MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11**, 591–611 doi: 10.1111/j.1755-0998.2011.03014.x

Current trends in microsatellite genotyping

E. GUICHOUX,*†‡ L. LAGACHE,*† S. WAGNER,*†§ P. CHAUMEIL,*† P. LÉGER,*† O. LEPAIS,*†¶ C. LEPOITTEVIN,*† T. MALAUSA,** E. REVARDEL,*† F. SALIN*† and R.J. PETIT*†
*INRA, UMR 1202 Biodiversity Genes & Communities, F-33610 Cestas, France, †Univ. Bordeaux, UMR1202 Biodiversity Genes & Communities, Bordeaux, F-33400 Talence, France, ‡Pernod Ricard Research Center, F-94000 Creteil, France, §Univ. Bonn, Steinmann Institut, D-53115 Bonn, Germany, ¶School of Biological and Environmental Sciences, University of Stirling, Stirling FK9 4LA, UK, **INRA, UMR 1301 IBSV INRA/UNSA/CNRS, F-06903 Sophia-Antipolis, France

Abstract

Microsatellites have been popular molecular markers ever since their advent in the late eighties. Despite growing competition from new genotyping and sequencing techniques, the use of these versatile and cost-effective markers continues to increase, boosted by successive technical advances. First, methods for multiplexing PCR have considerably improved over the last years, thereby decreasing genotyping costs and increasing throughput. Second, next-generation sequencing technologies allow the identification of large numbers of microsatellite loci at reduced cost in non-model species. As a consequence, more stringent selection of loci is possible, thereby further enhancing multiplex quality and efficiency. However, current practices are lagging behind. By surveying recently published population genetic studies relying on simple sequence repeats, we show that more than half of the studies lack appropriate quality controls and do not make use of multiplex PCR. To make the most of the latest technical developments, we outline the need for a well-established strategy including standardized high-throughput bench protocols and specific bioinformatic tools, from primer design to allele calling.

Keywords: binning, high throughput, multiplexing, nextgen sequencing, quality control, SSR

Received 5 October 2010; revision received 24 February 2011; accepted 7 March 2011

Introduction

At a time where radically new genome-wide approaches emerge to study genetic variation, it is important to recall that many questions in molecular ecology can be efficiently addressed with a limited number of highly polymorphic markers, such as microsatellites. Microsatellites, also known as simple sequence repeats (SSRs) or short tandem repeats (STRs), remain the most popular markers in population genetic studies (Fig. 1). They consist of motifs of one to six nucleotides repeated several times that have a characteristic mutational behaviour (Kelkar *et al.* 2010). As a consequence of their elevated mutation rates, SSRs are typically highly polymorphic: Different individuals exhibit variation manifested as repeat number differences. Microsatellites have been used increasingly since the late eighties for applications such as fingerprinting, parentage analyses, genetic mapping or genetic structure analyses (Ellegren 2004; Mittal & Dubey 2009; Jones *et al.* 2010). Their genomic distribution, evolutionary dynamics, biological function and practical utility have been the object of a very large body of research, as summarized in several review articles (Tautz & Schlötterer 1994; Jarne & Lagoda 1996; Schlötterer 1998; Chambers & MacAvoy 2000; Li *et al.* 2002; Dieringer & Schlötterer 2003; Ellegren 2004; Buschiazzo & Gemmell 2006; Chistiakov *et al.* 2006; Oliveira *et al.* 2006; Selkoe & Toonen 2006; Subirana & Messeguer 2008; Sun *et al.* 2009). Their advantages over single-nucleotide polymorphisms (SNPs), which tend to be used increasingly, include high allelic diversity and relative ease of transfer between closely related species (Box 1). However, SSRs have some drawbacks: a lengthy and costly development phase and a relatively low throughput because of difficulties for automation and data management, especially when compared to SNPs (Box 1). Hence, the continued use of microsatellites will probably depend on the possibility to overcome some of these limitations.

Recently, progresses in SSR development and genotyping have been made in several directions, suggesting that SSRs could remain relevant genetic markers,

Correspondence: Rémy J. Petit, Fax: 33 5 57 12 28 81; E-mail: petit@pierroton.inra.fr

at least for specific applications. First, the emergence of next-generation sequencing technologies means that identifying SSRs has become cheaper and faster. This trend is very recent, with the first reports appearing only in 2009 (Abdelkrim *et al.* 2009; Rasmussen & Noor 2009; Santana *et al.* 2009). Second, multiplexing microsatellites has become much easier. It can be accomplished through the co-amplification of multiple microsatellites in a single PCR cocktail, a procedure called true multiplexing. Alternatively, PCR products from multiple amplification reactions can be combined in a single lane, a procedure referred to as pseudo-multiplexing or poolplexing (Ghislain *et al.* 2004; Meudt & Clarke 2007). A blend of the two approaches is also possible. In true multiplex PCR (henceforth simply called multiplex), more than one target sequence are amplified by including more than one pair of primers in the reaction. The first successful attempt to multiplex PCR took place more than 20 years ago (Chamberlain *et al.* 1988). Since then, capillary electrophoresis equipments relying on automated laser-induced fluorescence DNA technology have facilitated the use of this technique (Butler *et al.* 2001, 2004). Loci with non-overlapping allele size ranges are labelled with the same fluorescent dye, whereas those with overlapping allele size ranges are labelled with different dyes and resolved individually because of the different characteristic emission spectrum of each dye, hence considerably expanding multiplexing potential. In addition, one of the dyes is used as an in-lane size standard, greatly improving the sizing precision of alleles. Multiplex PCR now forms the basis for many studies, on both diploid and polyploid species (Jewell *et al.* 2010; Raabova *et al.* 2010), reducing very significantly the cost and time of genetic analyses (Box 2). Important progresses have also been made in SSR data scoring, a critical and time-limiting step.

In this study, we survey a sample of the recent literature on SSR genotyping. We show that multiplexing many (≥8) SSRs is not yet commonplace, despite the potential for much higher levels of multiplexing (e.g. Hill *et al.* 2009). We continue by outlining the key steps necessary to develop accurate SSR multiplex. This involves paying attention to the whole process, from microsatellite identification to primer selection, data scoring and associated bioinformatics. We consider genotyping accuracy and troubleshooting and discuss areas where technical improvements of SSR genotyping are already possible and other areas where new developments would be important. We rely on our recent efforts to develop SSR multiplexes in forest trees for parentage analyses and population genetic surveys, during which we have reconsidered most steps to obtain high-quality data sets (Guichoux *et al.* 2011). Although several review articles on multiplex development already exist (Edwards & Gibbs 1994; Henegariu *et al.* 1997; Elnifro *et al.* 2000; Markoulatos *et al.* 2002; Wallin *et al.* 2002; Butler 2005a; Cryer *et al.* 2005), none of these papers has provided a complete overview of SSR identification, multiplex design and genotyping. In addition, the latest developments based on next-generation sequencing techniques postdate these studies. Here, we first review current practices in SSR genotyping studies and then consider the entire process of SSR genotyping, which ranges from SSR selection to data scoring and managing, while paying special attention to methods that help improve throughput and workflow, such as multiplexing.

A review of current practices

We surveyed a subset of the recent literature to examine current practices in terms of SSR genotyping. We checked 100 original journal articles relying on SSRs that had been published recently (in 2009–2010, see Data S1, Supporting Information) in the journal *Molecular Ecology*, along with associated primer notes, if needed. Among the 100 original studies, 69 deal with population structure and 31 with parentage or sibship analyses. The organisms studied were all diploid and involved vertebrates, invertebrates, fungi and plants (Table 2). On average, 564 individuals were surveyed at 11.6 nuclear SSR loci, with no major bias depending on the organism investigated. Most studies took advantage of an automatic capillary electrophoresis system (90%). Overall, less than half of the studies (42%) used true multiplexing. This result illustrates the still limited penetration of multiplexing technique in the field, despite the nearly universal availability of suitable equipment. Unfortunately, the frequency of pseudo-multiplexing could not be calculated as its use appears not to be systematically reported. The mean number of SSRs surveyed was 11.1 in studies without multiplexing and 12.3 in studies with multiplexing with an average of 3.9 loci (2-12) per multiplex. For those studies that used a specialized multiplex PCR buffer (e.g. Qiagen PCR Multiplex kit), the corresponding figures are 13.9 SSRs with 5.0 loci per multiplex. Therefore, researchers using multiplexing techniques tend to use more loci, either to address different questions requiring more markers or to produce higher-quality data sets for similar applications. Even higher levels of multiplexing are possible in the context of studies of non-model species, as 11 studies among the 100 surveyed relied on ≥8-plex. In fact, a few recent SSR studies have relied on very large (>20) multiplexes (e.g. Hill *et al.* 2009; Chen *et al.* 2010), whereas simultaneous PCR amplification of 35–40 PCR products is routinely achieved in the case of SNPs (e.g. Gabriel *et al.* 2009; Buggs *et al.* 2010), demonstrating that problems of primer competition can be overcome. The poor penetration of

Box 1 SSRs vs. SNPs

To evaluate current trends in genotyping methods, we searched the ISI Web of Knowledge database for papers citing SSRs or SNPs. The former have increased linearly since the early 1990s, whereas the latter have increased exponentially since the late 1990s (Fig. 1). Yet, papers citing SSRs still outnumbered those citing SNPs in 2009. Although this should change soon, the continued increase in studies relying on SSRs justifies efforts to improve their effectiveness.

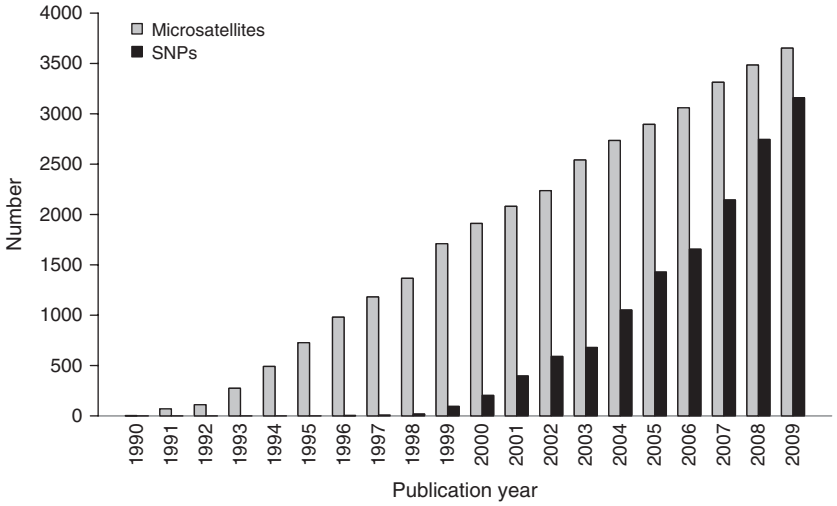

Fig. 1 Evolution of the number of studies relying on SSRs and SNPs since 1990.

Current popularity is not always the best guide to decide which markers to use (Schlötterer 2004). Instead, information on the relative advantages of each type of marker for various applications should help researchers embarking on new projects in molecular ecology. Following Morin *et al.* (2004), we provide here a brief summary of the relative merits of SSRs and SNPs, focusing successively on the intrinsic differences between the two markers and then on the technical aspects of their analysis.

There are two main differences between SSRs and SNPs. First, SNPs are more numerous than SSRs in the genome of most species. On average, in the human genome, there is one SNP every 100–300 bp (Thorisson *et al.* 2005), compared to one SSR locus every 2–30 kb (Webster *et al.* 2002), depending on how SSRs are defined (Kelkar *et al.* 2010). This can be important for genome-wide association studies but not necessarily for other applications. Second, the mutation rate per generation differs drastically between the two marker types. SSRs have mutation rates ranging from 10^{-3} to 10^{-4} per locus per generation (Ellegren 2000), compared to about 10^{-9} for SNPs, i.e. several orders of magnitude lower. As a consequence, SNPs are typically diallelic: In humans, <0.1% of SNPs are triallelic (Lai 2001). In contrast, SSR loci generally have high allelic richness, often in excess of 10 alleles. Below, we list the relative merits of SSRs and of SNPs to help researchers decide which type of markers is best suited for their needs.

(a) Advantages of SSRs over SNPs

- SSR loci above a certain number of repeats can be assumed to be polymorphic (Schlötterer 2004) whereas to identify SNPs, homologous regions must be sequenced from multiple chromosomes.
- SSRs have little ascertainment bias (the bias resulting from the choice of the initial panel of genotypes used to screen for polymorphisms) in contrast to SNPs (e.g. Li *et al.* 2008).
- The success rate of cross-amplification for SSRs in closely related species is typically higher than for SNPs (up to 50%, Sharma *et al.* 2007).
- SSR loci are more powerful than SNPs to detect mixtures (Clayton *et al.* 1998; Gill 2001).
- SSR accuracy is easy to assess because a larger proportion of errors can be detected in pedigree analyses when there are many alleles per locus; in contrast, for SNPs, which are typically diallelic, many errors will remain undetected when analysing pedigrees as they will be compatible with Mendelian segregation rules (Palsson *et al.* 1999).

• SSRs will be more useful for detecting recent population expansions than SNPs, because the accumulation of new mutations, which is the hallmark of population expansion, requires shorter time periods for rapidly evolving loci than for slowly evolving ones (Morin *et al.* 2004).

• For many applications, there is not much gain in using more loci after a certain threshold is reached. For instance, low error rates can be achieved in clonal identification using a few highly polymorphic loci. Moreover, using more than a few tens of loci might not be relevant as additional loci become non-independent because of linkage (Santure *et al.* 2010). In such cases, microsatellites represent a credible alternative. To help researchers decide on the best alternative, we provide indications from the literature on the number of SNPs needed to result in a power equivalent to that of one SSR for different applications (Table 1). The information originates mostly from simulation studies aiming at evaluating the relative power of different markers differing in allelic richness.

Table 1 Number of SNPs needed to result in a power equivalent to that of one SSR depending on the application

Application	Relative power of SSRs vs. SNPs	Comments	References
Linkage study, individual identification	2–3	Power proportional to heterozygosity H: $H_{SSR} \sim 2.H_{SNP}$	Kruglyak (1997), Waits *et al.* (2001), Seddon *et al.* (2005)
Parentage analysis	~5	This estimate was obtained using SNPs with minor allele frequency >0.2. Note also that with diallelic SNPs, a heterozygous genotype is a universal donor.	Glaubitz *et al.* (2003)
Genetic structure	4–12	SNPs have typically few private alleles as a consequence of the way they are identified, i.e. using a limited panel of genotypes; such private alleles are particularly useful to reconstruct genetic structure.	Rosenberg *et al.* (2003), Liu *et al.* (2005)
Association studies/Linkage disequilibrium	5–20	Expected power of genome-wide LD testing for the detection of a low-frequency disease variant, assuming SNPs have minor allele frequencies >0.2.	Ohashi & Tokunaga (2003)
Sibling reconstruction	∞	The 4-allele property states that no more than four alleles can be found in a full-sib family; this property cannot be used to reconstruct sibships with diallelic SNPs.	Berger-Wolf *et al.* (2007), Ashley *et al.* (2009), Wang & Santure (2009), Jones & Wang (2010)

(b) Drawbacks of SSRs over SNPs

• The large number of alleles per locus in SSRs implies that for accurate estimation of allelic frequencies, large sample sizes are needed, in contrast to SNPs.

• Spontaneous mutations are more likely to take place at SSRs than at SNPs within a given pedigree, potentially complicating parentage reconstruction when using SSRs (Ellegren 2000; Phillips *et al.* 2007; Borsting *et al.* 2009).

• The high rate of recurrent or backward mutation of SSRs makes them poor indicators of long-term population history (Li *et al.* 2002; Ellegren 2004; Morin *et al.* 2004; Schlötterer 2004).

• Variability at highly polymorphic microsatellite markers might not accurately reflect the underlying genomic diversity (Väli *et al.* 2008 but see Ljungqvist *et al.* 2010).

• Capillary gel electrophoresis coupled with fluorescence-based detection is the only commonly reported method for the assay of SSRs (Butler *et al.* 2001; Koumi *et al.* 2004). In contrast, SNPs are potentially amenable to typing through many techniques, including digital typing methods using chip technology, allowing the development of ultra-high-density methods (Syvänen 2005; McCarroll *et al.* 2008).

• With SSRs, there is a need to include common controls among studies and across time. In contrast, SNP studies can be replicated, performed in parallel across several laboratories and added to as samples become available without the need to calibrate results at each step in the process. To date, reduced portability of SSR data across laboratories has resulted in significant data use limitations (e.g. Hoffman *et al.* 2006).

• PCR amplicons are typically longer for SSRs than for SNPs, making it more difficult to study highly degraded DNA samples, such as faecal and other non-invasive samples, with SSRs than with SNPs (Seddon *et al.* 2005; Morin & McCarthy 2007; Sanchez & Endicott 2006).

In conclusion, the widespread adoption of SSRs lies in the power that they provide to solve biological problems, due in particular to their high allelic richness. In contrast, many disadvantages of SSRs are of a technical nature (Chambers & MacAvoy 2000). This suggests that SSRs could remain useful in the future if at least some of the technical problems identified are overcome (Glaubitz *et al.* 2003; Schlötterer 2004; Ryynänen *et al.* 2007; Matschiner & Salzburger 2009). In principle, using blocks of tightly linked SNPs and treating each haplotype as a separate allele could yield genotyping data with properties similar to those obtained with SSR loci (Jones *et al.* 2009). However, the incidence of missing data will probably be high, whereas compound genotyping errors will quickly increase as multiple PCRs are needed to type a single locus.

multiplexing, despite considerable potential, might be caused by the persistent belief that multiplexing greatly increases complexity or costs of microsatellite development (e.g. Neff *et al.* 2000), which dates from the early times of PCR multiplexing (Edwards & Gibbs 1994). Further results regarding the types of SSRs studied and the quality controls used (estimation of the frequency of null alleles and of error rates) are discussed below. In general, our survey illustrates the need for more standardized reporting of microsatellite studies. This would help monitor the developments in the field and better evaluate the quality of the data sets produced.

SSR selection

Source of sequence data

Microsatellite detection requires sequence data. Until recently, the only possibility to identify sequences harbouring SSR motifs was the screening of size-fractionated genomic DNA or of EST (expressed sequence tag) libraries (Zane *et al.* 2002). EST-SSRs are often reported to be less variable than genomic SSRs, being found in selectively more constrained regions of the genome (Gupta *et al.* 2003). They also have the disadvantage that amplicon sizes can differ from expectation, as a consequence of the undetected presence of introns in flanking regions (Varshney *et al.* 2005). However, this is balanced by several important advantages over genomic SSRs: (i) They should detect variation in the expressed portion of the genome, which might be of interest for studies of marker–trait associations; (ii) They can be developed at no cost from EST databases; and (iii) Once developed, these markers, unlike genomic SSRs, may work across a number of related species, because primers designed in flanking coding sequences are more likely to be conserved across species, resulting in high levels of transferability (Gupta *et al.* 2003; Pashley *et al.* 2006), especially if efforts are made to target conserved regions by using multiple alignments to design primers (Dawson *et al.* 2010).

Regardless of whether genomic or EST sequences are used for SSR detection, traditional laboratory methods involving cloning, cDNA library construction and Sanger sequencing remain costly and time-consuming (Squirrell *et al.* 2003; Pashley *et al.* 2006; Parchman *et al.* 2010). To remediate this, next-generation sequencing techniques have now started to be used to identify sequences harbouring SSR motifs in non-model species (Allentoft *et al.* 2009). The first successful attempts have allowed a two to five times cost reduction as well as a significant decrease in time expenditure compared to traditional microsatellite development (Abdelkrim *et al.* 2009; Santana *et al.* 2009; Castoe *et al.* 2010; Csencsics *et al.* 2010; Malausa *et al.* 2011). Methodological improvements, such as biotin-based enrichment in SSR motifs, are now being proposed in combination with next-generation sequencing, which should further boost these approaches (Malausa *et al.* 2011). Besides, these approaches generate millions of base pairs of genomic sequence that may be useful for both SSRs-related and SSRs-unrelated research.

Table 2 Characteristics of 100 original journal articles relying on SSRs published in the journal *Molecular Ecology* in 2009–2010. Values outlined in the text are in bold

Organisms studied (%)		Size of repeat units (%)	
Mammals	18	Di-nucleotides	**46**
Other invertebrates	16	Tri-nucleotides	13
Plants	15	Tetra-nucleotides	14
Arthropods	14	Imperfect	26
Amphibian and reptiles	12		
Birds	11	Null alleles check (%)	
Fungi	8		
Fish	6	Yes	**40**
		No	60
Multiplexing (%)			
1–4 markers	15	Error-rate measurement (%)	
5–8 markers	19	Yes	**26**
>8 markers	8	No	74
No	**58**		

Box 2 Cost-effectiveness of multiplex SSR typing

We have estimated the overall cost of SSR genotyping as a function of the degree of multiplexing, following Renshaw *et al.* (2006). The goal we set was the genotyping of up to 2500 samples at 24 microsatellites. Five strategies were considered: no multiplexing, 2-plex, 4-plex, 8-plex and 12-plex. Cost included consumables (plates, tips) and reagents (Qiagen Multiplex PCR kit, unlabelled primers, labelled primers, LIZ-600 size standard). Salary costs were based on those of an experienced research assistant in France. We conservatively assumed that in the absence of true multiplexing, pseudo-multiplexing was used by combining four loci marked with different fluorochromes in one lane.

The results (Fig. 2) show that even for a moderate number of samples (100), multiplexing is cost-effective (12-plex is eight times cheaper than simplex PCR). For completeness, this should be balanced with the cost of developing the multiplex. However, most of the work to develop and optimize SSR multiplex is actually represented by phases that are common to all SSR development projects. If primers have been selected with the objective of multiplexing in mind, the extra costs of multiplexing can amount to little more than 2–4 PCR tests for an 8-plex, depending on whether the concentration of some primers has to be optimized or some primers have to be replaced.

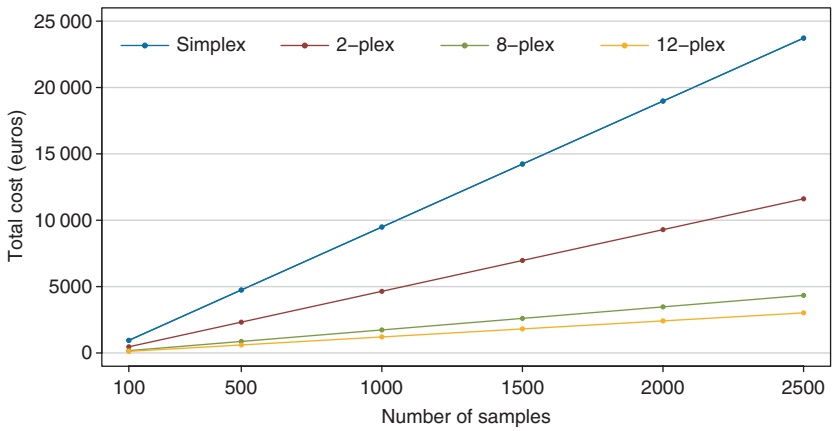

Fig. 2 Overall cost for genotyping 24 SSRs, depending on the multiplex strategy and the number of genotyped samples.

Other solutions to decrease costs:

The Qiagen Multiplex PCR Kit is the most widely cited commercial kit, with 25% of the papers we surveyed mentioning it. This commercial kit has a high cost per sample, but the final volume can be decreased to 5 μL (Lepais & Bacles 2010) with a final buffer concentration of 0.7× (Qiagen recommends 1×), without compromising reproducibility or specificity (Spathis & Lum 2008). This reduces the final cost to 0.13€ per sample (compared to 1.88€ with no optimization). Another solution to decrease the costs is to shift to 384 plates as these allow the use of even smaller volumes, down to 2 μL (Kenta *et al.* 2008). Finally, instead of relying on direct fluorescent labelling of primers, it is possible to use universal tailed primers, one for each fluorescence detection (Missiaggia & Grattapaglia 2006). Such a method allows the same level of marker multiplexing and accuracy in SSR genotyping attained in regular direct-labelled microsatellite fluorescent detection assays, while significantly reducing the costs. This procedure is particularly adapted when many SSRs need to be investigated on relatively few samples.

From transcriptome to whole genome shotgun sequencing for SSR detection. To optimize SSR detection with next-generation sequencing techniques, several strategies can be adopted, depending on the species' genome size, the abundance and nature of SSR motifs, and the sequencing coverage that can be achieved. For species harbouring large and complex genomes, such as conifers, direct approaches might be risky because of the large amount of repetitive sequences with no interest for SSR detection (Parchman *et al.* 2010). In this case, focusing on transcriptome—with the advantages and drawbacks previously discussed—can be more appropriate than whole genome shotgun sequencing. For genomes with a low frequency of SSRs, SSR enrichment techniques should be considered. Pyrosequencing of enriched libraries has proved efficient and cost-effective to isolate SSRs in non-model species (Santana *et al.* 2009; Malausa *et al.* 2011). Moreover, a test of this procedure on model species showed that distribution of the isolated markers across the genome satisfactorily reflects the actual distribution of

SSRs across the genome (Martin *et al.* 2010). If possible, informed choices about the motifs to target should be made, as this can greatly increase the number of useful SSR loci eventually identified (Santana *et al.* 2009; Dubut *et al.* 2010). To date, however, most studies (12 of the 15 articles relying on SSR detection with next-generation techniques that we identified, see Data S2, Supporting Information) have relied on whole genome shotgun sequencing, even when genome coverage was low (0.1× in Rasmussen & Noor 2009; 0.02× in Castoe *et al.* 2010) or when the genomes studied were known to have a low frequency of SSRs (Abdelkrim *et al.* 2009).

Read length. Interestingly, in all 15 studies published to date, the only sequencing technology used was the 454 pyrosequencing method of Roche. This technology generates the longest read length among the next-generation sequencing methods currently available. Hence, single reads can be used for SSR identification and primer design (Abbott *et al.* 2010). By circumventing the need for sequence assembly, this saves researchers from time-consuming bioinformatic steps. Software, such as MSAT-COMMANDER (Faircloth 2008) or QDD (Meglécz *et al.* 2010), has been created to identify SSRs from 454 sequence data, the first one being used in more than half of the studies. Despite this, read length remains a limiting factor: when the average read length is around 200 bp, up to two-thirds of the SSRs detected are too close to either fragment end to enable design of flanking PCR primers (Abdelkrim *et al.* 2009; Castoe *et al.* 2010; Csencsics *et al.* 2010; Lepais & Bacles 2010; Parchman *et al.* 2010). Such limitations should no longer be an issue because 454 technologies delivering >400 bp reads have now become available (Schuster 2008; Kircher & Kelso 2010). Such read lengths, in combination with the sequencing depth of the 454 technology, allow the design of a medium number of markers at sizes >300 bp (Malausa *et al.* 2011).

Advantages of next-generation sequencing. Hundreds or even thousands of SSR loci can be identified from a fraction of a single next-generation sequencing run (Tang *et al.* 2008; Boomer & Stow 2010; Castoe *et al.* 2010; Saarinen & Austin 2010). Moreover, if coverage is sufficient, shotgun data can be used to identify SSRs with unique primer sequences, which have a higher probability of producing successful locus-specific PCR amplification products (Castoe *et al.* 2010). Next-generation sequencing also provides preliminary information on SSR polymorphism, in particular if more than one genotype is sequenced. In our survey, only one study reported the use of more than one genotype at the sequencing stage, but available polymorphism data were not used to select candidate SSRs (Parchman *et al.* 2010). The low coverage attained in most of the studies probably precludes reliable detection of polymorphism. However, the throughput of sequencing technologies increases constantly, so we can expect higher genome coverage in the near future. Potentially, SSR polymorphism data should therefore become available very early on, which should in turn greatly facilitate SSR selection and optimization, at least if the necessary bioinformatic tools are accessible to the research team.

Choice of SSR type

Once sequence data harbouring candidate SSR loci have been obtained, a number of choices need to be made, as outlined below. Interestingly, the availability of large amounts of sequence data obtained from next-generation sequencing projects will allow stringent selection of the best markers, thereby greatly saving time in downstream optimizations.

Perfect or imperfect repeats. Microsatellites have been classified according to the type of repeat sequence as perfect (with simple repeats only) or imperfect (Urquhart *et al.* 1994). A common characteristic of imperfect repeats is that there is no more equivalency between fragment length and amplicon sequence: several sequences can correspond to a given length variant (e.g. Estoup *et al.* 1995). Choosing perfect motifs should ensure that microsatellite loci follow as much as possible the stepwise mutation model used in coalescent-based methods to infer demographic events (Estoup *et al.* 2001). Hence, preference should be given to perfect motifs (Gusmão *et al.* 2006). Yet, imperfect SSRs remain frequently used. In the 100 studies surveyed, 26% of the SSRs used were imperfect (Table 2).

Size of repeat unit. Microsatellite repeat units typically vary from one to six bases. Focusing on the shortest motifs (such as mono- or dinucleotide repeats) rather than on longer ones (≥trinucleotide repeats) should allow packing more loci on a given separation system, resulting in larger multiplexes. This can be important because sequencing machines used for SSR genotyping make use of no more than five fluorochromes, which severely limits the number of SSR loci that can be analysed simultaneously, given that allelic range size often reaches up to 50 or 100 bp and that amplicons measuring over 300 bp are rarely used (e.g. Hill *et al.* 2009; Chen *et al.* 2010). However, mononucleotide repeat SSRs can be difficult to accurately assay (Sun *et al.* 2006), so they are often eliminated at the outset (Kim *et al.* 2008). Among the 100 studies we surveyed, there was not a single case of mononucleotide repeat SSRs (Table 2) even if these markers have been used successfully in studies of chloroplast DNA variation in plants (Ebert & Peakall 2009), SSR-poor

fungi (Christians & Watt 2009) or in other circumstances where mononucleotide repeats are of special interest. In contrast, dinucleotide repeat SSRs were most frequently used. Unfortunately, dinucleotide repeats often show one or more 'stutter' bands (multiple PCR products from the same fragment that are typically shorter by one or a few repeats than the full-length product) (Chambers & MacAvoy 2000). This is attributed to enzyme slippage during amplification (slipped-strand mispairing), making allele designation difficult (Levinson & Gutman 1987; Meldgaard & Morling 1997), especially for heterozygotes with adjacent alleles. In contrast, tri-, tetra- or pentanucleotide repeats appear to be significantly less prone to slippage (Edwards *et al.* 1991). Hence, SSRs with core repeats three to five nucleotides long are sometimes preferred for forensic and parentage applications (Kirov *et al.* 2000; Cipriani *et al.* 2008). Note however that stutter bands, when not too strong, can be useful, by helping distinguish true alleles from artefacts (e.g. Schwengel *et al.* 1994). Note also that a few solutions have been proposed to overcome stuttering problems (Box 3).

Number of repeat units. The number of repeats has a critical effect on mutation behaviour to the point that it helps define which sequences actually represent microsatellites (Kelkar *et al.* 2010). As on average SSR loci with more repeats have higher mutation rates (Weber 1990; Ellegren 2000; Petit *et al.* 2005; Kelkar *et al.* 2008), selecting loci with sufficient number of repeats is necessary to ensure polymorphism. However, SSRs with numerous repeats have also some drawbacks, such as increased allele dropout (Kirov *et al.* 2000; Buchan *et al.* 2005) and increased stutter (Hoffman & Amos 2005). Moreover, SSRs with numerous repeats are characterized by large allelic range, so that fewer can be combined in a given multiplex. Hence, an intermediate number of repeats could represent a good compromise, by preserving most of the advantages of SSRs (multiallelic, high diversity) while avoiding some of their drawbacks caused by very high mutation rate (Box 1). For instance, van Asch *et al.* (2010) suggest to select tetranucleotide repeats having more than 11 but less than 16 repeats. The lower limit is based on reported higher mutation rate for alleles with ≥11 repeats, thus increasing the chance of identifying highly polymorphic loci. The upper limit was defined based on the assumption that alleles with more than 16 repeats have a higher probability of accumulating interrupted motifs that confound the interpretation of the results.

Primer design

Once the sequences harbouring repeat motifs have been identified, suitable primers must be chosen. To develop high-quality multiplexed SSRs, stringent selection of markers is necessary (Varshney *et al.* 2005). Primer pairs that amplify fragments of contrasted sizes (e.g. about 100, 200 and 300 bp) should be chosen to permit amplification of several non-overlapping markers with a single dye. Computer programs that simultaneously identify SSRs and design primers for multiplex exist (Kaplinski *et al.* 2005; Rachlin *et al.* 2005; Kraemer *et al.* 2009; Shen *et al.* 2010). Some of them search for suitable combinations of primer pairs for multiplex PCR and handle large data sets automatically. To ensure the success of co-amplification, it is critical to eliminate primers with potential primer–dimer interactions (Vallone & Butler 2004; van Asch *et al.* 2010). A local blast or dedicated tools such as Multiplex Manager (Holleley & Geerts 2009) or NetPrimer (Premier Biosoft International, USA) can be used for this purpose (Appendix 1).

For multiplexing, primer pairs should have similar annealing temperature range [58–60 °C has been considered to be optimal (Butler 2005a; Hill *et al.* 2009)]. If primers have been developed previously and have different melting temperatures, primer redesign should be considered before multiplexing. However, redesign should be restricted to specific cases, such as when available SSRs are in short supply or when the corresponding SSRs are of special interest. Another possibility to buffer annealing temperatures is to add some extra sequence to primers (e.g. 5'-ACGTTGGATG-3'), thereby bringing GC% closer to 50% (Ghebranious *et al.* 2005). The presence of nanosatellites (i.e. low-complexity sequences that are too short to qualify as microsatellites) in the amplicons should be avoided. Since nanosatellites are abundant, this reduces the size of flanking sequences available for design, which can be problematic when selecting primers that amplify longer amplicons. This has been taken into account in the computer program QDD designed to isolate microsatellite loci from libraries of thousands of DNA fragments (Meglécz *et al.* 2010).

Primer validation in simplex

It is important to fully validate primer pairs early in the development process, so as to avoid losing time later with inefficient primers or uninformative loci (Fig. 5). In particular, SSR loci presenting excessive stuttering, split peaks, null alleles, low heterozygote peak height ratios and other artefacts should be identified early on and discarded or primers redesigned (Box 3). For this purpose, SSRs need to be tested in simplex, e.g. using labelled M13-tails (Schuelke 2000). Briefly, the primer mix contains a forward primer that has a specific sequence at its 5' end (the M13-tail), a reverse primer and a universal fluorescent-labelled M13-tail. This technique is economic because the cost of direct fluorescent primer labelling is

Box 3 Problems arising during SSR amplification

A number of problems can arise during amplification. They can compromise allele calling and binning, resulting in increased error rates or extensive need for manual corrections, and should therefore be identified as early as possible (Figs 3 and 4):

(1) Low heterozygote peak height ratios (Fig. 3b). They are caused by mutations in the flanking region at primers binding sites, resulting in poor amplification of the corresponding allele. Possible solutions to avoid them are similar to those put forth for null alleles below.

(2) Stuttering or shadow bands (Fig. 3c). This corresponds to the amplification of PCR products that differ from the original template by one or a few repeats. This widespread phenomenon complicates the interpretation of electropherograms. Because of a strong bias towards contractions, stutter bands are typically shorter than the original fragment (Shinde *et al.* 2003). To reduce stuttering, one option is to decrease denaturation temperature to 83 °C (Olejniczak & Krzyzosiak 2006), another is to use new-generation polymerases, such as fusion enzymes (Fazekas *et al.* 2010). However, the best solution is to select loci that present reduced stuttering from the outset (e.g. O'Reilly *et al.* 2000). Note that M13-tails labelling can result in slight stuttering because of low melting temperature of this primer (53 °C), so if primers are first tested in simplex with an M13-tail, some improvements can be expected at the time of multiplexing.

(3) Split peaks (Fig. 3d). This is caused by the non-template addition of a nucleotide (generally an adenine) to PCR fragments by the *Taq* polymerase (Clark 1988; Esselink *et al.* 2003). When this adenylation is incomplete, it results in double peaks (the original fragment and an additional peak 1 bp longer corresponding to the adenylated fragment), thereby compromising automatic peak recognition, particularly for heterozygote genotypes with nearby alleles. The addition of a guanine base (G), a 'PIG-tail' (5'-GTTTCTT-3' or 5'-GTTT-3'), or longer (40 bp) sequences at the 5' end of the reverse (non-labelled) primer has been shown to promote full adenylation of some fragments during PCR (Brownstein *et al.* 1996; Binladen *et al.* 2007; Hill *et al.* 2009). However, according to our observations, PCR efficiency can decrease with such tailed primers. This can in some cases be compensated by increasing the number of amplification cycles, as shown for primers with M13-tails (de Arruda *et al.* 2010). Other suggestions to promote complete adenylation include the reduction in the amount of template DNA, down to 10 ng (Lederer *et al.* 2000; Butler 2005b), the decrease in primer concentration, the increase in *Taq* concentration (Fishback *et al.* 1999) or the use of alternative polymerases (Hu 1993; Vallone *et al.* 2008).

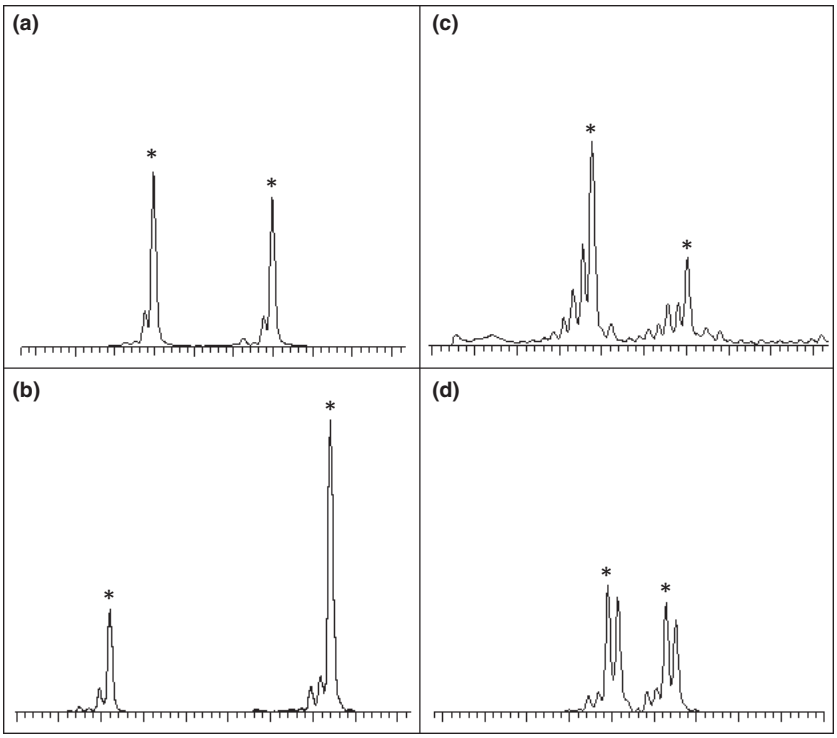

Fig. 3 Illustration of SSR profiles generated on capillary sequencer: correct profile (a), low heterozygote peak height ratios (b), excessive stuttering (c) and split peaks (d). Correct alleles are marked with asterisks.

(4) Null alleles (Figs 4a,b). These are non-amplifying alleles that result in an apparent homozygote when present in heterozygote state and in the lack of amplification when present in homozygote state. In the latter case, they can be confounded with reaction failure (Varshney *et al.* 2005). Null alleles are produced by mutations in the flanking region, at primer binding sites. When null alleles are present, observed banding patterns represent one of several possible true genotypes. While methods have been developed to mitigate this problem during data analysis (e.g. Wagner *et al.* 2006; Chapuis & Estoup 2007), the best approach is to avoid design primers in polymorphic regions, either using prior information on sequence variation (Meglécz *et al.* 2010) or by checking early on all candidate loci using Mendelian segregation analyses. In our laboratory, we use 12 or 24 progenies (one mother and seven of her open-pollinated progenies) representing one or two 96-well plates. The use of full-sib families (e.g. the mother, the father and six offspring) would be twice as informative by screening both the mother and the father for the presence of null alleles. If such approaches are not feasible, deviations from Hardy–Weinberg equilibrium proportions can be investigated (van Oosterhout *et al.* 2004). For large-scale population studies, markers should be validated on multiple populations to minimize null allele occurrence (Sinama *et al.* 2011). In the 100 studies that we surveyed, explicit tests of the presence of null alleles were reported in only 40% of the studies.

(5) Primer-dimers, artifactual bands (Fig. 4c) and triallelic patterns (Fig. 4d). These can be caused by the mispriming of primers (Brownie *et al.* 1997; Hill *et al.* 2009). Although the artefacts produced could be simply omitted during scoring if they do not interfere with allele calling, they may be a criterion for exclusion or redesign to facilitate automatic interpretation of electropherograms.

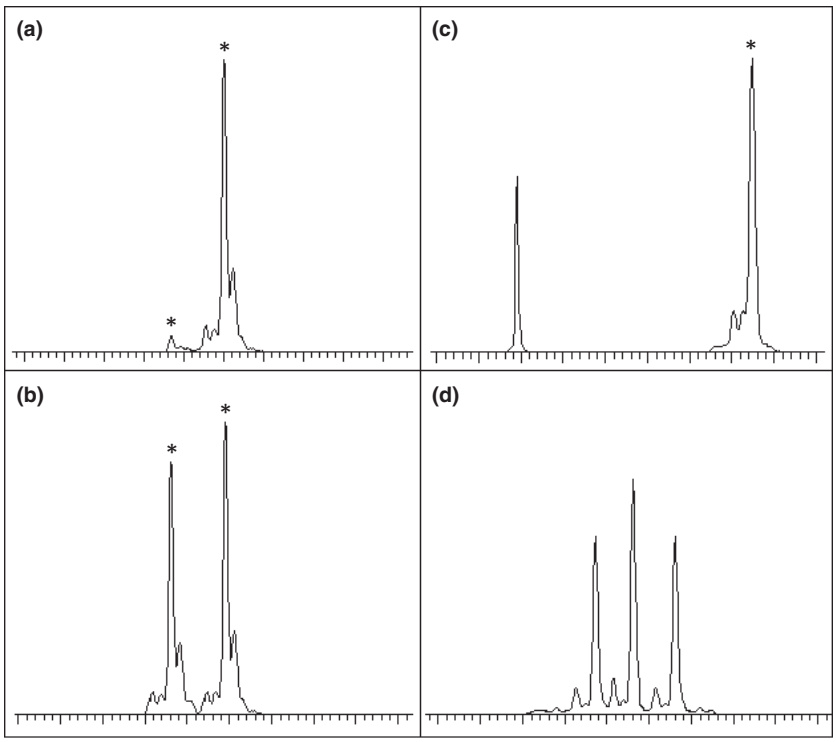

Fig. 4 Illustration of SSR profiles generated on capillary sequencer: weak allele before (a) and after (b) successful primer redesign, artifactual band (c) and triallelic pattern (d). Correct alleles are marked with asterisks.

typically five to ten times higher than the cost of the synthesis of an unlabelled primer (Hayden *et al.* 2008). However, the PCR conditions required for amplification using the M13-tailed primer method are often somewhat different from those optimal for amplification using standard length primers, which could create difficulties if the PCR protocol is tested in simplex with M13-tailed primers and then in multiplex with labelled primers but without M13-tail. In particular, M13-tails appear to decrease PCR efficiency, resulting in a need for additional PCR amplification cycles (de Arruda *et al.* 2010). The samples used for validation of the primers should be representative of the genetic diversity (i.e. originating from different populations) to identify most alleles early on (Sinama *et al.*

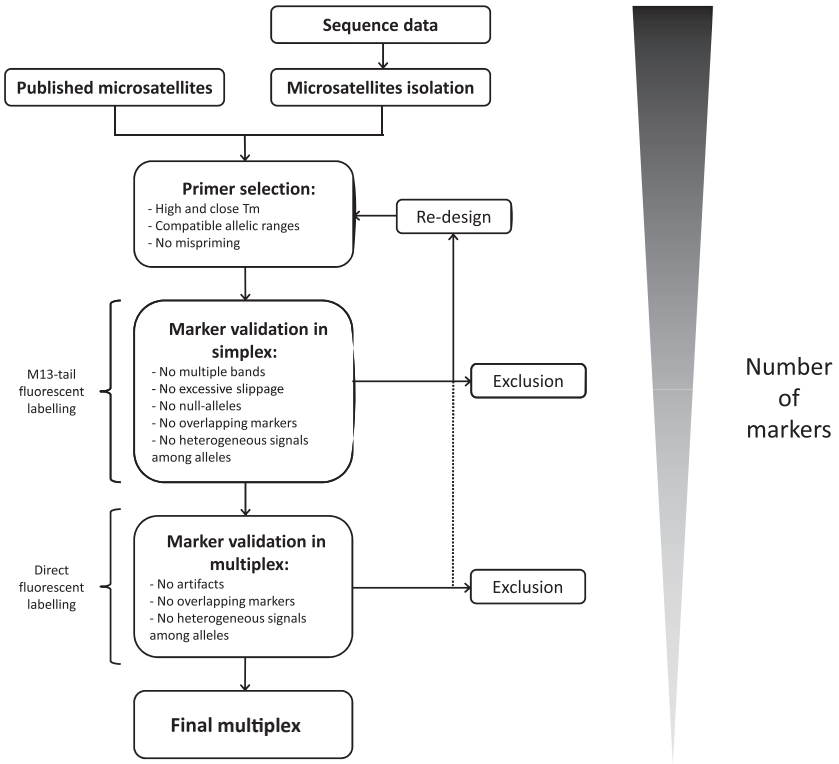

Fig. 5 One possible strategy for the development of multiplex SSRs suitable for high-throughput genotyping.

2011). This will minimize the risks to subsequently discover new alleles differing widely in size and overlapping with the allelic range of other loci labelled with the same fluorochrome, thereby compromising allele scoring. DNA pooling has been suggested as a cost-effective way to expedite this phase (Collins *et al.* 2000; Cryer *et al.* 2005).

The multiplexing phase

The throughput of standard (i.e. simplex) SSR analysis is low as it yields genotype information at only one locus per reaction. In contrast, multiplex PCR can boost genotyping by reducing laboratory work and consumption of expensive reagents without compromising test utility (Elnifro *et al.* 2000; Lederer *et al.* 2000; Galan *et al.* 2003; Renshaw *et al.* 2006 and see Box 2). Moreover, a reduced amount of DNA is needed to genotype a given number of loci (Karaiskou & Primmer 2008), even if for high levels of multiplexing, more DNA per reaction is necessary compared to standard simplex PCR (Chen *et al.* 2010). Another advantage is that multiplex PCR provides better indications on template quantity and quality (Edwards & Gibbs 1994). Potential problems in PCR include false negatives owing to reaction failure or false positives owing to contamination. In particular, complete PCR failure can be more easily distinguished from an informative no amplification. In view of these advantages, multiplexing SSRs should be a priority in all but the smallest SSR genotyping projects (Box 2).

The objective of the multiplexing phase is to combine all markers into the smallest number of reactions or select a subset of markers to design efficient and robust multiplexes, with each locus assigned a given florescent dye. A computer program (Multiplex Manager 1.0) has been developed to perform this task using prior marker information (Holleley & Geerts 2009). It minimizes the differences in annealing temperature and maximizes the spacing between markers, the heterozygosity and the number of alleles (Fig. 6).

Multiplex PCR is a sensitive technique. To obtain repeatable results, careful standardization of all steps is needed. In particular, DNA concentration should be standardized (e.g. Livingstone *et al.* 2009), if possible using automated pipetting robots. Although too little DNA can result in poor amplification, including imbalance among loci and allele dropout, too much DNA is generally more problematic. It can lead to off-scale fluorescent signal and to various PCR artefacts, such as imbalance among loci, incomplete adenylation of PCR products and enhanced strand-slippage or 'stutter' of various forms (Kline *et al.* 2005). The use of specialized multiplex PCR buffer (e.g. Qiagen PCR Multiplex kit) can help overcome some problems during PCR, particularly if a high level of multiplexing is targeted (Anonymous, 2002). In our survey, all studies with high level of multiplexing (≥8-plex) used

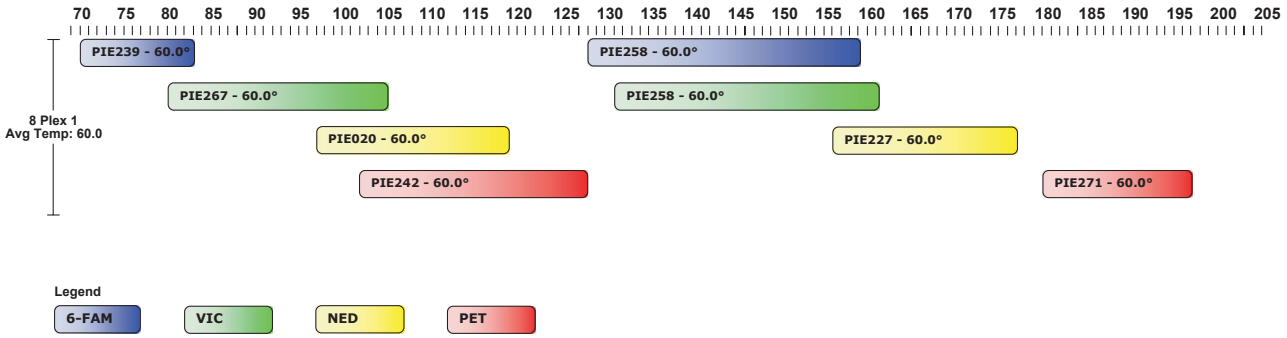

Fig. 6 Example of output obtained with Multiplex Manager software (Holleley & Geerts 2009). This software is used to identify combinations of markers suitable for multiplex reactions. In this example, for each of the eight SSRs, one of the four dyes (6-FAM, VIC, NED and PET) is assigned and the allele size range is provided along the main axis (in base pairs).

the Qiagen PCR Multiplex kit. This kit relies on a synthetic factor that allows efficient primer annealing and extension irrespective of primer sequence, by increasing the local concentration of primers at the DNA template and stabilizing specifically bound primers (Anonymous, 2002). Whereas excellent results have been obtained without resorting to the use of specialized multiplex buffers, by stringent optimization of all parameters (e.g. Hill *et al.* 2009), such buffers should be particularly useful when primers have different optimal annealing temperatures (Anonymous, 2002; Karaiskou & Primmer 2008). Touchdown PCR protocols can also be used to amplify heterogeneous SSR sets via progressively reducing annealing temperature in successive annealing cycles, so that the optimal annealing temperature of every primer pairs is matched at some point during PCR (Rithidech & Dunn 2003; Renshaw *et al.* 2006).

Even when stringent selection of SSRs has been performed on the basis of simplex PCR, problems can occur during the multiplexing step, in particular heterogeneous amplification of the different SSR loci (i.e. locus-to-locus imbalance). To limit this problem, primers should have similar annealing temperatures, as pointed out before. If differences are nevertheless observed following multiplexing, a first possibility is to increase the primer concentration for the weakest markers or alternatively decrease primer concentration for the strongest ones, and repeat the process to adjust locus-to-locus balance. Obtaining uniform amplification signal facilitates automatic reading of the electropherograms. Using different fluorochromes in multiplexes might also produce dye-induced mobility shift, which can lead to allele mis-scoring, with size differences between dyes (for the same allele) up to 3.7 bp (Sutton *et al.* 2011). Hence, strict quality control must be used to limit genotyping errors (see Measuring and reporting error rates section).

To increase the consistency of genetic profiling protocols, testing the quantity and quality of fluorescently labelled primers can be relevant. A simple method to assess primers quality on capillary electrophoresis system has been developed by checking profiles or fluorescence intensity in comparison with standards (Frasier & White 2008). This should help reduce variation in amplification among primer batches and among dyes. Another precaution is to limit the frequency of freeze–thaw cycles that can accelerate the breakdown of the dye attachment to the oligonucleotide, resulting in heterogeneous signals (Butler 2005a).

In general, for moderate multiplexing (≤8 loci), there is no need for extensive optimization if all precautions outlined in Fig. 5 are taken. In this respect, the situation has greatly changed compared to a few years ago when primer-to-template ratio, dNTP/$MgCl_2$ balance and PCR buffer concentration had to be carefully optimized and multiple rounds of changes in primer concentration were considered unavoidable (Henegariu *et al.* 1997; Markoulatos *et al.* 2002). However, for highly multiplexed sets (>12 SSRs), more advanced strategies might still be necessary. Hill *et al.* (2009) have proposed a method that relies on a core set of co-amplifying markers to which other primers are added one after another. If difficulties are encountered, the primer causing the problem is identified by successively adding each primer in the multiplex primer mix. However, intensive optimization such as that proposed by Hill *et al.* (2009) must only be considered in exceptional cases.

Sizing precision

Sizing precision is defined as the ability to reproducibly estimate fragment sizes from run to run on a given instrument (Moretti *et al.* 2001; Greenspoon *et al.* 2008). It is calculated by averaging the standard deviation of size estimates across alleles at each locus. Imprecise sizing directly translates into genotyping errors, especially when the spacing of alleles is minimal (Ghosh *et al.* 1997). For alleles 1 base apart, the tolerance level is normally set at a value near 0.2 bp. Precision depends on capillary length

and voltage as well as of the detection window and the detection integration time. It can also be affected by temperature fluctuations, polymer and capillary effects (Hartzell *et al.* 2003; Sgueglia *et al.* 2003) or by the type of fluorescent dye used (Hahn *et al.* 2001). Limiting variation in PCR conditions should also help (Ghosh *et al.* 1997).

'Allelic drift' is the tendency for true allele sizes to differ by a value slightly different from the known repeat length. At dinucleotide SSRs, for instance, the effective spacing between peaks of observed allele sizes has been shown to vary between 1.8 and 2.2 bp (Amos *et al.* 2007). Spacing of adjacent alleles decreases with increases in PCR product size, thereby reducing precision (Idury & Cardon 1997). The precision should however still be sufficient to distinguish reliably one base pair difference for fragments >300 bp (Koumi *et al.* 2004).

Allele calling and binning

Once large data sets of multiplexed SSR markers have been collected from capillary sequencing machines, the corresponding genotypes need to be read. There are two distinct steps in this process: true allele size calling, i.e. using decimal numbers, and binning, i.e. the conversion of alleles from real-valued DNA fragment sizes into discrete units to which an integer label is assigned (Idury & Cardon 1997).

The first step of the analysis is allele calling, i.e. identifying peaks that correspond to alleles and measuring the size of the corresponding fragments. Commercial software provided by constructors of capillary electrophoresis systems decreases analysis set-up time through automated correction of common genotyping problems, including saturated peaks, excessive baseline noise, voltage spikes caused by micro-air bubbles or debris in the laser path, and stutter peaks. However, depending on the quality of the markers, allele calling often necessitates additional manual editing. As this step can be labour intensive and can generate errors, it is important to select well-behaved markers at the outset, as emphasized before (Scandura *et al.* 2006).

The next step, allelic binning, is critical (Morin *et al.* 2010). In one comparative study, 83% of discrepancies between laboratories in scoring dinucleotide alleles were caused by arbitrary decisions in binning (Weeks *et al.* 2002). In another study, binning errors accounted for 21% to 40% of all errors (Ewen *et al.* 2000), confirming the necessity of well-established reading rules. Interestingly, in our survey, most authors (95%) used software with automatic binning module. We assume that these studies relied on user-friendly automated binning procedure (Appendix 1) and possibly on manual checks, rather than on direct analysis of raw fragment sizes, hence increasing risks of genotyping errors (Amos *et al.* 2007).

Because integer labels may not directly reflect the underlying allele sizes, raw allele sizes need to be stored for later reference and comparisons. One efficient and simple procedure is to export raw fragment size data to a spreadsheet and use it to compile cumulative frequency plots of size distributions (Jayashree *et al.* 2006). New bins for the inferred number of repeats can then be constructed around these distributions at places where discrete breaks in periodic size classes are evident. In this way, alleles that deviate from the expected periodicity of repeats (i.e. off-ladder microvariants) can be identified. Software has been designed for this step. ALLELOBIN and FLEXIBIN use least-squares minimization procedures and allow for allelic drift (Idury & Cardon 1997; Amos *et al.* 2007). TANDEM has been specifically designed for integration into population genetic and genomic workflows and requires no additional reformatting of data files (Matschiner & Salzburger 2009). MsatAllele is a computer package built on R to visualize and bin the raw microsatellite allele size distributions (Alberto 2009). It uses files exported from the open source electropherogram peak-reading program STRand. Genotype files with the resulting binned data can then be exported. In our laboratory, we developed an Excel macro, inspired from FlexiBin (Amos *et al.* 2007), Autobin (http://www4.bordeaux-aquitaine.inra.fr/biogeco/Ressources/Logiciels/Autobin), which automatically analyses raw data generated with commercial software (Appendix 1). The number of samples and loci is automatically detected, alleles in raw sizes are sorted and plotted to detect relevant gaps in size (Fig. 7), alleles are binned (with manual checking), and the whole data set is formatted for GENEPOP (Raymond & Rousset 1995) or STRUCTURE (Pritchard *et al.* 2000).

Thousands of data sets that could potentially be expanded as samples become available are regarded as lost because of the effort that would be required to validate congruence of genotypes from old and new data sets (Presson *et al.* 2008; Morin *et al.* 2009). To take advantage of past studies, specific software has been designed (ALLELOGRAM and MicroMerge). These two software programs can normalize and bin alleles from multiple data sources using a relatively small set of controls (Appendix 1). Binning can also be harmonized using reference genotypes and allelic ladders (Gill *et al.* 2001; LaHood *et al.* 2002; Rathmacher *et al.* 2009).

Measuring and reporting error rates

Error rates per locus and per individual should be systematically measured and reported in genotyping studies. In our survey, however, genotyping error rates were reported in only 26% of the studies. In genotyping studies relying on multiplexing, measuring error rates is

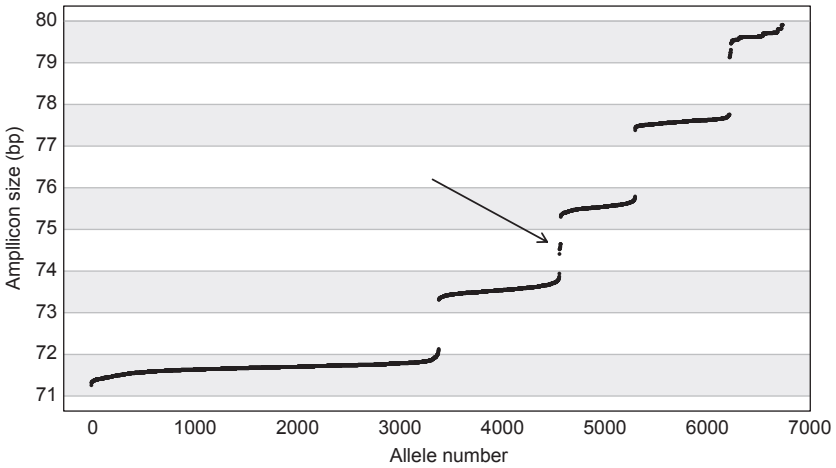

Fig. 7 Size distribution of 6762 alleles for one dinucleotide EST-SSR developed on oaks, achieved with the macro we developed. The arrow indicates the presence of an off-ladder microvariant found in 13 alleles that differs by one base pair from the expected periodicity of 2 bp. Analyses of segregating progenies have confirmed that this variant corresponds to a different allele.

particularly important (Luikart *et al.* 2008), because information on locus-specific error rates is necessary to improve multiplex assays. Genotyping error rates can be estimated by counting Mendelian inconsistencies in parent–offspring pairs or by counting mismatches between duplicated genotypes (Bonin *et al.* 2004; Hoffman & Amos 2005; Pompanon *et al.* 2005; DeWoody *et al.* 2006; Johnson & Haydon 2007). This second option can be further subdivided into two cases, depending on whether duplicated genotypes include or not a well-characterized control (i.e. concordance checking using standard reference genotypes vs. regenotyping of a random subset of genotypes). Clearly, none of these approaches allow the identification of all genotyping problems. For instance, in parent–offspring comparisons, not all errors result in Mendelian inconsistencies. Similarly, with duplicated samples, some problems, such as mutations or null alleles, cannot be identified (Ewen *et al.* 2000). When randomly regenotyping samples in the absence of reference sample, some errors might remain unnoticed, as when a heterozygous genotype is genotyped twice as a homozygote. Moreover, when the duplicated genotypes differ, the nature of the error can sometimes be difficult to establish. In particular, it might not be possible to distinguish between allelic dropout (failure to amplify one of the two alleles in heterozygotes) and false alleles (caused by polymerase errors) (Broquet & Petit 2004). This is unfortunate because the two classes of error affect analyses in different ways (Wang 2004; Hadfield *et al.* 2006). Hence, multiple strategies should be used whenever possible, concentrating on pedigree evaluation and regenotyping with reference samples. Nevertheless, from a practical point of view, regenotyping to get complete data set in multiplex surveys means that, as a by-product of this process, individuals will be genotyped several times at some of the loci, thereby providing more accurate error rate measurements. Software has been developed to estimate error rates and break them down into different categories (reviewed in Johnson & Haydon 2007).

Data management

The utility of genotyping techniques is only as good as one's ability to handle the flood of data produced from them. Managing genotyping data can indeed be challenging. In particular, because records for a particular sample might have to be revised over time, the management system must keep track of each DNA sample during the whole process. Genotyping data must be kept as raw data for future work (in the same laboratory or in another laboratory) to avoid laborious normalization work. Database management systems or Laboratory Information Management Systems (LIMS) specialized in genotyping data have been released to meet these demands (Li *et al.* 2001; Jayashree *et al.* 2006; van Rossum *et al.* 2010). Besides serving as workflow managers, these systems also provide visible quality checks and centralization of data, but their use is far from being commonplace.

Conclusions and perspectives

There are many applications in molecular ecology where 10–30 highly polymorphic markers such as SSRs would suffice to provide precise answers (Box 1). During the last years, considerable progresses have been made in SSR development and genotyping, including in associated bioinformatics. However, the efforts remain somewhat disparate, and current practices are lagging behind. As a consequence, SSR markers are not used to their full power, as shown by our survey of a sample of the recent literature. Hence, additional efforts to improve SSR isolation, multiplex genotyping and scoring remain critical.

The identification of SSR motifs has long been a bottleneck in studies involving non-model species for which

sequence data are not readily available. The use of next-generation sequencing techniques instead of cloning and conventional sequencing to obtain sequence data and identify SSRs in such species is just beginning and appears extremely promising. It provides the optimal conditions for subsequent multiplex development by detecting many potential SSRs. In fact, the throughput and cost-effectiveness of next-generation sequencing should allow researchers to be more selective in their choice of SSR loci. In particular, sequencing depth should provide sufficient data on sequence variation to focus on conserved regions flanking polymorphic SSR motifs for designing primers, considerably simplifying the whole process of marker testing.

The number of multiplexed markers could be increased, because there is no major limitation in combining up to 30 or 40 SSRs in a single PCR (Gabriel *et al.* 2009; Hill *et al.* 2009). Increasing the number of fluorochromes could also help. Multiplexing should not only increase throughput but also accuracy. The latter point might not be immediately obvious. However, designing a good multiplex is demanding, hence forcing researchers to take a number of precautions and to better evaluate candidate loci, which eventually benefits to the whole genotyping process. Better precision could also be achieved with new size standards or improved algorithms (Johansson *et al.* 2003). Automation, from DNA isolation to capillary electrophoresis, could be developed using appropriate robotics and high-throughput plate formats (384 or 1536 wells). Recently, laboratory-on-a-chip systems relying on microfluidic technology have been tested successfully for DNA amplification (Horsman *et al.* 2007; Sinville & Soper 2007; Greenspoon *et al.* 2008; Bienvenue *et al.* 2009; Liu & Mathies 2009; Petersen *et al.* 2009). Such systems potentially offer speed, automation, sensitivity and portability (Beyor *et al.* 2009). Completely different methods amenable to highly parallelized SSR assays might also emerge (e.g. Pettersson *et al.* 2006; Zajac *et al.* 2009).

With the outbreak of next-generation sequencing technologies, SSR genotyping could eventually be performed via sequencing of amplified fragments. The million reads obtained could make it possible to genotype hundreds of samples at thousands of loci, provided these samples can be identified prior to sequencing (e.g. with short ligated sequence tags). This would result in a drastic reduction in genotyping costs and a substantial improvement of data quality. Indeed, direct access to microsatellite motif sequence (rather than PCR product sizes) would reduce problems of homoplasy in data sets and avoid poor genotyping repeatability among laboratories using different equipments or reagents. However, such processes still need to be set up and must be associated with bioinformatic methods aiming at sorting sequences, correcting for sequencing errors and finally summarizing genotype information.

Acknowledgements

We are especially grateful to Christophe Boury for developing the robotics used in the frame of SSR genotyping and to Sarah Monllor for help with genotyping. We also thank Joëlle Chat, François Hubert and Stéphanie Mariette for their useful comments on the paper and Sophie Lefèvre for sharing with us her experience on the development of multiplexed SSRs in beech. The experience on genotyping was gained in our Genome-Transcriptome facility, which is part of the Functional Genomic Center of Bordeaux. We acknowledge financial support from the Aquitaine Region, from the EVOLTREE Network of Excellence supported by the 6th Framework Programme of the European Commission no. 016322 and from the LINKTREE project from the Eranet Biodiversa Programme (ANR-08-BDVA-006).

References

Abbott C, Ebert D, Tabata A, Therriault T (2010) Twelve microsatellite markers in the invasive tunicate, *Didemnum vexillum*, isolated from low genome coverage 454 pyrosequencing reads. *Conservation Genetics Resources*, **3**, 79–81.

Abdelkrim J, Robertson B, Stanton JA, Gemmell N (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. *BioTechniques*, **46**, 185–192.

Alberto F (2009) MsatAllele_1.0: an R package to visualize the binning of microsatellite alleles. *Journal of Heredity*, **100**, 394–397.

Allentoft M, Schuster SC, Holdaway R *et al.* (2009) Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. *BioTechniques*, **46**, 195–200.

Amos W, Hoffman JI, Frodsham A *et al.* (2007) Automated binning of microsatellite alleles: problems and solutions. *Molecular Ecology Notes*, **7**, 10–14.

Anonymous (2002) Multiplex PCR that simply works—the new QIAGEN Multiplex PCR Kit. *QIAGENews*, **5**, 14–16.

de Arruda M, Gonçalves E, Schneider M, da Costa da Silva A, Morielle-Versute E (2010) An alternative genotyping method using dye-labeled universal primer to reduce unspecific amplifications. *Molecular Biology Reports*, **37**, 2031–2036.

van Asch B, Pinheiro R, Pereira R *et al.* (2010) A framework for the development of STR genotyping in domestic animal species: characterization and population study of 12 canine X-chromosome loci. *Electrophoresis*, **31**, 303–308.

Ashley MV, Caballero IC, Chaovalitwongse W *et al.* (2009) KINALYZER, a computer program for reconstructing sibling groups. *Molecular Ecology Resources*, **9**, 1127–1131.

Berger-Wolf TY, Sheikh SI, DasGupta B *et al.* (2007) Reconstructing sibling relationships in wild populations. *Bioinformatics*, **23**, i49–i56.

Beyor N, Yi L, Seo TS, Mathies RA (2009) Integrated capture, concentration, polymerase chain reaction, and capillary electrophoretic analysis of pathogens on a chip. *Analytical Chemistry*, **81**, 3523–3528.

Bienvenue JM, Legendre LA, Ferrance JP, Landers JP (2009) An integrated microfluidic device for DNA purification and PCR amplification of STR fragments. *Forensic Science International Genetics*, **4**, 178–186.

Binladen J, Gilbert MTP, Campos PF, Willerslev E (2007) 5′-Tailed sequencing primers improve sequencing quality of PCR products. *BioTechniques*, **42**, 174–176.

Bonin A, Bellemain E, Eidesen PB *et al.* (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, **13**, 3261–3273.

Boomer J, Stow A (2010) Rapid isolation of the first set of polymorphic microsatellite loci from the Australian gummy shark, *Mustelus antarcticus* and their utility across divergent shark taxa. *Conservation Genetics Resources*, **2**, 393–395.

Borsting C, Rockenbauer E, Morling N (2009) Validation of a single nucleotide polymorphism (SNP) typing assay with 49 SNPs for forensic genetic testing in a laboratory accredited according to the ISO 17025 standard. *Forensic Science International Genetics*, **4**, 34–42.

Broquet T, Petit E (2004) Quantifying genotyping errors in noninvasive population genetics. *Molecular Ecology*, **13**, 3601–3608.

Brownie J, Shawcross S, Theaker J *et al.* (1997) The elimination of primer-dimer accumulation in PCR. *Nucleic Acids Research*, **25**, 3235–3241.

Brownstein MJ, Carpten D, Smith JR (1996) Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. *BioTechniques*, **20**, 1004–1010.

Buchan JC, Archie EA, Van Horn RC, Moss CJ, Alberts SC (2005) Locus effects and sources of error in noninvasive genotyping. *Molecular Ecology Notes*, **5**, 680–683.

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.

Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. *Bioessays*, **28**, 1040–1050.

Butler JM (2005a) Constructing STR multiplex assays. *Methods in Molecular Biology*, **297**, 53–65.

Butler JM (2005b) *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. Elsevier Academic Press, London.

Butler JM, Ruitberg CM, Vallone PM (2001) Capillary electrophoresis as a tool for optimization of multiplex PCR reactions. *Fresenius Journal of Analytical Chemistry*, **369**, 200–205.

Butler JM, Buel E, Crivellente F, McCord BR (2004) Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. *Electrophoresis*, **25**, 1397–1412.

Castoe TA, Poole AW, Gu WJ *et al.* (2010) Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Molecular Ecology Resources*, **10**, 341–347.

Chamberlain JS, Gibbs RA, Rainer JE, Nguyen PN, Thomas C (1988) Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic Acids Research*, **16**, 11141–11156.

Chambers GK, MacAvoy ES (2000) Microsatellites: consensus and controversy. *Comparative Biochemistry and Physiology—Part B: Biochemistry and Molecular Biology*, **126**, 455–476.

Chapuis MP, Estoup A (2007) Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution*, **24**, 621–631.

Chen JW, Uboh CE, Soma LR *et al.* (2010) Identification of racehorse and sample contamination by novel 24-plex STR system. *Forensic Science International Genetics*, **4**, 158–167.

Chistiakov DA, Hellemans B, Volckaert FAM (2006) Microsatellites and their genomic distribution, evolution, function and applications: a review with special reference to fish genetics. *Aquaculture*, **255**, 1–29.

Christians JK, Watt CA (2009) Mononucleotide repeats represent an important source of polymorphic microsatellite markers in *Aspergillus nidulans*. *Molecular Ecology Resources*, **9**, 572–578.

Cipriani G, Marrazzo MT, Di Gaspero G *et al.* (2008) A set of microsatellite markers with long core repeat optimized for grape (*Vitis* spp.) genotyping. *BMC Plant Biology*, **8**, 127.

Clark JM (1988) Novel non-templated nucleotide addition-reactions catalyzed by procaryotic and eukaryotic DNA-polymerases. *Nucleic Acids Research*, **16**, 9677–9686.

Clayton TM, Whitaker JP, Sparkes R, Gill P (1998) Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International*, **91**, 55–70.

Collins HE, Li H, Inda SE *et al.* (2000) A simple and accurate method for determination of microsatellite total allele content differences between DNA pools. *Human Genetics*, **106**, 218–226.

Cryer N, Butler D, Wilkinson M (2005) High throughput, high resolution selection of polymorphic microsatellite loci for multiplex analysis. *Plant Methods*, **1**, 3.

Csencsics D, Brodbeck S, Holderegger R (2010) Cost-effective, species-specific microsatellite development for the endangered dwarf bulrush (*Typha minima*) using next-generation sequencing technology. *Journal of Heredity*, **101**, 789–793.

Dawson DA, Horsburgh GJ, Küpper C *et al.* (2010) New methods to identify conserved microsatellite loci and develop primer sets of high cross-species utility—as demonstrated for birds. *Molecular Ecology Resources*, **10**, 475–494.

Dereeper A, Argout X, Billot C, Rami JF, Ruiz M (2007) SAT, a flexible and optimized Web application for SSR marker development. *BMC Bioinformatics*, **8**, 465.

DeWoody JA, Nason JD, Hipkins VD (2006) Mitigating scoring errors in microsatellite data from wild populations. *Molecular Ecology Notes*, **6**, 951–957.

Dieringer D, Schlötterer C (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Research*, **13**, 2242–2251.

Dubut V, Grenier R, Meglécz E *et al.* (2010) Development of 55 novel polymorphic microsatellite loci for the critically endangered *Zingel asper* L. (Actinopterygii: Perciformes: Percidae) and cross-species amplification in five other percids. *European Journal of Wildlife Research*, **56**, 931–938.

Ebert D, Peakall R (2009) Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. *Molecular Ecology Resources*, **9**, 673–690.

Edwards MC, Gibbs RA (1994) Multiplex PCR: advantages, development, and applications. *PCR Methods and Applications*, **3**, 65–75.

Edwards A, Civitello A, Hammond HA, Caskey CT (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics*, **49**, 746–756.

Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. *Trends in Genetics*, **16**, 551–558.

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

Elnifro EM, Ashshi AM, Cooper RJ, Klapper PE (2000) Multiplex PCR: optimization and application in diagnostic virology. *Clinical Microbiology Reviews*, **13**, 559–570.

Esselink GD, Smulders MJM, Vosman B (2003) Identification of cut rose (*Rosa hybrida*) and rootstock varieties using robust sequence tagged microsatellite site markers. *Theoretical and Applied Genetics*, **106**, 277–286.

Estoup A, Garnery L, Solignac M, Cornuet JM (1995) Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics*, **140**, 679–695.

Estoup A, Wilson IJ, Sullivan C, Cornuet J-M, Moritz C (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics*, **159**, 1671–1687.

Ewen KR, Bahlo M, Treloar SA *et al.* (2000) Identification and analysis of error types in high-throughput genotyping. *American Journal of Human Genetics*, **67**, 727–736.

Faircloth BC (2008) MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources*, **8**, 92–94.

Fazekas AJ, Steeves R, Newmaster SG (2010) Improving sequencing quality from PCR products containing long mononucleotide repeats. *BioTechniques*, **48**, 277–281.

Fishback AG, Danzmann RG, Sakamoto T, Ferguson MM (1999) Optimization of semi-automated microsatellite multiplex polymerase chain reaction systems for rainbow trout (*Oncorhynchus mykiss*). *Aquaculture*, **172**, 247–254.

Frasier TR, White BN (2008) Increased efficiency of genetic profiling through quantity and quality assessment of fluorescently labeled oligonucleotide primers. *BioTechniques*, **44**, 49–52.

Gabriel S, Ziaugra L, Tabbaa D (2009) SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Current Protocols in Human Genetics*, **2**, 1–18.

Galan M, Cosson JF, Aulagnier S *et al.* (2003) Cross-amplification tests of ungulate primers in roe deer (*Capreolus capreolus*) to develop a multiplex panel of 12 microsatellite loci. *Molecular Ecology Notes*, **3**, 142–146.

Ghebranious N, Ivacic L, Mallum J, Dokken C (2005) Detection of ApoE E2, E3 and E4 alleles using MALDI-TOF mass spectrometry and the homogeneous mass-extend technology. *Nucleic Acids Research*, **33**, e149.

Ghislain M, Spooner DM, Rodríguez F *et al.* (2004) Selection of highly informative and user-friendly microsatellites (SSRs) for genotyping of cultivated potato. *Theoretical and Applied Genetics*, **108**, 881–890.

Ghosh S, Karanjawala ZE, Hauser ER *et al.* (1997) Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers. FUSION (Finland-U.S. Investigation of NIDDM Genetics) Study Group. *Genome Research*, **7**, 165–178.

Gill P (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *International Journal of Legal Medicine*, **114**, 204–210.

Gill P, Brenner C, Brinkmann B *et al.* (2001) DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. *International Journal of Legal Medicine*, **114**, 305–309.

Glaubitz JC, Rhodes OE, DeWoody JA (2003) Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Molecular Ecology*, **12**, 1039–1047.

Greenspoon SA, Yeung SH, Johnson KR *et al.* (2008) A forensic laboratory tests the Berkeley microfabricated capillary array electrophoresis device. *Journal of Forensic Sciences*, **53**, 828–837.

Guichoux E, Lagache L, Wagner S, Léger P, Petit RJ (2011) Two highly-validated multiplex (12-plex and 8-plex) for species delimitation and parentage analysis in oaks (*Quercus spp.*). *Molecular Ecology Resources*, **11**, 578–585.

Gupta PK, Rustgi S, Sharma S *et al.* (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Molecular Genetics and Genomics*, **270**, 315–323.

Gusmão L, Butler JM, Carracedo A *et al.* (2006) DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *International Journal of Legal Medicine*, **120**, 191–200.

Hadfield JD, Richardson DS, Burke T (2006) Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework. *Molecular Ecology*, **15**, 3715–3730.

Hahn M, Wilhelm J, Pingoud A (2001) Influence of fluorophore dye labels on the migration behavior of polymerase chain reaction-amplified short tandem repeats during denaturing capillary electrophoresis. *Electrophoresis*, **22**, 2691–2700.

Hartzell B, Graham K, McCord B (2003) Response of short tandem repeat systems to temperature and sizing methods. *Forensic Science International*, **133**, 228–234.

Hayden MJ, Nguyen TM, Waterman A, Chalmers KJ (2008) Multiplex-ready PCR: a new method for multiplexed SSR and SNP genotyping. *BMC Genomics*, **9**, 80.

Henegariu O, Heerema NA, Dlouhy SR, Vance GH, Vogt PH (1997) Multiplex PCR: critical parameters and step-by-step protocol. *BioTechniques*, **23**, 504–511.

Hill CR, Butler JM, Vallone PM (2009) A 26plex autosomal STR assay to aid human identity testing. *Journal of Forensic Sciences*, **54**, 1008–1015.

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, **14**, 599–612.

Hoffman JI, Matson CW, Amos W, Loughlin TR, Bickham JW (2006) Deep genetic subdivision within a continuously distributed and highly vagile marine mammal, the Steller's sea lion (*Eumetopias jubatus*). *Molecular Ecology*, **15**, 2821–2832.

Holleley CE, Geerts PG (2009) Multiplex Manager 1.0: a cross-platform computer program that plans and optimizes multiplex PCR. *BioTechniques*, **46**, 511–517.

Horsman KM, Bienvenue JM, Blasier KR, Landers JP (2007) Forensic DNA analysis on microfluidic devices: a review. *Journal of Forensic Sciences*, **52**, 784–799.

Hu G (1993) DNA Polymerase-catalyzed addition of nontemplated extra nucleotides to the 3′ of a DNA fragment. *DNA and Cell Biology*, **12**, 763–770.

Idury RM, Cardon LR (1997) A simple method for automated allele binning in microsatellite markers. *Genome Research*, **7**, 1104–1109.

Jarne P, Lagoda PJL (1996) Microsatellites, from molecules to populations and back. *Trends in Ecology & Evolution*, **11**, 424–429.

Jayashree B, Reddy PT, Leeladevi Y *et al.* (2006) Laboratory Information Management Software for genotyping workflows: applications in high throughput crop genotyping. *BMC Bioinformatics*, **7**, 383.

Jewell MC, Frere CH, Prentis PJ, Lambrides CJ, Godwin ID (2010) Characterization and multiplexing of EST-SSR primers in *Cynodon* (Poaceae) species. *American Journal of Botany*, **97**, 99–101.

Johansson Å, Karlsson P, Gyllensten U (2003) A novel method for automatic genotyping of microsatellite markers based on parametric pattern recognition. *Human Genetics*, **113**, 316–324.

Johnson PCD, Haydon DT (2007) Software for quantifying and simulating microsatellite genotyping error. *Bioinformatics and Biology Insights*, **2007**, 71–75.

Jones B, Walsh D, Werner L, Fiumera A (2009) Using blocks of linked single nucleotide polymorphisms as highly polymorphic genetic markers for parentage analysis. *Molecular Ecology Resources*, **9**, 487–497.

Jones AG, Small CM, Paczolt KA, Ratterman NL (2010) A practical guide to methods of parentage analysis. *Molecular Ecology Resources*, **10**, 6–30.

Jones OR, Wang J (2010) COLONY: a program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, **10**, 551–555.

Kaplinski L, Andreson R, Puurand T, Remm M (2005) MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics*, **21**, 1701–1702.

Karaiskou N, Primmer C (2008) PCR multiplexing for maximising genetic analyses with limited DNA samples: an example in the collared flycatcher, *Ficedula albicollis*. *Annales Zoologici Fennici*, **45**, 478–482.

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research*, **18**, 30–38.

Kelkar YD, Strubczewski N, Hile SE *et al.* (2010) What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. *Genome Biology and Evolution*, **2**, 620–635.

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.

Kim TS, Booth J, Gauch H *et al.* (2008) Simple sequence repeats in *Neurospora crassa*: distribution, polymorphism and evolutionary inference. *BMC Genomics*, **9**, 31.

Kircher M, Kelso J (2010) High-throughput DNA sequencing—concepts and limitations. *Bioessays*, **32**, 524–536.

Kirov G, Williams N, Sham P, Craddock N, Owen MJ (2000) Pooled genotyping of microsatellite markers in parent-offspring trios. *Genome Research*, **10**, 105–115.

Kline MC, Duewer DL, Redman JW, Butler JM (2005) Results from the NIST 2004 DNA quantitation study. *Journal of Forensic Sciences*, **50**, 571–578.

Koumi P, Green HE, Hartley S *et al.* (2004) Evaluation and validation of the ABI 3700, ABI 3100, and the MegaBACE 1000 capillary array electrophoresis instruments for use with short tandem repeat microsatellite typing in a forensic environment. *Electrophoresis*, **25**, 2227–2241.

Kraemer L, Beszteri B, Gabler-Schwarz S *et al.* (2009) STAMP: extensions to the STADEN sequence analysis package for high throughput interactive microsatellite marker design. *BMC Bioinformatics*, **10**, 41.

Kruglyak L (1997) The use of a genetic map of biallelic markers in linkage studies. *Nature Genetics*, **17**, 21–24.

LaHood ES, Moran P, Olsen J, Grant WS, Park LK (2002) Microsatellite allele ladders in two species of Pacific salmon: preparation and field-test results. *Molecular Ecology Notes*, **2**, 187–190.

Lai E (2001) Application of SNP technologies in medicine: lessons learned and future challenges. *Genome Research*, **11**, 927–929.

Lederer T, Seidl S, Graham B, Betz P (2000) A new pentaplex PCR system for forensic casework analysis. *International Journal of Legal Medicine*, **114**, 87–92.

Lepais O, Bacles C (2010) Comparison of random and SSR-enriched shotgun pyrosequencing for microsatellite discovery and single multiplex PCR optimisation in *Acacia harpophylla* F. Muell. Ex Benth. *Molecular Ecology Resources*, DOI: 10.1111/j.1755-0998.2011.03002.x.

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, **4**, 203–221.

Li JL, Deng H, Lai DB *et al.* (2001) Toward high-throughput genotyping: dynamic and automatic software for manipulating large-scale genotype data using fluorescently labeled dinucleotide markers. *Genome Research*, **11**, 1304–1314.

Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, **11**, 2453–2465.

Li JZ, Absher DM, Tang H *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100–1104.

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

Liu P, Mathies RA (2009) Integrated microfluidic systems for high-performance genetic analysis. *Trends in Biotechnology*, **27**, 572–581.

Livingstone D, Freeman B, Tondo CL *et al.* (2009) Improvement of high-throughput genotype analysis after implementation of a dual-curve Sybr Green I-based quantification and normalization procedure. *HortScience*, **44**, 1228–1232.

Ljungqvist M, Åkesson M, Hansson B (2010) Do microsatellites reflect genome-wide genetic diversity in natural populations? A comment on Väli *et al.* (2008). *Molecular Ecology*, **19**, 851–855.

Luikart G, Zundel S, Rioux D *et al.* (2008) Low genotyping error rates and noninvasive sampling in bighorn sheep. *Journal of Wildlife Management*, **72**, 299–304.

Malausa T, Gilles A, Meglécz E *et al.* (2011) High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. *Molecular Ecology Resources*, DOI: 10.1111/j.1755-0998.2011.02992.x.

Markoulatos P, Siafakas N, Moncany M (2002) Multiplex polymerase chain reaction: a practical approach. *Journal of Clinical Laboratory Analysis*, **16**, 47–51.

Martin J-F, Pech N, Meglécz E *et al.* (2010) Representativeness of microsatellite distributions in genomes, as revealed by 454 GS-FLX Titanium pyrosequencing. *BMC Genomics*, **11**, 560.

Matschiner M, Salzburger W (2009) TANDEM: integrating automated allele binning into genetics and genomics workflows. *Bioinformatics*, **25**, 1982–1983.

McCarroll SA, Kuruvilla FG, Korn JM *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*, **40**, 1166–1174.

Meglécz E, Costedoat C, Dubut V *et al.* (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics*, **26**, 403–404.

Meldgaard M, Morling N (1997) Detection and quantitative characterization of artificial extra peaks following polymerase chain reaction amplification of 14 short tandem repeat systems used in forensic investigations. *Electrophoresis*, **18**, 1928–1935.

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

Missiaggia A, Grattapaglia D (2006) Plant microsatellite genotyping with 4-color fluorescent detection using multiple-tailed primers. *Genetics and Molecular Research*, **5**, 72–78.

Mittal N, Dubey A (2009) Microsatellite markers—A new practice of DNA based markers in molecular genetics. *Pharmacognosy Reviews*, **3**, 235–246.

Moretti TR, Baumstark AL, Defenbaugh DA *et al.* (2001) Validation of STR typing by capillary electrophoresis. *Journal of Forensic Sciences*, **46**, 661–676.

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

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

Morin PA, Manaster C, Mesnick SL, Holland R (2009) Normalization and binning of historical and multi-source microsatellite data: overcoming the problems of allele size shift with ALLELOGRAM. *Molecular Ecology Resources*, **9**, 1451–1455.

Morin PA, Martien KK, Archer FI *et al.* (2010) Applied conservation genetics and the need for quality control and reporting of genetic data used in fisheries and wildlife management. *Journal of Heredity*, **101**, 1–10.

Neff BD, Fu P, Gross MR (2000) Microsatellite multiplexing in fish. *Transactions of the American Fisheries Society*, **129**, 584–593.

Ohashi J, Tokunaga K (2003) Power of genome-wide linkage disequilibrium testing by using microsatellite markers. *Journal of Human Genetics*, **48**, 487–491.

Olejniczak M, Krzyzosiak WJ (2006) Genotyping of simple sequence repeats factors implicated in shadow band generation revisited. *Electrophoresis*, **27**, 3724–3734.

Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira MLC (2006) Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology*, **29**, 294–307.

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.

O'Reilly PT, Canino MF, Bailey KM, Bentzen P (2000) Isolation of twenty low stutter di- and tetranucleotide microsatellites for population analyses of walleye pollock and other gadoids. *Journal of Fish Biology*, **56**, 1074–1086.

Palsson B, Palsson F, Perlin M *et al.* (1999) Using quality measures to facilitate allele calling in high-throughput genotyping. *Genome Research*, **9**, 1002–1012.

Parchman T, Geist K, Grahnen J, Benkman C, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics*, **11**, 180.

Pashley CH, Ellis JR, McCauley DE, Burke JM (2006) EST databases as a source for molecular markers: lessons from *Helianthus*. *Journal of Heredity*, **97**, 381–388.

Petersen J, Poulsen L, Birgens H, Dufva M (2009) Microfludic device for creating ionic strength gradients over DNA microarrays for efficient DNA melting studies and assay development. *PLoS ONE*, **4**, e4808.

Petit RJ, Deguilloux M-F, Chat J *et al.* (2005) Standardizing for microsatellite length in comparisons of genetic diversity. *Molecular Ecology*, **14**, 885–890.

Pettersson E, Lindskog M, Lundeberg J, Ahmadian A (2006) Tri-nucleotide threading for parallel amplification of minute amounts of genomic DNA. *Nucleic Acids Research*, **34**, e49.

Phillips C, Fang R, Ballard D *et al.* (2007) Evaluation of the Genplex SNP typing system and a 49plex forensic marker panel. *Forensic Science International Genetics*, **1**, 180–185.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, **6**, 847–859.

Presson AP, Sobel EM, Pajukanta P *et al.* (2008) Merging microsatellite data: enhanced methodology and software to combine genotype data for linkage and association analysis. *BMC Bioinformatics*, **9**, 317.

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

Raabova J, Hans G, Risterucci A-M, Jacquemart A-L, Raspé O (2010) Development and multiplexing of microsatellite markers in the polyploid perennial herb, *Menyanthes trifoliata* (Menyanthaceae). *American Journal of Botany*, **97**, 31–33.

Rachlin J, Ding C, Cantor C, Kasif S (2005) MuPlex: multi-objective multiplex PCR assay design. *Nucleic Acids Research*, **33**, 544–547.

Rasmussen D, Noor M (2009) What can you do with 0.1x genome coverage? A case study based on a genome survey of the scuttle fly *Megaselia scalaris* (Phoridae). *BMC Genomics*, **10**, 382.

Rathmacher G, Niggemann M, Wypukol H *et al.* (2009) Allelic ladders and reference genotypes for a rigorous standardization of poplar microsatellite data. *Trees-Structure and Function*, **23**, 573–583.

Raymond M, Rousset F (1995) GENEPOP (Version 1.2)-Population-genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Renshaw MA, Saillant E, Bradfield SC, Gold JR (2006) Microsatellite multiplex panels for genetic studies of three species of marine fishes: red drum (*Sciaenops ocellatus*), red snapper (*Lutjanus campechanus*), and cobia (*Rachycentron canadum*). *Aquaculture*, **253**, 731–735.

Rithidech K, Dunn JJ (2003) Combining multiplex and touchdown PCR for microsatellite analysis. *Methods in Molecular Biology*, **226**, 295–300.

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

van Rossum T, Tripp B, Daley D (2010) SLIMS—a user-friendly sample operations and inventory management system for genotyping labs. *Bioinformatics*, **26**, 1808–1810.

Rozen S, Skaletsky H (1999) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, **132**, 365–386.

Ryynänen HJ, Tonteri A, Vasemägi A, Primmer CR (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.

Saarinen EV, Austin JD (2010) When technology meets conservation: increased microsatellite marker production using 454 genome sequencing on the endangered Okaloosa darter (*Etheostoma okaloosae*). *Journal of Heredity*, **101**, 784–788.

Sanchez JJ, Endicott P (2006) Developing multiplexed SNP assays with special reference to degraded DNA templates. *Nature Protocols*, **1**, 1370–1378.

Santana QC, Coetzee MPA, Steenkamp ET *et al.* (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. *BioTechniques*, **46**, 217–223.

Santure AW, Stapley J, Ball AD *et al.* (2010) On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology*, **19**, 1439–1451.

Scandura M, Capitani C, Iacolina L, Marco A (2006) An empirical approach for reliable microsatellite genotyping of wolf DNA from multiple noninvasive sources. *Conservation Genetics*, **7**, 813–823.

Schlötterer C (1998) Genome evolution: are microsatellites really simple sequences? *Current Biology*, **8**, 132–134.

Schlötterer C (2004) The evolution of molecular markers: just a matter of fashion? *Nature Reviews Genetics*, **5**, 63–69.

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology*, **18**, 233–234.

Schuster SC (2008) Next-generation sequencing transforms today's biology. *Nature Methods*, **5**, 16–18.

Schwengel DA, Jedlicka AE, Nanthakumar EJ, Weber JL, Levitt RC (1994) Comparison of fluorescence-based semi-automated genotyping of multiple microsatellite loci with autoradiographic techniques. *Genomics*, **22**, 46–54.

Seddon JM, Parker HG, Ostrander EA, Ellegren H (2005) SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. *Molecular Ecology*, **14**, 503–511.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, **9**, 615–629.

Sgueglia J, Geiger S, Davis J (2003) Precision studies using the ABI Prism 3100 Genetic Analyzer for forensic DNA analysis. *Analytical and Bioanalytical Chemistry*, **376**, 1247–1254.

Sharma PC, Grover A, Kahl G (2007) Mining microsatellites in eukaryotic genomes. *Trends in Biotechnology*, **25**, 490–498.

Shen Z, Qu W, Wang W *et al.* (2010) MPprimer: a program for reliable multiplex PCR primer design. *BMC Bioinformatics*, **11**, 143.

Shinde D, Lai Y, Sun F, Arnheim N (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Research*, **31**, 974–980.

Sinama M, Dubut V, Costedoat C *et al.* (2011) Challenges of microsatellite development in Lepidoptera: *Euphydryas aurinia* (Nymphalidae) as a case study. *European Journal of Entomology*, **108**, 261–266.

Sinville R, Soper SA (2007) High resolution DNA separations using microchip electrophoresis. *Journal of Separation Science*, **30**, 1714–1728.

Spathis R, Lum JK (2008) An updated validation of Promega's PowerPlex 16 System: high throughput databasing under reduced PCR volume conditions on Applied Biosystem's 96 capillary 3730xl DNA Analyzer. *Journal of Forensic Sciences*, **53**, 1353–1357.

Squirrell J, Hollingsworth PM, Woodhead M *et al.* (2003) How much effort is required to isolate nuclear microsatellites from plants? *Molecular Ecology*, **12**, 1339–1348.

Subirana JA, Messeguer X (2008) Structural families of genomic microsatellites. *Gene*, **408**, 124–132.

Sun X, Liu Y, Lutterbaugh J *et al.* (2006) Detection of mononucleotide repeat sequence alterations in a large background of normal DNA for screening high-frequency microsatellite instability cancers. *Clinical Cancer Research*, **12**, 454–459.

Sun JX, Mullikin JC, Patterson N, Reich DE (2009) Microsatellites are molecular clocks that support accurate inferences about history. *Molecular Biology and Evolution*, **26**, 1017–1027.

Sutton JT, Robertson BC, Jamieson IG (2011) Dye shift: a neglected source of genotyping error in molecular ecology. *Molecular Ecology Resources*, **11**, 514–520.

Syvänen A-C (2005) Toward genome-wide SNP genotyping. *Nature Genetics*, **37**, 5–10.

Tang J, Baldwin SJ, Jacobs JM *et al.* (2008) Large-scale identification of polymorphic microsatellites using an in silico approach. *BMC Bioinformatics*, **9**, 374.

Tautz D, Schlötterer C (1994) Simple sequences. *Current Opinion in Genetics & Development*, **4**, 832–837.

Thorisson GA, Smith AV, Krishnan L, Stein LD (2005) The international HapMap project web site. *Genome Research*, **15**, 1592–1593.

Toonen RJ, Hughes S (2001) Increased throughput for fragment analysis on an ABI PRISM 377 automated sequencer using a membrane comb and STRand software. *BioTechniques*, **31**, 1320–1324.

Urquhart A, Kimpton CP, Downes TJ, Gill P (1994) Variation in Short Tandem Repeat sequences—a survey of twelve microsatellite loci for use as forensic identification markers. *International Journal of Legal Medicine*, **107**, 13–20.

Väli Ü, 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.

Vallone PM, Butler JM (2004) AutoDimer: a screening tool for primer-dimer and hairpin structures. *BioTechniques*, **37**, 226–231.

Vallone PM, Hill CR, Butler JM (2008) Demonstration of rapid multiplex PCR amplification involving 16 genetic loci. *Forensic Science International Genetics*, **3**, 42–45.

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology*, **23**, 48–55.

Wagner AP, Creel S, Kalinowski ST (2006) Estimating relatedness and relationships using microsatellite loci with null alleles. *Heredity*, **97**, 336–345.

Waits LP, Luikart G, Taberlet P (2001) Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular Ecology*, **10**, 249–256.
Wallin JM, Holt CL, Lazaruk KD, Nguyen TH, Walsh PS (2002) Constructing universal multiplex PCR systems for comparative genotyping. *Journal of Forensic Sciences*, **47**, 52–65.
Wang J (2004) Sibship reconstruction from genetic data with typing errors. *Genetics*, **166**, 1963–1979.
Wang J, Santure AW (2009) Parentage and sibship inference from multilocus genotype data under polygamy. *Genetics*, **181**, 1579–1594.
Weber JL (1990) Informativeness of human (dC-dA)n · (dG-dT)n polymorphisms. *Genomics*, **7**, 524–530.
Webster MT, Smith NGC, Ellegren H (2002) Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 8748–8753.
Weeks DE, Conley YP, Ferrell RE, Mah TS, Gorin MB (2002) A tale of two genotypes: consistency between two high-throughput genotyping centers. *Genome Research*, **12**, 430–435.
Zajac P, Öberg C, Ahmadian A (2009) Analysis of Short Tandem Repeats by parallel DNA threading. *PLoS ONE*, **4**, e7823.
Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. *Molecular Ecology*, **11**, 1–16.

Supporting Information

Additional supporting information may be found in the online version of this article.

Data S1 Endnote library of 100 original journal articles relying on SSRs that had been published recently (in 2009–2010) in the journal *Molecular Ecology*.

Data S2 Endnote library of 15 original journal articles relying on SSR identification using next-generation sequencing techniques, published recently (2009–2010).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Appendix 1 Non-exhaustive list of software for microsatellites detection and genotyping

Software name	Licence	Functionalities	Type of program	Platforms	Reference
SSR detection and primers design					
AutoDimer	Free	Screening for primer-dimer and hairpins	Visual Basic standalone version or Web application	Platform independent	Vallone & Butler (2004)
Generunner	Commercial	Sequence analysis tool	Unknown	Windows	Hastings Software Inc.
MultiPlx	Free	PCR primer compatibility multiplexing	Web application	Linux/Windows/ Solaris	Kaplinski *et al.* (2005)
MSATCOMMANDER	Free	SSR marker detection and design	Python	Platform independent	Faircloth (2008)
NetPrimer	Free	Primer design and secondary structure analysis	Java	Mac/Windows	Premier Biosoft Int.
PolySSR	Free	SSR marker detection	Web application	Platform independent	Tang *et al.* (2008)
Primer3	Free	SSR marker design	Web application	Platform independent	Rozen & Skaletsky (1999)
QDD	Free	SSR marker detection and design	Perl	Linux/Windows	Meglécz *et al.* (2010)
SAT	Free	SSR analysis tool	Web application	Platform independent	Dereeper *et al.* (2007)
STAMP	Free	SSR marker design	Extension to the STADEN package	Platform independent	Kraemer *et al.* (2009)
Multiplexing					
Multiplex Manager	Free	Design and optimization of multiplex PCRs	C++	Linux/Mac/ Windows	Holleley & Geerts (2009)
Estimation of error rates					
MasterBayes	Free	Pedigree reconstruction, analysis and simulation	R package	Mac/Unix/ Windows	Hadfield *et al.* (2006)

Appendix 1 Continued

Software name	Licence	Functionalities	Type of program	Platforms	Reference
Pedant	Free	Estimation of maximum likelihood allelic dropout and false allele error rates	Delphi	Windows	Johnson & Haydon (2007)
PedManager	Free	Inheritance errors and more	Unix	Unix/Windows	Ewen *et al*. (2000)
Fragment calling					
GeneMapper	Commercial	Genotyping software package	Unknown	Windows	Applied Biosystems
GENOTYPER	Commercial	Genotyping software	Unknown	Windows	Applied Biosystems
Peak Scanner	Free	Genotyping software	Unknown	Windows	Applied Biosystems
STRand	Free	Analysis of DNA fragment length polymorphism	C++/Visual Basic	Windows	Toonen & Hughes (2001)
TrueAllele	Commercial	Genotyping software	Matlab	Mac/Unix/ Windows	None
Fragment binning and analysis					
ALLELOBIN	Free	Automated allele binning	C and Java	Unknown	Idury & Cardon (1997)
ALLELOGRAM	Free	Allele binning and normalization	Java	Mac/Unix/ Windows	Morin *et al*. (2009)
Decode-GT	Free	Quality measures for allele calling	Unknown	Mac/Unix/ Windows	Palsson *et al*. (1999)
FLEXIBIN	Free	Automated allele binning	Microsoft Visual Basic	Excel macro	Amos *et al*. (2007)
MsatAllele	Free	Automated allele binning	R package	Mac/Unix/ Windows	Alberto (2009)
MicroMerge	Free	Merging of microsatellite data sets	Unknown	Linux/Windows	Presson *et al*. (2008)
TANDEM	Free	Automated allele binning	Ruby	Mac/Unix/ Windows	Matschiner & Salzburger (2009)
AutoBin	Free	Automated allele binning	Microsoft Visual Basic	Excel macro	See text
Data Management					
GenoDB	Free	Manipulation of dinucleotide SSRs genotype data	Unknown	Unknown	Li *et al*. (2001)
SLIMS	Free	Sample-based LIMS	Web application	Platform independent	van Rossum *et al*. (2010)