Molecular Ecology (2003) **12**, 563–584

INVITED REVIEW

Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects

DE-XING ZHANG* and GODFREY M. HEWITT†

**State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, 19 Zhongguancun Road, Beijing 100080, PR China, †School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, UK*

Abstract

Population-genetic studies have been remarkably productive and successful in the last decade following the invention of PCR technology and the introduction of mitochondrial and microsatellite DNA markers. While mitochondrial DNA has proven powerful for genealogical and evolutionary studies of animal populations, and microsatellite sequences are the most revealing DNA markers available so far for inferring population structure and dynamics, they both have important and unavoidable limitations. To obtain a fuller picture of the history and evolutionary potential of populations, genealogical data from nuclear loci are essential, and the inclusion of other nuclear markers, i.e. single copy nuclear polymorphic (scnp) sequences, is clearly needed. Four major uncertainties for nuclear DNA analyses of populations have been facing us, i.e. the availability of scnp markers for carrying out such analysis, technical laboratory hurdles for resolving haplotypes, difficulty in data analysis because of recombination, low divergence levels and intraspecific multifurcation evolution, and the utility of scnp markers for addressing population-genetic questions. In this review, we discuss the availability of highly polymorphic single copy DNA in the nuclear genome, describe patterns and rate of evolution of nuclear sequences, summarize past empirical and theoretical efforts to recover and analyse data from scnp markers, and examine the difficulties, challenges and opportunities faced in such studies. We show that although challenges still exist, the above-mentioned obstacles are now being removed. Recent advances in technology and increases in statistical power provide the prospect of nuclear DNA analyses becoming routine practice, allowing allele-discriminating characterization of scnp loci and microsatellite loci. This certainly will increase our ability to address more complex questions, and thereby the sophistication of genetic analyses of populations.

Keywords: DHPLC, mitochondrial, microsatellite, nuclear SNP, rate of evolution, recombination and selection,

Received 17 November 2002; revision received 1 December 2002; accepted 1 December 2002

Introduction

In the last decade of the 20th century, we have seen a considerable impact of population-genetic studies on our understanding of evolutionary processes, population and species history, and even our own relationship to nature. Indeed, never before have we been able to understand so much about the evolution and natural history of our own species, *Homo sapiens*. Theoretical, analytical and methodological advances have revolutionized population-genetic research to such a degree that this discipline is no longer largely a debating field of mathematics and theory, but has become an explanatory science and is attracting more popular attention. This is by any standard a great achievement and will further promote the development of the field. Conceptually, this remarkably productive and successful period relied strongly on population-genetics theory established from the 1930s. The recent development of coalescent and phylogenetic theory has changed fundamentally the way we analyse and interpret molecular data. Methodologically, this blossoming is due largely

Correspondence: De-Xing Zhang. Fax: (+ 86) 10 62612962; E-mail: dxzhang@panda.ioz.ac.cn

to the invention of polymerase chain reaction (PCR) technology and the introduction of sensitive molecular markers, most notably using mitochondrial and microsatellite DNA. The effects of such technical advance are so profound that they have 'transformed the mainstream of population-genetics research from prospective to retrospective, from demonstration of principles to inference of events that happened in the past' (Fu & Li 1999). Further advances in the field will continue to be influenced greatly by technical developments.

In the large body of recent work employing molecular markers, the aforementioned two classes of DNA markers are clearly predominant. Taking the new discipline 'phylogeography', for example, as pointed by Avise (1998), some 70% of studies carried out thus far involved analyses of animal mitochondrial DNA. Similarly, although it is difficult to provide an exact figure, the majority of ongoing research projects using nuclear DNA markers involve microsatellite DNA. For instance, among the 1758 primary papers and primer notes published in the last 9 years in the journal *Molecular Ecology* we analysed, 29.8% and 42.5% are indexed with mitochondrial and microsatellite DNA markers, respectively. While mitochondrial DNA has proved to be powerful for genealogical and evolutionary studies of animal populations, and while microsatellite sequences are the most revealing DNA markers available so far for inferring population-genetic structure and dynamics, these markers have some important and unavoidable limitations. These limitations, on one hand, restrict further research development, and on the other hand strongly signal the need for other markers—markers that complement the aforesaid ones and have potential to advance further genetic analyses of populations. Justifiably, single copy nuclear polymorphic (scnp) DNA is becoming the marker of choice.

In this review, we intend to show that, although challenges still exist, recent progress in genomic studies and molecular biology technologies, and increase of statistical power provide a basis for nuclear DNA analyses to be widely applicable. They will allow allele-discriminating characterization of scnp loci and microsatellite loci, thus making genealogical data largely accessible. This will increase significantly our ability to address more difficult population-genetic questions, and lead the genetic analysis of populations into a new and flourishing era. Given limits on space, the complexity of the subject and the wide spectrum of audience, we will not discuss all aspects in detail; rather we attempt to make a balanced treatment on the present status and future trends of this field.

Current DNA markers and their limitations

There are a number of recent reviews dealing extensively with various molecular markers employed in genetic studies of populations, for example the volumes by Avise (1994, 2000), Karp *et al*. (1998), Goldstein & Schlötterer 1999), and journal reviews such as Estoup *et al*. (2002). Interested readers are directed to these sources. Here, we would like to provide a short overview of two classes of most popular DNA markers: mitochondrial DNA (mtDNA, see Box 1) which represents organellar DNA markers based on sequence analysis, and microsatellite DNA, which represents nuclear DNA markers based on fragment analysis. The advantages of these markers are already well known; it is thus more important to outline the problems associated with them.

Mitochondrial DNA markers

Depending on which organism(s) one is working on, mitochondrial DNA can be the cause of fortune, regret or headache. Its applicability is limited largely to metazoan animals. It is good fortune that many metazoan animals have mtDNA that possesses a set of characteristics which makes it an almost ideal molecular marker for evolutionary and population-genetic studies (Avise *et al*. 1987; Moritz *et al*. 1987; Harrison 1989; Avise 1991; Simon 1991). The contribution it has made to our understanding of evolution and natural history is enormous (see Avise 1994 for a fuller discussion).

In plants, except for a few peculiar groups (Palmer *et al*. 2000), the rate of evolution of mtDNA is the slowest among their three genomes (Wolfe *et al*. 1987). However, it manifests a very elevated recombination rate (Palmer & Herbon 1988). Consequently, mtDNA is practically of little use for population-genetic studies, which is regrettable for plant population biologists. The unavailability of suitable DNA markers in plants is the single most limiting factor in making molecular population-genetic studies in plants lag far behind those in animals (Schaal *et al*. 1998).

Nevertheless, its application in animals is not without problems. The recent demonstration of the presence of mitochondrial pseudogenes in the nuclear genome of a wide range of organisms is, for population studies, an unwanted reality (for reviews, see Zhang & Hewitt 1996a; Bensasson *et al*. 2001). Although there exist methods and techniques to reduce the interference of mtDNA pseudogenes in sample preparation and data analysis (see the aforementioned reviews and references therein), they have only limited resolving power. The effectiveness of using mtDNA in population-genetic studies has been greatly weakened by this. It has been shown that mitochondrial-like sequences can exist in high copy number with little difference among members in many organisms, sometimes group-wide as in the acridid grasshoppers and locusts, making mtDNA of little practical value for population-genetic studies in these groups (Zhang & Hewitt 1996b; Bensasson *et al*. 2000). The employment of mtDNA marker

Information Box 1. Terms, abbreviations and acronyms

Synonymous site: sites at which nucleotide substitutions do not cause amino acid changes. Many third codon positions are synonymous sites. Also referred to as silent sites.
Nonsynonymous site: sites at which nucleotide substitutions will lead to amino acid changes or create a stop codon. Also referred to as replacement sites.
Silent site: see synonymous site.
Replacement site: see nonsynonymous site.
Indel: abbreviation of insertion/deletion mutation.
ScnDNA: single copy nuclear DNA.
CpDNA: chloroplast DNA.
MtDNA: mitochondrial DNA.
Numt: nuclear mitochondrial-like sequences (nuclear copies of mtDNA, or nuclear mitochondrial pseudogenes).
Scnp: *s*ingle *c*opy *n*uclear *p*olymorphic (DNA, sequences, loci, markers).
SNP: single nucleotide polymorphism. The common name given to biallelic genetic variation. A piece of DNA may carry several SNPs. It is not a new type of polymorphic marker. Rather, it is a modern term given to a class of well-known DNA polymorphisms. Brookes (1999) define SNPs as 'single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s), wherein the least frequent allele has an abundance of 1% or greater'. SNPs in coding region are sometimes refered as cSNPs, those in noncoding regions as ncSNPs.
(π) the observed average number of nucleotide differences per site between two alleles in populations. Often referred to as nucleotide diversity.
(θ) the average number of segregating sites per nucleotide site based on the expected distribution of neutral variants in a panmictic population at equilibrium. Also referred to as the heterozygosity. This can be called the population mutation rate, because it is a measure of the nucleotide substitution per site per population.
Allele: refers to any variant of DNA sequence observed at a given locus (gene).
Haplotype: refers to the genotype of a set of linked loci on a single stretch of DNA.
D(T)GGE: denaturing (temperature) gradient gel electrophoresis.
SSCP: single strand conformation polymorphism.
RFLP: restriction fragment length polymorphism.
AFLP: amplified fragment length polymorphism.
RAPD: randomly amplified polymorphic DNA.
DHPLC: denaturing high performance liquid chromatography, also referred to as temperature-modulated heteroduplex Chromatography (TMHC), or recently 'DNA chromatography'.
SCAR: sequence characterized amplified region(s).

in such organisms without first being aware of this phenomenon can be a real headache. Ironically, the great value of mitochondrial DNA is nowhere better appreciated than in these situations where it can not be employed, as in cases where numerous nuclear pseudogenes exist very similar to the authentic mitochondrial sequences.

Apart from this taxonomic patchiness of applicability, mitochondrial DNA data on their own have some important limitations. First, mtDNA represents only a single locus. No matter how much effort we have spent on mtDNA as a marker and how much information we have collected from this source, we have looked only through a single window (among countless others) of evolution. This window reflects at best only the matrilineal history (mtDNA in animals is maternally inherited, with some exceptions, see Kondo *et al*. 1990; Gyllensten *et al*. 1991; Skibinski *et al*. 1994), which could well differ from that overall of populations or species. Therefore, the inference we make on species/population history is likely to be highly biased. Second, the effective population size of mtDNA is only a fourth of that of nuclear autosomal sequences, therefore mtDNA lineages have a much faster lineage sorting rate and higher allele extinction rate. The consequences of this are as follows: (i) evolutionary relationships could be oversimplified by mtDNA data; (ii) genetic diversity can be underestimated by mtDNA markers; (iii) uncertainty in genealogical analysis may increase due to the increased probability of more missing links in mitochondrial haplotypes; and (iv) remote population processes may not be detected correctly with mtDNA markers.

Microsatellite DNA markers

The popularity of microsatellite DNA markers among molecular population biologists is not surprising, considering the special features of these markers and the apparent reliability of data produced from them. For use in intraspecific analyses, microsatellite markers have overtaken mitochondrial and other DNA markers currently employed (see above). These simple repetitive sequences are present widely throughout eukaryote organisms

Table 1 Rate of evolution of various DNA markers

DNA markers	Organism	Average rate* (%/Myr)	Reference
Microsatellite DNA			
Dinucleotide (autosomal)	Human	5.6×10^{-4}	Weber & Wang (1993)
Tetranucleotide (autosomal)	Human	2×10^{-4}	Weber & Wang (1993)
	Human	1×10^{-3}	Xu *et al*. (2000)
	Human	2.1×10^{-3}	Brinkmann *et al*. (1998)
Combined (Y chromosome)	Human	2.8×10^{-3}	Kayser *et al*. 2000
Dinucleotide (Y chromosome)	Human	2.04×10^{-3}	Kayser *et al*. 2000
Tetranucleotide (Y chromosome)	Human	3.17×10^{-3}	Kayser *et al*. 2000
Dinucleotide	Green turtle	2×10^{-3}	FitzSimmons (1998)
	Lizard	1.84×10^{-3}	Gardner *et al*. (2000)
	Zebrafish	1.5×10^{-4}	Shimoda *et al*. (1999)
Combined	Alligator	1.73×10^{-3}	Davis *et al*. 2001
Combined	*Drosophila*	6.3×10^{-6}	Schug *et al*. (1997)
Dinucleotide	*Drosophila*	9.3×10^{-6}	Schug *et al*. (1998)
	Drosophila	6.3×10^{-6}	Schlötterer *et al*. (1998)
Trinucleotide (inferred)	*Drosophila*	1.5×10^{-6}	Schug *et al*. (1998)
Tetranucleotide (inferred)	*Drosophila*	1.1×10^{-6}	Schug *et al*. (1998)
Dinucleotide	Ant	7.2×10^{-4}	Crozier *et al*. (1999)
Trinucleotide $(TAA)_n$	Chickpea (var. Ghab 2)	1.0×10^{-2}	Udupa & Baum 2001
	Chickpea (var. Syrian Local)	3.9×10^{-3}	Udupa & Baum 2001
Other Nuclear DNA			
Nonsynonymous sites	Mammals	0.15	Li 1997
	Drosophila	0.38	Li 1997
	Plant (monocot)	0.014	Li 1997
Synonymous (silent) sites	Mammals	0.7	Li 1997
	Drosophila	3.12	Li 1997
	Drosophila	3.0	Rowan & Hunt (1991)
	Plant (monocot)	0.114	Li 1997
Intron	Mammals	0.7	Li 1997, Li & Graur (1991)
3′ nontranscribed region	*Drosophila*	2.0	Rowan & Hunt (1991)
Chloroplast DNA			
Nonsynonymous sites	Plant (Angiosperm)	0.004–0.01	Li 1997
Synonymous (silent) sites	Plant (Angiosperm)	0.024–0.116	Li 1997
Mitochondrial DNA			
Nonsynonymous sites	Plant (Angiosperm)	0.004–0.008	Li 1997
Synonymous (silent) sites	Plant (Angiosperm)	0.01–0.042	Li 1997
	Human	7.8	Horai *et al*. (1995)
Protein-coding regions	Mammals	2.0	Brown *et al*. (1979) Pesole *et al*. (1999)
	Drosophila	2.0	DeSalle *et al*. 1987
COI	*Alpheus* (shrimps)	1.4	Knowlton & Weigt (1998)
D-loop	Human	14	Horai *et al*. (1995)
	Human	17.5	Tamura & Nei (1993)
	Human	23.6	Stoneking *et al*. (1992)
	Human	260	Howell *et al*. (1996)
	Human	270	Parsons & Holland (1998)

*Mutation rate of microsatellite DNA is in unit of mutation per locus per meiosis; estimated here as the average values over several loci. Individual loci may have a much higher mutation rate, for example, 4.2×10^{-2} at the lizard tetranucleotide locus EST2 compared to the average 1.84×10^{-3} (Gardner *et al*. 2000).

investigated, having high levels of polymorphism (high mutation rates, Table 1), obeying Mendelian inheritance and following apparently simple modes of evolution. Their introduction to population-genetic studies has greatly advanced our ability to detect population-genetic structure, to test parentage and relatedness, to assess genetic diversity, and to study recent population history. Without doubt, these markers will continue to dominate related research fields in the foreseeable future.

However, on the whole, although our knowledge of these simple repetitive sequences is better than some 10 years ago when their potential for population-genetic studies was first discussed (Bruford & Wayne 1993), it is in fact still far from complete, in spite of the large number of publications indexed with 'microsatellite(s)'. This is largely because study on the molecular evolution of these sequences lags far behind their application. Their modes and patterns of molecular evolution and their mutation mechanisms need to be addressed by extensive and systematic sequence analysis of microsatellite alleles. Without some firm knowledge of these the use of microsatellite markers will not reach its full potential, and hence their role in the history of population-genetic studies will become secondary. The following issues deserve our attention when employing microsatellite sequences in population-genetic studies:

1 Evolutionary relationships among microsatellite alleles are complicated. Allele size difference may not be directly related to divergence. The assumption of all variation being due to the changes of the copy number of the repeat unit requires careful examination. Homoplasy has been frequently observed, and it is known that allele difference can be produced by both repeat number variation and base variation, with the latter occurring within the repeat regions as well as in the flanking nonrepeated regions (reviewed in Jarne & Lagoda 1996; Culver *et al*. 2001). Although some form of homoplasies (size homoplasy) may not represent a significant problem for estimating certain population parameters in genetic analysis (for review, see Estoup *et al*. 2002), it should not be so assumed a priori, particularly when no complementary data from other types of molecular markers are available for the study systems (Queney *et al*. 2001).

2 Mutation rate varies considerably among organisms and even between varieties (Table 1). Also, substantial interlocus and within-locus variation of mutation rates has been observed (e.g. Brinkmann *et al*. 1998; Crozier *et al*. 1999; Ellegren 2000a; Gardner *et al*. 2000; Xu *et al*. 2000). Because constraints appear to exist on allele sizes (e.g. Garza *et al*. 1995), mutation rate of microsatellite at a given locus is not even among alleles and not constant over time. Both mutation rate and the direction of mutation are found to be affected by allele length (Ellegren 2000a; Xu *et al*. 2000). However, there appears to exist considerable heterogeneity among microsatellite loci, mutation and evolution of these simple sequence repeats are possibly a dynamic and variable process (Brohede *et al*. 2002).

3 The neutrality of some microsatellite sequences is increasingly questionable. Kashi *et al*. (1997) speculated that these simple sequence repeats could be an abundant source of mutations contributing to quantitative trait variation. The conservation of some microsatellite loci across large evolutionary distance (i.e. their antiquity), shown in a number of studies (e.g. FitzSimmons *et al*. 1995; Rico *et al*. 1996; Ezenwa *et al*. 1998; Martin *et al*. 2002), on its own provides strong evidence against selective neutrality of these loci (conservation for up to a billion years has been documented, see Martin *et al*. 2002). While these are intriguing observations, employment of such conserved microsatellite loci as neutral markers in population-genetic studies should be very cautious before the reason for such conservation is understood. Evidence is accumulating rapidly that these simple repetitive sequences are far more complex than thought previously (for review see Ellegren 2000b), and they can be of functional importance. For example, it has been demonstrated recently that a sequence element Z1 containing $(CA/TG)_n$, the most abundant dinucleotide repeats in mammals, forms Z-DNA and inhibits promoter activity in the rat nucleolin (*Ncl*) gene (Rothenburg *et al*. 2001). Z1 has been found to be polymorphic with regard to dinucleotide repeat number. The five allelic variants identified show differences in their inhibitory ability and thus are capable of modulating promoter activity differently. It is also interesting to note that the Z1 element exhibits a high degree of conservation between rodent species (Bourbon *et al*. 1988; Rothenburg *et al*. 2001).

A notable shortcoming of the present microsatellite data is that ancestral information they contain is often ambiguous (multiple state allele), therefore genealogical patterns of relationships cannot be deduced with certainty. Data produced by other size-based methods, such as RFLP (restriction fragment length polymorphism), RAPD (randomly amplified polymorphic DNA), AFLP (amplified fragment length polymorphism), SSCP (single strand conformation polymorphism) and D/TGGE (denaturing/temperature gradient gel electrophoresis), have similar weakness. That is, like allozyme allele data, they are unordered, hence genealogies cannot be inferred easily.

An important task of genetic analyses of populations is to infer their evolutionary history and to deduce the consequence or impact of related demographic, ecological and climatic changes. To fulfil this, DNA sequence data are required. Sequence data available so far are almost exclusively mitochondrial, genealogical data from nuclear loci

Table 2 Polymorphic levels of nuclear DNA markers observed in various organisms

Organism	Nuclear locus	Nucleotide diversity (%)*	References
Human	49 loci	0.1	Li & Sadler (1991)
	Genome-wide	0.075	International SNP Map Working Group (2001)
Dolphins	4 loci (introns)	0.05–0.22	Hare *et al*. 2002
Birds	22 loci	0.23–0.25	Primmer *et al*. 2002
Fruit flies	24 loci	0.403 (c), 1.05 (nc)	Moriyama & Powell (1996)
D. melanogaster			
D. simulans	12 loci	0.799 (c), 1.877 (nc)	Moriyama & Powell (1996)
C. capitata	4 loci	0.727–1.477	Villablanca *et al*. (1998)
Arabidopsis thaliana	Various loci	0.21–1.09	Innan *et al*. (1996) Kawabe *et al*. (1997) Purugganan & Suddith (1998), (1999) Miyashita 2001 Le Corre *et al*. 2002
Rice	Genome-wide	0.67	Yu *et al*. (2002)
Pine	*Pal1*	0.17	Dvornyk *et al*. 2002

*Nucleotide diversity (%) refers to π (or θ) × 100; ‘c’ denotes ‘coding region’ and ‘nc’ ‘noncoding region’. Values shown in the table are values averaged over loci, except where a range is given.

are lacking. As we can see in the next section, single copy nuclear polymorphic sequences (hereafter referred as scnp sequences) exist commonly in the nuclear genome, ample fuel for a boom in population-genetic studies.

Scnp markers: availability, patterns and rate of evolution

Availability

Data from genome projects and pioneer studies in model organisms provide excellent examples, demonstrating the availability of scnp markers in the nuclear genome. The nucleotide diversity of the human nuclear genome is around 0.1% (Li & Sadler 1991; International SNP Map Working Group 2001). This means that there are approximately three million nucleotide differences between any two individuals. For population studies, where usually a large number of sequences are sampled, the expected percentage of polymorphic sites is 0.2–0.5% (Fu & Li 1999). Considering the relatively young history of modern human populations (less than 200 000 years from the most recent common ancestor, or less than 10 000 generations), the above value is encouraging. It raises confidence in the common availability of scnp loci in the nuclear genome of any eukaryotic organism. Indeed, other genome projects and population-genetic studies carried out in other model organisms, such as rice, *Drosophila* and *Arabidopsis*, confirm this finding. Data from the rice genome sequencing project revealed that the overall polymorphism level in rice is 0.67% (Yu *et al*. 2002). From a pooled analysis of available *Drosophila* nuclear DNA data, Moriyama & Powell (1996) concluded that for genes located in normal genomic regions the ‘amount of variation is sufficiently large that virtually every diploid individual is heterozygous at every locus’. The nucleotide diversity in *Drosophila* is about one order of magnitude higher than that in human. Even for the highly selfing weed *Arabidopsis thaliana*, a mean nucleotide diversity of 0.74% (π varies between 0.0021 and 0.0104) at several nuclear loci is observed (e.g. Innan *et al*. 1996; Kawabe *et al*. 1997; Purugganan & Suddith 1998, 1999; see Table 2), although genome-wide the frequency of SNPs is one in every 3.3 kb (International SNP Map Working Group 2001). Data on chromosome 1 of maize indicated that on average the frequency of SNP is one in every 104 base pairs (bp) between two randomly sampled sequences (Tenaillon *et al*. 2001). Our work on the desert locust *Schistocerca gregaria*, a highly migratory insect, employing a 3′ non-coding scnp marker, revealed that about 7% of nucleotide sites are segregating (Zhang & Hewitt 1996c; unpublished data). Table 2 summarizes polymorphic levels of nuclear DNA observed in various organisms.

Patterns of nuclear DNA variation

The data available so far show clearly that the distribution of polymorphic sites is not random along the nuclear genome nor within a gene. This unevenness is associated with differences in recombination rate, gene density in the genomic region, transmission pattern, selection strength and compositional pressure. Genomic regions with low recombination rates generally have reduced levels of polymorphisms (Begun & Aquadro 1992; Nachman *et al*.

Information Box 2. Classification of genic regions

A gene, from 5′ to 3′ ends, can be divided into different parts according to their function or location.

Exon: refers to the part(s) that can be transcribed into RNA. It may encode peptide, transfer RNA (tRNA) or ribosomal RNA (rRNA).

Intron: refers to the part(s) between exons. During RNA maturation, introns are removed from the precursor RNAs. Introns are mainly present in eukaryotic genes, and occasionally are found in prokaryotic (viral or organellar) tRNA and rRNA genes.

Non-coding region: refers to the total noncoding parts, i.e. the 5′ and 3′ noncoding sequences plus introns. It does not code for mRNA, rRNA, tRNA, etc. Parts of the first and the last exons are usually noncoding.

Transcribed region: refers to the parts that are transcribed to RNA sequences. Introns, parts of the 5′ and 3′ noncoding regions, as well as exons, constitute the transcribed region.

Nontranscribed region: also known as untranscribed region. Refers to the 5′ and 3′ noncoding sequences flanking the transcribed region. 5′ nontranscribed region sometimes are further divided into distal, central and proximal parts.

Translated region: refers to the part(s) that will be translated to peptide(s), i.e. beginning with the initiation codon ATG and ending at the stop codon. Exons are translated regions, except for parts of the first and the last exons in the gene.

Nontranslated region: usually refers to the 5′ and 3′ parts that are transcribed but not translated. It consists of 5′ untranslated region which is the transcribed part before the initiation codon ATG, and 3′ untranslated region which is the transcribed sequence after the stop codon.

Intergenic region: refers to genomic sequences between two neighbouring genes. This is usually defined as regions located more than 5 kb away from any (predicted) transcription unit.

1998; Nachman 2001; Lercher & Hurst 2002). Payseur & Nachman (2002) showed that nucleotide diversity and gene density in the surrounding genomic regions are negatively correlated in humans. Regions subject to strong balancing selection, such as genes involved in the immune defence system (e.g. the MHC loci, also known as HLA loci in man) or loci involved in disease resistance show the greatest diversity (Noël *et al*. 1999; Arabidopsis Genome Initiative 2000; International SNP Map Working Group 2001). Sex chromosomes appear to vary the least in man and *Drosophila*, due to possibly a combination of low recombination rate, reduced mutation rate and reduced effective population size (Moriyama & Powell 1996; International SNP Map Working Group 2001; Venter *et al*. 2001). Therefore, it is difficult to draw a general conclusion on the potential variability of different genomic or genic regions (intergenic, coding, noncoding introns and noncoding flanking regions, see Box 2) without considering the function and genomic location of the sequences in question.

While limited largely to the few model organisms, including man, *Drosophila* and *Arabidopsis*, scnp data from population-genetics studies shows that the substitution rate at synonymous sites is often higher than that of introns, and the rate of introns is higher than that of the immediate nontranslated flanking regions (Boxes 1 and 2). This characterization of rate differences sometimes leads to incorrect interpretation of such observations in the literature. An important issue concerning intraspecific variability of introns deserves clarification. Higher substitution rate, or higher nucleotide diversity at synonymous sites than in introns is often erroneously taken as evidence for less variability in introns than in nuclear coding regions (see review Caterino *et al*. 2000: 17). We explain here why this is not so.

For biologists studying molecular evolution of DNA sequences, the term ‘per site substitution rate’ (or nucleotide diversity in an intraspecific context) is employed as a measure of the rate of evolution. To facilitate comparison among different genic regions, genes and taxa, silent site substitution rate is used because these sites are apparently neutral or nearly neutral. However, for population biologists who are interested in comparing genetic variations among individuals and populations and who are looking for pieces of DNA capable of providing as much variation as possible with the minimum expense, silent site substitution rate is a quite obscure measure, because it cannot reflect the overall genetic variability of the region involved. The reason for this is as follows. For any normal protein coding region, the number of silent sites in the total sequence is always much smaller than the number of nonsilent sites. A reasonable ratio of silent sites to nonsilent sites can be expressed as 1:2.5 (in a random protein coding sequence, assuming that there is equal use of all codons and that at twofold degenerate sites, 50% of the changes are silent, the ratio is 1:2.66, see Ochman & Wilson 1987). Thus, in coding regions only about 28% of sites are the effective ‘silent sites’. Therefore, given the same length of the coding region as intron, unless the substitution rate at silent sites are more than 3.5-fold higher than that of intron, the coding

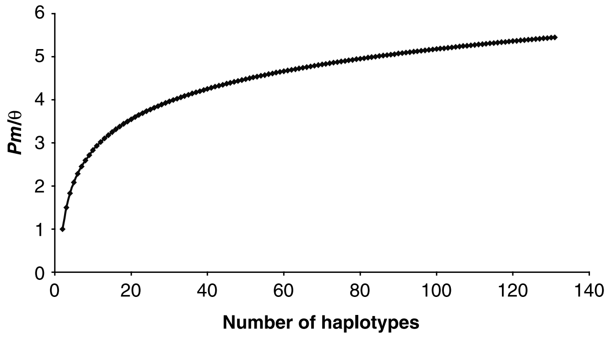

Fig. 1 Effect of the number of haplotypes sampled from populations on the correlation of *Pm* (the 'polymorphicity') and θ (the estimated population mutation rate, or the heterozygosity). See text for the definition of *Pm*.

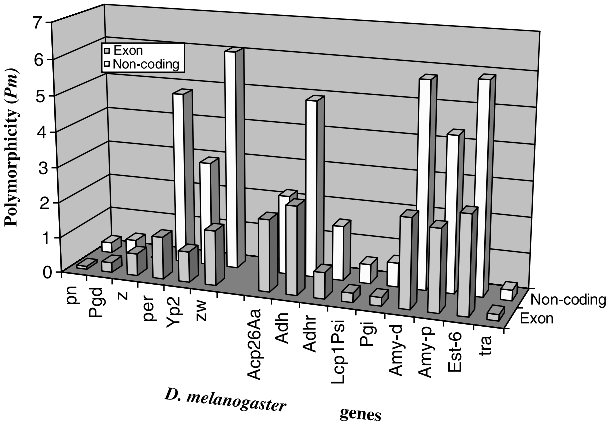

Fig. 2 Polymorphicity (*Pm*) of exon and noncoding regions of 15 *D. melanogaster* genes, of which six are X chromosome-linked genes and nine autosomal genes (data modified from Moriyama & Powell 1996). *Pm* of noncoding regions (light columns), which is largely contributed by introns, is clearly higher than that of exons (dark columns).

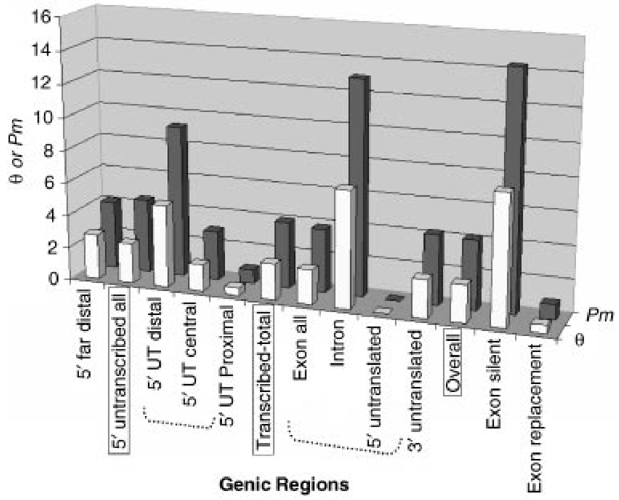

Fig. 3 Polymorphicity (*Pm*) and the population mutation rate (θ) of various genic regions of the *Est-6* gene of *Drosophila simulans* (data modified from Karotam *et al*. 1995). Columns in the front row are θ-values; columns at the back are *Pm*. It can be seen that *Pm* and θ are highly correlated. As shown in the figure, from left to right, a gene can be divided into 5′ far distal region, 5′ untranscribed region (which can further be fragmented to distal, central and proximal regions), and transcribed region (which comprises exons, introns, 5′ untranslated and 3′ untranslated regions) (see Box 2). 'Overall' refers to a gene as a whole, 'exon silent' and 'exon replacement' refer to the total of silent sites and replacement sites in all exons, respectively.

region is unlikely to harbour more polymorphic sites than the intron (of course subject to statistical error). Such a high rate at silent sites has hardly ever been observed. This means that, although silent site substitution rates of coding regions are often higher than that of introns, their overall variability is often lower than that of introns.

To simplify the issue, we propose to use the mean percentage of polymorphic sites observed in a DNA region as a rough measure of the potential variability, i.e. the 'polymorphicity' (denoted as *Pm*) of a DNA region. Clearly *Pm* is sensitive to sample size and the populations sampled, and is a measure of the frequency of segregating sites in populations. The relationship between *Pm* and the estimated population mutation rate (or heterozygosity, θ, see Box 1) can be expressed as

$$Pm = (\theta\cancel{/}100)\cdot a_{h-1}$$

where a_{h-1} is a modifier whose value depends on the number of haplotypes (alleles) observed (h), i.e.

$$a_{h-1} = \sum_{i=1}^{h-1} (1/i)$$

(Watterson 1975). Figure 1 shows the correlation of *Pm* and θ under various number of haplotypes sampled from populations. It can be seen that *Pm* is a fair indicator of the variability of a DNA region. Using this measure, the patterns of sequence variation of nuclear genes from various organisms are illustrated in Figs 2–4. Figure 2 shows the polymorphicity of exon and noncoding sequences from 15 *D. melanogaster* genes, of which six are X chromosome-linked genes and nine autosomal genes (data modified from Moriyama & Powell 1996). Two features emerge from this comparison: first, there is evident heterogeneity among genes for variability for both coding and noncoding regions, and these two are correlated, as noticed by Moriyama & Powell; second, in the majority of cases (14 of 15), polymorphicity (*Pm*) of noncoding regions is higher than that of exons. Because for most loci analysed here variability of the noncoding regions overall is largely contributed by introns, it suggests

doi: 10.1046/j.1365-294X.2003.01887.x

Corrigendum

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

On page 570, the correct formula describing the relationship between *Pm* and the estimated population mutation rate (or heterozygosity, i.e. the estimated average number of segregating sites per nucleotide site, θ) should be

$Pm = (\theta * 100) * a_{h-1}$

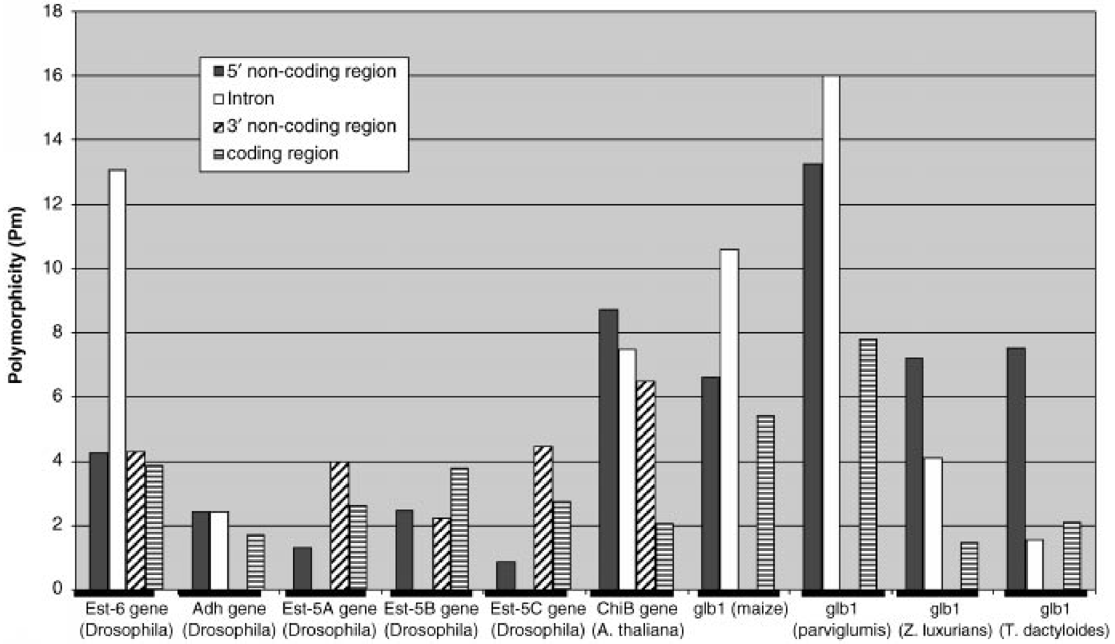

Fig. 4 Among gene variation of the relative variability (*Pm*) of different nuclear noncoding regions. Compared here are the 5′ noncoding region, introns, 3′ noncoding region and coding regions (data based on Kreitman & Hudson 1991; Karotam *et al*. 1995; Hilton & Gaut 1998; King 1998; Kawabe & Miyashita 1999).

that introns have a higher polymorphicity than exons. This is true for other *Drosophila* loci as well, such as the *runt* locus (Labate *et al*. 1999), *Est-5* locus (King 1998), *Est-6* and *Sod* loci (Karotam *et al*. 1995; Hudson *et al*. 1997; Balakirev *et al*. 1999), *Adh* locus (Kreitman & Hudson 1991), *Tpi* locus (Hasson *et al*. 1998), and also the *AdhA* locus in cottons (interspecific comparisons, Small & Wendel 2000), *ChiB*, *AP3* and *PI* loci in *Arabidospsis thaliana* (Kawabe & Miyashita 1999; Purugganan & Suddith 1999), *glb1* locus of maize (Hilton & Gaut 1998) (see Fig. 4). Large-scale analyses of the distribution frequency of SNPs in the *Arabidopsis*, human and rice genomes strongly support this conclusion: SNPs are found much more frequently in introns than coding regions, or intergenic regions (Arabidopsis Genome Initiative 2000; Venter *et al*. 2001; J. Wang, personal communication).

Figure 3 contrasts the polymorphicity and the population mutation rate (θ) of various genic regions using the *Est-6* gene of *Drosophila simulans* as an example (data modified from Karotam *et al*. 1995). It can be seen that for this locus the intron is the most variable part, followed by the 5′ distal untranscribed region, then the 5′ far distal untranscribed region and 3′ untranslated region, then the exons, and with the 5′ proximal untranscribed and 5′ untranslated noncoding regions being the least variable parts. Again, it is shown here that *Pm* and θ are highly correlated. The relative variability of various noncoding regions manifested in the *Est6* locus is, however, not universal; it varies considerably from gene to gene, as shown in Fig. 4.

The above data, taken together, suggest that introns and 5′ and 3′ flanking noncoding regions are also under varying degrees of functional constraint. Introns do not appear to be just nonfunctional junk sequences (Duret 2001). For example, it is well known that introns contain signal sequences important for splicing and even for regulation of transcription (for reviews, see Mattick & Gagen 2001; Zieler & Huynh 2002). Promoters and other regulatory elements are present at the 5′ flanking region of a gene, and sequences at the 3′ noncoding region are necessary for post-transcriptional processing, transcription and translation regulation, and/or stability of mRNA (Ross 1996; Wickens *et al*. 2002). Nevertheless, compared to the coding regions, in general introns and distal flanking regions could serve as reasonable markers for population-genetic analyses. Introns in particular, because of their general higher polymorphicity and the ease of primer design in the flanking coding sequences (normally well conserved among closely related taxa), will be probably the most used nuclear markers in the coming years.

Rate of evolution of nuclear DNA

The slow rate of evolution of nuclear DNA has often been considered a factor limiting its use in intraspecific studies. It is known widely that in many vertebrates the rate of evolution of single copy nuclear sequences is much lower than that of mitochondrial DNA (Brown *et al*. 1979, 1982). However, this is not a general rule. It has long been recognized in plants that nuclear DNA is the fastest evolving among the three genomes they harbour (Wolfe *et al*. 1987). In animals other than mammals, the rate of evolution of nuclear DNA is not in general lower than mtDNA. For example, the substitution rates of mtDNA and scnDNA in fruitfly do not appear to differ significantly (see Table 1; Caccone *et al*. 1988), possibly due to an increased substitution rate in the nuclear genome, which is eight times or more higher than that of primates (Sharp & Li 1989; Moriyama & Gojobori 1992). Table 1 lists the rate of evolution of various DNA markers currently employed in evolutionary studies.

It should be emphasized that substitution rate estimated from interspecific comparisons may not reflect the intraspecific variability of a given DNA region, and often it is an underestimate of that variability. The reason for this is simple, in the context of the neutral or nearly neutral theory of molecular evolution (Kimura 1983; Ohta 1992). Most interspecific differences represent fixed differences among the taxa compared. They are the final products of selection and drift on mutations, hence they represent only a small fraction of the mutations that ever occurred. In contrast, intraspecific polymorphisms represent an earlier transitional stage in the evolution of species. Many or most of the intraspecific differences observed are still on their way to fixation or extinction. That is, there has been less time for selection and drift to shape genetic differences in populations. Therefore, intraspecific polymorphisms represent a closer view of the raw mutations occurring in the populations. This leads us to think that intraspecific polymorphisms may be much more pronounced than usually thought.

Challenges faced in the applications of nuclear DNA markers

Nuclear DNA polymorphisms that exist widely in eukaryotic organisms provide virtually unlimited opportunities for studying the mechanisms of evolution. However, when working with nuclear DNA markers, we face challenges at almost every stage of a study. These include recombination, selection (non-neutrality), heterozygosity, insertion/deletion polymorphism, low divergence and polytomy, gene-specific variation in rate and history, PCR and sequencing difficulty, and so forth. Some of these do not occur with mitochondrial DNA markers and are therefore specific to nuclear markers.

1. Recombination

Recombination occurs frequently in nuclear genomes. Recombination rate varies considerably from locus to locus, influenced by various factors such as the chromosomal location of the locus involved or the structural features of the sequence. The information on evolutionary histories carried by DNA sequence haplotype data will be distorted by recombination, and if ignored will produce false inferred history. Where recombination had occurred, the evolutionary history of the sequences splits into a group of trees instead of being represented by a single tree (Posada 2001; Wiuf *et al*. 2001). Compared to the rate of spontaneous mutations (substitutions, small insertions and deletions), recombination rate is significant (e.g. $\approx 4 \times 10^{-7}$ for maize, Wang *et al*. 1999). It is not unusual to find regions where recombination rate is higher than mutation rate (Morgan & Strobeck 1979; Hilton & Gaut 1998; Wang *et al*. 1999). There are two kinds of strategies to deal with recombination. One is to employ nuclear regions with very low recombination rate, such as the fourth chromosome of *D. melanogaster* and the NRY region of human Y chromosome, or regions near centromeres and telomeres. However, it has been shown in *Drosophila*, mammals including man, and plants that there is a positive correlation between recombination rate and the level of polymorphism (Moriyama & Powell 1996; Dvorák *et al*. 1998; Stephan & Langley 1998; International Human Genome Sequencing Consortium 2001; reviewed in Payseur & Nachman 2002; but see also Payseur & Nachman 2000). This is probably a general phenomenon in eukaryotic organisms. Therefore, genomic regions with low or no recombination may not provide enough variation to address many questions.

The second strategy is to detect recombination events from the data set and then incorporate recombination into models of evolution during the analysis of the data. This is likely to become the mainstream of nuclear data analysis, as it allows a fuller exploration of the nuclear DNA resources. A body of literature already exists on this subject, developing a number of recombination detection methods (e.g. Stephens 1985; Hudson & Kaplan 1988; Maynard Smith 1992; Hein 1993; Templeton & Sing 1993; Innan *et al*. 1996; Maynard Smith & Smith 1998; Wiuf & Hein 1999; McGuire *et al*. 2000; Schierup & Hein 2000; Xu 2000; Strimmer *et al*. 2001; reviewed in Posada 2002). David Robertson at the University of Oxford has set up a website with links to many recombination detection programs available so far (*http://evolve.zoo.ox.ac.uk/Grinch/RAP_links.html*). The relative performance of different methods for detecting recombination from DNA sequences has been evaluated recently by several groups (Maynard Smith 1999; Wiuf *et al*. 2001; Posada 2002).

2. Selection (non-neutrality)

Levels and patterns of nucleotide variation of DNA sequences will be affected not only by their functional importance, but also by their genomic background. The aforementioned correlation between genetic diversity and rates of recombination is in fact the result of the interplay between selection and recombination. Population-genetics theory predicts that natural selection is less effective in regions with low recombination. In such regions, linkage between nucleotides is tight and a large region tends to act as a single unit of selection. Consequently, selection for or against a few, or even a single nucleotide substitution will affect the whole region involved. This is the well-known hitchhiking effect (Maynard Smith & Haigh 1974; for deleterious mutations, it is referred as background selection, Charlesworth *et al*. 1993). In contrast, in regions with high recombination rates, nucleotide sites or blocks of nucleotides serve as independent selection units. Consequently, selection on a single nucleotide site or a block of sites affects only those sites tightly linked to it and will have a much reduced regional effect, allowing an overall higher substitution rate. The implication of this selection-recombination interplay is important. It means that the apparent neutrality of a DNA sequence is not simply determined by its functional importance to the organism. Regions with no function, if tightly linked to a functionally important gene, will not follow the patterns of neutral evolution. Thus caution should be exercised when analysing data from nuclear loci, knowledge on the adjacent chromosomal regions is equally important for correct interpretation of the data.

3. Insertion/deletion polymorphism

A large proportion of nuclear DNA markers employed for population analysis would be noncoding regions, because they are usually more variable than the coding regions (in contrast to mitochondrial DNA markers where coding regions are mostly used because of the lack of introns). Apart from nucleotide substitutions, insertions/deletions (indels) often constitute a large part of the polymorphism detected. Thus, patterns of indels (or gaps in the aligned sequences) contain a sufficiently large amount of phylogenetic information that they should not be ignored. However, most phylogenetic inferring methods available so far do not use such gaps efficiently. Gaps are either ignored or treated as ambiguities. The use of information contained in gaps is an important goal for the full interpretation and exploration of polymorphism data from nuclear markers. Several groups have explored the use of gaps in sequence-based phylogenetic analyses, for example by treating them as a fifth character (Swofford 1993; Simmons & Ochoterena 2000; McGuire *et al*. 2001, and references therein); however, the underlying evolutionary processes responsible for indels are different from those for substitutions. It remains a challenge to work out how to properly handle indel characters.

4. Low divergence and polytomy

Two other features of intraspecific variation require special attention in nuclear data analysis. First, the level of intraspecific divergence is low. Although polymorphism exists in the nuclear genomes of eukaryotic organisms, the pairwise sequence divergence is in general small, being no more than 1–2% in most cases. Second, as summarized recently by Posada & Crandall (2001), intraspecific DNA evolution does not in general follow a bifurcating process; instead multifurcation is the norm, with the additional complications of reticulate relationships produced by recombination, parallel mutations and recurrent mutations (Bandelt *et al*. 1995; Smouse 1998, 2000). Therefore, traditional phylogenetic analysis methods are not the most appropriate for analysing intraspecific polymorphic data. For instance, with the low level of divergence, traditional phylogenetic analysis methods cannot resolve unambiguously the evolutionary relationships in the data set. Also, bootstrapping, the method commonly employed to assess cluster reliability, loses its power. For example, given an average pairwise divergence of 1%, one is unlikely to obtain a high bootstrap support value from a real data set. In addition, rare polymorphisms, such as singletons, are treated as noninformative from a parsimony perspective; a singleton is probably, however, an important signature of recent population expansion. With intraspecific data, allele divergence, allele frequency and allele number are key parameters which reflect different aspects of populations and their evolutionary history. For neutral loci, allele divergence and the associated genealogy reflect evolutionary events that occurred over a long period of time, and even from before speciation; allele frequencies are predicted to be correlated to the ages of alleles under the coalescent theory; the number of alleles in populations should be positively correlated to the sizes of populations. Clearly, traditional statistics (Wright *F*-statistics) alone are not sufficient to reveal all the information contained in intraspecific polymorphic nuclear DNA data. A number of strategies have been explored to more fully use such information, including the nested clade analysis method (Templeton *et al*. 1995; Templeton 1998) and the network analysis method (Bandelt *et al*. 1995, 1999). These methods, in combination with coalescent analysis, appear to be able to overcome some of the major hurdles (such as multifurcation, recombination) of the traditional phylogenetic analysis techniques, and capable of recovering more precisely patterns of past population processes (Hare 2001).

5. History of gene and history of populations

DNA markers are employed to reveal genetic structure, evolutionary history and the potential for evolution of populations. In reality, what has been inferred from the data is the structure and history of the DNA regions used. This is the so-called 'gene tree and species tree' problem (Avise 1989). For nuclear DNA, the problem is further complicated by the fact that different genomic regions may have different rates of evolution and have been shaped differently by different evolutionary forces. While some of the forces (e.g. random drift, population bottleneck) have a genome-wide effect, others (e.g. recombination, selection) may have only a regional influence. For a given nuclear marker, the correlation of gene tree with species tree (population tree) depends on the interplay among these forces. Where selection or recombination exerts a preponderant effect the recovery of the correct gene tree is itself problematic, not to mention the population tree. At present, except for a few model organisms, knowledge on the evolutionary and genetic contexts of the genomic regions employed as molecular markers is usually not available. The best strategy for most organisms is to carry out analysis using multiple independent nuclear markers. A coherent pattern between independent loci would suggest a more reliable inference of the population history.

6. Heterozygosity and allele discrimination

For outcrossing diploid organisms, heterozygosity is common in the nuclear genome. Individuals heterozygous at a given locus have two different alleles (haplotypes). Except in some particular situations (for examples, where sex-specific DNA markers are available, such as human Y-chromosome loci, or where haploid tissue can be used as the source of DNA, such as the megagametophytes of gymnospermous plants), perhaps the biggest technical challenge in the employment of nuclear DNA markers is to separate these alleles in the analysis, so that allele-specific characterization can be carried out. This problem (haplotype determination) will be addressed fully in the next section.

Current approaches for haplotype determination in nuclear DNA analyses

For most organisms studied so far, the nuclear DNA markers employed in population-genetic analysis are almost exclusively microsatellite DNA and ribosomal DNA (rDNA). This is simply because haplotype determination can, in practice, be performed easily for these markers. For example, microsatellite variation is considered to be largely due to copy number variation of the repeat unit, hence alleles from heterozygous individuals can be distinguished on a denaturing polyacrylamide gel; rDNA, although highly repetitive, is thought in general to follow concerted evolution, thereby being effectively haploid in nucleotide sequence (NB: both microsatellite homoplasy and nonconcerted evolution of rDNA have been documented, e.g. Schlötterer & Tautz 1994; Vogler & deSalle 1994; Buckler *et al.* 1997; Onyabe & Conn 1999; Culver *et al.* 2001). Single copy nuclear DNA or low copy nuclear DNA markers have not been widely employed for assessing intraspecific variation because of both the limited availability of such markers in most organisms and the difficulty for haplotype determination due to heterozygosity. Although not all evolutionary analyses require the phase of haplotype to be determined, such information is essential for the inference of many population processes. Various attempts to overcome the latter problem have been made, with some being more general than others. We summarize below reported techniques, experimental and statistical, for haplotype determination of nuclear markers.

Experimental approaches

Cloning of PCR products. While this is a universally applicable method for determining unambiguously the sequence of each haplotype of heterozygous individuals, some serious drawbacks render it practically less attractive. In particular, it is costly and labourious, and requires the analysis of several independent clones of each PCR product in order to reduce polymerase-replication errors and to pick up poorly represented alleles. Additionally, artefacts can be produced during the cloning step due to *in vitro* recombination upon transformation of bacterial cells with heteroduplex DNA (Tang & Unnasch 1995). Study cases using this strategy can be found in Palumbi & Baker (1994), Duda & Palumbi (1999), and LaForest *et al.* (1999).

Signal-intensity dependent inference. The two alleles at a locus in heterozygous individuals are often amplified with different efficiency during PCR (due to a template drift effect and/or replication advantage of one of the alleles), and thus are represented in different proportions in the final product. The less-represented allele should, as a consequence, give weaker signals at the heterozygous sites on sequencing gels. Some researchers rely on this principle to infer the two allelic sequences from a heterogeneous sequence. However, experience tells us that this is not a reliable method if the two alleles differ at more than two sites. This is because during the sequencing reaction signal intensities of the two alleles at each heterozygous position are subject to independent sampling variance. Therefore, a weaker signal at the first heterozygous position for allele A does not guarantee that its signal at, for instance, the fourth heterozygous position will also be weaker.

Using the 'allele-dropout-effect'. Template selection in a PCR reaction is a stochastic process apparently following the 'threshold model', which posits that template molecules must reach a threshold number for them to be amplified efficiently (Teng *et al.* 2001). If more than one templates exists, only the one above the threshold can be reliably amplified. Based on this, alleles from heterozygous individuals could be stochastically separated by carrying out PCR from suitably diluted genomic DNA template. Once one allele has been unambiguously determined the other allele can be inferred easily. The extreme of this method is to dilute genomic DNA to the extent that a given aliquot contains only a single molecule of the desired region. This has been shown to be feasible and effective (Ruano *et al.* 1990; Taberlet *et al.* 1996). We refer to this phenomenon as the allele-dropout-effect. Although this allele-dropout-effect can be used to achieve haplotype separation, it has limitations for population analyses, particularly when the sample size is large. Considerable time and effort is needed in order to resolve each sample, and overall the practice will not be cost-effective.

Allele-specific amplification. Taking advantage of the critical importance of perfect matches between the most 3′-end base of PCR primers and their templates, one can design allele-specific PCR primers to selectively amplify only one allele (e.g. Ruano & Kidd 1989). Although this may be an effective method for clinical screening of a given genetic variation, for population-genetic studies it is not practical. It requires previous information about allelic sequences. In addition, for highly polymorphic markers, many site-specific primers are needed, and even this does not guarantee that all haplotypes can be detected.

Physical isolation of hemizygous templates. Using flow cytometry, specific chromosomes can be sorted and thus used as hemizygous templates for PCR. Here only one allele is amplified, allowing allele-specific characterization. Chromosome sorting has been used in various applications, including chromosome-specific genomic library construction, chromosome painting, karyotyping, mapping of genes and genetic markers, etc. (for reviews, see Carter 1994; Dolezel *et al.* 1994). The use of this technique in population-genetic study faces the problem of large sample size, and is very demanding in techniques and equipment.

Genetic isolation of haplotype. In model organisms, such as *Drosophila*, laboratory breeding has been routinely used to establish isofemale lines which become effectively homozygous. Therefore, for these organisms haplotypes can be easily determined by direct sequencing of PCR amplified genomic regions (e.g. King 1998).

Haplotype separation by SSCP, DGGE and related techniques. The high resolving power of SSCP (single strand conformation polymorphism) and DGGE (denaturing gradient gel electrophoresis) for the detection of DNA sequence variation allows these and related techniques to serve as effective tools to separate heterozygous alleles on polyacrylamide gels. This has been demonstrated in several studies (Orti *et al.* 1997a, 1997b; Aldridge *et al.* 1998; Bagley & Gall 1998). While these techniques are not so difficult to set up in established laboratories (Sunnucks *et al.* 2000), they are nevertheless a challenge for many groups. In addition, employing these techniques to isolate alleles involves much manual work with toxic chemicals, and is prone to cross-contamination. Although offering the desired results, they may not be the ideal techniques for carrying out large-scale analyses.

Allele-specific sequencing using restriction enzyme and biotinylation (ASSURE B). Taking advantage of both the specific cleavage of heterozygous sites by restriction enzymes and the solid phase molecule separation technique offered by the biotin/magnetic streptavidin technology, the ASSURE B method allows the specific isolation of one of the alleles present in a heterozygous PCR product (Zhang & Hewitt 1996d). This method is rapid and simple, if suitable restriction sites are available. It works better for highly polymorphic markers. However, it requires polymorphic restriction site information. Furthermore, on its own it cannot generally resolve all haplotypes. Two factors may affect the cleavage efficiency when employing this technique (D. X. Zhang, unpublished data): (i) ratio of the two PCR primers and (ii) heteroduplex formation. Unbalanced primer ratios promote the production of single-stranded DNA in PCR products, and heteroduplexes may result in imperfect pairs at restriction sites, both leading to incomplete cleavage by restriction enzymes.

Resolving haplotypes with DHPLC (denaturing high performance liquid chromatography). An emerging technology, DHPLC (Oefner & Underhill 1995; Underhill *et al.* 1996, 1997), has a greater potential for effective nuclear DNA haplotype separation. DHPLC is an ion-pair reverse phase HPLC method. For mutation detection, the essence of DHPLC is heteroduplex analysis. Because of mismatches present in heteroduplexes, they have (i) different electrophoretic mobilities and (ii) different melting properties from homoduplexes. Therefore, under controlled experimental conditions, the presence of heteroduplexes can be detected. Unlike conventional heteroduplex analysis, which only takes advantage of the first property of heteroduplex molecules to distinguish heterozygous and homozygous, the technique of DHPLC can exploit both properties of heteroduplexes mentioned above and perform analysis under denaturing, partial denaturing or nondenaturing conditions, thus allowing size-dependant, conformation-dependant, or sequence-dependant separations. Therefore, the two alleles in a heterozygous sample can be separated and differentially recovered for subsequent analysis. It has been reported that DHPLC is able to discriminate single-stranded DNA molecules of identical size (< 100 nucleotides) that differ in a single base (Oefner 2000), and single base differences in double-stranded DNA fragment as large as 1.5 kb can be detected (O'Donovan *et al*. 1998).

The most prominent benefits of DHPLC technique are that (i) it is a highly automated technology (automation includes sample injection, gradient set-up, detection, sample elution and recovery), thus as long as DNA fragments can be differentiated, they can be efficiently and separately recovered, and (ii) recovered samples can be used directly for subsequent analyses such as PCR or sequencing. More studies in depth are needed to explore the potential of DHPLC for haplotype separation. If confirmed, it is likely to be the technique-of-choice for nuclear DNA analyses.

Statistical approaches

In theory, it is not necessary to resolve experimentally every haplotype drawn from populations; statistical procedures exist for inferring the unresolved ones from the determined haplotypes.

Clark (1990) proposed a parsimony algorithm to infer haplotype sequences. This approach does not require the physical separation of the two alleles from heterozygous individuals before sequencing. Where reliable, this is a shortcut to circumvent the problem. It works on sequences obtained from direct sequencing of PCR products of diploid individuals. When homozygous, the sequences are unambiguous; when heterozygous, the sequences contain ambiguities at positions where the two alleles differ in sequence. The algorithm uses the unambiguous sequences from homozygous individuals and/or single site heterozygotes as reference sequences to extract allelic sequences from the heterozygous sequences; it then uses the newly inferred allelic sequences as reference to extract more allelic sequences from the remaining ambiguous sequences until it reaches a point where all sequences have been resolved or no further advance can be made. Olsen & Schaal (1999) employed this method in their phylogeographical study of cassava (*Manihot esculenta*); Antunes *et al*. (2002) implemented it to determine the transferrin gene haplotypes in the Brown trout.

Several expectation–maximization (EM)-based algorithms were introduced recently (e.g. Excoffier & Slatkin 1995; Hawley & Kidd 1995; Long *et al*. 1995; Schaid *et al*. 2002) to carry out a maximum-likelihood estimation of haplotypes and haplotype frequencies. Like Clark's algorithm, they are based on the assumption of Hardy–Weinberg equilibrium; therefore, both methods should perform better for large samples. Factors affecting the performance of these algorithms include departure from Hardy–Weinberg equilibrium, the average heterozygosity per nucleotide site, the number of variable sites, frequency distribution of haplotypes, sample size, recombination rate and sequence reliability, with the last two factors being the most difficult to deal with. For example, practical experience tells us that when directly sequencing PCR products from heterozygous individuals, intensity of signal at heterozygous sites is not always strong enough for both bases to be called. If extreme bias for nucleotide composition exists, noise level can easily cover up weak signals from heterozygous bases, producing wrong sequence haplotypes. Fallin & Schork (2000) carried out a simulation study to examine the accuracy of the EM-based procedure for estimating haplotype frequency as a function of some of the above factors. They showed that (i) EM algorithm performs quite well under the conditions tested, (ii) estimation of the frequencies of rare haplotypes is subject to larger error, (iii) the excess of heterozygotes cause a considerably increase of estimation errors, and (iv) sampling errors appear to account for a major fraction of the inaccuracies.

More recently, Bayesian methods for haplotype phase reconstruction have also been developed. Stephens *et al*. (2001) reported that their Bayesian procedure based on coalescent theory (referred to as the 'PHASE method' after the name of the computer program developed by the authors, available at http://www.stats.ox.ac.uk/mathgen/software.html) outperforms both the parsimony and EM procedures mentioned above. This method can also estimate the uncertainty associated with each phase call, a valuable feature not present in the above methods. In an independent study that compares the performance of the PHASE method and EM-algorithm, Zhang *et al*. (2001) showed that overall there is no significant difference between the two methods, and for haplotype reconstruction using a real data set, error rates of PHASE and EM methods are 19% and 27%, respectively. Examples of application of the PHASE method can be found in Hull *et al*. (2001), Kaessmann *et al*. (2002) and Lazarus *et al*. (2002).

Niu *et al*. (2002) developed a novel Bayesian algorithm based on a Monte Carlo approach (referred to as the 'HAPLOTYPER method' after the name of their computer program, which is available at the following website: http://www.people.fas.harvard.edu/~junliu/Haplo/docMain.htm). HAPLOTYPER incorporates computational techniques to help the algorithm to escape from a local maximum and construct whole haplotypes from haplotype segments. The authors reported that HAPLOTYPER, in a comparative study, outperforms PHASE, EM algorithm and Clark's algorithm in all real data applications tested, and only in the coalescence-based simulation did PHASE perform better. It was shown that HAPLOTYPER is robust to the violation of Hardy–Weinberg equilibrium, to the presence of missing data and to occurrences of recombination hotspots.

Clearly, statistical power for haplotype reconstruction from partially resolved data has rapidly increased in recent years. Some authors have even claimed that sufficient accuracy in haplotype inference is available using statistical methods, and 'reconstructing haplotypes experimentally, or by genotyping additional family members, may be an inefficient use of resources' (Stephens *et al.* 2001). However, although statistical inference procedures may provide an attractive shortcut for haplotype reconstruction, they should not replace empirical approaches before their performance has been extensively tested and shown to be robust in various circumstances and with a large amount of empirical data.

For contemporary population-genetic approaches using nuclear DNA markers, the ideal technique for haplotype determination should be (i) easy to employ, (ii) able to fully separate haplotypes with high reliability and efficiency, (iii) minimal in the steps required for subsequent analyses such as PCR and sequencing, and (iv) manageable for automation of large-scale studies. None of the above techniques, when employed alone, can provide a satisfactory means for obtaining nuclear haplotype data in large-scale genetic analyses of populations. The highest resolving power could be achieved by a complementary employment of experimental approaches and statistical inference methods, and this seems to be the future trend.

Prospects of scnp nuclear DNA markers

For analysis with nuclear DNA markers, where should we start and how should we find scnp loci? This is not a problem for people working with model organisms, as a large number of scnp sequences and SNP markers are already available from genome sequencing projects. With nonmodel organisms, however, different strategies have to be employed depending on the situation. Because universally conserved or versatile nuclear primers are most welcome, we discuss this issue first.

'Versatile' nuclear primers for intraspecific studies

The development of conserved PCR primers has played a key role in promoting population-genetic studies at the molecular level in the past decade (Zhang & Hewitt 1997; Sunnucks 2000). Enormous success has been achieved with animal mitochondrial DNA primers (e.g. Kocher *et al.* 1989; Simon 1991; Simon *et al.* 1994). Because of this, it has been a common desire to find similarly conserved and useful nuclear DNA primers. It is fortunate enough that, with mitochondrial, and to a lesser extent chloroplast, genomes primers amplifying DNA regions highly variable at intraspecific level can be designed that are conserved among very diverged organisms. With the nuclear genome, however, the situation appears to be much more complicated. So far, only limited success has been achieved with primers amplifying the ribosomal ITS (internal transcribed spacer) regions (Hillis & Davis 1988; Williams *et al.* 1988; White *et al.* 1990; Hillis & Dixon 1991) — a fairly good conservation of primer sequences and therefore their cross-species applicability have been observed, but the intraspecific variability of the amplified regions is often low. Most other 'conserved' primers designed so far are intended for phylogenetic study, such as those that amplify genes encoding the elongation factor-1α (EF-1α), elongation factor-2 (EF-2), RNA polymerase II, dopa decarboxylase and phosphoenolpyruvate carboxykinase (Friedlander *et al.* 1994; Cho *et al.* 1995; Regier & Shultz 2001), or the *period* locus (Regier *et al.* 1998). Those primers that were designed with an intraspecific perspective have not been, in most cases, empirically tested across a significant taxonomic range, such as the plant primers by Strand *et al.* (1997), or primers amplifying introns of plant nitrate reductase gene (Howartha & Baum 2002). A general observation with nuclear primers is that frequently either they fail to amplify across distant taxonomic groups or the amplified sequences are not suitable for intraspecific studies. This is not surprising considering the situation with microsatellite loci most microsatellite primers are very organism-specific and only limited conservation among closely related species was observed.

Nuclear primers widely applicable for intraspecific studies need to satisfy the following requirements: (i) they should be evolutionarily well conserved across different taxonomic groups; (ii) their amplicons (i.e. target sequences amplified by the primers) should be of reasonable size; (iii) target sequences should be highly variable at the intraspecific level; and (iv) target sequences should be single-copy or low-copy in the nuclear genome. However, several processes of molecular evolution of nuclear DNA counteract the search for such 'versatile' nuclear primers. These include gene duplication or amplification, production of pseudogenes, intron losts, sliding or size change, etc. Fundamentally, the conflicting requirements for the candidate sequences to be evolutionarily conserved but also have high intraspecific variability are logically difficult to comply with. Recall that the variability of introns and that of the adjacent exons is correlated (Fig. 2; Moriyama & Powell 1996), thus the high variability of the amplified sequences tends to reduce the level of conservation of the primer sites. Consequently, the approach of searching for 'versatile primers' will not be so effective. Our practical experience also suggests that it will be extremely difficult at present to find many universally applicable scnp markers meeting our expectation from mitochondrial work. For a given project, the time and effort spent in testing the suitability of 'versatile primers' may not be more efficient than for developing up specific primers in the organism involved.

Introns as primary candidates

Although various noncoding and intergenic regions are all in general more variable than coding sequences, introns are the best candidates for scnp markers. This is because flanking regions of introns are exonic sequences which are evolutionarily more conserved and ideal places to place PCR primers. Also, as a structural characteristic, introns are commonly present in many genes of eukaryotic organisms. Palumbi & Baker (1994) described a strategy to find intron markers, the 'exon-primed, introns-crossing (EPIC)' method, which is generally applicable to any organism. With this strategy, Hassan *et al*. (2002) reported 17 pairs of introns-amplifying primers for fishes. Other examples of the application of EPIC strategy include Palumbi (1996) for invertebrates, Friesen *et al*. (1997) for vertebrates, Strand *et al*. (1997) for plants, among many others.

If the sequence of the 5′ or 3′ noncoding region of a gene is known (for example, from a characterized cDNA sequence or EST project), specific primers can be designed in these regions. Such primers are fairly more taxon-specific and may not be applicable even in closely related species (e.g. Zhang & Hewitt 1996c).

Microsatellite flanking regions

In organisms in which well-characterized microsatellite loci are available, microsatellite flanking regions are the ideal places to locate scnp markers. This is because (i) the majority of microsatellite loci employed in population-genetic studies should lie in noncoding genomic regions, their flanking regions are likely to be highly polymorphic in populations; (ii) it should have already been established that microsatellite loci employed in population-genetic studies are biallelic markers, that is, they are single-copy sequences. Therefore, so are their flanking sequences; (iii) as a natural extension, sequencing analysis of entire microsatellite loci will greatly increase the value of these markers while minimizing costs and labours. It provides an effective way to retrieve genealogical information from an existing framework. Similarly, sequence characterized amplified regions (SCARs, Paran & Michelmore 1993) derived from RAPD or AFLP analyses are good targets for scnp primers (e.g. McLenachan *et al*. 2000)

Genes involved in immune defence systems or disease resistance

Genes involved in immune defence systems (e.g. MHC loci in animals), mating-recognition systems (e.g. S-loci in plant) and disease resistance (e.g. the RPP5 complex locus in *Arabidopsis*) have pronounced intraspecific variability. Their high polymorphic levels make them attractive molecular markers for population-genetic study. However, these targets are believed to be under balancing selection. It is important to note that the evolutionary dynamics of balanced polymorphism are very different from those of neutral genetic variation, therefore one should be careful about the questions to be addressed and whether this is feasible with such markers. Richman (2000) has extensively discussed the evolution of balanced genetic polymorphism in a recent review.

Developing specific scnp markers from the bench

If there is no shortcut for finding scnp markers in an organism, one has to develop species-specific markers in the laboratory. This is in fact not as difficult as it may sound. The basic procedure for developing scnp markers alone has long been established and is no more difficult than for isolating microsatellite DNA loci. For example, Karl & Avise (1993) described a universally applicable method for isolating anonymous single-copy nuclear DNA (scnDNA). It involves the construction of small-insert partial genomic DNA library, and then screening of the recombinant clones for scnDNA with labelled total DNA as the probes. An alternative method was reported by Bagley *et al*. (1997), which employs a RAPD-SSCP-based approach and implements a prescreening step to first identify sequences that are polymorphic before carrying out the cloning procedures. For many organisms in which little sequence data are available, these methods could serve as the starting point of a nuclear DNA analysis project. The key element in scnp development is the verification of the biallelic status of the loci isolated.

SNP markers — a wonderland for most organisms

SNP refer to single nucleotide polymorphism, with its essence being biallelic genetic variation (see Box 1 for a working definition). Because of their simplicity in character state and the ease of large-scale automated detection, SNPs will become the marker-of-choice in nuclear DNA analysis of populations. These markers, however, will only be really valuable when they are employed in numbers. At present and in the near future, except for a few model organisms in which extensive genetic studies and genome sequencing projects are in place and thus a large numbers of genomewide SNPs have been identified, the unavailability of SNP markers in most organisms remains a big obstacle to their systematic employment in genetic study of populations.

Concluding remarks

Four major uncertainties for nuclear DNA analyses of populations were previously facing us, i.e. the availability of scnp markers for carrying out such analysis, technical laboratory hurdles for recovering haplotype data, difficulty in data analysis because of recombination, low divergence levels and intraspecific multifurcation evolution and the utility of scnp markers for addressing population-genetic questions. Although challenges still exist, these obstacles are now being removed. Doubt on the availability of scnp markers no longer exists, as shown by data from different genome sequencing projects and studies in model organisms. Different strategies are now available for mining such markers from nonmodel organisms. Haplotype recovery becomes much less a problem if we combine empirical study with statistical inference. A number of methods have already been established for detecting recombination in nuclear sequences, and the nested clade or network analysis in combination with coalescent approach promise a fuller exploration of the information contained in scnp data. The utility of scnp markers for addressing population-genetic questions has been demonstrated clearly in empirical studies, as recently reviewed by Hare (2001). What we need to overcome is the hidden conceptual obstacle that nuclear DNA analysis is too complicated to implement. The inclusion of nuclear DNA markers in evolutionary and population-genetic studies is indispensable for a better understanding of evolutionary processes that have occurred, and it will considerably extend our ability to infer the past beyond that so far allowed by other more popular DNA markers.

This trend to nuclear DNA analyses does not, however, mean that scnp markers are going to replace cytoplasmic and microsatellite DNA markers. Instead, the latter will continue to serve as powerful diagnostic tools for screening genetic variation. Scnp, mitochondrial and microsatellite DNA markers are complementary in the sense that they reveal different aspects of a complex story at different depth of perception. It is worth noting that, unlike mtDNA, widely applicable versatile PCR primers that amplify regions with high intraspecific variability are scarce for nuclear DNA. Therefore, as with microsatellite DNA markers, one often needs to specifically develop the scnp marker systems required for a particular project. Finally, the employment of several independent nuclear loci is a prerequisite for inferring a general answer for any population-genetic question.

Acknowledgements

This work was supported by NSFC, CAS and MOST grants to D-X Zhang (30025008, KSCX2-1-05, G2000016207), CAS 'Knowledge Innovation Program' and the CAS 'Bai Ren Ji Hua' professorship. Part of the manuscript was written during an academic visit of DXZ to the University of East Anglia in Spring 2001. Exchange visits by both authors were supported in part by the Royal Society. Our thanks to Dr Rieseberg for his encouragement and two anonymous referees for their helpful comments on an earlier version of the manuscript. Dr P. J. Oefner is gratefully thanked for providing a manuscript prior to publication. We thank Ya-Jie Ji for her assistance in the preparation of the manuscript.

References

Aldridge BM, McGuirk SM, Clark RJ *et al.* (1998) Denaturing gradient gel electrophoresis: a rapid method for differentiating *BoLA-DRB3* alleles. *Animal Genetics*, **29**, 389–394.

Antunes A, Templeton AR, Guyomard R, Alexandrino P (2002) The role of nuclear genes in intraspecific evolutionary inference: genealogy of the transferrin gene in the brown trout. *Molecular Biology and Evolution*, **19**, 1272–1287.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the fowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Avise JC (1989) Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution*, **43**, 1192–1208.

Avise JC (1991) Ten unorthodox perspectives on evolution prompted by comparative population-genetic findings on mitochondrial DNA. *Annual Review of Genetics*, **25**, 45–69.

Avise JC (1994) *Molecular Markers, Natural History and Evolution*. Chapman & Hall, New York.

Avise JC (1998) The history and purview of phylogeography: a personal reflection. *Molecular Ecology*, **7**, 371–379.

Avise JC (2000) *Phylogeography: the History and Formation of Species*. Harvard University Press, Cambridge.

Avise JC, Arnold J, Ball RM *et al.* (1987) Intraspecific phylogeography: the mitochondrial DNA bridge between population-genetics and systematics. *Annual Review of Ecology and Systematics*, **18**, 489–522.

Bagley MJ, Gall GAE (1998) Mitochondrial and nuclear DNA sequence variability among populations of rainbow trout (*Oncorhynchus mykiss*). *Molecular Ecology*, **7**, 945–961.

Bagley MJ, Medrano JF, Gall GAE (1997) Polymorphic molecular markers from anonymous nuclear DNA for genetic analysis of populations. *Molecular Ecology*, **6**, 309–320.

Balakirev ES, Balakirev EI, Rodriguez-Trelles F, Ayala FJ (1999) Molecular evolution of two linked genes, *Est-6* and *Sod*, in *Drosophila melanogaster*. *Genetics*, **153**, 1357–1369.

Bandelt H-J, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**, 37–48.

Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. *Genetics*, **141**, 743–753.

Begun DJ, Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, **356**, 519–520.

Bensasson D, Zhang DX, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology and Evolution*, **16**, 314–321.

Bensasson D, Zhang DX, Hewitt GM (2000) Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. *Molecular Biology and Evolution*, **17**, 406–415.

Bourbon HM, Prudhomme M, Amalric F (1988) Sequence and structure of the nucleolin promoter in rodents — characterization of a strikingly conserved CpG island. *Gene*, **68**, 73–84.

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *American Journal of Human Genetics*, **62**, 1408–1415.

Brohede J, Primmer CR, Moller A, Ellegren H (2002) Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. *Nucleic Acids Research*, **30**, 1997–2003.

Brookes AJ (1999) The essence of SNPs. *Gene*, **234**, 177–186.

Brown WM, George M Jr, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences USA*, **76**, 1967–1971.

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution*, **18**, 225–239.

Bruford MW, Wayne RK (1993) Microsatellites and their application to population-genetic studies. *Current Opinion in Genetics and Development*, **3**, 939–943.

Buckler ES IV, Ippolito A, Holtsford TP (1997) The evolution of ribosomal DNA: divergent paralogues and phylogenetic implications. *Genetics*, **145**, 821–832.

Caccone A, Amato GD, Powell JR (1988) Rates and patterns of scnDNA and mtDNA divergence within the *Drosophila melanogaster* subgroup. *Genetics*, **118**, 671–683.

Carter NP (1994) Cytogenetic analysis by chromosome painting. *Cytometry*, **18**, 2–10.

Caterino MS, Cho S, Sperling FAH (2000) The current state of insect molecular systematics: a thriving Tower of Babel. *Annual Review of Entomology*, **45**, 1–54.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Cho SW, Mitchell A, Regier JC *et al.* (1995) A highly conserved nuclear gene for low-level phylogenetics — elongation factor-1-alpha recovers morphology-based tree for heliothine moths. *Molecular Biology and Evolution*, **12**, 650–656.

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution*, **7**, 111–122.

Crozier RH, Kaufmann B, Carew ME, Crozier YC (1999) Mutability of microsatellites developed for the ant *Camponotus consobrinus*. *Molecular Ecology*, **8**, 271–276.

Culver M, Menotti-Raymond MA, O'Brien SJ (2001) Patterns of size homoplasy at 10 microsatellite loci in pumas (*Puma concolor*). *Molecular Biology and Evolution*, **18**, 1151–1156.

Davis LA, Glenn TC, Elsey RM, Dessauer HC, Sawyer RH (2001) Multiple paternity and mating patterns in the American alligator, *Alligator mississippiensis*. *Molecular Ecology*, **10**, 1011–1024.

DeSalle R, Freedman T, Prager EM, Wilson AC (1987) Tempo and mode of sequence evolution in mitochondrial-DNA of Hawaiian *Drosophila*. *Journal of Molecular Evolution*, **26**, 157–164.

Dolezel J, Lucretti S, Schubert I (1994) Plant chromosome analysis and sorting by flow cytometry. *Critical Reviews in Plant Sciences*, **13**, 275–309.

Duda TF, Palumbi SR (1999) Population structure of the black tiger prawn, *Penaeus monodon*, among western Indian Ocean and western Pacific populations. *Marine Biology*, **134**, 705–710.

Duret L (2001) Why do genes have introns? Recombination might add a new piece to the puzzle. *Trends in Genetics*, **17**, 172–175.

Dvorák J, Luo MC, Yang ZL (1998) Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing *Aegilops* species. *Genetics*, **148**, 423–434.

Dvornyk V, Sirvio A, Mikkonen M, Savolainen O (2002) Low nucleotide diversity at the *pal1* locus in the widely distributed *Pinus sylvestris*. *Molecular Biology and Evolution*, **19**, 179–188.

Ellegren H (2000a) Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature Genetics*, **24**, 400–402.

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

Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population-genetics analysis. *Molecular Ecology*, **11**, 1591–1604.

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, **12**, 921–927.

Ezenwa VO, Peters JM, Zhu Y *et al.* (1998) Ancient conservation of trinucleotide microsatellite loci in polistine wasps. *Molecular Phylogenetics and Evolution*, **10**, 168–177.

Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation–maximization algorithm for unphased diploid genotype data. *American Journal of Human Genetics*, **67**, 947–959.

FitzSimmons NN (1998) Single paternity of clutches and sperm storage in the promiscuous green turtle (*Chelonia mydas*). *Molecular Ecology*, **7**, 575–584.

Fitzsimmons NN, Moritz C, Moore SS (1995) Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. *Molecular Biology and Evolution*, **12**, 432–440.

Friedlander TP, Regier JC, Mitter C (1994) Phylogenetic information-content of 5 nuclear gene-sequences in animals — initial assessment of character sets from concordance and divergence studies. *Systematic Biology*, **43**, 511–525.

Friesen VL, Congdon BC, Walsh HE, Birt TP (1997) Intron variation in marbled murrelets detected using analysis of single-stranded conformational polymorphisms. *Molecular Ecology*, **6**, 1047–1058.

Fu Y-X, Li W-H (1999) Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theoretical Population Biology*, **56**, 1–10.

Gardner MG, Bull CM, Cooper SJB, Duffield GA (2000) Microsatellite mutations in litters of the Australian lizard *Egernia stokesii*. *Journal of Evolutionary Biology*, **13**, 551–560.

Garza JC, Slatkin M, Freimer NB (1995) Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Molecular Biology and Evolution*, **12**, 594–603.

Goldstein DB, Schlötterer C (1999) *Microsatellites: Evolution and Applications*. Oxford University Press, Oxford.

Gyllensten U, Wharton D, Josefsson A, Wilson AC (1991) Paternal inheritance of mitochondrial DNA in mice. *Nature*, **352**, 255–257.

Hare MP (2001) Prospects for nuclear gene phylogeography. *Trends in Ecology and Evolution*, **16**, 700–706.

Hare MP, Cipriano F, Palumbi SR (2002) Genetic evidence on the demography of speciation in allopatric dolphin species. *Evolution*, **56**, 804–816.

Harrison RG (1989) Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. *Trends in Ecology and Evolution*, **4**, 6–11.

Hassan M, Lemaire C, Fauvelot C, Bonhomme F (2002) Seventeen new exon-primed intron-crossing polymerase chain reaction amplifiable introns in fish. *Molecular Ecology Notes*, **2**, 334–340.

Hasson E, Wang IN, Zeng LW, Kreitman M, Eanes WF (1998) Nucleotide variation in the triosephosphate isomerase (*Tpi*) locus of *Drosophila melanogaster* and *Drosophila simulans*. *Molecular Biology and Evolution*, **15**, 756–769.

Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *Journal of Heredity*, **86**, 409–411.

Hein J (1993) A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, **36**, 396–405.

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

Hillis DM, Dixon MT (1991) Ribosomal DNA — molecular evolution and phylogenetic inference. *Quarterly Review of Biology*, **66**, 410–453.

Hilton H, Gaut BS (1998) Speciation and domestication in maize and its wild relatives: evidence from the *Globulin-1* gene. *Genetics*, **150**, 863–872.

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proceedings of the National Academy of Sciences USA*, **92**, 532–536.

Howartha DG, Baum DA (2002) Phylogenetic utility of a nuclear intron from nitrate reductase for the study of closely related plant species. *Molecular Phylogenetics and Evolution*, **23**, 525–528.

Howell N, Kubacka I, Mackey DA (1996) How rapidly does the human mitochondrial genome evolve? *American Journal of Human Genetics*, **59**, 501–509.

Hudson RR, Kaplan NL (1988) The coalescent process in models with selection and recombination. *Genetics*, **120**, 831–840.

Hudson RR, Saez AG, Ayala FJ (1997) DNA variation at the *Sod* locus *Drosophila melanogaster*: an unfolding story of natural selection. *Proceedings of the National Academy of Sciences USA*, **94**, 7725–7729.

Hull J, Ackerman H, Isles K *et al.* (2001) Unusual haplotypic structure of IL8, a susceptibility locus for a common respiratory virus. *American Journal of Human Genetics*, **69**, 413–419.

Innan H, Tajima F, Terauchi R, Miyashita TN (1996) Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. *Genetics*, **143**, 1761–1770.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–910.

International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.

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

Kaessmann H, Zöllner S, Gustafsson AC *et al.* (2002) Extensive linkage disequilibrium in small human populations in Eurasia. *American Journal of Human Genetics*, **70**, 673–685.

Karl SA, Avise JC (1993) PCR-based assays of Mendelian polymorphisms from anonymous single-copy nuclear-DNA — techniques and applications for population-genetics. *Molecular Biology and Evolution*, **10**, 342–361.

Karotam J, Boyce TM, Oakeshott JG (1995) Nucleotide variation at hypervariable esterase 6 isozyme locus of *Drosophila simulans*. *Molecular Biology and Evolution*, **12**, 113–122.

Karp A, Isaac PG, Ingram DS (1998) *Molecular Tools for Screening Biodiversity: Plants and Animals*. Chapman & Hall, London.

Kashi Y, King D, Soller M (1997) Simple sequence repeats as a source of quantitative genetic variation. *Trends in Genetics*, **13**, 74–78.

Kawabe A, Innan H, Terauchi R, Miyashita NT (1997) Nucleotide polymorphism in the acidic chitinase locus (*ChiA*) region of the wild plant *Arabidopsis thaliana*. *Molecular Biology and Evolution*, **14**, 1303–1315.

Kawabe A, Miyashita NT (1999) DNA variation in the basic chitinase locus (ChiB) region of the wild plant *Arabidopsis thaliana*. *Genetics*, **153**, 1445–1453.

Kayser M, Roewer L, Hedman M *et al.* (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *American Journal of Human Genetics*, **66**, 1580–1588.

Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

King LM (1998) The role of gene conversion in determining sequence variation and divergence in the *Est-5* gene family in *Drosophila pseudoobscura*. *Genetics*, **148**, 305–315.

Knowlton N, Weigt LA (1998) New dates and new rates for divergence across the Isthmus of Panama. *Proceedings of The Royal Society of London Series B-Biological Sciences*, **265**, 2257–2263.

Kocher TD, Thomas WK, Meyer A *et al.* (1989) Dynamics of mitochondrial-DNA evolution in animals — amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences USA*, **86**, 6196–6200.S.

Kondo R, Satta Y, Matsuura ET *et al.* (1990) Incomplete maternal transmission of mitochondrial DNA in *Drosophila*. *Genetics*, **126**, 657–663.

Kreitman M, Hudson RR (1991) Inferring the evolutionary histories of the Adh and Adh-dup loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics*, **127**, 565–582.

Labate JA, Biermann CH, Eanes WF (1999) Nucleotide variation at the *runt* locus in *Drosophila melanogster* and *Drosophila simulans*. *Molecular Biology and Evolution*, **16**, 724–731.

LaForest SM, Prestwich GD, Löfstedt C (1999) Intraspecific nucleotide variation at the pheromone binding protein locus in the turnip moth, *Agrotis segetum*. *Insect Molecular Biology*, **8**, 481–490.

Lazarus R, Klimecki WT, Palmer LJ *et al.* (2002) Single-nucleotide polymorphisms in the Interleukin-10 gene: differences in frequencies, linkage disequilibrium patterns, and haplotypes in three United States ethnic groups. *Genomics*, **80**, 223–228.

Le Corre V, Roux F, Reboud X (2002) DNA polymorphism at the FRIGIDA gene in *Arabidopsis thaliana*: extensive nonsynonymous variation is consistent with local selection for flowering time. *Molecular Biology and Evolution*, **19**, 1261–1271.

Lercher MJ, Hurst LD (2002) Human SNP variability and mutation rate are higher in regions of high recombination. *Trends in Genetics*, **18**, 337–340.

Li WH (1997) *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Li WH, Graur D (1991) *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Li WH, Sadler LA (1991) Low nucleotide diversity in man. *Genetics*, **129**, 513–523.

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics*, **56**, 799–810.

Martin AP, Pardini AT, Noble LR, Jones CS (2002) Conservation of a dinucleotide simple sequence repeat locus in sharks. *Molecular Phylogenetics and Evolution*, **23**, 205–213.

Mattick JS, Gagen MJ (2001) The evolution of controlled multitasked gene networks: the role of introns and other noncoding

RNAs in the development of complex organisms. *Molecular Biology and Evolution*, **18**, 1611–1630.

Maynard Smith J (1992) Analyzing the mosaic structure of genes. *Journal of Molecular Evolution*, **34**, 126–129.

Maynard Smith J (1999) The detection and measurement of recombination from sequence data. *Genetics*, **153**, 1021–1027.

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genetical Research*, **23**, 23–35.

Maynard Smith J, Smith NH (1998) Detecting recombination from gene trees. *Molecular Biology and Evolution*, **15**, 590–599.

McGuire G, Denham MC, Balding DJ (2001) Models of sequence evolution for DNA sequence containing gaps. *Molecular Biology and Evolution*, **18**, 481–490.

McGuire G, Wright F, Prentice MJ (2000) A Bayesian model for detecting past recombination events in DNA multiple alignments. *Journal of Computational Biology*, **7**, 159–170.

McLenachan PA, Stöckler K, Winkworth RC, McBreen K, Zauner S, Lockhart PJ (2000) Markers derived from amplified fragment length polymorphism gels for plant ecology and evolution studies. *Molecular Ecology*, **9**, 1899–1903.

Miyashita NT (2001) DNA variation in the 59 upstream region of the *Adh* locus of the wild plants *Arabidopsis thaliana* and *Arabis gemmifera*. *Molecular Biology and Evolution*, **18**, 164–171.

Morgan K, Strobeck C (1979) Is intragenic recombination a factor in the maintenance of genetic variation in natural populations? *Nature*, **277**, 383–384.

Moritz C, Dowling TE, Brown WM (1987) Evolution of animal mitochondrial DNA: relevance for population biology and systematics. *Annual Review of Ecology and Systematics*, **18**, 269–292.

Moriyama EN, Gojobori T (1992) Rates of synonymous substitution and base composition of nuclear genes in *Drosophila*. *Genetics*, **130**, 855–864.

Moriyama EN, Powell JR (1996) Intraspecific nuclear DNA variation in *Drosophila*. *Molecular Biology and Evolution*, **13**, 261–277.

Nachman MW (2001) Single nucleotide polymorphisms and recombination rate in humans. *Trends in Genetics*, **17**, 481–485.

Nachman MW, Bauer VL, Crowell SL, Aquadro CF (1998) DNA variability and recombination rates at X-linked loci in humans. *Genetics*, **150**, 1133–1141.

Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, **70**, 157–169.

Noël L, Moores TL, van der Biezen EA *et al.* (1999) Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of *Arabidopsis*. *Plant Cell*, **11**, 2099–2111.

O'Donovan MC, Oefner PJ, Roberts SC *et al.* (1998) Blind analysis of denaturing high-performance liquid chromatography as a tool for mutation detection. *Genomics*, **52**, 44–49.

Ochman H, Wilson AC (1987) Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *Journal of Molecular Evolution*, **26**, 74–86.

Oefner PJ (2000) Allelic discrimination by denaturing high-performance liquid chromatography. *Journal of Chromatography B*, **739**, 345–355.

Oefner PJ, Underhill PA (1995) Comparative DNA sequencing by denaturing high-performance liquid chromatography (DHPLC). *American Journal of Human Genetics*, **57** (Suppl.), A266.

Ohta T (1992) The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics*, **23**, 263–286.

Olsen KM, Schaal BA (1999) Evidence on the origin of cassava: phylogeography of *Manihot esculenta*. *Proceedings of the National Academy of Sciences USA*, **96**, 5586–5591.

Onyabe DY, Conn JE (1999) Intragenomic heterogeneity of a ribosomal DNA spacer (ITS2) varies regionally in the neotropical malaria vector *Anopheles nuneztovari* (Diptera: Culicidae). *Insect Molecular Biology*, **8**, 435–442.

Orti G, Hare MP, Avise JC (1997a) Detection and isolation of nuclear haplotypes by PCR-SSCP. *Molecular Ecology*, **6**, 575–580.

Orti G, Pearse DE, Avise JC (1997b) Phylogenetic assessment of length variation at a microsatellite locus. *Proceedings of the National Academy of Sciences USA*, **94**, 10745–10749.

Palmer JD, Adams KL, Cho Y *et al.* (2000) Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. *Proceedings of the National Academy of Sciences USA*, **97**, 6960–6966.

Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution*, **28**, 87–97.

Palumbi SR (1996) The polymerase chain reaction. In: *Molecular Systematics*, 2nd edn (eds Hillis DM, Moritz C, Mable BK), pp. 205–247. Sinauer, Sunderland.

Palumbi SR, Baker CS (1994) Contrasting population structure from nuclear introns sequences and mtDNA of humpback whales. *Molecular Biology and Evolution*, **11**, 426–435.

Paran I, Michelmore RW (1993) Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. *Theoretical and Applied Genetics*, **85**, 985–993.

Parsons TJ, Holland MM (1998) Mitochondrial mutation rate revisited: hot spots and polymorphism — response. *Nature Genetics*, **18**, 110.

Payseur BA, Nachman MW (2000) Microsatellite variation and recombination rate in the human genome. *Genetics*, **156**, 1285–1298.

Payseur BA, Nachman MW (2002) Gene density and human nucleotide polymorphism. *Molecular Biology and Evolution*, **19**, 336–340.

Pesole G, Gissi C, De Chirico A, Saccone C (1999) Nucleotide substitution rate of mammalian mitochondrial genomes. *Journal of Molecular Evolution*, **48**, 427–434.

Posada D (2001) Unveiling the molecular clock in the presence of recombination. *Molecular Biology and Evolution*, **18**, 1976–1978.

Posada D (2002) Evaluation of methods for detecting recombination from dna sequences: empirical data. *Molecular Biology and Evolution*, **19**, 708–717.

Posada D, Crandall KA (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology and Evolution*, **16**, 37–45.

Primmer CR, Borge T, Lindell J, Sætre GP (2002) Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. *Molecular Ecology*, **11**, 603–612.

Purugganan MD, Suddith JI (1998) Molecular population-genetics of the *Arabidopsis CAULIFLOWER* regulatory gene: nonneutral evolution and naturally occurring variation in floral homeotic function. *Proceedings of the National Academy of Sciences USA*, **95**, 8130–8134.

Purugganan MD, Suddith JI (1999) Molecular population-genetics of floral homeotic loci: departures from the equilibrium-neutral model at the *APETALA3* and *PISTILIATA* genes of *Arabidopsis thaliana*. *Genetics*, **151**, 839–848.

Queney G, Ferrand N, Weiss S, Mougel F, Monnerot M (2001) Stationary distributions of microsatellite loci between divergent population groups of the European rabbit (*Oryctolagus cuniculus*). *Molecular Biology and Evolution*, **18**, 2169–2178.

Regier JC, Fang QQ, Mitter C, Peigler RS, Friedlander TP, Solis MA (1998) Evolution and phylogenetic utility of the period gene in Lepidoptera. *Molecular Biology and Evolution*, **15**, 1172–1182.

Regier JC, Shultz JW (2001) Elongation factor-2: a useful gene for Arthropod phylogenetics. *Molecular Phylogenetics and Evolution*, **20**, 136–148.

Richman A (2000) Evolution of balanced genetic polymorphism. *Molecular Ecology*, **9**, 1953–1963.

Rico C, Rico I, Hewitt G (1996) 470 million years of conservation of microsatellite loci among fish species. *Proceedings of the Royal Society of London, Series B*, **263**, 549–557.

Ross J (1996) Control of messenger RNA stability in higher eukaryotes. *Trends in Genetics*, **12**, 171–175.

Rothenburg S, Koch-Nolte F, Rich A, Haag F (2001) A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proceedings of the National Academy of Sciences USA*, **98**, 8985–8990.

Rowan RG, Hunt JA (1991) Rates of DNA change and phylogeny from the DNA sequences of the alcohol dehydrogenase gene for five closely related species of Hawaiian. *Drosophila*. *Molecular Biology and Evolution*, **8**, 49–70.

Ruano G, Kidd KK (1989) Direct haplotyping of chromosomal segments from multiple heterozygotes via allele-specific PCR amplification. *Nucleic Acids Research*, **17**, 8392.

Ruano G, Kidd KK, Stephens JC (1990) Haplotype of multiple polymorphisms resolved by enzymatic amplification of single DNA molecules. *Proceedings of the National Academy of Sciences USA*, **87**, 6296–6300.

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

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics*, **70**, 425–434.

Schierup MH, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. *Genetics*, **156**, 879–891.

Schlötterer C, Ritter R, Harr B, Brem G (1998) High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. *Molecular Biology and Evolution*, **15**, 1269–1274.

Schlötterer C, Tautz D (1994) Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. *Current Biology*, **4**, 777–783.

Schug MD, Hutter CM, Wetterstrand KA *et al.* (1998) The mutation rates of di-, tri-, and tetranucleotide repeats in *Drosophila melanogaster*. *Molecular Biology and Evolution*, **15**, 1751–1760.

Schug MD, Mackay TFC, Aquadro CF (1997) Low mutation rates of microsatellite loci in *Drosophila melanogaster*. *Nature Genetics*, **15**, 99–102.

Sharp PM, Li WH (1989) On the rate of DNA sequence evolution in *Drosophila*. *Journal of Molecular Evolution*, **28**, 398–402.

Shimoda N, Knapik EW, Ziniti J *et al.* (1999) Zebrafish genetic map with 2000 microsatellite markers. *Genomics*, **58**, 219–232.

Simmons MP, Ochoterena H (2000) Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology*, **49**, 369–381.

Simon C (1991) Molecular systematics at species boundary: exploiting conserved and variable regions of the mitochondrial genome of animals via direct sequencing from amplified DNA. In: *Molecular Techniques in Taxonomy* (eds Hewitt GM, Johnston AWB, Young JPW), pp. 33–71. Springer-Verlag, Berlin.

Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P (1994) Evolution, weighting, and phylogenetic utility of mitochondrial gene-sequences and a compilation of conserved polymerase chain-reaction primers. *Annals of the Entomological Society of America*, **87**, 651–701.

Skibinski DOF, Gallagher C, Beynon CM (1994) Mitochondrial DNA inheritance. *Nature*, **368**, 817–818.

Small RL, Wendel JF (2000) Phylogeny, duplication, and intraspecific variation of *Adh* sequences in new world diploid cottons (*Gossypium* L., Malvaceae). *Molecular Phylogenetics and Evolution*, **16**, 73–84.

Smouse PE (1998) To tree or not to tree. *Molecular Ecology*, **7**, 399–412.

Smouse PE (2000) Reticulation inside the species boundary. *Journal of Classificattion*, **17**, 165–173.

Stephan W, Langley CH (1998) DNA polymorphism in Lycopersicon and crossing-over per physical length. *Genetics*, **150**, 1585–1593.

Stephens JC (1985) Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Molecular Biology and Evolution*, **2**, 539–556.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.

Stoneking M, Sherry ST, Redd AJ, Vigilant L (1992) New approaches to dating suggest a recent age for the human mtDNA ancestor. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, **337**, 167–175.

Strand AE, Leebens-Mack J, Milligan BG (1997) Nuclear DNA-based markers for plant evolutionary biology. *Molecular Ecology*, **6**, 113–118.

Strimmer K, Wiuf C, Moulton V (2001) Recombination analysis using directed graphical models. *Molecular Biology and Evolution*, **18**, 97–99.

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

Sunnucks P, Wilson ACC, Beheregaray LB, Zenger K, French J, Taylor AC (2000) SSCP is not so difficult: the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. *Molecular Ecology*, **9**, 1699–1710.

Swofford D (1993) *Phylogenetic Analysis Using Parsimony*, Version 3.1. Illinois Natural History Survey, Champaigne, IL.

Taberlet P, Griffin S, Goossens B *et al.* (1996) Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research*, **24**, 3189–3194.

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial-DNA in humans and chimpanzees. *Molecular Biology and Evolution*, **10**, 512–526.

Tang J, Unnasch TR (1995) Discriminating PCR artifacts using directed heteroduplex analysis (DHDA). *Biotechniques*, **19**, 902–905.

Templeton AR (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology*, **7**, 381–397.

Templeton AR, Routman E, Phillips CA (1995) Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics*, **140**, 767–782.

Templeton AR, Sing CF (1993) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analysis with cladogram uncertainty and recombination. *Genetics*, **134**, 659–669.

Tenaillon MI, Sawkins MC, Long AD *et al.* (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proceedings of the National Academy of Sciences USA*, **98**, 9161–9166.

Teng DHF, Hsu F, Peterson I *et al.* (2001) Template selection during manipulation of complex mixtures by PCR. *Biotechniques*, **30**, 868–877.

Udupa SM, Baum M (2001) High mutation rate and mutational bias at (TAA)n microsatellite loci in chickpea (*Cicer arietinum* L.). *Molecular Genetics and Genomics*, **265**, 1097–1103.

Underhill PA, Jin L, Lin AA *et al.* (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Research*, **7**, 996–1005.

Underhill PA, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL (1996) A pre-columbian Y chromosome-specific transition and its implications for human evolutionary history. *Proceedings of the National Academy of Sciences USA*, **93**, 196–200.

Venter JC, Adams DM, Myers EW *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1350.

Villablanca FX, Roderick GK, Palumbi SR (1998) Invasion genetics of the Mediterranean fruit fly: variation in multiple nuclear introns. *Molecular Ecology*, **7**, 547–560.

Vogler Ap DeSalle R (1994) Evolution and phylogentic information content of the ITS-1 region in the tiger beetle *Cicindela dorsalis*. *Molecular Biology and Evolution*, **11**, 393–405.

Wang RL, Stec A, Hey J, Lukens L, Doebley J (1999) The limits of selection during maize domestication. *Nature*, **398**, 236–239.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Weber JL, Wang C (1993) Mutation of human short tandem repeats. *Human Molecular Genetics*, **2**, 1123–1128.

White TJ, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: *PCR Protocols: A guide to Methods and Applications* (eds Innis MA, Gelfand DH, Sninsky JJ, White TJ), pp. 315–322. Academic Press, San Diego.

Wickens M, Bernstein DS, Kimble J, Parker R (2002) A PUF family portrait: 3′ UTR regulation as a way of life. *Trends in Genetics*, **18**, 150–157.

Williams SM, Debry RW, Feder JL (1988) A commentary on the use of ribosomal DNA in systematic studies. *Systematic Zoology*, **37**, 60–62.

Wiuf C, Christensen C, Hein J (2001) A simulation study of the reliability of recombination detection methods. *Molecular Biology and Evolution*, **18**, 1929–1939.

Wiuf C, Hein J (1999) The ancestry of a sample of sequences subject to recombination. *Genetics*, **151**, 1217–1228.

Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences USA*, **84**, 9054–9058.

Xu S (2000) Phylogenetic analysis under reticulate evolution. *Molecular Biology and Evolution*, **17**, 897–907.

Xu X, Peng M, Fang Z, Xu X (2000) The direction of microsatellite mutations is dependent upon allele length. *Nature Genetics*, **24**, 396–399.

Yu J, Hu S, Wang J, Wong GKS *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.

Zhang DX, Hewitt GM (1996a) Nuclear integrations: Challenges for mitochondrial DNA markers. *Trends in Ecology and Evolution*, **11**, 247–251.

Zhang DX, Hewitt GM (1996b) Highly conserved nuclear copies of the mitochondrial control region in the desert locust *Schistocerca gregaria*: some implications for population studies. *Molecular Ecology*, **5**, 295–300.

Zhang DX, Hewitt GM (1996c) The use of DNA markers in population-genetics and ecological studies of the desert locust *Schistocerca gregaria* (Orthoptera: Acrididae). In: *The Ecology of Agricultural Pests: Biochemical Approaches* (eds Symondson WOC, Liddell JE), pp. 213–230. Chapman & Hall, London.

Zhang DX, Hewitt GM (1996d) An effective method for allele-specific sequencing using restriction enzyme and biotinylation (ASSURE B). *Molecular Ecolgy*, **5**, 591–594.

Zhang DX, Hewitt GM (1997) Assessment of the universality and utility of a set of conserved mitochondrial COI primers in insects. *Insect Molecular Biology*, **6**, 143–150.

Zhang S, Pakstis AJ, Kidd KK, Zhao H (2001) Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. *American Journal of Human Genetics*, **69**, 906–912.

Zieler H, Huynh CQ (2002) Intron-dependent stimulation of marker gene expression in cultured insect cells. *Insect Molecular Biology*, **11**, 87–95.

De-Xing Zhang's research group at Zhongguancun, Beijing focuses on population dynamics of pest insects and phylogeography over China. Current work in the group involves population-genetic studies of locusts, cotton bollworm, pine moth, scorpion and frogs, using mitochondrial, microsatellite and nuclear polymorphic DNA markers. Godfrey Hewitt's research group at Norwich has for several years used molecular data in studies of hybrid zones, postglacial colonization, speciation and phylogeography.