Heredity (2006) 97, 296–303

www.nature.com/hdy

Simple allelic-phenotype diversity and differentiation statistics for allopolyploids

DJ Obbard, SA Harris and JR Pannell
Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

The analysis of genetic diversity within and between populations is a routine task in the study of diploid organisms. However, population genetic studies of polyploid organisms have been hampered by difficulties associated with scoring and interpreting molecular data. This occurs because the presence of multiple alleles at each locus often precludes the measurement of genotype or allele frequencies. In allopolyploids, the problem is compounded because genetically distinct isoloci frequently share alleles. As a result, analysis of genetic diversity patterns in allopolyploids has tended to rely on the interpretation of phenotype frequencies, which loses information available from allele composition. Here, we propose the use of a simple allelic-phenotype diversity statistic (H') that measures diversity as the average number of alleles by which pairs of individuals differ. This statistic can be extended to a population differentiation measure (F'_{ST}), which is analogous to F_{ST}. We illustrate the behaviour of these statistics using coalescent computer simulations that show that F'_{ST} behaves in a qualitatively similar way to F_{ST}, thus providing a useful way to quantify population differentiation in allopolyploid species.
Heredity (2006) **97,** 296–303. doi:10.1038/sj.hdy.6800862; published online 5 July 2006

Keywords: disomic inheritance; FDASH; F_{ST}; genetic differentiation; genetic diversity; polyploidy

Introduction

A primary aim of population genetics is the measurement of genetic diversity and the characterisation of its hierarchical distribution among individuals, populations, or groups of populations. For molecular markers with a clear genetic interpretation such as microsatellites, isozymes and DNA sequences, widely used measures of diversity include allelic richness (A), gene diversity (H_e – 'expected heterozygosity', see eg Hartl and Clark, 1997), and, for DNA sequence data, the proportion of pairwise site differences (π). Several different methods have been used to quantify genetic differentiation among groups, most notably differentiation statistics related to F_{ST} (Wright, 1951; Weir and Cockerham, 1984). Although the utility of F_{ST} may be limited for highly diverse loci (eg Nagylaki, 1998), and its interpretation in terms of simple population genetic models is questionable (Whitlock and McCauley, 1999), F_{ST} and related measures continue to be quoted almost universally in studies of population genetic structure.

F_{ST} and the other summary statistics cited above have been very widely employed to quantify patterns of genetic variation in diploid organisms. However, many organisms are polyploid. Indeed, although polyploidy is particularly common among plants, fish, and amphibians, it is also found among birds, mammals, and many invertebrates (Leitch and Bennett, 1997; Otto and Whitton, 2000; Legatt and Iwama, 2003).

Correspondence: DJ Obbard. Current address: Institute of Evolutionary Biology, University of Edinburgh, Ashworth Labs, Kings Buildings, West Mains Road, Edinburgh, Midlothian EH9 3JT, UK.
E-mail: darren.obbard@ed.ac.uk
Received 30 November 2005; accepted 5 June 2006; published online 5 July 2006

Polyploids are often conceptually divided into two groups: autopolyploids and allopolyploids (reviewed by Ramsey and Schemske, 2002). Autopolyploids are derived from the duplication of a single genome and – at least initially – there is no significant differentiation between duplicate genomes. Conversely, allopolyploids are derived from interspecific hybridisation, and therefore comprise two (or more) differentiated genomes. However, as species boundaries are rarely clear-cut, auto- and allopolyploidy actually represent opposite ends along a spectrum of intergenome differentiation, with 'hybrids' between differentiated lineages from within a single species forming the middle ground.

Polyploids can be further divided by their mode of inheritance. In newly formed autopolyploids, duplicated chromosomes do not have unique partners at meiosis. Chromosomes either pair at random or form multivalents (described as 'polysomic' inheritance), such that a tetraploid individual carrying alleles *ABCD* can form gametes *AB*, *AC*, *AD*, *BC*, *BD*, and *CD* (reviewed by Bever and Felber, 1992; Olson, 1997; Ronfort *et al*, 1998). Furthermore, recombination in multivalents can lead to sister chromatids segregating together, giving rise to *AA*, *BB*, *CC*, and *DD* alleles ('Double reduction', see discussion in Ronfort *et al*, 1998). Unlike newly formed autopolyploids, allopolyploids have differentiated pairs of chromosomes, which are often able to pair normally, as in the diploid progenitors (described as 'disomic' inheritance). As with auto- and allopolyploidy, disomic and polysomic inheritance constitute extremes from a continuum. This is both because allopolyploid hybrids between close relatives may allow for some multivalent formation, and because autopolyploids diploidise over time, eventually developing fully disomic inheritance (Wolfe, 2001; Ramsey and Schemske, 2002). Indeed,

intermediate modes of inheritance may be probabilistic, with particular loci having a higher or lower probability of pairing with a partially differentiated partner (the 'pairing preference', eg Wu *et al*, 2001).

Unfortunately, the quantification of genetic diversity and population differentiation for organisms with polyploid genomes can be much more difficult than for diploids (Figure 1). In a polyploid, and unlike the diploid case, multiple alleles can be present in more than one copy. For example, a diploid carrying alleles *A* and *B* at a locus must have one copy of each, whereas a tetraploid carrying alleles *A* and *B* ('allelic phenotype' *AB*) may have any one of three different genotypes: *AAAB*, *AABB*, or *ABBB*.

In some species, and particularly for low-order polyploids (eg tetraploids), it is possible to estimate this allele copy number ('dosage') on the basis of band intensity or electropherogram peak height (eg Arft and Ranker, 1998; Prober *et al*, 1998; Young *et al*, 1999; Hardy and Vekemans, 2001; Nassar *et al*, 2003). When this is the case, and inheritance is polysomic, the genotype follows directly from the allelic phenotype as it does in diploids. Consequently, extensions of standard diploid summary statistics such as H_e and F_{ST} can be used to quantify genetic diversity and differentiation (Nei, 1987; Ronfort *et al*, 1998; Thrall and Young, 2000), and computer programs are available to conduct analyses (SPAGEDi – Hardy and Vekemans, 2002; AUTOTET – Thrall and Young, 2000). However, in higher order polyploids, the genotype can rarely be inferred from gel banding patterns or electropherograms (eg Kahler *et al*, 1980; Krebs and Hancock, 1989; Brochmann *et al*, 1992).

A further difficulty arises when the target polyploid population displays disomic inheritance, because it is then often not clear which alleles are associated with which of the duplicate loci (homeologous loci, or 'isoloci'). This problem applies both to autopolyploids that have become diploidised, and to allopolyploids in which alleles from each of the two parental species segregate in a diploid manner at different isoloci (see Figure 1 for an illustration). In both cases, although the genetic segregation is effectively diploid, it is typically very difficult or impossible to know whether a particular allele, scored as a band on a gel, is segregating at which of two or more isoloci. A tetraploid genotype that produces two distinct bands on a gel, for example, may be homozygous for different alleles at each of its two isoloci, or heterozygous at one or both loci. At one extreme, all individuals may have the same heterozygous genotype (heterozygosity is 'fixed'; Figure 1c), whereas at the other extreme, several alleles may be shared among isoloci, making disomic inheritance superficially appear polysomic (Figure 1b); this latter situation has been termed 'cryptic disomy' (De Silva *et al*, 2005).

In situations where allele dosage can be scored for populations with disomic inheritance, the underlying allele frequencies can sometimes be estimated for different isoloci using the superficial genotypes, and these estimates used to calculate genetic diversity statistics (Waples, 1988; Hedrick *et al*, 1991; Bouza *et al*, 2001). Recently, De Silva *et al* (2005) have provided a sophisticated approach to estimating allele frequencies in both polysomic and disomic polyploids when allele dosage cannot be scored. However, in order to provide good estimates of allele frequencies with these methods, populations must be assumed to be at equilibrium, and independent estimates of the selfing rate are required.

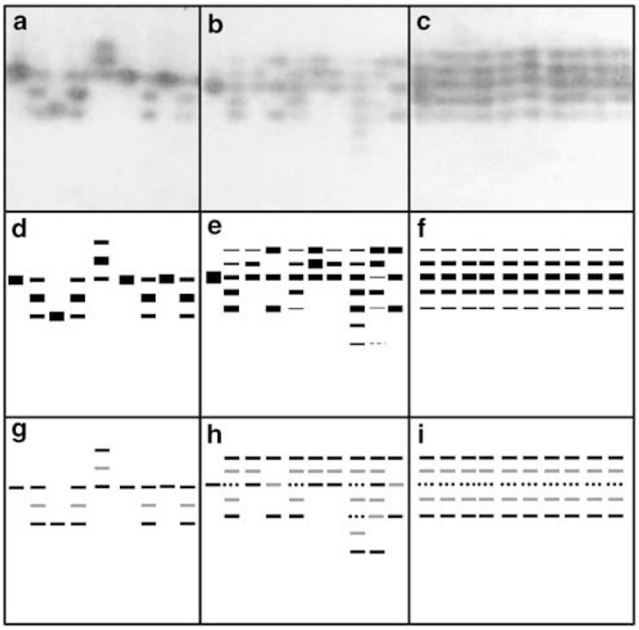

Figure 1 Isozyme banding patterns (allelic phenotypes) for glucose-6-phosphate isomerase (PGI, E.C. 5.3.1.9) in the weedy annual plant, *Mercurialis annua* L. (Euphorbiaceae). Left to right are: (**a, d, g**) diploid *M. annua*; (**b, e, h**) allohexaploid *M. annua* displaying cryptic disomy; and (**c, f, i**) allohexaploid *M. annua* showing fixed heterozygosity. The rows are: (**a–c**) gel photo; (**d–f**) interpretation of band presence and approximate intensity; and (**g–i**) allelic interpretation. PGI functions as a homodimer. Thus, when two alleles with different electrophoretic mobilities are present, three bands are visible because an interallele (hetero)dimer with intermediate mobility is formed (grey lines in the lower panels). For example, panel **a** shows three alleles (*f*, *m*, and *s*): lane one is genotype *mm*, lane two genotype *ms*, lane three is *ss*, and lane five is *fm*. When two different alleles are present in equal copy number (as in diploid heterozygotes), the three bands are expected to have intensity ratios of 1:2:1, as seen in panel a lanes two, four, five, seven, and nine. Hexaploid gels (**b** and **c**) may be considerably more complex. In panel **b**, lane one is a homozygote (with six copies of allele *m*), whereas lane four is a straightforward heterozygote (alleles *f* and *s*). Problems arise when more than two alleles are present, such as in panel **b**, lanes two and five, which are both heterozygous with alleles *f*, *m* and *s* (note that the *fs* heterodimer has identical mobility to the *mm* homodimer; superimposed hetero- and homodimers are indicated dotted lines in panels **g–i**). The difference between these two lanes is in their allele copy number; lane two has more copies of allele *s* than allele *f* (it may have genotype *fmmsss*), whereas lane five probably has more copies of allele *m* (eg *fmmmms*). Similarly, lanes six and seven also differ only in copy number; lane six has approximately equal numbers of alleles *f* and *m*, whereas lane seven is 'unbalanced' towards *m*. This illustrates the two fundamental problems met when scoring polyploid genotypes: first, it is difficult to know the exact dosage (is lane seven *ffmmmm* or *fmmmmm*?), and second, given that these individuals display disomic inheritance, it is impossible to assign alleles to isoloci (if lane seven does have alleles *ffmmmm*, is it genotype *ff*, *mm*, *mm*, or *fm*, *fm*, *mm*?). Our solution is to avoid both issues by merely recording which alleles an individual carries: its 'allelic phenotype'. Thus, in panel **b**, the allelic phenotypes are: *m*, *fms*, *fm*, *fs*, *fms*, *fm*, *fm*, *fmsv*, *fmv*, and *fs*, and in panel **c** all phenotypes are *fms*.

An alternative approach has been to interpret polyploid gel banding patterns as allelic phenotypes

(Figure 1), and to calculate simple summary statistics on the basis of gel phenotypic diversity without recourse to a full genetic interpretation (eg Jain and Singh, 1979; Gaur *et al*, 1980; Murdy and Carter, 1985; Bayer and Crawford, 1986; Chung *et al*, 1991; Brochmann *et al*, 1992; Rogers, 2000; Berglund and Westerbergh, 2001). With this approach, diversity can be measured in terms of the total number of different banding or allelic phenotypes in the population, or by calculating statistics similar to H_e (Nei's gene diversity, the probability that two alleles sampled at random are different) on the basis of allelic phenotype frequencies. Two such statistics have been widely used: H^{Phen}, which is calculated as one minus the sum of squared phenotype frequencies, and is thus analogous to H_e (Yunus *et al*, 1991; Meerts *et al*, 1998), and H^{SW}, which is a Shannon–Weaver diversity index of phenotypes (eg Jain and Singh, 1979; Gaur *et al*, 1980; Chung *et al*, 1991). Both these measures can be used to calculate population differentiation as the ratio of between-population to species-wide diversity, analogous to F_{ST}. However, because they treat gel phenotypes only as being either identical or different, they do not make use of all the information present on a gel, for example, they do not recognise the greater similarity of phenotypes that share more bands over those that share fewer.

Recently, Meirmans and van Tienderen (2004) and Meirmans (2004) have used several measures of inter-individual similarity, including (1) the number of steps to convert one phenotype into the other, and (2) a measure related to the Dice coincidence index (Dice, 1945), calculated as the number of shared bands (alleles) between two individuals, divided by the total number of bands present. Bruvo *et al* (2004) took a similar approach for microsatellite data using the number of stepwise mutations that separate allelic phenotypes. However, the behaviour of none of these measures has been compared with that of F_{ST}.

Here, we introduce a new simple measure of allelic-phenotype diversity devised for use in allopolyploid species. This diversity measure (denoted H') accounts for the fact that allelic phenotypes may share differing numbers of bands (alleles). Specifically, H' is defined as the average number of alleles by which pairs of individuals differ at a single locus; thus

$$H' = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k \in \{\text{alleles}\}} x_{ijk}$$

where n is the total number of individuals and x_{ijk} equals one if allele k is carried by either individual i or by individual j (but not by both), and is otherwise equal to zero. Although devised for use with allopolyploids, it may be possible also to apply this statistic to other forms of polyploid (see Discussion).

Based on this measure of diversity, we define a differentiation statistic, F'_{ST}, as $(H'_T - H'_S)/H'_T$. A computer program, 'FDASH', which uses allelic phenotype data to calculate the above statistics is available from the authors upon request. Below, we compare the statistical behaviour of H' and F'_{ST} with that of H^{Phen}, H^{SW}, and their associated differentiation statistics, ${}^{P}F_{ST}$ and ${}^{SW}F_{ST}$. Because it is not clear how statistics derived from allelic phenotype data will respond to demographic processes such as migration, we use coalescent simulations in a preliminary exploration of their behaviour under the simple 'island model' of population structure. Although the island model is an unrealistic caricature of a subdivided population (Whitlock and McCauley, 1999), the expectation for F_{ST} under a given migration rate is known, and it thus provides firm ground on which to test a new statistic. We also consider the extent to which the statistics are affected by the polyploid level and the degree of differentiation between isoloci, for example, the degree divergence between the parental species of an allopolyploid population.

Methods

Model

We assumed an island model of population structure. Under this simple model, a (meta)population is divided into d discrete subpopulations or demes, each of size N. Each generation, a proportion m of the individuals in each deme are replaced by migrants drawn randomly from the rest of the metapopulation. Migration is haploid, as by pollen in a diploid organism. For m much greater than u, the mutation rate to new alleles, the expected value of F_{ST} is $1/(1+4Nm)$; this provides a simple expectation against which to compare differentiation statistics using phenotype-based diversity measures under focus here. In an infinite-allele framework, the sharing of alleles between isoloci must be owing to common ancestry; for example, in an ancestor of the parental taxa (in the case of allopolyploidy) or before the onset of disomic inheritance (in the case of a now-diploidised autopolyploid).

We used coalescence-based simulations (eg Hudson, 1990; Nordborg and Donnelly, 1997; Nordborg, 2001) to compare genotype- and phenotype-based statistics in terms of their response to polyploid level, their deviation from the expectation of F_{ST}, and their variance. In particular, we followed (Wakeley, 2001; Wakeley and Aliacar, 2001) in separating the coalescent process into two parts: an evolutionarily rapid 'scattering phase', in which lineages coalesce within demes or migrate out of them, and a slow 'collecting phase', in which lineages from different demes first migrating into common demes and then eventually coalesce at their common ancestor. The collecting phase is just a neutral coalescent, with the effective population size scaled to account for population structure (Wakeley and Aliacar, 2001; Rousset, 2003).

We modelled populations of allopolyploids with disomic inheritance by explicitly recognising that the multiple genetically distinct pairs of homeologous loci or isoloci share a common ancestral locus in the distant past. Thus, we considered an initial sample of lineages at time zero consisting of $2x$-ploid individuals, with $x=2$ for tetraploids, $x=3$ for hexaploids, and so on, and recorded migration and coalescence events as the simulation proceeded backwards in time towards increasingly more inclusive common ancestors. The sample thus passed through the scattering phase and entered the collecting phase with x simultaneous coalescent processes, one for each independent isolocus. After a given point in time, coalescence was then allowed to occur between lineages from different isoloci. This threshold determines the extent to which isoloci share

alleles by descent; it corresponds either to the speciation event that separated the two parental species of the simulated allopolyploid population or to the point at which polysomic inheritance became disomic through diploidisation. If the threshold is ancient, isoloci will share no alleles and the markers will be effectively diploid (ie a paleopolyploid); in contrast, if the threshold is recent, then isoloci will share alleles, and banding patterns may look superficially like polysomic inheritance (cryptic disomy).

In each simulation run, a given number of $2x$-ploid individuals were sampled from each of several demes (see below). A genealogy for the sample was simulated, and a Poisson-distributed number of mutations was applied to each branch, with the parameter proportional to branch length and mutation rate, and the mutation process following assumptions of the infinite-alleles model (as appropriate, for example, for isozymes). Finally, the allelic state of each of the sampled alleles was identified, and diversity and differentiation statistics were calculated for the sample.

We calculated differentiation statistics based on phenotype frequencies, as described above. We also calculated a genotype-based estimate of F_{ST} (θ), following Weir (1996), with multilocus (ie multi-isolocus) estimates calculated as a ratio of averages. For comparison, the same genotype-based statistic was also calculated as if the polyploid had polysomic inheritance, that is, a single locus with four alleles rather than two isoloci with two alleles each (as described by Ronfort *et al*, (1998)). We calculated the expectation of F_{ST} for the island model as $1/(1+4Nm)$. To assess the quality of F_{ST} estimators, we used the mean square error of estimates, calculated as the sum of squared bias and the variance (ie bias^2 + var) (Balloux and Goudet, 2002).

Model parameters

For all simulations, the population comprised 500 demes each of 250 (polyploid) individuals, and samples consisted of 25 individuals drawn from each of 10 demes. For each parameter combination, the simulation was repeated 20 000 times to estimate statistic means and variances. To examine the effect of ploidy on H'_T, F'_{ST}, and the other phenotype-based statistics, we simulated diploids, tetraploids, and hexaploid populations with three different levels of divergence between isoloci. These were chosen such that low divergence ($0.01 \times 2N_e$ generations) resulted in most alleles occurring at all isoloci (cryptic disomy), and high divergence ($100 \times 2N_e$ generations) resulted in alleles almost never occurring at multiple isoloci (paleopolyploidy). Following an initial search of parameter space, we chose a migration rate ($m = 0.0062$) and a mutation rate ($\mu = 5.7 \times 10^{-5}$) that yielded numbers of observed alleles ($A = 1.99$) and values of population differentiation ($F_{ST} = 0.197$) corresponding to the means reported for isozymes in outcrossing plants (Hamrick and Godt, 1990). To examine the relative utility of genotype- and phenotype-based differentiation statistics across a range of migration rates, we ran the simulation for tetraploids only using a single level of divergence between isoloci ($2N_e$ generations) and the same parameters otherwise, with expected F_{ST} values between 0.025 and 0.995.

Results

The response to increasing polyploidy

As expected, the genotype-based genetic diversity statistic H, calculated as an average across isoloci for the whole population (ie Nei's gene diversity; allele dosage scored and alleles attributed to isoloci), did not vary with increasing polyploid level or increasing differentiation between isoloci (Figure 2a). In contrast,

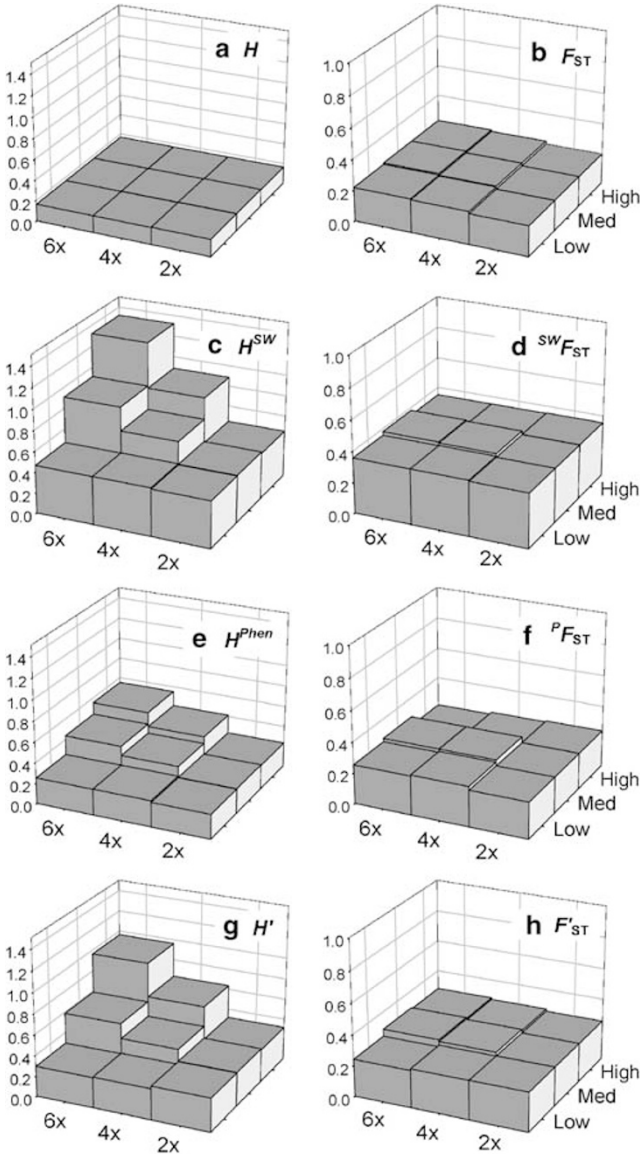

Figure 2 Genetic diversity (**a**, **c**, **e**, **g**) and differentiation (**b**, **d**, **f**, **h**) statistics for polyploids with disomic inheritance, under an island model of population structure. Samples of 250 individuals (25 each from 10 demes) were drawn from a structured population of 500 demes each of 250 individuals; values are the average of 20 000 replicates. Statistics were calculated from genotypic data (H; **a** and **b**), Shannon–Weaver diversity (H^{SW}; **c** and **d**), Phenotype frequencies (H^{Phen}; **e** and **f**) and allele differences (H'; **g** and **h**), and are plotted with respect to polyploid level ($2x$–$6x$) and differentiation between isoloci (low, medium, and high; see main text). Phenotype-based diversity increases with polyploid level and differentiation between isoloci, whereas at this intermediate migration rate, differentiation statistics are largely unaffected by polyploidy.

the genetic diversity statistics based on phenotype data (a function of diversity across multiple isoloci) increased with the level of polyploidy and the degree of differentiation between isoloci. This was true of both phenotype diversity measures calculated from phenotype frequencies (H^{SW} and H^{Phen}; Figure 2c and e, respectively), and the measure of diversity based on allele differences (H'; Figure 2g). The genotype-based differentiation statistic, θ (Weir and Cockerham, 1984), calculated across isoloci, did not vary greatly with polyploid level or differentiation between isoloci (Figure 2b). For this intermediate level of differentiation, there was almost no variation in the phenotypic differentiation statistics based on phenotype frequency ($^{SW}F_{ST}$ and $^{P}F_{ST}$), or allele differences (F'_{ST}), (Figures 2d, f, and h, respectively).

The response to increasing migration rate

Qualitative effects: Three of the differentiation statistics deviated qualitatively from the expected F_{ST} in their response to migration rate (Figure 3). The differentiation statistics calculated from allelic-phenotype diversity ($^{SW}F_{ST}$ and $^{P}F_{ST}$) did not approach zero as the migration rate increased (Figure 3a), although under the parameters explored here the discrepancy from expectation was extremely small for $^{P}F_{ST}$ (Figure 3a). This was not seen for F'_{ST} (Figure 3a, grey line). When the polysomic genotype-based estimate θ (Ronfort *et al*, 1998) was calculated, as might be performed erroneously in the case of cryptic disomy, it did not tend towards one as the migration rate decreased (Figure 3b).

Phenotypes in place of genotypes: The use of phenotypic data also led to quantitative differences in measures of differentiation. To examine the relative loss of information associated with the use of allelic phenotype data in place of genotype data, the phenotype-based differentiation statistic, F'_{ST}, was considered as an estimator of expected F_{ST}. As would be predicted, the tetraploid genotypic estimate, which requires alleles to be assigned to isoloci (two isoloci, dashed line in Figure 4) was always better than the diploid genotypic estimate (one locus, dot-dash line in Figure 4), reflecting the information gained from using two loci for the estimate in place of one. Under the parameters examined here, when differentiation was low (ie expected value of $F_{ST} < ca.$ 0.5) genotype-based multilocus θ appeared to be a better estimator of F_{ST} than was the phenotype-based F'_{ST}. However, when differentiation was high, F'_{ST} and θ were very similar (Figure 4).

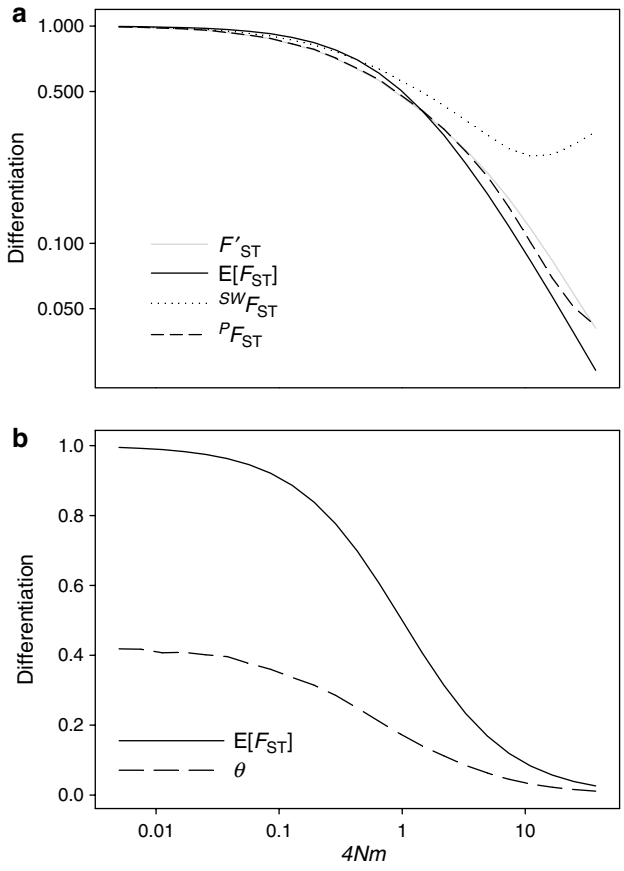

Figure 3 (**a**) To illustrate a qualitative deviation from expected F_{ST}, differentiation statistics based on phenotype frequencies were plotted against the migration parameter $4Nm$. A log scale is used for F_{ST} to highlight the fact that $^{SW}F_{ST}$ (and possibly also $^{P}F_{ST}$) is not asymptotic to zero as migration rates increase (see main text for details). This does not seem to happen for F'_{ST} (grey line), which was otherwise similar to $^{P}F_{ST}$. (**b**) As expected, a qualitative deviation is also seen if polysomic θ (Ronfort *et al*, 1998) is calculated for a disomic system (dashed line *versus* solid line). This is because apparent heterozygosity, actually owing to differences between isoloci, is treated as if it were genuine diversity, thereby inflating subpopulation diversity, and leading to low estimates of differentiation (see main text). For both (**a**) and (**b**), samples of 250 individuals (25 each from 10 demes) were drawn from an island model of population structure, with 500 demes, each of 250 tetraploid individuals, and values given are the average of 20 000 replicates, calculated on the basis of coalescent simulations.

Discussion

The principal aim of our study was to compare statistics based on phenotype frequencies (H^{Phen} and H^{SW}), with one based on phenotypic similarity (H'), and to illustrate the extent to which these statistics are informative about population structure. Our results suggest that allelic phenotype-based diversity statistics for polyploids with disomic inheritance may depend strongly on details such as the polyploid level (number of isoloci) and the differentiation between isoloci (Figure 2). In contrast, differentiation statistics did not appear to be strongly affected by polyploid level (Figure 2). Differentiation statistics, when calculated as though inheritance were polysomic, and when calculated from allelic phenotype diversity, differed qualitatively from the island-model expectation of F_{ST} (Figures 2 and 3a). Below, we discuss the likely reason for these effects, and the implications for quantifying diversity and differentiation in polyploids with disomic inheritance.

Appropriate models of inheritance

When fixed heterozygosity is identified for multiple loci in a polyploid population, it is clear that inheritance must be disomic (Figure 1c). However, when isoloci share a large proportion of their alleles, the great inter-individual variation in the number of distinct alleles can make gel banding patterns look superficially polysomic

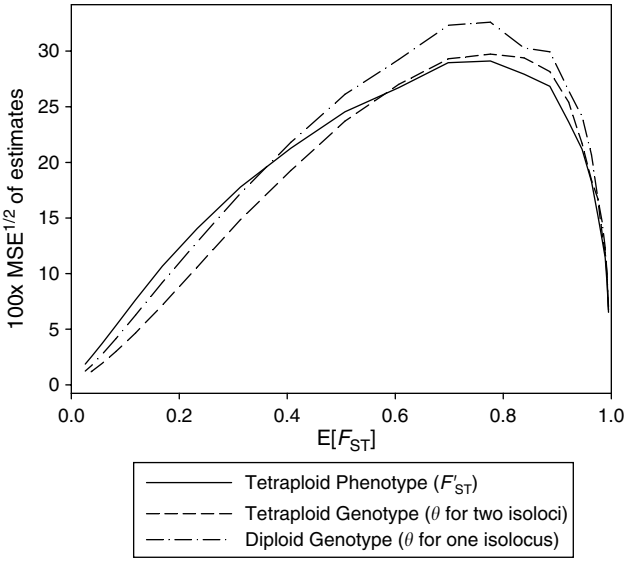

Figure 4 To illustrate quantitative deviations from expected F_{ST}, the mean square error of genotype-based (θ) and phenotype-based (F'_{ST}) differentiation statistics is plotted for a range of expected F_{ST} values (high to low migration rates). For high migration rates, allelic phenotype-based estimates of F_{ST} appear marginally worse than genotype-based estimates. Indeed, estimates are apparently worse than the single-locus case, despite the increase in information available from an additional isolocus (solid line *versus* dot-dash line). However, when the migration rate is low, phenotype- and genotype-based differentiation statistics are approximately equal in their ability to estimate F_{ST} (solid *versus* dashed line). Statistics were calculated from 20 000 replicates; samples were of 250 individuals (25 from each of 10 demes) drawn from an island model of population structure with 500 demes each of 250 tetraploid individuals. For other parameters, see Methods.

(ie disomic inheritance is cryptic; Figure 1b). It is tempting to analyse such data using computer packages intended for autopolyploids (eg SPAGEDi: Ronfort *et al*, 1998; Hardy and Vekemans, 2002), without confirming that inheritance is polysomic. However, this procedure is inappropriate, because the apparent excess of heterozygotes (owing to disomic inheritance) will artificially inflate within-population diversity (H_S), so that it is nonzero even when there are no differences between individuals within populations. Thus, the polyploid analogue of θ, calculated under the assumption that inheritance is polysomic (Ronfort *et al*, 1998), may be very small in a polyploid population with disomic inheritance, even when migration rates are almost zero (Figure 3b).

Utility of phenotype-based estimates of genetic diversity in disomic polyploids

Phenotype-based diversity statistics depend strongly on the number of isoloci (ie the polyploid level) and the degree of differentiation between isoloci (Figure 2). This is because phenotype-based diversity statistics simultaneously record the diversity at several duplicate isoloci. If isoloci shared no alleles, the overall phenotype diversity would be an additive function of diversity at each of the (diploid) isoloci, and thus should increase with the polyploid level (Figure 2). By contrast, genetic differentiation statistics do not vary much with polyploid level, given that other population parameters are the same (Figure 2). This is because differentiation statistics, such as F_{ST}, are essentially a ratio of within-population diversity to total diversity, and they will thus be affected relatively little by factors that simultaneously increase both. This means that, although direct comparisons of diversity statistics such as H' and H^{SW} cannot be made between polyploid levels, comparisons of differentiation statistics derived from them are likely to be informative.

Some of the phenotype-based differentiation statistics behave unexpectedly in response to migration, that is, they differ qualitatively from expected F_{ST} or genotype-based statistics (Figure 3). In particular, the differentiation statistic derived from the Shannon Weaver diversity of phenotype frequencies ($^{SW}F_{ST}$) does not approach zero with increasing migration. We believe this is an effect of finite sample size; if F_{ST} is considered as a standardised variance in allele frequencies between populations (eg Weir, 1996), the variance in phenotype frequencies will be larger for a given sample size than the variance in allele frequencies. This is because alleles will be distributed differently between individuals in different samples, and unless the sample is very large, many rare phenotypes will not be included. The effect is strong for $^{SW}F_{ST}$ because H^{SW} weights rare phenotypes disproportionately highly. However, there is also some suggestion that under high migration rates a small effect is seen for $^{P}F_{ST}$ (dashed line, Figure 3a). We therefore suggest that inference regarding relative migration rates, when based on differentiation statistics calculated from phenotype frequencies, should be treated with some caution as differentiation statistics may be appreciably greater than zero even under panmictic gene flow. F'_{ST}, the differentiation statistic based on allele differences, does not appear to suffer from this limitation.

Although polyploidy presents a number of challenges, the concomitant increase in the number of loci in principle has the potential to provide more information for making inferences about population processes, such as migration. It is therefore interesting to ask whether the information gain associated with the availability of more (iso)loci outweighs the information lost through the use of allelic phenotype data in place of genetic data. The results of our simulations suggest that gains do indeed tend to balance the losses. Thus, under the parameters we examined, there was an overall loss in information when migration rates were high (ie with low differentiation), but no appreciable loss when migration rates were low (Figure 4).

Conclusions

We have made a preliminary investigation of the behaviour of phenotype-based statistics in a simple island model, for outcrossing polyploids with disomic inheritance. Our study suggests that for many purposes, the diversity statistic H' is an informative way of summarising genetic diversity in disomic polyploids, and that the differentiation statistic derived from it (F'_{ST}) behaves in a very similar way to other more widely used differentiation statistics. Furthermore, F'_{ST} seems very little affected by polyploid level in disomic polyploids, so that comparisons between polyploid levels, and potentially among species that differ in ploidy, are most likely to be viable.

The statistics we have introduced here are, of course, purely descriptive. However, differences in diversity and differentiation might be tested for statistical significance using randomisation procedures. For example, by randomising population samples between 'treatments', we were able to infer a difference in H'_S and F'_{ST} between different sexual systems in allohexaploid *Mercurialis annua* (Obbard *et al*, 2006). For some of the polyploid statistics discussed here, this randomisation procedure has been implemented in our program 'FDASH' (available on request).

The statistics F'_{ST} and H' were devised for use in polyploids with disomic inheritance (eg most allopolyploids), thus our simulations were limited to a disomic model of inheritance and assumed an infinite-allele model of mutation. However, although information from allele frequencies will be lost, H' and F'_{ST} should also capture essential information regarding genetic diversity in polyploids with other modes of inheritance, such as polysomic polyploids, and cases for which allele sharing may not be owing to common ancestry, such as microsatellite markers (for an application see Refoufi and Esnault, 2006). In particular, it is worth noting that patterns of allele sharing under polysomic inheritance are superficially similar to disomic inheritance (as in Figure 4, 'low divergence'), so long as there is little divergence between isoloci. This would suggest that H' and F'_{ST} may be applicable to alternative modes of inheritance, although further work is required to test this. It will also important to determine how these statistics behave under more complex population models, such as those that include selfing or metapopulation processes. The coalescent approach adopted here will be ideally suited to making these extensions (Nordborg and Donnelly, 1997; Wakeley and Aliacar, 2001).

Acknowledgements

We thank Gil McVean for advice on coalescent simulation and for suggesting the diversity measure H'. We thank Pete Hollingsworth and two anonymous referees for valuable comments on this paper. DJO was funded by a Long Studentship from the Queens College, Oxford. SAH and JRP acknowledge support from the Natural Environment Research Council, UK.

References

Arft AM, Ranker TA (1998). Allopolyploid origin and population genetics of the rare orchid *Spiranthes diluvialis*. *Am J Bot* **85**: 110–122.

Balloux F, Goudet J (2002). Statistical properties of population differentiation estimators under stepwise mutation in a finite island model. *Mol Ecol* **11**: 771–783.

Bayer RJ, Crawford DJ (1986). Allozyme divergence among five diploid species of *Antennaria* (Asteraceae: Inuleae) and their allopolyploid derivatives. *Am J Bot* **73**: 287–296.

Berglund ABN, Westerbergh A (2001). Two postglacial immigration lineages of the polyploid *Cerastium alpinum* (Caryophyllaceae). *Hereditas* **134**: 171–183.

Bever JD, Felber F (1992). The theoretical population genetics of autopolyploidy. *Oxford Surv Evolut Biol* **8**: 185–217.

Bouza C, Castro J, Sanchez L, Martinez P (2001). Allozymic evidence of parapatric differentiation of brown trout (*Salmo trutta* L.) within an Atlantic river basin of the Iberian Peninsula. *Mol Ecol* **10**: 1455–1469.

Brochmann C, Soltis DE, Soltis PS (1992). Electrophoretic relationships and phylogeny of Nordic polyploids in *Draba* (Brassicaceae). *Plant Syst Evol* **182**: 35–70.

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

Chung MG, Hamrick JL, Jones SB, Derda GS (1991). Isozyme variation within and among populations of *Hosta* (Liliaceae) in Korea. *Syst Bot* **16**: 667–684.

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.

Dice LR (1945). Measures of the amount of ecologic association between species. *Ecology* **26**: 297–302.

Gaur PK, Lichtwardt RW, Hamrick JL (1980). Isozyme variation among soil isolates of *Histoplasma capsulatum*. *Exp Mycol* **5**: 69–77.

Hamrick JL, Godt MJW (1990). Allozyme diversity in plant species. In: Brown AHD, Clegg MT, Kahler AL, Weir BS (eds) *Plant Population Genetics, Breeding, and Genetic Resources*. Sinauer: Sunderland, MA. pp 43–63.

Hardy OJ, Vekemans X (2001). Patterns of allozyme variation in diploid and tetraploid *Centaurea jacea* at different spatial scales. *Evolution* **55**: 943–954.

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

Hartl DL, Clark AG (1997). *Principles of Population Genetics*. Sinauer Associates: Sunderland, MA.

Hedrick PW, Hutchinson ES, Mesler M (1991). Estimation of self-fertilization rate and allelic frequencies in diploidized tetraploids. *Heredity* **67**: 259–264.

Hudson RR (1990). Gene genealogies and the coalescent process. In: Futuyma DJ, Antonovics J (eds) *Oxford Surveys in Evolutionary Biology*. Oxford University Press: Oxford, **7**: 1–44.

Jain SK, Singh RS (1979). Population biology of *Avena*. VII. Allozyme variation in relation to the genome analysis. *Bot Gaz* **140**: 356–363.

Kahler AL, Allard RW, Krzakowa M, Wehrhahn CF, Nevo E (1980). Associations between isozyme phenotypes and environment in the slender Wild Oat (*Avena barbata*) in Israel. *Theor Appl Genet* **56**: 31–47.

Krebs SL, Hancock JF (1989). Tetrasomic inheritance of isoenzyme markers in the highbush blueberry, *Vaccinium corymbosum* L. *Heredity* **63**: 11–18.

Legatt RA, Iwama GK (2003). Occurrence of polyploidy in the fishes. *Rev Fish Biol Fish* **13**: 237–246.

Leitch IJ, Bennett MD (1997). Polyploidy in angiosperms. *Trends Plant Sci* **2**: 470–476.

Meerts P, Baya T, Lefebvre C (1998). Allozyme variation in the annual weed species complex *Polygonum aviculare* (Polygonaceae) in relation to ploidy level and colonizing ability. *Plant Syst Evol* **211**: 239–256.

Meirmans PG. (2004). GenoType and GenoDive: Users Manual. Retrieved July 1st 2004, from http://staff.science.uva.nl/~meirmans/softindex.html.

Meirmans PG, van Tienderen PH (2004). GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Mol Ecol Notes* **4**: 792–794.

Murdy WH, Carter MEB (1985). Electrophoretic study of the allopolyploidal origin of *Talinum teretifolium* and the specific status of *T. appalachianum* (Portulacaceae). *Am J Bot* **72**: 1590–1597.

Nagylaki T (1998). Fixation indices in subdivided populations. *Genetics* **148**: 1325–1332.

Nassar JM, Hamrick JL, Fleming TH (2003). Population genetic structure of Venezuelan chiropterophilous columnar cacti (Cactaceae). *Am J Bot* **90**: 1628–1637.

Nei M (1987). *Molecular Evolutionary Genetics*. University Press: New York, Columbia.
Nordborg M (2001). Coalescent Theory. In: Balding DJ (ed) *Handbook of Statistical Genetics*. John Wiley, NY. pp 179–212.
Nordborg M, Donnelly P (1997). The coalescent process with selfing. *Genetics* **146**: 1185–1195.
Obbard DJ, Harris SA, Pannell JR (2006). Sexual systems and population genetic structure in an annual plant: testing the metapopulation model. *Am Nat* **167**: 354–366.
Olson MS (1997). Bayesian procedures for discriminating among hypotheses with discrete distributions: Inheritance in the tetraploid *Astilbe biternata*. *Genetics* **147**: 1933–1942.
Otto SP, Whitton J (2000). Polyploid incidence and evolution. *Annu Rev Genet* **34**: 401–437.
Prober SM, Spindler LH, Brown AHD (1998). Conservation of the grassy white box woodlands: effects of remnant population size on genetic diversity in the allotetraploid herb *Microseris lanceolata*. *Conserv Biol* **12**: 1279–1290.
Ramsey J, Schemske DW (2002). Neopolyploidy in flowering plants. *Annu Rev Ecol Syst* **33**: 589–639.
Refoufi A, Esnault MA (2006). Genetic diversity and population structure of *Elytrigia pycnantha* (Godr.) (Triticeae) in Mont Saint–Michel Bay using microsatellite markers. *Plant Biol* **8**: 234–242.
Rogers DL (2000). Genotypic diversity and clone size in old-growth populations of coast redwood (*Sequoia sempervirens*). *Can J Bot-Rev Can Bot* **78**: 1408–1419.
Ronfort J, Jenczewski E, Bataillon T, Rousset F (1998). Analysis of population structure in autotetraploid species. *Genetics* **150**: 921–930.
Rousset F (2003). Effective size in simple metapopulation models. *Heredity* **91**: 107–111.
Thrall PH, Young A (2000). AUTOTET: A program for analysis of autotetraploid genotypic data. *J Hered* **91**: 348–349.
Wakeley J (2001). The coalescent in an island model of population subdivision with variation among demes. *Theor Popul Biol* **59**: 133–144.
Wakeley J, Aliacar N (2001). Gene genealogies in a metapopulation. *Genetics* **159**: 893–905.
Waples RS (1988). Estimation of allele frequencies at isoloci. *Genetics* **118**: 371–384.
Weir BS (1996). *Genetic Analysis II*. Sinauer: Sunderland, MA.
Weir BS, Cockerham CC (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.
Whitlock MC, McCauley DE (1999). Indirect measures of gene flow and migration: *F*-ST not equal 1/(4*Nm*+1). *Heredity* **82**: 117–125.
Wolfe KH (2001). Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* **2**: 333–341.
Wright S (1951). The genetical structure of populations. *Ann Eugenics* **15**: 323–354.
Wu RL, Gallo-Meagher M, Littell RC, Zeng ZB (2001). A general polyploid model for analyzing gene segregation in outcrossing tetraploid species. *Genetics* **159**: 869–882.
Young AG, Brown AHD, Zich FA (1999). Genetic structure of fragmented populations of the endangered daisy *Rutidosis leptorrhynchoides*. *Conserv Biol* **13**: 256–265.
Yunus AG, Jackson MT, Catty JP (1991). Phenotypic polymorphism of 6 enzymes in the Grasspea (*Lathyrus sativus* L.). *Euphytica* **55**: 33–42.