MOLECULAR ECOLOGY

Molecular Ecology (2010)

NEWS AND VIEWS

COMMENT

Do microsatellites reflect genome-wide genetic diversity in natural populations? A comment on Väli *et al.* (2008)

MARCUS LJUNGQVIST, MIKAEL ÅKESSON and BENGT HANSSON

Department of Animal Ecology, Ecology Building, Lund University, S-223 62 Lund, Sweden

Abstract

A recent study by Väli et al. (2008) highlights that microsatellites will often provide a poor prediction of the genome-wide nucleotide diversity of wild populations, but does not fully explain why. To clarify and stress the importance of identity disequilibrium and marker variability for correlations between multilocus heterozygosity and genome-wide genetic variability, we performed a simple simulation with different types of markers, corresponding to microsatellites and SNPs, in populations with different inbreeding history. The importance of identity disequilibrium was apparent for both markers and there was a clear impact of marker variability.

Keywords: genetic diversity, heterozygosity, identity disequilibrium, inbreeding, microsatellites

Received 20 October 2009; revision received 3 December 2009; accepted 7 December 2009

Introduction

A main aim in conservation genetics is to unravel the link between genetic diversity and population viability (Frankham *et al.* 2002). To reach this goal we need to understand the causes and consequences of genetic variation in natural populations, which requires transparent methods to quantify genetic diversity within and between populations. On a genome-wide scale, genetic variation is distributed over thousands of genes and non-coding genomic regions. Ideally, evaluations of the genetic status of populations would include data on every aspect of genetic variation, including (i) the quantitative genetic variation of adaptive traits (Reed & Frankham 2001, 2003); (ii) the molecular genetic variation at critical functional loci (Hughes 1991; Miller & Lambert 2004); and (iii) the genetic variation over genome-wide distributed coding and non-coding loci (Hansson & Westerberg 2002; Slate *et al.* 2004). For practical reasons, we are still unable to quantify this diversity of genetic variation in studies of natural populations. Instead, we need to rely on proxies of the overall genetic variation, often inferred from a limited set of molecular markers. Microsatellites have been the marker of choice in many conservation genetic studies over the last decade (Frankham 1995; Beaumont & Bruford 1999; Coltman & Slate 2003).

A recent study by Väli *et al.* (2008) highlights that microsatellites will often provide a poor prediction of the nucleotide diversity elsewhere in the genome of wild populations, which can have important implications for conservation genetic studies. Väli *et al.* (2008) investigated to what extent a set of microsatellites (10–27 loci) explained the nucleotide diversity at 10 intronic loci (in total, approximately 5.5 kb) within and between eight mammalian populations of four species; grey wolves (*Canis lupus*), coyotes (*Canis latrans*), Eurasian lynx (*Lynx lynx*) and wolverine (*Gulo gulo*). In brief, they found a strong positive correlation ($r^2 = 0.70$) between microsatellite heterozygosity and nucleotide diversity at the population level, but very weak correlations at the individual level within populations ($r^2 \leq 0.09$). Also, they noted that estimates of nucleotide diversity varied 30-fold (7.1×10^{-5}–2.1×10^{-3}) between populations, whereas microsatellite multilocus heterozygosity showed a 1.4-fold difference (0.54–0.78). They concluded that variability at microsatellite markers does not necessarily reflect the underlying genomic diversity in the wild. Similar conclusions have been reached previously (e.g. Hedrick 2001; Pemberton 2004; Slate *et al.* 2004) and we share the view that multilocus heterozygosity could be a poor predictor of genetic diversity in many population scenarios (Hansson & Westerberg 2002).

However, the reasons for why microsatellite data are often poor predictors of genetic variation in natural populations are not fully explained in Väli *et al.* (2008). Several important explanations are thoroughly discussed, e.g. the fundamental difference in the underlying mutation processes of repetitive and non-repetitive DNA (Hedrick 2001; Ellegren 2004) and the ascertainment bias introduced by selecting the most polymorphic microsatellite loci (Pardi *et al.* 2005; Brandström & Ellegren 2008), while other explanations are not vetted. Most importantly, what is not clearly pointed out is that an initial requirement for any correlation between multilocus heterozygosity and genome-wide genetic variability is the existence of variation in the degree of genome-wide diversity within the population (or between populations if more than one population is sampled). In other words, a requirement for genome-wide heterozygosity–diversity correlations is that the population shows 'identity disequilibrium', i.e. non-random associations of diploid genotypes between loci (Weir & Cocker-

Correspondence: Marcus Ljungqvist, Fax: 46462224719;
E-mail: marcus.ljungqvist@zooekol.lu.se

ham 1969; Chakraborty 1981; Lynch & Walsh 1998; Hansson & Westerberg 2002). Identity disequilibrium may arise for several reasons, for example, due to partial inbreeding (matings among kin or self-fertilization), admixture and genetic drift.

Väli *et al.* (2008) suggested that single nucleotide polymorphisms (SNPs) and resequencing approaches are superior to microsatellites (with their unique mutation patterns and potential ascertainment bias) in predicting genetic diversity in the context of conservation genetics. In this commentary, we would like to clarify that this is not necessarily true, since correlations between multilocus heterozygosity and genome-wide variability are not expected at identity equilibrium, regardless of marker choice. Moreover, in the presence of identity disequilibrium, previous studies have shown that heterozygosity at more variable markers, such as microsatellites, will provide a much stronger prediction of genome-wide genetic diversity, than heterozygosity at markers with little variation, such as SNPs (Slate *et al.* 2004; see also Csilléry *et al.* 2006); and we would like to highlight that conservation genetic studies using SNPs will require a substantial effort in terms of number of genotyped loci.

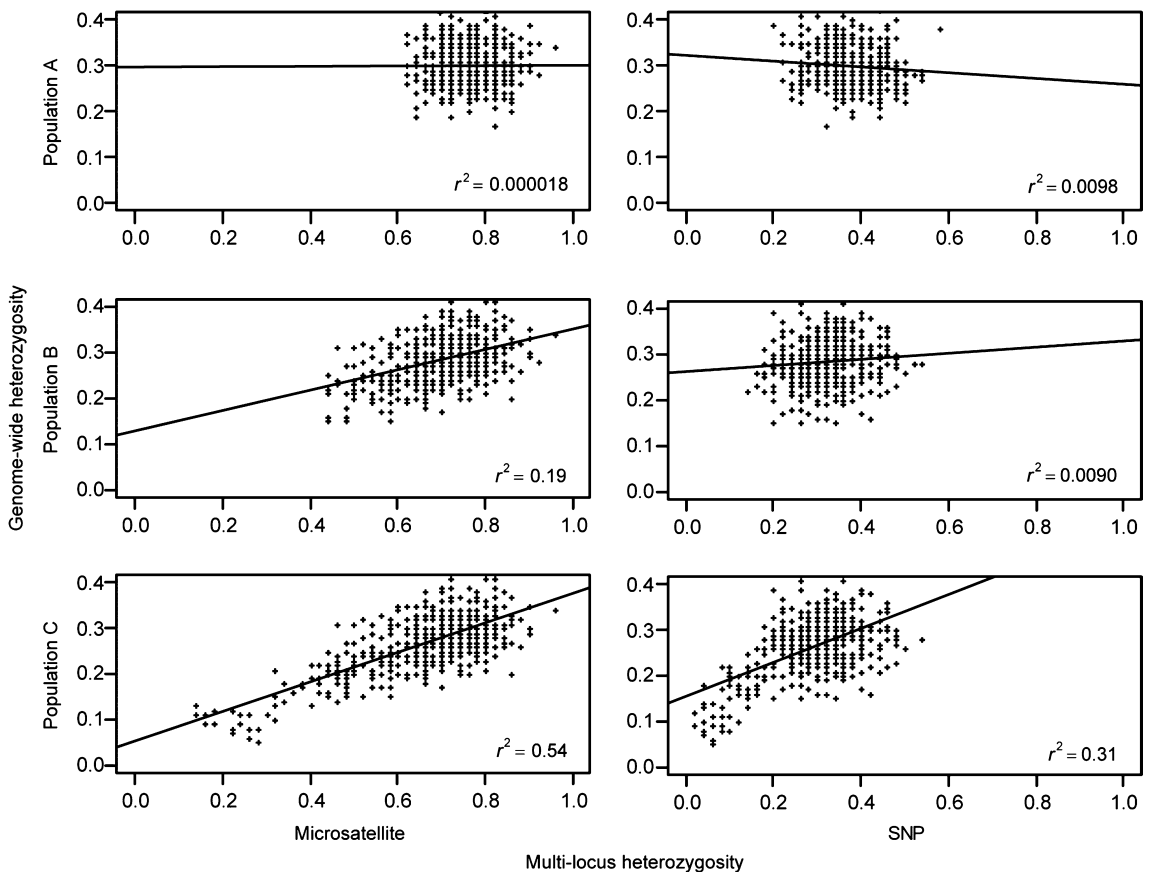
Results and discussion

To clarify and to stress the importance of identity disequilibrium and marker variability for correlations between multilocus heterozygosity (MLH) and one measure of genetic variability, genome-wide heterozygosity (GWH), we performed a simple simulation where heterozygosity at two types of markers, corresponding to microsatellites and SNPs, respectively, were tested for correlation with GWH in populations with different levels of identity disequilibrium (Box 1). By simulating different levels of inbreeding we generated three populations (A–C) that differed substantially in the degree of identity disequilibrium and GWH. At one extreme, a largely outbred population [population A; average (range) inbreeding coefficient, $f = 0.012$ (0–0.13)] and, at the other extreme, populations harbouring both highly inbred (genome-wide homozygous) and outbred (genome-wide heterozygous) individuals [population B: $f = 0.095$ (0–0.32); population C: $f = 0.16$ (0–0.74); Box 1]. Thus, our three simulated populations differed in the degree of variation in the inbreeding coefficient [$\sigma^2(f) = 7.0 \times 10^{-5}$, 6.2×10^{-3}, 2.7×10^{-2}, respectively] and genome-wide heterozygosity [$\sigma^2(\mathrm{GWH}) = 1.9 \times 10^{-3}$, 2.3×10^{-3} and 4.1×10^{-3}, respectively]; consequently, they differed in the degree of identity disequilibrium.

The importance of identity disequilibrium for generating correlations between MLH and GWH was apparent for both microsatellites and SNPs (Box 1). Population A with weak identity disequilibrium, exhibited no correlation between MLH and GWH ($r^2 < 0.01$; Box 1), whereas the populations with stronger identity disequilibrium showed moderate to strong correlations between MLH and GWH (population B: $r^2 = 0.01$–0.19; population C: $r^2 = 0.31$–0.54; Box 1). Consequently, only when there is substantial identity disequilibrium, MLH at different sets of loci will correlate strongly to genetic variability. In line with this reasoning, Väli *et al.* (2008) found a clear correlation between microsatellite MLH and nucleotide diversity ($r^2 = 0.70$) when all eight mammalian populations were analysed simultaneously (see fig. 1a, b in Väli *et al.* 2008). These populations show striking differences in their population history (Väli *et al.* 2008), which lead to substantial identity disequilibrium in the merged sample. Strong correlations have also been found in other between populations studies, e.g. in Atlantic salmon (*Salmo salar*) with $r^2 = 0.42$ for microsatellite and SNP heterozygosity (Ryynänen *et al.* 2007). Furthermore, in Väli *et al.*'s study, the strongest within-population correlation between microsatellite and SNP heterozygosity was noticed in the Scandinavian wolf population (see fig. 2a in Väli *et al.* 2008). This wolf population is characterized by frequent incestuous inbreeding and occasional outbreeding due to immigration (Vila *et al.* 2003; Bensch *et al.* 2006) and as a consequence pronounced partial inbreeding and identity disequilibrium (Liberg *et al.* 2005; Bensch *et al.* 2006). We do not know whether the other populations (showing no correlation between microsatellite and SNP heterozygosity) in Väli *et al.* (2008) are close to identity equilibrium, but if they are, no or very weak correlations between marker heterozygosity and genome-wide variability are expected (see population A; Box 1).

There was a pronounced impact of marker variability in our simulated data sets: the microsatellite-based MLH provided a much stronger prediction of GWH than did SNP-based MLH (Box 1). For example, in population C, the population with the highest variation in inbreeding, the proportion of variance in GWH explained by microsatellite MLH was 0.54, whereas the corresponding value for the SNPs was 0.31 (Box 1). Similar results have been found in previous simulation studies (Slate *et al.* 2004). Therefore, to study one aspect of the genetic variation, namely the genome-wide variation, these results would suggest the use of highly variable markers like microsatellites, unless a substantially higher number of SNPs can be screened to compensate for their low variability. Following Slate *et al.* (2004; equation 4), we estimated that approximately five times more SNPs (with 38% heterozygosity) than microsatellites (with 78% heterozygosity) would be required to achieve similar correlations between MLH and f, where f is used as a proxy of GWH, even in the most genetically variable of our populations (Box 2). The high number of SNPs necessary to achieve sufficient power certainly needs to be taking into account when designing conservation genetic studies. However, recent molecular techniques and bioinformatics tools make SNPs increasingly accessible and screenable compared to just a few years ago (e.g. Ryynänen *et al.* 2007; Kenta *et al.* 2008; Stapley *et al.* 2008; Slate *et al.* 2009), which may compensate for the need of genotyping larger numbers of SNPs.

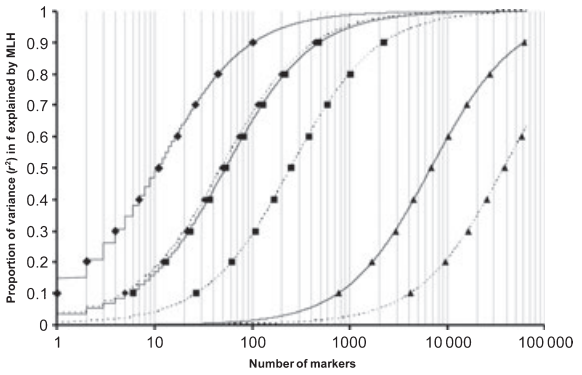
Finally, we would like to point out that other aspects of the genetic variation than the genome-wide variability may be preferentially studied at the sequence level rather than



Box 1. The correlation between MLH and GWH in populations with different degree of variation in GWH using microsatellites and SNPs

Individual-based population simulations were conducted in MatLab 7 (MathWorks, version 7.8.0.347, R2009a). Individuals in the founder population were created by randomly drawing alleles from specified allele frequency distributions. Populations were run for 50 generations and different population scenarios were implemented to create populations with the same number of individuals (500), but with different variance in the inbreeding coefficient (f) and GWH [population A: $\sigma^2(f) = 7.0 \times 10^{-5}$, $\sigma^2(\text{GWH}) = 1.8 \times 10^{-3}$; population B: $\sigma^2(f) = 6.2 \times 10^{-3}$, $\sigma^2(\text{GWH}) = 2.3 \times 10^{-3}$; population C: $\sigma^2(f) = 2.7 \times 10^{-2}$, $\sigma^2(\text{GWH}) = 4.0 \times 10^{-3}$]. A total of 200 independently segregating loci were modelled, of which 100 bi-allelic SNPs represented genome-wide loci and 50 bi-allelic SNPs and 50 penta-allelic microsatellites represented markers. In outbred individuals ($f = 0$), the genome-wide SNPs were selected to have an expected mean heterozygosity of 32%, the marker SNP a mean of 38% and the microsatellite markers a mean of 78%. In the end of the simulation, each individual was scored for GWH and SNP- and microsatellite-heterozygosity. From pedigree information of each population we calculated each individual's f. The main results are that the correlation between MLH and GWH increases with increasing variation in f and GWH and with the variability of the markers. These patterns remained when different numbers of loci were used to calculate GWH (50 and 500, respectively; results not shown).

with highly variable markers, such as microsatellites. It has been suggested that the genetic variation at specific, critical genes, such as the major histocompatibility complex, should be the focus in conservation genetic studies (Hughes 1991; Miller & Lambert 2004). Resequencing approaches of the target genes *per se*, using techniques described by Väli *et al.* (2008), would be preferable as they will provide direct data of the amount of genetic variation for these potentially important genes and it would also make it possible to evaluate past and ongoing selection at the molecular level by using various sequence-based genomic approaches (Nielsen 2005; Nielsen *et al.* 2005). The amount of genetic variation at specific genes can also be studied through correlated patterns at linked markers, caused by hitch-hiking processes. This is done with outlier approaches, where a large number of randomly selected



Box 2. The association between the proportion of variance in the inbreeding coefficient (*f*) explained by mean MLH and the number of markers used to calculate MLH for SNPs and microsatellites in three different population scenarios

We used equation 4 in Slate *et al.* (2004) to estimate the proportion of variance in *f* explained by (i) SNPs markers with a heterozygosity level of 38% (dashed lines); and (ii) microsatellite markers with 78% heterozygosity (solid lines), respectively, for the three populations [A (triangles), B (squares) and C (diamonds)] modelled in Box 1]. The main result is that the proportion of variance in *f* explained by MLH increases with increasing number of markers, increasing marker heterozygosity and increasing genetic variation in the population. To achieve, for instance, an r^2-value of 0.5 between *f* and MLH in these populations, approximately 4.4–5.6 times more SNPs than microsatellites would be required.

markers are typed and then evaluated for deviances from neutral expectations to find spots of importance for selection and adaptation (Hemmer-Hansen *et al.* 2007; Wood *et al.* 2008). In addition, in genetic model organisms for which genetic maps are available it is also possible to study the degree of genetic variation and other signatures of selection on a local chromosomal scale by genotyping linked markers; e.g. selecting markers that are tightly linked to genes of particular interest (Sutter *et al.* 2007). It is likely that SNPs will in most circumstances provide a better prediction of the amount of genetic variation at tightly linked genes than will microsatellites, because the high mutation rates of microsatellites may lead to allelic heterogeneities and decay of linkage disequilibrium to critical loci (Hästbacka *et al.* 1992; Kruglyak 1999; Slatkin 2008). An exception could be when selection has been very recent. Then, variable markers, such as microsatellites, may provide superior signal of linked genetic diversity compared to SNPs.

Rapid advancements in molecular biology and bioinformatics during the last years make it possible to carry out large-scale genotyping of markers and functional genes in large population data sets (Ellegren & Sheldon 2008). This is good news since the structure and demographic history of populations can be studied with increasing precision, potentially by using a combination of microsatellite, SNP and other types of genetic data, which may provide crucial knowledge to increase our understanding of the role and importance of genetic variation in conservation biology.

References

Beaumont MA, Bruford MW (1999) Microsatellites in conservation genetics. In:*Microsatellites: Evolution and Application* (eds Goldstein DB, Schlötterer C). pp. 165–182, Oxford University Press, Oxford.

Bensch S, Andrén H, Hansson B *et al.* (2006) Selection for heterozygosity gives hope to a wild population of inbred wolves. *PLoS ONE*, **1**, e72.

Brandström M, Ellegren H (2008) Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. *Genome Research*, **18**, 881–887.

Chakraborty R (1981) The distribution of the number of heterozygous loci in an individual in natural populations. *Genetics*, **98**, 461–466.

Coltman DW, Slate J (2003) Microsatellite measures of inbreeding: a meta-analysis. *Evolution*, **57**, 971–983.

Csilléry K, Johnson T, Beraldi D *et al.* (2006) Performance of marker-based relatedness estimators in natural populations of outbred vertebrates. *Genetics*, **173**, 2091–2101.

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, **5**, 435–445.

Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural populations. *Nature*, **452**, 169–175.

Frankham R (1995) Conservation genetics. *Annual Review of Genetics*, **29**, 305–327.

Frankham R, Ballou JD, Briscoe DA (2002) *Introduction to Conservation Genetics*. Cambridge University Press, Cambridge.

Hansson B, Westerberg L (2002) On the correlation between heterozygosity and fitness in natural populations. *Molecular Ecology*, **11**, 2467–2474.

Hästbacka J, de la Chapelle A, Kaitila I *et al.* (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics*, **2**, 204–211.

Hedrick PW (2001) Conservation genetics: where are we now? *Trends in Ecology and Evolution*, **16**, 629–636.

Hemmer-Hansen J, Nielsen EE, Gronkjaer P *et al.* (2007) Evolutionary mechanisms shaping the genetic population structure of marine fishes; lessons from the European flounder (*Platichthys flesus* L.). *Molecular Ecology*, **15**, 3104–3118.

Hughes AL (1991) MHC polymorphism and the design of captive breeding programs. *Conservation Biology*, **5**, 249–251.

Kenta T, Gratten J, Haigh NS *et al.* (2008) Multiplex SNP-SCALE: a cost-effective medium-throughput single nucleotide polymorphism genotyping method. *Molecular Ecology Resources*, **8**, 1230–1238.

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics*, **22**, 139–144.

Liberg O, Andren H, Pedersen HC *et al.* (2005) Severe inbreeding depression in a wild wolf (*Canis lupus*) population. *Biological Letters*, **1**, 17–20.

Lynch M, Walsh B (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates. Inc., Sunderland.

Miller HC, Lambert DM (2004) Genetic drift outweighs balancing selection in shaping post-bottleneck major histocompatibility complex variation in New Zealand robins (Petroicidae). *Molecular Ecology*, **13**, 3709–3721.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Nielsen R, Williamson S, Kim Y *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Research*, **15**, 1566–1575.

Pardi F, Sibly RM, Wilkinson MJ, Whittaker JC (2005) On the structural differences between markers and genomic AC microsatellites. *Journal of Molecular Evolution*, **60**, 688–693.

Pemberton J (2004) Measuring inbreeding depression in the wild: the old ways are the best. *Trends in Ecology and Evolution*, **19**, 613–615.

Reed DH, Frankham R (2001) How closely correlated are molecular and quantitative measures of genetic variation? *A meta-analysis. Evolution*, **55**, 1095–1103.

Reed DH, Frankham R (2003) Correlation between population fitness and genetic diversity. *Conservation Biology*, **17**, 230–237.

Ryynänen HJ, Tonteri A, Vasemägi A *et al.* (2007) A comparison of biallelic markers and microsatellites for the estimation of population and conservation genetic parameters in Atlantic salmon (*Salmo salar*). *Journal of Heredity*, **98**, 692–704.

Slate J, David P, Dodds KG *et al.* (2004) Understanding the relationship between the inbreeding coefficient and multilocus heterozygosity: theoretical expectations and empirical data. *Heredity*, **93**, 255–265.

Slate J, Gratten J, Beraldi D *et al.* (2009) Gene mapping in the wild with SNPs: guidelines and future directions. *Genetica*, **136**, 97–107.

Slatkin M (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, **9**, 477–485.

Stapley J, Birkhead TR, Burke T, Slate J (2008) A linkage map of the zebra finch Taeniopygia guttata provides new insights into avian genome evolution. *Genetics*, **179**, 651–667.

Sutter NB, Bustamante CD, Chase K *et al.* (2007) A single IGF1 allele is a major determinant of small size in dogs. *Science*, **316**, 112–115.

Väli U, Einarsson A, Waits L, Ellegren H (2008) To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Molecular Ecology*, **17**, 3808–3817.

Vila C, Sundqvist AK, Flagstad O *et al.* (2003) Rescue of a severely bottlenecked wolf (*Canis lupus*) population by a single immigrant. *Proceedings in Biological Science*, **270**, 91–97.

Weir BS, Cockerham CC (1969) Group inbreeding with two linked loci. *Genetics*, **63**, 711–742.

Wood HM, Grahame JW, Humphray S *et al.* (2008) Sequence differentiation in regions identified by a genome scan for local adaptation. *Molecular Ecology*, **17**, 3123–3135.

doi: 10.1111/j.1365-294X.2010.04522.x