Genes, Genomes and Genomics ©2012 Global Science Books

Microsatellite Markers in Plants and Insects Part II: Databases and *in Silico* Tools for Microsatellite Mining and Analyzing Population Genetic Stratification

Tracie M. Jenkins[1*] • Ming Li Wang[2] • Noelle A. Barkley[2]

[1] Department of Entomology, University of Georgia, USA
[2] USDA-ARS, PGRCU, 1109 Experiment Street, Griffin, GA 30223-1797, USA

Corresponding author: * **jenkinst@uga.edu**

ABSTRACT

Nucleotide sequence information available in searchable sequence databases and the free *in silico* software with which to extract and analyze microsatellite data continues to grow at a rapid rate across eukaryote taxa. The sheer amount of information available means that a comprehensive or exhaustive review of databases and free bioinformatic tools lies beyond the purview of any journal review. The purpose of this review is therefore to provide targeted information aimed at helping the insect and plant biologist effectively utilize *in silico* resources to find, navigate and analyze empirically derived data from sequence databases. The objectives are threefold. First, since the basic characteristics of microsatellites make them the markers of choice for studies of genetic structure that underlie adaptation and evolution, these will be delineated. Second, because sequence databases are increasingly mined for microsatellites, the major databases are discussed, as well as, available programs for *in silico* mining of sequence databases to retrieve microsatellites for a species of interest. Lastly, a general review is given of population genetics software for *in silico* genetic analyses of microsatellite data to determine population genetic structure, phylogenetic relationships, and genetic diversity in a species of interest.

Keywords: databases, *in silico* data analysis, microsatellites, phylogeny, population structure

CONTENTS

INTRODUCTION

In 1981, population genetics got an analytical boost when the sequence analysis of the human β globin locus revealed microsatellites (Miesfeld *et al*. 1981; Spritz 1981), which are now known to be powerful DNA markers (Wang *et al*. 2006) and pervasive class of repetitive DNA (Bhargava and Fuentes 2009). Since then this marker (Wang *et al*. 2006) has significantly contributed to research in population genetics (Zhang and Hewitt 2003). They have provided insights into the effects of gene flow on the genetic structuring of populations (Jarne and Lagoda 1996; Chambers and Macaboy 2000; Jenkins *et al.* 2002; Edh *et al*. 2007; Roratto *et al*. 2008) with implications for the evolutionary processes underlying population genetics and conservation biology (Balloux and Lugon-Moulin 2002; Lawson and Zhang 2006). Also, since plants and herbivorous insects are "inexorably intertwined" (Futuyma and Agrawal 2009), microsatellites provide a tool with which to study insect-plant dynamics within an evolutionary and community ecology paradigm.

Microsatellites are non-randomly distributed (Li *et al*. 2002) in coding, non-coding and regulatory regions (Wang *et al*. 1994; Li *et al*. 2002, 2004; Zhang *et al*. 2004; Lawson and Zhang 2006; Hisano *et al*. 2007) in eukaryotic nuclear (Goldstein and Schlotterer 1999; Toth *et al*. 2000; Roy *et al*. 2004; Legendre *et al*. 2009) and organelle (Cato and Richardson 1996; Rajendrakumar *et al*. 2008) genomes. Thus, since they are located in transcribed regions of the genome, e.g. expressed sequence tags (ESTs) and open reading frames (ORF) (Morgante *et al*. 2002), they are well suited for studying gene function, regulation, and recombination phenomena (Biet *et al*. 1999; Lawson and Zhang 2006; Guo *et al* 2009).

High through-put sequencing, the by-product of rapid growth biotechnology (Wang *et al*. 2009) and high through-put, low cost machinery, has directly contributed to the exponential expansion of sequence data in databases. *In silico* (e.g. performed on the computer), software has therefore been developed as an aid for the identification of individual sequences deposited in these databases which contain microsatellite repeats. Once the appropriate desired sequences are identified from *in silico* analysis, microsatellite-targeting and species-specific primers can be designed from websites

Received: 16 November, 2010. Accepted: 4 May, 2011

Invited Review

Table 1 Alphabetical order of all topics and corresponding websites referred to in the body of the paper.

Topic	Website
Allain Andry	http://www.genomicslawreport.com/index.phy/tag/whole-genome-sequencing
Bioinformatics tools	http://softlinks.amnh.org/microsatellites.html ; http://courses.washington.edu/fish543/software.htm
CIB-DDBJ	http://www.cib.nig.ac.jp
CMD	http://www.cottonssr.org
DDBJ	http://www.ddby.nig.ac.jp
DDBJ overview	http://www.www.ddby.nig.ac.jp.introe.html
EMBL	http://www.ebi.ac.uk/embl
EMBL overview	http://www.embl.de/aboutus/general_information/mission/index.html
Eukaryotic Genome-Sequence Initiatives	http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome
GenBank	http://www.ncbi.nlm.nih.gov/genbank
GenBank funding	http://www.lanl.gov/orgs/pa/News/112100.html
GOBASE	http://gobase.bcm.umontreal.ca
IGGP	http://www.vitaceae.org/index.php/International_Grape_Genome_Program
InSatDB	http://www.cdfd.org.in/insatdb
INSDC	http://www.insdc.org; http://www.ddbj.nig.ac.jp/intro-e.html
Mayer C (2006-2010)	http://www.rub.de/spezzoo/cm/cm_phobos.htm
MICROSAT	http://hpgl.stanford.edu/projects/microsat/programs
NAR Database Summary Papers Category List URL	http://www.oxfordjournals.org/nar/database/a/
NCBI	http://www.ncbi.nlm.nih.gov
NCBI text-based *Entrez*	http://www.ncbi.nlm.nih.gov/Database
NCBI eukaryotic genome-sequence initiatives	http://www.ncbi.nlm.hin.gov/sites/entrez?db=genome
Nucleic Acids Research	http://nar.oxfordjournals.org
Paul Evans Library of Fruit Science	http://library.missouristate.edu/paulevans/grapegen.shtml
Phylogenetic analysis, evolution	http://evolution.genetics.washington.edu/phylip/software.html ; http://softlinks.amnh.org/microsatellites.html
Primer3	http://frodo.wi.mit.edu/primer3
RepeatMasker Open-3.0	http://www.repeamasker.org
SilkSatDb	http://www.cdfd.org.in/silksatdb
Small Genomes Microsatellite Database	http://www.genomics.ceh.ac.uk/cgi-bin/sgmd/index.cgi
STRUCTURE	http://www.pritch.bsd.uchicago.Edu.software/structure2_1.html
WHC	http://wheat.pw.usda.gov/ggpages/ssr/WMC

such as Primer3 to amplify polymorphic loci. This synergy among biotechnology, sequence database expansion, and the growth of *in silico* analyses has resulted in techniques for developing microsatellites faster and more cost-effectively (Abdelkrim *et al*. 2009) than conventional methods, which rely on the creation and screening of genomic libraries enriched for repeat motifs (Zane *et al*. 2002, a review). As whole eukaryotic genome-sequence initiatives (**Table 1**, NCBI eukaryotic genome-sequence initiatives), or species specific database "genome projects" have become routine with next generation sequencing and so are *in silico* protocols designed to simply search or mine entire genome databases (Jayashree *et al*. 2006; Sharma *et al*. 2007; Mikheyev *et al*. 2010) in order to retrieve microsatellites or other desired genes/targets.

Microsatellites, whether developed conventionally or through *in silico* methods, are currently relatively easy to generate and evaluate. Polymerase chain reaction (PCR) products are observed as bands on an electrophoresis gel; and, as such are highly adaptable to computer managed high throughput biotechnologies (Schlipalius *et al*. 2001; Wang *et al*. 2009) which generate and organize large datasets necessary to minimize statistical error. Multiple primer sets are often combined or multiplexed into a single reaction (Tang *et al*. 2003; Dzialuk *et al*. 2005) because each microsatellite forward and reverse primer set is locus-specific (Zhang and Hewitt 2003) or specific for a unique conserved genome region (Li *et al*. 2004) 5' and 3' of the repeat motif. This phenomenon further facilitates high throughput (Schlipalius *et al*. 2001; Wang *et al*. 2006, 2009a, 2009b; Raabova *et al*. 2010) computer managed genotyping (Wang *et al*. 2009a) necessary for in depth population analyses. These characteristics among other inherent phenomena make microsatellites applicable to population genetics studies (Arthofer *et al*. 2007).

Microsatellites are generally variable among individuals within and between populations (Goldstein and Pollock 1997). But, allele length appears to be constrained (Schlotterer 1988; Nauta and Weissing 1996; Goldstein and Pollock 1997; Colson and Goldstein 1999) and under the influence of selection (Li *et al*. 2000; Morgante *et al*. 2002) gene conversion, or nonreciprocal recombination (Richard and Pâques 2000). They are also subject to size homoplasy in nuclear (Curtu *et al*. 2004; Barkley *et al*. 2009) and organelle (Hale *et al*. 2004) genomes, i.e. alleles that are the same size but are not homologous or the result of common ancestry (not identical by descent), but arose independently by parallel or convergent mutations in different ancestors. Size homoplasy, if significant, has been shown to affect the interpretation of the phylogeny and population structure (Viard *et al*. 1998) results produced from *in silico* analysis of microsatellite data. This caveat to the use of microsatellites (Curtu *et al*. 2004) means that they should be verified by sequencing alleles prior to being used in population genetic studies to rule out homoplasy (Anmarkrud *et al*. 2008), a topic which will be discussed in more detail in a later review.

This is not intended to be an exhaustive or comprehensive review of databases and free population genetics software. The exponential growth of nucleotide sequence information across eukaryote taxa in the last few years as well as the *in silico* software capable of analyzing it makes this impractical in the extreme. The objectives of this review are therefore to: define the basic characteristics of microsatellites which make them the markers of choice for studies of genetic structure that underlie population structure and adaptation; guide researchers to web databases, which can be used to mine for microsatellites; and, help the researcher locate free population genetics software for *in silico* analyses of genetic structure. This review is also not meant to endorse any web-based site or software package. But, it is meant to be a bridge from which to navigate the vast databases and *in silico* analysis resources available to the insect and plant biologist at the beginning of the 21st century.

MICROSATELLITES CHARACTERISTICS AND DIVERSIFICATION

Microsatellites are often referred to as simple sequence repeats (SSRs) or short tandem repeats (STRs) because they consist of a core motif generally from one to six base pairs (bp) (Zane *et al*. 2002) repeated consecutively in tandem arrays (**Table 2**). They are ubiquitous across eukaryotic genomes (Jarne and Lagoda 1996) to include nuclear introns, exons, and promoters (Reviewed in Goldstein and Pollock 1997; Proven *et al*. 2001; Legendre *et al*. 2009), non-recombinant chloroplast (Edh *et al*. 2007; McGrath *et al*. 2007) and mitochondrial (Estoup *et al*. 1993; Sia *et al*. 2000; Rajendrakumar *et al*. 2008) genomes. Other than being abundant, microsatellites are locus-specific, codominant, inherited in a Mendelian fashion, hypervariable length polymorphism (Ellegren 2004; Grasela and McIntosh 2005; Chambers *et al*. 2007), which can be observed under electrophoresis as a mobility differential (Schlipalius *et al*. 2001).

Due to high mutation rates which are variable from repeat to repeat motif, ranging from 10- to 10,000-fold (10^{-3} and 10^{-6} per cellular generation), they are generally higher than those of non-repetitive regions (Verstrepen *et al*. 2005). These characteristics make them good markers for studying the population structure of agronomically important crop species (Decroocq *et al*. 2003; Wang *et al*. 2006; Barkley *et al*. 2007;Wang *et al*. 2007; Weng *et al*. 2007; Barkely *et al*. 2009; Wang *et al*. 2009a), potential feedstock species (Wang *et al*. 2009b), and insects which impact agricultural resources (Kim and Sappington 2005b; Miao *et al*. 2005) and disease transmission (Oliveira *et al*. 1998; Rongnoparut *et al*. 1999; Archak *et al*. 2006; Pizarro *et al*. 2008).

The exact cause of microsatellite instability continues to be studied, but has yet to be verified and appears to be length-limited (Calabrese *et al*. 2001; Lia and Sun 2003). The elevated mutation rate (Bagshaw *et al*. 2008) of repeat motifs contributes to the high length polymorphism (Calabrese *et al*. 2001), as well as, to the variation across genomic locations in eukaryotes. Microsatellite mutability has been attributed to stepwise mutations (Legendre *et al*. 2009) resulting from DNA polymerase slippage during replication and repair (Tautz and Schlötterer 1994; Lia and Sun 2003). This causes the newly created DNA strand to contain an expanded or contracted section of the repeat array (Kruglyak *et al*. 1998). Unequal crossing over during meiotic recombination has also been attributed to expansion and contraction of repeat motifs (Ellegren 2004; Bhargava and Fuentes 2009), and, therefore, also contributes to taxa diversity.

Microsatellites are found in coding and noncoding regions of eukaryotic genomes; although, they appear more abundantly in noncoding regions (Hancock 1995; Metzgar *et al*. 2000; Parida *et al*. 2009). They are particularly high in areas of eukaryote genomes that serve to maintain and create the centromeric, telomeric, and subtelomeric chromosome regions, the heterochromatin (Palomeque and Lorite 2008). Although functional roles of microsatellites in these areas continue to be studied, they may have a role in segregation of chromosomes (Palomeque and Lorite 2008) and recombination (Petes 2001). The nature of microsatellite tandem array structure makes them prone to insertions or deletions (indels). These indels result in frameshifts, which could affect gene expression and phenotypes (Caburet *et al*. 2005; Stranger *et al*. 2007) in coding and promoter regions of importance, which ultimately may influence adaptation and evolution (Moxon and Wills 1999; Li *et al*. 2004). A microsatellite length polymorphism in a gene sequence would frequently affect gene expression through effects on transcription (Albanése *et al*. 2001; Iglesias *et al*. 2004). For example, single amino acid repeat length polymorphisms or homopolymeric runs, e.g. polyglutamine (encoded as CAG repeat), have been shown to affect protein-protein interactions (Gerber *et al*. 1994; Perutz *et al*. 1994) with transcription factors (Dechering *et al*. 1998; Huntley and Golding 2006). It has also been suggested that microsatellites in coding regions foster rapid genetic responses to environmental pressures that result in phenotypic adaptations and evolution (Rando and Verstrepen 2007). Changes in repeat length within genes are also linked to many neurological, neurodegenerative, and neuromuscular diseases in humans (Pearson *et al*. 2005).

Repetitive motifs do not appear to occur by chance (Li *et al*. 2004; Bagshaw *et al*. 2008). Thus, these tandemly repeated monomeric units may have a role in gene regulation (Contente *et al*. 2002; Bagshaw 2008) and meiotic recombination (Schultes and Szostak 1991; Kirkpatrick *et al*. 1999; Benet *et al*. 2000). Cardle *et al*. (2000) also found that the frequency of microsatellites in plant genomes varied as it did for insects (Archak *et al*. 2007). Microsatellites seem to be no more polymorphic when located near regions with high levels of recombination known as meiotic hot spots (Gerton *et al*. 2000; Petes 2001, a review) than when located in regions with low levels of recombination known as meiotic cold spots (Richard and Dujon 2006). Motif may, however, be significant. A recent study reported that recombination rates increased when motifs consisted of 50% or more of A or T, e.g. AG, TC, CA, TG; but, decreased when motifs consisted of only A and T or G and C, e.g. AT, TA, GC or CG (Guo *et al*. 2009). Thus, these hypervariable repeats seem non-randomly distributed (Zhang *et al*. 2004) and particularly significant in light of the most common repeats in insects and plants.

Microsatellites are abundant in most plant and insect genomes (Thorén *et al*. 1995; Varshney *et al*. 2005; Palomeque and Lorite 2008; Pannebakker *et al*. 2010) as well as transferable cross-species (Smith *et al*. 2005; Ellis and Burke 2007; Augustinos *et al*. 2008) simple by lowering primer annealing temperatures in a PCR protocol (Barbará *et al*. 2007). This transferability, however, is disparately applicable in plants and insects and apparently dependent on genome size (Garner 2002). Barbará *et al* (2007) expanded earlier studies (Primmer *et al*. 1996; Rosetto 2001; Primmer *et al*. 2005) by gathering evidence of cross-species microsatellite transfer from the literature and determined that cross-species transfer does have some limitations. Their data showed that eudicot cross-species transfer was more likely than monocot cross-species transfer; and, vertebrate taxa were 63% more likely to have microsatellite cross-species transfer than invertebrate taxa (311 vertebrates: 114 invertebrates). Microsatellite cross-species transferability, although disproportionately allocated, is still a significant technology. It can be used to illuminate the processes of population stratification leading to speciation as well as study the interactions between and among populations (Noor and Feder 2006) across environments.

Genes coding for proteins are transcribed as mature

Table 2 Selected reviews on microsatellite topics.

Review focus	References
Microsatellites	
Biotechnology	Wang *et al*. 2009
Mutation process	Goldstein and Pollock 1997
EST databases	Jongeneel 2000
Satellite DNA in insects	Palomeque and Lorite 2001
Null alleles	Dakin and Avise 2004
Meiotic recombination	Petes 2001
Strategies for isolation	Zane *et al*. 2002
Population differentiation	Balloux *et al*. 2002
Distribution, function, mutational mechanisms	Li *et al*. 2002
Evolution	Ellegren 2004
EST-SSRs	Ellis and Burke 2007
Features and properties	Mittal and Dubey 2009
***In silico* analyses**	
Software	Labate 2000
Computer programs	Excorrier and Heckel 2002
Bioinofrmatics software	Gilbert 2004
From *in vivo* to *in silico*	Di Ventura *et al*. 2006
Mining for microsatellites	Sharma *et al*. 2007

mRNA or the transcriptome, which represent a small component of a taxon's genome (Jongeneel 2000, a review). Through reverse transcriptase (RT) technology, these coding regions are turned into a complementary DNA (cDNA) library. Single unverified regions called expressed sequence tag (EST) sequences are generally obtained from the 5' and 3' ends of each cloned cDNA. Since EST-derived microsatellites originated from transcribed regions of the genome with greater evolutionary continuity, it was not surprising that they and their microsatellites were conserved and, therefore, transferable across and between taxa (Bouck and Vision 2007). Even though studies have shown EST-derived microsatellite transferability limited beyond genus (Pashley *et al*. 2006), as 'single-pass' cDNA sequence databases have grown so concordantly have EST databases such as the National Center for Biotechnology Information (NCBI) EST database (dbEST; Boguski *et al*. 1993) and other EST databases (Pashley *et al*. 2006). This means that EST databases continues to be a resource from which high quality and transferable microsatellites can be mined and developed for many species of interest (Jongeneel 2000, a review; Palomeque and Lorite 2008).

Intergenic regions of arthropods and vascular plants appear to have an excess of AAC and AAG trinucleotide repeats when compared to introns of the same taxon (Tóth *et al*. 2000). Although AC/TG has been reported for mangrove species (Maguire *et al*. 2000), the most common overall plant motif was AA/TT, followed by AT/TA and CT/GA, respectively (Tóth *et al*. 2000). A database search indicates that these three motifs together composed approximately 75% of all microsatellites with a length of six or more repeats (Lagercrantz *et al*. 1993). AT/TA is generally the most common (Marriage *et al*. 2009), especially with a length of 20 or more bp and AAG/CTT was the most abundant trinucleotide repeats with a length of more than six repeats (Lagercrantz *et al*. 1993; Cardle *et al*. 2000). Cardle *et al*. (2000) found that in plant genomic DNA A/T repeats comprised 32% of repeats, AT/TA 16%, AAG/TTC 14% and AG/TC 10%. When EST databases were mined, Cardle *et al*. (2000) found that of all microsatellites, AAG/TTC was the most common plant repeat, ATC/TAG was the next trinucleotide motif followed by AG/TC and A/T at 29, 17, 20 and 10%, respectively. Thus, when genomic databases are mined AT/TA is the most common plant motif, but when EST plant databases are mined the AG/TC motif is more frequent than the AT/TA motif (Zhang *et al*. 2004; Marriage *et al*. 2009).

The most abundant trinucleotide motif in legumes, wheat, and other crop species was TAA/ATT followed by GAA/CTT, which was consistent with previous studies (Lagercrantz *et al*. 1993; Cardle *et al*. 2000). Mining EST databases recovered TC/AG as being the most prevalent dinucleotide in wheat, rice, maize, and soybean. Considering that the source was limited to coding regions, this is likely a biased sample. The dinucleotide repeats in the mitochondrial genome of rice, however, was made up of about 48% AG/CT repeats (Rajendrakumar *et al*. 2008). Like plants, insects are AT rich (Archak *et al*. 2006; Palomeque and Lorite 2008, a review). AT/TA is a prevalent dinucleotide in arthropods generally with AGC, AAC, AAT being the dominant trinucleotides, respectively (Zane *et al*. 2002, a review). In arthropods generally, dinucleotide repeats are predominant in introns and intergenic regions. AG repeats were found mostly in intergenic regions and AT repeats in introns. GT/CA repeats appear scarce in both plants and insects; whereas, this motif comprises 20% of simple repeats in humans (Lagercrantz *et al*. 1993).

Because microsatellites are codominant, they can distinguish between heterozygotes and homozygotes in populations. This is singularly important because individual progeny and their population genetic structure can be effectively studied using microsatellite markers. Gene flow, e.g. alleles exchanged between or among populations across an adaptive landscape (Wright 1932), is dependent on the variables of ecology and biological history to include resource exploitation and mating system (Balloux and Lugon-Moulin 2002; Lourmas *et al*. 2007). Gene flow is also influenced by selection-dependent genetic structure due to epigenetic or "self-guiding" mechanisms which manipulate phenotypic variability in response to changing selective pressures (Rando and Verstrepen 2007). Spatial and temporal distributions within populations can be affected by reducing gene flow. They could deviate from Hardy-Weinberg proportions because subpopulations may vary in allele frequencies (Yang 1998). This phenomenon can be indirectly measured with large microsatellite datasets generated by present technology in conjunction with calculation of *F*-statistics (fixation indices) from the microsatellite data. [Fixation indices are measures of heterozygosity in individuals (H_I), subpopulations (H_S) and total population (H_T) relative to Hardy-Weinberg expectations and can range in value from 0 (no differentiation) to 1 (complete differentiation): $F_{ST} = (H_T - H_S)/H_T$ – measure of genetic differentiation among subpopulations; $F_{IS} = (H_S - H_I)/H_S$ - measure of genetic inbreeding within a subpopulation; $F_{IT} = (H_T - H_I)/H_T$ – measure of heterozygosity of individuals relative to the total population]. High gene flow between populations increases genetic variability and effective population size, but decreases local adaptation due to population panmixis, i.e. if the migration rate (m) is more than the selection coefficient (s), then selection will have a negligible effect on allele frequency divergence among populations (Stofer 1999). This in turn decreases the effects of genetic drift and increases the phenotypes for selection (Barton and Hewitt 1985; Balloux and Lugon-Moulin 2002).

Insect herbivores and the plants on which they feed display a wealth of diversity in morphology, adaptation, ecology, and genetics due to millions of years of divergence and diversification. Thus, as plant populations continue to diversify and evolve from selection pressures so too will herbivorous insect populations (Futuyma and Agrawal 2009). Microsatellites can illuminate the population fine genetic structure (Schrey *et al*. 2008; Yao and Akimoto 2009) and gene flow (Chaix *et al*. 2003), which affects the heterozygosity, and thereby, the adaptive potential of these populations through genome evolution (Tóth *et al*. 2000). This is why they have been used to study the genetic structure and diversity of agriculturally important insect (Kim *et al*. 2008) and plant (Olsen and Schaal 2001; Li *et al*. 2003; Lia *et al*. 2007; Arakaki *et al*. 2010) populations in order to illuminate possible adaptive responses (Rudmann-Maurer *et al*. 2007) vital in this time of climate change (Hochkirch and Damerau 2009; Horning and Cronn 2009).

Microsatellites have also been used in plant and insect population studies to understand biological phenomena which serve to genetically structure plants (Zhang *et al*. 2010) and insect populations (Carletto *et al*. 2009). They have been used to identify genes under selective pressure as a result of crop domestication (Olsen and Schaal 2001; Vigouroux *et al*. 2002, 2003), to develop linkage maps, particularly in search of resistance genes (Akkaya *et al*. 1995; Cregan *et al*. 1999; Miao *et al*. 2005), to study insects which affect human health (Norris *et al*. 2001; Bataille *et al*. 2009), to develop insect pest control strategies and for studying Mendelian inheritance (Schipalius *et al*. 2001; Miao *et al*. 2005). They are also uniquely suited to study natural and anthropogenic determinants (Rudman-Maurer *et al*. 2007) of population genetic structure (Ballous and Lugon-Moulin 2002; Van't Hof *et al*. 2007; Wang *et al*. 2009a).

Although microsatellite development for certain insect genera, *Lepidoptera* (Nève and Meglécz 2000; Zhang *et al*. 2004) and *Aedes* (Fagerberg *et al*. 2001), has proven difficult because of the nature of the repeat motifs, these markers continue to be the marker of choice (Avise and Ball 1990; Goldstein and Pollock 1997; Fisher *et al*. 2000; Richard and Thorpe 2001; Symonds and Lloyd 2003; Holzer *et al*. 2006; Weng *et al*. 2007; Ross and Shoemaker 2008) for insects (Harr *et al*. 2000; Temu *et al*. 2004; Ross *et al*. 2008), "the most diverse group of organisms on Earth"

(Carletto *et al*. 2009), and plants on which all animal life depends (Powell *et al*. 1996; Oliveira *et al*. 1998; Roder *et al*. 1998; Rongnoparut *et al*. 1999; Temnykh *et al*. 2001; Chaix *et al*. 2003; Ritschel *et al*. 2004; Li *et al*. 2005; Miao *et al*. 2005; Varshney *et al*. 2005; You *et al*. 2005; Debout *et al*. 2007; Edh *et al*. 2007; Zheng *et al*. 2007; Lia *et al*. 2007; Wang *et al*. 2007; Exeler *et al*. 2008; Kobayashi 2008; Pérez de Rosas *et al*. 2008; Pizarro *et al*. 2008; Ross and Shoemaker 2008; Schrey *et al*. 2008; Takahashi *et al*. 2008; Carletto *et al*. 2009; Ijaz and Khan 2009; Yao and Akimoto 2009).

Microsatellites are often transferable, e.g. can be amplified across taxa (Barbará *et al*. 2007; Wang *et al*. 2009a), especially in insect (Huttunen and Schotterer 2002; Kim and Sappington 2005a; Smith *et al*. 2005) and plant (Rossetto 2001; Varshney *et al*. 2005; Wang *et al*. 2005; Barkley *et al*. 2007; Wang *et al*. 2007; Gong *et al*. 2008; Tang *et al*. 2010) species. Thus "*in silico* mining" (reviewed in Sharma *et al*. 2007), of genome-sequence or microsatellite databases (Aishwarya and Sharma 2008) is cost effective in terms of time, money, and ease of application. Bioinformatics provides an efficient way to facilitate microsatellite marker discovery which, because of the marker's inherent characteristics (Wang *et al*. 2009a), can be applied to population genetic studies, i.e. research on genetic diversity, gene flow (Noor and Feder 2006; Barbará *et al*. 2007) phylogenetics (Arévalo *et al*. 2004; Ochieng *et al*. 2007; Yao *et al*. 2008), and population structure (Barkley *et al*. 2006) among closely related, sympatric, or fragmented populations (Nishimura *et al*. 2005; Barbará *et al*. 2007).

SEQUENCE DATABASES: A VALUABLE RESOURCE FOR MICROSATELLITE *IN SILICO* MINING

Insect and plant biologist are increasingly realizing that successful research is inextricably linked to the web and *in silico* analyses (Jenkins *et al*. 2007, 2009). Scientists have witnessed an explosion of DNA sequence data during the last decade (Refer to Higgs and Attwood 2007, pp 81-88), which has facilitated the development of free software for *in silico* analysis of high through-put microsatellite data. A plethora of internet databases, including genome sequences, are now regularly mined for microsatellite markers using robust, user-friendly, open access programs and databases (Prasad *et al*. 2005; Blenda *et al*. 2006; Archak *et al*. 2007; Yasodha *et al*. 2008) which are experiencing explosive growth (Thiel *et al*. 2003; Aishwarya and Sharma 2007; Wang *et al*. 2009a). Genomic and EST databases (Varshney *et al*. 2002; Crane 2007) with their enormous amounts of sequence information from which SSRs can be mined (Sreenu *et al*. 2003; Prasad *et al*. 2005; Aishwarya *et al*. 2007; Kim *et al*. 2008; McWilliam *et al*. 2009) coupled with computer programs capable of analyzing large repetitive datasets continue to sustain the popularity of this highly informative and versatile codominant marker (Schlötterer 1998) by "bridging the gap between a large body of experimental data and useful mathematical models" (Ventura *et al*. 2006). Free web-based computational or bioinformatic tools for *in silico* analyses are also expanding (Excoffier and Heckel 2006). These resources enable the researcher to design, identify, generate, or analyze large microsatellite datasets (Aishwarya *et al*. 2007) for spatial and temporal insights into population structure and genetic diversity within the framework of population genetic and evolutionary theory (Gilbert 2004; Johnson and Haydon 2007a, 2007b).

The International Nucleotide Sequence Database Collaboration (INSDC) (**Table 1**, INSDC) includes the three primary sequence databases in common use today which partner so closely that all new and updated database entries are exchanged among the groups on a daily basis (Higgs and Attwood 2007, p. 82) (**Fig. 1**). These INSDC collaborative databases include the DNA data Bank of Japan (DDBJ), the European Nucleotide Sequence Database (EMBL), and GenBank in the United States (**Fig. 1**).

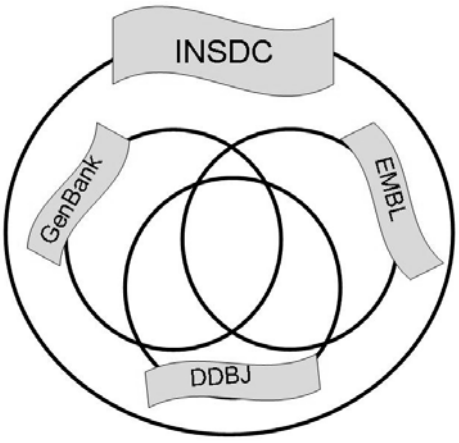

Fig. 1 Cartoon of INSCS organization showing the collaboration among DDBJ, EMBL and GenBank databases (refer to text for explanation).

DDBJ (**Table 1**, DDBJ); Sugawara *et al*. 2008) was established to meet the needs of Japanese researchers. DDBJ began collaborating with GenBank and EMBL in 1982 (Stoesser *et al*. 1998, 2003). It operates from the Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ) in Mishima, Japan, which began operations in 1995 (**Table 1**, CIB-DDBJ). DDBJ is the only nucleotide sequence database for Asia that is certified to collect DNA sequence and then issue accession numbers acceptable to INSDC and the international community (**Table 1**, DDBJ overview).

EMBL (**Table 1**, EMBL) is an inter-governmental organization with 20 member states and one associate state. There are five laboratories. The first and main laboratory was initiated in 1978 and is located in Heidelberg, Germany. The other four laboratories are in Hinxton, UK (the European Bioinformatics Institute, EBI), Grenoble, France, Hamburg, Germany, and Monterotondo, Italy. The overall purpose of EMBL is "to promote molecular biology across Europe and to provide an attractive alternative to the United States as a workplace for Europe's leading young molecular biologists". It is part of the European Nucleotide Archive (ENA) (Cochrane *et al*. 2008), which was established in 1980 and is maintained by the European Bioinformatics Institute (EMI) (Stoesser *et al*. 1998, 2003). The umbrella collaborative group, INSDC, is governed by an advisory committee of nine members, three from the US, three from Europe and three from Japan and constitutes the major repository for sequences generated in laboratories across the world (**Fig. 1**).

Historically, GenBank began with Walter Goad and others of the Theoretical Biology and Biophysics Group at Los Alamos National Laboratory in Los Alamos, NM, USA. They established the Los Alamos Sequence Database in 1979; and, in 1982 this database morphed into GenBank, two years after the EMBL Data Library was established (Stoesser *et al*. 2003). In 1982, GenBank, DDBL, and EMBL (**Table 1**) began to share data and have been doing so efficiently and consistently for 25 years. This overlap provides all sequence information to users in a single source greatly facilitating the ease to mine data. Funding for GenBank comes from the National Institutes of Health (NIH), the National Science Foundation (NSF), the Department of Energy (DOE), and the Department of Defense (DOD) (**Table 1**, GenBank funding). GenBank is accessible through the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, Maryland 20894, USA (**Table 1**, NCBI)). End-users of GenBank can easily search and retrieve data from the multiple databases including microsatellite and nucleotide sequence databases with NCBI's text-based *Entrez* system (**Table 1**, NCBI text-based *Entrez*). By the end of 1983 there were more than 2,000

Table 3 A list of frequently used software programs which can be utilized for microsatellite identification from sequence data along with their respective website.

Name	Website	Reference
CID	http://www.shrimp.ufscar.br/cid/index.php	Freitas *et al*. 2008
Microsatellite Repeats Finder	http://biophp.org/minitools/microsatellite_repeats_finder/	Benson 1999
MISA	http://pgrc.ipk-gatersleben.de/misa/	
MRepatt	http://alggen.lsi.upc.es/cgi-bin/search/mrepatt/mrepatt.pl	
MSATCOMMANDER	http://code.google.com/p/msatcommander/	Faircloth 2008
Phobos	http://www.ruhr-uni-bochum.de/spezzoo/cm/cm_phobos.htm	Mayer 2006-2010
Poly	http://www.bioinformatics.org/poly/wiki/	Bizzaro and Marx 2003
QDD	http://gsite.univ-provence.fr/gsite/Local/egee/dir/meglecz/QDD.html	Meglécz *et al*. 2010
RepeatMasker	http://www.repeatmasker.org	Smit *et al*. 1996-2004
Sputnik	http://espressosoftware.com/sputnik/index.html	
SSR Finder	http://www.maizemap.org/bioinformatics/SSRFINDER/SSR_Finder_Download.html	
SSRIT	http://acorn.cshl.org/db/searches/ssrtool	
SSR Locator	http://minerva.ufpel.edu.br/~lmaia.faem/	da Maia *et al*. 2008
Tandem Repeats Finder	http://tandem.bu.edu/trf/trf.html	Benson 1999
TROLL	http://finder.sourceforge.net/	Castelo *et al*. 2002
WebSat	http://wsmartins.net/websat/	Martins *et al*. 2009

sequences in GenBank. It presently has over 61 million sequences representing over 100,000 unique taxa. GenBank has doubled in size approximately every 18 months (Benson *et al*. 2007) a phenomenon largely due to direct submissions by individual scientist as well as from sequencing project collaborations such as the International Anopheles Genome Project, the Mouse Genome Sequencing Consortium (MGSC) (Stoesser *et al*. 2003) the Soybean Genome Database (SoyGD) (Archak *et al*. 2007) among others.

Besides the INSDC databases, which continue to grow exponentially (Stoesser *et al*. 2003), open access (OA) journals published online are excellent sources for database information. One of the most accessible and up-to-date OA journals is *Nucleic Acids Research* (NAR). It has published an Annual Database Issue since 1993 (Galperin and Cochrane 2009). The objective of the NAR Annual Database Issue is to keep researchers abreast of new and enduring database collections. The 2010 issue and its accompanying complete database list and summaries are available online at the *Nucleic Acids Research* web site (**Table 1**, *Nucleic Acids Research*) and at the 2010 NAR Database Summary Papers Category List URL (**Table 1**, NAR Database Summary Papers Category List URL), which includes immediate access to GenBank, DDBJ, and EMBL. When NAR's annual Database issue was first published in 1993 it included 24 database papers (Galperin and Cochrane 2009). The 2009 NAR Database issue had 179 papers which described 95 new databases. The NAR online Molecular Biology Database Collection now has 1,170 publically available databases representing international contributions (Galperin and Cochrane 2009). All sequences are deposited into GenBank for all to access and mine.

The international scientific community has cooperated on open access database creation, maintenance, and usage to the extent that these database collections are experiencing exorbitant growth. Due to the growth in sequences deposited to the three major databases (GenBank, EMBL, and DDBJ) there was a need in the community to develop taxa specific databases to manage sequence information. Some of these new and expanding databases include: the internationally supported microsatellite databases such as the Small Genomes Microsatellite Database (includes organelles) (**Table 1**, Small Genomes Microsatellite Database)), insects e.g., insect database (InSatDb) (Archak *et al*. 2006) (**Table 1**, InSatDb) and silkmoth microsatellite database (SilkSatDb) (**Table 1**, SilkSatDb), and plants e.g., wheat (WHC) (**Table 1**, WHC), Cotton (CMD) (**Table 1**, CMD), the International Grape Genome Program (IGGP) (**Table 1**, IGGP) formed by the genetics international research community with projects funded by the National Science Foundation (NSF) in the US and listed among other *Vitis* spp. databases on the Paul Evans Library of Fruit Science website, which is maintained by the Missouri State University Libraries (**Table 1**, Paul Evans Library of Fruit Science). This is not an exhaustive list of all the taxa specific databases that exist on the web, but a sample of some specific to plant or insect species.

Insights into population structure are often gleaned from organelle data. Thus, organelle databases can also be mined for possible microsatellites or SSRs. An example of such a database is GOBASE (**Table 1**, GOBASE) (O'Brien *et al*. 2006), which is now in its 21st release (O'Brien *et al*. 2009). It contains all published mitochondrion-encoded sequences and chloroplast-encoded sequences from a wide collection of eukaryote taxa. GOBASE has over 910,000 mitochondrial sequences and over 250,000 chloroplast sequences representing 737,000 and 174,000 genes, respectively (O'Brien *et al*. 2009). These data were culled mostly from GenBank releases. Furthermore, GOBASE has recently added three reference bacteria genomes (O'Brien *et al*. 2006) from which SSRs can be mined: the gamma-proteobacterium, *Escherichia coli* K12, the alpha-proteobacterium, *Rickettsia prowazekii* and *Nostoc* sp. A truncated gamma-proteobacterium is part and parcel of insect mutualistic symbioses widely found among hymenopteran, hemipteran and other insect orders (O'Neill *et al*. 1992; Thao *et al*. 2000; reviewed in Gil Latorre and Moya 2004; Moran *et al*. 2005; Kikuchi *et al*. 2007; Allen *et al*. 2009). *Rickettsia prowazekii* was likely an ancestor of mitochondria and *Nostoc* sp. appears to be the ancestor of chloroplasts. Presently GOBASE and scientists at NCBI are making the GOBASE content an "auxiliary to GenBank" (O'Brien *et al*. 2009) and, therefore, available to INSDC collaborators to search for SSRs.

The first human nuclear genome was sequenced in 2003 after 12 years of work and cost over $3 billion. Today the cost to sequence an individual's nuclear genome is approaching $1000.00, according to an article posted by Allain Andry (**Table 1**, Allain Andry). Because the price of whole genome sequencing continues to become cheaper with next generation sequencing technology, genetic tests for specific genes linked to cancer, other diseases and life issues are being developed. We hear of a new era of "personalized medicine" in which drugs and therapies will be prescribed / targeted based on an individual patient's specific genes or whole genome.

Because the price of whole genome sequencing continues to become cheaper with next generation sequencing technology, the explosion of whole genomic data will continue to rise which will aid in the ease of development of microsatellites for plants and insects. Many of the major crops or model crop species such as *Arabidopsis* have their whole genome publicly available to the entire scientific community. Access to whole genome data is advantageous to researchers working with these taxa in that *in silico* programs can be easily employed to mine sequence infor-

Table 4 A list of some of the common programs for use in *in silico* analyses with microsatellite data. (This is not a comprehensive list of all programs available and all of their functions for analyses. Only functions relative to microsatellite analyses are listed in the table see associated websites to review full list of functions for each program.). Refer to **Table 1**, under Topic Bioinformatics, website for softlinks for a more exhaustive list.

Program	Free-ware	Reference	Website	Analyses
Arelequin	Yes	Excoffier and Lischer 2010	http://cmpg.unibe.ch/software/arlequin35/	AMOVA, F-statistics, Gene diversity, LD, Hardy Weinberg, allele frequencies, expected heterozygosity, population differentiation, Garza-Williamson index, Mantel test, genetic distance
GENEPOP	Yes	Rousset 2008	http://genepop.curtin.edu.au/	Hardy Weinberg, Linkage Disequilibrium, F statistics, Nm, and gene diversity
LAMARC	Yes	Kuhner 2006	http://evolution.genetics.washington.edu/lamarc/index.html	Population size, population growth rate, Ne, recombination and migration rates
Micro-Checker	Yes		http://www.microchecker.hull.ac.uk/	Checks for null alleles and scoring errors in microsatellite data
Migrate-N	Yes	Beerli and Palczewski 2010	http://popgen.sc.fsu.edu/Migrate/Migrate-n.html	N_E, migration rates between populations, Likelihood-ratio tests, Bayesian inference or Maximum Likelihood inference
NTSysPC	No		http://www.exetersoftware.com/cat/ntsyspc/ntsyspc.html	PCA, genetic distance, Neighbor-Joining, UPGMA
PAUP	No	Swofford 1998	http://paup.csit.fsu.edu/	NJ, UPGMA tree construction, bootstrapping
PHYLIP	Yes	Felsenstein 2005	http://evolution.genetics.washington.edu/phylip.html	NJ, UPGMA tree construction, bootstrapping
POPTREE2	Yes	Takezaki *et al*. 2010	http://www.med.kagawa-u.ac.jp/~genomelb/takezaki/poptree2/index.html	Calculates genetic distances, heterozygosity, number of alleles, and Gst, NJ or UPGMA tree constructing, bootstrapping
Power Marker	Yes	Liu and Muse 2005	http://statgen.ncsu.edu/powermarker/index.html	Calculates heterozygosity, PIC, Hardy-Weinberg equilibrium, gene diversity, linkage disequilibrium, F statistics, bootstrapping, NJ and UPGMA tress
RSTCALC	Yes	Goodman 1997	http://www.biology.ed.ac.uk/research/institutes/evolution/software/rst/rst.html	Rst-genetic differentiation, Nm-number of migrants, genetic distance calculation, bootstrapping
STRUCTURE	Yes	Pritchard *et al*. 2000	http://pritch.bsd.uchicago.edu/structure.html	Infers population structure, assigns individuals to populations using a Bayesian approach, estimates allele frequency in each population

mation for various microsatellites motifs (**Table 3**) as well as be used for numerous other research applications. However, there are still many minor plant and insect species that have limited sequence or no sequence data available in publicly available databases. These minor taxa may only have a few hundred deposited sequences or less, which makes mining sequence data for microsatellites impractical. This can be a serious limitation of working with minor taxa with three main ways to overcome this hurdle. The first method is to develop a microsatellite enriched library (Takundwa *et al*. 2010) for the taxa of interest and utilize the library to reveal microsatellite repeats. The second method, which relies on transportability between species, is to use microsatellites developed in closely related species (if available), on the species of interest (Wang *et al*. 2009a). The third method is to have the taxa of interest sequenced or wait until it is sequenced by other researchers and deposited into a public database.

GENETIC SOFTWARE FOR *IN SILICO* ANALYSES

According to a review by Zhang and Hewitt (2003), "microsatellite sequences are the most revealing DNA markers available so far for inferring population structure and dynamics." This suggests that the field of population genetics has benefited enormously from the synergy between population genetic theory and the development and application of DNA multi-locus marker technology. The insect and plant scientist studying genetic structure must understand the genetics of the microsatellites they use, as well as, how to generate large datasets with the markers using the latest biotechnologies (Wang *et al* 2009a). Further, they must also know how to download and use free *in silico* software on the internet for analyzing microsatellites (Benson 1999) so that datasets can be collected and analyzed in a timely manner (Kim *et al*. 2008; Rajendrakumar *et al*. 2008).

The massive internet resources for microsatellite discovery or data analyses has therefore become indispensable to the insect and plant biologist of the 21st century who desires to generate large datasets in order to perform phylogenetic analysis or evaluate historical and spatial population genetic structure from the stand point of its deviation from the Hardy-Weinberg equilibrium model. High throughput sequencing machines have made the generation of large microsatellite datasets routine (Wang *et al*. 2009a) and, therefore, insights into fine population structure accessible. Large, genome-based microsatellite datasets (Lovin *et al*. 2009) can also be effectively collected and analyzed in a timely manner through *in silico* analyses using computer freeware (**Table 4**). Just as the number and content of online databases have exploded so have online open access to freely downloadable bioinformatics tools for mining and analyzing these datasets (reviewed in Labate 2000 and Excoffier and Heckel 2006) (**Tables 1, 3** and **4**, Bioinformatics tools).

Since large datasets can be analyzed *in silico* from plastid and nuclear genomes in population-specific plant species, inter- and intraspecific genetic diversity as well as insect and other mediated seed and pollen gene flow for important agricultural crops, horticultural species and forage and turf grass species have been illuminated (Chaix *et al*. 2003; You *et al*. 2005; Edh *et al*. 2007; McGrath *et al*. 2007). Population genetic theory and *in silico* analyses of microsatellite data from insect vectors of disease (Lovin *et al*. 2009) have revealed that gene flow in these insect populations has been more extensive than previously thought and may only be limited by geographic barriers. If microsatellites are uniquely suited to population genetics studies, and they appear to be, then *in silico* population estimates are uniquely suited to these large microsatellite datasets.

The speed with which data generation and analysis can now be done also necessitates quicker communication of results to the scientific community. This need has given rise to open access journals such as the Public Library of Sciences (PLOS) which, because of their fast turn-around from submission to publishing, are increasingly being supported by the scientific community (Gitschier 2009). Furthermore, there are now journals, such as *In Silico Biology*, dedicated to research articles on the latest *in silico* analyses, modeling, and simulations, which help scientists' keep abreast of the

latest in data analyses. Population genetic analyses benefits from the integration of large, biologically based, empirical datasets with theoretical population models through this *in silico* or computer science phenomenon (Ventura *et al*. 2006) for two reasons. First, larger datasets provide more insights into the selection pressures and stochastic dynamics within and between populations. Secondly, population genetics is highly theoretical and model driven. Thus, *in silico* analyses provides a computer platform for the integration of knowledge building from empirically generated observations and theoretical population genetics concepts (Ventura *et al*. 2006).

Genetic structure *in silico* analyses

Populations of unknown genetic structure, e.g. cryptically stratified or structured due to differences in ancestry are challenging to elucidate for most genetic studies (Marchini *et al*. 2004; Alexander *et al*. 2009). The inherent difficulty lies in developing adequate mathematical models that can sufficiently replicate the true complexity of nature. Many *in silco* methods designed to identify how populations are genetically structured and infer gene flow by estimating the ancestry of individuals within subpopulations are now being used. Two basic approaches exist for evaluating population structure, global ancestry estimates which is model based and algorithmic ancestry estimates which are algorithmic based (Alexander *et al*. 2009). *In silico* stratification analyses of multilocus genotypes sampled from a population with unknown structure are capable of estimating not only the ancestry of individuals, but also identify the subpopulations to which they belong. While not advocating for one program over another we will focus on STRUCTURE (Pritchard *et al.* 2000) since, with over 4,800 citations (ISI Web of Knowledge) for its first version (Pritchard *et al*. 2000), it is the most widely used clustering software at the time of this writing (Kaeuffer *et al*. 2007).

STRUCTURE (Pritchard *et al*. 2000) identifies the presence of subpopulations by using a model-based clustering algorithm (Bayesian inference) on multilocus genotypes. It generates clusters or groups based on transient Hardy-Weinberg disequilibrium (HWD) and linkage disequilibrium (LD) resulting from admixture between populations (Kaeuffer *et al*. 2007). Therefore, it assumes that loci fit Hardy-Weinberg equilibrium and linkage equilibrium within a subpopulation. The enhanced version of STRUCTURE (STRUCTURE 2.1; Falush *et al*. 2003) improved clustering results by combining the admixture model (Pritchard and Wen 2004) with user defined map distances between markers (Kaeuffer *et al*. 2007). Overestimates of clusters, however, can occur due to 'strong' linkage disequilibrium or deviations from Hardy-Weinberg equilibrium (Falush *et al*. 2003) and using different combinations of loci can affect the number of inferred subpopulations (Kaeuffer *et al*. 2007).

STRUCTURE analysis can reveal the presence (or absence) of population structure, examine hybrids, identify migrants, and evaluate admixed individuals based on allele frequencies all of which can help elucidate population variation and ancestry in plants and insects. STRUCTURE probabilistically assigns individuals to populations or multiple populations if their genotype indicates admixture (Pritchard *et al.* 2000). One of the main advantages of this program is that it does not assume a specific mutation model, and thus, can be utilized for most linked and unlinked genetic markers including microsatellites. Further, STRUCTURE can identify subpopulations (or clusters) with or without predetermined user-defined population information (**Fig. 2**). This program does allow for missing data points which can often occur when collecting large microsatellite data sets (large sets of markers and/or taxa) or when employing microsatellite markers on divergent genera/species.

The program STRUCTURE has been extensively used since its inception to evaluate population structure of plants. A few examples of its use in plants are discussed. STRUCTURE analysis was employed in a large study of 260 maize inbred lines which were assayed for variation from 94 microsatellite loci. This analysis identified five populations that corresponded to the major breeding groups and identified some lines displaying mixed origin (Liu *et al.* 2003). In a separate maize study, STRUCTURE analysis was employed to evaluate gene flow and genetic contribution from teosinte to Mexican maize. This report demonstrated that Mexican maize at higher elevations had a modest contribution of gene flow from teosinte; whereas, maize in lower elevations had less of a genetic contribution from teosinte (Matsuoka *et al.* 2002). Kwak and Gepts (2009) used microsatellite loci to examine the major gene pools of common bean (*Phaseolus vulgaris*). The STRUCTURE results demonstrated that the Mesoamerican gene pool had higher proportions of non-hybrid accessions compared to the Andean gene pool. Furthermore, the population structure was consistent with ecogeographic racial structure within gene pools and a subdivision between Mesoamerican and Andean gene pools (Kwak and Gepts 2009). Population structure was also used to evaluate the US sorghum germplasm collection. Population analysis identified four subgroups from the genotypes of 96 accessions and partitions among groups were well correlated with geographic locations that these accessions either were collected or originated (Wang *et al*. 2009b).

In Barkley *et al*. (2006), STRUCTURE analysis elucidated hybrid origin in multiple citrus accessions which had limited passport data. Taxonomy of *Citrus* has long been debated since many cultivated species are derived from natural hybridization of the ancestral forms. This analysis identified five clusters in a population of 370 citrus accessions. The five populations represented all of the ancestral species and citrus relatives, while the remaining species were hybrids among the naturally occurring forms (**Fig. 2**). Furthermore, this study demonstrated that some accessions previously believed to be non-hybrids were actually hybrids or hybrid derivatives. It also confirmed the ancestry of known hybrids. In addition, the structure data provided an alternative approach to evaluate varieties with questionable passport data, which ultimately led to better understanding of these accessions origin and improved management of the germplasm (Barkley *et al.* 2006).

As population genetics continues to evolve into a "data-driven discipline" (Pool *et al*. 2010) genome-wide multilocus datasets continue to push the analytical limits of *in silico* analyses (Price *et al*. 2006; Davidson et al. 2009; Pool *et al*. 2010). STRUCTURE is time consuming on large datasets since the user has to specify different values of K (number of populations) and allow the program to run until the data has converged. EIGENSTRAT (Price *et al*. 2006), and other similar methodologies (Zhang *et al*. 2009), employs an algorithm-based ancestry estimates which uses principal component analysis (PCA) to model population structure or stratification. Cited 666 times (ISI Web of Knowledge) it is a fast and efficient way to analyze genotypes on a genome-wide scale (Price *et al*. 2006). Because PCA is less parametric than model based ancestry estimates, it can provide low-dimensional projections of the data that describe the aggregate variation among genotypes (Alexander *et al.* 2009). Furthermore, PCA is considerably faster on large scale data sets compared to model based ancestry estimates, it provides an initial test for the presence of population structure in a data set, and PCA does not purport to force individuals into distinct subpopulations where stratification may or may not actually exist (Patterson *et al*. 2006). An advantage PCA can have for a researcher studying population stratification is that the results from algorithm based analyses can provide a default for the number of clusters (K) to infer in the program STRUCTURE (Patterson *et al*. 2006). This would reduce the total time required to run the program since the number of iterations to convergence in model based ancestry estimates increase as the number of subpopulations set by the user increases and/or the value of K chosen poorly supports the data (Ale-

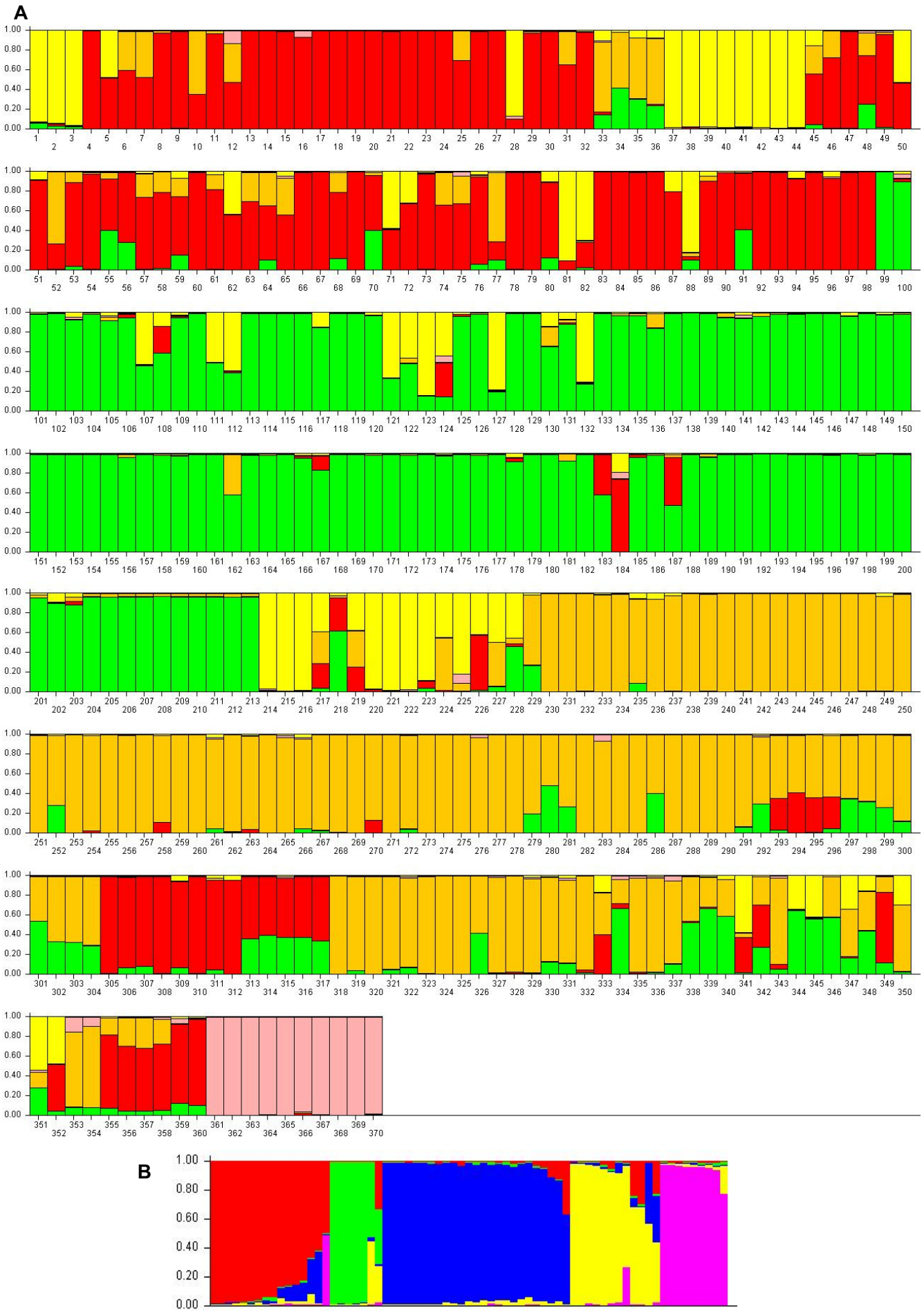

Fig. 2 (**A**) Structure data from a population of 370 *Citrus* and two related genera showing individual gene history (Barkley *et al*. 2006). (**B**) Bar plot of the genetic composition of subterranean termite subpopulation inbred genotypes represented by five colors from a population representing a single introduction of 10 years longevity and generated by STRUCTURE 2.0 using the admixture model (Jenkins, unpublished data).

xander *et al*. 2009). Overall, if researchers employ model based and algorithm based ancestry estimates to a data set, population structure or the lack of population structure should be evident in either analyses (Patterson *et al*. 2006) and this overlap in results from analyzing data utilizing two distinct models leads to enhanced support of "true" stratification in a data set.

Phylogenetics: Genealogy and relationships in plants and insects

Phylogenetics, which can be defined only in an evolutionary context, as the relatedness of a species or genera relative to a common ancestor has been used in all branches of biology and several related fields. These diagrams (trees) illustrate genealogical descent and mutational change throughout time among lineages. Phylogenies display historical relationships, not similarities; although, closely related taxa tend to be similar because of a recent shared common ancestor (Baum *et al.* 2005). Applications involving phylogenies include studies in eukaryotes and prokaryotes alike from studying the relationships between insects and disease, investigating the origin of human populations, revealing insights of the origin of land plants, evaluating plant and animal interactions, identifying relationships between pathogens, to determining if a dentist infected patients with HIV (Armbruster 1992; Leitner *et al.* 1996; Huelsenbeck and Rannala 1997; Soltis and Soltis 2000; Bataille *et al*. 2009; Llewellyn *et al*. 2009) along with numerous other applications. Phylogenies also can be employed to address ecological issues such as evaluating biodiversity and conservation priorities given that close relatives of a species going extinct are likely to be a high risk for extinction (Mace *et al.* 2003). Generally speaking, phylogenetic reconstruction has been a dominant force in the last 25-30 years of research that has made a huge impact in all facets of biology and related fields. A list of programs available for phylogenetic analysis of microsatellite data is included in **Table 4** and more information can also be found in **Table 2** (phylogenetic analysis, evolution) and/or at the specific website for each respective program. It is beyond the scope of this review to detail steps for utilization of the various programs mentioned in the text; but, we aim to guide users to alternative programs for data analysis and demonstrate some of the potential uses in plants and insects).

Even though phylogenetic inference is a fairly common practice, development of trees representing the relationships among animals, plants, fungi, protists, as well as the numerous undescribed species are still poorly understood or entirely unknown (Mayr 2001) since so many species still have limited genetic information collected. However, the explosion of genomic data in the past 15-20 years and identification of new species are helping to fill in these unknown gaps. Reconstructing relationships relies on mathematical methods to infer the past from features of contemporary species with the fossil record as support to these inferences (Delsuc *et al.* 2005). Molecular phylogenetics, most commonly used today, utilizes differences in molecules, such as DNA or microsatellite markers, to supply information on the relationships among taxa and provides the basis for the majority of trees developed. Phylogenies can be created for either extant or extinct individuals as long as molecular data can be directly collected or indirectly obtained (i.e., GenBank) from the individuals being analyzed. Overall, one of the most important impacts of phylogenetics is the ability to clarify intra- and interspecific relationships among organisms and understand their evolutionary history.

Plants display a wealth of diversity in morphology, adaptation, ecology, and genetic composition, due to millions of years of divergence and diversification, which should be characterized to understand the mechanisms through which this diversity arises (Schaal *et al.* 1998). Microsatellite derived phylogenies have been extensively used to evaluate genetic diversity and inter- and intraspecific relationships in many plant and insect species. Phylogenetic relationships are determined based on a calculated genetic distance (sequence conservation or diversification) in their evolutionary history and reflect the relatedness of a group of individuals. Therefore, everything makes more sense in the light of a phylogeny (Soltis and Soltis 2000). This tool has been successful to help clarify relationships, distinguish individuals, reclassify samples, evaluate population structure, examine the geographic distribution of a species (phylogeography), and support or revise current taxonomic classification. A few selected examples of phylogenies for plants and insects in the current literature and their impact will be discussed.

Phylogeography can be defined as the concurrent evaluation of geographic distributions and phylogenetic relationships of a set of individuals. This process allows the evaluation of a species' evolutionary history over space and time (Templeton 2004). Further, phylogeography can be employed to assess genetic exchange among populations and distinguish genetic variation caused by gene flow from variation derived from common ancestry (Schaal *et al.* 1998). One potential drawback is that there is no means to establish if enough individuals or geographical sites have been sampled to ensure that the pattern did not arise by chance alone (Templeton 2004). However, this approach has been used to successfully evaluate gene flow and population structure in plants and insects. For example, six microsatellite loci were used to examine the phylogeography of lodgepole pine, which demonstrated geographical clustering of the recent clades, further supporting the hypothesis of rapid expansion of pines followed by local population differentiation (Marshall *et al.* 2002). Microsatellites were also used to assess geographic origin and relatedness in *Arabidopsis thaliana*, which showed general congruence with a few exceptions. Some of the exceptions could be due their reticulate evolutionary history (Symonds and Lloyd 2003). An endangered tree in China, *Fraxinus mandshurica*, was evaluated with microsatellites for genetic diversity and spatial structure, which illustrated that intra-population diversity significantly decreased with latitude and no clear geographic genetic structure was identified (Hu *et al.* 2008). Unfortunately, phylogeographical studies do not always show a clear relationship between similar phylogenetic relationships with similar geographic origin. The two common reasons this occurs is that the genealogy is incorrect and more markers are required or the phylogenetic relationships are correct, thus, an alternative explanation must be sought (Symonds and Lloyd 2003).

Molecular marker research studies have helped to clarify, validate, or change the current dogma of assumed relationships among plants. An extensive study of potato (*Solanum* spp.) landraces and wild progenitors with microsatellite markers established a need for reclassifying cultivated potatoes into four species. This study suggested that ploidy which was important traditionally as an indicator to categorize potato accessions was a poor character to employ to classify individuals (Spooner *et al.* 2007). Microsatellites from organelle genomes (mitochondria and chloroplast) were utilized to evaluate the phylogenetic relationships of rice, sorghum, maize, and wheat (Rajendrakumar *et al.* 2008). This analysis verified that rice and sorghum were closely related (phylogenetically), while wheat was more distant to rice and sorghum, which helped validate the synteny between these related grass genomes (Rajendrakumar *et al.* 2008).

Interspecific genetic diversity and crop domestication analyses are also common results of microsatellite derived phylogenies. Rice, which is a staple crop around the world, showed ample genetic diversity with an average polymorphism information content (PIC value) of 0.707 and a clear demarcation in the phylogenetic relationships between landraces, cultivars, and wild relatives (Ram *et al.* 2007). Microsatellites were also used to assess genetic variability in 75 avocado accessions which revealed a deficit of heterozygotes in most loci due to a positive fixation index (*F*),

departure from Hard-Weinberg expectations, and clustering of accessions into three major groups in the phylogenetic tree (Alcaraz and Hormaza 2007). Maize, an important agronomic crop, has been extensively evaluated with microsatellite markers. One study evaluated 260 inbred lines with 94 microsatellites demonstrated that the phylogenetic relationships and a model based clustering analysis were congruent with pedigree information (Liu *et al.* 2003). A different microsatellite derived phylogeny study demonstrated that maize, contrary to proposed opinion to explain its high genetic diversity, had a single domestication event from teosinte (wild relative) about 9,000 years ago (Matsuoka *et al.* 2002). A monophyletic origin of cultivated pearl millet (*Pennisetum glaucum*) was also identified by employing microsatellites, phylogeny, and population structure analyses (Oumar *et al*. 2008).

Applying microsatellite markers and phylogenetics are indispensable tools for managing plant germplasm collections because the marker data can provide information on the diversity or homogeneity of a species of interest. Genetic diversity and phylogenetic relationships were evaluated from germplasm collections such as a temperate bamboo (Barkley *et al.* 2005), a citrus variety collection (Barkley *et al.* 2006), sorghum (Wang *et al.* 2006, 2009b), and a cultivated and wild peanut collection (Barkley *et al.* 2007). Inter- and intraspecific phylogenetic analysis led to the identification of a contaminated bamboo plot identified via phylogenetic analysis of multiple taxa (Barkley *et al.* 2005), which was subsequently validated by morphology and the plot was purged of the contamination. A study of genetic diversity and phylogenetic relationships in citrus demonstrated that there are only a few naturally occurring forms of citrus, while the remaining species arose through hybridization events from the ancestral species (Barkley *et al.* 2006). These studies focused on germplasm diversity and management identified contaminated plots, putative parentage of hybrid accessions, genetic diversity, disease resistance, population structure, and tentatively classified accessions into botanical varieties. Furthermore, this technology can also help researchers indicate potential needs for expansion of a germplasm particular collection based on a lack of diversity within a species.

CONCLUDING REMARKS

In silico analysis of phylogenetic relationships, genetic diversity, and population stratification in plants and insects has become common practice since the discovery of microsatellites and availability of computers and free software programs to handle these types of data sets. Data mining and analysis of sizeable microsatellite data sets would be impractical without the aid of *in silico* analysis from mining the large number of sequences currently deposited in publically available sequence databases for microsatellites to employing software programs to calculating complex model based ancestry estimations on multilocus genotypes. As time goes on, computation will continue to become more powerful, sequence databases will continue to expand, and the need for more complexity in mathematical models to analyze data will continue to be realized. Numerous programs currently exist for *in silico* data analysis and mining. This review is not meant to support or endorse any particular program; but, provides some resources and information for a researcher to consider, navigate, and utilize. The user should, however, select *in silico* programs in which the underlying models and output can be easily interpreted and help them to elucidate principles, solve questions, and achieve their research goals for their species of interest.

ACKNOWLEDGEMENTS

We are grateful to Jessica Norris for reviewing and editing this paper.

REFERENCES

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

Aishwarya V, Grover A, Sharma PC (2007) EuMicroSatdb: A database for microsatellites in the sequenced genomes of eukaryotes. *BMC Genomics* **8**, 225

Aishwarya V, Sharma PC (2008) UgMicroSatdb: Database for mining microsatellites from unigenes. *Nucleic Acids Res*earch **36**, D53-D56

Akkaya MS, Shoemaker CR, Specht JE, Bhagwat AA, Cregan PB (1995) Integration of simple sequence repeat DNA markers into a soybean linkage map. *Crop Science* **35**, 1439-1445

Albanése V, Biguet NF, Kiefer H, Bayard E, Mallet J, Meloni R (2001) Quantitative effects on gene silencing by allelic variation at a tetranucleotide microsatellite. *Human Molecular Genetics* **10**, 1785-1792

Alcaraz ML, Hormaza JI (2007) Molecular characterization and genetic diversity in an avocado collection of cultivars and local Spanish genotypes using SSRs. *Hereditas* **144**, 244-253

Allen JM, Light JE, Perotti MA, Braig HR, Reed DL (2009) Mutational meltdown in primary endosymbionts: selection limits Muller's ratchet. *PLoS ONE* **4**, e4969

Alexander D, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**, 1655-1664

Anmarkrud JA, Kleven O, Bachmann L, Lifjeld JT (2008) Microsatellite evolution: Mutations, sequence variation, and homoplasy in the hypervariable avian microsatellite locus *HrU10*. *BMC Evolutionary Biology* **8**, 138

Arakaki M, Soltis DE, Soltis PS, Speranza PR (2010) Characterization of polymorphic microsatellite loci in Haageocereus (Trichocereeae, Cactaceae). *American Journal of Botany* **97**, e17-e19

Archak S, Meduri E, Kumar PS, Nagaraju J (2007) InSatDb: A microsatellite database of fully sequenced insect genomes. *Nucleic Acids Research* **35**, D36-9

Arévalo E, Zhu Y, Carpenter JM, Strassmann JE (2004) The phylogeney of the social wasp subfamily Polistinae: Evidence from microsatellite flanking sequences, mitochondrial COI sequence, and morphological characters. *BMC Evolutionary Biology* **4**, 8

Armbruster WS (1992) Phylogeny and the evolution of plant-animal interactions. *Bioscience* **42**, 12-20

Arthofer W, Schlick-Steiner BC, Steiner FM, Avtzis DN, Crozier RH, Stauffer C (2007) Lessons from a beetle and an ant: Coping with taxon-dependent differences in microsatellite development success. *Journal of Molecular Evolution* **65**, 304-307

Augustinos AA, Stratikopoulos EE, Drosopoulou E, Kakani EG, Mavragani-Tsipidou P, Zacharopoulou A, Mathiopoulos KD (2008) Isolation and characterization of microsatellite markers from the olive fly, Bactrocera oleae, and their cross-species amplification in the Tephritidae family. *BMC Genomics* **9**, 618

Avise JC, Ball RM (1990) Principles of genealogical concordance in species concepts and biological taxonomy. In: Gutuyama D, Antonovics J (Eds) *Oxford Surveys in Evolutionary Biology 7*, Oxford University Press, Oxford, England, pp 45-76

Bagshaw ATM, Pitt JPW, Gemmell NJ (2008) High frequency of microsatellites in *S. cerevisiae* meiotic recombination hotspots. *BMC Genomics* **9**, 49

Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with microsatellite markers. *Molecular Ecology* **11**, 155-165

Barbará T, Palma-Silva C, Paggi GM, Bered F, Fay MF, Lexer C (2007) Cross-species transfer of nuclear microsatellite markers: Potential and limitations. *Molecular Ecology* **16**, 3759-3767

Barkley NA, Newman ML, Wang ML, Hotchkiss MW, Pederson GA (2005) Assessment of the genetic diversity and phylogenetic relationships of a temperate bamboo collection by using transferred EST-SSR markers. *Genome* **48**, 731-737

Barkley NA, Roose, ML, Krueger RR, Federici CT (2006) Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (SSRs). *Theoretical and Applied Genetics* **112**, 1519-1531

Barkley NA, Dean RE, Pittman RN, Wang ML, Holbrook CC, Pederson GA (2007) Genetic diversity of cultivated and wild-type peanuts evaluated with M13-tailed SSR markers and sequencing. *Genetic Research Cambridge* **89**, 93-106

Barkley NL, Krueger R, Federici CT, Roose ML (2009) What phylogeny and gene genealogy analyses reveal about homoplasy in citrus microsatellite alleles. *Plant Systematics and Evolution* **282**, 71-86

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics* **16**, 113-148

Bataille A, Cunningham AA, Cedeño V, Patiño, Constantinou A, Kramer LD, Goodman SJ (2009) Natural colonization and adaptation of a mosquito species in Galápagos and its implications for disease threats to endemic wildlife. *Proceedings of the National Academy of Sciences USA* **106**, 10230-10235

Baum DA, Smith SD, Donovan SS (2005) Evolution. The tree-thinking challenge. *Science* **310**, 979-980

Beerli P, Palczewski M (2010) Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics* **185**, 313-326

Benet A, Mollá G, Azorín F (2000) d(GA.TC)$_n$ microsatellite DNA sequences enhance homologous DNA recombination in SV40 minichromosomes. *Nucleic Acids Research* **28**, 4617-4622

Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2007) GenBank. *Nucleic Acid Research* **36**, D25-D30

Bhargava A, Fuentes FF (2009) Mutational dynamics of microsatellites. *Molecular Biotechnology* **44**, 250-266

Biet E, Sun J, Dutreix M (1999) Conserved sequence preference in DNA binding among recombination proteins: An effect of ssDNA secondary structure. *Nucleic Acids Research* **27**, 596-600

Bizzaro JW, Marx KA (2003) Poly: A quantitative analysis tool for simple sequence repeat (SSR) tracts in DNA. *BMC Bioinformatics* **4**, 22

Blenda A, Scheffler J, Scheffler B, Palmer M, Lacape M, Yi JZ, Jesudurai C, Jung S, Muthukumar S, Yellambalase P, Ficklin S, Staton M, Eshelman R, Ulloa M, Saha S, Burr B, Liu S, Zhang T, Fang D, Pepper A, Kumpatla S, Jacobs J, Tomkins J, Cantrell R, Main D (2006) CMD: A cotton microsatellite database resource for *Gossypium* genomics. *BMC Genomics* **7**, 132

Boguski MS, Lowe TM, Toistoshev CM (1993) dbEST-database for "expressed sequenced tags." *Nature Genetics* **4**, 332-333

Bouck A, Vision T (2007) The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology* **16**, 907-924

Caburet S, Cocquet J, Vaiman D, Veitia RA (2005) Coding repeats and evolutionary agility. *Bioessays* **27**, 581-587

Calabrese PP, Durrentt RT, Aquadro CF (2001) Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. *Genetics* **159**, 839-852

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* **156**, 847-854

Carletto J, Lombaert E, Chavigny P, Brévault T, Lapchin L, Vanlerberghe-Masutti F (2009) Ecological specialization of the aphid *Aphis gossypii* Glover on cultivated host plants. *Molecular Ecology* **18**, 2198-2212

Castelo AT, Martins W, Gao GR (2002) TROLL-tandem repeat occurrence locator. *Bioinformatics* **4**, 634-636

Cato SA, Richardson TE (1996) Inter- and intraspecific polymorphism at chloroplast SSR loci and the inheritance of plastids in *Pinus radiata* D. Don. *Theoretical and Applied Genetics* **93**, 587-592

Chaix G, Gerber S, Razafimaharo V, Vigneron P, Verhaegen D, Hamon S (2003) Gene flow estimation with microsatellites in a Malagasy seed orchard of *Eucalyptus grandis*. *Theoretical and Applied Genetics* **107**, 705-712

Chambers EW, Meece JK, McGowan JA, Lovin DD, Hemme RR, Chadee DD, McAbee K, Brown SF, Knudson DL, Severson DW (2007) Microsatellite isolation and linkage group identification in the yellow fever mosquito *Aedes aegypti*. *Heredity* **98**, 202-210

Chambers GK,MacAvoy ES (2000) Microsatellites: Consensus and controversy. *Comparative Biochemistry and Physiology – Part B: Biochemistry and Molecular Biology* **26**, 455-476

Cochrane G, Akhtar R, Addebert P, Althorpe N, Baldwin A, Bates K, Bhattacharyya S, Bonfield J, Bower L, Browne P, Castro M, Cox T, Demiralp F, Eberhardt R, Faruque N, Hoad G, Jang M, Kulikova T, Labarga A, Leininen R, Leonard S, Lin Q, Lopez R, Lorenc D, McWilliams H, Mukherjee G, Nardone F, Plaister S, Robinson S, Sobhany S, Vaughan R, Wu D, Zhu W, Apweiler R, Hubbard T, Birney E (2007) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL nucleotide sequence database. *Nucleic Acid Research* **36**, D5-D12

Colson I, Goldstein DB (1999) Evidence for complex mutations at microsatellite loci in *Drosophila*. *Genetics* **152**, 617-627

Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M (2002) A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nature Genetics* **30**, 315-320

Crane CH (2007) Patterned sequence in the transcriptome of vascular plants. *BMC Genomics* **8**, 173

Cregan PB, Jarvik T, Bush AL, Shoemaker RC, Lark KG, Kahler AL, Kaya N, Van Toai TT, Lohnes DG, Chung J, Specht JE (1999) An integrated genetic linkage map of the soybean. *Crop Science* **39**, 1464-1490

Curtu AL, Finkeldey R, Gailing O (2004) Comparative sequencing of a microsatellite locus reveals size homoplasy within and between European oak species (*Quercus* spp.). *Plant Molecular Biology Reporter* **22**, 339-346

Dakin EE, Avise JC (2004) Microsatellite null alleles in parentage analysis. *Heredity* **93**, 504-509

da Maia CL, Palmieri DA, de Souza VQ, Kopp MM, de Carvalho FIF, de Oliveira AC (2008) SSR Locator: Tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *International Journal of Plant Genomics* **2008**, 412696

Davison D, Pritchard JK, Coop G (2009) An approximate likelihood for genetic data under a model with recombination and population splitting. *Theoretical Population Biology* **75**, 331-344

Debout GDG, Ventelon-Debout M, Emerson BC, Yu DW (2007) PCR primers for polymorphic microsatellite loci in the plant-and *Azteca ulei cordiae* (Formicidae: Dolichoderinae). *Molecular Ecology Notes* **7**, 607-609

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* **6**, 361-375

Dechering KKJ, Cuelenaere K, Konings RNH, Leunissen JAM (1998) Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Research* **26**, 4056-4062

Decroocq V, Favé MG, Hagen L, Borenave L, Decroocq S (2003) Development and transferability of apricot and grape EST microsatellite markers across taxa. *Theoretical and Applied Genetics* **106**, 912-922

Dzialuk A, Chybicki I, Burczyk J (2005) PCR multiplexing of nuclear microsatellite loci in *Quercus* species. *Plant Molecular Biology Reporter* **23**, 121-128

Edh K, Widén B, Ceplitis A (2007) Nuclear and chloroplast microsatellites reveal extreme population differentiation and limited gene flow in the Aegean endemic *Brassica cretica* (Brassicaceae). *Molecular Ecology* **16**, 4972-4983

Ellegren H (2004) Microsatellites: Simple sequences with complex evolution. *Nature Review* **5**, 435-445

Ellis JR, Burke JM (2007) EST-SSRs as a resource for population genetic analyses. *Heredity* **99**, 125-132

Estoup A, Solignac M, Harry M, Cornuet JM (1993) Characterization of (GT)n and (CT)n microsatellites in two insect species: *Apis mellifera* and *Bombus terrestris*. *Nucleic Acids Research* **25**, 1427-1431

Excoffier L, Heckel G (2006) Computer programs for population genetics data analysis: A survival guide *Nature Reviews Genetics* **7**, 745-758

Exeler N, Kratochwil A, Hochkirch A (2008) Strong genetic exchange among populations of a specialist bee, *Andrena vaga* (Hymenoptera: Andrenidae). *Conservation Genetics* **9**, 1233-1241

Fagerberg AJ, Fulton RE, Black WC (2001) Microsatellite loci are not abundant in all arthropod genomes: analysis in the hard tick, *Ixodes scapularis* and the yellow fever mosquito, *Aedes aegypti*. *Insect Molecular Biology* **10**, 225-236

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

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

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle

Fisher MC, Koenig G, White TJ, Taylor JW (2000). A test for concordance between the multilocus genealogies of genes and microsatellites in the pathogenic fungus *Coccidioides immitis*. *Molecular Biology and Evolution* **17**, 1164-1174

Freitas PD, Martins DS, Galetti PM (2008) CID: A rapid and efficient bioinformatic tool for the detection of SSRs from genomic libraries. *Molecular Ecology Resources* **8**, 107-108

Futuyma DJ, Agrawal AA (2009) Macroevoltuion and the biological diversity of plants and herbivores. *Proceedings of the National Academy of Sciences USA* **106**, 18054-18061

Galperin MY, Cochrane GR (2009) Nucleic Acids Research annual database issue and the NAR online molecular biology database collection in 2009. *Nucleic Acid Research* **37**, D1-D4

Garner TWJ (2002) Genome size and microsatellites: The effect of nuclear size on amplification potential. *Genome* 45, 212-215.

Gerber HP, Seipel K, Georgiev O, Höfferer M, Hug M, Rusconi S, Schaffner W (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263**, 808-811

Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO, Petes TD (2000) Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences USA* **97**, 11383-11390

Gil R, Latorre A, Moya A (2004) Bacterial endosymbionts of insects: insights from comparative genomics. *Environmental Microbiology* **6**, 1109-1122.

Gilbert D (2004) Software review: Bioinformatics software resources. *Briefings Bioinformatics* **5**, 300-304

Gitschier J (2009) You say you want a revolution: An interview with Pat Brown. *PLOS Genetics* **5 (7)**, e1000560

Goldstein DB, Pollock DD (1997) Launching microsatellites: A review of mutation processes and methoids of phylogenetic inference. *Journal of Heredity* **88**, 335-342

Goldstein DB, Schlotterer C (1999) *Microsatellites: Evolution and Applications*, Oxford University Press, Oxford, 352 pp

Gong L, Stift G, Kofler R, Pachner M., Lelley T (2008) Microsatellites for the genus *Cucurbita* and an SSR-based genetic linkage map of *cucurbita pepo* L. *Theoretical and Applied Genetics* **117**, 37-48

Goodman SJ (1997) Rst Calc: A collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. *Molecular Ecology* **6**, 881-885

Grasela JJ, McIntosh AH (2005) Cross-species investigation of *Helicoverpa*

armigera microsatellites as potential markers for other related species in *Helicoverpa-Heliothis* complex. *Journal of Insect Science* **5**, 47

Guo WJ, Ling J, Li P (2009) Consensus features of microsatellite distribution: Microsatellite contents are universally correlated with recombination rates and are preferentially depressed by centromeres in multicellular eukaryotic genomes. *Genomics* **93**, 323-331

Hale ML, Borland AM, Gustafsson MHG, Wolff K (2004) Causes of size homoplasy among chloroplast microsatellites in closely related *Clusia* species. *Journal of Molecular Evolution* **58**, 182-190

Hancock JM (1995) The contribution of slippage-like processes to genome evolution. *Journal of Molecular Evolution* **41**, 1038-1047

Harr B, Zangerl B, Schlötterer (2000) Removal of microsatellite interruptions by DNA replkication slippage: Phylogenetic evidence from Drosophila. *Molecular Biology and Evolution* **17**, 1001-1009

Higgs PG, Attwood TK (2007) *Bioinformatics and Molecular Evolution*, Blackwell Publishing Malden, MA, 384 pp

Hisano H, Sato S, Isobe S, Shigemi S, Wada T, Matsuno A, Fkujishiro T, Yamada M, Nakayama S, Nakamura Y, Watanabe S, Harada K, Tabata S (2007) Characterization of the soybean genome using EST-derived microsatellite markers. *DNA Research* **14**, 271-281

Hochkirch A, Damerau M (2009) Rapid range expansion of a wing-dimorphic bush-cricket after the 2003 climatic anomaly. *Biological Journal of the Linnean Society* **97**, 118-127

Holzer B, Chapuisat M, Kremer N, Finet C, Keller L (2006) Unicoloniality, recognition and genetic differentiation in a native Formica ant. *Journal of Compilation* **19**, 2031-2039

Horning ME, Cronn RC (2009) Development of variable microsatellite loci and range-wide characterization of nuclear genetic diversity in the important dryland shrub antelope bitterbrush (*Purshia tridentata*). *Journal of Arid Environments* **73**, 7-13

Hu LJ, Uchiyama K, Shen HL, Saito Y, Tsuda Y, Ide Y (2008) Nuclear DNA microsatellites reveal genetic variation but a lack of phylogeographical structure in an endangered species, *Fraxinus mandshurica*, across North-east China. *Annals of Botany* **102**, 195-205

Huelsenbeck JP, Rannala B (1997) Phylogenetic methods come of age: testing hypotheses in an evolutionary context. *Science* **276**, 227-232

Huntley MA, Golding GB (2006) Selection and slippage creating serine homopolymers. *Molecular Biology and Evolution* **23**, 2017-1025

Huttunen S, Schötterer C (2002) Isolation and characterization of microsatellites in Drosophila virilis and their cross species amplification in members of the *D. virilis* group. *Molecular Ecology Notes* **2**, 593-597

Iglesias AR, Kindlund E, Tammi M, Wadelius C (2004) Some microsatellites may act as novel polymorphic *cis*-regulatory elements through transcription factor binding. *Gene* **341**, 149-165

Ijaz S, Khan IA (2009) Molecular characterization of wheat germplasm using microsatellite markers. *Genetics and Molecular Research* **8**, 809-815

Jarne P, Lagoda JL (1996) Microsatellites, from molecules to populations and back. *TREE* **11**, 424-429

Jayashree B, Punna P, Prasad P, Bantte K, Tom Hash C, Chandra S, Hoisington DA, Varshney PK (2006) A database of simple sequence repeats from cereal and legume expressed sequence tags mined in silico: Survey and evaluation. *In Silico Biology* **6**, 607-620

Jenkins TM, Newman ML, Forschler BT (2002) Subterranean termite movements and relationships over time: a genetic characterization. In: Zhai J, Robinson WM, Jones SC (Eds) *Proceedings of the 4th International Conference on Urban Pests*, Pocahontas Press, Inc., Charleston, SC, pp 95-102

Jenkins TM, Jones SC, Lee C-Y, Forschler BT, Chen Z, Lopez-Martinez G, Gallagher NT, Brown G, Neal M, Thistleton B, Kleinschmidt S (2007) Phylogeography illuminates maternal origins of exotic *Coptotermes gestroi* (Isoptera: Rhinotermitidae). *Molecular Phylogenetics and Evolution* **42**, 612-621

Jenkins TM, Braman SK, Chen Z, Eaton, TD, Pettis CV, Boyd DW (2009) Insights into flea beetle (Coleoptera: Chrysomelidae: Galerucinae) host specificity from concordant mitochondrial and nuclear DNA phylogenies. *Annals of the Entomological Society of America* **102**, 386-395

Johnson PCK, Haydon DT (2007a) Maximum likelihood estimation of alleleic dropout and false allele error rates from microsatellite genotypes in the absence of reference data. *Genetics* **175**, 827-842

Johnson PCK, Haydon DT (2007b) Software for quantifying and simulating microsatellite genotyping error. *Bioinformatics and Biology Insights* **1**, 71-75

Jongeneel CV (2000) Searching the expressed sequence tag (EST) databases: Panning for genes. *Briefings in Bioinformatics* **1**, 76-92

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

Kikuchi Y, Hosokawa T, Fukatsu T (2007) Insect-microbe mutualism without vertical transmission: A stinkbug acquires a beneficial gut symbiont from the environment every generation. *Applied Environmental Microbiology* **73**, 43089-4316

Kim KS, Ratcliffe ST, French BW, Liu L, Sappington TW (2008) Utility of EST-derived SSRs as population genetic markers in a beetle. *Journal of Heredity* **99**, 112-124

Kim KS, Sappington TW (2005a) Polymorphic microsatellite loci from the western corn rootworm (Insecta: Coleoptera: Chrysomelidae) and cross-amplification with other *Diabrotica* spp. *Molecular Ecology Notes* **5**, 115-117

Kim KS, Sappington TW (2005b) Genetic structuring of western corn rootworm (Coleoptera: Chrysomelidae) populations in the United States based on microsatellite loci analysis. *Environmental Entomology* **34**, 494-503

Kirkpatrick DT, Wang YH, Dominska M, Griffith JD, Petes TD (1999) Control of meiotic recombination and gene expression in yeast by a simple repetitive DNa sequence that excludes nucleosomes. *Molecular Cell Biology* **19**, 7661-7671

Kobayashi T (2008) Development of polymorphic microsatellite markers for the sorghum plant bug, *Stenotus rubrovittatus* (Heteroptera: Miridae). *Molecular Ecology Resources* **8**, 690-691

Kruglyak S, Durrett RT, Schug MD, Aquadro CF (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences USA* **95**, 10774-10778

Kuhner MK (2006) LAMARC 2.0: Maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* **22**, 768-770

Kwak M, Gepts P (2009) Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae). *Theoretical and Applied Genetics* **118**, 979-992

Labate JA (2000) Software for population genetic analyses of molecular marker data. *Crop Science* **4**, 1521-1528

Lagercrantz U, Ellegren H, Andersson L (1993) The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. *Nucleic Acids Research* **21**, 1111-1115

Lawson MJ, Zhang LQ (2006) Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biology* **7**, 1-11

Legendre M, Pocket N, Pak T, Verstrepen KJ (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Research* **17**, 1787-1796

Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J (1996) Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proceedings of the National Academy of Sciences USA* **93**, 10864-10869

Li YC, Fahima T, Korol AB, Peng J, Röder MS, Kirzhner V, Beiles A, Nevo E (2000) Microsatellite diversity correlated with ecological-edaphic and genetic factors in three microsites of wild emmer wheat in North Israel. *Molecular Biology and Evolution* **17**, 851-862

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, Sjukste TG, Röder, Ganal MW (2003) Development and genetic mapping of 127 new microsatellite markers in barley. *Theoretical and Applied Genetics* **107**, 1021-1027

Li Y-C, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: Structure, function, and evolution. *Molecular Biology and Evolution* **21**, 991-1007

Li M, Shen L, Xu A, Miao X, Hou C, Sun P, Zhang Y, Huang Y (2005) Genetic diversity among silkworm (*Bombyx mori* L., Lep., Bombycidae) germplasms revealed by microsatellites. *Genome* **48**, 802-810

Lia VV, Sun F (2003) The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular Biology and Evolution* **20**, 2123-2131

Lia VV, Confalonieri VA, Ratto N, Hernández JAC, Alzogaray AMM, Poggio L, Brown TA (2007) Microsatellite typing of ancient maize: Insights into the history of agriculture in southern South America. *Proceedings of the Royal Society of London. Series B* **74**, 545-554

Liu K, Muse SV (2005) PowerMarker: Integrated analysis environment for genetic marker data. *Bioinformatics* **21**, 2128-2129

Liu K, Goodman M, Muse S, Smith JS, Buckler E, Doebley J (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* **165**, 2117-2128

Llewellyn MS, Lewis MD, Acosta N, Yeo M, Carrasco HJ, Segovia M, Vargas J, Torrico F, Miles MA, Gaunt MW (2009) *Trypanosoma cruzi* IIc: Phylogenetic and phylogeographic insights from sequence and microsatellite analysis and potential impact on emergent Chagas disease. *PLoS Neglected Tropical Disease* **3**, e510

Lourmas M, Kjellberg F, Dessard H, Joly HI, Chevallier M-H (2007) Reduced density due to logging and its consequences on mating system and pollen flow in the African mahogany *Entandrphragma cylindricum*. *Heredity* **99**, 151-160

Lovin DD, Washington KO, deBruyn B, Hemme RR, Mori A, Epstein SR, Harker BW, Streit TG, Severson DW (2009) Genome-based polymorphic microsatellite development and validation in the mosquito *Aedes aegypti* and application to population genetics in Haiti. *BMC Genomics* **10**, 590

Mace GM, Gittleman JL, Purvis A (2003) Preserving the tree of life. *Science* **300**, 1707-1709

Maguire TL, Edwards KJ, Saenger P, Henry R (2000) Characterisation and analysis of microsatellite loci in a mangrove species, *Avicennia marina* (Forsk) Vierh. (Avicenniaceae). *Theoretical and Applied Genetics* **101**, 279-285

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of

human population structure on large genetic association studies. *Nature Genetics* **36**, 512-517

Marriage TN, Hudman S, Mort ME, Kelly JK (2009) Direct estimation of the mutation rate at dinucleotide microsatellite loci in *Arabidopsis thaliana* (Brassicaceae). *Heredity* **103**, 310-317

Marshall HD, Newton C, Ritland K (2002) Chloroplast phylogeography and evolution of highly polymorphic microsatellites in lodgepole pine (*Pinus contorta*). *Theoretical and Applied Genetics* **104**, 367-378

Martins WS, Lucas DCS, Neves KFS, Bertioli DJ (2009) WebSat - a web software for microsatellite marker development. *Bioinformation* **3**, 282-283

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez GJ, Buckler E, Doebley J (2002) A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences USA* **99**, 6080-6084

Mayer C (2006-2010) http://www.rub.de/spezzoo/cm/cm_phobos.htm

Mayr E (2001) *What Evolution is*, Basic Books, New York, 318 pp

McGrath S, Hodkinson TR, Barth S (2007) Extremely high cytoplasmic diversity in natural and breeding populations of *Lolium* (Poaceae). *Heredity* **99**, 531-544

McWilliam H, Valentin F, Goujon M, Li W, Narayanasamy M, Martin J, Miyar T, Lopez R (2009) Web services at the European Bioinformatics Institute-2009. *Nucleic Acids Research* **37**, W6-W10

Meglécz E, Costedoat C, Dubut V, Gilles A, Malausa T, Pech N, Martin J-F (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics* **26**, 403-404

Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Research* **10**, 72-80

Miao XX, Xu SJ, Li MH, Li MW, Huang JH, Dai FY, Marino SW, Mills DR, Zeng P, Mita K, Jia SH, Zhang Y, Liu WB, Xiang H, Guo QH, Xu AY, Kong XY, Lin HX, Shi YZ, Lu G, Zhang X, Huang W, Yasukochi Y, Sugasaki T, Shimada T, Nagaraju J, Xiang ZH, Wang SY, Goldsmith MR, Lu C, Zhao GP Huang YP (2005) Simple sequence repeat-based consensus linkage map of *Bombyx mori*. *Proceedings of the National Academy of Sciences USA* **102**, 16303-16308

Miesfeld R, Krystal M, Arnheim N (1981) A member of a new repeated sequence family which is conserved throughout the eukaryotic evolution is found between the human globin genes. *Nucleic Acids Research* **9**, 5931-5947

Mikheyev AS, Vo T, Wee B, Singer MC, Parmesan C (2010) Rapid microsatellite isolation from a butterfly by *De Novo* transcriptome sequencing: performance and a comparison with AFLP-derived distances. *PLoS ONE* **5**, e11212

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

Moran NA, Tran P, Gerardo NM (2005) Symbiosis and insect diversification: an ancient symbiont of sap-feeding insects from the bacterial phylum Bacteroidetes. *Applied Environmental Microbiology* **71**, 8802-8810

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics* **30**, 194-200

Moxon ER, Wills C (1999) DNA microsatellites: Agents of evolution. *Scientific American* **280**, 94-99

Nauta MJ, Weissing FJ (1996) Constraints on allele size at microsatellite loci: Implications for genetic differentiation. *Genetics* **143**, 1021-1030

Nève G, Meglécz E (2000) Microsatellite frequencies in different taxa. *Trends in Ecology and Evolution* **15**, 376-377

Nishimura S, Hinomoto N, Takafuji A (2005) Gene flow and spatio-temporal genetic variation among sympatric populations of *Tetranychus kanzawai* (Ascari: Tetranychidae) occurring on different host plants, as estimated by microsatellite gene diversity. *Experimental Applied Acarology* **35**, 59-71

Noor MAF, Feder JL (2006) Speciation genetics: evolving approaches. *Nature Review Genetics* **7**, 851-861

Norris DE, Shurtleff AC, Touré YT, Lanzaro GC (2001) Microsatellite DNA polymorphism and heterozygosity among field and laboratory populations of *Anopheles gambiae* S.S. (Diptera: Culicidae). *Journal of Medical Entomology* **38**, 336-340

O'Brien EA, Ahkang Y, Yang L, Wang E, Marie V, Lang BF, Burger G (2006) GOBASE – a database of organelle and bacterial genome information. *Nucleic Acids Research* **34**, D697-D699

O'Brien EA, Zkhang Y, Wang E, Marie J, Badejoko W, Long BF, Burger G (2009) GOBASE: An organelle genome database. *Nucleic Acid Research* **37**, D946-D950

Ochieng JW, Steane DA, Iadiges PY, Baverstock PR, Henry RJ, Shepherd M (2007) Microsatellites retain phylogenetic signals across genera in eucalypts (Myrtaceae). *Genetics and Molecular Biology* **30**, 1125-1134

Oliveira RP, Broude NE, Macedo AM, Cantor CR, Smith CL, Pena SDJ (1998) Probing the genetic population structure of *Trypanosoma cruzi* with polymorphic microsatellites. *Proceedings of the National Academy of Sciences USA* **95**, 3776-3780

Olsen KM, Schaal BA (2001) Microsatellite variation in cassava (*Manihot esculenta*, Euphorbiaceae) and its wild relatives: Further evidence for a southern Amazonian origin of domestication. *American Journal of Botany* **88**, 131-142

O'Neill SL, Giordano R, Colbert AME, Karr TL, Robertson HM (1992) 16S rRNA phylogeneticanalysis of the bacterial endosymbionts associated with cytoplasmic incompatibility in insects. *Proceedings of the National Academy of Sciences USA* **89**, 2699-2702

Oumar I, Mariac C, Pham JL, Vigouroux Y (2008) Phylogeny and origin of pearl millet (*Pennisetum glaucum* [L.] R. Br) as revealed by microsatellite loci. *Theoretical and Applied Genetics* **117**, 489-497

Palomeque T, Lorite P (2008) Satellite DNA in insects: A review. *Heredity* **100**, 564-573

Pannebakker BA, Niehuis O, Hedley A, Gadau J, Shuker DM (2010) The distribution of microsatellites in the *Nasonia* parasitoid wasp genome. *Insect Molecular Biology* **19**, 91-98

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

Parida SK, Dalal V, Singh AK, Singh NK, Mohapatra T (2009) Genic non-coding microsatellites in the rice genome: Characterization, marker design and use in assessing genetic and evolutionary relationships among domesticated groups. *BMC Genomics* **10**, 140

Patterson N, Prices AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics* **2**, 2074-2093

Pearson CE, Edamura KN, Cleary JD (2005) Repeat instability: Mechanisms of dynamic mutations. *Nature Reviews Genetics* **6**, 729-742

Pérez de Rosas AR, Segura EL, Fichera L, García BA (2008) Macrogeographic and microgeographic genetic st4ructure of the Chagas' disease vector *Triatoma infestans* (Hemiptera: Reduviidae) from Catamarca, Argentina. *Genetica* **133**, 247-260

Perutz MF, Johnson T, Suzuki M, Finch JT (1994) Glutamine repeats as polar zippers: Their possible role in inherited neurodegenerative diseases. *Proceedings of the National Academy of the Sciences USA* **91**, 5355-5358

Petes TD (2001) Meiotic recombination hot spots and cold spots. *Nature Reviews* **2**, 360-369

Pizarro JC, Gilligan LM, Stevens L (2008) Microsatellites reveal a high population structure in *Triatoma infestans* from Chuquisaca, Bolivia. *PLoS Neglected Tropical Diseases* **2**, e202

Pool JE, Hellmann I, Jensen JD, Nielson R (2010) Population genetics inference from genomic sequence variation. *Genome Research* **20**, 291-300

Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. *Trends in Plant Science* **1**, 215-222

Prasad MD, Muthulakshmi M, Arunkumar KP, Madhu M, Sreenu VB, Pavithra V, Bose B, Nagarajaram HA, Mita K, Shimada T, Nagaraju J (2005) SilkSatDb: A microsatellite database of the silkworm, *Bombyx mori*. *Nucleic Acids Research* **33**, D403-D406

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904-909

Primmer CR, Møller AP, Ellegren H (1996) A wide-range survey of cross-species microsatellite amplification in birds. *Molecular Ecology* **5**, 365-378

Primmer CR, painterJN, Koskinen MT, Palo JU, Merilä J (2005) Factors affecting avian cross-species microsatellite amplification. *Journal of Avian Biology* **36**, 348-360

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

Pritchard JK, Wen W (2004) Documentation for the STRUCTURE software Version 2.3.3. http://pritch.bsd.uchicago.edu/structure.html

Proven J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: New tools for studies in plant ecology and evolution. *Trends in Ecological Evolution* **16**, 142-147

Raabova J, Hans G, Risterucci AM, Jacquemart AL, Raspe O (2010) Development and multiplexing of microsatellite markers in the polyploid perennial herb, *Menyanthes trifoliata* (Menyanthaceae). *American Journal of Botany* **97**, e31-e33

Rajendrakumar P, Biswall AK, Balachandran SM, Sundaram RM (2008) *In silico* analysis of microsatellites in organellar genomes of major cereals for understanding their phylogenetic relationships. *In Silico Biology* **8**, 87-104

Ram SG, Thiruvengadam V, Vinod KK (2007) Genetic diversity among cultivars, landraces and wild relatives of rice as revealed by microsatellite markers. *Journal of Applied Genetics* **48**, 337-345

Rando OJ, Verstrepan KJ (2007) Timescales of genetic and epigenetic inheritance. *Cell* **128**, 655-668

Richard GF, Dujon B (2006) Molecular evolution of minisatellites in hemiascomycetous yeasts. *Molecular Biology and Evolution* **23**, 189-202

Richard GF, Pâques F (2000) Mini- and microsatellite expansions: The recombination connection. *EMBO Reports* **1**, 122-126

Richard M, Thorpe RS (2001) Can microsatellites be used to infer phylogenies? Evidence from population affinities of the Western Canary Island lizard (*Gallotia galloti*). *Molecular Phylogenetics and Evolution* **20**, 351-360

Ritschel PS, de Lima Lins TC, Tristan RL, Buso GSC, Buso JA, Ferreira ME (2004) Development of microsatellite markers from an enriched genomic library for genetic analysis of melon (*Cucumis melo* L.). *BMC Plant Biology* **4**, 9

Roder MS, Korzun V, Wendehake K, Plaschke J, Tixier M-H, Leroy P, Ganal MW (1998) A microsatellite map of wheat. *Genetics* **149**, 2007-2023

Rongnoparut P, Sirichotpakorn N, Rattanarithikul R, Yaicharoen S, Linthicum KJ (1999) Estimates of gene flow among *Anopheles maculatus* populations in Thailand using microsatellite analysis. *American Journal of Tropical Medical Hygiene* **60**, 508-515

Roratto PA, Buchmann D, Santos S, Bartholomei-Santos ML (2008) PCR-mediated recombination in development of microsatellite markers: Mechanism and implications. *Genetics and Molecular Biology* **31**, 58-63

Rosetto M (2001) Sourcing of SSR markers from related plant species. In: Henry RJ (Ed) *Plant Genotyping: The DNA Fingerprinting of Plants*, CAB International, Oxford, UK, pp 211-224

Ross KG, Shoemaker DD (2008) Estimation of the number of founders of an invasive pest insect population: The fire ant *Solenopsis invicta* in the USA. *Proceedings of the Royal Society of London, Section B* **275**, 2231-2240

Roy CB, Nazeer MA, Saha T (2004) Identification of simple sequence repeats in rubber (*Hevea brasiliensis*). *Current Science* **87**, 807-811

Rousset F (2008) Genepop'007: A complete reimplementation of the Genepop software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106

Rudmann-Maurer K, Weyand A, Fischer M, Stöcklin J (2007) Microsatellite diversity of the agriculturally important alpine grass *Poa alpina* in relation to land use and natural environment. *Annals of Botany* **100**, 1249-1258

Schaal BA, Hayworth DA, Olsen KM, Rauscher JT, Smith WA (1998) Phylogeographic studies in plants: Problems and prospects. *Molecular Ecology* **7**, 465-474

Schlipalius DI, Waldron J, Carroll BJ, Collins PJ, Ebert PR (2001) A DNA fingerprinting procedure for ultra high-throughput genetic analysis of insects. *Insect Molecular Biology* **10**, 579-585

Schlötterer C (1988) Genome evolution: Are microsatellites really simple sequences? *Current Biology* **8**, R132-R134

Schrey NM, Schrey AW, Heist EJ, Reeve JD (2008) Fine-scale genetic population structure of Southern pine beetle (Coleoptera: Curculionidae) in Mississippi forests. *Environmental Entomology* **37**, 271-276

Schultes NP, Szostak JW (1991) A poly(dA.dT) tract is a component of the recombination initiation site at the *ARG4* locus in *Saccharomyces cerevisiae*. *Molecular Cell Biology* **11**, 322-328

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

Sia EA, Butler CA, Dominska M, Greenwell P, Fox TD, Petes TD (2000) Analysis of microsatellite mutations in the mitochondrial DNA of *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences USA* **97**, 250-255

Smit AFA, Hubley R, Green P (1996-2004) RepeatMasker Open-3.0 RepeatMasker Open (3.0) http://www.repeatmasker.org

Smith JL, Keyghobadi N, Matrone MA, Escher RL, Fonseca DM (2005) Cross-species comparison of microsatellite loci in the *Culex pipiens* complex and beyond. *Molecular Ecology Notes* **5**, 697-700

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

Spooner DM, Nunez J, Trujillo G, Herrera Mdel R, Guzman F, Ghislain M (2007) Extensive simple sequence repeat genotyping of potato landraces supports a major reevaluation of their gene pool structure and classification. *Proceedings of the National Academy of Sciences USA* **104**, 19398-19403

Spritz RA (1981) Duplication/deletion polymorphism 5′ – to the human beta globin gene. *Nucleic Acids Research* **9**, 5037-5047

Sreenu VB, Alevoor V, Nagaraju AJ, Nagarajaram HA (2003) MICdb: Database of prokaryotic microsatellites. *Nucleic Acids Research* **31**, 106-108

Stoesser G, Baker W, van den Broek A, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, Nardone F, Stoehr P, Tuli MA, Tzouvara K, Vaughan R (2003) The EMBL nucleotide sequence database: major new developments. *Nucleic Acids Research* **31**, 17-22

Stoesser G, Moseley MA, Sleep J, McGowran M, Garcia-Pastor M, Sterk P (1998) The EMBL nucleotide sequence database. *Nucleic Acids Research* **26**, 8-15

Stofer A (1999) Gene flow and endangered species translocation: A topic revisited. *Biological Conservation* **87**, 173-180

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-853

Sugawara H, Ogasawara O, Okubo K, Gojobori T, Tateno Y (2008) DDBJ with new system and face. *Nucleic Acids Research* **36**, D22-D24

Swofford DL (1998) PAUP* Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland

Symonds VV, Lloyd AM (2003) An analysis of microsatellite loci in *Arabidopsis thaliana*: Mutational dynamics and application. *Genetics* **165**, 1475-496

Takahashi J, Itoh M, Shimizu I, Ono M (2008) Male parentage and queen mating frequency in the bumblebee *Bombus ignitus* (Hymenoptera: bombinae). *Ecological Research* **23**, 937-942

Takezaki N, Nei M, Tamura K (2010) POPTREE2: Software for constructing population trees from allele frequency data and computing other population statistics with windows interface. *Molecular Biology and Evolution* **27**, 747-752

Takundwa M, Chimwamurombe PM, Kunert K, Cullis CA (2010) Isolation and characterization of microsatellite repeats in Marama bean (*Tylosema esculentum*). *African Journal of Agricultural Research* **5**, 561-566

Tang SX, Kishore VK, Knapp SJ (2003) PCR-multiplexes for a genome-wide frame-work of simple sequence repeat marker loci in cultivated sunflower. *Theoretical and Applied Genetics* **107**, 6-19

Tang DQ, Lu JJ, Fang W, Zhang S, Zhou MB (2010) Development, characterization and utilization of GenBank microsatellite markers in *Phyllostachys pubescens* and related species. *Molecular Breeding* **25**, 299-311

Tautz D, Schlötterer C (1994) Simple sequences. *Current Genetic Developments* **4**, 832-837

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. *Genome Research* **11**, 1441-1452

Templeton AR (2004) Statistical phylogeography: Methods of evaluating and minimizing inference errors. *Molecular Ecology* **13**, 789-809

Temu EA, Hunt RH, Coetzee M (2004) Microsatellite DNA polymorphism and heterozygosity in the malaria vector mosquito *Anopheles funestus* (Diptera: Culicidae) in east and southern. *Africa Acta Tropica* **90**, 39-49

Thao ML, Moran NA, Abbot P, Brennan EB, Burckhardt DH, Baumann P (2000) Cospeciation of Psyllids and their primary prokaryotic endosymbionts. *Applied Environmental Microbiology* **66**, 2898-2905

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics* **106**, 411-422

Thorén P, Paxton RJ, Estoup A (1995) Unusually high frequency of (CT)n and (GT)n microsatellite loci in a yellowjacket wasp, *Vespula rufa* (L.) (Hymenoptera: Vespidae). *Insect Molecular Biology* **4**, 141-148

Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Research* **10**, 9670981

Van't Hof AE, Brakefield PM, Saccheri IJ, Zwann BJ (2007) Evolutionary dynamics of multilocus microsatellite arrangements in the genome of the butterfly *Bicyclus anynana*, with implications for other Lepidoptera. *Heredity* **98**, 320-328

Varshney RK, Thiel T, Langridge NSP, Graner A (2002) *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell and Molecular Biology Letters* **7**, 537-546

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

Ventura BD, Lemerle C, Michalodimitrakis K, Serrano L (2006) From *in vivo* to *in silico* biology and back: A review. *Nature* **443**, 527-533

Verstrepen KJ, Jansen A, Lewitter F, Fink GR (2005) Intragenic tandem repeats generate functional variability. *Nature Genetics* **37**, 986-990

Vigouroux Y, Jaqueth JS, Matsuoka Y, Smith OS, Beavis WD, Smith JSC, Doebley J (2002) Rate and pattern of mutation at microsatellite loci in maize. *Molecular Biology and Evolution* **19**, 1251-1261

Vigouroux Y, Matsuoka Y, Doebley J (2003) Directional evolution for microsatellite size in maize. *Molecular Biology and Evolution* **20**, 1480-1483

Viard F, Franck P, Dubois MP, Estoup A, Jarne P (1998) Variation of microsatellite size homoplasy across electromorphs, loci, and populations in three invertebrate species. *Journal of Molecular Evolution* **47**, 42-51

Wang ML, Barkley NA, Yu J-K, Dean RE, Newman ML, Sorrells ME, Pederson GA (2005) Transfer of simple sequence repeat (SSR) markers from major cereal crops to minor grass species for germplasm characterization and evaluation. *Plant Genetic Research* **3**, 45-57

Wang ML, Dean R, Erpelding J, Pederson G (2006) Molecular genetic evaluation of sorghum germplasm differing in response to fungal diseases: Rust (*Puccinia purpurea*) and anthracnose (*Collectotrichum graminicola*). *Euphytica* **148**, 319-330

Wang ML, Morris JB, Barkley NA, Dean RE, Jenkins TM, Pederson GA (2007) Evaluation of genetic diversity of the USDA *Lablab purpureus* germplasm collection using simple sequence repeat markers. *Journal of Horticultural Science and Biotechnology* **82**, 571-578

Wang ML, Barkley NA, Jenkins TM (2009a) Microsatellite markers in plants and insects: Applications of biotechnology. *Genes, Genomes, and Genetics* **3 (Special Issue 1)**, 54-67

Wang ML, Zhu C, Barkley NA, Chen Z, Erpelding JE, Murray SC, Tuinstra MR, Tesso T, Pederson GA, Yu J (2009b) Genetic diversity and population structure analysis of accessions in the US historic sweet sorghum collection. *Theoretical and Applied Genetics* **120**, 13-23

Wang Z, Weber JL, Zhong G, Tanksley SD (1994) Survey of plant short tandem DNA repeats. *Theoretical and Applied Genetics* **88**, 1-6

Weng Y, Azhaguvel P, Michels Jr. GJ, Rudd JC (2007) Cross-species transferability of microsatellite markers from six aphid (Hemiptera: Aphididae) species and their use for evaluating biotypic diversity in two cereal aphids. *Insect Molecular Biology* **16**, 613-622

Wright S (1932) The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In: *Proceedings of the Sixth International Congress on Genetics*, pp 355-366

Yang RC (1998) Estimating hierarchical F-statistics. *Evolution* **52**, 950-956

Yasodha R, Sumathi R, Chezhian P, kavitha S, Ghosh M (2008) Eucalyptus

microsatellites mined *in silico*: Survey and evaluation. *Journal of Genetics* **87**, 21-25

Yao I, Akimoto S-I (2009) Seasonal changes in the genetic structure of an aphid-ant mutualism as revealed using microsatellite analysis of the aphid *Tuberculatus quercicola* and the ant *Formica yessensis*. *Journal of Insect Science* **9**, 9

Yao X, Ye Q, Fritsch PW, Cruz BC, Huang H (2008) Phylogeny of *Sinojackia* (Styracaceae) based on DNA sequence and microsatellite data: Implication for taxonomy and conservation. *Annals of Botany* **101**, 651-659

You G, Zhang X, Wang L (2005) An estimation of the minimum number of SSR loci needed to reveal genetic relationships in wheat varieties: Information from 96 random accessions with maximized genetic diversity. *Molecular Breeding* **14**, 397-406

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: A review. *Molecular Ecology* **11**, 1-16

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

Zhang L, Yuan D, Yu S, Li Z, Cao Y, Miao Z, Qian H, Tang K (2004) Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana*. *Bioinformatics* **20**, 1081-1086

Zhang J, Niyogi P, McPeek MS (2009) Laplacian eigenfunctions learn population structure. *PLoS One* **4 (12)**, e7928

Zhang M, Wang H, Dong Z, Qi B, Xu K, Liu B (2010) Tissue culture-induced variation at simple sequence repeats in sorghum (*Sorghum bicolor* L.) is genotype-dependent and associated with down-regulated expression of a mismatch repair gene, MLH3. *Plant Cell Reports* **29**, 51-59

Zheng YJ, Feng SH, Guo EZ, Zheng RZ, Chen J, Zhu ZW, Zhuang JY (2007) Construction and testing of a primary microsatellite database of major rice varieties in China. *Rice Science* **14**, 247-255