 doi: 10.1111/j.1471-8286.2007.02019.x

TECHNICAL REVIEW

Sequencing breakthroughs for genomic ecology and evolutionary biology

MATTHEW E. HUDSON
Department of Crop Sciences, University of Illinois, Urbana, 334 NSRC, 1101 W. Peabody Blvd., IL 61801, USA

Abstract

Techniques involving whole-genome sequencing and whole-population sequencing (metagenomics) are beginning to revolutionize the study of ecology and evolution. This revolution is furthest advanced in the Bacteria and Archaea, and more sequence data are required for genomic ecology to be fully applied to the majority of eukaryotes. Recently developed next-generation sequencing technologies provide practical, massively parallel sequencing at lower cost and without the requirement for large, automated facilities, making genome and transcriptome sequencing and resequencing possible for more projects and more species. These sequencing methods include the 454 implementation of pyrosequencing, Solexa/Illumina reversible terminator technologies, polony sequencing and AB SOLiD. All of these methods use nanotechnology to generate hundreds of thousands of small sequence reads at one time. These technologies have the potential to bring the genomics revolution to whole populations, and to organisms such as endangered species or species of ecological and evolutionary interest. A future is now foreseeable where ecologists may resequence entire genomes from wild populations and perform population genetic studies at a genome, rather than gene, level. The new technologies for high throughput sequencing, their limitations and their applicability to evolutionary and environmental studies, are discussed in this review.

Keywords: DNA sequencing, ecology, evolution, genomic, metagenomics

Received 17 June 2007; revision accepted 4 September 2007

Introduction

Model eukaryotes such as *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Arabidopsis thaliana* were among the first organisms to have completely sequenced genomes. As a result, extensive and well-developed molecular and bioinformatics tools are available for researchers on these organisms. The concentration of resources on these organisms has led to very successful molecular ecology and evolutionary research on model eukaryotes and their close relatives, such as *Drosophila pseudoobscura* (Richards *et al*. 2005) or *Arabidopsis lyrata* (Wright *et al*. 2003). Studies in many species of Bacteria and Archaea now routinely use whole-genome sequence information. However, only in a small number of model eukaryotes is it currently possible to understand processes of adaptation on a whole-genome scale. Comparative genomics between closely related species (such as models and their relatives) has great promise to elucidate basic mechanisms of evolution and selection at the genomic level. Our understanding of whole-genome evolution and ecology is likely to be further transformed by whole-genome sequencing of many individuals (re-sequencing) from model species and their close relatives. However, such data will have limited relevance to many of the most important species in global ecology and evolutionary biology, which are generally distantly related to model organisms. Genome sequencing in nonmodel organisms has the potential to also be greatly enhanced by developments in sequencing technologies. The growing interest in human genome resequencing, together with developments in nucleic acid chemistry, nanotechnology and microscopy, has led to a new generation of sequencing and genotyping technologies. The most radical recent development has been technologies that sequence DNA very fast and cheaply, in short segments. These new methods are currently driving down sequencing costs and increasing capacity at an

Correspondence: Matthew Hudson, Fax: 217 3338046; E-mail: mhudson@vivc.edu

unprecedented rate (Fig. 1). This review explores new, low-cost sequencing technologies, their limitations and their likely impact on the study of evolution and ecology both in model and nonmodel organisms.

Why sequence genomes for environmental and evolutionary biology?

Genomics and molecular biology have already revolutionized the study of ecology and evolution. For example, the ability to identify individuals by their DNA (Jeffreys *et al*. 1985), to perform taxonomic studies on DNA sequence information (Cann *et al*. 1987; Woese *et al*. 1990), to study gene flow within wild populations (e.g. Ellstrand *et al*. 1999) and to identify unculturable microorganisms from their DNA sequence (reviewed by Handelsman 2004) have all allowed great advances in their respective fields and in many cases had much broader impacts on society. However, all these techniques are based on the analysis of a tiny fragment of an organism's total gene complement, usually one gene or variable region. Analysis of selection and evolution at a particular locus such as the small subunit ribosomal RNA (ss rRNA) gene is a powerful tool, but is limited in its capacity to resolve the evolutionary history of some organisms. For example, rRNA genes can be horizontally transferred between species (Yap *et al*. 1999). Complications such as this can complicate the interpretation of single-gene phylogenies, but be resolved by genome-wide studies. In addition, analysis based on a single gene will always give a restricted picture of how an organism arrived at its current genotype, or how a population is constructed at a genetic level.

It is difficult to overstate the potential of genomics for discovery in evolutionary and environmental biology. Genomic sequence data are now accumulating very rapidly in public databases (Fig. 1). The availability of genomic sequence data for a given species allows a wide range of powerful techniques to be applied to understanding the genetics and adaptation of that species. These techniques include large-scale genotyping on a known physical template, rapid positional genetics, comparative genomics, microarray analysis, proteomics and the identification of orthologous genes and entire paralogous gene families. While *de novo* whole-genome sequencing is still relatively difficult and expensive, transcriptome sequencing can provide many of the tools of interest to evolutionary biologists. Using next-generation transcriptome sequencing tools, Toth *et al*. (2007) addressed a long-standing evolutionary behavioural question in a venerable model, *Polistes metricus*. Prior to the study, *Polistes* had essentially no genomic data resources, demonstrating how next-generation technologies have the potential to bring genomics into new areas of biology.

It is perhaps the comparison of entire genomes, especially of closely related species or individuals, which is currently

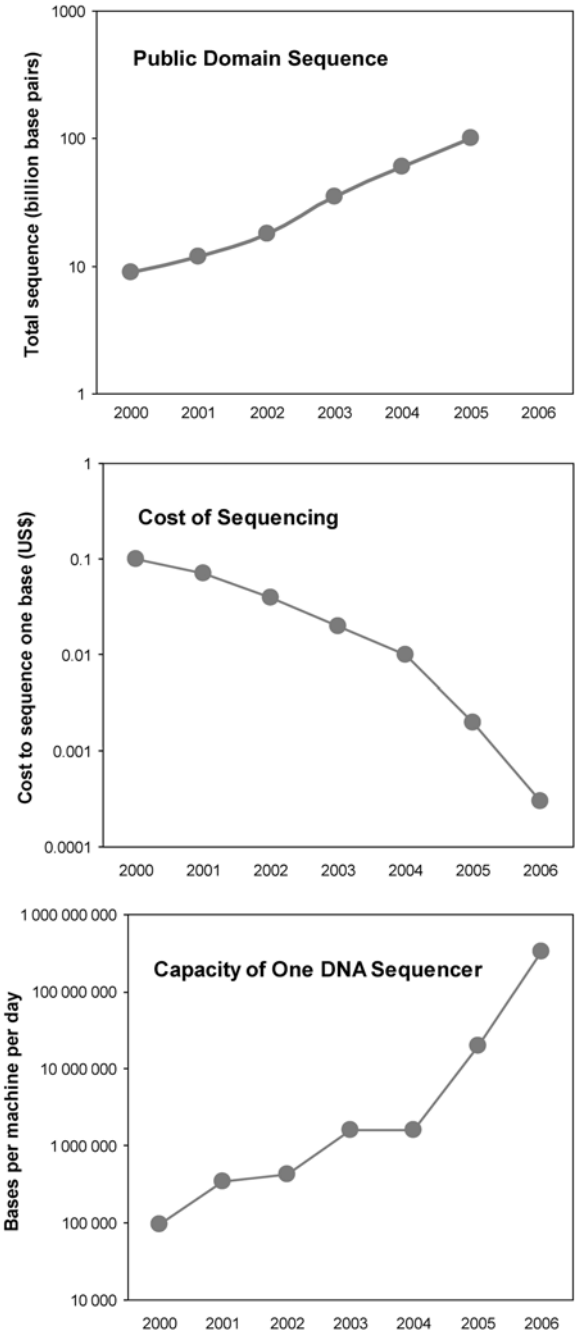

Fig. 1 Advances in DNA sequencing technology and their effect. Upper Graph: the increasing amount of DNA sequence data in the public domain (total deposited in International Nucleotide Sequence Databases) from 2000 to 2005. Centre Graph: the cost per base of DNA sequenced for genome sequencing projects from 2000 to 2006. Lower Graph: the capacity of the highest-throughput commercially available DNA sequencer from 2000 to 2006. Note that the 454/Roche sequencing technology was introduced in 2005 and the Solexa/Illumina 1G sequencing technology in 2006, and that the 2005 publicly available sequence was essentially all generated using Sanger technology. *Y* axis scales are logarithmic.

generating the most enthusiasm among scientists interested in evolution and populations. Comparative genomics has the potential to directly deliver causative genes and alleles for traits using association mapping, to allow population genetics on a whole-genome scale, and to uncover the genomic events that lead to the formation of new species. Whole genome comparison has already generated some important and often counterintuitive results. For example, comparison of human and chimpanzee genomes revealed that fewer of the current genomic differences between these organisms are single-base substitutions than are large, segmental duplications (Cheung *et al*. 2005). However, even though humans and chimps are relatively closely related, analysing evolutionary events as they happen on a whole-genome scale requires the resequencing of whole genomes from multiple individuals of the same species, and even the same population. Resequencing of closely related individuals adds significant complexity to the problems of both sequencing and data analysis, but also has tremendous potential (Hirschhorn & Daly 2005). The availability of multiple genome sequences makes many advances in traditional and quantitative genetics also possible, such as the ability to fine-map and characterize quantitative trait loci rapidly, and to perform association genetics across entire genomes without the need for extensive genotyping. In areas such as plant or animal breeding, the emerging ability of next-generation sequencing technology to resequence whole genomes of valuable genetic lines could ultimately make the entire genomic makeup of the breeder's stock of genetic diversity accessible from a computer. Using similar techniques, this technology has great potential to resolve the processes underlying natural selection in addition to artificial selection.

Sequencing DNA from whole populations, rather than individuals, commonly referred to as metagenomics (Handelsman 2004) represents a field which is rapidly expanding due to falling costs of DNA sequencing. Metagenomics began using highly conserved sequences such as the ss rRNA to identify the species present in a population (Handelsman 2004). With falling costs of Sanger sequencing, it became possible to sequence randomly sheared fragments of entire genomes from microbial populations (Venter *et al*. 2004; Tringe *et al*. 2005). With the availability of affordable, massively parallel sequencing using next-generation technology, metagenomic sequencing approaches can be greatly increased in their power to resolve rare species using next-generation sequencing technology (Sogin *et al*. 2006), and random sequencing approaches can be expanded and applied in greater depth (Angly *et al*. 2006; Turnbaugh *et al*. 2006). While such experiments are still expensive, the newer technologies discussed in this review are likely to become the methods of choice for sequencing applications such as metagenomics and bioprospecting (Schloss & Handelsman 2003; Huse *et al*. 2007).

In molecular ecology, genome resequencing also holds great promise to achieve a number of important goals. These goals are to determine the underlying genetic polymorphisms that confer phenotypic traits on a high-throughput basis, and to understand genetic structure and variation within a population. Reduced cost resequencing, while driven by commercial research for the human genomics market, is likely also to revolutionize the study of diversity in an ecological and evolutionary context. These changes are likely to bring great opportunities and challenges to genomics, genetics and computational biology in the near future, by greatly increasing the need for new methods of quantitative analysis for multiple, whole-genome data sets.

Applications of genomic technologies

De novo *whole-genome sequencing*

To apply the full power of genomics to any given species, it is first necessary to generate a complete genome sequence of one individual (*de novo* sequencing). The new DNA sequencing technologies discussed in this review are eventually likely to have a large impact on *de novo* sequencing, but have primarily been developed for, and are better suited to, the purpose of whole-genome resequencing (discussed later).

Conventional *de novo* whole-genome sequencing uses one of two approaches. One, the oldest and best established, is to build a 'tiling path' of large-insert bacterial vector clones across the genome, then sequence the clones, in a 'clone-by-clone' approach. This method still gives by far the best genome coverage and sequence quality for large, complex genomes (i.e. eukaryotic haploid genomes larger than 100 million bp). Applying new-generation methods to the sequencing of these individual clones may lead to significant lowering of the costs of this method. However, much time and labour is needed to build the tiling path of clones, and the clone-by-clone method is thus likely to remain relatively expensive.

The second established method is termed whole-genome shotgun sequencing (Fleishmann *et al*. 1995). In this method, rather than building a tiling path of clones, clones are sequenced randomly and the whole-genome sequence assembled using a powerful computer. This method is now the method of choice for sequencing smaller, less complex genomes such as those of Bacteria and Archaea. It has more recently been successfully applied to the sequencing of much larger genomes. Shotgun sequencing is generally significantly cheaper than the clone-by-clone approach, since the time and effort required to build the tiling path is avoided. While shotgun sequencing produces excellent results from bacterial genomes, large, complex genomes generally still contain many unsequenced gaps after a shotgun sequencing project. These gaps can be closed if necessary by a manual 'finishing' process similar to the clone-by-clone approach;

however, this process substantially reduces the cost advantage of shotgun sequencing. Drawbacks of whole-genome shotgun sequencing are discussed in detail by Green (1997); although many problems have since been ameliorated by better strategies and computational tools.

Shotgun sequencing of large genomes relies on relatively long sequence reads being available from both ends of a DNA clone, such as a plasmid, fosmid or Bacterial Artificial Chromosome. These paired end reads are often termed mate pairs or pairwise ends. Mate pairs constrain the number of places where any given sequence read can be placed, since the mate pair of that read must also fit into the assembly, at a suitable distance consistent with the size of the DNA insert in the vector. The availability of mate pairs is thus essential for the computational assembly of larger, more complex genomes (Siegel *et al*. 2000). While, in principle, the new DNA sequencing technologies can be applied to *de novo* shotgun genome sequencing, their application is currently limited to very small genomes (i.e. Bacterial, Archaeal and viral genomes smaller than 10 million bp) by a combination of relatively short sequence reads and restrictions on mate pairs.

Expressed sequence tags and transcriptome sequencing

Where a *de novo* whole-genome sequencing project is not yet feasible in a given organism for reasons of cost or technical difficulty, an alternative is an expressed sequence tag (EST) project, which focuses the project on protein-coding sequence by sequencing only mRNA (Adams *et al*. 1991). A traditional EST project produces sequences of 500–700 bp, anchored at one or both ends of each mRNA. While an EST project does not produce more than a fraction of the information available from a whole-genome sequence, it can produce a 'tag' of sequence from the protein coding region of most genes and, thus, allows the creation of many genomic tools, for example expression microarrays. EST projects are likely to become much cheaper in the very near future, thanks to the use of next-generation sequencing tools to produce the sequence of large numbers of cDNAs quickly and cheaply. Using next-generation technology, a form of EST sequencing sometimes referred to as transcriptome sequencing (where each mRNA is sequenced in its entirety in random fragments and assembled computationally) may be a more suitable strategy than traditional EST sequencing (where each sequence is anchored at the 3′ or 5′ end of the transcript). These methods can be used as a readily accessible 'entry point' for genomics for species with no existing genome resources (Toth *et al*. 2007).

Whole-genome resequencing

Once a fully finished whole-genome sequence is available, it is relatively straightforward to generate complete sequences of multiple genotypes of the same species (resequencing) in order to understand the degree of genetic variation within populations, and the processes of selection on a genomic scale. The sequencing of the human genome has led to extreme interest in the variation between the genome sequences of individual humans, and how these variations might lead to diagnosis or cure for inherited tendencies to disease (Hirschhorn & Daly 2005). Much of this intraspecies variation is in the form of single nucleotide polymorphisms (SNPs), which are straightforward to detect and to assay (see below). Insertion and/or deletion of sequence (indels), genomic rearrangements (such as segmental duplications, inversions and translocations), copy number polymorphisms caused by local duplication and other structural variations are also common types of variation between genomes, but less attention has been focused on these, primarily because they can be harder to detect. In order to analyse the complete set of sequence variants in a particular individual, it is necessary to sequence the entire human genome again for that individual, known as genome resequencing (Shendure *et al*. 2004). The cost of resequencing the human genome using conventional, Sanger technology was around US$50 million as of 2003 [National Human Genome Research Institute (NHGRI) 2003]. However, costs of genomic sequencing have been falling dramatically over the last 10 years (Fig. 1, Shendure *et al*. 2004) and the current cost of such a project using Sanger sequencing is probably less than US$25 million. Recently, it was announced that James Watson's genome had been resequenced for 'less than $1m' using next-generation (in this case 454, see below) technology (Check 2007). Knowledge from such resequencing experiments can permit directed resequencing at specific loci (i.e. known SNPs) at much lower cost than whole-genome resequencing.

Since true whole-genome resequencing is necessary in order to discover all the differences between genomes, known and unknown, the NHGRI has targets to reduce resequencing costs by two and four orders of magnitude by 2010 and 2020, respectively (NHGRI 2003). Whole-genome resequencing does not necessarily require any assembly step, since the sequence reads can in principle be aligned to a template of a previously assembled whole-genome sequence for that organism. The shotgun approach can therefore be used without any requirement for mate pairs, or for long reads. However, the sequence reads must be long enough to be unique in the genomic regions of interest, and it can be necessary to perform an assembly of the resequenced genome to detect genomic rearrangements, rather than just SNPs. The resequencing approach can also be applied to mRNA, and such 'transcriptome resequencing' is a low-cost alternative for complex eukaryotic genomes. This approach focuses polymorphism detection on the key protein-coding regions of the genome, but cannot detect changes in most regulatory regions. Other approaches to reduce the amount of sequence required in resequencing projects

include 'reduced representation' resequencing, which uses biochemical methods to remove highly repetitive DNA before sequencing (Paterson 2006). The details of the latest technologies for SNP genotyping, genome and transcriptome sequencing and resequencing are discussed below.

SNP genotyping and genetic marker technologies

Direct detection of DNA sequence has been used to determine genotype by geneticists, forensic scientists, and plant and animal breeders for some time (Johnson 2004; Butler 2005). The use of any genetic marker in a combination of individuals is essentially genome resequencing, but only at one specific locus, base or allele. A major disadvantage of older molecular marker technologies is that they are usually based on unknown, randomly probed or amplified genomic sequence regions, and they cannot be targeted at specific SNPs, which can be discovered in large numbers using resequencing methods such as those discussed in this review. SNP genotyping technology is discussed below since it is so closely interrelated with genome resequencing technology.

Low- and medium-throughput SNP technologies

Low-throughput SNP detection strategies have been available for some time. One example from many such approaches, which is still widely used in the model plant community, is the gel-based Cut Amplified Polymorphic Sequence (CAPS) approach. The CAPS method requires that the SNP alters a restriction enzyme recognition site (Konieczny & Ausubel 1993) and can thus be detected by polymerase chain reaction (PCR) followed by restriction digestion. An extension of CAPS is the dCAPS approach, which can be used to engineer in a restriction site where one did not previously exist, making this approach applicable to any SNP in principle (Neff *et al*. 1998). CAPS and dCAPS are low throughput SNP assays, but they also require very little specialized equipment and reagents. Several proprietary PCR-based assays for SNP detection are now available which, unlike CAPS, do not require gel electrophoresis and thus allow higher throughput (e.g. the TaqMan SNP assay; De La Vega *et al*. 2005). However, none of the methods that require a PCR to be set up for every SNP to be assayed have the depth or throughput necessary for genotyping on a truly whole-genome scale. While some multiplexing is possible with newer PCR-based methods, higher-throughput technologies are necessary for whole-genome SNP genotyping.

In *ecotilling*, rather than using a particular marker to follow a gene, the researcher begins with a gene (usually defined using genome sequencing) which is a likely mediator of a phenotype of interest (Comai *et al*. 2004). The ecotilling method then allows the rapid detection of natural sequence variants in this gene within a population of genotypes, by PCR amplification of a population of alleles followed by mismatch detection using the CEL1 endonuclease to cleave DNA at polymorphic sites and gel electrophoresis. Variant alleles can then be followed using molecular markers and associated with particular traits of adaptive significance. A similar approach, the 'candidate gene approach', uses standard Sanger DNA sequencing technology to resequence a target gene in many individuals from a population in parallel (Tabor *et al*. 2002). The resulting haplotype information is then used to determine whether specific alleles can be associated with traits of interest. Both ecotilling and candidate gene sequencing allow a complete haplotype to be determined across a specified region of the genome, and are thus very powerful for gene-specific applications. However, both of these technologies are directed towards a single gene or region, and are thus not whole-genome genotyping technologies.

High-throughput microarray-based SNP methods

Another approach to genome-scale polymorphic sequence discovery and detection is based on differential hybridization of variant sequences to short-oligonucleotide microarrays (Tillib & Mirzabekov 2001; Borevitz *et al*. 2003; Borevitz *et al*. 2007). These approaches have the potential to process many thousands of loci in parallel, at a relatively low cost per locus, using a standard or custom microarray to resequence a genome at many specific, variable loci. These methods have many applications, present and future, in evolutionary and ecological genomics (Shiu & Borevitz 2006). The disadvantages of microarray-based resequencing methods can include a relatively low reliability of individual probes and relatively low sensitivity. The 'probe sets' of the Affymetrix expression arrays used by many researchers (in organisms where specialized SNP arras are not available) must be split and used as individual probes, removing a key advantage of the Affymetrix expression technology. Low reliability necessitates high levels of replication (either multiple microarrays, or custom-designed arrays with several probes per SNP) and complex statistical analysis. In addition, individual microarrays are expensive — while the cost per locus is acceptably low, the cost per individual to be genotyped is high compared to other marker-based approaches. However, microarrays have significant potential for the discovery of markers between diverse genotypes where no complete genome sequence is available, but a high-quality short-oligonucleotide microarray is in production (at the time of writing, this includes many crop plants such as barley, wheat, maize or soybean — see Rostoks *et al*. 2005). When combined with dCAPS, or other low-cost SNP genotyping technology, the polymorphisms discovered using standard expression arrays can be converted into useful markers. It is important to distinguish between the use of expression microarrays for polymorphism discovery

(as by Rostoks *et al*. 2005) and the use of custom-designed genotyping arrays (where each probe interrogates a known SNP). If many SNPs in an organism are known in detail, as in the human genome, a custom microarray can be created such that each feature on the array can interrogate one known SNP. Such SNP arrays permit very large scale genotyping studies (over 500 000 SNPs per individual) which can provide genotyping on a truly whole-genome scale in highly complex genomes (e.g. The Wellcome Trust Case Control Consortium 2007).

Other high-throughput SNP methods

Several other high-throughput methods for SNP genotyping on a whole-genome scale are now available in addition to microarray methods. All these methods can be used to genotype thousands of individuals at thousands of loci (Tsuchihashi & Dracopoli 2002). The latest bead-based methods for 'Whole Genome Genotyping' such as the Infinium assay (Gunderson *et al*. 2006) are capable of querying over 100 000 SNPs from one array. These methods have advantages over microarray-based methods as the sensitivity is often higher, and individual genotype calls can be more reliable and hence require less replication and statistical analysis.

Comparison of single-site resequencing technologies

Genotyping technologies can be divided into those that can discover previously unknown polymorphisms (such as *ecotilling* and the Borevitz array-based method) and those that give a genotype result at one or more known SNPs (such as dCAPS or the Infinium assay). The majority of newer technologies are aimed at known-SNP genotyping, since all SNP discovery methods must compete with ever-cheaper DNA sequencing. The main difference between known-SNP genotyping methods is that they lie on a spectrum between very low throughput assays and very high throughput assays. Low throughput assays (such as dCAPS) generally have a short development time, and low equipment and reagent costs, but a very high labour cost per SNP genotyped. These assays are thus well suited for applications such as fine genetic mapping in defined mapping populations, where rapid development and low cost is more important than throughput. High throughput assays generally have higher setup costs, but a much lower time and cost per SNP genotyped for large-scale efforts. Such assays are therefore suited to whole-genome genetic characterization, particularly when many individuals from a population must be genotyped. In the case of the highest throughput assays, such as the Affymetrix SNP arrays or the Infinium assay, such methods represent a currently viable and cheaper, if less complete, alternative to true whole-genome resequencing. These methods are practical where the polymorphisms within a population are already characterized in detail, and a large-scale project is justified (e.g. Borevitz *et al*. 2007; Sladek *et al*. 2007; The Wellcome Trust Case Control Consortium 2007). However, such methods can only genotype known SNPs, and true reseq-uencing is necessary to detect many types of mutation including unknown SNPs, insertion or deletion events (indels), and genomic rearrangements.

True sequencing and resequencing technologies

True sequencing technologies produce a 'read' of consecutive, sequenced bases from a DNA molecule. These technologies have no requirement for the DNA sequence or its variants to be known in advance. True DNA sequencing is hence essential for *de novo* sequencing or for the genome-wide identification of genetic variation, rather than the genotyping of known variants.

Whole-genome Sanger sequencing

All completed eukaryotic genomes, and the great majority of *de novo* sequencing projects currently in production, have used the long-established Sanger dideoxy technology with fluorescent dye-labelled terminators (Franca *et al*. 2002). This method will not be discussed in detail since it is so widely known and applied. It delivers reliable, high accuracy sequence in relatively long contiguous reads (> 700 bp) which can be mate-paired across known distances without loss of read length by the use of bacterial vectors with inserts of different sizes. The main drawbacks of this method are the barriers to further lowering cost and increasing throughput. A draft whole-genome sequence of a complex higher plant or mammal (e.g. a primate with three billion bases of DNA) requires 8× coverage in mate-paired reads and, thus, is still likely to cost around US$25 million in total (given a current estimate of cost per base of US 0.1 cent) and take several years. Despite constantly and rapidly falling costs, this technology will probably be too expensive in the near term to resequence multiple complete genomes for environmental or evolutionary biology research projects. While this is still the method of choice for *de novo* sequencing, a substantial investment must be made to perform *de novo* Sanger-based sequencing on any organism of interest.

'Next generation' sequencing

The limitations on how far the cost of whole-genome Sanger sequencing can fall are (i) the necessity to separate elongation products by size before scanning, requiring one capillary or gel lane per sample, and (ii) the need for the production of clonal populations of molecules by the use of *Escherichia coli* grown on nutrient media, which is labour- and space-intensive. The latter requirement can potentially be

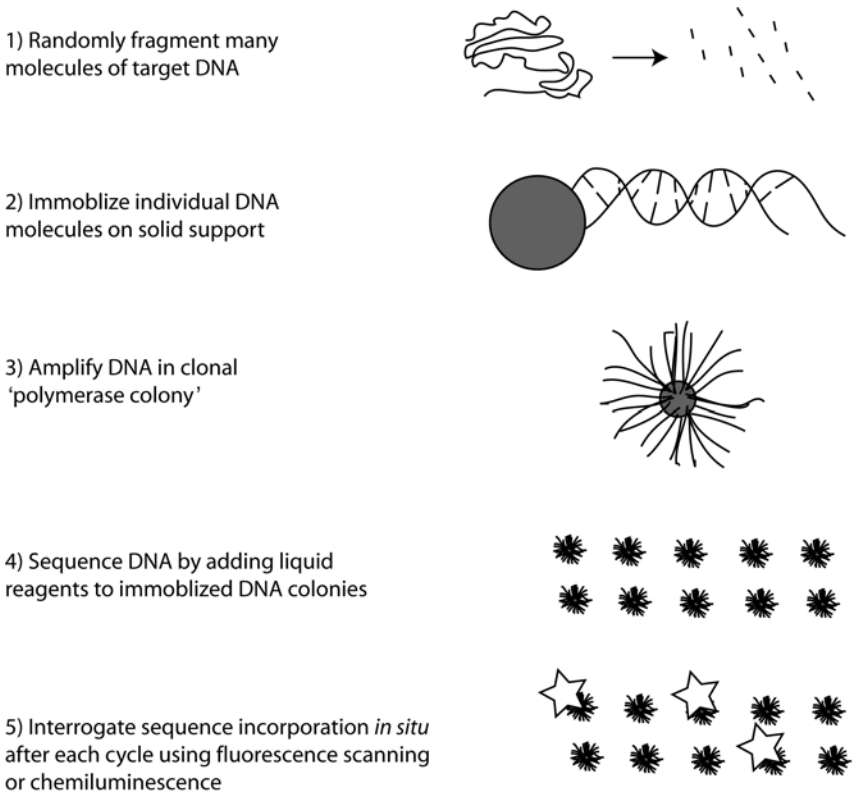

Fig. 2 A generalized description of the steps common to next-generation genome sequencing technologies. All these technologies involve genomic DNA random fragmentation, immobilization of single molecules on a solid support (a bead or planar solid surface), amplification by PCR, and subsequent *in situ* biochemical interrogation of the template DNA at each base in turn.

reduced by the use of PCR-based methods (although currently *E. coli* cloning is still used for all whole-genome sequencing), and the reaction costs can be reduced further, for example, by performing the sequencing reactions in very small volumes (Smailus *et al*. 2005), but the fundamental restrictions on how cheap Sanger sequencing can be made remain. For some time, many groups have been working on alternative technologies to increase the rapidity and/or throughput of DNA sequencing (Shendure *et al*. 2004; Kling 2005). New technologies, some already commercially available, circumvent some of the limiting requirements of Sanger sequencing and can thus be multiplexed to a much greater extent. The 'next-generation' sequencing methods now commercially available vary from many of those that are still under development (Shendure *et al*. 2004; Kling 2005) and share a common overall concept (Fig. 2).

For all next-generation genomic DNA sequencing methods commercially available as of August 2007, genomic DNA is randomly sheared, as it would be for creation of a conventional *E. coli* plasmid or phage library. Individual DNA molecules are then immobilized on a solid support, which can be a microscopic bead (in which case one molecule is affixed to each bead) or a macroscopic support such as a flow cell or slide (in which case many molecules are arrayed randomly on the support). These individual DNA molecules are then amplified using the PCR; in the case of bead-based methods, this is done in an emulsion phase where the beads are protected from cross-contamination by the barrier of an immiscible solvent. The result is a series of polymerase colonies, often called 'polonies' (Shendure *et al*. 2004) or 'clusters', which are clonally identical DNA molecules either attached to a single bead or attached to a localized region on a solid support. In bead-based methods, the beads are then either themselves immobilized on a planar support, or placed in individual microscopic wells—for non-bead-based methods the polymerase colonies are generated *in situ*. Once a planar array of polymerase colonies is produced, the sequencing chemistry (which varies between the several competing technologies, see below) is applied directly to the molecules on the support. Rather than separating elongation products, the sequence is interrogated at every base, by the use of either fluorescence or chemiluminescence to directly detect the incorporation of a base-specific chemical probe (Fig. 2). It is in this sequencing chemistry that the three 'next generation' sequencing methods now available have their greatest variation. All three methods have the potential to resequence multiple genotypes of complex eukaryotes. The currently available methods are discussed below, in the order in which they became available.

Pyrosequencing and 454 sequencing

The first of the 'next-generation' sequencing technologies to hit the mainstream market, 454 sequencing was developed by 454 Life Sciences (Branford, CT, USA; http://www.454.com/). 454 Life Sciences was originally a division of

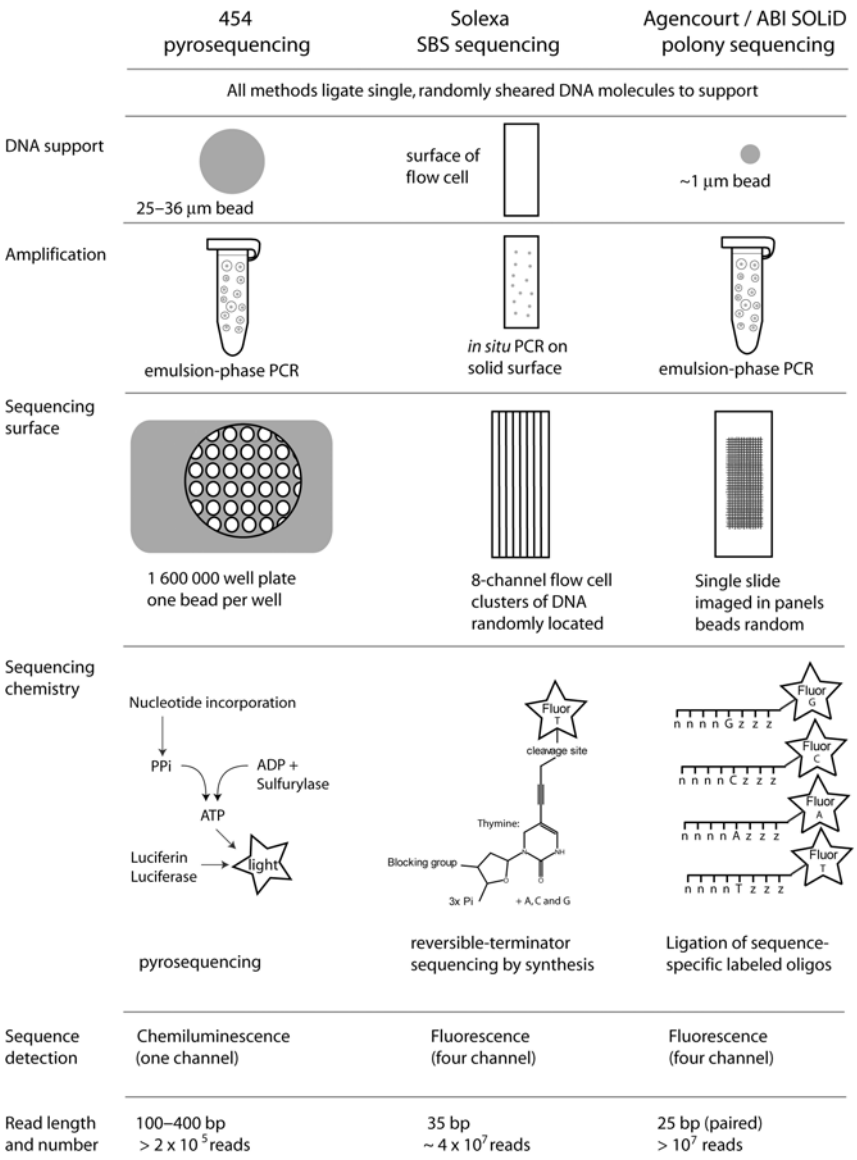

Fig. 3 A description of the key features of, and differences between, the three commercially available next-generation sequencing methods. The major steps in each procedure are arranged in the order in which they are performed by the operator or sequencing instrument. All three technologies share a common workflow, but differ greatly in the type of solid support used and the chemistry used to interrogate the DNA base pairs.

Curagen corporation and is now part of Roche. The method has now been in widespread use for almost 2 years, and is based on the technology known as pyrosequencing. Pyrosequencing is fundamentally different in concept to Sanger sequencing, in that it uses pyrophosphate release as a method for detection of base incorporation (Ronaghi *et al*. 1996, 1998; Rongahi 2001). Pyrosequencing has the significant advantage over Sanger sequencing that it requires no gels or capillaries to separate extension products by size, and that base incorporation can be detected in real time. In the 454 method, emulsion-phase PCR is first used to amplify randomly sheared DNA fragments linked to beads. The beads are then immobilized in a 1.6-million-well picotitre plate, and addition of nucleotides detected in a polymerase-mediated elongation by the release of pyrophosphate (Fig. 3). The standard 3730XL Sanger sequencing machines can perform 96 simultaneous reads, whereas the first-generation 454 Life Sciences system, known as the GS20, performed around 200 000 simultaneous reactions (Margulies *et al*. 2005). The current generation 'FLX' instrument performs around 400 000 simultaneous reactions.

As the best established of the new-generation methods, over 70 publications have so far resulted from purely 454-based sequencing, while the other methods discussed later have one or two publications at the time of writing. Because of this extensive data set, the disadvantages of the 454 approach are well known. The sequence accuracy from the original GS20 instrument was originally described as much lower than for Sanger technology, especially for homopolymer sequence regions. This limitation results from the main advantage of the method — that it detects the incorporation of unmodified, unlabelled bases in real time (Fig. 3).

Consequently, a run of a single nucleotide (e.g. six 'A' residues) will produce a single flash of chemiluminescence as all six residues are incorporated near-simultaneously. The only way to determine how many 'A's were in the run is to measure the intensity of the flash — this can discriminate runs of one, two and three bases relatively reliably but becomes increasingly inaccurate for longer homopolymeric stretches. However, as our understanding of the likely error types generated by 454 technology grows, so methods are being developed to filter low-quality base calls and increase accuracy (Huse *et al*. 2007).

The first-generation GS20 instrument delivered about 20 million bp of sequence per run, at a cost of approximately US$5000 per run (reagents only: excluding instrument and staffing costs), making the cost per base approximately 10-fold cheaper than contemporary Sanger sequencing. These instruments are now obsolete and have been largely replaced by the FLX system, although essentially all publications still use data from the GS20 system at the time of writing. The FLX system has per-base costs at least twofold lower as a result of longer sequence reads and more reads per run (around 400 000), but since Sanger sequencing also is constantly falling in cost the relative cost differential will probably be maintained. Both the GS20 and the FLX complete a run in a few hours. The main improvement in the newer instrument is in read length and error rate. Read length is up to over 200 bp in the currently available instruments, 400 000 reactions are performed per run, and raw base error rate is reportedly down to below 0.5% (compared to 4% in one study for the original GS20, however, see Huse *et al*. 2007 for a discussion of GS20 error models which can reduce the GS20 rate to well below 0.5%). FLX error data are estimates supplied by Roche, since no published FLX data set is available at the time of writing. There are no insuperable technical barriers to increasing read length enough for 454 to compete directly with Sanger sequencing (500+ bp) making 454 potentially poised to succeed Sanger as the most cost-effective method for *de novo* sequencing of eukaryotic genomes. It has already supplanted Sanger sequencing for many *de novo* sequencing projects of bacterial and archaeal genomes (e.g. Smith *et al*. 2007).

The availability of mate pairs — where two reads are separated by a known distance — is the main continuing advantage of Sanger sequencing over the 454 method for *de novo* sequencing of eukaryotic genomes. These paired reads are necessary for the assembly of large, complex genomes, and the ability of Sanger sequencing to place two reads of 700+ bp at a known, variable distance is still unmatched. Current methods for generating mate pairs exist for 454 sequencing, but they make the effective read length still shorter relative to Sanger methods, offsetting the advantage of paired reads. Despite these disadvantages, the 454 pyrosequencing technology is the best-proven alternative to Sanger methods, and also currently the next-generation technology that produces the greatest read lengths. It is thus currently the most suitable next generation technology for *de novo* genome sequencing and EST projects. However, the current cost of 454 sequencing, at several hundreds of thousands of dollars for a resequenced eukaryotic genome, is likely to limit its use for whole-genome ecology and evolution work.

Solexa/Illumina 1G SBS technology

An emerging rival technology to that produced by 454 Life Sciences is the technology developed by Solexa (Hayward, CA, USA) and now marketed by Illumina as the '1G Genome Analysis System' (http://www.illumina.com). The Illumina 1G system uses 'sequencing by synthesis' (SBS) technology and promises to deliver at least 1000 million bp per run (one gigabase or Gb), double for a mate pair run, at a cost per run two- to threefold lower than 454 sequencing (which delivers 100 million bases per run). The potential cost per base sequenced for this system can thus be estimated at approximately one-twentieth to one-thirtieth that of 454 sequencing, excluding the cost of the instrument itself. Such an advance has the potential to open entirely new fields to high-throughput sequencing. This technology has been exploited for mapping protein–DNA interactions (Johnson *et al*. 2007) and is primarily designed for whole-genome resequencing. The instrument generates read lengths of 35 bp, which can be extended to 50 bp with lower quality and longer run times. It takes 3 days to complete one standard run. Illumina 1G sequencing thus produces substantially shorter reads than 454, and the instrument takes longer to produce one run, but with the advantage of a much lower per-base cost.

Illumina 1G SBS technology employs a method more closely related to Sanger sequencing than 454 technology in that it employs terminator nucleotides which are incorporated by DNA polymerase (Fig. 3). Unlike the terminators used in Sanger sequencing, the four-colour Illumina terminators are reversible, allowing the continuation of polymerization after incorporation of a fluorophore. They also have removable fluorescent labels, allowing deactivation of the label from the previously incorporated base. Random genomic DNA fragments are immobilized on a solid surface, and after solid-phase amplification, sequence is determined by DNA synthesis using the reversible terminator chemistry (Fig. 3) and four-channel fluorescent scanning. While SBS technology lacks the limitation of 454 in resolution of homopolymeric regions, and coverage and accuracy are comparable to or better than 454 according to the manufacturers, the length of the sequence reads is likely to remain much shorter. The extremely short reads make the suitability of SBS for *de novo* genome sequencing questionable, although the cost advantage of this method makes it likely that attempts at this will be made soon. The limitation on read

length is the chemical reaction yield on the successive reactions needed to complete each sequenced base. Each base sequenced requires sequential addition, deprotection and fluor removal steps for the reversible terminators; while the DNA polymerase addition step is very efficient, the other steps have a yield restriction comparable to those for nucleotide synthesis. If, at each base, the combined yield of all the organic chemical reactions required to sequence that base is 95%, then less than 18% of the signal will remain from that polymerase cluster by the time base 35 is sequenced. The signal: noise ratio therefore rapidly becomes too low to permit the sequencing of long reads. The recent addition of the capability for mate-pair reads in Illumina 1G sequencing will double the number of bases available from a single run to 2 Gb (although double the run time to 6 days). The availability of mate pairs will also greatly increase the utility of this technology for sequencing that requires assembly, such as *de novo* transcriptome sequencing or the detection of complex genomic rearrangements in a resequencing project. Even with mate pairs, the current Illumina technology is likely still to fall short of the requirements of *de novo* whole-genome eukaryotic sequencing. However, for the key target of whole-genome resequencing where a near-identical genome is already complete, Illumina 1G SBS is currently the leading contender.

Agencourt/AB SOLiD technology

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) technology is not yet commercially available, but is currently being used by collaborators of the developers, Applied Biosystems (www.appliedbiosystems.com). The technology was originally developed for commercial applications by Agencourt Personal Genomics (Beverly, MA, USA) which subsequently was acquired by Applied Biosystems. SOLiD is based on 'polony sequencing' ligation chemistry for which instructions are available online using relatively standard laboratory reagents and a fluorescence microscope (note that all three technologies described above use a form of the polymerase colony or cluster concept, but 'polony sequencing' refers specifically to the sequencing chemistry developed by Shendure *et al*. 2005). The commercial version of the technology to be supplied by Applied Biosystems offers higher throughput than polony sequencing and also automation, dedicated equipment and technical support. This method is similar in some ways to the 454 Life Sciences approach, in that DNA is immobilized on beads and amplified using emulsion phase PCR before the direct sequencing of polony DNA on a solid surface (Fig. 3). However, the sequence detection chemistry of this method is fundamentally different from Sanger, pyrosequencing or Illumina 1G, as it uses DNA ligase and ligation rather than DNA polymerase and synthesis to interrogate sequence. In the original polony sequencing chemistry (Shendure *et al*. 2005) labelled, base-specific primers are used to create a sequence-specific ligation that interrogates a single base (Fig. 3). In the commercial AB SOLiD application each base will be interrogated twice, using a system of 'two base encoding'. This system increases the confidence of SNP detection but could complicate the use of standard software to align sequences — thus AB intend to provide freely available software to process this form of data.

The number of base pairs per run is currently quoted at about 2 Gb (identical to a mate-pair Illumina run), and costs are estimated to be similar to Illumina 1G sequencing. A single run takes about 4 days for the instrument to complete currently (a mate pair Illumina run currently takes 6 days). The read length of the first-generation SOLiD system will be fixed at either two mate-paired 25 bp reads, or individual 35 bp reads. As with Illumina 1G technology, read length is limited by yield, and also by other factors such as the sensitivity of the ligation assay over longer molecules. Like Illumina 1G technology and 454, the system also allows the use of a mate pair strategy (in the case of SOLiD, the mate pair can be up to 10 kb apart). Flexible paired ends improve the utility of the sequence data for genome assembly purposes, but the short reads are nonetheless likely to represent a disadvantage when compared to 454 technology for purposes such as *de novo* genome sequencing.

Other methods

While the above three methods are available to researchers at the time of writing, several other sequencing methods are under various stages of development. One such method, Single Molecule Sequencing, has received a great deal of attention and has potential to further increase throughput. In many respects this technology can resemble the previously described technologies, in that DNA is immobilized on a solid surface and fluorescently labelled nucleotides are sequentially added and scanned. However, unlike the three above methods, the method developed by Helicos Biosciences (Cambridge, MA, USA) does not incorporate an amplification step; rather, single, individual DNA molecules are directly sequenced. Various single-molecule and other sequencing and resequencing methods are under development in academic laboratories, and at several companies, including Biotage, Helicos, Li-Cor, Microchip Biotechnologies, Nanofluidics, Nanogen, Network Biosystems and Visigen. The technologies under development include sequence determination by drawing DNA through nanopore sensors, sequencing by hybridization, microfluidics, microelectronics, sequencing by atomic force microscopy and several competing versions of the single-molecule technology described above (Shendure *et al*. 2004; Kling 2005).

Comparison of sequencing technologies

As cost is the main factor in determining the most suitable sequencing technology for any project, current rough estimates are provided, with the additional caveat that these costs may not represent true current costs and may also change rapidly. Currently, Sanger sequencing at major centres costs around US 0.1 cent per base, 454 sequencing costs a little over 0.01 cent per base, and Illumina 1G sequencing costs around 0.0005 cent per base. Costs for SOLiD when commercialized are likely to be in the same range as Illumina 1G sequencing.

With all next-generation sequencing technology, data collection, storage and analysis costs must also be borne in mind when calculating the overall cost of sequencing. When so much sequence is produced so quickly and cheaply, computer resources and informatics personnel can become a significant part of the budget for a project. Availability of tried and tested computational infrastructure and software can also influence the choice of an appropriate technology for a particular project. For all next-generation technologies, bioinformatics software is not yet optimized to the level that researchers are accustomed to for Sanger sequencing. This is particularly true of the very-short-read technologies at the time of writing. Assembly algorithms and software necessary for *de novo* genome assembly with short-read mate pairs have still to be developed, making this a field of intense activity in bioinformatics research. Since bioinformatics is such a major part of any sequencing project, this must be borne in mind when selecting the optimal technology for a particular project.

Which, if any, of the available sequencing technologies is appropriate for a particular project depends not only on the cost per base delivered by that technology, but on whether the technology will allow the completion of the project, and, if so, the depth of sequencing required for the envisaged goal. Depth of sequencing is generally calculated as the number of genome equivalents that must be sequenced, with 1 × coverage referring to the sequencing of one genome equivalent, 2× of two genome equivalents and so on. The required depth for either *de novo* sequencing or resequencing is always much greater than 1×, and depends on many factors, chiefly sequence read length, sequence quality and, for some applications, whether mate pairs are available. The properties of the genome under investigation are also a factor, since repetitive content and recent genomic duplication will favour the use of longer read and mate-pair technologies. Also the relative percentage of the nucleotides G and C in the genome can affect sequence quality and read length, and this effect will vary according to the technology selected.

For *de novo* sequencing using mate paired Sanger sequencing, 8× coverage is generally considered necessary, and 10× coverage preferable, for assembly of a complex genome (Siegel *et al*. 2000). A common drawback of all of the next-generation sequencing methods is that the length of the sequence reads, and often also the accuracy of the base calls, can be substantially lower than achieved by state-of-the-art Sanger methods. While the cost savings of this technology over Sanger sequencing are potentially huge, it must be emphasized that the lower cost for all next-generation technologies must be offset to some extent by the need for greater depth of coverage. Because errors can be solved when the same base is sequenced multiple times, and shorter reads tend to generate more gaps in coverage, it is generally accepted that sequencing must be performed in greater depth using these methods (i.e. each base must be covered by many reads).

For *de novo* sequencing using 200 bp, non-mate-paired reads as produced by the 454 FLX, a naive calculation using the methods of Lander & Waterman (1988) suggests that 15× coverage would be sufficient for a draft assembly of a less complex (minimally repetitive) genome. However, imperfect sequence quality will increase the necessary depth, as will the presence of repeats and homopolymers, so a sequence depth of around 30× is generally recommended. Nonetheless, despite the increased need for sequence depth, a genome which is small and simple enough to be assembled using 454 sequence data can potentially be completed much more quickly and cheaply with 454 than with Sanger sequencing. Any repetitive sequence longer than 200 bp will, however, likely produce insoluble problems in assembly without the availability of mate-paired reads. In principle, 25 or 35 bp mate-paired reads as produced by AB SOLiD or Illumina 1G technology could also be used to assemble genomes *de novo*, if depth of coverage of 30× or more were available. However, reads this short will likely cause insoluble problems in assembly from the repeat stretches in many genomes, and the impact of low sequence quality will be even greater from short reads. The availability of mate pairs may alleviate this problem to the extent that small, minimally repetitive genomes (such as some bacteria) could be assembled with high (> 30×) coverage, as potentially could genomes from an organism with a fully sequenced close relative that could provide an assembly template. With SOLiD, Illumina 1G or 454, an investigator would be able to use the paired information of two reads that are separated by a known distance (which can be flexible, up to 10 kb) by the use of a Type IIs restriction enzyme along with other manipulations to cut and recircularize DNA. The Type IIs method is currently supported by 454 and will be supported by SOLiD, it is also possible using the Illumina 1G technology but no manufacturer-supported kit is currently available. However, this method involves some loss of read length and thus reduces total base pairs per run. Thus, with 454 and SOLiD, some read length is lost by using the mate-pair strategy, and cost per base increases. With Illumina 1G

another option is offered, with mate pair distance limited to a maximum of around 600 bp by sequencing both ends of the bridge PCR product. Unlike the Type IIs methods, no sequence data or read length is lost by using this mate-pair strategy and thus Illumina 1G provides mate pairs with no read length penalty, but with a pair-distance restriction which reduces some of the advantages of mate pairs.

For EST or transcriptome sequencing, Sanger sequencing remains the current standard, but several groups have successfully used 454 technology to generate EST or sheared transcriptome sequences (e.g. Cheung *et al*. 2006; Emrich *et al*. 2007; Toth *et al*. 2007). In the case of *de novo* EST or transcriptome sequencing, more sequence depth is always an advantage, since it allows better coverage of rare mRNA sequences. Conversely, even a relatively small, inexpensive amount of next-generation sequencing can produce the sequence of several thousand genes from an organism with no existing genomic resources (Toth *et al*. 2007). Very short read technologies such as Illumina 1G or Solexa may not allow effective *de novo* assembly of transcripts that are not deeply covered or do not have a closely related template as well as longer read technologies such as 454, even with mate pairs. Resolution of genes with very similar sequences will also be more challenging with very short read technologies. However, given the depth of coverage one sequence run using very short read methods will provide, many thousands of genes from a complex organism will likely be covered with sufficient depth for effective assembly, and this is likely to be very useful for transcriptome resequencing with a species or in closely related species.

For whole-genome resequencing, the necessary coverage can be estimated using an adaptation of the equation of Clarke & Carbon (1976), and depends on the size of the genome and the acceptable probability of any given sequence being 'missed' by the resequencing project, but not on read length. Read length is, however, important because the reads must be long enough to be unique in the genome. In practice, even 25-bp reads are generally unique in the genomic regions of most interest to biologists, but problems arise in larger genomes with closely related gene families, recent whole-genome duplications and repetitive regions, because some of the reads cannot then be unambiguously assigned to one locus. Short reads therefore leave significant fractions of complex genomes that cannot be resequenced, regardless of depth of coverage, and the shorter the reads are, the worse the problem. Shorter reads also lead to a different pattern of resequencing coverage, with more (but shorter) gaps. These factors significantly reduce evenness of coverage with short read technologies. Mate pairs can help to alleviate this problem by providing unique positions for many more reads — mate pairs also help to resolve changes such as large deletions and rearrangements.

Despite potential drawbacks, very short read technologies (Illumina and potentially SOLiD) are much better suited to resequencing than *de novo* sequencing. If it is desired that the probability of missed sequence be sufficiently low that none of the genome be missed by the sequencing project, then approximately 20× coverage is necessary for genomes in the gigabase size range, regardless of read length (but given the above caveat that regions which are not unique may still be missed). However, if a 0.001 probability of missing any given polymorphism is considered acceptable, then coverage as low as 7× may be adequate if the sequence data is of sufficient quality. While a resequencing experiment with 7× coverage may therefore be considered inadequate for discovery of a single disease-causing SNP thought to lie in the unique coding regions of the human genome, it could well be acceptable in a survey of ecological diversity.

Conclusions

The cost, labour and time required to perform sequencing on the scale needed for whole-genome analysis has fallen significantly as a result of the recent development of new technologies, and this trend is likely to continue. This change is already impacting the cost of whole-genome resequencing where a genome sequence is already available. It is likely also to have a huge impact on the biology of nonmodel organisms, such as those central to most evolutionary and environmental biology. However, short-read technologies are only gradually becoming accepted by the genomics community, mostly because early technologies had problems with sequence quality. Additionally, short read methods have serious drawbacks for *de novo* sequencing. Currently, only small fragments of complex eukaryotic genomes could be assembled using short read sequences. This low-cost sequence data can still be useful, as genome surveys can be performed using next-generation technology to provide insight into genomes with otherwise limited available data (Swaminathan *et al*. 2007), and transcriptome sequences can be cheaply and quickly obtained (Cheung *et al*. 2006; Emrich *et al*. 2007; Toth *et al*. 2007). However, the most powerful use of short-read sequence data is comparison to other genomes that are sequenced at high quality, and initially therefore next-generation sequencing is likely to be of greatest benefit to genomics in fully sequenced species and their close relatives. Nonetheless, the impact of next-generation sequencing is already being felt in taxa distantly related to the classic molecular model organisms (Toth *et al*. 2007).

For *de novo* genome sequencing of complex eukaryotes, Sanger methods are likely to remain the method of choice for a short while at least, because of accuracy and read length, and because costs are likely to fall further with competition from next-generation technologies. For bacterial *de novo* genome sequencing, 454 is already making significant progress in replacing conventional sequencing (Smith *et al*. 2007). The combination of a relatively low-coverage

sequence using conventional Sanger sequencing with mate-paired reads, combined with high-coverage, short read sequencing, has the potential to allow whole-genome assembly as well as lower *de novo* genome sequencing costs. This approach may provide an intermediate stage in the transition to next-generation sequencing; however, combined Sanger and 454 does not currently offer the dramatic cost savings that exclusive application of next-generation methods does.

An immediately applicable and inexpensive way in which next-generation sequencing can speed genomics in non-model complex eukaryotes is by transcriptome sequencing. Transcriptome sequences using 454 technology are likely to emerge soon for many more eukaryotes of evolutionary and/or environmental interest (M. E. Hudson & G. E. Robinson, unpublished). While such sequences are not replacements for a whole-genome sequence, they allow the large-scale analysis of the evolution of gene sequences and expression, as well as the application of tools such as microarrays for gene expression analysis. Using such methods, costs for *de novo* transcriptome sequencing are likely to fall to the extent that most species will become feasible targets for such a project. In contrast, true *de novo* whole-genome sequencing is likely to be restricted in the immediate future to a few widely studied eukaryotes with large research communities. Mass *de novo* whole-genome sequencing of environmentally and ecologically important species will require still more advanced and lower cost technologies than those currently on the market. Given the current rate of progress, however, such technologies may be available in the near future.

Whole-genome resequencing, while not applicable to organisms without an existing whole-genome sequence, is likely to produce significant advances in the study of populations of model organisms and their close relatives. Currently, of the commercially available and priced methods for whole-genome resequencing (454 and Illumina at the time of writing), the Illumina 1G technology is likely to prove much more cost-effective at current pricing and output, since the approximately 20-fold lower per-base cost is likely to outweigh other advantages of 454 (chiefly longer reads). AB SOLiD is aimed to compete head-on with Illumina technology. Whichever technology wins the comparison on cost-effectiveness, the competition between the two is likely to further drive down resequencing costs. The 454 method is likely to remain much more expensive in cost per base than Illumina 1G or AB SOLiD. However, for whole-genome resequencing, it is likely to have substantial advantages for any experiment where the underlying genome sequence is not fully characterized or is highly repetitive.

For the routine analysis of diversity, SNPs are likely to remain the method of choice, since SNP polymorphisms in single-copy regions of the genome are widely thought of as the 'low hanging fruit' that cause much phenotypic variation and can track haplotypes at high resolution and relatively low cost. For SNP discovery on a whole-genome or transcriptome basis, the Illumina 1G and potentially SOLiD methods are likely to produce the lowest cost, highest density SNP discovery data of the currently available technologies. Methods such as the Infinium assay can then provide a still lower-cost, high-throughput way to characterize the genomes of very large numbers of individuals, making projects achievable that otherwise could not be contemplated. If the promises of the manufacturers are met, these methods are very likely to revolutionize the entire field of genetics in organisms with existing whole-genome sequences. The use of transcriptome sequencing and resequencing provides an immediate means by which these developments can be expanded into nonmodel organisms without existing genome sequences.

The coming era of genomic evolutionary and environmental biology is likely to bring ecologists, evolutionary, computational and molecular biologists closer together in terms of their techniques and knowledge. The burden of experimental difficulty is likely to be shifted from the laboratory-based research that currently dominates genomics, and towards the genetic, environmental and evolutionary interpretation of whole-genome information. The complexity and sheer amount of genomic data will be unprecedented, and will require significant advances in bioinformatics and computational biology to allow effective experimental design, data analysis and interpretation. The next few years will be an exciting time in evolutionary and environmental biology.

Acknowledgements

Thanks are due to Drs Karen Kaczorowski (Purdue University), Gene Robinson (University of Illinois), Lei Du (454/Roche), Christian Haudenschild (Solexa/Illumina) and Brandon Blakey (Applied Biosystems), for useful suggestions and critical reading of the manuscript. Three anonymous reviewers also improved the manuscript significantly through helpful and detailed comments. Research in Matthew Hudson's laboratory is funded by the US Department of Agriculture, the US National Science Foundation and the State of Illinois.

References

Adams MD, Kelley JM, Gocayne JD *et al*. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

Angly FE, Felts B, Breitbart M *et al*. (2006) The marine viromes of four oceanic regions. *PLoS Biology* **4**, e368.

Borevitz JO, Hazen SP, Michael TP *et al*. (2007) Genome wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA*, **104**, 12057–12062.

Borevitz JO, Liang D, Plouffe D *et al*. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Research*, **13**, 513–523.

Butler J (2005) *Forensic DNA Typing*, 2nd edn. Elsevier, Amsterdam, The Netherlands.

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature*, **325**, 31–36.

Check E (2007) James Watson's genome sequenced. *Nature* online 10.1038/news070528-10.

Cheung F, Haas BJ, Goldberg SMD, May GD, Xiao Y, Town CD (2006) Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics*, **7**, 272.

Clarke L, Carbon J (1976) A colony bank containting synthetic Col E1 hybrid plasmids representative of the entire E. coli genome. *Cell*, **9**, 91–99.

Comai L, Young K, Till BJ *et al*. (2004) Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. *Plant Journal*, **37**, 778–786.

De La Vega FM, Lazaruk KD, Rhodes MD, Wenz MH (2005) Assessment of two flexible and compatible SNP genotyping platforms: Taqman® SNP Genotyping Assays and the Snplex™ Genotyping System. *Mutation Research*, **573**, 111–135.

Ellstrand NC, Prentice HC, Hancock JF (1999) Gene flow and introgression from domesticated plants into their wild relatives. *Annual Review of Ecology and Systematics*, **30**, 539–563.

Emrich SJ, Barbazuk WB, Schnable PS (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, **17**, 69–73.

Fleishmann RD, Adams MD, White O *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Franca LTC, Carrilho E, Kist TBL (2002) A review of DNA sequencing technologies. *Quarterly Reviews of Biophysics*, **35**, 169–200.

Green P (1997) Against a whole-genome shotgun. *Genome Research*, **7**, 410–417.

Gunderson KL, Steemers FJ, Ren H *et al*. (2006) Whole-genome genotyping. *Methods in Enzymology*, **410**, 359–376.

Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, **68**, 669–685.

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nature Reviews. Genetics*, **6**, 95–108.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively-parallel DNA pyrosequencing. *Genome Biology*, **8**, R143.

Jeffreys AJ, Wilson V, Thein SL (1985) Individual-specific 'fingerprints' of human DNA. *Nature*, **316**, 76–79.

Johnson GR (2004) Marker-assisted selection. *Plant Breeding Reviews*, **24**, 293–309.

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein–DNA interactions. *Science*, **316**, 1497–1502.

Kling J (2005) The search for a sequencing thoroughbred. *Nature Biotechnology*, **23**, 1333–1335.

Konieczny A, Ausubel FM (1993) A procedure for mapping Arabidopsis mutations using codominant ecotype-specific PCR-based markers. *Plant Journal*, **4**, 403–410.

Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

Margulies M *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

National Human Genome Research Institute (2003) *Concept Papers for Two New DNA Sequencing Technology Development Programs*. http://www.genome.gov/11008124.

Neff MM, Neff JD, Chory J, Pepper AE (1998) dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in Arabidopsis thaliana genetics. *Plant Journal*, **14**, 387–392.

Paterson AH (2006) Leafing through the genomes of our major crop plants: strategies for capturing unique information. *Nature Reviews. Genetics*, **174**, 174–184.

Richards S, Liu Y, Bettencourt BR *et al*. (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene and *cis*-element evolution. *Genome Research*, **15**, 1–18.

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, **242**, 84–89.

Ronaghi M, Uhlen M, Nyren P (1998) A sequencing method based on real-time pyrophosphate. *Science*, **281**, 363–365.

Rongahi M (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Research*, **11**, 3–11.

Rostoks N, Borevitz JO, Hedley PE *et al*. (2005) Single-feature polymorphism discovery in the barley transcriptome. *Genome Biology*, **6**, R54.

Schloss PD, Handelsman J (2003) Biotechnological prospects from metagenomics. *Current Opinion in Biotechnology*, **14**, 303–310.

Shendure J, Mitra RD, Varna C, Church GM (2004) Advanced sequencing technologies: methods and goals. *Nature Reviews. Genetics*, **5**, 335–344.

Shendure J, Porreca GJ, Reppas NB *et al*. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728–1732.

Shiu S-H, Borevitz JO (2006) The next generation of microarray research: applications in evolutionary and ecological genomics. *Heredity* (online 8 November 2006; doi 10.1038/sj.hdy.6800916).

Siegel AF, van den Engh G, Hood L, Trask B, Roach JC (2000) Modeling the feasibility of whole-genome shotgun sequencing using a pairwise end strategy. *Genomics*, **68**, 237–246.

Sladek R, Rocheleau G, Rung J*et al*. (2007) A genome–wide association study identifies novel risk loci for type 2 diabetes. *Nature*, **445**, 881–885.

Smailus DE, Marziali A, Dextras P, Marra MA, Holt RA (2005) Simple, robust methods for high-throughput nanoliter-scale DNA sequencing. *Genome Res*, **15**, 1447–1450.

Smith MG, Gianoulis TA, Pukatzki S *et al*. (2007) New insights into *Acinetobacter baumanii* pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. *Genes & Development*, **21**, 601–614.

Sogin ML, Morrison HG, Huber JA *et al*. (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proceedings of the National Academy of Sciences, USA*, **103**, 12115–12120.

Swaminathan K, Varala K, Hudson ME (2007) Global repeat discovery and estimation of genomic copy number in a large, complex genome using a 454 sequence survey. *BMC Genomics*, **8**, 132.

Tabor HK, Risch NJ, Myers RM (2002) Candidate gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews. Genetics*, **3**, 391–397.

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature*, **447**, 661–678.

Tillib SV, Mirzabekov AD (2001) Advances in the analysis of DNA sequence variations using oligonucleotide microchip technology. *Current Opinion in Biotechnology*, **12**, 53–58.

Toth AL, Varala K, Newman TC *et al*. (2007) Wasp brain gene expression supports an evolutionary link between maternal behavior and eusociality. *Science*, doi: 10.1126/science.1146647.

Tringe SG, von Mering C, Kobayashi A *et al*. (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.

Tsuchihashi Z, Dracopoli NC (2002) Progress in high throughput SNP genotyping methods. *Pharmacogenomics Journal*, **2**, 103–110.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 21–28.

Venter JC, Remington K, Heidelberg JF *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.

Woese CR, Kandler O, Wheelis ML. (1990) Towards a natural system of organisms: proposal for the domains archaea bacteria, and eucarya. *Proceedings of the National Academy of Sciences, USA*, **87**, 4576–4579.

Wright SI, Lauga B, Charlesworth D (2003) Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. *Molecular Ecology*, **12**, 1247–1263.

Yap WH, Zhang Z, Wang Y (1999) Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon. *J Bacteriol*, **181**, 5201–5209.