without double reduction. This approach is implemented in PolySat (Clark & Jasieniuk 2011) and in GenoDive in a modified form (Meirmans & Van Tienderen 2004). A level of selfing can also be introduced in this estimate to improve the estimation of allele frequencies in inbred populations (not implemented in GenoDive). Detailed simulations of the consequences of implementing any of the approaches to circumvent dosage uncertainty have not yet been conducted, which would be required to assess what types of biases might result from the various strategies. Moreover, it should be realized that any method to assign an allelic state will obviously lead to a bias in cases for which the unknown allelic state is a null allele or an artefact arising during the PCR process.

Population structure

Bayesian clustering methods such as implemented in structure for the analysis of population structure (Falush *et al.* 2003) and in InStruct for simultaneous analysis of population structure and inbreeding rates (Gao *et al.* 2007) are popular methods in population genetics. The principle of Bayesian clustering is to assign individuals to one or more clusters such that deviation from HW equilibrium is minimized (Pritchard *et al.* 2000; Falush *et al.* 2003). Although initially developed for diploids, the programmes Structure and InStruct accommodate (auto)tetraploid data and allow joint analysis of different ploidy levels. Bayesian clustering has been used to infer the assignment of polyploid individuals to structured subpopulations (Lo *et al.* 2009; Shimizu-Inatsugi *et al.* 2009; Vanderpoorten *et al.* 2011; Tsuchimatsu *et al.* 2012). For example, Lo *et al.* (2009) used structure on a data set of 13 microsatellite loci to investigate possible gene flow and evolutionary relationships between sexually reproducing diploid and polyploid (triploid and tetraploid) populations reproducing by pseudogamous gametophytic apomixis of two species of hawthorns (*Crataegus suksdorfii* and *Crataegus douglasii*; Rosaceae). Due to a lack of genetic structuring (supported by the absence of isolation by distance) in tetraploid apomictic populations of *C. douglasii*, they concluded that there was either substantial gene flow among populations or that the populations originated from the same set of founders. In contrast, populations of the mixed-ploidy species *C. suksdorfii* clustered according to the ploidy level, suggesting a reduction of gene flow between cytotypes in this species.

However, the application of Bayesian clustering based on HW equilibrium in polyploids comes with a number of potential problems. Potential issues could mainly arise due to violations of the basic assumption of random mating within clusters. This problem is not specific to polyploids, but many polyploids frequently show a shift to inbreeding (reviewed in Mable 2004b) and asexual reproduction (Tomiuk & Loeschcke 1992; Dufresne & Hebert 1995; Stenberg *et al.* 2003; Aguilera *et al.* 2007; Lo *et al.* 2009; Vergilino *et al.* 2009; Neiman *et al.* 2011) and are often associated with novel habitats at range edges (e.g. Hijmans *et al.* 2007; Parisod *et al.* 2010). Because selfing, asexual reproduction and fast population growth cause departures from HW equilibrium, each of these cases represents a violation of the core assumptions of structure, which may produce either spurious population clustering or a lack of population structuring, depending on the genetic variability (Pritchard *et al.* 2000). It is currently unknown how seriously inference of population structure can be affected by violation of the underlying assumptions. Again there is a strong need for simulations that explicitly test each of the potential causes of bias in polyploids, simultaneously addressing the potential effect of null alleles and departures from polysomic inheritance.

Population differentiation

Quantifying population differentiation is among the main goals of population genetic analysis. Measures of population differentiation and partitioning of variance such as F_{ST} (or analogs) are therefore routinely reported in diploids. The main principles of *F*-statistics are extendible to autopolyploids with polysomic inheritance (e.g. Hardy *et al.* 2001; Andreakis *et al.* 2009). However, the previously identified problem of dosage uncertainty often prevents calculation of accurate allele and genotype frequencies. As such frequencies are needed to assess fixation indices (Box 2), it is frequently impossible to calculate *F*-statistics for autopolyploids.

In rare cases, where allele and genotype frequencies can be inferred for polyploids, remaining issues with using F_{ST} or related indices in polyploids include potential violations of the assumptions of HW equilibrium and polysomic inheritance. GenoDive (Meirmans & Van Tienderen 2004) and adegenet (Jombart 2008; Jombart & Ahmed 2011) have options to test deviation from HW equilibrium in polyploids. Simulating tetraploid genotype data, Meirmans & Van Tienderen (2013) demonstrated that assuming tetrasomic inheritance for a marker that in reality is inherited disomically may overestimate the expected within population heterozygosity and underestimate the estimation of divergence between populations as measured by Nei's G_{ST} (Nei 1987), G''_{ST} (Hedrick 2005) and Jost's *D* (Jost 2008). *Rho-st* (Tachida & Yoshimaru 1996; Ronfort *et al.* 1998) proved to be the only measure of population differentiation that was independent of the ploidy level, selfing rate and double reduction rate, and appeared unbiased by the type of inheritance (Meirmans & Van Tienderen 2013). This led the authors to recommend *Rho-st* as the preferred statistic for assessing population differentiation in polyploids with unknown segregation (see Box 2). However, they also warned that *Rho* is analogous to the 'correlation between truly outcrossed mates' defined for diploids in Tachida & Yoshimaru (1996) and cannot be interpreted as directly equivalent to an F_{ST} estimate because the values that *Rho* takes are comparable to expected F_{ST} values for haploid populations. Encouragingly, though, Meirmans & Van Tienderen (2013) also found that violating the assumption of tetrasomic inheritance does not bias other more standard F_{ST} measures too much, as long as there are sufficient intergenomic recombination events (around one event per generation). It is this type of simulation work that holds most promise for assessing potential biases that might result from lack of knowledge of inheritance patterns and allelic dosage when making population genetic inferences based on polyploid data sets.

Box 3

Inter individual and inter population distance/similarity indices

In this box, we provide an overview of the formulae underlying the distance and similarity indices discussed in Genetic Distance Based Analyses. They represent indices that are frequently applied to polyploids and that users are likely to encounter in software packages that accommodate polyploid data. For each of the indices below, it is important to realize that most have been formulated as similarities (Simple Match, Jaccard, Lynch, Kosman), but some as distances (Bruvo, and all the interpopulation measures). Similarities and distances are related in a relatively simple manner: *similarity* 1 *distance*. The general use of distance and similarity indices in population genetics has been reviewed by Kosman & Leonard (2005) and will not be dealt with in detail. Here, we focus predominantly on their extensions to polyploid data, and we indicate the software programs that implement these extensions.

Simple match index (squared Euclidian distance)

Software allowing calculation for polyploid data: none

The simple match index (*M*) calculates the similarity between two (in principle haploid) individuals based on the multilocus presence/absence data. It is calculated as $M(i_1,i_2) = (n \quad b \quad c)/n$ or $M(i_1,i_2) = (a + d)/n$ (Sneath & Sokal 1973), in which $n = a + b + c + d$ and is the length of the presence/absence (1,0) vector for all individuals under consideration, *a* is the number of shared band presences among i_1 and i_2, *b* is the number of bands present in i_1 and absent in i_2, *c* is the number of bands absent in i_1 and present in i_2, and *d* is the number of shared band absences. Because it is based on presence and absence, it can be applied independently of ploidy level. For both diploids and polyploids, the index should be calculated per locus and subsequently averaged over all loci. As shared absences contribute to similarity, the simple match index increases with marker diversity, making it mainly applicable to closely related individuals (Kosman & Leonard 2005).

Jaccard and Dice (Lynch) similiarity indices

Software allowing calculation for polyploid data: PolySat

The Jaccard similarity index is calculated as $J(i_1,i_2) = a/(a + b + c)$ and the Dice similarity index as $D(i_1,i_2) = 2a/(2a + b + c)$ (Legendre & Legendre 1998), in which *a* corresponds to bands shared between individuals i_1 and i_2, *b* corresponds to the presence in i_1 and absence in i_2, and *c* corresponds to the absence in i_1 and presence in i_2. The main difference between the Jaccard and Dice indices lies in the weight given to shared bands. This is twice as large for the Dice index, which works out to be the equivalent of the similarity index independently developed by Lynch (1990). Both the Jaccard and Dice/Lynch index can be readily calculated for dominant and co-dominant diploid and polyploid data by calculating the index per locus and subsequently averaging over all loci. They mainly differ from the simple match index in that the shared absence of bands does not contribute to similarity. This makes their use unrestricted with regard to the expected relatedness of the analysed individuals, although with highly variable markers the risk of homoplasy may lead to overestimation of similarity (Kosman & Leonard 2005).

Kosman & Leonard´s similarity index

Software allowing calculation for polyploid data: none

Kosman & Leonard (2005) questioned the consistency of the Jaccard/Dice/Simple Match indices when analysing diploid or higher ploidy data. They based this on an apparent inconsistency when more than two alleles are present at a single locus. For example, in a diploid case with three alleles at a single locus (A, B and C), the similarity between genotypes AC and CC gives a Jaccard-similarity of 1/2 (Dice: 2/3), whereas the similarity between AB and AC gives a similarity of 1/3 and 1/2, respectively. Kosman & Leonard (2005) argued that since in both comparisons one allele is shared, the genotype pairs should have the same similarity and proposed an index (which we dub the Kosman Leonard index). This is calculated as a/q, in which a corresponds to the number of shared alleles and x to the ploidy. The Kosman Leonard similarity between AC and CC thus equals 1/2, just like the similarity between AB and AC. For a tetraploid, the similarity between AAAA and AAAB equals 3/4 (three of four alleles shared), between AAAA and AABB 1/2 (two of four alleles shared), and between AAAA and ABBB 1/4 (one out of four alleles shared). One disadvantage is that the index can only be calculated for complete co-dominant genotypes, whereas determining the dosage is one of the main challenges in polyploids.

Smouse and Peakall interpopulation distance

Software allowing calculation for polyploid data: GenoDive

Smouse & Peakall (1999) proposed a distance specifically designed for co-dominant markers in diploids that was adapted for polyploid individuals in the GenoDive software (Meirmans & Van Tienderen 2004). The distance is based on a geometric space with r vertices, where each vertex is represented by each homozygous genotype, the distance between them for diploid organisms equals 2, and heterozygotes are positioned midway between the respective homozygotes. So, using this framework for a locus with three alleles (A, B and C) in diploids, the distances between AA and BB and between AA and AB are equal to 2 and 1, respectively, and the distance between AA and BC is $\sqrt{3}$. For polyploids, the distances are difficult to summarize verbally, but the following matrix shows the Smouse and Peakall distances calculated by GenoDive for tetraploid genotypes, with ABCD as reference:

	ABCD	AABC	ABBC	ABCC	ABBB	AABB	AAAB	AAAA	AAAE	AAEF	AFGH	EFGH	EEEE
ABCD	0	1	1	1	3	2	3	6	4	3	3	4	10

The Smouse and Peakall distance has been criticized for having a poor biological rationale even for relationships between diploid organisms (Kosman & Leonard 2005), and as this criticism also applies to relationships between polyploids (why would ABCD be more distant from AAAA than from EFGH?), its use should be avoided. Besides that, it is a disadvantage that the index can only be calculated for complete co-dominant genotypes.

Nei's interpopulation distance

Software allowing calculation for polyploid data: popdist, GenoDive, SPaGeDi, adegenet, atetra, StAMPP

Nei (Nei 1972) proposed an inter-population distance:

$$D_s = \quad \ln(J_{12}/\sqrt{J_1 \cdot J_2})$$

with J_1 and J_2 corresponding to the arithmetic means of the probabilities of identity of two randomly chosen alleles in populations 1 and 2, respectively, and J_{12} corresponding to the arithmetic means of the probabilities of identity of a randomly chosen allele in population 1 and a randomly chosen allele in population 2. This measure has not been tested in simulation studies using polyploid populations and so it is not known whether the mode of inheritance assumed will bias the estimates. As it is based on estimation of allele frequencies, it will suffer from difficulties with resolving complete genotypes in polyploids.

Tomiuk and Loeschke interpopulation distance

Software allowing calculation for polyploid data: popdist

Tomiuk & Loeschcke (1991) proposed a distance D_{TLG} based on the frequency of shared phenotype/genotype classes between populations with mixed ploidy level:

$$D_{\mathrm{TLG}} = \quad \ln(\sqrt{I_1 \cdot I_2})$$

with I_1 and I_2 corresponding to the genetic identities of two populations, 1 and 2, and their common ancestral population, following the equation:

$$\frac{z_{X1}+z_{X2}}{I_X} \quad \frac{z_{X3}(I_X^{n-1} \quad (1 \quad I_X)^{n-1})}{1 \quad I_X^n \quad (1 \quad I_X)^n} \quad \frac{z_{X4}}{1 \quad I_X}=0$$

where n is the ploidy level of population X and z_{X1}, z_{X2}, z_{X3} and z_{X4} represent the observed frequencies of: (i) homozygotes found in population X whose alleles are found in both populations; (ii) heterozygotes found in population X whose alleles are found in both populations; (iii) heterozygotes that have at least one allele present in both populations and at least one allele that is not observed in the other population; and (iv) phenotypes/genotypes carrying exclusively allele(s) found only in population X. This distance is only useful for studies of populations where private alleles occur (i.e. an allele that is present in only one of the two populations being compared).

Tomiuk's band sharing measurement

Software allowing calculation for polyploid data: POPDIST

Tomiuk *et al.* (2009) proposed another interpopulation distance measure for polyploids, the Band Sharing Measurement (D_{BSM}), largely inspired by the inter-population distance of Nei (1972) but that does not take into account the redundancy of alleles in partial heterozygotes (in other words, it is based on allelic phenotypes rather than genotypes). It allows estimation of distances between subpopulations with different ploidy levels even if these subpopulations have no private alleles. However, this measure does not behave linearly with ploidy level increase or when populations are closely related (Tomiuk *et al.* 2009).

Bruvo distance

Software allowing calculation for polyploid data: GENODIVE *and* POLYSAT

Bruvo *et al.* (2004) proposed a distance for microsatellites that takes the mutational process into consideration. It assumes a stepwise mutation model and calculates a matrix of distances between pairs of individuals, in which the distance (d) is calculated as d = 1 $2^{-|x|}$, in which x is number of repeat differences, so that the distance approaches 1 as the number of repeat differences increases. Its main advantage is that it can be used for calculating distances regardless of ploidy level. Although its application is much simpler when complete genotypes are available, it can deal with dosage uncertainty. It does so by averaging over all possible allelic constitutions. In cases of mixed ploidy, the method assumes autopolyploidization and assigns one or more 'virtual alleles' to the individuals with the lower ploidy level. For each paired comparison, it will do this in two steps. The first step represents a scenario of 'genome loss': that one or more alleles of the higher ploidy level were lost in the lower ploidy. Hence, it assigns the value of the virtual allele of the lower ploidy level to represent each of the different alleles of the higher ploidy level, calculates d for each situation and calculates the average d over each of the genome loss scenarios. The second step represents a scenario of 'genome addition': that one or more alleles of the lower ploidy level were duplicated. Hence, it assigns the value of the virtual allele of the lower ploidy level to represent each of the different alleles of the lower ploidy level, in all possible combinations and calculates the average d over each of the genome addition scenarios. Finally, the distance between the two individuals that differ in ploidy level is calculated as the sum of the average distance for the two scenarios, divided by the higher of the two ploidy levels (k_{max}): $d_{\text{different ploidy}}$ = ($d_{\text{'genome loss'}}$ + $d_{\text{'genome addition'}}$)/$k_{\text{max.}}$

The Bruvo distance is not implemented in the same way in GENODIVE and POLYSAT. In GENODIVE, the value of the 'virtual allele' can be set manually from 0 to infinite. In the POLYSAT package, the value of the 'virtual allele' is infinite by default, so that the geometric distance between any allele and a virtual allele is always 1.

Example: As a hypothetical example, we compared 10 diploid and triploid single locus genotypes (AA, BB, CC, AB, AC, BC, AB , AC , BC and ABC, in which A, B and C are alleles that differ by one mutational step, and ' ' corresponds to the unknown allele). Depending on the index (Jaccard or Bruvo, with an infinite value for the virtual allele) and software (GENODIVE or POLYSAT), the relative distances between diploid and triploid genotypes changes (Figure 1). This is due to the fact that the Jaccard distance does not consider allelic dosage in ambiguous polyploid genotypes, resulting in a Jaccard distance of 0 between diploid heterozygotes (e.g. AB) and corresponding triploid partial heterozygotes (AB) (Fig. 1A,B). For the Bruvo distances in this specific example, POLYSAT differs from GENODIVE, because POLYSAT accounts for allelic dosages in ambiguous polyploid genotypes (resulting in a distance >0 between diploids and triploids with the same alleles; Fig. 1C), whereas GENODIVE does not (resulting in a distance of 0 between diploids and triploids with the same alleles; Fig. 1B).This simple theoreti cal experiment shows that the choice of distance index and software used to estimate the relationships between diploid and polyploid organisms can strongly affect the conclusions reached.

Genetic distance–based analyses

Distance or similarity indices are a common tool in population genetics; for example, to assess population differentiation, diversity within populations, isolation by distance and for clustering approaches (reviewed by Kosman & Leonard 2005). There are several distance/similarity measures (Box 3), most of which were not specifically developed for polyploids, but which can be applied to polyploid data.

For example, the simple-match similarity coefficient has been used to differentiate populations of the tetraploid marram grass based on AFLPs (*Amnophila arenaria*; Hol *et al.* 2008) and to assess isolation by distance patterns in hexaploid and enneaploid (9x) individuals of a tallgrass species (*Andropogon gerardii*; Rouse *et al.* 2011). Its calculation for polyploid data does not require modification of the formula for haploids and the index can include mixed ploidies (Kosman & Leonard 2005). Other distance measures, such as the Jaccard and Dice similarity indices (Lynch 1990; Legendre & Legendre 1998), Kosman and Leonard's similarity index (Kosman & Leonard 2005), and Smouse and Peakall's distance (Smouse & Peakall 1999) can also be applied to polyploid data and mixed-ploidy data. However, each of these distances, like any other summary statistic using phenotypes instead of genotypes (Obbard *et al.* 2006), suffer from a loss of information due to the fact they do not take into account the allele dosage in polyploid heterozygotes. In addition, they have a poor genetic rationale (Clark & Jasieniuk 2011) and could lead to biases in interpretation, especially in cases of mixed ploidy. The potential for such bias is best illustrated through a thought experiment. For a diploid and a triploid, both with dominant genotype AB and AB, the distance can vary from 0 to 0.33, depending on the method used to calculate the distances. Methods collapsing the allelic data into dominant phenotypes result in a distance of 0, as both the diploid and triploid will be treated as AB. On first sight, this may not be so bad, as all this reflects is that the diploid AB and triploid AB share all their alleles, which is true (regardless of whether the triploid is AAB or ABB). In addition, if the unknown allele in the triploid is not actually A or B, but a null allele, the zero distance would be an underestimate of the real distance. As the probability of null alleles could increase with increasing ploidy, the estimated distances between diploids and higher ploidy levels would on average be underestimated. Some methods developed for diploids can take into account complete genotypes, such as Kosman and Leonard's Similarity Index (Kosman & Leonard 2005) and Nei's interpopulation distance (Nei 1972), but they retain the difficulty of resolving dosage in polyploids.

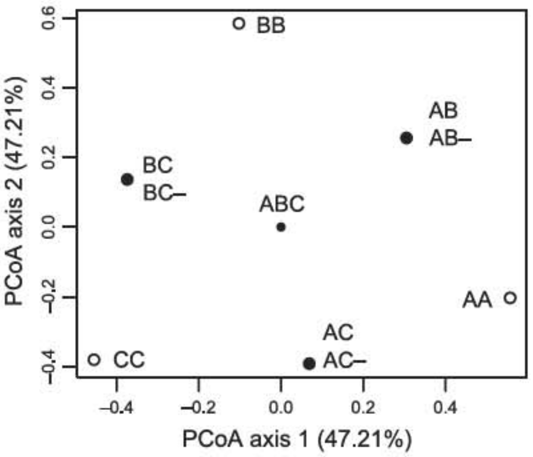
Some interpopulation indices have specifically been designed to study relationships between polyploid populations and between populations of different ploidy levels (see Box 3). The Tomiuk and Loeschke distance (Tomiuk & Loeschcke 1991) estimates interpopulation distance based on proportions of different classes of genotypes (i.e. homozygotes vs. different types of heterozygotes, which is often reduced to phenotypes if the dosage is unknown). The Band Sharing Measurement (Tomiuk *et al.* 2009) does the same based on the sharing of common alleles. An alternative to the indices based on the shared presence and/or absence is the Bruvo distance (Bruvo *et al.* 2004). It was specifically developed for polyploids and calculates distances between co-dominant microsatellite genotypes based on the assumption that slipped-strand mispairing is the main driver of length variation among alleles. It has been implemented in GenoDive and PolySat, which differ in the way allelic dosage in partial heterozygotes is assessed this has a strong effect on the distance calculation (Box 3). The Bruvo distance has been used, for example, to differentiate between clonal tetraploid genotypes of hawthorns (Crataegus; Rosaceae; Lo *et al.* 2009) and between closely related octoploid subspecies of *Atriplex* sp. (Sampson & Byrne 2012). The Bruvo distance may also lead to an overestimation of the genetic distance between individuals with different ploidy levels and may falsely group individuals with the same ploidy level together, especially in the case of autopolyploids or allopolyploids from closely related parents (Clark & Jasieniuk 2011). Hence, the parameters used to calculate the Bruvo distance, in particular those related to allele dosage, have to be set with caution (Meirmans & Van Tienderen 2004; Clark & Jasieniuk 2011), and the Bruvo distance should only be used for microsatellite loci for which it is reasonable to assume a stepwise mutation model. Given the general difficulty of determining dosage in polyploids, the distance/similarity indices that do not require full genotypes are most useful for the analysis of polyploids, despite their suboptimal use of genetic information.

Multivariate analyses

Multivariate and cluster analyses such as principal component analysis (PCA; Pearson 1901; Hotelling 1933) or principal coordinate analysis (PCoA; Gower 1966) can be used to visualize genetic distances among individuals. Their lack of any underlying population genetics-based assumptions, such as HW equilibrium, make multivariate approaches independent of ploidy level, and therefore suitable to analyse polyploid data as well as mixed-ploidy data (reviewed in Jombart *et al.* 2009). In polyploids, multivariate analyses have been used to

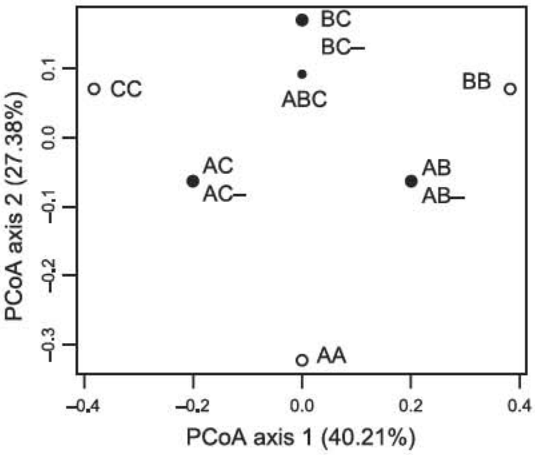
(a)

	AA	BB	CC	AB	AC	BC	AB–	AC–	BC–
BB	1.00								
CC	1.00	1.00							
AB	0.50	0.50	1.00						
AC	0.50	1.00	0.50	0.67					
BC	1.00	0.50	0.50	0.67	0.67				
AB–	0.50	0.50	1.00	0.00	0.67	0.67			
AC–	0.50	1.00	0.50	0.67	0.00	0.67	0.67		
BC–	1.00	0.50	0.50	0.67	0.67	0.00	0.67	0.67	
ABC	0.67	0.67	0.67	0.33	0.33	0.33	0.33	0.33	0.33



(b)

	AA	BB	CC	AB	AC	BC	AB–	AC–	BC–
BB	0.50								
CC	0.75	0.50							
AB	0.25	0.25	0.63						
AC	0.38	0.50	0.38	0.25					
BC	0.63	0.25	0.25	0.38	0.25				
AB–	0.25	0.25	0.63	0.00	0.25	0.38			
AC–	0.38	0.50	0.38	0.25	0.00	0.25	0.25		
BC–	0.63	0.25	0.25	0.38	0.25	0.00	0.38	0.25	
ABC	0.50	0.50	0.50	0.33	0.33	0.33	0.33	0.33	0.33



(c)

	AA	BB	CC	AB	AC	BC	AB–	AC–	BC–
BB	0.50								
CC	0.75	0.50							
AB	0.50	0.50	0.75						
AC	0.50	0.75	0.50	0.25					
BC	0.75	0.50	0.50	0.38	0.25				
AB–	0.67	0.67	0.83	0.33	0.50	0.58			
AC–	0.67	0.83	0.67	0.50	0.33	0.50	0.17		
BC–	0.83	0.67	0.67	0.58	0.50	0.33	0.25	0.17	
ABC	0.67	0.67	0.67	0.33	0.33	0.33	0.33	0.33	0.33

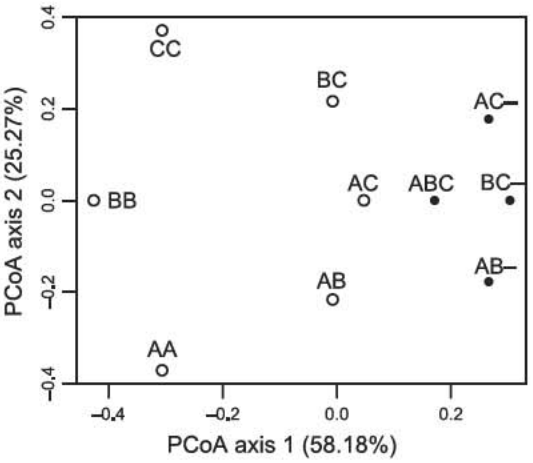


Fig. 1 Jaccard and Bruvo genetic distance between ten simulated diploid and triploid genotypes (AA, BB, CC, AB, AC, BC, AB , AC , BC and ABC) and the corresponding Principal Coordinate Analyses. (a) Jaccard distance calculated by hand; (b) Bruvo distance with ploidy level variation as an infinitely large mutation using (calculated by GenoDive); (c) Bruvo distance with ploidy level variation as an infinitely large mutation (calculated by Polysat). In the left panels, shaded cases show distances between diploid heterozygotes and the corresponding ambiguous triploid partial heterozygotes. In the right panel, large open symbols represent only diploid genotypes and small solid symbols represent only triploid genotypes; large solid symbols represent both diploid and triploid genotypes.

infer evolutionary and genetic relationships either between populations or between individual genotypes (Vergilino *et al.* 2011).

Other multivariate analyses such as K-means clustering (Hartigan & Wong 1979) and Discriminant Analysis of Principal Components (DAPC; Jombart *et al.* 2010) allow clustering of polyploid and mixed-ploidy level populations using SNPs (including datasets from high-throughput genotyping) and SSR data. The DAPC method, which may use K-means as a priori clustering algorithms, implemented in ADEGENET (Jombart 2008; Jombart & Ahmed 2011), provides an interesting alternative to STRUCTURE and INSTRUCT software as it does not require that populations are in HW equilibrium and can high handle a large amount of data (Jombart *et al.* 2010). However, as for the other multivariate analyses, the reduction of genetic information to interindividual or interpopulation distances represents a substantial loss of information. Therefore, methods that make use of genotype information are in principle more powerful, and to be preferred over multivariate approaches. Nevertheless, multivariate approaches can provide an attractive visual complementation to other methods and are the method of choice in cases where other methods are inappropriate due to violation of assumptions (such as random mating). The main issue with using multivariate analyses in polyploids is the calculation of the underlying distance matrices, which should be chosen with caution according to the marker used and the type of the multivariate analysis used (Jombart *et al.* 2009).

Custom model-based analyses to analyse complex scenarios

Population geneticists have provided custom models to test different hypotheses about demographic and evolutionary history, such as bottleneck events or change in mode of reproduction in diploid populations (Beaumont *et al.* 2010; Csilléry *et al.* 2010). Few simulation software or coalescent models have included features specific to polyploid organisms, such as larger effective population size at the gene level or the possibility of polysomic inheritance (but see Arnold *et al.* 2012). The few studies that have analysed polyploid organisms with a custom model-based approach have taken advantage of particular features of the organisms studied, such as disomic inheritance or high selfing rate, to analyse their data and avoid the difficulties inherent to polyploidy. For example, using coalescent-based models on numerous microsatellite loci and nuclear sequences, Jakobsson *et al.* (2006) tested the hypothesis of a unique origin of the highly selfing allotetraploid species *Arabidopsis suecica* from hybridization between the highly selfing species *A. thaliana* and the obligate outbreeding species *A. arenosa*. They assessed different scenarios changing the number of founders and the time of origin, taking into account the mode of reproduction of the different species and then accepted the model suggesting a unique origin of the polyploid species *A. suecica*, as previously proposed by Säll *et al.* (2003). St-Onge *et al.* (2012) used a coalescent-based model on the basis of nucleotide variation of 14 nuclear genes to test whether Shepherd's purse (*Capsella bursa-pastoris*), a polyploid species with disomic inheritance, had an allo- or autopolyploid origin. Sequences of each gene copy from the tetraploid species *C. bursa-pastoris* were amplified using homeolog-specific primers and compared with corresponding sequences from the diploid, and potentially parental, species *C. grandiflora and C. rubella*. St-Onge *et al.* (2012) first compared the number of fixed differences in the homeologous genomes A and B in *C. bursa-pastoris* and in the genomes of *C. grandiflora* and *C. rubella*, and the number of shared polymorphism between them. The high number of fixed differences between the homeologous genomes of *C. bursa-pastoris* that are shared with the genome of *C. grandiflora* is consistent with a scenario of speciation by autopolyploidization of *C. bursa-pastoris*. They then used coalescent-based simulations and an Approximate Bayesian Computation (ABC) model to estimate that the gene copies in *C. bursa-pastoris* diverged before the speciation process leading to the formation of the diploids *C. grandiflora* and *C. rubella* and so were not able to reject the hypothesis of autopolyploidization of *C. bursa-pastoris,* followed by the divergence of gene copies following polyploidization. Although Jakobsson *et al.* (2006) and St-Onge *et al.* (2012) used diploid-based models and simulation programmes to test their hypotheses, specific models including polysomic inheritance are under development. For example, a coalescent model for autotetraploid populations with tetrasomic inheritance, in which partial selfing (as well as double reduction) can be simulated, has been provided recently that may be useful to test different evolutionary scenarios (Arnold *et al.* 2012). Such approaches are promising, as polyploid or ploidy-mixed populations may have different modes of reproduction, rates of mutation or demographic history and these different parameters may be modelled and tested independently using custom-based models.

Conclusions

The evolution of polyploidy is a fascinating topic and many insights have been obtained from investigation of population genetic processes through the analysis of various types of molecular markers. However, dosage uncertainty and unknown segregation patterns result in difficulties in calculating observed and expected allele frequencies. This affects the ability to apply standard population genetic models to investigate population structure in polyploid organisms and additional custom-based models should be developed to take such factors into account. One of the major problems in the analysis of single copy markers remains our inability to reliably determine allelic configurations in polyploids. This prevents estimation of heterozygosity, which is at the heart of population genetic theory in diploids. New genomic tools offer great promise to unravel population genetics questions in polyploids given the astounding number of markers that are becoming available. While NGS approaches still bear some old (determining allelic dosage, gene copy number and paralogous sequences) and new (potential biases associated with PCR-based techniques, difficult assembly and annotation due to the presence of multiple gene copies, lack of models to filter errors that can incorporate dosage uncertainty and incomplete genotypes) problems when working with polyploid genomes, we anticipate that rapid developments in both sequencing technology and computational approaches to statistical inference should dramatically reshape the field of polyploid population genetics in the near future. For both new and old markers, we recommend that more simulation-based studies should be conducted to assess the sensitivity of population genetic analyses to potential biases caused by uncertainty in genotypes and modes of inheritance. We hope that this review will stimulate the development of new theory and practical approaches for analysing complex data sets involving extensive gene duplication and 'flexible' modes of inheritance.

Acknowledgements

FD acknowledges a grant from the Natural Sciences and Engineering Research Canada (NSERC). RV is supported by a fellowship from the NSERC CREATE program: Aquatic ecosystem health. We thank Karim Gharbi for many fruitful discussions on the current challenges of applying deep sequencing approaches to polyploid genomes. We also thank Maurine Neiman and two anonymous reviewers for valuable suggestions that have substantially improved the manuscript.

References

Aguilera X, Mergeay J, Wollebrants A, Declerck S, De Meester L (2007) Asexuality and polyploidy in *Daphnia* from the tropical Andes. *Limnology and Oceanography*, **52**, 2079 2088.

Ainouche ML, Jenczewski E (2010) Focus on polyploidy. *New Phytologist*, **186**, 1 4.

Ainouche ML, Baumel A, Salmon A (2004) *Spartina anglica* CE Hubbard: a natural model system for analysing early evolutionary changes that affect allopolyploid genomes. *Biological Journal of the Linnean Society*, **82**, 475 484.

Ainouche M, Fortune P, Salmon A *et al.* (2009) Hybridization, polyploidy and invasion: lessons from *Spartina* (Poaceae). *Biological Invasions*, **11**, 1159 1173.

Allendorf FW (1978) Protein polymorphism and the rate of loss of duplicate gene expression. *Nature*, **272**, 76 78.

Allendorf FW, Danzmann RG (1997) Secondary tetrasomic segregation of MDH B and preferential pairing of homeologs in rainbow trout. *Genetics*, **145**, 1083 1092.

Altshuler D, Pollara VJ, Cowles CR *et al.* (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513 516.

Andreakis N, Kooistra WHCF, Procaccini G (2009) High genetic diversity and connectivity in the polyploid invasive seaweed *Asparagopsis taxiformis* (Bonnemaisoniales) in the Mediterranean, explored with microsatellite alleles and multilocus genotypes. *Molecular Ecology*, **18**, 212 226.

Arnold B, Bomblies K, Wakeley J (2012) Extending coalescent theory to autotetraploids. *Genetics*, **192**, 195 204.

Arnold B, Corbett Detig R, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179 3190.

Auer H, Newsom DL, Nowak NJ *et al.* (2007) Gene resolution analysis of DNA copy number variation using oligonucleotide expression microarrays. *BMC Genomics*, **8**, 111.

Aversano R, Ercolano MR, Caruso I *et al.* (2012) Molecular tools for exploring polyploid genomes in plants. *International Journal of Molecular Sciences*, **13**, 10316 10335.

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

Baumel A, Ainouche M, Bayer R, Ainouche A, Misset M (2002) Molecular phylogeny of hybridizing species from the genus *Spartina* Schreb. (Poaceae). *Molecular Phylogenetics and Evolution*, **22**, 303 314.

Beaumont MA (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013 2029.

Beaumont MA, Nielsen R, Robert C *et al.* (2010) In defence of model based inference in phylogeography. *Molecular Ecology*, **19**, 436 446.

Bennett J (1968) Mixed self and cross fertilization in a tetrasomic species. *Biometrics*, **24**, 485 500.

Bensch S, Åkesson M (2005) Ten years of AFLP in ecology and evolution: why so few animals? *Molecular Ecology*, **14**, 2899 2914.

Bever JD, Felber F (1992) The theoretical population genetics of autopolyploidy. *Oxford Surveys in Evolutionary Biology*, **8**, 185 217.

Birchler JA, Veitia RA (2007) The gene balance hypothesis: from classical genetics to modern genomics. *Plant Cell*, **19**, 395 402.

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M (2000) Extensive duplication and reshuffling in the Arabidopsis genome. *The Plant Cell*, **12**, 1093 1101.

Bogart JP (1980) Evolutionary implications of polyploidy in amphibians and reptiles. In: *Polyploidy* (ed. Lewis WH), pp. 341 378. Plenum Press, New York, New York.

Bourret V, Kent MP, Primmer CR *et al.* (2012) SNP array reveals genome wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Molecular Ecology*, **22**, 532 551.

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, **422**, 433 438.

Brenchley R, Spannagl M, Pfeifer M *et al.* (2012) Analysis of the bread wheat genome using whole genome shotgun sequencing. *Nature*, **491**, 705 710.

Brochmann C, Soltis PS, Soltis DE (1992) Recurrent formation and polyphyly of Nordic polyploids in *Draba* (Brassicaceae). *American Journal of Botany*, **79**, 673 688.

Brockman W, Alvarez P, Young S *et al.* (2008) Quality scores and SNP detection in sequencing by synthesis systems. *Genome Research*, **18**, 763 770.

Bruvo R, Michiels NK, D'Souza TG, Schulenburg H (2004) A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. *Molecular Ecology*, **13**, 2101 2106.

Brysting AK, Mathiesen C, Marcussen T (2011) Challenges in polyploid phylogenetic reconstruction: a case story from the arctic alpine *Cerastium alpinum* complex. *Taxon*, **60**, 333 347.

Buggs R (2008) Towards natural polyploid model organisms. *Molecular Ecology*, **17**, 1875.

Buggs RJA (2013) The consequences of polyploidy and hybridisation for transcriptome dynamics unravelling gene expression of complex crop genomes. *Heredity*, **110**, 97 98.

Buggs R, Doust A, Tate J *et al.* (2009) Gene loss and silencing in *Tragopogon miscellus* (Asteraceae): comparison of natural and synthetic allotetraploids. *Heredity*, **103**, 73 81.

Buggs RJ, Chamala S, Wu W *et al.* (2010) Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular Ecology*, **19**, 132 146.

Buggs RJ, Renny Byfield S, Chester M *et al.* (2012) Next generation sequencing and genome evolution in allopolyploids. *American Journal of Botany*, **99**, 372 382.

Bundock PC, Eliott FG, Ablett G *et al.* (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotechnology Journal*, **7**, 347 354.

Bundock PC, Casu RE, Henry RJ (2012) Enrichment of genomic DNA for polymorphism detection in a nonmodel highly polyploid crop plant. *Plant Biotechnology Journal*, **10**, 657 667.

Burnier J, Buerki S, Arrigo N, Kupfer P, Alvarez N (2009) Genetic structure and evolution of Alpine polyploid complexes: *Ranunculus kuepferi* (Ranunculaceae) as a case study. *Molecular Ecology*, **18**, 3730 3744.

Caballero A, Quesada H (2010) Homoplasy and distribution of AFLP fragments: an analysis in silico of the genome of different species. *Molecular Biology and Evolution*, **27**, 1139 1151.

de Carvalho JF, Poulain J, Da Silva C *et al.* (2013) Transcriptome de novo assembly from next generation sequencing and comparative analyses in the hexaploid salt marsh species *Spartina maritima* and *Spartina alterniflora* (Poaceae). *Heredity*, **110**, 181 193.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171 182.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124 3140.

Chelaifa H, Monnier A, Ainouche M (2010) Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species *Spartina* × *townsendii* and *Spartina anglica* (Poaceae). *New Phytologist*, **186**, 161 174.

Chen J, Wang J, Tian L *et al.* (2004) The development of an Arabidopsis model system for genome wide analysis of polyploidy effects. *Biological Journal of the Linnean Society. Linnean Society of London*, **82**, 689.

Clark LV, Jasieniuk M (2011) polysat: an R package for polyploid microsatellite analysis. *Molecular Ecology Resources*, **11**, 562 566.

Cockerham CC (1973) Analyses of gene frequencies. *Genetics*, **74**, 679 700.

Crow KD, Stadler PF, Lynch VJ, Amemiya C, Wagner GP (2006) The "fish specific" Hox cluster duplication is coincident with the origin of teleosts. *Molecular Biology and Evolution*, **23**, 121 136.

Csilléry K, Blum MGB, Gaggiotti OE, François O (2010) Approximate Bayesian computation (ABC) in practice. *Trends in Ecology and Evolution*, **25**, 410 418.

Culling MA, Janko K, Boron A *et al.* (2006) European colonization by the spined loach (*Cobitis taenia*) from Ponto Caspian refugia based on mitochondrial DNA variation. *Molecular Ecology*, **15**, 173 190.

Dakin E, Avise J (2004) Microsatellite null alleles in parentage analysis. *Heredity*, **93**, 504 509.

Danzmann RG, Bogart JP (1982) Evidence for a polymorphism in gametic segregation using a malate dehydrogenase locus in the tetraploid treefrog *Hyla versicolor*. *Genetics*, **100**, 287 306.

De Silva HN, Hall AJ, Rikkerink E, McNeilage MA, Fraser LG (2005) Estimation of allele frequencies in polyploids under certain patterns of inheritance. *Heredity*, **95**, 327 334.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**, 1 38.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next generation DNA sequencing data. *Nature Genetics*, **43**, 491 498.

Dufresne F, Hebert PDN (1995) Polyploidy and clonal diversity in an arctic cladoceran. *Heredity*, **75**, 45 53.

Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next generation sequencing. *TAG Theoretical and Applied Genetics*, **126**, 1 11.

Egan AN, Schlueter J, Spooner DM (2012) Applications of next generation sequencing in plant biology. *American Journal of Botany*, **99**, 175 185.

Esselink GD, Nybom H, Vosman B (2004) Assignment of allelic configuration in polyploids using the MAC PR (microsatellite DNA allele counting peak ratios) method. *TAG Theoretical and Applied Genetics*, **109**, 402 408.

Evans BJ, Kelley DB, Tinsley RC, Melnick DJ, Cannatella DC (2004) A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution. *Molecular Phylogenetics and Evolution*, **33**, 197 213.

Evans BJ, Kelley DB, Melnick DJ, Cannatella DC (2005) Evolution of RAG 1 in polyploid clawed frogs. *Molecular Biology and Evolution*, **22**, 1193 1207.

Evans BJ, Bliss SM, Mendel SA, Tinsley RC (2011) The Rift Valley is a major barrier to dispersal of African clawed frogs (*Xenopus*) in Ethiopia. *Molecular Ecology*, **20**, 4216 4230.

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

Everett M, Miller M, Seeb J (2012) Meiotic maps of sockeye salmon derived from massively parallel DNA sequencing. *BMC Genomics*, **13**, 521.

Ewing B, Hillier L, Wendl MC, Green P (1998) Base calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome Research*, **8**, 175 185.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479 491.

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.

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574 578.

Fawcett JA, Maere S, Van de Peer Y (2009) Plants with double genomes might have had a better chance to survive the Cretaceous Tertiary extinction event. *Proceedings of the National Academy of Sciences USA*, **106**, 5737 5742.

Fay M, Cowan R, Leitch I (2005) The effects of nuclear DNA content (C value) on the quality and utility of AFLP fingerprints. *Annals of Botany*, **95**, 237 246.

Ferris SD, Whitt GS (1977) Loss of duplicate gene expression after polyploidization. *Nature*, **265**, 258 260.

Flagel LE, Wendel JF (2010) Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytologist*, **186**, 184 193.

Flagel L, Udall J, Nettleton D, Wendel J (2008) Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. *BMC Biology*, **6**, 16.

Gaeta RT, Pires JC, Iniguez Luy F, Leon E, Osborn TC (2007) Genomic changes in resynthesized *Brassica napus* and their effect on gene expression and phenotype. *The Plant Cell Online*, **19**, 3403 3417.

Gao H, Williamson S, Bustamante CD (2007) A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics*, **176**, 1635 1651.

García Verdugo C, Calleja JA, Vargas P *et al.* (2013) Polyploidy and microsatellite variation in the relict tree *Prunus lusitanica* L.: how effective are refugia in preserving genotypic diversity of clonal taxa? *Molecular Ecology*, **22**, 1546 1556.

Gautier M, Gharbi K, Cezerd T *et al.* (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165 3178.

Geiringer H (1949) Chromatid segregation of tetraploids and hexaploids. *Genetics*, **34**, 665.

Gholami M, Bekele WA, Schondelmaier J, Snowdon RJ (2012) A tailed PCR procedure for cost effective, two order multiplex sequencing of candidate genes in polyploid plants. *Plant Biotechnology Journal*, **10**, 635 645.

Gibson G (2002) Microarrays in ecology and evolution: a preview. *Molecular Ecology*, **11**, 17 24.

Gower JC (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325 338.

Gregory TR, Mable BK (2005) Polyploidy in animals. In: *The Evolution of the Genome* (ed. Gregory TR), pp. 427 517. Elsevier Academic Press, Burlington, Massachusetts.

Griffin PC, Robin C, Hoffmann AA (2011) A next generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to *Poa* grasses. *BMC Biology*, **9**, 19.

Guichoux E, Lagache L, Wagner S *et al.* (2011) Current trends in microsatellite genotyping. *Molecular Ecology Resources*, **11**, 591 611.

Guo YP, Vogl C, Van Loo M, Ehrendorfer F (2005) Hybrid origin and differentiation of two tetraploid *Achillea* species in East Asia: molecular, morphological and ecogeographical evidence. *Molecular Ecology*, **15**, 133 144.

Haldane JBS (1930) Theoretical genetics of autopolyploids. *Journal of Genetics*, **22**, 359 372.

Hale MC, McCormick CR, Jackson JR, DeWoody JA (2009) Next generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics*, **10**, 203.

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618 620.

Hardy OJ, De Loose M, Vekemans X, Meerts P (2001) Allozyme segregation and inter cytotype reproductive barriers in the polyploid complex *Centaurea jacea*. *Heredity*, **87**, 136 145.

Hartigan JA, Wong MA (1979) Algorithm AS 136: a K Means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**, 100 108.

Hedrén M, Fay MF, Chase MW (2001) Amplified fragment length polymorphisms (AFLP) reveal details of polyploid evolution in *Dactylorhiza* (Orchidaceae). *American Journal of Botany*, **88**, 1868 1880.

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

Hegarty MJ, Barker GL, Wilson ID *et al.* (2006) Transcriptome shock after interspecific hybridization in Senecio is ameliorated by genome duplication. *Current Biology*, **16**, 1652 1659.

Hegarty MJ, Barker GL, Brennan AC *et al.* (2008) Changes to gene expression associated with hybrid speciation in plants: further insights from transcriptomic studies in Senecio. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 3055 3069.

Hegarty M, Barker G, Brennan A *et al.* (2009) Extreme changes to gene expression associated with homoploid hybrid speciation. *Molecular Ecology*, **18**, 877.

Hijmans RJ, Gavrilenko T, Stephenson S *et al.* (2007) Geographical and environmental range expansion through polyploidy in wild potatoes (*Solanum* section Petota). *Global Ecology and Biogeography*, **16**, 485 495.

Hillis DM, Davis SK (1988) Ribosomal DNA: intraspecific polymorphism, concerted evolution, and phylogeny reconstruction. *Systematic Zoology*, **37**, 63 66.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Hohenlohe P, Amish S, Catchen J, Allendorf F, Luikart G (2011) Next generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117.

Hol WHG, van der Wurff AWG, Skøt L, Cook R (2008) Two distinct AFLP types in three populations of marram grass (*Ammophila arenaria*) in Wales. *Plant Genetic Resources: Characterization and Utilization*, **6**, 201 207.

Holland PWH, Garcia Fernàndez J, Williams NA, Sidow A (1994) Gene duplications and the origins of vertebrate development. *Development*, **1994**, 125 133.

Hollister JD, Arnold BJ, Svedin E *et al.* (2012) Genetic adaptation associated with genome doubling in autotetraploid *Arabidopsis arenosa*. *PLoS Genetics*, **8**, e1003093.

Holloway AK, Cannatella DC, Gerhardt HC, Hillis DM (2006) Polyploids with different origins and ancestors form a single sexual polyploid species. *The American Naturalist*, **167**, E88 E101.

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417.

Huang S, Sirikhachornkit A, Su X *et al.* (2002) Genes encoding plastid acetyl CoA carboxylase and 3 phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proceedings of the National Academy of Sciences USA*, **99**, 8133 8138.

Husband BC, Sabara HA (2003) Reproductive isolation between autotetraploids and their diploid progenitors in fireweed, *Chamerion angustifolium* (Onagraceae). *New Phytologist*, **161**, 703 713.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254 267.

Jackson S, Chen ZJ (2010) Genomic and expression plasticity of polyploidy. *Current Opinion in Plant Biology*, **13**, 153 159.

Jaillon O, Aury JM, Noel B *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463 467.

Jakobsson M, Hagenblad J, Tavaré S *et al.* (2006) A unique recent origin of the allotetraploid species *Arabidopsis suecica*: evidence from nuclear DNA markers. *Molecular Biology and Evolution*, **23**, 1217 1231.

Jannoo N, Grivet L, David J, D'Hont A, Glaszmann JC (2004) Differential chromosome pairing affinities at meiosis in polyploidy sugarcane revealed by molecular markers. *Heredity*, **93**, 460 467.

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

Jombart T, Ahmed IØ (2011) adegenet 1.3 1: new tools for the analysis of genome wide SNP data. *Bioinformatics*, **27**, 3070 3071.

Jombart T, Pontier D, Dufour A (2009) Genetic markers in the playground of multivariate analysis. *Heredity*, **102**, 330 341.

Jombart T, Devillard Sb, Balloux Fo (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.

Jørgensen MH, Ehrich D, Schmickl R, Koch MA, Brysting AK (2011) Interspecific and interploidal gene flow in Central European *Arabidopsis* (Brassicaceae). *BMC Evolutionary Biology*, **11**, 346.

Jørgensen MH, Lagesen K, Mable BK, Brysting AK (2012) Using high throughput sequencing to investigate the evolution of self incompatibility genes in the Brassicaceae: strategies and challenges. *Plant Ecology & Diversity*, **5**, 473 484.

Jost L (2008) GST and its relatives do not measure differentiation. *Molecular Ecology*, **17**, 4015 4026.

Jupe F, Witek K, Verweij W *et al.* (2013) Resistance gene enrichment sequencing (RenSeq) enables re annotation of the NB LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *The Plant Journal*, **76**, 530 544.

Kamiri M, Stift M, Sraiti I, Costantino G, *et al.* (2011) Evidence for non disomic inheritance in a Citrus interspecific tetraploid somatic hybrid between *C. reticulata* and *C. limon* using SSR markers and cytogenetic analysis. *Plant Cell Reports*, **30**, 1415 1425.

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, **428**, 617 624.

Kolář F, Fér T, Štech M *et al.* (2012) Bringing together evolution on serpentine and polyploidy: spatiotemporal history of the diploid tetraploid complex of *Knautia arvensis* (Dipsacaceae). *PLoS ONE*, **7**, e39988.

Koning Boucoiran CF, Gitonga VW, Yan Z, Dolstra O, *et al.* (2012) The mode of inheritance in tetraploid cut roses. *Theoretical Applied Genetics*, **125**, 591 607.

Kosman E, Leonard KJ (2005) Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Molecular Ecology*, **14**, 415 424.

Koukalova B, Moraes AP, Renny Byfield S *et al.* (2010) Fall and rise of satellite repeats in allopolyploids of *Nicotiana* over 5 million years. *New Phytologist*, **186**, 148 160.

Krak K, Caklova P, Chrtek J, Fehrer J (2013) Reconstruction of phylogenetic relationships in a highly reticulate group with deep coalescence and recent speciation (*Hieracium*, Asteraceae). *Heredity*, **110**, 138 151.

Kreitman M (1987) Molecular population genetics. *Oxford Surveys in Evolutionary Biology*, **4**, 38 60.

Lai K, Duran C, Berkman PJ *et al.* (2012) Single nucleotide polymorphism discovery from wheat next generation sequence data. *Plant Biotechnology Journal*, **10**, 743 749.

Lampert K, Schartl M (2008) The origin and evolution of a unisexual hybrid: *Poecilia formosa*. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 2901 2909.

Landergott U, Naciri Y, Schneller JJ, Holderegger R (2006) Allelic configuration and polysomic inheritance of highly variable microsatellites in tetraploid gynodioecious *Thymus praecox*. *Theoretical and Applied Genetics*, **113**, 453 465.

Legendre P, Legendre L (1998) *Numerical Ecology*, 2nd English edn. Elsevier Science, Amsterdam.

Leitch A, Leitch I (2008) Genomic plasticity and the diversity of polyploid plants. *Science*, **320**, 481 483.

Lo EYY, Stefanović S, Dickinson TA (2009) Population genetic structure of diploid sexual and polyploid apomictic hawthorns (*Crataegus*; Rosaceae) in the Pacific Northwest. *Molecular Ecology*, **18**, 1145 1160.

Ludwig A, Belfiore NM, Pitra C, Svirsky V, Jenneckens I (2001) Genome duplication events and functional reduction of ploidy levels in sturgeon (*Acipenser*, *Huso* and *Scaphirhynchus*). *Genetics*, **158**, 1203 1215.

Luttikhuizen PC, Stift M, Kuperus P, Van Tienderen PH (2007) Genetic diversity in diploid vs. tetraploid *Rorippa amphibia* (Brassicaceae). *Molecular Ecology*, **16**, 3544 3553.

Lynch M (1990) The similarity index and DNA fingerprinting. *Molecular Biology and Evolution*, **7**, 478 484.

Lynch M (2009) Estimation of allele frequencies from high coverage genome sequencing projects. *Genetics*, **182**, 295 301.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151 1155.

Ma J X, Li Y N, Vogl C, Ehrendorfer F, Guo Y P (2010) Allopolyploid speciation and ongoing backcrossing between diploid progenitor and tetraploid progeny lineages in the *Achillea millefolium* species complex: analyses of single copy nuclear genes and genomic AFLP. *BMC Evolutionary Biology*, **10**, 100.

Mable BK (2004a) Polyploidy and self compatibility: is there an association? *New Phytologist*, **162**, 803 811.

Mable BK (2004b) 'Why polyploidy is rarer in animals than in plants': myths and mechanisms. *Biological Journal of the Linnean Society*, **82**, 453 466.

Mable B, Beland J, Di Berardo C (2004) Inheritance and dominance of self incompatibility alleles in polyploid *Arabidopsis lyrata*. *Heredity*, **93**, 476 486.

Mable BK, Alexandrou MA, Taylor MI (2011) Genome duplication in amphibians and fish: an extended synthesis. *Journal of Zoology*, **284**, 151 182.

Madlung A (2013) Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity*, **99**, 372 382.

Markwith SH, Stewart DJ, Dyer JL (2006) TETRASAT: a program for the population analysis of allotetraploid microsatellite data. *Molecular Ecology Notes*, **6**, 586 589.

Masterson J (1994) Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science*, **264**, 421 424.

Mavarez J, Audet C, Bernatchez L (2009) Major disruption of gene expression in hybrids between young sympatric anadromous and resident populations of brook charr (*Salvelinus fontinalis* Mitchill). *Journal of Evolutionary Biology*, **22**, 1708 1720.

Mayer KF, Martis M, Hedley PE *et al.* (2011) Unlocking the barley genome by chromosomal and comparative genomics. *The Plant Cell Online*, **23**, 1249 1263.

McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next generation DNA sequencing data. *Genome Research*, **20**, 1297 1303.

McQuown E, Gall G, May B (2002) Characterization and inheritance of six microsatellite loci in lake sturgeon. *Transactions of the American Fisheries Society*, **131**, 299 307.

Meirmans PG, Hedrick PW (2011) Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources*, **11**, 5 18.

Meirmans PG, Van Tienderen PH (2004) Genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, **4**, 792 794.

Meirmans PG, Van Tienderen PH (2013) The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity*, **110**, 131 137.

Menken SBJ, Smit E, Den Nijs HCM (1995) Genetical population structure in plants: gene flow between diploid sexual and triploid asexual dandelions (*Taraxacum* section Ruderalia). *Evolution*, **49**, 1108 1118.

Meudt HM, Clarke AC (2007) Almost forgotten or latest practice? AFLP applications, analyses and advances. *Trends in Plant Science*, **12**, 106 117.

Mraz P, Spaniel S, Keller A *et al.* (2012) Anthropogenic disturbance as a driver of microspatial and microhabitat segregation of cytotypes of *Centaurea stoebe* and cytotype interactions in secondary contact zones. *Annals of Botany*, **110**, 615 627.

Munzbergová Z, Šurinová M, Castro S (2013) Absence of gene flow between diploids and hexaploids of *Aster amellus* at multiple spatial scales. *Heredity*, **110**, 123 130.

Nei M (1972) Genetic distance between populations. *American Naturalist*, **106**, 283 292.

Nei M (1973) Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences USA*, **70**, 3321 3323.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York, New York.

Neiman M, Paczesniak D, Soper DM, Baldwin AT, Hehman G (2011) Wide variation in ploidy level and genome size in a New Zealand freshwater snail with coexisting sexual and asexual lineages. *Evolution*, **65**, 3202 3216.

Obbard DJ, Harris SA, Pannell JR (2006) Simple allelic phenotype diversity and differentiation statistics for allopolyploids. *Heredity*, **97**, 296 303.

Ogden R, Gharbi K, Mugue N *et al.* (2013) Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. *Molecular Ecology*, **22**, 3112 3123.

Ohno S (1970) *Evolution by Gene Duplication*, Springer Verlag, New York, New York.

O'Neill EM, Schwartz R, Bullock CT *et al.* (2013) Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology*, **22**, 111 129.

Otto SP, Whitton J (2000) Polyploid incidence and evolution. *Annual Review of Genetics*, **34**, 401 437.

Parisod C, Salmon A, Zerjal T *et al.* (2009) Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in *Spartina*. *New Phytologist*, **184**, 1003 1015.

Parisod C, Holderegger R, Brochmann C (2010) Evolutionary consequences of autopolyploidy. *New Phytologist*, **186**, 5 17.

Parkin IA, Clarke WE, Sidebottom C *et al.* (2010) Towards unambiguous transcript mapping in the allotetraploid *Brassica napus*. *Genome*, **53**, 929 938.

Parsons P (1959) Some problems in inbreeding and random mating in tetrasomics. *Agronomy Journal*, **51**, 465 467.

Paun O, Stuessy TF, Horandl E (2006) The role of hybridization, polyploidization and glaciation in the origin and evolution of the apomictic *Ranunculus cassubicus* complex. *New Phytologist*, **171**, 223 236.

Pearson K (1901) On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**, 559 572.

Pembleton LW, Cogan NO, Forster JW (2013) StAMPP: an R package for calculation of genetic differentiation and structure of mixed ploidy level populations. *Molecular Ecology Resources*, **13**, 946 952.

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

Piednoel M, Aberer AJ, Schneeweiss GM *et al.* (2012) Next generation sequencing reveals the impact of repetitive DNA across phylogenetically closely related genomes of Orobanchaceae. *Molecular Biology and Evolution*, **29**, 3601 3611.

Pignatta D, Dilkes BP, Yoo SA *et al.* (2010) Differential sensitivity of the *Arabidopsis thaliana* transcriptome and enhancers to the effects of genome doubling. *New Phytologist*, **186**, 194 206.

Postlethwait JH, Woods IG, Ngo Hazelett P *et al.* (2000) Zebrafish comparative genomics and the origins of vertebrate chromosomes. *Genome Research*, **10**, 1890 1902.

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

Ptacek MB, Gerhardt HC, Sage RD (1994) Speciation by polyploidy in treefrogs: multiple origins of the tetraploid, *Hyla versicolor*. *Evolution*, **48**, 898 908.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.

R Development Core Team (2004) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3 900051 07 0. Available from: http://www.R project.org.

Roder M, Plaschke J, Konig S (1995) Abundance, variability and chromosomal location of microsatellites in wheat. *Molecular Genomics and Genetics*, **246**, 327 333.

Ronfort Jl, Jenczewski E, Bataillon T, Rousset Fo (1998) Analysis of population structure in autotetraploid species. *Genetics*, **150**, 921 930.

Rouse MN, Saleh AA, Seck A *et al.* (2011) Genomic and resistance gene homolog diversity of the dominant tallgrass prairie species across the US Great Plains precipitation gradient. *PLoS ONE*, **6**, e17641.

Rowe H, Renaut S, Guggisberg A (2011) RAD in the realm of next generation sequencing technologies. *Molecular Ecology*, **20**, 3499 3502.

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics*, **15**, 174 175.

Saitoh K, Chen W J, Mayden RL (2010) Extensive hybridization and tetrapolyploidy in spined loach fish. *Molecular Phylogenetics and Evolution*, **56**, 1001 1010.

Sall T, Jakobsson M, Lind Halldén C, Halldén C (2003) Chloroplast DNA indicates a single origin of the allotetraploid *Arabidopsis suecica*. *Journal of Evolutionary Biology*, **16**, 1019 1029.

Salmon A, Ainouche ML, Wendel JF (2005) Genetic and epigenetic consequences of recent hybridization and polyploidy in *Spartina* (Poaceae). *Molecular Ecology*, **14**, 1163.

Salmon A, Flagel L, Ying B, Udall JA, Wendel JF (2010) Homoeologous nonreciprocal recombination in polyploid cotton. *New Phytologist*, **186**, 123 134.

Salmon A, Udall JA, Jeddeloh JA, Wendel J (2012) Targeted capture of homoeologous coding and noncoding sequence in polyploid cotton. *G3: Genes Genomes Genetics*, **2**, 921 930.

Saltonstall K (2003) Microsatellite variation within and among North American lineages of *Phragmites australis*. *Molecular Ecology*, **12**, 1689 1702.

Sampson JF, Byrne M (2012) Genetic diversity and multiple origins of polyploid *Atriplex nummularia* Lindl. (Chenopodiaceae). *Biological Journal of the Linnean Society*, **105**, 218 230.

Sánchez Navarro B, Jokela J, Michiels NK, D'Souza TG (2013) Population genetic structure of parthenogenetic flatworm populations with occasional sex. *Freshwater Biology*, **58**, 416 429.

Schmickl R, Jørgensen MH, Brysting AK, Koch MA (2008) Phylogeographic implications for the North American boreal arctic *Arabidopsis lyrata* complex. *Plant Ecology and Diversity*, **1**, 245 254.

Seeb J, Pascal C, Grau E *et al.* (2011a) Transcriptome sequencing and high resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids. *Molecular Ecology Resources*, **11**, 335 348.

Seeb JE, Carvalho G, Hauser L *et al.* (2011b) Single nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Molecular Ecology Resources*, **11**, 1 8.

Segraves K, Thompson J, Soltis P, Soltis D (1999) Multiple origins of polyploidy and the geographic structure of *Heuchera grossulariifolia*. *Molecular Ecology*, **8**, 253 262.

Serang O, Mollinari M, Garcia AAF (2012) Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. *PLoS ONE*, **7**, e30906.

Shimizu Inatsugi RIE, Lihovà J, Iwanaga H *et al.* (2009) The allopolyploid Arabidopsis kamchatica originated from multiple individuals of *Arabidopsis lyrata* and *Arabidopsis halleri*. *Molecular Ecology*, **18**, 4024 4048.

Shiu S, Borevitz J (2006) The next generation of microarray research: applications in evolutionary and ecological genomics. *Heredity*, **100**, 141 149.

Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139**, 457 462.

Slatkin M, Barton NH (1989) A comparison of three indirect methods for estimating average levels of gene flow. *Evolution*, **43**, 1349 1368.

Slotte T, Holm K, McIntyre LM, Lagercrantz U, Lascoux M (2007) Differential expression of genes important for adaptation in *Capsella bursa pastoris* (Brassicaceae). *Plant Physiology*, **145**, 160 173.

Smouse PE, Peakall R (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, **82**, 561 573.

Sneath PA, Sokal RR (1973) *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. H. Freeman and Co., San Francisco, California.

Soltis DE, Soltis PS (2000) Contributions of plant molecular systematics to studies of molecular evolution. *Plant Molecular Biology*, **42**, 45 75.

Soltis DE, Soltis PS, Ness BD (1989) Chloroplast DNA variation and multiple origins of autopolyploidy in *Heuchera micrantha* (Saxifragaceae). *Evolution*, **43**, 650 656.

Soltis DE, Soltis PS, Pires JC *et al.* (2004) Recent and recurrent polyploidy in *Tragopogon* (Asteraceae): cytogenetic, genomic and genetic comparisons. *Biological Journal of the Linnean Soci ety*, **82**, 485 501.

Sommer S, Courtiol A, Mazzoni CJ (2013) MHC genotyping of non model organisms using next generation sequencing: a new methodology to deal with artefacts and allelic dropout. *BMC Genomics*, **14**, 542.

Stebbins GL Jr (1947) Types of polyploids: their classification and significance. *Advances in Genetics*, **1**, 403 429.

Stenberg P, Saura A (2013) Meiosis and its deviations in poly ploid animals. *Cytogenetic and Genome Research*, **140**, 185 203.

Stenberg P, Lundmark M, Knutelski S, Saura A (2003) Evolu tion of clonality and polyploidy in a weevil system. *Molecu lar Biology and Evolution*, **20**, 1626 1632.

Stift M, Berenos C, Kuperus P, Van Tienderen PH (2008) Segre gation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to *Rorippa* (yellow cress) microsatellite data. *Genetics*, **179**, 2113 2123.

Stift M, Bregman R, Oostermeijer JGB, van Tienderen PH (2010) Other tetraploid species and conspecific diploids as sources of genetic variation for an autotetraploid. *American Journal of Botany*, **97**, 1858 1866.

Stock M, Lamatsch DK (2013) Trends in polyploidy research in animals and plants. *Cytogenetics Genome Research*, **140**, 71 312.

Stock M, Steinlein C, Lamatsch DK, Schartl M, Schmid M (2005) Multiple origins of tetraploid taxa in the Eurasian *Bufo viridis* subgroup. *Genetica*, **124**, 255 272.

St Onge KR, Foxe JP, Li J *et al.* (2012) Coalescent based analysis distinguishes between allo and autopolyploid origin of Shepherd's Purse (*Capsella bursa pastoris*). *Molecular Biology and Evolution*, **29**, 1721 1733.

Sunnucks P (2000) Efficient genetic markers for population biology. *Trends in Ecology & Evolution*, **15**, 199 203.

Suomalainen E, Saura A, Lokki J (1987) *Cytology and Evolution in Parthenogenesis*. CRC Press, Boca Raton, Florida.

Swofford DL, Olsen GJ, Waddel PJ, Hillis DM (1996) Phyloge netic inference. In: *Molecular Systematics*, 2nd edn (eds Hillis D. M., Moritz C. & Mable B. K.), pp. 407 514. Sinauer Associates, Sunderland, Massachusetts.

Tachida H, Yoshimaru H (1996) Genetic diversity in partially selfing populations with the stepping stone structure. *Hered ity*, **77**, 469 475.

Talavera Ma, Navarro Sampedro L, Ortiz PL, Arista M (2013) Phylogeography and seed dispersal in islands: the case of *Rumex bucephalophorus* subsp. *canariensis* (Polygonaceae). *Annals of Botany*, **111**, 249 260.

Tate J, Soltis D, Soltis P (2005) Polyploidy in plants. In: *The Evolution of the Genome* (ed. Gregory RT). Elsevier, San Diego, California.

Tomiuk J, Loeschcke V (1991) A new measure of genetic iden tity between populations of sexual and asexual species. *Evo lution*, **45**, 1685 1694.

Tomiuk J, Loeschcke V (1992) Evolution of parthenogenesis in the *Otiorhynchus scaber* complex. *Heredity*, **68**, 391 397.

Tomiuk J, Guldbrandtsen B, Loeschcke V (2009) Genetic similarity of polyploids: a new version of the computer program POPDIST (version 1.2.0) considers intraspecific genetic differentiation. *Molecular Ecology Resources*, **9**, 1364 1368.

Tsigenopoulos CS, Rab P, Naran D, Berrebi P (2002) Multiple origins of polyploidy in the phylogeny of southern African barbs (Cyprinidae) as inferred from mtDNA markers. *Hered ity*, **88**, 466 473.

Tsuchimatsu T, Kaiser P, Yew C L, Bachelier JB, Shimizu KK (2012) Recent loss of self incompatibility by degradation of the male component in allotetraploid *Arabidopsis kamchatica*. *PLoS Genetics*, **8**, e1002838.

Tuskan GA, Difazio S, Jansson S *et al.* (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596 1604.

Twyford AD, Ennos RA (2012) Next generation hybridization and introgression. *Heredity*, **108**, 179 189.

Van Dijk P, Bakx Schotman T (1997) Chloroplast DNA phylog eography and cytotype geography in autopolyploid *Plantago media*. *Molecular Ecology*, **6**, 345 352.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) Micro checker: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes*, **4**, 535 538.

Van Oosterhout C, Weetman D, Hutchinson W (2006) Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. *Molecular Ecology Notes*, **6**, 255 256.

Van Puyvelde K, Van Geert A, Triest L (2010) ATETRA, a new software program to analyse tetraploid microsatellite data: comparison with TETRA and TETRASAT. *Molecular Ecology Resources*, **10**, 331 334.

Vanderpoorten A, Hardy OJ, Lambinon J, Raspé O (2011) Two reproductively isolated cytotypes and a swarm of highly inbred, disconnected populations: a glimpse into Salicornia's evolutionary history and challenging taxonomy. *Journal of Evolutionary Biology*, **24**, 630 644.

Vergilino R, Belzile C, Dufresne F (2009) Genome size evolu tion and polyploidy in the *Daphnia pulex* complex (Clado cera: Daphniidae). *Biological Journal of the Linnean Society*, **97**, 68 79.

Vergilino R, Markova S, Ventura M, Manca M, Dufresne F (2011) Reticulate evolution of the *Daphnia pulex* complex as revealed by nuclear markers. *Molecular Ecology*, **20**, 1191 1207.

Vijay N, Poelstra JW, Kunstner A, Wolf JB (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA seq experiments. *Molecular Ecology*, **22**, 620 634.

Voorrips R, Gort G, Vosman B (2011) Genotype calling in tetra ploid species from bi allelic marker data using mixture mod els. *BMC Bioinformatics*, **12**, 172.

Waller DM, Knight SE (1989) Genetic consequences of outcross ing in the cleistogamous annual, Impatiens capensis. II. Outcrossing rates and genotypic correlations. *Evolution*, **43**, 860 869.

Wang J, Tian L, Lee H S *et al.* (2006) Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. *Genetics*, **172**, 507 517.

Wang Y, Zeng X, Iyer NJ *et al.* (2012) Exploring the switchgrass transcriptome using second generation sequencing technol ogy. *PLoS ONE*, **7**, e34225.

Wang N, Thomson M, Bodles WJA *et al.* (2013) Genome sequence of dwarf birch (*Betula nana*) and cross species RAD markers. *Molecular Ecology*, **22**, 3098 3111.

Weir B, Cockerham CC (1984) Estimating F statistics for the analysis of population structure. *Evolution*, **38**, 1358 1370.

Wendel JF, Schnabel A, Seelanan T (1995) Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proceedings of the National Academy of Sciences USA*, **92**, 280 284.

Whitlock MC (2011) G_{ST} and D do not replace F_{ST}. *Molecular Ecology*, **20**, 1083 1091.

Winfield MO, Wilkinson PA, Allen AM *et al.* (2012) Targeted re sequencing of the allohexaploid wheat exome. *Plant Biotechnology Journal*, **10**, 733 742.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114 138.

Wright S (1946) Isolation by distance under diverse systems of mating. *Genetics*, **31**, 39 59.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323 354.

Wright S (1965) The interpretation of population structure by F statistics with special regard to systems of mating. *Evolution*, **19**, 395 420.

Wu S B, Wirthensohn MG, Hunt P, Gibson JP, Sedgley M (2008) High resolution melting analysis of almond SNPs derived from ESTs. *Theoretical and Applied Genetics*, **118**, 1 14.

Wu L L, Cui X K, Milne RI, Sun Y H, Liu J Q (2010) Multiple autopolyploidizations and range expansion of *Allium przewalskianum* Regel. (Alliaceae) in the Qinghai Tibetan Plateau. *Molecular Ecology*, **19**, 1691 1704.

Yoo M, Szadkowski E, Wendel J (2013) Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity*, **110**, 171 180.

You FM, Huo N, Deal KR *et al.* (2011) Annotation based genome wide SNP discovery in the large and complex *Aegilops tauschii* genome using next generation sequencing without a reference genome sequence. *BMC Genomics*, **12**, 59.

Zhang DX, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, **12**, 563 584.

The initial plan for the review came from discussions between F.D., R.V. and B.K.M. at Evolution 2011 in Ottawa, with M.S. subsequently contributing valuable suggestions for improvement of content and structure. All authors contributed extensively to conceptual content, manuscript structure, and writing.