Evolutionary Conservation Genetics
Jacob Höglund

Print publication date: 2009
Published to Oxford Scholarship Online: May 2009
DOI: 10.1093/acprof:oso/9780199214211.001.0001

Ecological genomics
Jacob Höglund
DOI:10.1093/acprof:oso/9780199214211.003.0007

Abstract and Keywords
This chapter reviews advances in genomics and their applications to conservation genetics. Topics discussed include whole-genome sequencing (WGS), assembly and annotation of the massive amount of sequence data produced by genomic studies, evolutionary and ecological analyses, genomics in conservation, and genomic studies of non-model species.

Keywords: genomics, whole-genome sequencing, conservation, phylogenetics, gene expression

The first decade of the twenty-first century has been called the age of ‘omics’. The now famous word ending was first used in genomics but now transcriptomics, proteomics, and the like are also used. The first whole genome sequenced was that of Haemophilus influenzae, the genome sequence of which was published in 1995 (Fleischman et al. 1995). By the end of 2007 more than 700 completely sequenced genomes were available (see GOLDTM, the Genomes Online Database, v 2.0 at www.genomesonline.org) and more than 3000 whole-genome sequencing (WGS) projects are on the way. Most of the published genomes are bacterial. In eukaryotes, WGS projects are focusing primarily on fungi, protists, and plants but other
taxa are also being subjected to WGS. Sequencing of more than 40 mammalian species and six bird species is now in progress (Segelbacher and Höglund 2008). WGS projects focus primarily on so-called model organisms for genetic and physiological research and on species of economical or agricultural interest. However, some whole-genome projects have been chosen because the species have a phylogenetic position that makes the data gathered useful in comparative genomic projects.

It is easy to get carried away by the technical advances and the landmark findings that are reported on a regular basis in the weekly science journals. No doubt, the availability of whole-genome information opens new research fields which can be of interest not only in model species, but can also be of potential use in related species. As more and more species become sequenced, the comparison of genomes will enable the identification functional DNA regions in ecologically interesting species (Travers et al. 2007, Piertney and Webster 2008, Wheat 2008). Whether these kinds of data and the techniques they allow to be employed will ever be of much use in the study of endangered or rare species is less certain.

Ecological and evolutionary applications of genomics are still in their infancies and it is far too early to make any good guesses about future research directions. At the moment, the financial costs for a WGS exceed what is usually spent on an average endangered species and I doubt that any but a few conservation flagship species will be subjected to WGS. However, as I hope will become apparent in (p.120) this chapter, some of the tools and some of the techniques may also be applied to species about which there is or will be only limited genomic information.
7.1 WGS

Genomics is the science of whole genomes. The focus of interest is on the properties of entire and completely sequenced genomes, such as genomic architecture, size of genomes, number of genes, gene order, and synteny (Pagel and Pomiankowski 2007). Typical questions are: how is the genetic information compartmentalized? What is the extent of regulatory genes? What is the extent of informative versus junk DNA? Are there any transposomal elements present? Although interesting and perfectly valid research foci, it is clear that these are quite far from the concerns of the average conservation biologist. However, given genomic information, a number of facts useful to conservation may be extracted. I will review these applications later on in this chapter but first I will briefly describe the techniques for gathering the data and types of analyses involved to get to this first step.

The first thing to do in any genomic survey is, of course, to gather masses of sequence data. In model organisms, the template for such studies are often a few isolates or an inbred line of the species of interest. Until recently, WGS projects used traditional Sanger sequencing which, depending on the organism, involved massive cloning and sequencing of mega-base pairs. Even for modest-sized genomes, these projects took years to complete and involved huge consortia. Currently, there are tremendous advances in fast and inexpensive sequencing technologies (Wheat 2008). Recently it has been established that alternatives to Sanger sequencing, such as parallel pyrosequencing, can provide deep coverage of eukaryotic genomes and transcriptomes (Bainbridge et al. 2006, Cheung et al. 2006, Weber et al. 2007, Vera et al. 2008). These major technological developments make it feasible to collect genomic data from non-model species.

Whereas it may not be feasible to invest heavily in genome sequence resources for every species or natural population of interest in conservation biology, so-called expressed sequence tags (ESTs) are a relatively inexpensive genomic resource that can be developed for almost any organism regardless of sequencing strategy (Bouck and Vision 2007). ESTs are single-read sequences produced from partial sequencing of an mRNA pool. Reverse transcriptase is used to produce cDNA, which is then cloned into a vector library and sequenced. EST libraries thus provide a snapshot of the transcribed mRNA population within a given set of tissues, developmental stages,
environmental conditions, and genotypes (Rudd 2003, Dong et al. 2005). Previously, transcriptome data was obtained by Sanger sequencing of ESTs (Adams et al. 1991). Despite improvements to Sanger (p.121) sequencing over the past 30 years, this methodology is still labour-intensive and expensive. By contrast, a single 8-hour sequencing run using 454 pyrosequencing (or similar techniques) can generate mega-bases of DNA sequence and does not involve any cloning step (Margulies et al. 2005). Parallel pyrosequencing yields randomly fragmented sequencing reads that, if sufficiently abundant, may span the entire transcriptome (if based on mRNA) or genome (if based on DNA).

The average conservation biology laboratory is probably equipped to perform Sanger sequencing in house or to prepare samples so that they can be sequenced commercially at reasonable cost. However, at the time of writing, it is necessary for most conservation biologists who are contemplating parallel pyrosequencing to team up with a specialist laboratory that has the equipment and knowledge to perform such studies.

7.2 What to do with the data? Assembly and annotation
Regardless of method, genomic studies produce masses of sequence data and the assembly and annotation of such data are not trivial tasks. As an example, de novo assembly of a eukaryote transcriptome using 454 pyrosequencing data has established the utility of gathering such data in an ecologically well-studied but genomically unknown species, the Glanville fritillary, *Melitaeia cinxia* (Vera et al. 2008), as outlined below.

There are a number of specialized statistical tools and software packages available for this assembly stage (see Wheat 2008 for a review). In short, raw sequences are filtered so that low-quality reads are taken out of the data; the remaining ones are entered into the assembly. Next the high-quality reads are aligned and overlapping sequences are combined into contiguous sequences (so-called contigs). Non-overlapping reads are left as singletons.

The Glanville fritillary study used two normalized complementary DNA collections from about 80 individuals collected in the study area, including larvae, pupae, and adults. Using 454 sequencing they produced 608053 ESTs of which 518079 exceeded the minimal quality-standard filtering
and entered the assembly. The ESTs were of a mean length of 110 nucleotides. This assembled into 48354 sets of overlapping DNA segments (contigs) and 59943 single reads. For quality-control purposes they also used Sanger sequencing to obtain 3888 sequence reads from Glanville fritillary cDNA libraries. With this technique they found 364 contigs with an average length of 574bp. In general they found good agreement between 454 ESTs and ESTs obtained by traditional methods.

The authors then compared their data with sequences already banked in Internet databases from other organisms like *Drosophila* species, the genomically well-studied silkworm *Bombyx mori*, and the butterfly *Heliconis erato* and confirmed the accuracy of the sequencing and assembly. These comparisons allowed the authors to find about 9000 unique genes and more than 6000 additional unannotated contigs. These unannotated contigs were confirmed to be expressed genes by microarray analyses. The average depth of the coverage was 6.5-fold, meaning that any sequence in the transcriptome was sequenced about six times. This example shows the utility of genetically well-studied genomic reference species for the annotation part of the process.
7.3 What to do with the data? Evolutionary and ecological analyses
When there is an annotated assembly the next issue becomes: what to do with it? As noted above simple descriptive statistics of a genome or a transcriptome do not aid a conservation project very much. Fortunately, whole-genome or transcriptome sequencing enables functional genomic studies. Such studies have so far by necessity been applied mainly to a few model organisms. However, the conservatism in gene organization and in the sequences of functional genes give hope that such tools may also be used in non-model species. Some, so-called house-keeping genes, code for gene products that are involved in cell-physiological processes that have been retained and maintained in many life forms for millions of years. It is already obvious that the findings in model organism research may be applied to closely related species; for example, findings in Arabidopsis thaliana can be applied to and tested in other species in the genus. Similarly, findings in studies of the domestic chicken, Gallus gallus, can be applied to other galliforms. Recently quantitative trait loci (QTL) for female comb size were shown to be non-randomly associated with female reproductive investment in domestic chicken (Wright et al. 2008). Whether such results can be applied to non-models and how phylogenetically distant they can be from a given model organism depends on the conservation of the particular sequences under study.

At present there are very few large-scale genome studies on ecologically relevant non-model organisms in which questions about adaptive genetic diversity in natural populations can be addressed. This problem may be exemplified by birds. Birds are very well known ecologically since there has been a long research tradition of ecological studies. In birds, the species with the most information on genomic architecture and gene sequence are the galliform domestic chicken (www.ncbi.nlm.nih.gov/projects/genome/guide/chicken/) and the passerine zebra finch (http://songbirdgenome.org/). However, these model systems may be poor indicators of genomic resources in other species as chickens have gone through multiple generations of domestication, and passerines are evolutionarily quite distinct from other bird taxa.

(p.123) Under the present biodiversity crisis conservation biologists need tools to define taxa and prioritise populations that need protection in the face of limited resources. It has
been claimed that these so-called management or evolutionarily significant units need to be defined as populations and lineages that are demographically and hence evolutionarily independent. How best to define such units is unclear (Crandall et al. 2000). Beaumont and Balding (2004) highlighted that ‘Hitherto, the degree of adaptive divergence between populations has been determined genetically by some measure of distinctiveness—for example the possession of reciprocal monophyly in mitochondrial sequences. However, this distinctiveness may, particularly if only based on [mitochondrial] DNA or a few nuclear markers, largely reflect the vagaries of demographic history. What is needed is to be able to quantify the distinctiveness of populations in terms of their local adaptation …, which may also only involve a few genes, but genes with key functional roles’ (see also Luikart et al. 2003).

The issue then boils down to the identification and localization of the genes underlying fitness differences and adaptive divergence in natural populations (Luikart et al. 2003, Vasemägi and Primmer 2005, Butlin 2008, Piertney and Webster 2008). One way to do this would be to test whether it is possible to find genetic variation that correlates with fitness in natural populations in genes that have a known function in a genomic reference species (i.e. a candidate gene approach). However, with this approach it is impossible to find new genes of functional importance in non-models. At the other extreme are studies using genome scans to detect signs of selection (e.g. Cork and Purugganan 2005; Fig. 7.1). Purifying and diversifying selection on polygenic traits can be expected to produce predictable patterns of allelic variation at the underlying loci underlying a QTL, and the locus-specific effects of selection should therefore be detectable against stochastic variability of the rest of the genome (Storz 2005). Vasemägi and Primmer (2005) called for the use of a set of complementary research strategies to find functionally important loci (Box 7.1).

It is clear that there no single strategy to cover the whole way of understanding the genetic basis of ecologically important traits. Basically there are two rather fundamentally divergent research traditions that need to be merged. On the one hand researchers in ecological genetics have long been using the quantitative genetic tools of plant and animal breeders. As such, additive genetic variation in the form of heritabilities has
been established for many important life-history traits in wild populations (see Chapter 2). New statistical advances in quantitative genetics have revitalized the study of quantitative genetics in natural populations (Frentiu et al. 2008, Ovaskainen et al. 2008). However, when it comes to identifying the genomic regions and the genes underlying the variation in these life-history traits, less progress has been made (but see examples below). In going from genes to ecology, the research traditions of molecular genetics, functional (p.124) genetics, and genomics need to be incorporated. In this research tradition there has been a focus on a few model organisms that are often poorly characterized from an ecological standpoint. It goes without saying that none of the model organisms belong to the category of threatened species about which conservation biologists are concerned.

Box 7.1 Methods to detect functionally important genetic variation

The methods have been categorized on a scale from bottom-up to top-down approaches depending on the focus of the research along the genotype–phenotype pathway (after Vasemägi and Primmer 2005).

Single-locus and sequence-based ‘neutrality’ tests

Listed below are statistical tests designed to test whether a particular DNA sequence have evolved under a neutral

![Figure 7.1 Nucleotide diversity for dimorphic genes in Arabidopsis (from Cork and Purugganan 2005, reprinted with permission from the publisher).]
model or under stabilizing or balancing selection. Example of tests used are listed below.

- Deviations from the expected Hardy-Weinberg genotypic proportions within a population (Watt and Dean 2000).

- The Ewens–Watterson test uses the allele frequency distributions and tests if there is more linkage disequilibrium and less genetic variation in the particular region than is expected in a neutral marker.

- Tests to detect evidence of selection in the past such as the Hudson–Kreitman–Aguadé (HKA) test and Tajima's D test are based on the distribution of sequenced alleles and/or the level of sequence variability (Watterson 1977, Hudson et al. 1987, Tajima 1989, Fu and Li 1993, Fu 1996, and Fay and Wu 2000).

- Tests based on the non-synonymous and synonymous substitution ratio (dN/dS or KA/KS) and McDonald–Kreitman-type test (Hughes and Nei 1988, McDonald and Kreitman 1991). The McDonald–Kreitman test is based on the observation that if the observed variation is neutral, then the rate of substitution between species and the amount of variation within species are both a function of the mutation rate. Thus the ratio of non-synonymous to synonymous fixed differences between species should be the same as the ratio of non-synonymous to synonymous polymorphisms within species.

Reviews of statistical tests that can be used to test for selection on DNA sequences are found in Kreitman and Akashi (1995), Kreitman (2000), Otto (2000), Ford (2002), and Garrigan and Hedrick (2003).

Tests of dN/dS ratios are reviewed in Nielsen (1997), Yang and Nielsen (2000), and Bierne and Eyre-Walker (2003). Some of these tests may explicitly examine which amino acid sites that have been subjected to selection (see examples in Nielsen and Yang 1998, Yang et al. 2000, Suzuki 2004, and Massingham and Goldman 2005).

**Multiple-marker-based ‘neutrality’ tests**
Information from many loci may be used to test whether any loci deviate from a neutral null distribution of variation. The studies of amplified fragment length polymorphisms (AFLPs) and environmental variation in periwinkles and common frogs are examples of this (Wilding et al. 2001, Bonin et al. 2006).

The test of Lewontin and Krakauer (1973) examines the variation among populations for given loci. Theoretically, all loci are subjected to the same amount of genetic drift and gene flow. Thus the expected variance over populations should be the same. However, differential selection among populations increases the variance. On the other hand, purifying selection over populations decreases the variance. This test have been used to identify outlier loci (see Luikart et al. 2003, Storz 2005).

**QTL mapping of mRNA expression variation**

Linkage mapping of mRNA transcripts are used to identify particular regions of the genome that are associated with variation in gene expression levels (Jansen and Nap 2001, Doerge 2002). This method requires that microarrays are developed for the study species or a reasonably closely related species and thus this method may be of limited value for conservation biologists.

**Allele-specific mRNA expression analysis**

Estimation of expression levels of alternative alleles within heterozygous individuals based on polymorphism in the transcribed region of a gene (Buckland 2004, Knight 2004, Yan and Zhou 2004).

**QTL mapping of protein expression variation**

Linkage mapping and protein expression analysis are used to identify particular regions of the genome that are associated with variation in quantitative and qualitative protein expression levels within a pedigree (Gorg et al. 2004). This approach, as with the previous one, suffers limitations such as a requirement of a large amount of pedigree material and a reasonable amount of fresh tissue and may thus be of limited value in conservation studies. However, the study of gene regulation has proven important in studies of various stressors such as drought.
(de Vienne et al. 2001), which is of direct relevance for conservation.

**Environmental association analysis**
This analysis estimates significant associations between environmental variables and specific alleles. Such may be taken as evidence for directional selection affecting a particular locus. Studies can be temporal (by following cohorts in time) or spatial, even at small spatial scales (Johannesson et al. 1995).

**QTL analyses (linkage mapping)**
If there is genetic linkage map information (i.e. knowledge where and on which chromosomes markers are positioned), pedigree material to trace the segregation of the markers, and phenotypic data of individuals in a pedigree, it is possible to tests for association between markers and certain ecological traits of interest (Erickson et al. 2004, Slate 2005). This technique has been extensively used in model organisms and domesticated species (Andersson and Georges 2004).

**Admixture mapping**
This is a similar technique to the one above but here the association among and between populations and their experimental backcrosses are used to identify traits and genomic regions that are distributed non-randomly (Rieseberg and Buerkle 2002, McKeigue 2005).
Association analysis (linkage disequilibrium mapping)
Tests of a non-random association of a phenotypic trait of interest within families or populations and a certain genotype (or haplotype) by utilizing the non-random occurrence of alleles at linked loci, known as linkage disequilibrium (LD). This approach requires that a large part of the genome is covered by a large set of genetic markers. For example, in humans it has been suggested that 1 million random single nucleotide polymorphisms (SNPs) are needed to provide reasonable whole-genome coverage for association studies (Hirschhorn and Daly 2005). Clearly this is at present unreasonable for most species of conservation concern. However, most studies of non-model species using this approach focus on LD between a limited number of candidate genes. For example, associations between major histocompatibility complex (Mhc) genes and immune response have been identified (reviewed in Bernatchez and Landry 2003 and Garrigan and Hedrick 2003).

Luikart et al. (2003) introduced the concept of population genomics, which was defined as ‘the simultaneous study of numerous loci or genome regions to better understand the roles of evolutionary processes (such as mutation, random genetic drift, gene flow and natural selection) that influence variation across genomes and populations.’ They proposed the following research strategy: step 1, sample as many individuals from as many populations as possible; step 2, genotype as many loci as possible, preferably with an even spread throughout the genome; step 3, test for outlier loci (such as loci that have a greater- or smaller-than-average $F_{ST}$ values); step 4, on the neutral loci, compute evolutionary or demographic parameters without using outlier loci. On the candidate (adaptive) loci, test for causes of outlier behaviour (for example, selection) and use adaptive information for biodiversity conservation or evolutionary inferences (Fig. 7.2).

As an example of a study using the research outline by Luikart et al. (2003), studies from my own research group of willow grouse, Lagopus lagopus, may (p.128)
serve as an example. We sequenced 18 autosomal protein-coding loci from approximately 15–18 individuals in four populations (S. Berlin, M. Quintela, and J. Höglund, unpublished results). From these sequences we retrieved more than 100 independently segregating single nucleotide polymorphisms (SNPs; see below). We found unusually high levels of nucleotide diversity in Scandinavian willow grouse as well as very little population structure among localities that were up to 1647km apart. None of the loci diverged from neutral expectations. There were also low levels of linkage disequilibrium, even within the genes, and the population recombination rate was high, indicative of an old panmictic population, where recombination has had time to break up large haplotype blocks. Compared with the silent nucleotide diversity at third codon position, (p.129) the non-synonymous nucleotide diversity was low, which is in agreement with effective purifying selection, possibly due to the large effective population size. In birds nucleotide-level variation is poorly characterized; the domesticated chicken and a few passerine species are exceptions in this respect (Backström et al. 2006). However, these studies suggest that bird nucleotide diversity is high (approximately $10^{-3}$) and a likely explanation for this is the generally higher effective population sizes compared with mammals. Our findings in the willow grouse need to be repeated in other bird species and such studies should increase the number of both synonymous and non-synonymous SNPs. At present the non-synonymous substitutions are far too few to address any relevant questions about patterns of adaptive genetic variation in willow grouse or any other bird.

Figure 7.2 Flow chart indicating the steps in a population genomic investigation (from Luikart et al. 2003, reprinted with permission from the publisher).
species. With more markers and more species studied it will be possible to make general conclusions of levels of genetic variation in natural bird populations and to test the hypotheses about the distribution of adaptive and neutral variation. If local adaptation is important in birds it is predicted that local populations and subspecies will show higher levels of differentiation in adaptive genes than in neutral. If this is the case, threatened bird populations should be managed accordingly.

7.4 Genomics in conservation
Genomic applications and techniques are likely to become more and more prevalent as genomic data from endangered and related species become available. As noted previously in this chapter it is not likely that data from endangered species will lie at the forefront of ecological genetics but experimental conservation genetic studies are often conducted on genetically well-known species such as *Drosophila* (e.g. Bijlsma et al. 1999, 2000). In these circumstances genomic resources and techniques may turn out to be useful.

7.4.1 SNP detection and genotyping
SNPs have a number of features that make them desirable genetic markers in studies of genetic variation in natural populations. In a review of the use of SNPs in conservation studies, Morin and coworkers (2004) listed a number of applications number of applications where SNPs are likely to become useful. While SNPs are less polymorphic than the major alternative microsatellites, they are more abundant throughout the genomes. Furthermore SNPs evolve by a simpler mutational process as compared to microsatellites. The short stretches of repetitive DNA sequences of microsatellites evolve when the endogenous DNA polymerase in the cell makes a replication error and either misincorporates or mistakenly removes a copy (so-called DNA slippage). These stepwise mutations are much more likely than many other types of mutation (in the order of $10^{-3}$ (p.130) instead of $10^{-6}$ per genome and generation). The mutation process of SNPs, on the other hand, is simpler, as they evolve mainly by point mutation. By being less variable, more SNPs are needed than microsatellites to achieve resolution power in many applications. However, by being less variable the problems of homoplasy (the same allelic state evolving more than once in the sample) can effectively be avoided.
Protocols for developing microsatellite markers are quite straightforward and affordable for any conservation genetic study anticipating studying genetic variation. However, microsatellites seem to be rare and hard to develop in genomes of some organisms (e.g. butterflies and arthropods). Briefly, microsatellites are found by cutting up the genomic DNA from the study species with restriction enzymes and ligating this DNA into bacteria using a phagemid vector. This genomic library is then probed with a synthetic oligonucleotide mirroring a repeat sequence. The bacterial clones are allowed to grow and clones with positive inserts are sequenced. This will allow detection of not only the repeat sequence but also flanking regions around the repeat. In the next step primers in the flanking regions close to the repeat sequence are designed. With the aid of the primers, the target DNA can be manifolded in a PCR to screen allelic length variation in a large number of individuals.

As indicated above SNP discovery is a more complicated process and several strategies exist (see Morin et al. 2004, Slate et al. 2008 for reviews). One method is referred to as exon priming intron crossing (EPIC). Even in the absence of previous sequence data, primers can be designed by aligning of sequences in exons of protein-coding genes in related species that have been sequenced. A pair of primers are made of which the forward primer is designed in one exon and the backward primer in an adjacent exon so that the intron in between can be amplified and sequenced. By sequencing intronic DNA the chances of discovering segregating SNPs are maximized. A related approach is to use core anchor tagged sequences (CATS). This is a set of primers originally developed for gene mapping in diverse set of organisms and they were therefore chosen to be as conservative as possible. Some but not all of these primer pairs yield PCR products, mostly within coding genes (Lyons et al. 1997).

The second main approach to find SNPs is to sequence random genomic fragments in a limited number of individuals and the aligning the sequences. SNPs are found by identifying segregating sites in the alignment. Random clones from genomic DNA libraries may be sequenced or existing EST databases may be mined for SNPs using a number of different
computer programs. Obviously 454 sequencing will be useful in this endeavour.

SNP genotyping can be performed in house, but in a recent review outsourcing to specialized laboratories was recommended (Slate et al. 2008). A number of strategies and platforms for SNP genotyping are available and the choice of these depends on the number of samples and SNPs. High-throughput genotyping is made possible by the fact that most SNPs are biallelic, which can be utilized to streamline the genotyping to accommodate a large number of samples and many SNPs fast and cost-effectively.

SNPs can be used for estimating genetic variation. It is believed that by using a large number of SNPs a better and more representative estimate of genomic diversity may be obtained. However, this increased precision comes at a cost. Reliable estimates of genome-wide variation required four to ten times more biallelic amplified fragment length polymorphism (AFLP) markers with multi-allelic markers (Mariette et al. 2002). However, dominant AFLP markers are less informative than codominant biallelic SNP markers and thus fewer SNPs may be required compared with AFLP loci.

SNPs can be used in identifying individuals and to reveal parentage and relatedness. There are already established techniques for this using microsatellites and AFLPs, and it is not clear whether these applications would be improved by using SNPs. On the other hand, there are no indications that such analyses would be worsened by using SNPs.

Similarly, SNPs may be used in estimates of population structure. Estimating $F_{ST}$ from microsatellites can be problematic. Hedrick (2005) showed that the theoretical upper limit of an $F_{ST}$ estimate is not 1 (as established for biallelic loci); instead, the upper limit of a multilocus estimate ($G_{ST}$) may be considerably lower and Hedrick suggested a correction to remedy this effect. Since most SNPs chosen for analysis are biallelic this problem may be less significant for these markers.

Estimating changes in past population size as reviewed in Chapter 4 is one area in which SNPs will not be an
improvement. This is because these tests are more powerful with a higher the number of alleles at a locus.

7.4.2 QTL mapping of functionally important loci

As reviewed in Chapter 2, quantitative trait variation always has a genetic component (large or small depending on the trait and circumstances). It has been argued that traits subject to natural selection, and thus representing adaptations to local conditions (Chapter 5), are of particular importance in conservation. Finding and characterizing the molecular basis of QTLs is therefore potentially important.

The application where SNPs are used to their best advantage is in mapping of QTLs, because SNPs can be typed on a much larger scale and are much more abundant than microsatellites. Using SNPs, any genomic location may thus be analysed (Slate et al. 2008). Mapping means going from QTL location to finding candidate genes and is seemingly a simple process. Mapping experiments in model organisms such as laboratory rats, tomato, and cattle have shown that complex traits in inbred organisms may have a simple genetic architecture, with (p.132) only a handful of chromosomal regions that associate with any given trait (Flint and Mott 2001). However, we are a long way from providing similar results in non-model species. Even in model species it is difficult to find an statistical association between a marker locus and a trait. Markers typically are linked to an approximately 30cM region of a given chromosome and thus even if a candidate gene is found within that region there may be several candidates and so finding the responsible mutation, the quantitative trait nucleotide, is hard and requires complex research protocols and large sample sizes. At the very least, to map complex traits in natural populations a linkage map is required so that the chromosomal position of the markers is known. Linkage maps have started to appear in some non-models (Hansson et al. 2005) but are likely to be rare in the average threatened species subjected to a conservation genetic study.

7.4.3 Differential gene expression

Microarrays are generally used to quantify differences in global gene-expression patterns between groups of individuals. In conservation genetics, microarrays can be used to screen the transcriptome for genes that might be differentially expressed in relation to specific treatments or
coming from different populations (e.g. a threatened and a viable one). Therefore, this is a tool for detecting candidate genes for further study (e.g. Whitehead and Crawford 2006).

A classical microarray is a sample of small dots of known DNA (whole genes or parts of genes), usually collected on a glass plate or a silicon chip, to which cDNA or RNA can hybridized in a dot specific manner. The hybridizing material is labelled with fluorescent dyes and hybridized to the DNA dots. If the DNA sequence on the chip has been expressed, a corresponding sequence is present in the test sample and the relative expression levels can be read as relative intensities of the dyes for each dot. Several statistical tests have been utilized to identify differentially expressed genes from two different EST libraries (reviewed by Ruijter et al. 2002).

Kristensen et al. (2005) used a microarray to study expression differences among experimental groups of fruit flies, Drosophila melanogaster, subjected to various levels of inbreeding. They showed that inbreeding changed transcription levels for a number of genes. The genes that were differentially expressed were disproportionately involved in metabolism and stress responses, for example heat-shock proteins (Hsp), which are chaperones involved in folding and unfolding of intracellular proteins and macromolecules. Such genes are also upregulated during physiological stress and ageing. The results of this experiment suggest that inbreeding acts like an environmental stress factor.

As exemplified by this Drosophila study, microarray studies may be useful in studies relevant to conservation biology. However, microarrays are only available for a few organisms and although they may be constructed for any species it is not realistic that they will be developed for any but a few specialized studies (e.g. Vera et al. 2008). Microarrays developed for a model species may be tried in related threatened species (e.g. Abzhanov et al. 2006).

7.4.4 Phylogenetics

Genomic resources combined with phylogenetics provide insight in studies of phylogeographic patterns and intraspecific phylogenies. Such studies address whether recognized subspecies have been separated and if so for how long. A phylogenetic approach also provides better estimates of past effective population sizes (Edwards et al. 2007). These
Ecological genomics are important parameters for understanding local adaptation. How long and how much separation is needed for divergence among populations? These issues are of vital importance for inferring evolutionarily significant units and therefore defining management units. Multilocus genealogical approaches are still uncommon in phylogeography and historical demography, which have been dominated by microsatellite markers and chloroplast and mitochondrial DNA.

Theoretical studies of the coalescent process show that gene trees are not the same as species trees (Nichols 2001) and to estimate a species phylogeny the information from many genes are needed. For example, a research protocol was outlined by Liu and Pearl (2007) and Edwards et al. (2007) for estimating species trees as distinct from gene trees. In the so-called BEST approach, a Bayesian method for estimating species trees (Liu and Pearl 2007), vectors of gene trees are first estimated using a tree generated from a constrained set of preliminary species trees. These posterior distributions of gene tree vectors will provide the raw data for (maximum likelihood and) Bayesian estimation of phylogeny, population divergence times, and ancestral population sizes by exploring species tree space and maximizing the likelihood of gene tree vectors using the coalescent model of Rannala and Yang (2003). Ancestral population sizes can also be estimated with a method that uses gene tree–species tree conflicts in multilocus data sets (Nei 1987). Even incompletely resolved nuclear gene trees, when summed over multiple loci, can provide a strong signal for inference of demographic history (Jennings and Edwards 2005, Edwards et al. 2007).

With a subspecies tree and divergence time estimates it is possible to study how genes subjected to natural selection behave as compared with neutral gene and species trees. In the case of mammalian Mhc genes it has been observed that species share alleles that date back to before the split of the lineages (Edwards et al. 1995). The common explanation is that heterozygous genotypes are more fit, and hence extinction of alleles due to genetic drift is reduced drastically, resulting in long persistence times of alleles across speciation events. Long persistence times are also predicted by frequency-dependent models and other variants of balancing selection (Vekemans and Slatkin 1994). By contrast, some studies have found more divergence in Mhc alleles between populations than for neutral genes, suggesting strong
divergent selection, perhaps as responses to different habitats or parasite faunas, between species (Miller et al. 1997, Ekblom et al. 2007, Saether et al. 2007).
7.5 Genomic studies of non-model species

Despite the difficulties in applying molecular genetics to ecologically well-known non-model species, a few studies have made exceptional advances in this field. In the following I review a few of these studies.

Bonin and coworkers (2006) used AFLPs to screen genetic variation along an altitude gradient of populations of the common frog, *Rana temporaria*, in France. Among a large set of markers they identified four that were more differentiated (higher $F_{ST}$ values) than would be expected by random genetic drift (Fig. 7.3). They then subdivided their data into a neutral data set containing all the loci which behaved according to neutral expectations and a an altitude data set containing the outlier loci. When calculating phylogenetic relationships with the neutral data set, populations that were geographically close clustered together, as expected. However, with the altitude data, populations clustered according to altitude and not geographic distance. This strongly suggests that the genomic regions containing the AFLP sites in the altitude data have been subjected to similar forms of selection and may contain functional loci that have responded to selection for life at high altitudes. The genetic architecture of the common frog is, however, not established and no candidate functional loci have been suggested so far. Similar studies also using AFLPs have found outlier loci affected by selection on microgeographic differences among intertidal snails, *Littorina saxatilis* (Wilding *et al.* 2001, Butlin 2008).

In a series of papers, the genetic architecture behind morphological differences among benthic and marine forms of three-spined sticklebacks, *Gasterosteus aculeatus*, have been revealed (Peichel *et al.* 2004, Albert and Schluter 2004, Colosimo *et al.* 2004, Shapiro *et al.* 2004). Using a range of quantitative and molecular genetic techniques and experimental crosses, a complete linkage map has been published and the number of linkage groups equals the number of chromosomes in the species. With this map the researchers have shown that two important morphological differences among benthic and marine sticklebacks, armour plates and gill raker number, map to independent chromosomal regions and thus are controlled by different sets of genes. Furthermore, three aspects of skeletal armour are
controlled by only a few chromosomal regions suggesting that only a few major genes play an important role in shaping these differences (Fig. 7.4).

(p.135)

Similarly, shape differences are best described by a geometric model of adaptation which states that a few major QTLs with a major effect are involved together with many genes with small pleiotropic effects. The major gene suggested to be involved and which maps to one of the markers with a major effect is the EDA gene. In humans ‘the EDA gene provides instructions for making a protein called ectodysplasin A. This protein is part of a signalling pathway that plays an important role in development before birth. Specifically, it is critical for interactions between the ectoderm and the mesoderm embryonic cell layers. In the early embryo, these cell layers form the basis for many of the body’s organs and tissues’ (NIH Genetics Home Reference; http://ghr.nlm.nih.gov).

Another major difference between marine and benthic sticklebacks is the pelvic reduction that occurs in freshwater benthic forms. Previous studies had suggested that pelvic

Figure 7.3 Plot of $F_{ST}$ values against heterozygosity estimates comparing a high- and a low-altitude population. Each dot indicates an AFLP marker. The lower, intermediate, and higher lines represent the 5, 50, and 95% confidence intervals, respectively. Outlier loci are pointed out by arrows and referred to by numbers (from Bonin et al. 2006, reprinted with permission from the publisher).
structures protect sticklebacks against gape-limited, soft-mouthed predators by presenting a lacerating defensive structure, increasing the effective diameter of the fish and resisting compressive forces during predator manipulation and chewing. However, several freshwater stickleback populations (p.136) have evolved complete or partial loss of the pelvic skeleton, perhaps in response to local absence of predatory fish (Shapiro et al. 2004). It was shown that the pelvic reduction was controlled by one major and four minor QTLs. The major gene involved was Pitx1, a gene expressed specifically during hindlimb development in mice, and which is required for normal hindlimb development in traditional vertebrate model systems. However, the sticklebacks did not show changes in Pitx1 protein sequence. Instead, pelvic-reduced sticklebacks showed site-specific regulatory changes in Pitx1 expression, with reduced or absent expression in pelvic and caudal fin precursors. It was thus suggested that regulatory mutations in major developmental control genes may provide a mechanism for generating rapid skeletal changes in natural populations, while preserving the essential roles of Pitx1 in other processes. The adaptive radiation of the Darwin’s finches is one of the textbook examples of how underlying differences in ecological conditions may cause divergent selection on beak morphologies and thus drive the speciation process in a group of birds of common origin (Lack 1947, Grant 1986). The genetic architecture of beak morphology have been studied in

---

**Figure 7.4** Mapping the genetic basis of lateral plate reduction in populations of three-spined sticklebacks. Dots show the geographic origins of the populations studied. AA, Aa, and aa refer to genotypes at Gac4174 (a microsatellite marker) near the major plate locus (Colosimo et al. 2004, reprinted with permission from the publisher).
domestic chicken (Wu et al. 2004) and Abzhanov and coworkers (2004) studied one candidate gene, bone morphogenetic protein 4 (Bmp4), previously identified in chicken as a major gene involved in beak morphogenesis, in a set of species of Darwin's finches. By (p.137) performing comparative analysis of expression patterns in the different species, they identified variation in the level and timing of Bmp4 expression that correlated with variation in beak morphology. During development, Bmp4 is strongly expressed in a broad distal-dorsal domain in the mesenchyme of the upper beak prominence in species with large and broad beaks. The authors speculate that differences in the cis-regulatory elements of Bmp4 may underlie the distinct expression patterns found. This study is thus a classic example of the candidate gene approach (Box 7.1) in which a gene identified in a model (in this case the chicken) is shown also to be involved in the development of similar morphologies in non-model species.

Figure 7.5 Phylogeny of Geospiza finches from the Galapagos islands with the different beak morphologies superimposed. The gene CaM is differentially expressed in the distal-ventral domain in the mesenchyme of the large-beaked species (from Abzhanov et al. 2006, reprinted with permission from the publisher).
In another study Abzhanov and coworkers used another approach to study the genetic background of differences in beak morphology in the same group of birds. By using a chicken microarray they were able to show that another protein, calmodulin (CaM), is differently expressed in Darwin’s finches and that expression levels correlate with beak morphology (Fig. 7.5, Abzhanov et al. 2006). Calmodulin is a protein that binds and activates certain enzymes that trigger a signal which eventually turns specific genes on or off. Again this result suggests that regulatory genes and gene products play a major role in the evolution of divergent morphologies.

7.6 Conclusions
This chapter has reviewed the advances in genomics and their applications to conservation genetics. Tying back to Chapter 4 and the discussion on invasive species, Lee (2002) stated: ‘the utility of genomic approaches for determining invasion mechanisms [are elucidated], through analysis of gene expression, gene interactions, and genomic rearrangements that are associated with invasion events.’ She emphasized the utility of exploring genomic characteristics of invasive species, such as genes, gene complexes, and epistatic interactions, that promote invasive behaviour and concluded that such information could yield insights into the relationship between genetic architecture and rate of evolution, and evolutionary and ecological factors which confer invasion success. There is thus great hope in the new technologies opened up by the advances in genomics. Conservation and evolutionary biologists will now doubt lag behind the specialists working on model organisms, but the new techniques already have and will continue to have a major impact on studies in both conservation and evolution. Nevertheless, it is only by also studying adaptations in the field that we can gain a deeper understanding of how life forms have evolved and how we should best preserve biodiversity for future generations.
Access brought to you by: