INVI TED R EVIE W

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
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 (scp) 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 scp 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 of 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 make 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
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 cannot 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

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

*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 x 10^{-5} at the lizard tetranucleotide locus EST2 compared to the average 1.84 x 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)m, 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.
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; Miyashita 2001; Le Corre et al. 2002), 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. 1996).
Information Box 2. Classification of genic regions

<table>
<thead>
<tr>
<th>Type</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon</td>
<td>The part(s) that can be transcribed into RNA.</td>
<td>Initiation codon ATG and ending at the stop codon.</td>
</tr>
<tr>
<td>Introns</td>
<td>The part(s) between exons. During RNA maturation, introns are removed from the precursor RNAs.</td>
<td>Genes, and occasionally are found in prokaryotic (viral or organellar) tRNA and rRNA genes.</td>
</tr>
<tr>
<td>Intergenic region</td>
<td>Refers to genomic sequences between two neighboring genes.</td>
<td>Non-coding region as intron, unless the substitution rate at silent sites is often erroneously taken as evidence for less variability in introns than in nuclear coding regions.</td>
</tr>
<tr>
<td>Transcribed region</td>
<td>Refers to the parts that are transcribed to RNA sequences.</td>
<td>Nontranslated region: also known as untranslated region.</td>
</tr>
<tr>
<td>Flanking regions</td>
<td>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.</td>
<td>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.</td>
</tr>
</tbody>
</table>

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.

Translational: 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.

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 in 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
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 $P_m$) of a DNA region. Clearly $P_m$ is sensitive to sample size and the populations sampled, and is a measure of the frequency of segregating sites in populations. The relationship between $P_m$ and the estimated population mutation rate (or heterozygosity, $\theta$, see Box 1) can be expressed as

$$P_m = (\theta/100) \cdot a_{b-1}$$

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

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

(Watterson 1975). Figure 1 shows the correlation of $P_m$ and $\theta$ under various number of haplotypes sampled from populations. It can be seen that $P_m$ 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). $P_m$ of noncoding regions (light columns), which is largely contributed by introns, is clearly higher than that of exons (dark columns).

Correlated, as noticed by Moriyama & Powell; second, in the majority of cases (14 of 15), polymorphicity ($P_m$) 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
Corrigendum


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, $\theta$) should be

$$Pm = (\theta * 100) * a_{h-1}$$
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 Arabidopsis 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 untranslated region, then the 5′ far distal untranslated region and 3′ untranslated region, then the exons, and with the 5′ proximal untranslated 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 scmDNA 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 polymorphism, 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, 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/RAPLinks.html](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 clad 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 multifurcations, 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 non-denaturing 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 ‘HAPLOTPYER 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). HAPLOTPYER incorporates computational techniques to help the algorithm to escape from a local maximum and construct whole haplotypes from haplotype segments. The authors reported that HAPLOTPYER, 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 HAPLOTPYER 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 non-model 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.

© 2003 Blackwell Publishing Ltd, Molecular Ecology, 12, 563–584
Species (e.g. Zhang & Hewitt 1996c).

Specific and may not be applicable even in closely related in these regions. Such primers are fairly more taxon-sequence or EST project), specific primers can be designed in animals), mating-recognition systems (e.g. S-loci in Genes involved in immune defence systems (e.g. MHC loci resistance

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

© 2003 Blackwell Publishing Ltd, Molecular Ecology, 12, 563–584
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


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.