The Journal of Experimental Biology 209, 1518-1525
Published by The Company of Biologists 2007
doi:10.1242/jeb.001370

Advanced sequencing technologies and their wider impact in microbiology

Neil Hall

School of Biological Sciences, Biosciences Building, Crown Street, University of Liverpool, Liverpool L69 7ZB, UK

e-mail: neilhall@liv.ac.uk

Accepted 13 March 2007

Summary

In the past 10 years, microbiology has undergone a revolution that has been driven by access to cheap high-throughput DNA sequencing. It was not long ago that the cloning and sequencing of a target gene could take months or years, whereas now this entire process has been replaced by a 10 min Internet search of a public genome database. There has been no single innovation that has initiated this rapid technological change; in fact, the core chemistry of DNA sequencing is the same as it was 30 years ago. Instead, progress has been driven by large sequencing centers that have incrementally industrialized the Sanger sequencing method. A side effect of this industrialization is that large-scale sequencing has moved out of small research labs, and the vast majority of sequence data is now generated by large genome centers. Recently, there have been advances in technology that will enable high-throughput genome sequencing to be established in research labs using bench-top instrumentation. These new technologies are already being used to explore the vast microbial diversity in the natural environment and the untapped genetic variation that can occur in bacterial species. It is expected that these powerful new methods will open up new questions to genomic investigation and will also allow high-throughput sequencing to be more than just a discovery exercise but also a routine assay for hypothesis testing. While this review will concentrate on microorganisms, many of the important arguments about the need to measure and understand variation at the species, population and ecosystem level will hold true for many other biological systems.

Glossary available online at
http://jeb.biologists.org/cgi/content/full/210/9/1518/DC1

Key words: comparative genomics, microbe, mutation screening, metagenomics, genome sequencing, bacteria, microorganism.

Introduction

Is there anything left to sequence?

The first bacterial genome, that of *Haemophilus influenzae*, was published in 1995 (Fleischmann et al., 1995). This was the first sequence of a free-living species to be completely decoded. The genome was sequenced at The Institute for Genomic Research using the Whole Genome Shotgun (WGS) method. The data from this project, which included 1 830 137 bp of DNA and 1743 predicted genes, laid out, for the first time, the full genetic complement of a bacterial organism. Within 5 years of this publication, numerous other bacteria were sequenced, including *Mycobacterium tuberculosis* (Cole et al., 1998), one of the most important human bacterial pathogens, *Escherichia coli* (Blattner et al., 1997) and the first archaeon, *Archaeoglobus fulgidus* (Klenk et al., 1997). Since then, eukaryotic microbes have been sequenced, such as the malaria parasite *Plasmodium falciparum* (Gardner et al., 2002a; Gardner et al., 2002b; Hall et al., 2002; Hyman et al., 2002) and yeast (Goffeau et al., 1997). These sequences, along with the large genomes of mammals such as human (Lander et al., 2001), mouse (Waterston et al., 2002) and chimpanzee (Mikkelsen et al., 2005), have led to the massive expansion of sequence data available today.

It is clear that genome sequencing has spearheaded a revolution in the biological sciences by allowing the study of molecular processes in the context of complete cellular systems, thus leading to the concept of 'systems biology'. Genome sequence is also the foundation of the 'omics' technologies such as proteomics and transcriptomics (such as microarrays). Despite its success, a casual observer of the genomics field might easily believe that there was no requirement for more genome sequencing, as almost all of the major model organisms and important human and animal pathogens have been sequenced. I will argue in this review that sequencing has yet to reach its full potential as a tool for discovery and hypothesis testing. I will draw upon three examples where the potential of new technologies has been, or

soon will be, demonstrated: comparative genomics, mutation screening and metagenomics. I will start by describing briefly what the technologies are.

Old and new sequencing technologies

The Sanger sequencing method (Sanger et al., 1977) has been the workhorse technology for DNA sequencing for almost 30 years. This method relies on synthesizing DNA on a single-stranded template while randomly incorporating chain terminators. This generates a range of different fragment sizes that correspond to the positions of the terminators. The older methods would require four reactions per template (one for each base: G, A, T and C), each reaction having a different base as a terminator. The reactions are then run on a gel to identify the size of each fragment. Improvements were made in the 1990s with the use of different colored fluorescent dyes to label terminators (Prober et al., 1987; Smith et al., 1986), so that all of the terminators can be incorporated in a single reaction. The first sequencing machines used this technology in combination with devices to automatically read fragments as they were separated on a polyacrylamide gel. Later, the gels were replaced by capillaries, which simplified the separation step and increased the length of reads (Madabhushi, 1998). In the past 10 years, the average length of a sequencing read has increased from around 450 bp to 850 bp. Despite these technological advances in the Sanger method, whole-genome sequencing is predominantly carried out at large dedicated genome centers that can each house up to 100 sequencing machines and that have the capacity to run them >10 times per day and 365 days per year, using highly automated template preparation pipelines. Without such an infrastructure in place, the cost and workload of generating enough sequencing to decode even a relatively small genome are highly prohibitive.

Recent developments in enzymology, imaging and microfluidics may offer a new approach to sequencing that could yield a massive increase in capacity while removing the need for the huge infrastructure required today. In this review, I will not give an exhaustive list of new technologies but I will describe a few of the published techniques that appear most promising. These can be separated into two approaches: sequencing with amplification and single-molecule sequencing. Fig. 1 gives an overview of some of the different sequencing strategies.

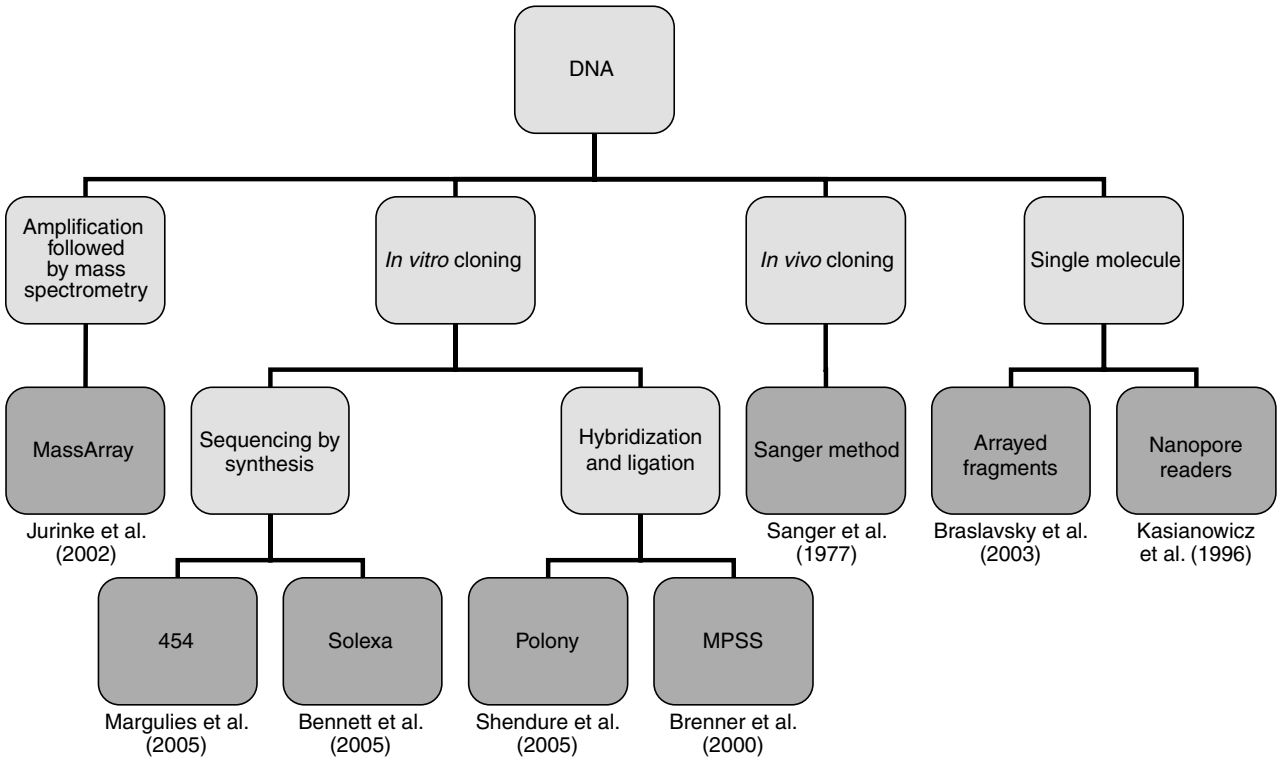

Fig. 1. An overview of current and emerging technologies for genomic sequencing. Sequencing can be classified into four main strategies: *in vitro* cloning, *in vivo* cloning, amplification and mass spectrometry, and single-molecule approaches. The mass spectrometry and single-molecule approaches are still either very specialized or in the developmental stages, although mass spectrometry methods such as the MassArray method is commonly used for single nucleotide polymorphism (SNP) analysis (Jurinke et al., 2002). *In vivo* cloning followed by Sanger sequencing is the workhorse method of most current genome sequencing projects. The *in vitro* cloning technologies can be further divided into methods that employ sequencing by synthesis, such as the 454 and Solexa methods, or those that use hybridization and ligation of oligonucleotides, such as MPSS (massively parallel signature sequencing) and polony methods.

New technologies for sequencing with amplification

The first step in most sequencing processes is to amplify the DNA. This is necessary because measuring biochemical processes at a single-molecule resolution is so technically challenging. In the Sanger method, this is usually done by cloning the DNA into a plasmid and growing clones; however, this has its pitfalls as DNA is a biologically active molecule, hence there are inherent biases against certain stretches of DNA that have physical properties that do not replicate well in *E. coli* or that code for toxic compounds. The two methods I will discuss here are the Margulies et al. method (Margulies et al., 2005), also known as 454 sequencing after 454 Life Sciences (Branford, CT, USA), which has commercialized it, and the Shendure et al. method (Shendure et al., 2005), also known as polony sequencing (Fig. 2). Both have developed high-throughput strategies for *in vitro* amplification that are very cheap and also get around the inherent biases of *in vivo* methods.

454 sequencing is, at the time of writing, the only new sequencing technology that has been widely deployed. The 454 method is similar to the polony method in that it involves massively parallel sequencing by synthesis on a solid support.

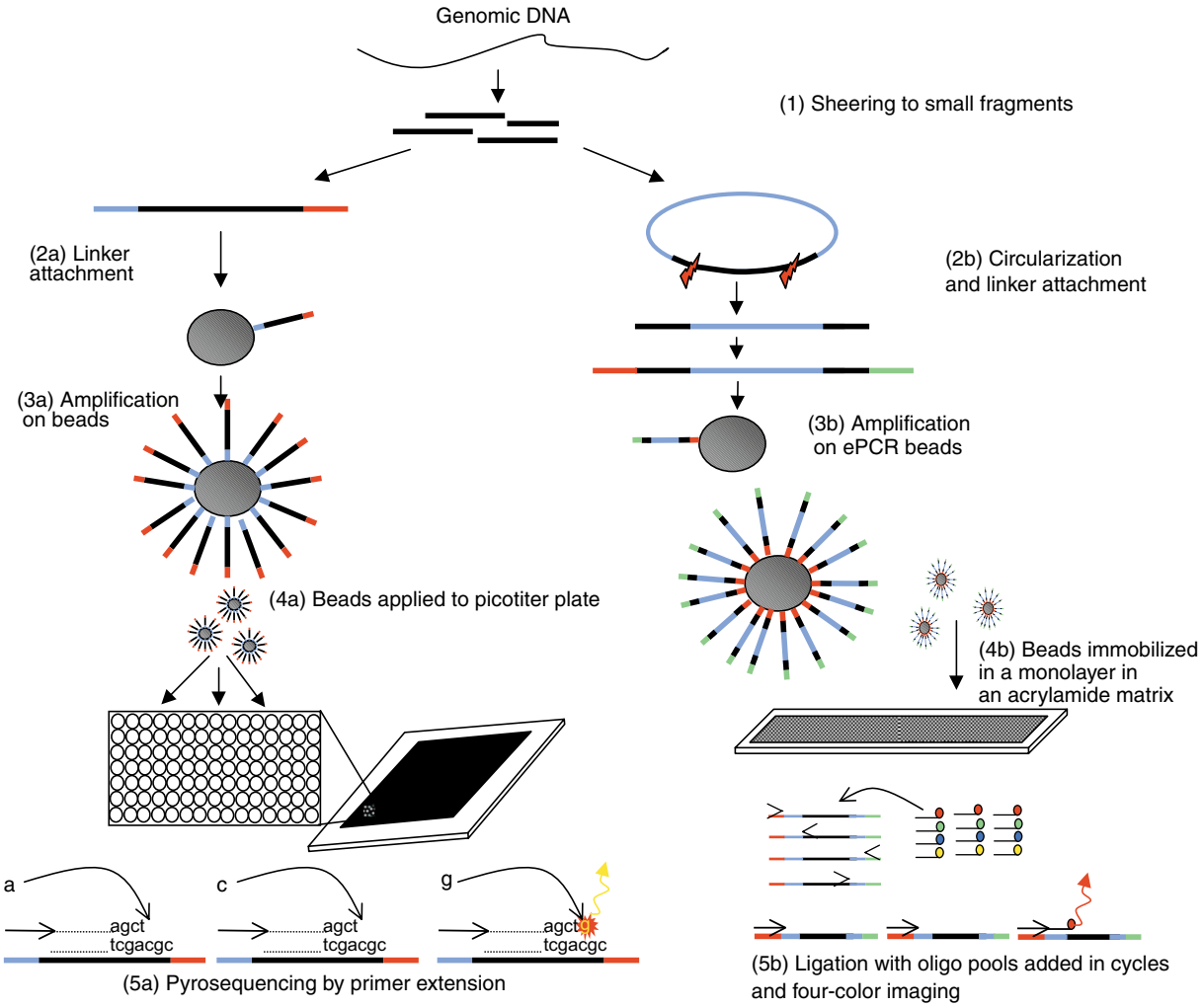

Fig. 2 Outline of the 454 and polony sequencing process. Both systems first fragment the genomic DNA (Step 1) and then use a process of *in vitro* cloning followed by amplification. The 454 process is shown on the left and Polony sequencing is shown on the right. In the 454 protocol, the linkers are ligated onto the ends of the DNA (Step 2a). Polony sequencing involves circularization followed by linearization and the addition of linkers to generate two fragments with a spacer between them and linkers at the end (Step 2B). Both processes then attach the *in vitro* clones to beads and carry out PCR in an emulsion mixture to generate beads with many clonal copies of the target fragments (Step 3a/3b). For the sequencing step, the beads must be immobilized in a single layer to allow imaging in an environment that enables the reaction reagents to be flowed across them. In the case of 454 sequencing, a picotiter plate is used, in which most cells will contain a single bead (Step 4a). The polony method immobilizes the beads in an acrylamide matrix in a dense monolayer (Step 4b). The methods are very similar up until the point of the sequencing reaction; in the case of 454 sequencing, a DNA synthesis reaction from a single sequencing primer is carried out. Bases are flowed across the picotiter plate one at a time and incorporation is detected by the release of light (Step 5a). The polony method uses ligation to anchor primers, which can be annealed in one of four positions. In each cycle, a population of degenerate nonomers, which have been fluorescently labeled, is added to the monolayer, and only complimentary oligos will anneal and ligate to the anchor primer.

The method allows reads as long as 250 bp (and the maximum read length is expected to increase further in the coming year) and is therefore at least approaching the read lengths obtainable through traditional methods. Margulies et al. have devised a scalable, highly parallel two-step sequencing approach (Margulies et al., 2005). The first step involves shearing the genome and attachment of oligonucleotides, a process that circumvents the need for generating a clone library. Adapters are ligated to the fragments and these are bound to beads and captured in the droplets of an oil-emulsion PCR reaction mixture. PCR amplification in each droplet results in each bead carrying 10 million copies of a unique DNA template. In the second step, a modified pyrosequencing (Ronaghi et al., 1996) protocol is carried out, in which nucleotide incorporation is detected by the release of inorganic pyrophosphate and the generation of photons.

Polony sequencing involves an *in vitro* library construction step that generates two paired genomic tags in a linear molecule separated by a universal linker and a universal tag on either end. Millions of these molecules are circularized using the linker ends and amplified in-parallel in a single reaction tube by a process of emulsion PCR using beads containing primers to the universal tags (very similar to the 454 method). The beads are then immobilized on a flow cell for sequencing. An unusual aspect of the polony technique is that it does not use primer extension replication for the sequencing stage but instead relies on the hybridization and ligation of oligonucleotides. First, an anchor primer is hybridized to one of the universal sequences, and then degenerate nonamers, which are labeled using fluorescent dyes, are hybridized to the template and then ligated to the anchor primer. The pools of nonamers are structured so that the base in the degenerate position corresponds to the color of the fluorescent dye labeling it. The nonamers will only ligate if the sequence is complementary to the bases adjacent to the anchor primer, therefore the sequence of the template can be derived. The sequence generated by this technique is very accurate and also benefits from having paired reads. A single run can generate around 30 Mb of sequence, with an estimated cost per kilobase of raw sequence that is 10-fold less than conventional sequencing. The disadvantage of this technique is the short read length, which is currently 26 bp per amplicon (13 bp per tag). The polony method has now been taken on by Applied Biosystems (Foster City, CA, USA). They have adapted the method so it is capable of 50 bp reads and generating >1 Mb of sequence in a single run. The technology (now named SOLiD) is expected to be brought to market in 2007.

Another method for massively parallel sequencing by synthesis from amplified fragments has been recently developed by a company called Solexa (Bennett, 2004; Bennett et al., 2005). Solexa sequencing differs from polony or 454 sequencing as it amplifies the DNA on a solid surface followed by synthesis by incorporation of modified nucleotides linked to colored dyes. Solexa sequencing will not be covered in depth here as (at the time of writing) the methodology has not been published in detail. However, as this review goes to press, Solexa have released their first instrument that is capable of sequencing over 1 Gb in a single run and is likely to have a major impact on the genomics field.

Single-molecule sequencing

Many of the problems, and inherent errors, of DNA sequencing result from the fact that thousands or millions of amplified templates are assessed in a single reaction. It would be far better to read DNA in the same way as cells do; as single molecules. The first published report of single-molecule sequencing was by the lab of Stephen Quake (Braslavsky et al., 2003). This method involves hybridizing target DNA to complimentary primers that are streptavidin–biotin bound to a silica surface. The primers are then extended by the addition of Cy3- and Cy5-labeled nucleotides; as each base is added, the incorporation is captured using a camera mounted on a microscope. A limitation of this technology is that it generates short reads, which at the time of publication was 5 bp; however, this technology has been taken up by a company (Helicos Biosciences Corporation, Cambridge, MA, USA) who are reporting much longer reads. This method is highly parallel, and on a 25 mm square it would be possible to sequence 12 million templates simultaneously, so, even with 5 bp reads, each ‘run’ would generate 60 million bases of information.

One other method of single-molecule sequencing that is in the very early stages of development involves ‘reading’ DNA as it is passed through a nanopore (Kasianowicz et al., 1996; Storm et al., 2005a; Storm et al., 2005b). This would not involve an enzymatic extension reaction of any kind but instead the physical properties of the molecule would be read as the bases wind through a tiny pore. In theory, this method would have no limit on read length and, hence, if the technical hurdles are overcome it could revolutionize how genome sequencing is achieved.

Read length, read quality and read pairs

When considering how a sequencing technology can be used for specific purposes, it is important to consider three quality measures: read length, read quality and read pairing. If reads are very short, then they are of limited use for *de novo* assembly of complete genomes. Although some simple bacterial genome assemblies have been carried out on reads of less than 50 bp, for the vast majority of genomes, assembly would be impossible. The ability to generate read pairs is also vital for assembly of large genomes as it allows distant regions of the genome to be linked. In Sanger sequencing, this is achieved by cloning large inserts and taking reads from both ends, but this is problematic for most new technologies. Short, single reads are still very useful for comparative studies where the aim is to identify single nucleotide polymorphisms (SNPs) or larger differences between a reference genome and a newly sequenced genome. This type of study requires high-quality reads and hence the error rate for any method used should be low.

Currently, Sanger sequencing outperforms all of the new technologies in these metrics of quality. Hence, efforts are underway to incorporate Sanger sequencing data into 454 sequence assemblies to improve the consensus quality. Because the reads and error distribution for new technologies are very different from Sanger methods, the tools needed to process them and assemble them are different. This means, frustratingly, that it is very difficult to mix Sanger sequencing reads with other types of reads and assemble them together, although some progress has been made in this direction (Goldberg et al., 2006; Wicker et al., 2006).

Comparative genomics: the need for more *de novo* genome sequencing

The fact that there are 279 complete bacterial genomes in the public databases (at the time of writing) sounds impressive, but recent estimates suggest that there could be 10^7 distinct bacterial taxa in only 10 g of pristine soil (Curtis and Sloan, 2005; Gans et al., 2005); it therefore follows that for the vast majority of microbes there is no genome sequence data at all. For the few 'lucky' species that have been selected for genomic analysis, there is usually only one reference genome.

For a few pathogenic microbes, multiple species have been sequenced, and the data from these studies have revealed that a single reference genome, while useful, may only give a snapshot of the genetic makeup of a species. A recent study of group B *Streptococcus* strains (Tettelin et al., 2005) revealed that, as each new strain was sequenced, new genes were discovered such that, after sequencing eight genomes, approximately 33 novel genes were discovered from each additional genome. This has led to the concept of the 'Pan-genome', which refers to the full gene repertoire contained within a species. The Pan-genome theory predicts that any bacterial species will be made up of a core set of genes that is found in all individuals and a dispensable set of genes that may or may not be present in any particular individual (Medini et al., 2005; Tettelin et al., 2005). This phenomenon seems to be applicable to most other microorganisms examined, and subtractive hybridization studies of *E. coli* suggest that up to 25% of the genome is specific to individual strains (Fukiya et al., 2004). By sequencing more and more individuals, the scale of the Pan-genome can be estimated. So, for *Bacillus anthracis*, no more new genes were identified after four species were sequenced whereas for group B *Streptococcus* and *E. coli* it is estimated that the number of strains needed to survey the Pan-genome is at least in the hundreds and effectively may be infinite. An important finding from this work is that for many species, the dispensable gene set may be significantly larger than the core genome. Therefore, a single genome may give a very poor representation of the genetic potential of the species. When predicting the chance of emergence of drug resistance or new virulent forms of pathogens, knowledge of the complete genetic complement of the species is far more important than the genetic complement of an individual.

Not only do more genomes allow for the discovery of more genes but they also help us to understand how genes and genomes are evolving, as this can provide clues to gene function. Pathogen genes that are interacting with the host are often subject to positive selection (and therefore appear to be evolving rapidly). Genome-wide molecular evolution studies have been applied to various pathogens such as *Plasmodium* (Hall et al., 2005), *Trypanosoma* (El-Sayed et al., 2005), *Borrelia* (Qiu et al., 2004) and many other species. These studies depend on tracing the pattern of mutations that occur in synonymous and non-synonymous sites by aligning orthologous genes in closely related species. The more genomes that can be aligned, the more accurate this analysis is. The studies to date have used up to four genomes at a time but as sequencing becomes more affordable it will be possible to scale this analysis up to look at tens or hundreds of genomes at a time.

Mutation screening

Genome sequencing is not yet being routinely used as a hypothesis-testing technology. The reason that we are limited in our ability to use genomic data is that a single reference genome does not provide enough data to allow correlations between genotypes and phenotypes. For example, the *Haemophilus influenzae* genome is only a single data point so we can't correlate the sequence to a phenotype. If genomes from say 100 strains of *H. influenzae* were sequenced, one could test hypotheses about which genes were linked to drug resistance, virulence or transmissibility, etc. However, there is a technology gap between the questions we would like to ask and what is feasible with current methods. To sequence 100 *Haemophilus* genomes (let alone 100 human genomes) would be completely impractical using traditional Sanger-based techniques and there is a requirement for new methods to allow genomics to address complex genetic questions.

One of the most obvious applications of cheaper, more high-throughput genome sequencing of microbes is for mutation screening. This may be carried out at the population level, to identify associations between phenotypes and genotypes, or in lab-generated strains, to identify SNPs or larger mutations that have given rise to selected phenotypes. Currently, there are a number of platforms that allow SNP screening using microarrays but these require the array to be pre-designed and they will not resolve large genomic changes such as insertions or inversions relative to the reference sequence. Recent work on experimentally evolved species has demonstrated how new sequencing methods can be used to track mutations that have been acquired in the laboratory.

Shendure et al. used polony sequencing to screen an evolved strain of an *E. coli* auxotroph (Shendure et al., 2005). The sequencing was able to identify a number of SNPs as well as larger deletions and inversions. This work demonstrated that, despite the small amount of data obtained per clone (26 bp), it was possible to identify large-scale rearrangements in the

genome and align fragments to identify SNPs. In a similar study of the cooperative bacterium *Myxococcus xanthus* (Velicer et al., 2006), a laboratory-evolved strain that had been selected for a cheating phenotype and reselected for a cooperative phenotype was shotgun sequenced using 454 sequencing technology. The 454 sequence was able to identify point mutations in the evolved strain compared with the reference strain, which could then be associated with the changes in phenotype (as well as identifying errors in the reference).

While whole-genome sequencing may still be prohibitively expensive for detection of point mutations, we may expect prices to fall for these new technologies, as they have in the past for Sanger sequencing. Due to their small genome size, microbes will be in the first wave of organisms to be studied this way and we can expect direct whole-genome sequencing to replace many other forward genetic techniques for the study of very specific traits.

Metagenomics

Metagenomics, or community genomics, is an approach aimed at analyzing the genomic content of microbial communities living in any particular niche such as the human gut or the soil. The problem of studying the microbial composition of an environmental sample is one that has baffled microbiologists for some time. The challenge is confounded by the sheer diversity of microbes that are present in even the most extreme environments, along with the fact that only a small proportion of the species are actually culturable. Genomic analysis has been used to circumvent these problems as it can allow the analysis of non-culturable organisms, and molecular phylogenetic analysis can be used to study the taxonomic diversity of the organisms present. The added advantage of genomic methods is that the analysis of gene content will also give an indication of the metabolic potential of an environment.

Metagenomic studies have been applied already to human environments such as the human gut (Breitbart et al., 2003; Gill et al., 2006; Manichanh et al., 2006; Zhang et al., 2006), environmental samples such as soil (Bertrand et al., 2005; Lim et al., 2005; Mills et al., 2006) and the ocean (Breitbart et al., 2004; Culley et al., 2006; DeLong et al., 2006; Sogin et al., 2006; Venter et al., 2004). These studies have provided interesting findings in terms of the metabolic capability and taxonomic diversity of the microbes inhabiting these environments. The major goal of these metagenomic studies is not only to find new biological species and systems but also to be able to identify biomarkers that can be used to classify the type of processes that occur in specific environments. For example, what processes and species are more commonly found in a diseased gut compared with a healthy one? Or which species or processes associate with polluted as opposed to pristine environments?

A major problem with this preliminary work is that the diversity is probably not fully sampled because of the complexity of the environments studied. It has been recently estimated that close to 10^7 distinct bacterial species inhabit a 10 g soil sample (Curtis and Sloan, 2005; Curtis et al., 2002; Gans et al., 2005); this is a species diversity two orders of magnitude higher than previous estimates. If each of these species had an average genome size of 3–5 Mb, this would mean that a single sample would contain the equivalent of 1000 human genomes. Even if the species were present in equal amounts then a large sequencing center would have to dedicate its entire resource for years to sample all of the genomes present. Unfortunately, the problem is still more complex than that; the new higher estimate is based on the finding that there is greater diversity in the low-abundance species that are masked by a less diverse group of high-abundance species. Hence, current studies only scrape the surface of the full diversity and most of the low-abundance species in the environments are not sampled at all. New highly parallel sequencing technologies offer a cost-effective solution to this problem as they can generate much more sequence than traditional methods. However, there are limitations to their utility because non-Sanger methods have shorter read lengths and are therefore more difficult to assemble. Two recent studies using 454 pyrosequencing have demonstrated the power of new sequencing technologies for this type of analysis: one analyzing the massive diversity in the oceans (Sogin et al., 2006) and the other analysing a low-complexity environment (Edwards et al., 2006).

The first study set out to measure the number of species in the Earth's ocean biosphere by using massively parallel sequencing to sufficiently sample the low-abundance taxa in order to make more accurate estimate of their diversity (Sogin et al., 2006). Using the 454 pyrosequencing technology, 118 000 amplicons were sequenced that spanned the V6 hypervariable region of the ribosomal RNA (rRNA) from bacteria collected at different depths and locations of the Atlantic and Pacific oceans. The resulting sequences were compared to a database of all known V6 regions in order to place them phylogenetically. Clustering of these sequences defined Operational Taxonomic Units (OTUs). In each sample, over 1000 OTUs were identified, and in the most sampled environment over 3000 OTUs were identified. In no environment did rarefaction analysis suggest that the sampling had reached a plateau, as the number of OTUs identified increased almost linearly with the sequencing of new tags. Although the authors of this study made specific efforts to control for sequencing errors, it is possible that some of the diversity observed was caused by the inherent base calling errors that occur in 454 sequencing reads, and the findings of this study should therefore be verified by other methods. Although this study was insufficient for measuring diversity, it still demonstrated the inadequacy of other methods and will increase estimates of natural diversity further.

In the second study, two water samples from adjacent sites that differed significantly in their chemistry and geology were

analyzed (Edwards et al., 2006). 454 sequencing was used to generate random sequence from each sample. Over 35 Mb of sequence was generated from both samples in short reads and therefore the challenge was to be able to analyze these data to identify processes and taxonomic groups that would allow a comparison of the microbial diversity in the two environments. The 16S reads that were present in the sample were used to identify the species present; this demonstrated that the oxygenated environment had a much higher species diversity than the oxygen-poor environment. This result was verified by using Sanger sequencing of an rRNA library from each sample.

In addition to looking at species, Edwards et al. also analyzed the metabolic potential of the different communities by automatically assessing gene function by homology searches of sequence reads against a metabolic database (Edwards et al., 2006). Using this analysis they identified processes that were significantly overrepresented in one sample relative to the other. This study was able to focus on biological processes as well as diversity, as the environments in question were far less complex that the ocean environment studies by Sogin et al. (Sogin et al., 2006). However, as the technologies used become faster and cheaper, it may be possible to deeply sequence complex environments. These studies are not only limited by sequencing, however, and there will need to be improvements in genomic assembly and annotation in order to analyze the data generated.

Conclusion

Genome sequencing has provided us with powerful insights into the genetic make-up of the microbial world and has spearheaded a host of revolutionary technologies, such as microarrays and proteomics, that have transformed the field of microbiological research. Yet DNA sequencing has only scratched the surface of the genetic diversity present in the real world. There are a number of new technologies that are now in development that promise to reinvigorate the genomics field as they massively increase throughput while markedly decreasing the cost of DNA sequencing.

Importantly, these technologies will enable researchers to undertake the process of genomic sequencing in a single operation using bench-top instruments. This will democratize a technology that, until now, has largely been the preserve of large genome centers. It is hoped that once this process can be viewed as an assay – in the same way that we view a microarray experiment – whole-genome sequencing will be applied to a host of new questions, such as genotype association studies, mutation screening, evolutionary studies and environmental profiling.

It may be that the term ‘post-genomics’ has been prematurely inserted into the scientific lexicon and we are in fact on the cusp of a genome sequencing renaissance.

I am grateful to Ian Paulson for his help and advice in writing this manuscript.

References

Bennett, S. (2004). Solexa Ltd. *Pharmacogenomics* **5**, 433-438.

Bennett, S. T., Barnes, C., Cox, A., Davies, L. and Brown, C. (2005). Toward the 1,000 dollars human genome. *Pharmacogenomics* **6**, 373-382.

Bertrand, H., Poly, F., Van, V. T., Lombard, N., Nalin, R., Vogel, T. M. and Simonet, P. (2005). High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. *J. Microbiol. Methods* **62**, 1-11.

Blattner, F. R., Plunkett, G., III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F. et al. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1474.

Braslavsky, I., Hebert, B., Kartalov, E. and Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. USA* **100**, 3960-3964.

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P. and Rohwer, F. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185**, 6220-6223.

Breitbart, M., Felts, B., Kelley, S., Mahaffy, J. M., Nulton, J., Salamon, P. and Rohwer, F. (2004). Diversity and population structure of a near-shore marine-sediment viral community. *Proc. Biol. Sci.* **271**, 565-574.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M. et al. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630-634.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., III et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537-544.

Culley, A. I., Lang, A. S. and Suttle, C. A. (2006). Metagenomic analysis of coastal RNA virus communities. *Science* **312**, 1795-1798.

Curtis, T. P. and Sloan, W. T. (2005). Microbiology. Exploring microbial diversity – a vast below. *Science* **309**, 1331-1333.

Curtis, T. P., Sloan, W. T. and Scannell, J. W. (2002). Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci. USA* **99**, 10494-10499.

DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N. U., Martinez, A., Sullivan, M. B., Edwards, R., Brito, B. R. et al. (2006). Community genomics among stratified microbial assemblages in the ocean’s interior. *Science* **311**, 496-503.

Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., Saar, M. O., Alexander, S., Alexander, E. C., Jr and Rohwer, F. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7**, 57.

El-Sayed, N. M., Myler, P. J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H., Worthey, E. A., Hertz-Fowler, C. et al. (2005). Comparative genomics of trypanosomatid parasitic protozoa. *Science* **309**, 404-409.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M. et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512.

Fukiya, S., Mizoguchi, H., Tobe, T. and Mori, H. (2004). Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray. *J. Bacteriol.* **186**, 3911-3921.

Gans, J., Wolinsky, M. and Dunbar, J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**, 1387-1390.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S. et al. (2002a). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511.

Gardner, M. J., Shallom, S. J., Carlton, J. M., Salzberg, S. L., Nene, V., Shoaibi, A., Ciecko, A., Lynn, J., Rizzo, M., Weaver, B. et al. (2002b). Sequence of *Plasmodium falciparum* chromosomes 2, 10, 11 and 14. *Nature* **419**, 531-534.

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M. and Nelson, K. E. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355-1359.

Goffeau, A., Aert, R., Agostini-Carbone, M. L., Ahmed, A., Aigle, M., Alberghina, L., Albermann, K., Albers, M., Aldea, M., Alexandraki, D. et al. (1997). The yeast genome directory. *Nature* **387(Suppl.)**, 1-105.

Goldberg, S. M., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., Khouri, H., Kravitz, S. A., Lauro, F. M. et al. (2006). A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci. USA* **103**, 11240-11245.

Hall, N., Pain, A., Berriman, M., Churcher, C., Harris, B., Harris, D., Mungall, K., Bowman, S., Atkin, R., Baker, S. et al. (2002). Sequence of *Plasmodium falciparum* chromosomes 1, 3-9 and 13. *Nature* **419**, 527-531.

Hall, N., Karras, M., Raine, J. D., Carlton, J. M., Kooij, T. W., Berriman, M., Florens, L., Janssen, C. S., Pain, A., Christophides, G. K. et al. (2005). A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science* **307**, 82-86.

Hyman, R. W., Fung, E., Conway, A., Kurdi, O., Mao, J., Miranda, M., Nakao, B., Rowley, D., Tamaki, T., Wang, F. et al. (2002). Sequence of *Plasmodium falciparum* chromosome 12. *Nature* **419**, 534-537.

Jurinke, C., van den Boom, D., Cantor, C. R. and Koster, H. (2002). The use of MassARRAY technology for high throughput genotyping. *Adv. Biochem. Eng. Biotechnol.* **77**, 57-74.

Kasianowicz, J. J., Brandin, E., Branton, D. and Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. USA* **93**, 13770-13773.

Klenk, H. P., Clayton, R. A., Tomb, J. F., White, O., Nelson, K. E., Ketchum, K. A., Dodson, R. J., Gwinn, M., Hickey, E. K., Peterson, J. D. et al. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390**, 364-370.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Lim, H. K., Chung, E. J., Kim, J. C., Choi, G. J., Jang, K. S., Chung, Y. R., Cho, K. Y. and Lee, S. W. (2005). Characterization of a forest soil metagenome clone that confers indirubin and indigo production on *Escherichia coli*. *Appl. Environ. Microbiol.* **71**, 7768-7777.

Madabhushi, R. S. (1998). Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions. *Electrophoresis* **19**, 224-230.

Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., Nalin, R., Jarrin, C., Chardon, P., Marteau, P. et al. (2006). Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55**, 205-211.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z. et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.

Medini, D., Donati, C., Tettelin, H., Masignani, V. and Rappuoli, R. (2005). The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589-594.

Mikkelsen, T., Hiller, L. W., Eichler, E. E., Zody, M. C., Jaffe, D. B., Yang, S. P., Enard, W., Hellmannm, I., Linbald-toh, K. and Altheide, T. K. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87.

Mills, D. K., Entry, J. A., Voss, J. D., Gillevet, P. M. and Mathee, K. (2006). An assessment of the hypervariable domains of the 16S rRNA genes for their value in determining microbial community diversity: the paradox of traditional ecological indices. *FEMS Microbiol. Ecol.* **57**, 496-503.

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A. and Baumeister, K. (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336-341.

Qiu, W. G., Schutzer, S. E., Bruno, J. F., Attie, O., Xu, Y., Dunn, J. J., Fraser, C. M., Casjens, S. R. and Luft, B. J. (2004). Genetic exchange and plasmid transfers in *Borrelia burgdorferi* sensu stricto revealed by three-way genome comparisons and multilocus sequence typing. *Proc. Natl. Acad. Sci. USA* **101**, 14150-14155.

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

Sanger, F., Nicklen, S. and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D. and Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. and Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M. and Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl. Acad. Sci. USA* **103**, 12115-12120.

Storm, A. J., Chen, J. H., Zandbergen, H. W. and Dekker, C. (2005a). Translocation of double-strand DNA through a silicon oxide nanopore. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **71**, 051903.

Storm, A. J., Storm, C., Chen, J., Zandbergen, H., Joanny, J. F. and Dekker, C. (2005b). Fast DNA translocation through a solid-state nanopore. *Nano. Lett.* **5**, 1193-1197.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S. et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. USA* **102**, 13950-13955.

Velicer, G. J., Raddatz, G., Keller, H., Deiss, S., Lanz, C., Dinkelacker, I. and Schuster, S. C. (2006). Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor. *Proc. Natl. Acad. Sci. USA* **103**, 8107-8112.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W. et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66-74.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562.

Wicker, T., Schlagenhauf, E., Graner, A., Close, T. J., Keller, B. and Stein, N. (2006). 454 sequencing put to the test using the complex genome of barley. *BMC Genom.* **7**, 275.

Zhang, T., Breitbart, M., Lee, W. H., Run, J. Q., Wei, C. L., Soh, S. W., Hibberd, M. L., Liu, E. T., Rohwer, F. and Ruan, Y. (2006). RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* **4**, e3.

The Journal of Experimental Biology 210, 1492-1496
Published by The Company of Biologists 2007
doi:10.1242/jeb.02783

Glossary of terms

This section is designed to help readers adapt to the complex terminology associated with contemporary molecular genetics, genomics and systems biology. Fuller descriptions of these terms are available at http://www.wikipedia.org/

Ab initio prediction	methods used to predict the potential genes encoded in the genome, which are trained on datasets made of known genes, and used computationally to predict coding regions out of genome without the aid of cDNA sequence. Although their performance is improving, these algorithms perform very poorly on non-protein coding genes.
Annotation	as applied to proteins, DNA sequences or genes. The storage of data describing these entities (protein/gene identities, DNA motifs, gene ontology categorisation, etc.) within a biological database. Active projects include FlyBase and WormBase. See Gene ontology.
Assembly	the process of aligning sequenced fragments of DNA into their correct positions within the chromosome or transcript.
cDNA	complementary DNA. This is DNA synthesised from a mature mRNA template by the enzyme reverse transcriptase. cDNA is frequently used as an early part of gene cloning procedures, since it is more robust and less subject to degradation than the mRNA itself.
ChIP	chromatin immunoprecipitation assay used to determine which segments of genomic DNA are bound to chromatin proteins, mainly including transcription factors.
Chip	see Microarray.
ChIP-on-chip	use of a DNA microarray to analyse the DNA generated from chromatin immunoprecipitation experiments (see ChIP).
cis-acting	a molecule is described as *cis*-acting when it affects other genes that are physically adjacent, on the same chromosome, or are genetically linked or in close proximity (for mRNA expression, typically a promoter).
Collision-induced dissociation	a mechanism by which molecules (e.g. proteins) are fragmented to form molecular ions in the gas phase. These fragments are then analysed within a mass spectrometer to provide mass determination.
Connectivity	a term from graph theory, which indicates the number of connections between nodes or vertices in a network. Greater connectedness between nodes is generally used as a measure of robustness of a network.
CpG islands	regions that show high density of ‘C followed by G’ dinucleotides and are generally associated with promoter elements; in particular, stretches of DNA of at least 200 bp with a C–G content of 50% and an observed CpG/expected CpG in excess of 0.6. The cytosine residues can be methylated, generally to repress transcription, while demethylated CpGs are a hallmark of transcription. CpG dinucleotides are under-represented outside regulatory regions, such as promoters, because methylated C mutates into T by deamination.
Edge	as in networks. Connects two nodes (or vertices) within a system. These concepts arise from graph theory.
Enhancer	a short segment of genomic DNA that may be located remotely and that, on binding particular proteins (*trans*-acting factors), increases the rate of transcription of a specific gene or gene cluster.
Epistasis	a phenomenon when the properties of one gene are modified by one or more genes at other loci. Otherwise known as a genetic interaction, but epistasis refers to the statistical properties of the phenomenon.

eQTL	the combination of conventional QTL analysis with gene expression profiling, typically using microarrays. eQTLs describe regulatory elements controlling the expression of genes involved in specific traits.
EST	expressed sequence tag. A short DNA sequence determined for a cloned cDNA representing portions of an expressed gene. The sequence is generally several hundred base pairs from one or both ends of the cloned insert.
Exaptation	a biological adaptation where the current function is not that which was originally evolved. Thus, the defining (derived) function might replace or persist with the earlier, evolved adaptation.
Exon	any region of DNA that is transcribed to the final (spliced) mRNA molecule. Exons interleave with segments of non-coding DNA (introns) that are removed (spliced out) during processing after transcription.
Gene forests	genomic regions for which RNA transcripts, produced from either DNA strand, have been identified without gaps (non-transcribed genomic regions). Conversely, regions in which no transcripts have ever been detected are called ‘gene deserts’.
Gene interaction network	a network of functional interactions between genes. Functional interactions can be inferred from many different data types, including protein–protein interactions, genetic interactions, co-expression relationships, the co-inheritance of genes across genomes and the arrangement of genes in bacterial genomes. The interactions can be represented using network diagrams, with lines connecting the interacting elements, and can be modelled using differential equations.
Gene ontology (GO)	an ontology is a controlled vocabulary of terms that have logical relationships with each other and that are amenable to computerised manipulation. The Gene Ontology project has devised terms in three domains: biological process, molecular function and cell compartment. Each gene or DNA sequence can be associated with these annotation terms from each domain, and this enables analysis of microarray data on groups of genes based on descriptive terms so provided. See http://www.geneontology.org
Gene set enrichment analysis	a computational method that determines whether a defined set of genes, usually based on their common involvement in a biological process, shows statistically significant differences in transcript expression between two biological states.
Gene silencing	the switching-off of a gene by an epigenetic mechanism at the transcriptional or post-transcriptional levels. Includes the mechanism of RNAi.
Genetic interaction (network)	a genetic interaction between two genes occurs when the phenotypic consequences of a mutation in one gene are modified by the mutational status at a second locus. Genetic interactions can be aggravating (enhancing) or alleviating (suppressing). To date, most high-throughput studies have focussed on systematically identifying synthetic lethal or sick (aggravating) interactions, which can then be visualised as a network of functional interactions (edges) between genes (nodes).
Genome	a portmanteau of gene and chromosome, the entire hereditary information for an organism that is embedded in the DNA (or, for some viruses, in RNA). Includes protein-coding and non-coding sequences.
Heritability	phenotypic variation within a population is attributable to the genetic variation between individuals and to environmental factors. Heritability is the proportion due to genetic variation usually expressed as a percentage.
Heterologous hybridization	the use of a cDNA or oligonucleotide microarray of probes designed for one species with target cRNA/cDNAs from a different species.
Homeotic	the transformation of one body part to another due to mutation of specific developmentally related genes, notably the *Hox* genes in animals and *MADS-box* genes in plants.
Hub	as in networks. A node with high connectivity, and thus which interacts with many other nodes in the network. A hub protein interacts with many other proteins in a cell.

Hybridisation	the process of joining (annealing) two complementary single-stranded DNAs into a single double-stranded molecule. In microarray analysis, the target RNA/DNA from the subject under investigation is denatured and hybridised to probes that are immobilised on a solid phase (i.e. glass microscope slide).
Hypomorph	in genetics, a loss-of-function mutation in a gene, but which shows only a partial reduction in the activity it influences rather than a complete loss (cf. hypermorph, antimorph, neomorph, etc).
Imprinting	a phenomenon where two inherited copies of a gene are regulated in opposite ways, one being expressed and the other being repressed.
Indel	insertion and deletion of DNA, referring to two types of genetic mutation. To be distinguished from a 'point mutation', which refers to the substitution of a single base.
Interactome	a more or less comprehensive set of interactions between elements within cells. Usually applied to genes or proteins as defined by transcriptomic, proteomic or protein–protein interaction data.
Intron	see Exon.
KEGG	The Kyoto Encyclopedia of Genes and Genomes is a database of metabolic and other pathways collected from a variety of organisms. See http://www.genome.jp/kegg
Metabolomics	the systematic qualitative and quantitative analysis of small chemical metabolite profiles. The metabolome represents the collection of metabolites within a biological sample.
Metagenomics	the application of genomic techniques to characterise complex communities of microbial organisms obtained directly from environmental samples. Typically, genomic tags are sequence characterised as markers of each species to inform on the range and abundance of species in the community.
Microarray	an arrayed set of probes for detecting molecularly specific analytes or targets. Typically, the probes are composed of DNA segments that are immobilised onto the solid surface, each of which can hybridise with a specific DNA present in the target preparation. DNA microarrays are used for profiling of gene transcripts.
Model species	a species used to study particular biological phenomena, the outcome offering insights into the workings of other species. Usually, the selection is based on experimental tractability, particularly ease of genetic manipulation. For the geneticist, it is an organism with inbred lines where sibs will be >98% identical (i.e. *Drosophila*, *Caenorhabditis elegans* and mice). For genomic science, it refers to a species for which the genomic DNA has been sequenced.
miRNA	a category of novel, very short, non-coding RNAs, generated by the cleavage of larger precursors (pri-miRNA). These short RNAs are included in the RNA-induced silencing complex (RISC) and pair to the 3′ ends of target RNA, blocking its translation into proteins (in animals) or promoting RNA cleavage and degradation (in plants).
mRNA	a protein-coding mRNA containing a protein-coding region (CDS), preceded by a 5′ and followed by a 3′ untranslated region (5′ UTR and 3′ UTR). The UTRs contain regulatory elements. A full-length cDNA contains the complete sequence of the original mRNA, including both UTRs. However, it is often difficult to assign the starting–termination positions for protein synthesis unambiguously. A cDNA containing the entire CDS is often considered acceptable for bioinformatic and experimental studies requiring full-length cDNAs.
ncRNA	non-coding RNA is any RNA molecule with no obvious protein-coding potential for at least 80 or 100 amino acids, as determined by scanning full-length cDNA sequences. It includes ribosomal (rRNA) and transfer RNAs (tRNA) and is now known to include various sub-classes of RNA, including snoRNA, siRNA and piRNA. Just like the coding mRNAs, a large proportion of ncRNAs are transcribed by RNA polymerase II and are large transcripts. A description of the many forms of ncRNA can be found at http://en.wikipedia.org/wiki/Non-coding_RNA.

Node	as in networks. Objects linked by edges to create a network.
PCR	polymerase chain reaction. A molecular biology technique for replicating DNA *in vitro*. The DNA is thus amplified, sometimes from very small amounts. PCR can be adapted to perform a wide variety of genetic manipulations.
piRNA	Piwi-interacting RNA. A class of RNA molecules (29–30 nt long) that complex with Piwi proteins (a class of the Argonaute family of proteins) and are involved in transcriptional gene silencing.
PMF	peptide mass fingerprinting. An analytical technique for protein identification in which a protein is fragmented using proteases. The resulting peptides are analysed by mass spectrometry and these masses compared against a database of predicted or measured masses to generate a protein identity.
Polyadenylation	the covalent addition of multiple A bases to the 3′ tail of an mRNA molecule. This occurs during the processing of transcripts to form the mature, spliced molecule and is important for regulation of turnover, trafficking and translation.
Post-source decay	in mass spectrometry. The fragmentation of precursor molecular ions as they accelerate away from the ionisation source of the mass spectrometer. All precursor ions leaving the ion source have approximately the same kinetic energy, but fragmentation results in smaller product ions that can be distinguished from precursor ions using a ‘reflectron’ by virtue of their lower kinetic energies.
Post-translational modification	the chemical modification of a protein after synthesis through translation. Some modifications, notably phosphorylation, affect the properties of the protein, offering a means of regulating function.
Principal component analysis (PCA)	a technique for simplifying complex, multi-dimensional datasets to a reduced number of dimensions, the principal components. This procedure retains those characteristics of the data that relate to its variance.
Promoter	a regulatory DNA sequence, generally lying upstream of an expressed gene, which in concert with other often distant regulatory elements directs the transcription of a given gene.
Proteome	the entire protein complement of an organism, tissue or cell culture at a given time.
Quantitative trait	inheritance of a phenotypic property or characteristic that varies continuously between extreme states and can be attributed to interactions between multiple genes and their environment.
qPCR	quantitative real-time PCR, sometimes called real-time PCR. A more quantitative form of RT-PCR in which the quantity of amplified product is estimated after each round of amplification.
QTL	quantitative trait loci. A region of DNA that contains those genes contributing to the trait under study.
RISC	RNA-induced silencing complex. A protein complex that mediates the double-stranded RNA-induced destruction of homologous mRNA.
RNAi	RNA interference or RNA-mediated interference. The process by which double-stranded RNA triggers the destruction of homologous mRNA in eukaryotic cells by the RISC.
RT-PCR	reverse transcription–polymerase chain reaction. A technique for amplifying a defined piece of RNA that has been converted to its complementary DNA form by the enzyme reverse transcriptase. See qPCR.
siRNA	small interfering RNA, or silencing RNA. A class of short (20–25 nt), double-stranded RNA molecules. It is involved in the RNA interference pathway, which alters RNA stability and thus affects RNA concentration and thereby suppresses the normal expression of specific genes. Widely used in biomedical research to ablate specific genes.

snoRNA	small nucleolar RNA. A sub-class of RNA molecules involved in guiding chemical modification of ribosomal RNA and other RNA genes as part of the regulation of gene expression.
SNP	single nucleotide polymorphism. A single base-pair mutation at a specific locus, usually consisting of two alleles. Because SNPs are conserved over evolution, they are frequently used in QTL analysis and in association studies in place of microsatellites, and in genetic fingerprinting analyses.
SSH	suppressive subtractive hybridisation. A powerful protocol for enriching cDNA libraries for genes that differ in representation between two or more conditions. It combines normalisation and subtraction in a single procedure and allows the detection of low-abundance, differentially expressed transcripts, such as those involved in signalling and signal transduction.
Structural RNAs	a class of non-coding RNA, long known to have a structural role (for instance, the ribosomal RNAs), transcribed by RNA polymerase I or III.
Systems biology	treatment of biological entities as systems composed of defined elements interacting in defined ways to enable the observed function and behaviour of that system. The properties of the systems are embedded in a quantitative model that guides further tests of systems behaviour.
TATA-boxes	sequences in promoter regions constituted by TATAAA, or similar variants, which were considered the hallmark of Promoters. Recent data show that they are present only in the minority of promoters, where they direct transcription at a single well-defined location some 30 bp downstream of this element.
trans-acting	a factor or gene that acts on another unlinked gene, a gene on a separate chromosome or genetically unlinked usually through some diffusible protein product (for mRNA expression, typically a transcription factor).
Transcript	an RNA product produced by the action of RNA polymerase reading the sequence of bases in the genomic DNA. Originally limited to protein-coding sequences with flanking UTRs but now known to include large numbers of products that do not code for a protein product.
Transcriptome	the full set of mRNA molecules (transcripts) produced by the system under observation. Whilst the genome is fixed for a given organism, the transcriptome varies with context (i.e. tissue source, ontogeny, external conditions or experimental treatment).
Transgene	a gene or genetic material that has been transferred between species or between organisms using one of several genetic engineering techniques.
Transinduction	generation of transcripts from intergenic regions. At least some such products do not relate to a definable promoter or transcriptional start site.
Transposon	sequences of DNA able to move to new positions within the genome of a single cell. This event might cause mutation at the site of insertion. Also called ‘mobile genetic elements’ or ‘jumping genes’.
Transvection	an epigenetic phenomenon arising from the interaction between one allele and the corresponding allele on the homologous chromosome, leading to gene regulation.
TUs	transcriptional units. Used to group all of the overlapping RNA transcripts that are transcribed from the same genomic strand and share exonic sequences.
UTR	untranslated region. Regions of the mRNA that lie at either the 3′ or 5′ flanking ends of the molecule (i.e. 3′ UTR and 5′ UTR). They bracket the protein-coding region and contain signals and binding sites that are important for the regulation of both protein translation and RNA degradation.