

REVIEW

Applications of next generation sequencing in molecular ecology of non-model organisms

R Ekblom and J Galindo

Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK

As most biologists are probably aware, technological advances in molecular biology during the last few years have opened up possibilities to rapidly generate large-scale sequencing data from non-model organisms at a reasonable cost. In an era when virtually any study organism can 'go genomic', it is worthwhile to review how this may impact molecular ecology. The first studies to put the next generation sequencing (NGS) to the test in ecologically well-characterized species without previous genome information were published in 2007 and the beginning of 2008. Since then several studies have followed in their footsteps, and a large number are undoubtedly under way. This review

focuses on how NGS has been, and can be, applied to ecological, population genetic and conservation genetic studies of non-model species, in which there is no (or very limited) genomic resources. Our aim is to draw attention to the various possibilities that are opening up using the new technologies, but we also highlight some of the pitfalls and drawbacks with these methods. We will try to provide a snapshot of the current state of the art for this rapidly advancing and expanding field of research and give some likely directions for future developments.

Heredity (2011) **107**, 1–15; doi:10.1038/hdy.2010.152; published online 8 December 2010

Keywords: ecological genomics; 454 sequencing; NGS; digital transcriptomics; SNP

Introduction

Several very different technologies constitute next generation sequencing (NGS), each of which has its own set of characteristics (Table 1). NGS rapidly generates huge amounts of sequence data in a very cost-effective way, and molecular ecologists are now starting to take advantage of this democratization of sequencing and embracing the discipline of 'ecological genomics' (Gilad *et al.*, 2009). By shifting the realms of genomics from laboratory-based studies of model-species towards studies of natural populations of ecologically well-characterized organisms, researchers can now start to address important ecological and evolutionary questions on a scale and precision that was unrealistic only a few years ago. We will not go into any technical details of NGS because this has already been extensively reviewed elsewhere (Hudson, 2008; Morozova and Marra, 2008; Shendure and Ji, 2008).

NGS can follow either a genomic or a transcriptomic route (Figure 1). In the latter, complementary DNA (cDNA) (see Glossary) is produced from the mRNA of a specific tissue or life stage. By this approach, data will be obtained on nucleotide variation as well as transcriptome (see Glossary) characteristics and gene expression levels. NGS allows for nucleotide variation profiling and large-scale discovery of genetic markers, which in turn will aid in the pursuit of the genetic basis of ecologically important

phenotypic variation through quantitative trait loci (QTL) mapping (see Glossary) or genome-wide association studies (GWAS). Studies of population history and demography, genetic structure and inference of relatedness will also be greatly improved. Genome-wide scans (for example, outlier analysis; see Glossary) and comparative genomics using NGS data will provide better chances of identifying loci under selection. Finally studies of gene regulation, DNA–protein interactions and epigenetics are also facilitated by the use of NGS, enabling molecular ecologists to venture into these areas of growing interest.

Genomic surveys in non-model species are much aided if there are genomic resources available for a related species (referred to as genomic reference species; see Glossary), for example, for assembly and functional annotation purposes (Wheat, 2010). With a growing number of species with sequenced genomes, this approach will be feasible for many non-model species. Such 'genome-enabled taxa' include a large number of species with ecological and/or conservation relevance (Kohn *et al.*, 2006). The increased number of whole genome sequencing (WGS) projects means that more and more ecological model organisms, for example, *Daphnia* (Eads *et al.*, 2007), *Mimulus* (Wu *et al.*, 2007) and sticklebacks (*Gasterosteus*; Hohenlohe *et al.*, 2010) are becoming important genetic models (Mitchell-Olds *et al.*, 2008). At the same time, more molecular ecology studies are focusing on natural variation and adaptation in classical genetic model species, or close relatives of these, like *Drosophila* (Nolte and Schlötterer, 2008), *Mus* (Teeter *et al.*, 2008) and *Arabidopsis* (Metcalf and Mitchell-Olds, 2009), thus closing the gap between model and non-model organisms from this end as well. In the following paragraphs, we will outline various applications of NGS

Correspondence: Current address: Dr R Ekblom, Department of Population Biology and Conservation Biology, Uppsala University, Norbyvägen 18 D, Uppsala SE-75256, Sweden.

E-mail: robert.ekblom@ebc.uu.se

Received 4 March 2010; revised 10 September 2010; accepted 2 November 2010; published online 8 December 2010

Table 1 Currently available NGS technologies and their characteristics

Technology	Sequencing method	Major advantages for studies of non-model organisms	Major disadvantages for studies of non-model organisms
Roche 454	Pyrosequencing	Relatively long reads enables assembly of contigs even in the absence of a reference genome	Relatively few reads results in shallower coverage of sequencing. High error rate, especially in homopolymers
Illumina/Solexa	Sequencing by synthesis	Very deep coverage because of large number of reads gives accurate measurements of gene expression levels	Short read length means that a reference genome is desirable for assembly
ABI SOLiD	Sequencing by ligation	Large number of reads and very deep coverage. Low error rate because of duplicate sequencing of each base pair	Short read length means that a reference genome is desirable for assembly. Data in colour space complicates downstream analysis and bioinformatics applications
Helicos tSMS	Single-molecule sequencing	Sequencing of single molecules means that no amplification step is needed, reducing bias in studies of expression levels or allele frequencies	New and still untested in studies of non-model organisms

Abbreviation: NGS, next generation sequencing.

For more details about the specific techniques, see, for example, Hudson (2008).

that will aid molecular ecologists and evolutionary biologists in their research. We will briefly summarize recent studies in non-model organisms (Table 2) that have taken advantage of NGS to answer important ecological and evolutionary questions and also briefly discuss various future prospects of this 'genomic revolution' in molecular ecology.

Transcriptome characterization

Currently the most common application of NGS in non-model species is transcriptome characterization (see Table 2). By this we mean generally describing what genes are expressed in a certain tissue, life stage or organism as well as functional characterization of these. Here, cDNA is synthesized by reverse transcription of mRNA and then sequenced. The first study describing the transcriptome in a non-model species through NGS was performed on the wasp *Polistes metricus* (Toth *et al.*, 2007). Here, genome information from the related honey bee was used as template (see Glossary) for mapping the reads (see Glossary) from 454 sequencing and for downstream analysis. In the Glanville fritillary butterfly (*Argynnis aglaja*), however, the transcriptome was assembled *de novo* (see Glossary) without the help of a closely related reference genome (Vera *et al.*, 2008). The fritillary butterfly is a text-book example of a species with complex meta-population dynamics (Saccheri *et al.*, 1998), and the aim of the genomics approaches recently employed in this system is to understand the genetics behind the variation in dispersal and colonization abilities seen between individuals. Since these first ground-breaking studies, sequencing and successful assembly of the transcriptome via 454 sequencing has been performed in a number of non-model organisms (Table 2). The first (to the best of our knowledge) study to use Illumina/Solexa sequencing data in a non-model species used a combination of *de novo* assembly and genomic reference species mapped assembly to study the transcriptome of the polyploid plant *Pachycladon enysii* (Collins *et al.*, 2008). However, since then, complete *de novo* assembly of Illumina/Solexa data has also been accomplished (Birol *et al.*, 2009).

Most studies characterizing transcriptomes so far have been very descriptive by nature, but they provide an

important starting point and a valuable resource for further analysis and ecological applications (Ellegren, 2008). For example, these sequences may be used as an assembly template (reference sequence) for further in-depth transcriptome re-sequencing and surveys of genetic variation. They may also be used to develop molecular markers, create targeted sequencing assays or to construct microarrays (see Glossary) for gene expression profiling, and to study alternative splicing (Harr and Turner, 2010), a phenomenon likely to be involved in processes of adaptation and speciation.

After the novel transcriptome has been annotated using a genomic reference species or publicly available sequence databases (for example, Genbank, Ensembl and UniProt), it can be used as a starting point for more detailed functional characterization, such as annotation using gene ontology (see Glossary) databases. As an example of this, (Dassanayake *et al.*, 2009) used transcriptome characterization of two species of mangrove to investigate convergent evolution of gene expression. Shared transcriptomic profiles between species may of course be a result of a common evolutionary origin or a joint distribution (Nuzhdin *et al.*, 2004). However, in the case of the two mangrove species investigated by Dassanayake *et al.*, this is not likely to explain the pattern seen. These two mangroves are not evolutionary related to each other but belong to very different lineages (one of the species is more related to *Arabidopsis* and the other more related to *Populus* than they are to each other). They also inhabit distinctly different areas, one being neotropical and the other Indo-West Pacific in the distribution. Despite of these differences between the two mangroves studied, they shared many functional characteristics of their transcriptomes (but differed substantially from *Arabidopsis* and *Populus*), probably resulting from parallel adaptation to a similar environment. Researchers might also be interested in functional comparisons between different sexes, life stages or tissues within the same species. For example, studying which transcripts are tissue specific or if certain pathways are overrepresented in a specific development stage can have important evolutionary implications. As a result of the cDNA library construction methods and the currently available read lengths, there is a bias of the transcripts towards the 3'-end and full-length transcripts are difficult to reach.

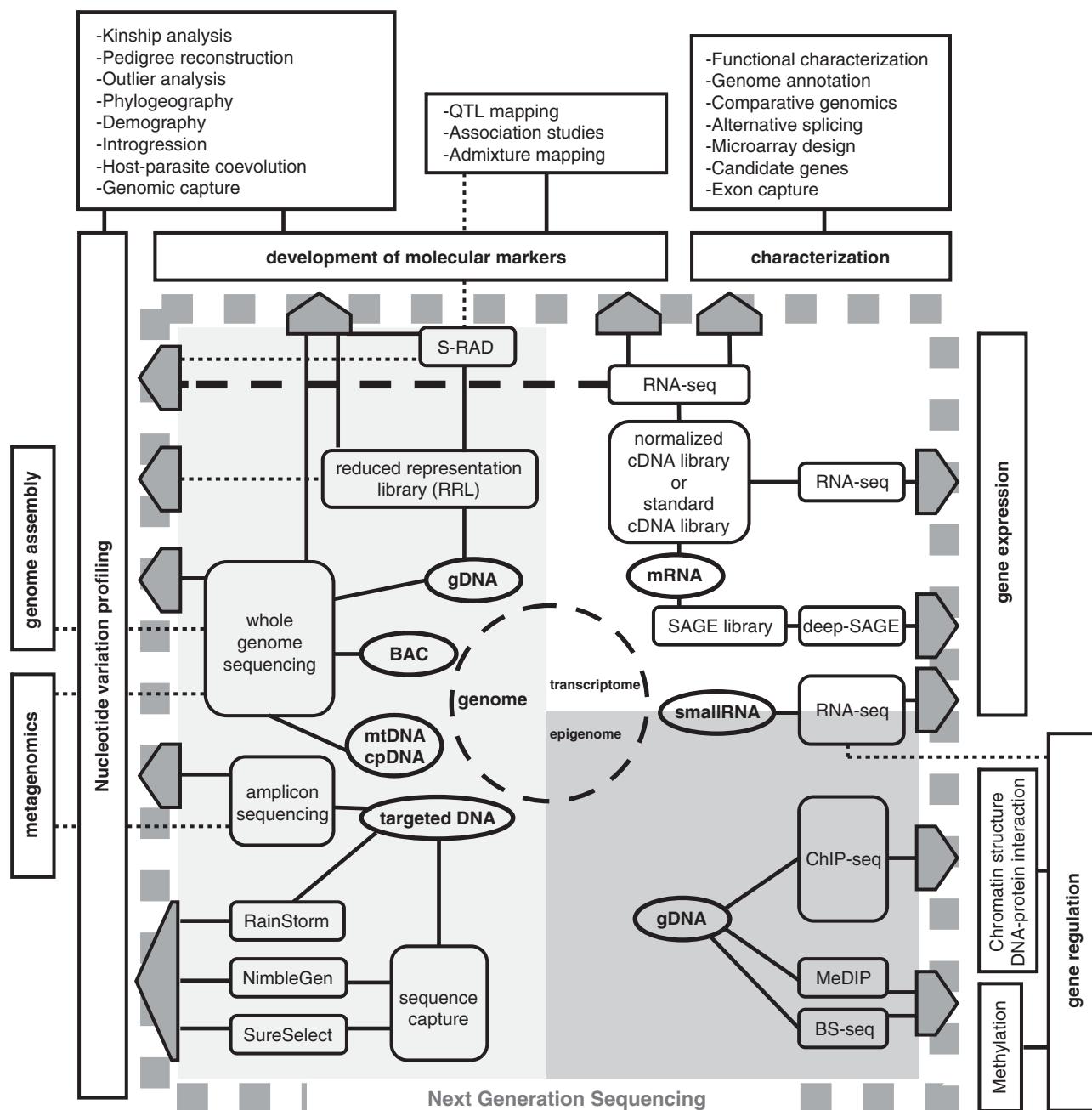


Figure 1 Scheme showing the workflow from sample to applications of NGS in molecular ecology. We considered three different sources of genetic variation, shown in the central circle of the diagram and indicated by different background shades: genome (light grey), transcriptome (white) and epigenome (grey). Genomic DNA (gDNA), mitochondrial or chloroplast DNA (mtDNA, cpDNA), vectors (for example, BAC), mRNA, non-coding small RNAs or target regions of the genome (targeted DNA) are the samples regarded as starting material (marked with thick circles). Different steps that can be performed before or during sequencing are shown within the grey dotted square representing NGS. gDNA samples, for example, can be used to produce reduced representation libraries (RRLs) and these can be either sequenced directly or used to generate S-RAD markers (sequenced restriction-site associated DNA). Targeted DNA can be generated in several different ways: PCR through amplicon sequencing, microdroplet PCR with RainStorm (RainDance Technologies, Lexington, MA, USA) or through sequence capture (NimbleGen, Roche; SureSelect, Agilent Technologies). mRNA can be used in a deep-serial analysis of gene expression (SAGE) approach or RNA-seq can be performed. gDNA can also be used in epigenetic studies of DNA–protein interactions through ChIP-seq or for studying methylation patterns through BS-seq (high-throughput bisulfite sequencing). The main applications (placed outside of the grey dotted ‘NGS’ square) of NGS are gene regulation, expression, transcriptome characterization, development of molecular markers (SNPs, microsatellites, InDels), nucleotide profiling and genome assembly. Each one of these main applications can then be divided further into many interesting areas of molecular ecology (kinship analysis, QTL mapping, and so on). An exceptional case is the use of S-RAD markers that can be used directly in mapping studies without the need of genotyping new developed molecular markers.

This is currently a problem when studying gene structure or when trying to sequence a complete gene (for example, candidate genes).

For many transcriptome characterization studies, it is preferable to have as broad a representation of the transcriptome as possible. This can be accomplished by

Table 2 Some early case studies for various applications of NGS in non-model organisms

Study organism	Platform	Transcriptome characterization	Expression level	Genetic variation	SNP finding	Microsat finding	Microarray construction	Candidate gene finding	Whole genome assembly	Reference
Wasp (<i>Polistes metricus</i>)	454	X	X					X		Toth <i>et al.</i> , 2007
Glanville fritillary butterfly (<i>Melitaea cinxia</i>)	454	X			X	X	X			Vera <i>et al.</i> , 2008; Wheat, 2010
Garden pea (<i>Pisum sativum</i>)	454	X								Bräutigam <i>et al.</i> , 2008
Bigmouth bass (<i>Micropterus salmoides</i>)	454						X			Garcia-Reyero <i>et al.</i> , 2008
Tobacco hornworm (<i>Manduca sexta</i>)	454	X							X	Zou <i>et al.</i> , 2008
Chick pea (<i>Cicer arietinum</i>)	454	X	X							Molina <i>et al.</i> , 2008
Oomycete (<i>Pythium ultimum</i>)	454	X					X			Cheung <i>et al.</i> , 2008
Gum tree (<i>Eucalyptus grandis</i>)	454	X			X	X				Novaes <i>et al.</i> , 2009
Red mangrove (<i>Rhizophora mangle</i>)	454	X								Dassanayake <i>et al.</i> , 2009
Looking-glass mangrove (<i>Heritiera littoralis</i>)	454	X								Dassanayake <i>et al.</i> , 2009
Plant (<i>Pachycladon enysii</i>)	Solexa	X								Collins <i>et al.</i> , 2008
Copperhead snake (<i>Agkistrodon contortrix</i>)	454						X			Castoe <i>et al.</i> , 2009
Lake sturgeon (<i>Acipenser fulvescens</i>)	454	X			X	X				Hale <i>et al.</i> , 2009
Eelpout (<i>Zoarces viviparous</i>)	454	X	X					X	X	Kristiansson <i>et al.</i> , 2009
California poppy (<i>Eschscholzia californica</i>)	454	X								Wall <i>et al.</i> , 2009
Avocado tree (<i>Persea americana</i>)	454	X								Wall <i>et al.</i> , 2009
Chestnut trees (<i>Castanea spp.</i>)	454	X	X						X	Barakat <i>et al.</i> , 2009
Condor (<i>Gymnogyps californianus</i>)	454	X							X	Romanov <i>et al.</i> , 2009
Moa birds (<i>Dinornithiformes</i>)	454					X				Allentoft <i>et al.</i> , 2009
Flesh fly (<i>Sarcophaga crassipalpis</i>)	454	X			X	X	X			Hahn <i>et al.</i> , 2009
Poplar leaf beetle (<i>Chrysomela tremulae</i>)	454	X							X	Pauchet <i>et al.</i> , 2009
Fungus (<i>Fusarium circinatum</i>)	454						X			Santana <i>et al.</i> , 2009
Sirex woodwasp (<i>Sirex noctilio</i>)	454						X			Santana <i>et al.</i> , 2009
Nematode (<i>Deladenus siricidicola</i>)	454					X				Santana <i>et al.</i> , 2009
Roughfruit amaranth (<i>Amaranthus tuberculatus</i>)	454					X			X	Lee <i>et al.</i> , 2009
Coral (<i>Acropora millepora</i>)	454	X			X			X	X	Meyer <i>et al.</i> , 2009
Bank vole (<i>Myodes glareolus</i>)	454			X						Babik <i>et al.</i> , 2009; Kloch <i>et al.</i> , 2010
Sweet wormwood (<i>Artemisia annua</i>)	454	X		X						Wang <i>et al.</i> , 2009a
Sugarcane (<i>Saccharum officinarum</i>)	454				X				X	Bundock <i>et al.</i> , 2009
Emperor scorpion (<i>Pandinus imperator</i>)	454	X							X	Roeding <i>et al.</i> , 2009
Apple maggot (<i>Rhagoletis pomonella</i>)	454	X	X		X	X			X	Schwarz <i>et al.</i> , 2009
Scuttle fly (<i>Megaselia scalaris</i>)	454					X				Rasmussen and Noor, 2009
Giant panda (<i>Ailuropoda melanoleura</i>)	Solexa				X	X			X	Li <i>et al.</i> , 2010
Tobacco hornworm (<i>Manduca sexta</i>)	454	X							X	Pauchet <i>et al.</i> , 2010
Blue mussel (<i>Mytilus galloprovincialis</i>)	454	X	X						X	Craft <i>et al.</i> , 2010
Stickleback (<i>Gasterosteus aculeatus</i>)	Solexa				X	X			X	Hohenlohe <i>et al.</i> , 2010
Birds (10 different species)	454	X		X						Künstner <i>et al.</i> , 2010
Cichlid fish (<i>Amphilophus spp.</i>)	454	X		X						Elmer <i>et al.</i> , 2010
Trees (10 different species)	Solexa			X						Cannon <i>et al.</i> , 2010
Whitefish (<i>Coregonus spp.</i>)	454			X	X	X				Renaut <i>et al.</i> , 2010
Moscow salsify (<i>Tragopogon porrifolius</i>)	454/ Solexa		X	X	X					Buggs <i>et al.</i> , 2010
Great tit (<i>Parus major</i>)	Solexa				X				X	van Bers <i>et al.</i> , 2010
Lake trout (<i>Salvelinus namaycush</i>)	454	X	X					X		Goetz <i>et al.</i> , 2010
Crow (<i>Corvus corone</i>)	454	X	X	X						Wolf <i>et al.</i> , 2010
Postman butterfly (<i>Heliconius melpomene</i>)	454	X							X	Ferguson <i>et al.</i> , 2010

Abbreviations: NGS, next generation sequencing; SNP, single-nucleotide polymorphism.

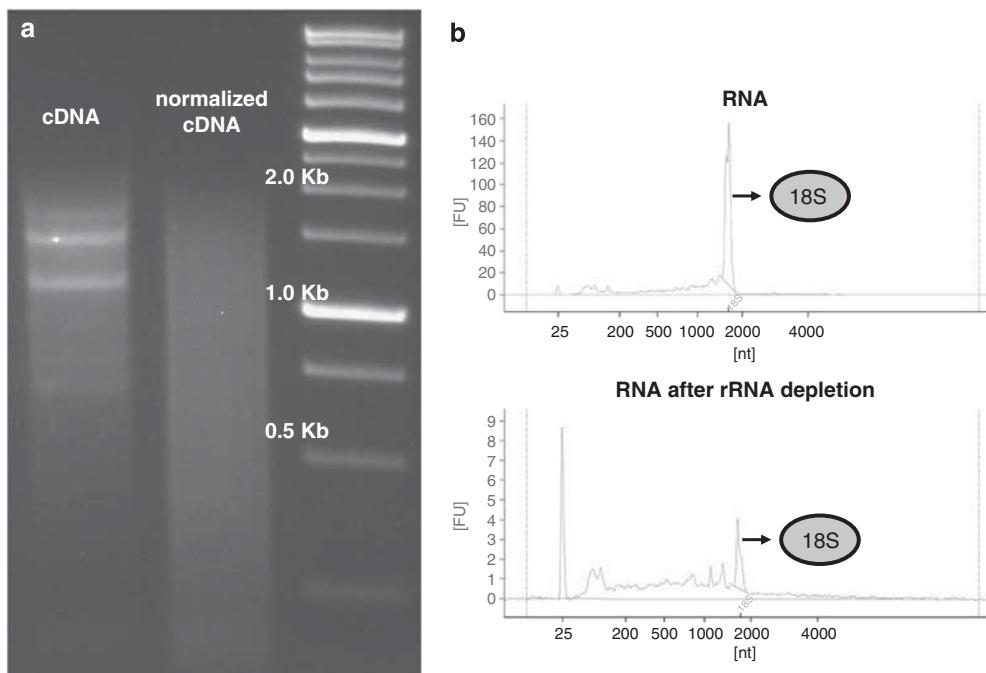


Figure 2 cDNA library normalization and ribosomal RNA (rRNA) depletion from total RNA. (a) Agarose gel showing the same double-strand cDNA sample before and after normalization. Normalization decreases the prevalence of highly abundant transcripts (seen as distinctive bands in the un-normalized sample) and equalizes mRNA concentrations in the cDNA library. It will increase the number of genes covered, but unless the sequencing quantity is high, fewer genes will be fully covered. Normalization consequently increases the coverage of most of the sequenced transcript. There is also an increase in the gene discovery rate in a normalized cDNA library, enhancing the identification and analysis of rare transcripts (Cheung *et al.*, 2006). Several normalization methods are reviewed in Bogdanova *et al.* (2009) but duplex-specific nuclease (DSN) normalization (Zhubilov *et al.*, 2004) has been widely used in recent years. (b) Analysis of RNA with an Agilent 2100 Bioanalyzer (Agilent Technologies) using a pooled RNA sample and the same sample after ribosomal RNA (rRNA) depletion with RiboMinus RNA-seq kit (Invitrogen). Concentrations are showed in fluorescence units (FU) and size in basepairs (nucleotides, nt). Notice the dramatic change of scale of the y axis, which is due to the reduced amount of 18s rRNA in the depleted sample.

pooling of several tissues, life stages and/or individuals when producing the cDNA library (Hahn *et al.*, 2009). A complementary way of increasing the breadth of the transcriptome coverage is normalization of the cDNA library (Figure 2a) and the depletion of ribosomal RNA (Figure 2b). The effect of this is to reduce the representation of very common transcripts (Zhubilov *et al.*, 2004). Although gene discovery may still only be marginally more efficient than for non-normalized libraries (Hale *et al.*, 2009), normalization will increase the depth of coverage for most transcripts, which is very valuable for nucleotide variation profiling and single-nucleotide polymorphism (SNP) discovery. However, normalization procedure will introduce biases in the relative representation of different transcripts, making estimates of gene expression and allele frequencies from such data less reliable. Nevertheless, normalized cDNA libraries have still been useful in some cases to this end (for example, Schwarz *et al.*, 2009; Ekblom *et al.*, 2010). When planning the experiment it is also valuable to note that there is a trade off between the cost of normalization and the cost of sequencing (Wall *et al.*, 2009). As sequencing technology keeps improving and the quantity of sequence yield increases, cDNA normalization may soon be superfluous for transcriptome characterization, as a single sequencing run will be able to cover the complete transcriptome of the sampled tissue, regardless of the distribution of transcript abundances.

Gene expression profiling

In gene expression profiling, the aim is not only to characterize what genes are expressed but also to investigate the specific level (absolute or relative) of gene expression. Traditionally this has been accomplished using microarrays (Kammenga *et al.*, 2007), and has thus mainly been restricted to model species with previous genome information. However, as mentioned above, microarrays can now be constructed for non-model species using data from NGS transcriptome profiling (Garcia-Reyero *et al.*, 2008), or used in a cross-specific way if developed in related genome reference species (Bar-Or *et al.*, 2007). Development of microarrays through NGS considerably reduces both the cost and the effort involved.

An attractive alternative to microarrays for gene expression studies in non-model organisms is sequencing-based expression profiling, or digital transcriptomics (Murray *et al.*, 2007). The rationale behind this approach is that the representation of specific sequences derived from deep cDNA sequencing is proportional to the amount of RNA from the gene in the original sample (t Hoen *et al.*, 2008). Basically, there are two different versions of this kind of analysis. Either more or less random parts of whole transcripts are sequenced directly (RNA-Seq; see Glossary), or specific parts of transcripts are cut out using restriction enzymes and then sequenced. Restriction enzyme-based methods, for

example deep serial analysis of gene expression (see Glossary) (Nielsen *et al.*, 2006) always outputs the same sequence tag for a given transcript, facilitating data analysis. The utility of this approach for surveying gene expression in a non-model organism was recently verified in a study of drought stress responses of chick pea (*Cicer arietinum*) roots (Molina *et al.*, 2008). The main drawback is that the tags need to be mapped to a reference sequence. In contrast, RNA-seq methods produce novel cDNA sequences across the whole range (or more random parts) of the transcript (Nagalakshmi *et al.*, 2008; Wilhelm and Landry, 2009; Wang *et al.*, 2009b). They have the benefit of determining gene expression patterns and characterizing the transcriptome at the same time. Thus, this approach also enables gene characterization, molecular marker finding and detection of splice variants among other applications (Simon *et al.*, 2009).

Digital transcriptomic approaches typically do not suffer from high background noise and cross-hybridizations as are common for microarrays. Furthermore these approaches are more efficient at detecting very rare transcripts and variation in highly expressed genes (because of higher resolution) and the analyses therefore have a greater dynamic range. Another advantage of digital transcriptomics over microarrays is the ability to detect expression levels in previously unknown genes (Nielsen *et al.*, 2006). In a recent study, there was a very high correlation between the number of 454 reads mapping to a gene and microarray determined gene expression of it (Kristiansson *et al.*, 2009), thus validating the RNA-Seq methodology for non-model organisms. The correlation in this study was surprisingly strong considering that the cDNA library used for the digital transcriptomic analysis was normalized. Gene expression profiling using an RNA-Seq approach was also performed in a study of blight resistance in different chestnut species (*Castanea* spp.). Several different genes were found to be differentially expressed in infected versus healthy tissues indicating a function in pathogen response (Barakat *et al.*, 2009). This study shows the strength of genomic approaches for addressing questions related to pathogen resistance and host-parasite co-evolution. Another study investigated the genomics of speciation by comparing expression profiles between samples from a hybrid zone of two subspecies of the crow (*Corvus corone*) with samples from a pure population of one of these subspecies (Wolf *et al.*, 2010). The investigators concluded that there is strong divergence between these in terms of gene expression levels, although there seem to be very little genetic sequence divergence in this system, and suggest that variation in expression may prove to be an important factor in the early stages of speciation processes.

NGS allows for gene expression studies in species without previous genomic resources, thus RNA-Seq is, for example, a very promising application for the study of adaptation. Determining which set of genes are differentially expressed because of adaptation to certain environmental conditions represents an important question in molecular ecology and represents a first step to better understand the genetics behind the regulatory mechanisms involved. Although RNA-Seq represents a favourable strategy in many studies of gene expression it becomes an expensive route when many different sets of conditions and/or populations need to be tested for gene

expression (for example, gene expression for different temperature conditions). In these cases, a more economic strategy would be using transcriptome information from NGS to design a microarray and then use these microarrays to study gene expression. It is, however, likely that in the future individual sample-tagging approaches and increased efficiencies of RNA-Seq can overcome this downside.

Candidate gene finding

Information about specific genes of interest can be mined from NGS transcriptome or genome data of non-model organisms using coding nucleotide or protein sequence information from genomic reference species. For example, (Toth *et al.*, 2007) focused on candidate genes for food provisioning and foraging behaviour in their study of a primitively eusocial wasp (*Polistes metricus*). By 454 sequencing of the transcriptome and aligning the reads to the honey bee genome, they were able to annotate these candidate genes and perform follow up real-time quantitative PCR experiments to study differential gene expression in different social casts of the wasp species. The results supported the hypothesis that provisioning behaviour is linked to the evolution of eusociality. In another early study, several hundred genes related to immunity and defence responses were identified in the tobacco hornworm (*Manduca sexta*). This species is a well-known model for insect physiological processes but before this 454 based sequencing study there were very limited genomic resources. This data thus represent an important stepping stone in understanding the functional basis of insect pathogen resistance (Zou *et al.*, 2008).

The candidate gene approach has also been widely used in conservation genetics (Höglund, 2009; Primmer, 2009). With the goal of conserving functionally important genetic information, studies have used genetic structure in ecologically relevant loci to identify taxonomic units of conservation interest (Hedrick, 2004; Ekblom *et al.*, 2007). NGS has great potential to open up this approach to conservation genetics to more species and include analyses of a larger number of potentially important genes. The eelpout (*Zoarces viviparous*) is a fish species commonly used in environmental monitoring and ecotoxicological studies. It recently had its transcriptome characterized using 454 sequencing and a number of biomarker genes for ecotoxicology were specifically identified (Kristiansson *et al.*, 2009), thus providing an important tool for future studies of the genetic basis for physiological responses to pollutant exposures.

When performing transcriptome sequencing with the aim of detecting candidate genes it is important to make sure that RNA from the right tissues and/or life stages are used. Genes under positive selection (therefore likely to be genes of interest) are expressed in a more tissue-specific manner compared with evolutionary conserved genes (Zhang and Li, 2004; Ekblom *et al.*, 2010) and could thus easily be missed if coverage is too low.

Whole genome sequencing (WGS)

A vast majority of studies using NGS are re-sequencing already fully described genomes (Wheeler *et al.*, 2008). As this is not possible for non-model organisms, it will

not be discussed further here. However, as was recently demonstrated by work on the giant panda (*Ailuropoda melanoleuca*), NGS may also be utilized for *de novo* WGS of large and complex genomes (Li *et al.*, 2010). Although more genomes will be sequenced through NGS and released in the near future, the costs, expertise and infrastructure required for data collection, analysis and output handling for this kind of application are still beyond reach for most molecular ecology research groups. However, large research centres will be carrying out an impressive amount of WGS projects in the next few years. One example is the 1000 plant and animal reference genomes project (http://ldl.genomics.org.cn/page/introduction_A&P.jsp), which aims to sequence 1000 economical and scientific important species in 2 years.

An alternative approach to WGS in species without a characterized reference genome may be to use NGS to generate a large amount of sequence data, and to analyse this without attempting a full genome assembly. This approach was taken in a study of mammoth genomics, in which previously generated Sanger sequence data and the sequenced genome of the African Elephant (*Loxodonta africana*) was used in annotation and analysis of 454 sequence reads (Miller *et al.*, 2008). Recently, a large amount of genome sequence data was produced by Illumina/Solexa sequencing of the great tit (*Parus major*). The investigators used this to assemble about 2.5% of the genome, after using a reduced representation library (see Glossary) strategy to increase the coverage of the sequenced fraction of the genome, which was then used for downstream SNP discovery (van Bers *et al.*, 2010). NGS can also be used to sequence bits of genomes after cloning these into bacterial artificial chromosomes (BACs). However, the assembly of 454 sequencing data of eight different BACs from Atlantic salmon (*Salmo salar*) was found to be problematic (Quinn *et al.*, 2008). The increase in the length of the sequences obtained with the current technologies in combination with the use of paired end sequencing will improve the quality of this type of assemblies in the future (DiGuistini *et al.*, 2009; Rounseley *et al.*, 2009). Targeted NGS of BACs will be a valuable resource for studying chromosomal regions involved in adaptation (Baxter *et al.*, 2010).

Targeted sequencing

The NGS applications reviewed so far generate an impressive amount of sequence data, but in terms of population genetics (for example, variation between individuals and populations) the amount of information is limited. For such applications, it is more informative to use NGS to sequence a limited number of targeted loci. By decreasing the number of targets, the coverage is considerably increased, and consequently more valuable information for population analyses is obtained. Targets for sequencing can be obtained either using PCR or genetic capture techniques before sequencing (Mamanova *et al.*, 2010). The targeted regions can represent individual independent loci (for example, exons) or a long stretch of genomic DNA. In order to apply either of these approaches, the sequence of the target regions needs to be known.

NGS of PCR products to target specific loci is generally referred to as amplicon sequencing (see Glossary) (Peng

and Zhang, 2009). This approach has been very useful when addressing population genetic and evolutionary questions for large functional groups of genes in model-organisms like *Drosophila* (Obbard *et al.*, 2009). By specifically tagging the PCR primers for each individual, this methodology can be used for high coverage genotyping of the loci of interest in a large number of samples (Binladen *et al.*, 2007). Another possibility is pooling of individuals before the sequencing step. A case study with high ecological relevance on a non-model organism is the recent survey of major histocompatibility complex class II (see Glossary) variation in the bank vole (*Myodes glareolus*). Using only a fraction of a 454 sequencing run, the investigators were able to genotype 96 individuals for this complicated multi locus gene (Babik *et al.*, 2009). As indicated by this study, great care needs to be taken to reduce problems with sequencing errors producing artificial alleles. But after dealing with these, it was shown that the NGS approach is able to detect alleles that are present at low frequency in the PCR product (and therefore could not be detected using conventional genotyping). Variation in these major histocompatibility complex loci was found to be related to prevalence of specific parasitic nematodes. These relationships were also population specific (Kloch *et al.*, 2010), thus providing a potential genetic basis for local adaptation. One of the problems of amplicon sequencing is that PCR products need to be carefully standardized to the same concentrations to avoid overrepresentation of a certain locus or a certain population during the sequencing.

Another popular application of amplicon NGS is so called 'barcoding' studies. Here, a small variable part of the genome (usually from mitochondria or chloroplasts) is amplified from unidentified or complex samples, and sequenced using NGS. The sequence information is then used to identify the species present in the sample (Valentini *et al.*, 2009). Studies using this approach have, for example, investigated the diet of a variety of animals by sequencing faecal samples (Deagle *et al.*, 2009), and characterized the extinct mammalian fauna using ancient DNA from frozen tundra sediment (Haile *et al.*, 2009). NGS barcoding approaches have also been extensively utilized to study meta-genomics of micro organism communities (Álvarez *et al.*, 2009; Buée *et al.*, 2009; Andersson *et al.*, 2010).

Finally, sequence capture methods such as NimbleGen arrays (Roche) (Hodges *et al.*, 2007; Okou *et al.*, 2007) and the SureSelect platform (Agilent Technologies, Santa Clara, CA, USA) (Gnirke *et al.*, 2009) are starting to become increasingly demanded by molecular ecologists. These capture techniques coupled with NGS generate high coverage sequence data from targeted DNA (for example, many independent fragments or a sequence of DNA of tens of kb) in several individuals or populations (that is, pool of individuals). These methodologies will allow molecular ecologists to study the sequence divergence between populations, morphs or species for hundreds of genes and gene families simultaneously, at a very reasonable cost and effort compared with previously available techniques. A drawback of the sequence capture approach is the need for a genome reference to mask repeated regions. Also, each individual or population will be hybridized independently, considerably increasing the cost of the methodology when

many samples are studied. A future improvement of these approaches will be the use of tagged samples, wherein multiple individuals/populations can be hybridized simultaneously. These approaches are predicted to change the way we think about phylogeography, demography and conservation genetics, by massively increasing the number of loci studied. This improvement will require the development of new software tools to make the analyses feasible from a computational point of view.

Large-scale identification and development of molecular markers

One of the most important application of NGS within ecological and population genetics is the development of molecular markers on a large scale. NGS generally allows for cost-effective and rapid identification of hundreds of microsatellite loci and thousands of SNPs, even if only a fraction of a sequencing run is used. This will, for example, facilitate QTL mapping studies and will increase the quality of outlier- and structure analyses. Massively increasing the number of markers will enable researchers not only to get better precision in population genetic studies (Novembre *et al.*, 2008), QTL and linkage disequilibrium (LD) mapping projects (Slate *et al.*, 2009) and kinship assignments (Santure *et al.*, 2010), but also to pursue topics such as historical demographic patterns, introgression and admixture (Jakobsson *et al.*, 2008).

As demonstrated by several recent studies (see Table 2) molecular markers can be developed on a large scale, in almost any organism from either transcriptome or genome sequences. One advantage of developing markers from transcriptome sequences is that those will be associated with functional genes, thus increasing the interest of these for studies of adaptation (Imelfort *et al.*, 2009). For example, expressed sequence tag (EST)-linked microsatellites typically occur in the untranslated regions (5'- and 3'-UTRs; Primmer, 2009). These loci are predicted to have a higher probability (compared with neutral genetic markers) of showing signatures of selection (Vasemagi *et al.*, 2005), and they can also have a functional significance in regulating gene expression and function. Microsatellite loci can be detected within NGS data sets with programs like msatcommander (Faircloth, 2008) and MSatFinder (Thurston and Field, 2005). SNPs (located in the coding region or in the UTRs) can also be discovered using many different approaches (for an updated list of specific software available for various applications see: http://seqanswers.com/wiki/Special:BrowseData/Bioinformatics_application). Both data from individually sequenced samples or samples of pooled individuals are useful to this end (Futschik and Schlötterer, 2010).

One obvious problem with SNP detection from NGS data is that sequencing errors will show a very similar signature to low-frequency SNP alleles. In most SNP discovery algorithms, this drawback is dealt with by including a minimum depth of sequencing at the place of interest to call a SNP as well as a minimum number of reads with the minor allele that is >1 (often 2 or 3). New theoretical work is also aiming to statistically distinguish true polymorphisms from sequencing errors

(see, for example, Lynch, 2009). GigaBayes (Hillier *et al.*, 2008) and VarScan (Koboldt *et al.*, 2009) represent examples of programs for SNP detection. GigaBayes calculates the probability that a polymorphism represent a true SNP or a sequencing error, for this calculation the program uses a Bayesian approach, taking into account the alignment depth, the base call in each sequence, the base composition in the region and the expected *a priori* polymorphism rate. With the advent of NGS, the use of SNPs in molecular ecology studies is predicted to expand dramatically. Once identified these markers can be typed in a large number of individuals using a wide variety of platforms (reviewed in Slate *et al.*, 2009). InDel (insertion–deletion) markers can be detected with similar software as the SNPs and they will probably become a popular kind of molecular marker in future population genetic studies, as they are easy and cheap to identify and score (Väli *et al.*, 2008).

Genomic DNA sequencing can also be used to generate genetic markers on a large scale. Sometimes low-coverage 454 genome sequencing is enough to identify several thousand microsatellite loci, like in the copperhead snake (*Agristodon contortrix*; Castoe *et al.*, 2009). Microsatellites have also been identified from museum samples of extinct moa species (*Dinornithiformes*) using a similar approach (Allentoft *et al.*, 2009). However, in order to increase the coverage of sequencing (and the efficiency of polymorphism detection), it may be very helpful to employ a genome complexity reduction technique (Santana *et al.*, 2009). Reduced representation libraries are obtained through restriction digestion of the DNA followed by an electrophoresis of the digestion and a size selection step (Van Tassell *et al.*, 2008). Another complexity reduction technique is the S-RAD markers (sequenced restriction-site-associated DNA; Baird *et al.*, 2008) (see Glossary), which also represent a very promising tool for SNP discovery from DNA samples (Hohenlohe *et al.*, 2010). S-RAD libraries are prepared by digesting genomic DNA with a restriction enzyme. Individually tagged adaptors are then ligated to the fragments and the samples are pooled. After this, physical shearing of the ligated DNA and a size selection are performed. The resulting library is sequenced through Illumina/Solexa paired-end sequencing and the tags are used to identify each sample after the sequencing. Many molecular ecology research groups are now investing their efforts in obtaining S-RAD markers in their study species. Library preparation is relatively easy and the reagents (for example, adaptors) can be used, *a posteriori*, for other organisms by the same research group. The data analysis involved might be the most difficult step, but once it is up and running in a research group it can be applied to almost any species.

Nucleotide variation profiling

Identifying the genes involved in ecologically important phenotypic variation is a major goal in ecological genomics (Feder and Mitchell-Olds, 2003). This has previously been accomplished by screening a large number of molecular markers (such as, SNPs or amplified fragment length polymorphisms; see Glossary) for outliers. As outlined above, such markers can now be developed on a large scale using NGS. Importantly, the

deep coverage provided in pools of samples (genomic DNA or cDNA) by NGS can also be used for directly screening genomic variation, bypassing the SNP genotyping step. This kind of approach may thus provide a shortcut in studies such as those trying to integrate population genomic and quantitative genetic approaches (Stinchcombe and Hoekstra, 2007). Such methodology has proven very efficient for outlier type analyses (see below) but may also be applicable in other kinds of population genetic studies. Tagging of samples, which is necessary for investigating genetic variation on an individual level, will be facilitated by the use of a recently launched highly automated procedure (Lennon *et al.*, 2010). The trade off between sequence depth of individual samples and number of samples sequenced also needs to be considered, as the calculations of allele frequencies and detection of low-frequency SNPs will be severely hampered if there is insufficient coverage of individual SNPs. It should also be noted that for SNP detection and estimation of population biology parameters, sequencing pools of individuals may be more effective than individual tagging of sequences (Futschik and Schlötterer, 2010).

A recent study used transcriptome 454 sequencing estimated SNP allele frequencies in two sympatric whitefish species (*Coregonus* spp.) to identify candidate markers for follow-up studies trying to determine the genetic basis of speciation and adaptation (Renaut *et al.*, 2010). Until recently, such studies were performed through QTL mapping, genome scans and microarrays for gene expression. Renaut *et al.* were able to corroborate the success of the methodology because results from these three analyses, which were previously performed in the same populations, were concordant with the new NGS study. Another study using a similar method to study the genetic basis for ecological speciation and adaptation was performed on two different host races of the apple maggot fly, *Rhagoletis pomonella* (Schwarz *et al.*, 2009). After performing SNP detection in their contigs, (see Glossary) they determined allele frequencies for each host race and those SNPs presenting significant differences were claimed as candidates for being involved in speciation. A slightly different approach was recently used to analyse the transcriptomes of two divergent ecotypes of the marine snail *Littorina saxatilis* undergoing ecological speciation (Galindo *et al.*, 2010). Here, allele frequencies for both ecotypes were calculated and neutral simulations were used to detect outlier SNPs. Some of these SNPs were found in genes related with shell formation and energetic metabolism, both functions that are very important in the adaptation of these ecotypes. At this stage, the main drawbacks of this type of analysis is the variance in coverage between SNPs, especially the low coverage of many of the SNPs. Additional problems are pooling of the RNA samples because of overrepresentation of some samples over the others (for example, not accurate normalization of the concentrations) or of some transcripts (for example, high expression levels of one transcript in certain samples).

Another recent study tried to determine the genetic basis of local adaptation in *Arabidopsis lyrata* using whole genome resequencing on the Illumina/Solexa platform. Divergent nucleotide polymorphisms between soil types were detected from allele frequency differences, and a

sliding window approach was used to identify outlier F_{ST} values between populations. Genes responsible for local adaptation to serpentine soils were detected after functional annotation and loci involved in heavy metal detoxification as well as calcium and magnesium transport pathways were overrepresented in markers with high divergence between soil types (Turner *et al.*, 2010). This kind of analysis will represent the outlier analysis of the future. As mentioned in the previous section, S-RAD markers can also be used to determine the SNP allele frequencies without additional genotyping because samples can be tagged individually. This strategy is thus also suitable for outlier analysis and genetic mapping (Baird *et al.*, 2008). A recent study using S-RAD markers in sticklebacks (Hohenlohe *et al.*, 2010) has identified regions across the genome showing signatures of selection between oceanic and freshwater populations.

When generating genomic data from a number of different species, NGS may be used to study molecular evolution (such as d_N/d_S ratios; see Glossary) on a large scale, an approach that has hereto been restricted to model organisms. Such a comparative genomics analysis has recently been performed in a study of avian genome evolution using 454 sequence data from 10 different bird species (Künstner *et al.*, 2010). Another interesting example studied two sympatric species of crater lake cichlids (Elmer *et al.*, 2010). After 454 sequencing of the transcriptome, d_N/d_S ratios were inferred to detect genes with signatures of divergent selection. Follow-up studies on these genes might reveal new insights about ecological speciation and adaptive radiation.

Epigenetics

Epigenetics is generally defined as the study of trait variation that does not come from changes in the DNA sequence, but rather involve other kinds of genetic modifications (such as patterns of DNA methylation and histone posttranslational modifications). Epigenetic modifications are primarily important because of their role in regulation of gene expression (Simon *et al.*, 2009). Traditionally, these phenomena have mainly been studied because of their importance in cancer biology and in regulation of development. As many ecologically important traits are also likely to be influenced by epigenetic variation (Bossdorf *et al.*, 2008), we include a short discussion of this topic here as a likely direction for future research on non-model organisms. Epigenetic modifications have mainly been studied on a large scale using microarray-based approaches. The one approach that has been applied so far in non-model organisms is methylation-dependent amplified fragment length polymorphism (Salmon *et al.*, 2005). NGS has, however, opened up a large range of different high-throughput methods for epigenetic surveys and some of these are also likely to be applicable to ecologically relevant systems (Hurd and Nelson, 2009; Simon *et al.*, 2009).

One major type of epigenetic modification that has been an important research focus for many years is the methylation of specific cytidine residues in the DNA. Generally, methylation has been studied using bisulphite treatment. Un-methylated cytidine residues in the DNA are converted to uracil after treatment with bisulphite,

whereas methylated cytidine are protected from this conversion. Sequencing of these regions can then pinpoint the specific methylated nucleotides. In recent years, NGS has been utilized to characterize DNA methylation patterns genome-wide, using an application known as ultra-deep bisulphite DNA sequencing (BS-Seq) (Taylor *et al.*, 2007; Cokus *et al.*, 2008). Methylated fractions of genomes can also be sequenced using a combination of NGS and methyl-DNA immunoprecipitation (MeDIP). By using this approach, detailed methylation maps of the genome will become available (Pomraning *et al.*, 2009), and by comparing these for different samples we will be able to address the importance of methylation for a large range of ecologically and evolutionary important questions, like the genetic architecture behind differential gene expression because of natural selection.

Another important determinant of transcription levels of genes is the structure of DNA packing, together with histone proteins, into nucleosomes (chromatin packing). This chromatin structure can also be studied using a high-throughput sequencing approach (Johnson *et al.*, 2006). The histone proteins themselves, particularly the N-terminal tails, are also subject to a large number of posttranslational modifications, such as methylation and acetylation of specific amino acid residues (Kouzarides, 2007). These can be studied on a large scale using a ChIP-Seq approach with Illumina/Solexa sequencing (Barski *et al.*, 2007). ChIP-Seq technology can also be used to study a large range of other DNA-protein interactions (Hurd and Nelson, 2009) including the identification of binding sites for transcription factors (Bhinge *et al.*, 2007). To the best of our knowledge, neither of these interesting approaches has yet been applied to non-model organisms, but such studies are undoubtedly under way.

Some general considerations regarding planning and data analysis

NGS provides a very cost-effective way to generate large amounts of sequence data, but a single sequencing run is still a considerable expense for most small labs working with non-model organisms. It is thus crucial to carefully evaluate whether the methodology will be able to answer the relevant biological questions asked. It is also important to consider expenses, skills and infra-structure needed for sample preparation and data analysis. The sheer volume of sequences produced by these new technologies constitutes a genuine challenge for data storage and analysis. For many applications, such as *de novo* assembly and downstream analysis, computing power may also be an issue. Also, as there are currently major advances in algorithm and software development for NGS analysis, and it may be a good idea for molecular ecologists to liaise with a bioinformatics research group. An alternative is to outsource parts of the data analysis to the sequencing facility.

There are several important considerations during the analysis and data handling steps of NGS data. This is currently a very active field of research and software are being developed and refined to deal with the special problems faced in NGS analyses (Pepke *et al.*, 2009). The first crucial steps in the data analyses are trimming (see

Glossary) and assembly of the reads. Several software exist (both freely available open source software and commercial programs) to perform *de novo* assembly of sequence reads, some of which also perform the pre-assembly trimming (http://seqanswers.com/wiki/Special:BrowseData/Bioinformatics_application). If there is already a reference genome (or transcriptome) sequence available, the reads may be mapped directly onto this, without the need for previous assembly (Trapnell and Salzberg, 2009). Much care is needed to avoid or remove mis-assemblies as these will introduce bias to downstream analyses and applications. A completely novel strategy for handling of NGS data was recently introduced by Cannon *et al.* (2010). They analysed the occurrence of specific complex short read motifs produced by Illumina/Solexa sequencing directly, without any previous assembly of the reads. Using this approach, they were able to study comparative genomics of several non-model tree species. Arguably, this approach will greatly facilitate short read sequence analyses in non-model organisms.

It is important to note that NGS technologies will not always be the optimal way to generate the sequence data of interest to molecular ecologists. If a particular candidate gene or gene family is of interest, then specific methods targeting these may be more cost effective than NGS (but see also discussion above for ways to utilize NGS in targeting specific genes). There is no guarantee that these genes will be present in the NGS data set even if suitable tissues are sampled for RNA extraction. Also, even if traces of the gene of interest are found it may not be straight forward to utilize this to get the full-length sequence, as there are no clones to go back to and sequence separately, as in traditional sequencing approaches (Wheat, 2010). Also, for many population genetic surveys, more traditional methods like marker typing or amplified fragment length polymorphism genome scans may prove to be sufficient to answer the questions of interest (at a significantly smaller cost than a large NGS-based study). For small labs with limited funding it may prove efficient to use only a small fraction of an NGS run to generate a small amount of background data and use this for more targeted studies of genes or regions of interest.

Finally, an important consideration is which sequencing platform to use (see Table 1). This will depend on the questions addressed and on the available genomic resources in the study species or in related species. Choosing a single platform is always a trade off between read length and sequence output (number of bases). Short read platforms (Illumina/Solexa and ABI SOLiD) provide more data than Roche 454, but at lower length of individual reads (although these are now increasing the read length, thus enabling *de novo* assembly). Until now, the 454 has been the most extensive method used (Table 2) because of the advantage of the relatively long reads produced with this technology. However, 454 sequencing presents slightly higher rates of sequencing errors compared with competing technologies (see Table 1), especially in homopolymers (see Glossary) (Huse *et al.*, 2007). Simulation experiments have shown that when there is no reference genome available, a combination of different NGS technologies may be most cost effective for transcriptome characterization (Wall *et al.*, 2009).

Future prospects

Given the current rate of technological advances in this field it is difficult to speculate very far into the future. Nevertheless, we will try to outline some improvements that are likely to be waiting round the corner. One major advantage of NGS in ecological studies is the small amounts of genetic material needed for analysis, making these technologies suitable for analyses of endangered species wherein non-invasive sampling is needed and for studies of ancient DNA. Future developments are aiming at decreasing this amount even further, down to the possibility of conducting sequence-based expression analysis on the scale of individual cells (Simon *et al.*, 2009). Also, as currently available technologies will continue to improve, these will be able to produce more and longer sequence reads, as well as decrease the sequence error rates. Together with improvements in data analysis algorithms, this will give higher quality assemblies of both genome and transcriptome data.

Within the applications of NGS that we have reviewed here, transcriptome characterization and gene expression profiling are most widely used until now (see Table 2). These approaches represent the first steps in more complex studies in which the availability of hundreds of genetic markers allows for QTL mapping or genome scans. In the near future, we expect that the number of studies taking advantage of genomic resources generated in this initial 'boom' of NGS will dominate over, for example, RNA-Seq. We anticipate the prices for sequencing and tagging might drop considerably enabling more population genomic type analyses with multiple samples and more precise and unbiased estimates of for example demographic parameters. Longer sequence reads will also improve downstream analysis applications because large haplotype blocks including several linked polymorphisms will become available. With time and with an increase in NGS studies, the goal of the research projects will not only be to detect signatures of selection but also to focus on the genetic architecture, the regulation and the history of selection. These topics are also directly linked to conservation biology. Until recently, conservation genetics projects generally studied a small number of neutral molecular markers to show the variation and/or heterozygosity of a population. NGS enables the shift to 'conservation genomics' (Ouborg *et al.*, 2010) wherein hundreds of genes can be study simultaneously. Some of these may be involved in important phenotypic variation and this is relevant from the conservation point of view, because such variation may be important to maintain within the population.

The most important emerging NGS technique might be single-molecule sequencing (Gupta, 2008). Applying this so called 'Next-next generation sequencing' (or 'Third generation sequencing') will eliminate the need of amplification during the sequencing reaction. This will not only be more cost effective and remove sequencing errors produced in the amplification step, but will also reduce bias in detecting expression levels of individual genes or alleles. Several platforms are currently being developed for this kind of analysis. At the time of writing only the Helicos tSMS system (Harris *et al.*, 2008) is commercially available but others (like the SMRT technology of Pacific Biosciences) are due to be launched

later this year. Taking things even one step further is the recently introduced method of direct sequencing of RNA through a modification of the Helicos tSMS protocol (Ozsolak *et al.*, 2009). This eliminates errors and biases in cDNA synthesis and thus provides a very accurate representation of gene expression levels.

As the cost of sequencing drops even further and the amount of data produced increases, there will be new demands for novel analysis methods and infrastructure. The future bottlenecks are more likely to be at the bioinformatics end rather than in producing the sequences (Schuster, 2008). Furthermore, there will likely be a large demand for molecular ecologists trying to make biological sense of all the gathered genomics data. It is probable that we will need radically new approaches for data storage and sharing as currently available databases might be unable to cope with the rapid generation of new sequencing data. We predict that keeping researchers from drowning in this data flood will be one of the major challenges in years to come. By bringing the realm of genomics into reach for studies of non-model organisms, NGS is currently radically changing our way of conducting genetic research, and it will continue to do so in the foreseeable future. As has been outlined in this review, this revolution is enhancing the scale that population genetic research can be conducted and bringing new objectives into reach of molecular ecologists.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We wish to thank Roger Butlin, Jon Slate, Gavin Horsburgh, Eleanor Jones and three anonymous reviewers for providing very valuable comments on earlier versions of this paper. The authors were funded by the European commission through a Marie Curie transfer of knowledge grant (MAERO).

References

- Allentoft ME, Schuster SC, Holdaway RN, Hale ML, McLay E, Oskam C *et al.* (2009). Identification of microsatellites from an extinct moa species using high-throughput (454) sequence data. *Biotechniques* **46**: 195–200.
- Alvarez LA, Exton DA, Timmis KN, Suggett DJ, McGenity TJ (2009). Characterization of marine isoprene-degrading communities. *Environ Microbiol* **11**: 3280–3291.
- Andersson AF, Riemann L, Bertilsson S (2010). Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *ISME J* **4**: 171–181.
- Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009). New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Mol Ecol Resour* **9**: 713–719.
- Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA *et al.* (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**: e3376.
- Bar-Or C, Czosnek H, Kolai H (2007). Cross-species microarray hybridizations: a developing tool for studying species diversity. *Trends Genet* **23**: 200–207.
- Barakat A, DiLoreto DS, Zhang Y, Smith C, Baier K, Powell WA *et al.* (2009). Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea*

- mollissima*) in response to the chestnut blight infection. *BMC Plant Biol* **9**: 51.
- Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z et al. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823–837.
- Baxter SW, Nadeau NJ, Maroja LS, Wilkinson P, Counterman BA, Dawson A et al. (2010). Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in the *Heliconius melpomene* clade. *PLoS Genet* **6**: e1000794.
- Bhinge AA, Kim J, Euskirchen GM, Snyder M, Iyer VR (2007). Mapping the chromosomal targets of STAT1 by sequence tag analysis of genomic enrichment (STAGE). *Genome Res* **17**: 910–916.
- Binladen J, Gilbert MTP, Bollback JP, Panitz F, Bendixen C, Nielsen R et al. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* **2**: e197.
- Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G et al. (2009). *De novo* transcriptome assembly with ABYSS. *Bioinformatics* **25**: 2872–2877.
- Bogdanova EA, Shagina IA, Mudrik E, Ivanov I, Amon P, Vagner LL et al. (2009). DSN depletion is a simple method to remove selected transcripts from cDNA populations. *Mol Biotechnol* **41**: 247–253.
- Bossdorf O, Richards CL, Pigliucci M (2008). Epigenetics for ecologists. *Ecol Lett* **11**: 106–115.
- Bräutigam A, Shrestha RP, Whitten D, Wilkerson CG, Carr KM, Froehlich JE et al. (2008). Low-coverage massively parallel pyrosequencing of cDNAs enables proteomics in non-model species: comparison of a species-specific database generated by pyrosequencing with databases from related species for proteome analysis of pea chloroplast envelopes. *J Biotechnol* **136**: 44–53.
- Buée M, Reich M, Murat C, Morin E, Nilsson RH, Uroz S et al. (2009). 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist* **184**: 449–456.
- Buggs RJA, Chamala S, Wu W, Gao L, May GD, Schnable PS et al. (2010). Characterization of duplicate gene evolution in the recent natural allopolyploid *Tragopogon miscellus* by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Mol Ecol* **19** (s1): 132–146.
- Bundock PC, Elliott FG, Ablett G, Benson AD, Casu RE, Aitken KS et al. (2009). Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotechnol J* **7**: 347–354.
- Cannon CH, Kua C-S, Zhang D, Harting JR (2010). Assembly free comparative genomics of short-read sequence data discovers the needles in the haystack. *Mol Ecol* **19** (s1): 147–161.
- Castoe TA, Poole AW, U WG, Koning APJD, Daza JM, Smith EN et al. (2009). Rapid identification of thousands of copperhead snake (*Agiistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. *Mol Ecol Resour* **10**: 341–347.
- Cheung F, Haas B, Goldberg S, May G, Xiao Y, Town C (2006). Sequencing *Medicago truncatula* expressed sequenced tags using 454 life sciences technology. *BMC Genomics* **7**: 272.
- Cheung F, Win J, Lang J, Hamilton J, Vuong H, Leach J et al. (2008). Analysis of the *Pythium ultimum* transcriptome using Sanger and pyrosequencing approaches. *BMC Genomics* **9**: 542.
- Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD et al. (2008). Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**: 215–219.
- Collins LJ, Biggs PJ, Voelckel C, Joly S (2008). An approach to transcriptome analysis of non-model organisms using short-read sequences. *Genome Informat* **21**: 3–14.
- Craft JA, Gilbert JA, Temperton B, Dempsey KE, Ashelford K, Tiwari B et al. (2010). Pyrosequencing of *Mytilus galloprovincialis* cDNAs: tissue-specific expression patterns. *PLoS ONE* **5**: e8875.
- Dassanayake M, Haas JS, Bohnert HJ, Cheeseman JM (2009). Shedding light on an extremophile lifestyle through transcriptomics. *New Phytologist* **183**: 764–775.
- Deagle BE, Kirkwood R, Jarman SN (2009). Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Mol Ecol* **18**: 2022–2038.
- DiGuistini S, Liao N, Platt D, Robertson G, Seidel M, Chan S et al. (2009). *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol* **10**: R94.
- Eads BD, Andrews J, Colbourne JK (2007). Ecological genomics in *Daphnia*: stress responses and environmental sex determination. *Heredity* **100**: 184–190.
- Ekblom R, Balakrishnan CN, Burke T, Slate J (2010). Digital gene expression analysis of the zebra finch genome. *BMC Genomics* **11**: 219.
- Ekblom R, Sæther SA, Jacobsson P, Fiske P, Sahlman T, Grahn M et al. (2007). Spatial pattern of MHC class II variation in the great snipe (*Gallinago media*). *Mol Ecol* **16**: 1439–1451.
- Ellegren H (2008). Sequencing goes 454 and takes large-scale genomics into the wild. *Mol Ecol* **17**: 1629–1631.
- Elmer KR, Fan S, Gunter HM, Jones JC, Boekhoff S, Kuraku S et al. (2010). Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. *Mol Ecol* **19** (s1): 197–211.
- Faircloth BC (2008). msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. *Mol Ecol Resour* **8**: 92–94.
- Feder ME, Mitchell-Olds T (2003). Evolutionary and ecological functional genomics. *Nat Rev Genet* **4**: 649–655.
- Ferguson L, Lee SF, Chamberlain N, Nadeau N, Joron M, Baxter S et al. (2010). Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the *HmYb/Sb* locus. *Mol Ecol* **19** (s1): 240–254.
- Futschik A, Schlötterer C (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* **186**: 207–218.
- Galindo J, Grahame JW, Butlin RK (2010). An EST-based genome scan using 454 sequencing in the marine snail *Littorina saxatilis*. *J Evol Biol* **23**: 2004–2016.
- Garcia-Reyero N, Griffitt RJ, Liu L, Kroll KJ, Farmerie WG, Barber DS et al. (2008). Construction of a robust microarray from a non-model species largemouth bass, *Micropterus salmoides* (Lacepede), using pyrosequencing technology. *J Fish Biol* **72**: 2354–2376.
- Gilad Y, Pritchard JK, Thornton K (2009). Characterizing natural variation using next-generation sequencing technologies. *Trends Genet* **25**: 463–471.
- Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**: 182.
- Goetz F, Rosauer D, Sitar S, Goetz G, Simchick C, Roberts S et al. (2010). A genetic basis for the phenotypic differentiation between siscowet and lean lake trout (*Salvelinus namaycush*). *Mol Ecol* **19** (s1): 176–196.
- Gupta PK (2008). Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol* **26**: 602–611.
- Hahn DA, Ragland GJ, Shoemaker DD, Denlinger DL (2009). Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. *BMC Genomics* **10**: 234.
- Haile J, Froese DG, MacPhee RDE, Roberts RG, Arnold LJ, Reyes AV et al. (2009). Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proc Natl Acad Sci USA* **106**: 22352–22357.
- Hale MC, McCormick CR, Jackson JR, DeWoody A (2009). Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative

- merits of normalization and rarefaction in gene discovery. *BMC Genomics* **10**: 203.
- Harr B, Turner LM (2010). Genome-wide analysis of alternative splicing evolution among *Mus* subspecies. *Mol Ecol* **19** (s1): 228–239.
- Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I et al. (2008). Single-molecule DNA sequencing of a viral genome. *Science* **320**: 106–109.
- Hedrick PW (2004). Recent developments in conservation genetics. *Forest Ecol Manag* **197**: 3–19.
- Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D et al. (2008). Whole-genome sequencing and variant discovery in *C elegans*. *Nat Meth* **5**: 183–188.
- Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW et al. (2007). Genome-wide *in situ* exon capture for selective resequencing. *Nat Genet* **39**: 1522–1527.
- Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* **6**: e1000862.
- Hudson ME (2008). Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol Ecol Notes* **8**: 3–17.
- Hurd PJ, Nelson CJ (2009). Advantages of next-generation sequencing versus the microarray in epigenetic research. *Brief Funct Genomic Proteomic* **8**: 174–183.
- Huse S, Huber J, Morrison H, Sogin M, Welch D (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8**: R143.
- Höglund J (2009). *Evolutionary Conservation Genetics*. Oxford University Press: Oxford.
- Imelfort M, Duran C, Batley J, Edwards D (2009). Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnol J* **7**: 312–317.
- Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**: 998.
- Johnson SM, Tan FJ, McCullough HL, Riordan DP, Fire AZ (2006). Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res* **16**: 1505–1516.
- Kammenga JE, Herman MA, Ouborg NJ, Johnson L, Breitling R (2007). Microarray challenges in ecology. *Trends Ecol Evol* **22**: 273–279.
- Kloch A, Babik W, Bajer A, Sinski E, Radwan J (2010). Effects of an MHC-DRB genotype and allele number on the load of gut parasites in the bank vole *Myodes glareolus*. *Mol Ecol* **19** (s1): 255–265.
- Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER et al. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**: 2283–2285.
- Kohn MH, Murphy WJ, Ostrander EA, Wayne RK (2006). Genomics and conservation genetics. *Trends Ecol Evol* **21**: 629–637.
- Kouzarides T (2007). Chromatin modifications and their function. *Cell* **128**: 693–705.
- Kristiansson E, Asker N, Förlin L, Larsson DJ (2009). Characterization of the *Zoarces viviparus* liver transcriptome using massively parallel pyrosequencing. *BMC Genomics* **10**: 345.
- Künstner A, Wolf JBW, Niclas B, Whitney O, Balakrishnan CN, Day L et al. (2010). Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. *Mol Ecol* **19** (s1): 266–276.
- Lee RM, Thimmapuram J, Thinglum KA, Gong G, Hernandez AG, Wright CL et al. (2009). Sampling the waterhemp (*Amaranthus tuberculatus*) genome using pyrosequencing technology. *Weed Sci* **57**: 463–469.
- Lennon N, Lintner R, Anderson S, Alvarez P, Barry A, Brockman W et al. (2010). A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. *Genome Biol* **11**: R15.
- Li R, Fan W, Tian G, Zhu H, He L, Cai J et al. (2010). The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**: 311–317.
- Lynch M (2009). Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* **182**: 295–301.
- Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A et al. (2010). Target-enrichment strategies for next-generation sequencing. *Nat Meth* **7**: 111–118.
- Metcalf CJ, Mitchell-Olds T (2009). Life history in a model system: opening the black box with *Arabidopsis thaliana*. *Ecol Lett* **12**: 593–600.
- Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK et al. (2009). Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFLx. *BMC Genomics* **10**: 219.
- Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM et al. (2008). Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456**: 387–390.
- Mitchell-Olds T, Feder M, Wray G (2008). Evolutionary and ecological functional genomics. *Heredity* **100**: 101–102.
- Molina C, Rotter B, Horres R, Udupa S, Besser B, Bellarmino L et al. (2008). SuperSAGE: the drought stress-responsive transcriptome of chickpea roots. *BMC Genomics* **9**: 553.
- Morozova O, Marra MA (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**: 255–264.
- Murray D, Doran P, MacMathuna P, Moss A (2007). In silico gene expression analysis—an overview. *Mol Cancer* **6**: 50.
- Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–1349.
- Nielsen KL, Hogg AL, Emmersen J (2006). DeepSAGE—digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucl Acids Res* **34**: e133.
- Nolte V, Schlötterer C (2008). African *Drosophila melanogaster* and *D simulans* populations have similar levels of sequence variability, suggesting comparable effective population sizes. *Genetics* **178**: 405–412.
- Novaes E, Drost DR, Farmerie WG, Pappas Jr GJP, Grattapaglia D, Sederoff RR et al. (2009). High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**: 312.
- Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A et al. (2008). Genes mirror geography within Europe. *Nature* **456**: 98.
- Nuzhdin SV, Wayne ML, Harmon KL, McIntyre LM (2004). Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol Biol Evol* **21**: 1308–1317.
- Obbard DJ, Welch JJ, Kim K-W, Jiggins FM (2009). Quantifying adaptive evolution in the *Drosophila* immune system. *PLoS Genet* **5**: e1000698.
- Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME (2007). Microarray-based genomic selection for high-throughput resequencing. *Nat Meth* **4**: 907–909.
- Ouborg NJ, Pertoldi C, Loeschke V, Bilsma R, Hedrick PW (2010). Conservation genetics in transition to conservation genomics. *Trends Genet* **26**: 177–187.
- Ozsak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P et al. (2009). Direct RNA sequencing. *Nature* **461**: 814–818.
- Pauchet Y, Wilkinson P, van Munster M, Augustin S, Pauron D, ffrench-Constant RH (2009). Pyrosequencing of the midgut transcriptome of the poplar leaf beetle *Chrysomela tremulae* reveals new gene families in Coleoptera. *Insect Biochem Mol Biol* **39**: 403–413.
- Pauchet Y, Wilkinson P, Vogel H, Nelson DR, Reynolds SE, Heckel DG et al. (2010). Pyrosequencing the *Manduca sexta*

- larval midgut transcriptome: messages for digestion, detoxification and defence. *Insect Mol Biol* **19**: 61–75.
- Peng H, Zhang J (2009). Commercial high-throughput sequencing and its applications in DNA analysis. *Biologia* **64**: 20–26.
- Pepke S, Wold B, Mortazavi A (2009). Computation for ChIP-seq and RNA-seq studies. *Nat Meth* **6**: S22–S32.
- Pomraning KR, Smith KM, Freitag M (2009). Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods* **47**: 142–150.
- Primmer CR (2009). From conservation genetics to conservation genomics. *Ann NY Acad Sci* **1162**, 357–368.
- Quinn N, Levenkova N, Chow W, Bouffard P, Boroevich K, Knight J et al. (2008). Assessing the feasibility of GS FLX pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics* **9**: 404.
- Rasmussen DA, Noor MAF (2009). What can you do with 0.1 × genome coverage? A case study based on a genome survey of the scuttle fly *Megaselia scalaris* (Phoridae). *BMC Genomics* **10**: 382.
- Renaut S, Nolte AW, Bernatchez L (2010). Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (*Coregonus* spp. *Salmonidae*). *Mol Ecol* **19** (s1): 115–131.
- Roeding F, Borner J, Kube M, Klages S, Reinhardt R, Burmester T (2009). A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). *Mol Phylogenet Evol* **53**: 826–834.
- Romanov MN, Tuttle EM, Houck ML, Modi WS, Chemnick LG, Korody ML et al. (2009). The value of avian genomics to the conservation of wildlife. *BMC Genomics* **10** (Suppl 2): S10.
- Rounsley S, Marri PR, Yu Y, He R, Sisneros N, Goicoechea JL et al. (2009). *De novo* next generation sequencing of plant genomes. *Rice* **2**: 35–43.
- Saccheri I, Kuussaari M, Kankare M, Vikman P, Fortelius W, Hanski I (1998). Inbreeding and extinction in a butterfly metapopulation. *Nature* **392**: 491–493.
- Salmon A, Ainouche ML, Wendel JF (2005). Genetic and epigenetic consequences of recent hybridization and polyploidy in Spartina (Poaceae). *Mol Ecol* **14**: 1163–1175.
- Santana QC, Coetzee MPA, Steenkamp ET, Mlonyeni OX, Hammond GNA, Wingfield MJ et al. (2009). Microsatellite discovery by deep sequencing of enriched genomic libraries. *Biotechniques* **46**: 217–223.
- Santure AW, Stapley J, Ball AD, Birkhead TR, Burke T, Slate J (2010). On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigree zebra finch population typed at 771 SNPs. *Mol Ecol* **19**: 1439–1451.
- Schuster SC (2008). Next-generation sequencing transforms today's biology. *Nat Meth* **5**: 16–18.
- Schwarz D, Robertson H, Feder J, Varala K, Hudson M, Ragland G et al. (2009). Sympatric ecological speciation meets pyrosequencing: sampling the transcriptome of the apple maggot *Rhagoletis pomonella*. *BMC Genomics* **10**: 633.
- Shendure J, Ji H (2008). Next-generation DNA sequencing. *Nat Biotech* **26**: 1135–1145.
- Simon SA, Zhai J, Nandety RS, McCormick KP, Zeng J, Mejia D et al. (2009). Short-read sequencing technologies for transcriptional analyses. *Ann Rev Plant Biol* **60**: 305–333.
- Slate J, Gratten J, Beraldí D, Stapley J, Hale M, Pemberton J (2009). Gene mapping in the wild with SNPs: guidelines and future directions. *Genetica* **136**: 97–107.
- Stinchcombe JR, Hoekstra HE (2007). Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* **100**: 158–170.
- t Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, de Menezes RX et al. (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucl Acids Res* **36**: 1–11.
- Taylor KH, Kramer RS, Davis JW, Guo J, Duff DJ, Xu D et al. (2007). Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res* **67**: 8511–8518.
- Teeter KC, Payseur BA, Harris LW, Bakewell MA, Thibodeau LM, O'Brien JE et al. (2008). Genome-wide patterns of gene flow across a house mouse hybrid zone. *Genome Res* **18**: 67–76.
- Thurston MI, Field D (2005). Msatfinder: detection and characterisation of microsatellites. Distributed by the authors at <http://www.genomics.ceh.ac.uk/msatfinder/>. CEH Oxford: Oxford.
- Toth AL, Varala K, Newman TC, Miguez FE, Hutchison SK, Willoughby DA et al. (2007). Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science* **318**: 441–444.
- Trapnell C, Salzberg SL (2009). How to map billions of short reads onto genomes. *Nat Biotech* **27**: 455–457.
- Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV (2010). Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nat Genet* **42**: 260–263.
- Valentini A, Pompanon F, Taberlet P (2009). DNA barcoding for ecologists. *Trends Ecol Evol* **24**: 110–117.
- Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, Liang H et al. (2009). Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* **10**: 347.
- van Bers NEM, Oers KV, Kerstens HHD, Dibbits BW, Croijmans RPMA, Visser ME et al. (2010). Genome-wide SNP detection in the great tit *Parus major* using high throughput sequencing. *Mol Ecol* **19** (s1): 89–99.
- Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT et al. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Meth* **5**: 247–252.
- Wang W, Wang Y, Zhang Q, Qi Y, Guo D (2009a). Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics* **10**: 465.
- Wang Z, Gerstein M, Snyder M (2009b). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57–63.
- Vasemagi A, Nilsson J, Primmer CR (2005). Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar* L.). *Mol Biol Evol* **22**: 1067–1076.
- Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I et al. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol* **17**: 1636–1647.
- Wheat C (2010). Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica* **138**: 433–451.
- Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872–876.
- Wilhelm BT, Landry J-R (2009). RNA-Seq—quantitative measurement of expression through massively parallel RNA sequencing. *Methods* **48**: 249–257.
- Wolf JBW, Bayer T, Haubold B, Schilhabel M, Rosenstiel P, Tautz D (2010). Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Mol Ecol* **19** (s1): 162–175.
- Wu CA, Lowry DB, Cooley AM, Wright JK, Lee YW, Willis JH (2007). Mimulus is an emerging model system for the

- integration of ecological and genomic studies. *Heredity* **100**: 220–230.
- Väli Ü, Brandström M, Johansson M, Ellegren H (2008). Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genet* **9**: 8.
- Zhang L, Li W-H (2004). Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol Biol Evol* **21**: 236–239.
- Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB et al. (2004). Simple cDNA normalization using kamchatka crab duplex-specific nucleic acid. *Nucl Acids Res* **32**: e37.
- Zou Z, Najar F, Wang Y, Roe B, Jiang H (2008). Pyrosequence analysis of expressed sequence tags for *Manduca sexta* hemolymph proteins involved in immune responses. *Insect Biochem Mol Biol* **38**: 677–682.

Appendix

Glossary

AFLP	Amplified fragment length polymorphism; PCR-based method for detection and large-scale genotyping of polymorphisms in genomic DNA
Amplicon sequencing	Sequencing of fragments that have first been amplified specifically using a PCR approach
Contig	A set of overlapping NGS reads making up a longer stretch of sequence
cDNA	complementary DNA; DNA synthesized <i>in vitro</i> from an mRNA template in a reaction catalysed by the enzyme reverse transcriptase.
<i>De novo</i> assembly	Aligning and merging fragments of DNA sequences (reads) into contigs without using previous knowledge of the sequence
d_N/d_S ratio	Also known as K_a/K_s or ω ; the rate of non-synonymous nucleotide substitutions divided by the rate of synonymous nucleotide substitutions. Can be used as an indicator of the selection pressure acting on a protein-coding gene
GO	Gene ontology; a defined set of terms representing gene product properties on three different levels (molecular function, cellular component and biological process)
GRS	Genomic reference species; the evolutionarily closest or most relevant species, compared with the species of interest, for which genomic resources exists
Homopolymer	A stretch of DNA having several of the same nucleotide together (for example, CCCC or AAAAAAA).
MHC	Major histocompatibility complex; a genetic region containing several genes involved in vertebrate immune functions, some of which are extremely polymorphic in most species
Microarray	A method for measuring gene expression by hybridising cDNA to a chip with a large number of microscopic spots with DNA oligonucleotides
Outlier analysis	A method to detect candidate loci for selection and differentiation by scanning the genome for regions departing from neutral expectations (for example, elevated F_{ST} values)
QTL mapping	Mapping of quantitative trait loci; a statistical analysis designed to investigate the genetic basis of continuous phenotypic variation
Read	An individual short piece of sequence output from an NGS platform
RNA-Seq	Also known as whole transcriptome shotgun sequencing (WTSS); NGS of cDNA in order to get information about a samples' RNA content (both nucleotide sequences and expression levels)
RRL	Reduced representation library; a DNA library created from only a certain fraction of the genome
SAGE	Serial analysis of gene expression; a technique to investigate gene expression, by measuring the amount of mRNA in a sample of interest in the form of small tags that correspond to fragments of the different transcripts
S-RAD	Sequenced restriction-site-associated DNA; a method combining SNP discovery and genotyping on a large scale using NGS
Templated assembly	Also referred to as mapping; assembly of NGS reads by aligning these to an already known template, for example, a full genome sequence or gene prediction database
Transcriptome	The set of all RNA molecules produced in a population of cells
Trimming	A crucial initial step in the analysis of NGS data. Removal of poor quality sequence, poly-A tails, adaptors, tags and primer sequences