

Molecular technologies for biodiversity evaluation: Opportunities and challenges

New technologies for detecting variation in DNA complement traditional methods in biodiversity.

Angela Karp¹, Keith J. Edwards¹, Mike Bruford², Stephan Funk², Ben Vosman³, Michele Morgante⁴, Ole Seberg⁵, Antoine Kremer⁶, Pierre Bourso⁷, Peter Arcander⁸, Diethard Tautz⁹, and Godfrey M. Hewitt¹⁰

Better information on the degree and distribution of genetic variation is essential for developing more efficient ways of evaluating and conserving biodiversity. At present, an array of molecular techniques is available to detect diversity at the DNA level¹, but the application of these techniques—so that they provide useful information and not simply data—depends critically on the analysis method employed (see “Analytical tools for molecular data”). In general, questions of genetic diversity can be addressed at the species, population, and within-population levels².

The species level

The identification of taxonomic units and the determination of the uniqueness of species is essential information for conservation. Questions at this level include: Does a particular isolate represent a species, subspecies, or race? Is it a hybrid? If it is a species, how unique is it? Molecular techniques are potentially relevant to all these questions. They can provide information that helps in defining the distinctiveness of species and their rank-

ing according to the number of close relatives and their phylogenetic position³. Molecular markers also have much to offer to the resolution of problems concerning hybridization and polyploidy. Sequence data provide the most accurate information for questions of this type, as sequences are the only molecular markers that contain a record of their own history. In addition to revealing the groupings of individuals into different classes, appropriate analyses based on sequence data (or restriction site data) can provide hypotheses on the evolutionary relationships between the different categories. One important caveat regarding the interpretation of such data is that the information it provides relates to the evolutionary history of the sequence (gene) in question, which may be separate from that of the organism carrying it. A straightforward, but time-consuming way to avoid this difficulty would be to collect information on the genealogies of many independent sequences. Fortunately, studies so far suggest that data from mitochondrial (mt) DNA analysis, and 1–2 nuclear sequences from critical taxa, may suffice, as most species comparisons reveal quite high levels of divergence.

Although arbitrary, semiarbitrary, and other multilocus profiling techniques have been (and are) used to provide information for answering questions at the species level, we would argue strongly against this because of data limitations in allelic assignment, dominance and homology. In principle, these limitations are not insurmountable, provided that sufficient preliminary pedigree analysis is carried out to determine indepen-

dence and mode of inheritance, and that sample sizes are large enough⁴. But in many biodiversity studies this is not possible due to sampling problems or financial and time constraints. Sequence tagged microsatellites (STMS) and minisatellites, in contrast, constitute a single locus with (usually) many different, codominant alleles. Identity and assignment of alleles is thus not a problem.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Courtesy Stephan Funk and Mike Bruford,
Institute of Zoology, London

Figure 1. Automated single sequence repeat (SSR) genotyping allows rapid DNA fingerprinting of organisms. Such markers are highly informative for characterizing plant and animal genetic resources.

Their high mutation rate does mean, however, that the accuracy with which true homology can be inferred for different genotypes becomes questionable over large genetic distances because of the increasing possibility of homoplasy. Although the presence or absence of a given STMS locus can be used as phylogenetic information, it is otherwise difficult to envisage the use of STMS in the reconstruction of phylogenies.

The population level

Below the species level, we are concerned with identifying how many different classes are present, determining the genetic similarities among the classes and their evolutionary relationships with wild relatives, and identi-

¹Department of Agricultural Sciences, IACR-Long Ashton Research Station, Bristol, BS18 9AF, UK. ²Institute of Zoology, Regents Park, London NW1 4RY, UK. ³CPRO-DLO, PO Box 16, Wageningen 6700 AA, The Netherlands.

⁴University of Udine, Via Fragnana, 209, I-33100 Udine, Italy. ⁵Botanisk Laboratorium, Gothersgade 140, DK-1123 Copenhagen, Denmark. ⁶Laboratoire de Genetique et Amelioration, INRA Station de Recherches de Bordeaux-Cestas, BP 45 Gazinet, Pierroton, France. ⁷CNRS URA, University of Montpellier II, 34095 Montpellier Cedex 5, France.

⁸Institute of Population Biology, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark.

⁹Zoologisches Institut der Universität München, Luisenstr 14, 8000 München 2, Germany. ¹⁰Biological Sciences, University of East Anglia, Norwich, NR4 7JJ, UK.

fying specific traits of interest. Much *ex situ* conservation, germplasm and breeding line management involves questions of this kind.

A variable number of tandem repeats (VNTR) fingerprints, amplified fragment length polymorphisms (AFLPs), and all arbitrary primed approaches [RAPDs, ISSR, DAMD, etc; see "Lexicon of molecular marker technologies"] produce multilocus profiles that are good for distinguishing between closely related genotypes. Their major applications are thus in establishing identities, determining parentage, fingerprinting genotypes, and in distinguishing genotypes below the species level. The difficulty of achieving robust profiles in arbitrary primed approaches such as RAPD does, however, make their reliability for "typing/fingerprinting" questionable. For the same reasons, band profiles are problematic for use in databases.

Questions concerning how many different classes are present and the estimation of genetic distances between them could, in principle, be tackled using any of the molecular techniques outlined in "Lexicon of molecular marker technologies." The choice will depend upon such factors as the anticipated level of polymorphism (e.g., where diversity is low, highly polymorphic markers are required, whereas the choice is wider for more diverse material) and the operational and financial resources available (e.g., RAPDs are less resource intensive than AFLPs). Caution should always be exercised, however, if information on the distribution of the markers is not known. Estimates of genetic distance between individuals (similarity or distance) may be affected by several factors: First, the number of markers used; second, the distribution of markers in the genome; and third, the nature of the evolutionary mechanisms underlying the variation measured.

Genome coverage is expected to affect the variance only in the presence of linkage dise-

quilibrium, in which case equally spaced markers will give a better estimate than randomly distributed ones. In the case of linkage equilibrium, marker distribution is less important. This is true for most natural populations of outcrossing organisms (animals, trees, etc.), but may not be the case for selfing species, or those under strong selective pressure because of breeding. Further caution is required if classes are to be ranked in terms of evolutionary history, for reasons outlined previously.

For the location of specific traits, molecular markers that are widely distributed in the genome are required. The development of dense genetic maps, and strategies such as bulked segregant analysis, have greatly facilitated the identification of markers linked to agronomic traits. Although restriction fragment length polymorphisms (RFLPs) are attractive because of their robustness and codominance, PCR-based assays are necessary for application to the extensive sample sizes that need to be screened. Whatever the marker, it will only be of use as long as the linkage to the trait is maintained when changing from one genetic background to another. The limited extent to which genetic maps can, in detail, be transferred among crosses portends the difficulties that may have to be faced. STMS could provide the means to produce "index maps," in which the markers are easily

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Courtesy Gitte Petersen and Ole Seberg,
Botanical Institute, Copenhagen, Denmark

Figure 2. A dendrogram constructed from molecular marker data reveals diversity patterns within resource collections (barley is shown here), facilitating both management of the collection and user access. See "Analytical tools for molecular data."

transferable between crosses and their map position is unambiguously defined.

Natural populations

Population questions are fundamental to *in situ* conservation and include the following: How are populations of given species distributed? Are they widespread or isolated in small patches? Are they genetically distinct from one another? How much genetic variation is there? Is there gene flow among them, and how is the genetic variation distributed among populations?

Although many molecular techniques

Lexicon of molecular marker techniques

Molecular marker techniques can be grouped into general categories, depending upon whether or not the assays are PCR-based and whether arbitrary/semiarbitrary primers for unknown sequences, or specifically designed primers for known sequences, are used. Non-PCR-based methods include RFLP analysis, in which DNA is digested with restriction enzymes and the resulting fragments are separated by gel electrophoresis, transferred to a filter by Southern blotting, and probes are hybridized to the filter⁹. Hybridization to genomic DNA with probes for hypervariable regions composed of tandem repeats [known as "microsatellites," or simple sequence repeats (SSRs); see Fig. 1], where

the basic repeat unit is around 2–8 base pairs in length, and "minisatellites" for longer repeat units of approximately 16–100 base pairs) gives multilocus patterns that can resolve variation at the levels of populations and individuals¹⁰. This last technique is often referred to as VNTRs, or oligonucleotide fingerprinting.

With the development of PCR, the necessity for probe hybridization steps could be avoided. Multiple arbitrary amplicon profiling (MAAP)¹¹ uses single "arbitrary" primers (purchasable from commercial companies) in the PCR reaction and results in the amplification of several discrete DNA products. MAAP includes RAPDs (random amplified polymorphic DNA), in which the amplification products are separated on agarose gels in the presence of ethidium bromide and visualized under ultraviolet light, and AP-PCR

(arbitrary primed PCR) and DAF (DNA amplification fingerprinting), which differ from RAPDs in primer length, stringency conditions, and fragment detection. The newer technique of amplified fragment length polymorphism (AFLP)¹² is essentially intermediate between RFLP and RAPD and involves restriction digestion of the genomic DNA, followed by selective PCR amplification of the restricted fragments. The amplified products are usually separated on a sequencing gel and can be visualized after exposure to X-ray film, or by fluorescent labeling. In directed amplification of minisatellite-region DNA (DAMD), VNTR core sequences, such as M13, are used as primers in PCR reactions. In single primer amplification reaction (SPARs), primers are based on microsatellite core motifs. Another technique, inter-simple

have been applied to questions of this kind, the most useful are codominant, single locus markers.

Information on the extent and distribution of diversity will assist in the development of efficient collecting and sampling strategies and in the identification of centers of diversity. For effective conservation, management principles have to be established^{5,6} (see Fig. 3). Here, information on genetic diversity is needed to define appropriate geographical scales for monitoring and management, to establish gene flow mechanisms, and to identify the origin of individuals (e.g., to determine the role of migration). A prerequisite for conservation is the identification of populations with inde-

Courtesy Christoph Sperisen, Urs Büchler, and Gabor Mátvás, Swiss Federal Institute of Snow and Landscape, Switzerland.

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Figure 3. The use of molecular markers enables the structure and history of diversity of a species (in this case, Norway Spruce) to be tracked. This knowledge is important for the management of populations to maintain diversity and for understanding the processes, dynamics, and biological function of biodiversity in natural and agricultural ecosystems.

pendent evolutionary histories and the ability to assess the conservation value of populations from an evolutionary or phylogenetic perspective. Furthermore, in the management of populations, demographic factors, such as mating systems, inbreeding depression, effective population size, and population subdivision, may be of equal importance to genetic factors⁷. Because the demographic history of a population is reflected in its genetic composition, molecular markers can provide important information on demography, provided that the data quality of different markers are taken into account⁶. STMS and sequences (haplotypes) are the markers of choice here, although the

levels of polymorphism detectable in some sequences may be insufficient to yield useful information for other than the most divergent populations.

Population diversity

Information on who breeds with whom and on the identity of individuals with respect to their parents is important for the management of small numbers of individuals in ex situ collections. Multilocus profiling approaches can provide extremely useful information for questions of this kind⁸. Provided the analysis is carried out properly (i.e., it is known that the bands in the fingerprint occur independently and there is no linkage

disequilibrium) relatedness can be accurately estimated from band sharing coefficients for the identification of individuals (e.g., in forensics) or relatives (e.g., in mating behavior and paternity exclusion).

The future

Although molecular techniques are already available for application to biodiversity evaluation, the current technologies all suffer some technical and theoretical limitations. There is a tradeoff between different types of marker with regard to their use for diversity assessments. Techniques that generate multilocus profiles provide information on numerous (presumably) dispersed loci, although the information on a single locus is low. Conversely, sequenc-

ing and STMS are limited in loci coverage, but they are extremely informative for the locus concerned. Methods based on random (anonymous) markers have proved useful in restricted and specific applications, such as relatedness analyses or cultivar/strain identification. Even in these cases, however, more accurate answers to the same questions can be obtained with reliable markers at individual loci.

Importantly, molecular methods are useful, not only in biodiversity measurement, but also in biodiversity management. Their use makes it possible to obtain an unprecedented understanding of the processes and dynamics of biodiversity, its evolution, and

sequence repeat amplification (ISSR), involves the anchoring of designed primers to a subset of microsatellites and results in the amplification of the regions between two closely spaced, oppositely oriented, SSRs. Primers based on microsatellite (random amplified microsatellite polymorphism), transposon or interspersed repeat sequences (REP-PCR) may also be used.

To generate diversity data from specific sequences, such as genes, it is necessary to have knowledge of the sequence surrounding the target to design specific primer pairs. There are three sources of potential sequences for a PCR-targeted approach: The chloroplast (cpDNA), mitochondrial (mtDNA)^{5,13} and nuclear (nDNA)¹⁴ genomes. These differ in their mode of inheritance, evolutionary rates, and recombination, all of which have important con-

sequences in terms of their use in diversity studies. A targeted PCR approach is applicable to minute amounts of DNA from extremely small samples, e.g., single pollen grains, tiny leaf fragments, and even fossils. Sequencing the amplified fragment will potentially resolve all possible differences, and the data from the aligned sequences of different individuals can then be compared. Gel systems, such as TGGE (thermal gradient gel electrophoresis), DGGE (denaturing gradient gel electrophoresis), single-strand conformational polymorphism (SSCP), and heteroduplex formation, provide sensitive assays for detecting variations down to a single base-pair and can be used to reduce the number of samples that need to be sequenced although, in practice, the generation of good TGGE and DGGE gels may be more laborious than sequencing. In the tech-

nically simpler method of PCR-RFLP, or cleaved amplified polymorphic sequence (CAPS), the amplified product is digested with a restriction enzyme and the products visualized on an agarose gel.

If SSR loci are cloned and sequenced, primers to the flanking regions can be designed to produce STMS^{15,16}. STMS provide attractive markers because each primer pair usually identifies a single locus which, because of the high mutation rate of SSRs, is often multiallelic. It is common to run STMS on sequencing gels, where single repeat differences can be resolved and thus all possible alleles detected. Minisatellites are generally very difficult to clone by virtue of their size; however, if they can be isolated with sufficient flanking sequence for primer design, they provide single locus markers similar to STMS, but even more polymorphic.

FEATURE

natural preservation—provided the right markers are chosen. All major advances in the field of population genetics and evolution have come from detailed studies of specific markers with well-known properties in terms of transmission, position in the genome, and mode of mutation. Of the current technologies, the marker systems that contribute most to this are STMS and sequences.

Current limitations lie in the number of well-defined markers available. Three points are relevant: First, random, or arbitrary amplification can be used as a first step toward the identification of single locus markers; second, considerable progress has been made in the field of genome mapping and sequencing of entire genomes, and a wealth of information of new gene and genomic sequences is thus being gathered; third, efficient retrieval systems for the isolation of large numbers of microsatellites from plant and animal

genomes are now available.

Much could be gained from a convergence between genetic mapping and diversity studies. Where possible, markers should be chosen according to their distribution to ensure that marker sampling errors are not committed. Thus far, most molecular markers have been used in an anonymous manner—often it is not known where they are located in the genome, whether they are in coding or noncoding regions, or linked to major genes, or even sometimes whether they are in the nuclear or cytoplasmic genomes. Clearly, more information is needed to enable the classification of markers into different categories, for example, on the basis of mode of transmission, or evolution with respect to different selective pressures. Research in this area needs also to include theoretical investigations on both the influences of different marker properties and con-

siderations of effective sampling strategies within genomes as well as at the individual, population, and geographic scales.

Finally, more facilities need to be devoted to microsatellite cloning and sequencing to enable researchers with access to the best data. Sharing and compilation of such data will, however, require the development of new bioinformatics methods adapted to the specific nature of polymorphism data. An interesting and useful byproduct of data from genome sequencing projects would be the preparation of a bank of primers of various types of organisms that would be accessible (at low or no cost) to anyone interested in applying molecular technologies to biodiversity.

Acknowledgments

The authors thank Arnd Hoeveler, European Commission, DGXII Biotechnology, for helpful discussions. The financial support of the following EC DGXII Biotechnology grants is acknowledged: BIO2-920476; BIO2-920486; BIO2-930373; BIO2-930295.

Analytical tools for molecular data

It is essential to understand the way in which molecular data are analyzed¹. Shared bands are scored as presence/absence and converted into similarity (or dissimilarity) measurements depending on the statistical method used (e.g., simple matching, Jaccard, Dice etc.). Such measures of "genetic distance" are an important way of expressing divergence (or similarity) between sequences, individuals, or taxa. All possible comparisons between the entities screened are used to construct a matrix of pairwise distances that are analyzed using clustering algorithms such as UPGMA (unweighted pair-group method using arithmetic averages) and neighbor-joining, or principle coordinate analysis (PCO). The results are presented as phenograms or principle coordinate plots that, respectively, provide graphic representations of the similarity between groups of entities or operational taxonomic units (OTUs). Nucleotide and restriction site data can be analyzed using such phenetic approaches (based on measures of overall distance/similarity), but they are also appropriate for cladistic analysis, which places samples together, not because of their high genetic similarity, but because they share a given marker(s)¹⁷. The resultant dendrograms (called cladograms; see Fig. 2) are reconstructions of phylogeny, and parsimony or maximum likelihood approaches can be used to select the best tree. The OTUs in such studies are usually selected representatives of higher taxa, or species, but cladistic methods may be used whenever recombinations between the OTUs is negligible, for example, for well-isolated con-specific populations and organellar genes. Arbitrary-primed data

(e.g., RAPDs) are not usually regarded as suitable for cladistic analyses because of problems of allelic assignment of bands and homology⁴.

For investigation of diversity in natural populations, information on gene and allele frequencies is more relevant. Here, data are computed using population genetic statistics, the most common of which are *F*-statistics, which describe correlations between alleles at different levels of sampling (in the whole population, in subpopulations and between subpopulations)¹⁸. Estimates from these statistics are based on a form of hierarchical analysis of variance of allele frequencies at these different levels of subdivision of the sample. The parameter of intersubpopulation correlation (F_{st}) expresses the level of differentiation between subpopulations per generation and can serve to calculate the average number of migrants between subpopulations per generation that would give the same level of differentiation in an ideal population with isotropic migration between subpopulations. This approach poses serious problems in the definition of the parameters when numerous alleles are present at each locus and when high mutation rates cause frequent homoplasy (i.e., when identically scored alleles are not identical by descent), as in the case of microsatellites. A new approach has been proposed that incorporates information on the degree of similarity of the alleles¹⁹. Several other indices, including K_{st} and N_{st} , which are analogous to Wright's *F*-statistics are available for use with sequence data²⁰ and, recently, cladistic analytical approaches have been used on sequence data to reveal other population processes, including historical patterns of colonization and bottleneck events.

1. Karp, A. et al. Molecular techniques in the assessment of botanical diversity. *Ann. Bot.* **78**:143–149.
2. Avise, J.C. 1994. *Molecular Markers, Natural History and Evolution*. Chapman & Hall, London.
3. Vane-Wright, R.I. et al. 1991. What to protect—Systematics and the agony of choice. *Biol. Conserv.* **55**:235.
4. Clark, A.G. and Lanigan, C.M.S. 1993. Prospects for estimating nucleotide divergence with RAPDs. *Mol. Biol. Evol.* **10**:1096–1111.
5. Moritz, C. 1994. Applications of mitochondrial DNA analysis in conservation: A critical review. *Mol. Ecol.* **3**:401–411.
6. Milligan, B.G. et al. 1994. Conservation genetics: Beyond the maintenance of marker diversity. *Mol. Ecol.* **3**:423–435.
7. Lande, R. 1988. Genetics and demography in biological conservation. *Science* **241**:1451–1460.
8. Bruford, M.W. et al. 1992. Single locus and multi locus DNA fingerprinting. In *Molecular Genetic Analysis of Populations: A Practical Approach*. Hoelzel, A.R. (ed). IRL Press, Oxford.
9. Helentjaris, T. et al. 1985. Restriction fragment polymorphisms as probes for plant diversity and their development as tools for applied plant breeding. *Plant Mol. Biol.* **5**:109–118.
10. Jeffreys, A.J. et al. 1985. Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**:67–73.
11. Caetano-Anolles, G. 1994. MAAP—A versatile and universal tool for genome analysis. *Plant Mol. Biol.* **25**:1011–1026.
12. Vos, P. et al. 1995. AFLP: A new technique for DNA fingerprinting. *Nucl. Acids Res.* **23**:4407–4414.
13. Demesure, B. et al. 1995. A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. *Mol. Ecol.* **4**:129–131.
14. Hillis, D.M. and Dixon, M.T. 1991. Ribosomal DNA: Molecular evolution and phylogenetic inference. *Quart. Rev. Biol.* **66**:411–453.
15. Tautz, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA. *Nucl. Acids Res.* **17**:6463.
16. Morgante M. and Olivieri, A.M. 1993. PCR-amplified microsatellites as markers in plant genetics. *Plant J.* **3**:175–182.
17. Hillis, D.M. and Mable B.K. 1996. Applications of molecular systematics. The state of the field and a look to the future, pp. 515–543, in *Molecular Systematics*, Edn 2. Hillis D.M., Moritz C., and Mable B.K. (eds). Sinauer Associates, Sunderland, MA.
18. Weir, B.S. and Cockerham, C.C. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**:1358–1370.
19. Goldstein, D.B. et al. 1995. An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**:463–471.
20. Lynch, M. and Crease, T.J. 1990. The analysis of population survey data on DNA sequence variation. *Mol. Biol. Evol.* **7**:377–394.