Application of 'next-generation' sequencing technologies to microbial genetics

Daniel MacLean, Jonathan D. G. Jones and David J. Studholme

Abstract | New sequencing methods generate data that can allow the assembly of microbial genome sequences in days. With such revolutionary advances in technology come new challenges in methodologies and informatics. In this article, we review the capabilities of high-throughput sequencing technologies and discuss the many options for getting useful information from the data.

De novo assembly
Construction of longer sequences, such as contigs or genomes, from shorter sequences, such as sequence reads, without prior knowledge of the order of the reads or reference to a closely related sequence.

New sequencing technologies that enable the acquisition of gigabases of sequence information in just a few days bring with them new problems that need to be overcome[1,2]. Sequencing technologies have applications in genome sequencing, metagenomics, epigenetics and discovery of non-coding RNAs and protein-binding sites (reviewed in REFS 3–14). The plethora of options for generating and dealing with the data might appear bewildering, but there are essentially two types of problems: alignment problems (for which a previously sequenced reference sequence is available) and de novo assembly problems (for which no reference sequence is available). Each of the technologies strikes a different balance between cost, read length, data volume and rate of data generation. Consequently, the choice of bioinformatics solution depends on the biological application and on the type of sequencing technology used to generate the data. The first wave of next-generation sequencing technologies aimed to resequence genomes in a shorter time and at a lower cost than traditional Sanger sequencing. The Solexa GA platform and 454 GS20 pyrosequencing (see Further information), which were developed by Illumina and Roche, respectively, generated reads of around 36 and 100 nucleotides, respectively. These short reads could be adequate for resequencing applications, but it was widely assumed that they would be too short for *de novo* assembly. Since the introduction of these technologies, the ambition and scope of applications have increased enormously, culminating in large-scale metagenomic and evolutionary analysis of tens of species simultaneously. In this Review article, we briefly introduce the key technologies, survey their applications in microbiology and discuss the bioinformatics options for successfully meeting both alignment and *de novo* sequencing challenges. We focus mainly on the technologies developed by Roche and Illumina, as they were the first to become widely available and so dominate the current literature. However, the field is moving fast and competing platforms are beginning to emerge.

The Sainsbury Laboratory, Norwich NR4 7UH, United Kingdom.
Correspondence to D.J.S.
e-mail: david.studholme@tsl.ac.uk
doi:10.1038/nrmicro2088
Published online
23 February 2009

High-throughput sequencing technologies

Technological aspects of sequencing technologies have been extensively reviewed elsewhere[3,4,8–10,11,13], so are only briefly summarized here.

The first next-generation high-throughput sequencing technology, the 454 GS20 pyrosequencing platform, which was developed by Roche, became available in 2005. The GS20 platform has now been replaced by the GS FLX platform. Illumina released Solexa GA in early 2007, and more recently, SOLiD and Heliscope were released by Applied Biosystems and Helicos, respectively, whereas Pacific Biosciences is currently developing a single-molecule sequencer, which is expected to be released in 2010 (see Further information). These methods all have different underlying biochemistries (FIG. 1). With the exception of Heliscope and the method developed by Pacific Biosciences, both of which sequence single molecules, all of these technologies sequence populations of amplified template-DNA molecules. The 454 GS FLX and SOLiD methods begin by ligating oligonucleotide adaptors to the DNA and immobilizing the ligation products onto beads[15]. The beads are placed into a water–oil emulsion and DNA is amplified by PCR. In 454 GS FLX pyrosequencing, the beads that carry the DNA template are then isolated in specially synthesized fibre-optic picolitre volume wells. A complementary strand is synthesized by the sequential addition of one species of dNTP and DNA polymerase in the presence of a chemiluminescent enzyme, such as luciferase. Incorporation of a nucleotide into the complementary strand releases pyrophosphate,

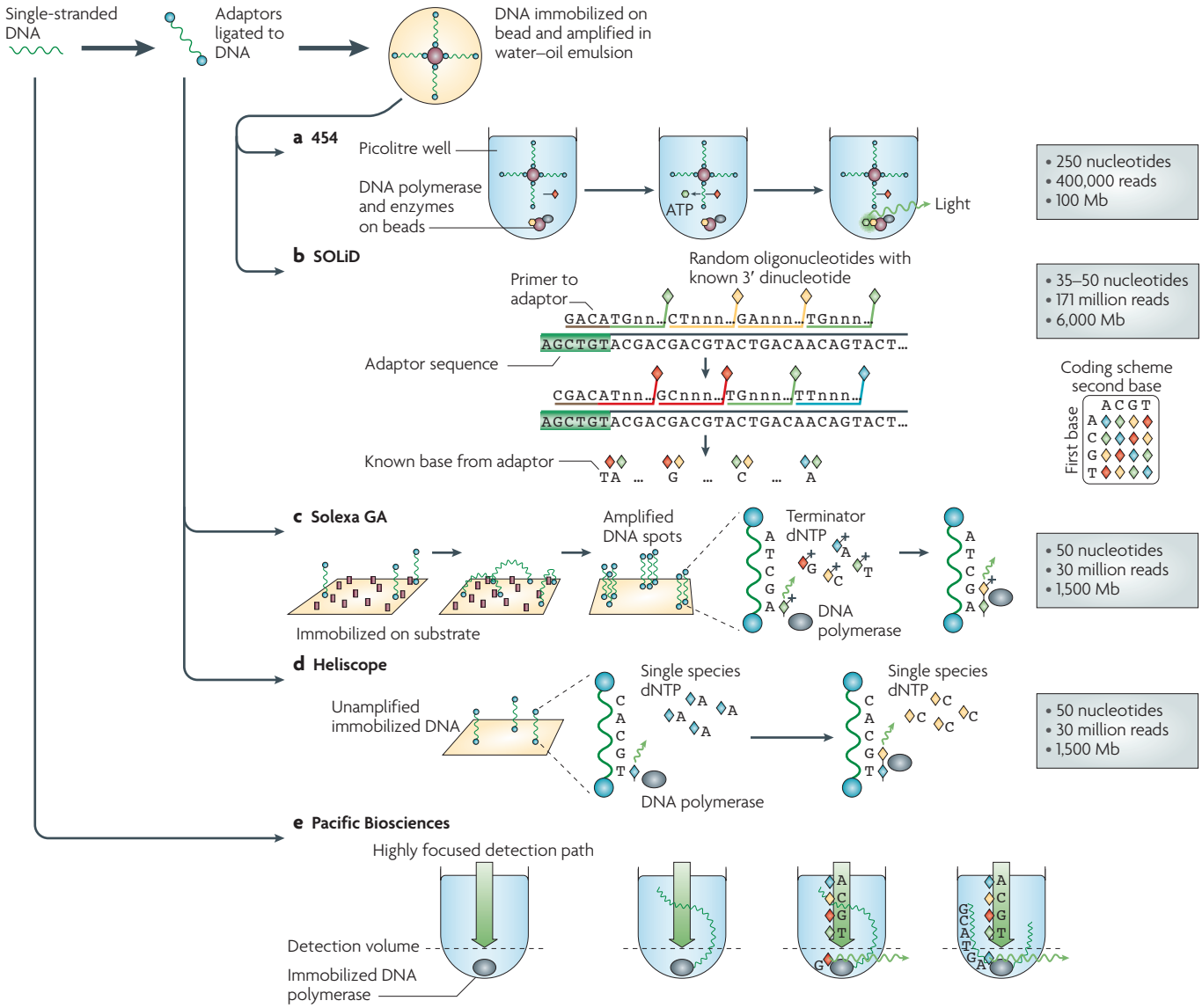


Figure 1 | **High-throughput sequencing technologies.** The read length, the number of reads and the total amount of sequence generated in a typical run are indicated for each of the new-generation high-throughput sequencing technologies. **a** | 454 GS FLX pyrosequencing. Oligonucleotide adaptors are ligated to fragmented DNA and immobilized to the surface of microscopic beads before PCR amplification in an oil–droplet emulsion. Beads are isolated in picolitre wells and incubated with dNTPs, DNA polymerase and beads bearing enzymes for the chemiluminescent reaction. Incorporation of a nucleotide into the complementary strand releases pyrophosphate, which is used to produce ATP. This, in turn, provides the energy for the generation of light. The light emitted is recorded as an image for analysis. **b** | SOLiD sample preparation is similar to that of 454 pyrosequencing. After amplification, the beads are immobilized onto a custom substrate. A primer that is complementary to the adaptor sequence (green), random oligonucleotides with known 3′ dinucleotides and a corresponding fluorophore are hybridized sequentially along the sequence and image data collected. After five repeats, the complementary strand is melted away and a new primer is added to the adaptor sequence, ending at a position one nucleotide upstream of the previous primer. Second-strand synthesis is repeated, allowing two-colour encoding and double reading of each of the target nucleotides. Repeats of these cycles ensure that nucleotides in the gap between known dinucleotides are read. Knowledge of the first base in the adaptor reveals the dinucleotide using the colour-space scheme. In other words, knowing that the last adaptor nucleotide is T and the colour is red means that the first base to be sequenced must be A. Knowing that the first base is A and the colour is green means that the next base must be C and so on. **c** | For Solexa GA sequencing, adaptors are ligated onto DNA and used to anchor the fragments to a prepared substrate. Fold-back PCR results in isolated spots of DNA of a large enough quantity that the amassed fluorophore can be detected. Terminator nucleotides and DNA polymerase are then used to create complementary-strand DNA. Images are collected at the end of each cycle before the terminator is removed. **d** | Heliscope sequencing immobilizes unamplified DNA with ligated adaptors to a substrate. Each species of dNTP with a bright fluorophore attached is used sequentially to create second-strand DNA; a 'virtual terminator' prevents the inclusion of more than one nucleotide per strand and cycle, and background signal is reduced by removal of 'used' fluorophore at the start of each cycle. **e** | The new sequencing method developed by Pacific Biosciences occurs in zeptolitre wells that contain an immobilized DNA polymerase. DNA and dNTPs are added for synthesis. Fluorophores are cleaved from the complementary strand as it grows and diffuse away, allowing single nucleotides to be read. Continuous detection of fluorescence in the detection volume and high dNTP concentration allow extremely fast and long reading.

which can be used to make ATP to provide energy for the chemiluminescent enzyme reaction. The light produced is proportional to ATP availability. The light is recorded and analysed computationally, eventually generating reads with a mean length of up to 400 nucleotides. Unlike the other methods, in 454 GS FLX sequencing, a lack of reversible terminator nucleotides means that more than one nucleotide can be incorporated into the complementary strand in one cycle. Consequently, this method has some problems resolving homopolymeric stretches of sequence, as runs of *n* nucleotide are occasionally read as *n*-1 nucleotides. In the SOLiD method, amplified DNA on beads is deposited onto a glass slide and then subjected to sequential hybridization with short random oligonucleotides containing known 3′ dinucleotides and a corresponding fluorophore. After five cycles, the synthesized DNA is melted away from the template, and the process is repeated. During this second cycle, synthesis starts at the nucleotide on the template immediately upstream of where synthesis began in the previous cycle. Repeated cycles allow for multi-colour-encoding of each base in the DNA sequence, thereby reducing errors in the sequence and allowing for detection of complicated genomic variation, such as single-nucleotide polymorphisms (SNPs). SOLiD can generate reads of 35–50 nucleotides. The Solexa GA method begins by ligating adaptors to the target DNA, then attaching the ligated product at one end to a microfluidics chip. Amplification is achieved by fold-back PCR *in situ* and sequencing is carried out by incorporating fluorescently labelled, reversible terminator nucleotides into the strand with a DNA polymerase. Imaging is carried out at the end of the synthesis cycle, after which the terminator is removed and a new cycle begins. Typically, reads of 36 nucleotides are generated, although read lengths of 70 nucleotides are expected to be officially supported by the end of 2008 and some laboratories are already using customized protocols to generate reads of up to 100 nucleotides.

Solexa GA, SOLiD and Heliscope use an amplification step, primarily to increase levels of signal compared with background noise. However, amplification inevitably introduces bias, such that the distribution of sequence reads on the template sequence is neither uniform nor random. This results in 'hot spots' and 'cold spots' of artificially deep or shallow coverage, respectively. Furthermore, if individual molecules fall out of synchronization ('out of phase'), simultaneous stepwise sequencing of a population of molecules leads to inaccuracy. The newest technologies, Heliscope and the new method developed by Pacific Biosciences, sequence single molecules. Single-molecule approaches promise to deliver consistently low error rates by avoiding amplification-associated bias, intensity averaging and phasing or synchronization problems[16,17]. Using the Heliscope method, oligonucleotide fragments of 100–200 bases are first attached along a proprietary substrate within a microfluidics flow cell. Nucleotides that bear a bright fluorophore (which increases detectability and removes the need for amplification of the template DNA) are introduced one species at a time and incorporated by DNA polymerase to the growing complementary strand. Images are then recorded and analysed to identify which nucleotide was incorporated into which growing strand, before the cycle is repeated with a different species of nucleotide. Currently, reads have a final length of 35 nucleotides. The method developed by Pacific Biosciences is the most revolutionary: it uses single molecules of DNA polymerase immobilized in zeptolitre (10^{-21} litre) wells to incorporate labelled nucleotides into a complementary strand from a single-stranded template. The wells, only nanometres in diameter and depth, allow for high concentrations of nucleotides, resulting in fast, reagent-efficient synthesis and a highly focused detection volume in the well. Imaging is carried out continuously; the detection volume, which contains the DNA polymerase, is monitored and each newly incorporated base can be read in a short time. The use of phospho-linked, rather than base-labelled, nucleotides means that the preceding fluorophore is removed after incorporation of the next base and diffuses out of the detection volume, resulting in low background noise and high accuracy. Pacific Biosciences claim that this system can generate reads that are thousands of nucleotides long, at the expense of overall throughput. This technology is also expected to suffer from insertion and deletion errors caused by the action of the DNA polymerase itself; continuous reading means that if the polymerase selects a nucleotide, but does not incorporate it, an extra base can be read. Incorporation of a nucleotide can also occur too quickly for the machine to read, resulting in deletion errors.

All the technologies discussed above have limits on the lengths of the reads that can be generated. Common limiting factors include the degrading effects of lasers on DNA and enzymes and the effects of cyclical washes, which slowly reduce the amount of DNA sequenced. The phasing problem limits read length too. This problem occurs because some of the molecules are not incorporated during one or more of the cycles. These errors build up cumulatively through the cycles. All the technology providers mentioned above are constantly improving protocols and reagent mixtures to increase read lengths.

Applications for microbiology

Each of the available sequencing technologies has its particular strengths and weaknesses. FIGURE 2 summarizes the factors that need to be considered when choosing the appropriate platform. Currently, the Genomes OnLine Database (see Further information) lists 921 completed genome-sequence projects most of which are microbial. However, each of these genome sequences provides only a glimpse of the real genetic make-up of microbial populations. In many bacterial species, as well as the 'core' set of genes found in all individuals, there are large numbers of 'dispensable' genes that can vary enormously between closely related species, pathovars and strains[18]. Work on model organisms often uses multiple laboratory strains, with different phenotypes, implying that there are underlying genetic differences. Therefore, understanding the genetic basis for pathovar- and strain-specific differences in pathogenicity, ecology and other phenotypic characteristics requires additional sequence data.

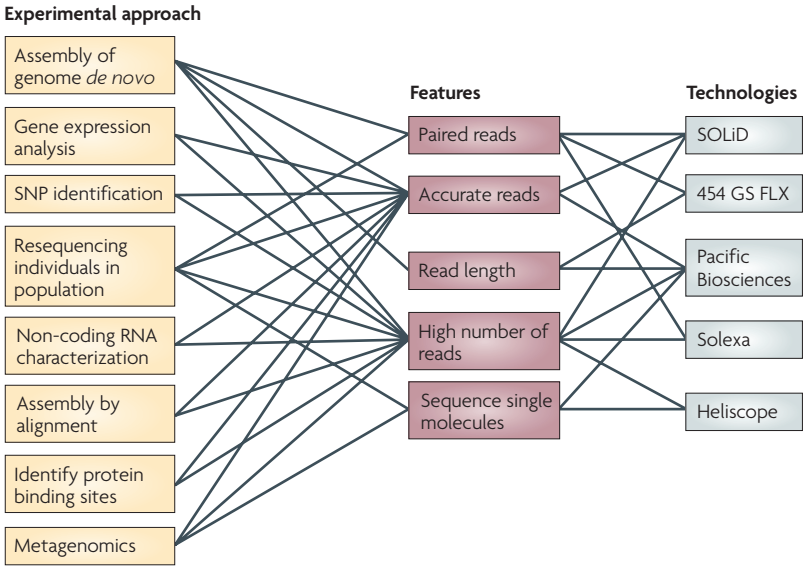


Figure 2 | **Selecting a technology for an experiment.** The yellow boxes represent different experimental approaches. The lines from each experimental approach lead to a 'wish list' of features (brown boxes) that are especially useful for the successful execution of an experiment or analysis. The features are then linked to the technologies that provide them. SNP, single-nucleotide polymorphism.

One obvious approach is to determine the complete genome sequence of additional strains. Until recently, such sequencing was almost entirely restricted to large genome centres, as it was not economically or logistically feasible for individual laboratories to determine the complete sequence of even a bacterial genome. However, this changed with the introduction of the 454 GS20 sequencer in 2006. The 'socially motile' soil bacterium *Myxococcus xanthus* was one of the first microorganisms to be subjected to genome sequencing using 454 GS20 technology[19]. Approximately 2.5 million reads (with an average length of around 100 nucleotides) were generated from 12 sequencing runs, and were assembled *de novo* into 104 contigs, leaving 104 gaps in the genome sequence. Of these gaps, 88 were filled by capillary sequencing of PCR products, leaving just 16 repeat-containing regions unassembled. The 454 GS FLX platform allowed the generation of read lengths of around 250 nucleotides, which, together with the recent introduction of paired-end reads, allowed bacterial genome sequencing to become even more tractable. More recently, read lengths have even reached around 400 nucleotides, using Roche's new titanium reagents.

'Finishing' a genome-sequencing project is a costly and time-consuming process that consists of gap closure in the assembly, and a series of rigorous quality-control steps and resolution of errors[20]. For many applications, a draft genome-sequence assembly is sufficient and there is no need to invest in finishing. Stiens and colleagues[21] used a draft genome assembly based on 454 sequencing to identify several unexpected genomic differences between the previously sequenced laboratory strain of *Sinorhizobium meliloti* RM102 and a related field isolate, SM11. Following a sudden epidemic or bioterrorist attack, draft sequencing enables the identification of genetic differences that underlie particular traits, such as antibiotic resistance or increased virulence[22]. For a clinical strain of *Francisella tularensis* subsp. *holartica*, DNA extraction to sequencing using the 454 GS20 platform took just 12 days[22]. In our laboratory, we have recently generated draft genomes from several plant pathogenic bacteria using the Solexa GA platform.

Most microbial species cannot be cultured, and even within a species there can be huge variations in genotype (and consequently phenotype) owing to genetic plasticity[18]. Therefore, sampling signature genes, such as 16S ribosomal RNA, does not give much insight into the metabolic activities of a microbial community. This problem has recently been addressed by the emerging field of metagenomics. Metagenomics is a 'brute force' approach[6], whereby total DNA from a microbial and/or viral population is sequenced and compared with all previously sequenced genes. The high-throughput capability offered by next-generation sequencing methods makes them attractive for such an approach[12]. Furthermore, traditional approaches to metagenomic sequencing have required an initial cloning step. This was problematic, because many cloned sequences are not stably maintained in the host (typically *Escherichia coli*), leading to biases in the repertoire of sequences. Recent metagenomics studies have exploited 454 pyrosequencing, which provides longer sequence reads than the alternatives. A notable example of such a study is a survey of microbial populations from nine distinct environments, including underground mine waters, marine and freshwater, coral, fish, terrestrial animals and mosquitoes[23]. As paired reads are now available for 454 GS FLX, Solexa and SOLiD platforms, and read lengths are increasing for the higher-throughput technologies, such as Solexa, it is likely that they will continue to play a part in future metagenomics projects.

The *de novo* assembly of sequence reads is not always necessary when comparing closely related strains; cataloguing polymorphisms relative to a reference genome sequence is often a satisfactory goal. The main goal of resequencing projects is generally to identify SNPs and other types of polymorphism, such as short insertions and deletions (collectively called indels). SNP discovery is essential for genetic mapping in eukaryotic model organisms that have large, complex genomes. An important international initiative is underway to resequence 1,001 naturally occurring strains (accessions) of the model plant *Arabidopsis thaliana*[24] using Solexa sequencing (see Further information for a link to the 1001 Genomes Project). For most of the 119 Mb *A. thaliana* genome, accessions can be compared by aligning sequence reads against a reference genome. This has already yielded more than 800,000 SNPs and nearly 80,000 short indels between 3 accessions of *A. thaliana*. However, several megabases of genome sequence are sufficiently diverged between accessions that alignment is not possible. For these regions, *de novo* sequence assembly is required[24].

In principle, similar approaches could be useful for the sequencing of eukaryotic microorganisms. Tag-based approaches can improve the efficiency of SNP detection without the need to fully sequence the entire genome

Contig
A fragment of genome sequence derived by assembling shorter sequence reads into larger constructs on the basis of overlap between the sequence reads.

Paired-end read
A sequence read known to come from a genomic region within a limited number of nucleotides of another. The extra information puts constraints on how far apart the reads can be placed during assembly or alignment, allowing more accurate placement and construction of contigs.

to saturation. Sequence polymorphisms often result in the disruption of restriction endonuclease sites, enabling the detection of SNPs by microarray-based comparison of restriction sites between strains or individuals. Baird and colleagues[25] extended this approach by sequencing restriction site-associated sequence tags using Solexa. Using this method, they successfully mapped a mutant that was methylation deficient to three genetic loci in the ascomycete fungus *Neurospora crassa* significantly faster than possible using classical methods of genetic mapping.

Although there is less interest in the use of SNPs as markers in genetic mapping in prokaryotes than in eukaryotes, exploration of genetic diversity between closely related bacteria could be important for other reasons. Holt and colleagues[26] used 454 GS FLX and Solexa sequencing to explore genetic diversity between isolates of *Salmonella enterica* subsp. *enteric* serovar Typhi, the causative agent of typhoid fever. This species is monomorphic, so it was necessary to resequence the entire genome to maximize the chances of finding isolate-specific genetic differences. The authors identified 1,787 SNPs, which allowed them to infer the recent evolutionary history of the species with high resolution. They then found evidence of weak stabilizing selection by measuring the ratio of synonymous and non-synonymous SNPs at the polymorphic loci. Relative rates of synonymous and non-synonymous mutations are informative about selective pressures, as non-synonymous mutations are subject to any selective pressures, whereas synonymous mutation rates are assumed to be independent of selection, reflecting underlying mutation rates. A rate of non-synonymous mutation that is significantly higher than the rate of synonymous mutation indicates positive selection. Conversely, a lower rate of non-synonymous mutation can indicate purifying selection.

Although apparently not important in *S.* Typhi, in many other pathogens, positive selection and diversifying selection can be a signature of genes encoding products that directly interact with the host organism[27]. For example, phytopathogenic oomycetes, such as *Hyaloperonospora arabidopsidis* and *Phytophthora infestans*, encode arsenals of so-called effectors. These secreted proteins are delivered directly into the host apoplast or cytoplasm, where they can subvert defences of the plant cell. The draft genome sequences of *Phytophthora* species contain hundreds of candidate genes with the potential to encode effectors[28]; the challenge is to prioritize and select candidates for detailed study. One approach is to sequence multiple strains and identify candidate genes in which there is allelic diversity in natural populations and non-synonymous nucleotide substitution rates are higher than the background rate of synonymous substitutions. This approach complements other selection criteria, such as secretion and induced expression in the host[27].

Besides genetic diversity in natural populations, there may be important genetic differences between isolates of a single laboratory strain. Until recently, the isogenicity of bacterial strains had not been examined. Using the Solexa GA platform, Srivatsan and colleagues[29] were able to detect several polymorphisms between different isolates of the same strain of *Bacillus subtilis* (up to 31 nucleotides per genome). Similar approaches have been taken to investigate the genetic basis of extensive drug resistance in *Mycobacterium tuberculosis* and may in the future become routine for tracking bacterial epidemiology and diagnosis[30].

The ability to detect genetic differences between highly similar genomes at such high resolution enables rapid mapping of mutations. A *B. subtilis* JH642 *relA*-deficient strain displays a severe growth defect phenotype, and it is possible to isolate spontaneous mutants that suppress this phenotype. When the complete genomes of two such spontaneous mutants were sequenced, Srivatsan and colleagues successfully identified two separate mutated sites that were responsible for the suppressor phenotype[29]. Directly sequencing the mutant genome (using Solexa) is less laborious and costly than genetic mapping, especially for polygenic phenotypes. In principle, the direct sequencing approach could also be applied to larger, eukaryotic genomes.

Sequencing of expressed sequence tags (ESTs) is a robust means of gene discovery and identification of transcripts involved in specific biological processes and can yield quantitative information about transcript abundance. However, sequencing tens or hundreds of thousands of ESTs using traditional capillary-based methods is expensive. SAGE (serial analysis of gene expression) and related methods represent a refinement of EST sequencing that reduces cost per transcript by reducing the amount of sequence collected for each transcript[31]. For quantitative transcript profiling, microarrays have been a viable alternative. The long reads and moderately high throughput of 454 sequencing makes it an attractive option for sequencing ESTs; for example, a recent study used a combination of 454 GS FLX and capillary sequencing to discover 4,689 unique candidate transcripts from the phytopathogenic oomycete *Pythium ultimum*[32].

Several recent papers describe the application of Solexa GA and SOLiD technologies to deeply sequence the transcriptomes of mammals[33,34], yeast[35] and plants[36]. These two platforms generate sequence data at a faster rate and a cheaper per-nucleotide cost than either traditional Sanger capillary methods or 454 GS FLX; it is now feasible to produce tens of millions of independent EST sequence reads (rather than tens of thousands). The Solexa GA and SOLiD methods, which are coming to be known as RNA–Seq[37], can be used quantitatively over five orders of magnitude and are not dependent on previous knowledge of transcribed sequences. RNA–Seq seems to be a serious contender that could replace microarrays for quantitative transcript profiling in eukaryotes[38].

Several recent publications[7,14] use the term 'census methods' for quantitative transcription profiling by RNA–Seq. This term describes the principle of directly sequencing complex mixtures of nucleic acids to determine their content without the need for cloning or hybridizing to arrays. Unlike microarray-based methods, the output from census-based methods is digital

and varies linearly over several orders of magnitude, and therefore is not limited by the dynamic range of a plate reader. Other applications of census methods include the discovery of protein-binding sites, profiling and the development of short RNA (sRNA) and epigenetic studies.

Chromatin immunoprecipitation using microarrays (ChIP-on-chip) is a high-throughput method for discovering protein-binding sites in DNA on a whole-genome scale[39]. Even though ChIP-on-chip involves hybridization of immunoprecipitated protein-bound DNA onto a microarray, it is now possible to directly determine the bound DNA sequence using next-generation sequencing technologies with all the usual advantages of census-based methods[14]. This approach, dubbed ChIP–Seq, has already been used to discover transcription factor binding sites in humans[40] and doubtless will soon be applied to microorganisms.

Next-generation sequencing methods are already making an impact in epigenetics. Epigenetic phenomena include methylation and other covalent modifications of DNA and chromatin proteins. Several studies successfully combined high-throughput sequencing with bisulphite treatment of DNA to identify methylated regions of DNA in, for example, human[41] and plant[36,42] cells, and ChIP–Seq is beginning to be used for genome-scale studies of histone modifications[39,43]. The pioneering studies have focused on higher eukaryotes; however, these methods should be equally applicable to eukaryotic microorganisms, such as apicomplexans, for which complete genome sequences are known and which may be unusually reliant on epigenetic mechanisms[44]. Epigenetics is also of relevance in host–pathogen interactions; for example, HIV preferentially integrates close to specific histone modification sites in the host genome[45].

Micro-RNAs (miRNAs) and other short RNA species that have lengths of approximately 21–24 nucleotides have attracted enormous interest in the past few years and are implicated in post-transcriptional regulation of numerous biological processes, especially development. The availability of high-throughput sequencing has led to the discovery of thousands of new species of non-coding sRNAs, especially in plants. The short read length of these sequencing platforms is not a limitation for sRNA discovery. Until recently, it was widely assumed that miRNAs were of importance only in complex multicellular organisms. However, deep 454 sequencing recently revealed that the unicellular green alga *Chlamydomonas reinhardtii* expresses several miRNAs, suggesting that miRNA-mediated regulation may be ancient and evolved in primitive, unicellular organisms[46].

I got my sequence data — now what?

Some applications, such as those that detect sRNAs using Solexa GA, yield sequence reads that still contain adaptor-derived sequences. Furthermore, some protocols include 'barcode' sequences in the adaptor that can be used to identify the source of the sequence read in multiplex samples. Once any adaptor and barcode sequences have been removed, and poor-quality sequence reads have been eliminated, the next step in the bioinformatics analysis is usually either *de novo* sequence assembly or alignment against a reference genome sequence. Comprehensive lists of relevant software can be found on the SEQanswers website (see Further information) and in REF. 11.

Software for* de novo *sequence assembly. The goal of *de novo* sequence assembly is to reconstruct contigs from a set of partially overlapping sequence reads in the absence of any other reference sequence. Ideally, a contig should span an entire chromosome, but in practice, most 'completely sequenced' eukaryotic genomes contain regions that cannot be assembled. There are several well-established algorithms and software packages for assembly of capillary sequence reads. However, these are not suitable for assembling reads generated by the new technologies. First, the distributions of sequencing errors are different, making the existing error models inappropriate. Second, these algorithms rely on long overlaps between sequence reads; such long overlaps do not exist in short sequence reads. Finally, the algorithms do not always scale well and are unable to deal with hundreds of thousands or millions of reads. Fortunately, a new cohort of assembly algorithms have been developed during the past 2 years that perform remarkably well even with reads as short as 36 nucleotides.

All of these assembly algorithms use the mathematical concept of a graph: a set of vertices or nodes that can be connected by edges. In the SHARCGS[47], SSAKE[48] and VCAKE[49] algorithms, as well as Roche's proprietary 454 assembly software, Newbler, each vertex in the graph is an observed sequence read, whereas the edges represent overlaps between read sequences. Each edge can be ascribed a weight, which essentially represents the abundance of that sequence read in the input data. Ideally, there should be a single linear path through the graph that represents the 'true' assembled sequence. In practice, the graph usually contains ambiguities, such as loops and branches, which result from repeat sequences and sequencing errors. A refinement of this graph-based approach is the de Bruijn graph, in which the edges are not whole-sequence reads, but are k-mers (also called l-tuples in some publications). In a de Bruijn graph, each edge is a k-mer that has been observed in the input data and implicitly represents a series of overlapping k-mers that overlap by a length of k–1 (BOX 1). This approach is more computationally efficient than the simple read-based graph approach when dealing with large numbers of reads. The de Bruijn paradigm was previously implemented in the EULER assembler[50] to assemble capillary sequence reads. However, it has been most useful for the assembly of short reads, and has been used in Velvet[51], Euler–SR[52], Edena[53] and ALLPATHS[54]. The algorithms differ mainly in their strategies for dealing with sequencing errors and resolving ambiguities.

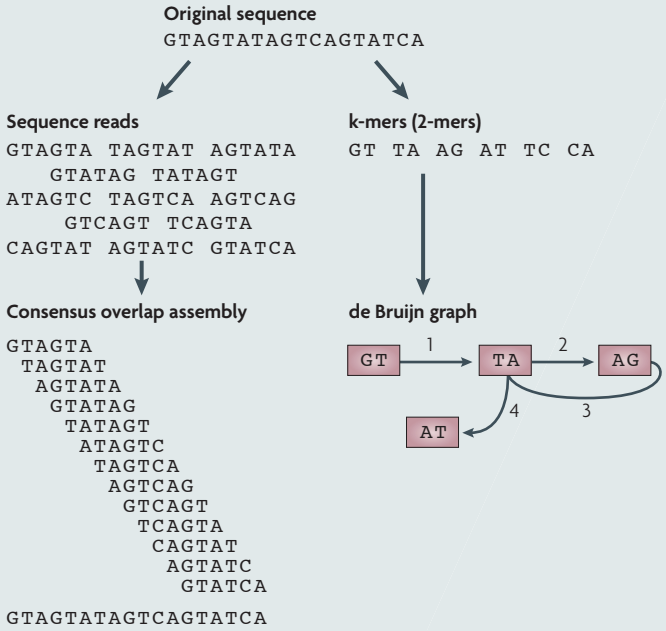
Theoretical and empirical results are beginning to indicate that the use of paired reads can overcome the limitations of short sequence reads for assembly[54]. The Velvet and SSAKE assembly programs can use paired

Epigenetics
The study of inherited changes in gene function that cannot be explained by changes in DNA sequence.

de Bruijn graph
In mathematics, a network structure is properly called a graph. The entities that are connected are called nodes and the connections are called edges. A de Bruijn graph is a graph in which the nodes are sets of symbols (similarly to the nucleotides in a sequence read) and the edges represent overlaps between the symbols. This is a convenient way to represent data, such as overlapping sequence reads.

k-mer
A piece of nucleotide sequence of length k. A k-mer is usually used to indicate a computationally selected subsequence of an experimentally derived sequence, such as a read or a genome.

Box 1 | **Overlap consensus assembly and de Bruijn graph assembly**



Whole genome shotgun assembly (WGSA) is the cornerstone of all genomics. WGSA has traditionally been carried out by fragmenting genomic DNA and cloning the fragments into vectors, such as BACs (bacterial artificial chromosomes). These vectors are then sequenced using the Sanger method and the resulting sequence is assembled computationally. Many programs have been developed to carry out WGSA, including ARACHNE, Phusion, PCAP, Atlas, TIGR and Phrap, which use the overlap consensus method, and Euler, which uses the de Bruijn graph method.

The overlap consensus assembly method uses the overlap between sequence reads to create a link between them. The contig is eventually formed by reading along the links as far as possible. Of course, multiple reads, errors, repeats and other ambiguities mean that there are multiple forks in the path through the links, and it is the differences in the ways of navigating through these ambiguities that differentiates the different methods. With short reads, overlap consensus assembly suffers from two main problems. First, the short read length means the overlaps must be calculated over a large proportion of the read to retain accuracy, and second, the huge number of reads increases the number of links, so that the contig path is difficult to compute. Some short-read assemblers that are based on this method include SSAKE, VCAKE, SHARCGS and Roche's proprietary 454 assembler, Newbler.

The de Bruijn graph approach circumvents the problems of overlap consensus assembly. Rather than using the reads 'as is' and trying to link them, the k-mers (all subsequences of length *k* within the reads) are computed and the reads are represented as a path through the k-mers. Such a paradigm handles redundancy better than the overlap consensus approach and makes the computation of paths more tractable.

In consensus overlap assembly, overlaps between reads are used to create a consensus sequence. In a de Bruijn graph, all reads are broken into k-mers and a path between the k-mers is calculated. This can be envisaged in the figure by starting at node GT and following the path that corresponds to the step number. From GT, move to TA down the path labelled 1, from TA move to AG down the path labelled 2 and so on. Assemblers that use this approach have been released recently, including a modification of the long-read assembler Euler, ALLPATHS and Edina. It is likely that the de Bruijn graph assemblers will improve vastly. Furthermore, the use of paired-read technologies will be better handled by de Bruijn graph assemblers, thereby creating potential for vast improvement in assemblies and alignments. Reports from primary papers that describe these methods support this view and indicate that the de Bruijn graph assemblers put together bigger contigs and can handle sequencing errors and complex genomes better than their counterparts.

reads as input and exploit their additional information content to increase the average length of the resulting contigs. Butler and colleagues have proposed an algorithm called ALLPATHS[54], which can assemble simulated paired 30-nucleotide reads at 80 times deep genome coverage from bacterial genomic reads into single contigs that represent the entire chromosome and provide high-quality assemblies of genomes up to 40 Mb in length.

Because these assembly methods have been developed only recently, it would be unwise to trust their output blindly until they have been validated using real experimental sequence data and the resulting assemblies have been compared with positive controls. Even in the absence of a reference sequence, it is possible to detect some types of misassembly, and a freely available software package (amosvalidate) has been developed to highlight errors in eukaryotic and prokaryotic genome assemblies, both finished and draft[55]. It would be advisable to submit assemblies to such examination before making inferences about genomic structure or gene order. An assembly-viewing tool, such as EagleView[56], can aid inspection and quality checking of candidate assemblies.

An additional complication is that the performance of each of the assembly methods discussed above can be sensitive to input parameter values, as well as the quality and quantity of the input data. It is usually impossible to predict the optimal parameter values *ab initio* and instead they must be determined empirically. In our laboratory, we tested each assembly method on a dataset of nucleotide reads generated using the Solexa 36 platform that represented simulated paired 30-nucleotide reads at 40 times deep genome coverage of a previously sequenced 6 Mb bacterial genome using a wide range of parameter values and both paired reads and unpaired assemblies[57]. We assessed the resulting assemblies for length and accuracy of each contig. The 'best' results were obtained using Velvet, which assembled 96% of the genome with an error rate of 0.33% per nucleotide. Half of the genome was covered by contigs of at least 8 kb. The 4% of the genome that was absent from the assembly was enriched for repeated sequences and non-coding RNA genes. By using paired reads, N_{50} could be increased to 165 kb. Encouragingly, more than 93% of the annotated protein-coding genes recovered from these assemblies were intact, full length and had no errors.

The sequencing field is moving fast and the assembly software packages are under active development, so are likely to improve further still. For example, the algorithm for exploiting paired-read information was substantially revised for Velvet[51] version 0.7 compared with version 0.6, yielding a tenfold increase in N_{50} contig length[57]. Our preliminary data from Solexa GA sequencing of genomic DNA from the phytopathogenic oomycete *Albugo candida* suggest that it might even be possible to construct a draft assembly of the gene complement of a eukaryotic genome that is more than 40 Mb in length (E. Kemen, D.J.S. and J.D.G.J., unpublished observations).

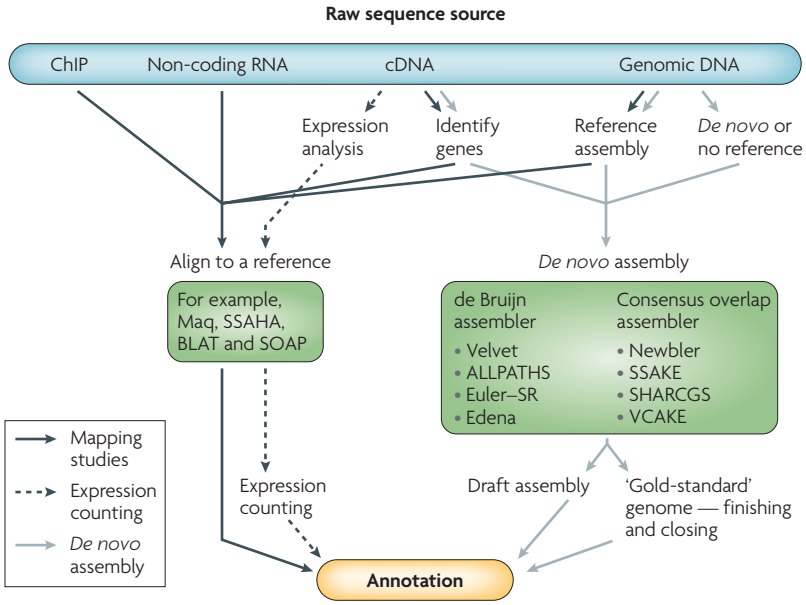


Figure 3 | **Road map for planning software solutions for experiments with different data sources and different goals.** Sequence reads derived from ChIP, non-coding RNA and cDNA sequencing experiments are aligned to a reference sequence before expression counting and final annotation. Sometimes, a cDNA sequence can be assembled *de novo* before these steps. Genome sequence reads may be aligned if a reference is available, but if not assembly *de novo* can still be carried out.

Software for sequence alignment. For applications such as SNP–discovery, ChIP–Seq, RNA–Seq and sRNA discovery, the first step, alignment of sequence reads against a reference sequence, which is usually a previously sequenced genome, is crucial. The best-known pairwise alignment algorithm, BLAST[58], uses a seed-and-extend strategy, whereby alignments that are extended form short, identical seed substrings. This approach does not work well for short query sequences that contain mismatches (owing to polymorphism or sequencing errors) because there might be no seed sequence from which to extend the alignment. Another problem is the high number of reads that need to be aligned: BLAST can be unacceptably slow when aligning tens of millions of short reads to a large genome. BLAT[59] and SSAHA[60] offer significantly faster alignment than BLAST, largely by holding an index of the query database in random access memory. BLAT was primarily intended for use in mapping ESTs against a reference genome, and works well for query sequences of at least ~100 nucleotides long that are nearly identical to the reference database; it can therefore be useful for aligning 454 pyrosequencing reads to a genome. By judicious choice of parameter values, SSAHA can be used to accurately and exhaustively map hundreds of thousands of short sequences, such as miRNAs (which are 21–24 nucleotides long), against eukaryotic genome sequences.

One of the authors of SSAHA has written a proprietary short-read alignment tool called ELAND, which Illumina supplies to its customers. However, many researchers prefer to use open-source products because they are freely available, allowing analyses to be repeated by multiple investigators, and because the underlying algorithms can be scrutinized, discussed and modified by the whole community. A number of open-source solutions have been released recently, including MAQ[61], GMAP[62], RMAP[63], PatMan[64], SHRiMP (see Further information), SOAP[65], SeqMap[66] and PASH[67]. Each of these software solutions has developed distinct, ingenious solutions. For example, SOAP indexes the reference rather than the query using two-bit-per-base encoding, which saves memory and speeds up searches, whereas PASH redefines the central search paradigm, indexing with a complex data structure called a positional hash.

One of the most important developments exploits quality information to inform confidence in an alignment. This idea has been implemented in the SHRiMP, MAQ and RMAP packages. At least two factors can lead to incorrect alignment. First, all genomes contain repeated sequences or at least nearly exact repeated sequences. This means that a sequence read could map equally well to more than one genomic site. Furthermore, the introduction of one or two sequencing errors or mutations could lead to placement of the read at the wrong site. The likelihood of a sequencing error can be estimated from the base-calling confidence scores for each of the read's nucleotides. To address this problem, for each sequence read, MAQ calculates the probability that that it has been aligned to the wrong place. This alignment score is a function of the base-calling confidence score and the number of alternative equally likely alignment sites in the reference genome. Alignment confidence scores must be taken into account when making inferences from alignments in applications such as SNP–discovery and ChIP–Seq. As is the case for sequence assembly, short-read alignment software is under active development, and the feature set and performance of each of the programs is likely to improve. Already, at least two alignment programs (MAQ and SOAP) can give improved accuracy by aligning mate pairs of reads. However, as yet, no comprehensive comparison of speed and accuracy has been performed for these alignment programs.

Software for downstream analysis. After alignment, census-based methods, such as ChIP–Seq and RNA–Seq require additional quantitation steps (FIG. 3). Freely available data-analysis programs and pipelines are already available; for example, FindPeaks[68], ERANGE[34] and QuEST[69]. Several of the alignment programs have built-in SNP-finding modules (for example, MAQ, Mosaik and SSAHA_SNP).

To make best use of such data, it is often helpful to display sequence data that are superimposed on an interactive genome browser; for example, based on the Generic Genome Browser[70]. One example is the *Chlamydomonas reinhardtii* silencing RNA database (see Further information), which displays data from the 454 GS FLX pyrosequencing of sRNA in *C. reinhardtii*. Although important datasets should eventually be integrated into central databases, such as Ensembl (see Further information), custom databases such as this one can be invaluable for individual laboratories and their collaborators who wish

N_{50}
A measure of contig length. If all contigs generated in an assembly are placed end to end in order of length (longest first), then the N_{50} is the length of the contig that, when added, causes the total length of the chain to exceed half of the length of the genome being sequenced. The longer the contigs are the longer the contig that would break this barrier.

BLAST
(Basic local alignment and search tool). A computer program for finding sequences in a database that have identity to a query sequence. BLAST has been available for years, and is the most widely used search tool.

MIGS
(Minimum information about a genome sequence). A proposed metadata standard that aims to capture essential species, the source of the strain and other phylogenetic and experimental data about a sequenced organism. Such data collection facilitates the cataloguing and searching of species in large-scale databases.

Finished genome
A genome sequence that has been shotgun sequenced and subjected to post-assembly procedures, such as long PCR, to close the gaps that occur between contigs.

to analyse their data before publication. Unfortunately, many small laboratories lack the informatics skills and infrastructure to effectively manage and analyse large datasets. However, help may soon be at hand thanks to projects such as EMAAS–Seq[71], a portal for RNA–Seq and ChIP–Seq that has an integrated genome browser[72] and is designed to be user friendly enough to be used by biologists who lack extensive bioinformatics expertise (S. Butcher, personal communication).

Community standards for data sharing

For decades, most laboratories have supported the free exchange of sequence data by depositing sequences in public sequence repositories. The advent of high-throughput sequencing data has created new challenges for the useful recording of experimental information and storage of sequences. The GSC (genomics standards consortium; a consortium of groups, including the DNA Data Bank of Japan, the European Bioinformatics Institute, the European Molecular Biology Laboratory, the Joint Genome Institute and the National Center for Biotechnology Information) and The Sanger Institute have proposed standards, such as MIGS, for the description of genome sequences[73]. MIGS holds information about origin, pathogenicity, host, growth conditions, estimated genome size and characteristics relating to growth, habitat, taxonomy and genetics. GEO (gene expression omnibus) and MGED (microarray and gene expression database) accept expression data from microarray platforms, together with metadata that describe the experiment descriptions as defined by the MIAME (minimum information about a microarray experiment) specification. These databases also accept sequence census data generated by next-generation methods. Since June 2008, the European Nucleotide Archive at the European Bioinformatics Institute (EMBL–EBI; see Further information) has begun to accept sequence reads generated by next-generation technologies. The emerging standard format for sequence deposition is sequence read format (a description is provided by SourceForge.net; see Further information), whereas the standards for metadata are still evolving.

Concluding remarks

In the short time that high-throughput sequencing technologies have been available, they have already made a huge impact on microbiology, providing a rapid and cost-effective means of generating draft genomes and widening access to genomics to individual laboratories, as well as the large genome-sequencing centres. For many purposes, a gold-standard finished genome is an expensive and unnecessary extravagance. Often, only a draft genome or transcriptome sequence that reveals the protein-coding potential of an organism is needed. This is certainly the case in molecular plant pathology, for which these methods enable us to discover a plethora of novel effectors and other virulence factors that are encoded in unsequenced genomes. The development of paired-read sequencing and the recent progress in bioinformatics of sequence assembly have been the key enabling factors.

Genome resequencing has already made an impact on microbiology. As was the case for microarrays a decade or so ago, the range of applications of the new sequencing technologies has expanded well beyond their original scope. Powerful approaches, such as ChIP–Seq, are making their mark in studies of higher animals, and it is surely only a question of time before they become almost routine in the study of microorganisms.

There is no sign of an end to the rapidly increasing rate of data generation. Applied Biosystems, Illumina and Roche's 454 Life Sciences have all recently announced upgrades to their respective systems that will increase read length and throughput while reducing cost per base. Illumina have recently begun supporting 45-nucleotide reads and are expected to increase read length to 70 nucleotides in the near future and to at least 100 nucleotides by the end of 2009.

Meanwhile, new contenders will soon join the three currently leading vendors. Pacific Biosciences have released some specifications for their sequencer, which they expect to launch in 2010. They claim that it will generate sequence reads of similar length to those of Sanger sequencing, which is capillary based. The Pacific Biosciences sequencer will empower the user to choose how to balance the trade-off between read length and volume of data.

These developments will contribute to exciting opportunities for microbiologists, but also bring new challenges in storing, transferring and analysing enormous datasets. Microbiology laboratories are going to become increasingly reliant on bioinformatics and information technology. This problem will become steadily more acute and will pose particular challenges as these methods are directed at metagenomic exploration of microbial diversity.

Note added in proof

An important development in the use of combinations of next-generation sequencing technologies has recently been made. Two publications have described a promising approach to *de novo* sequence assembly that combines Illumina and 454 sequencing technologies[74,75].

1. Pop, M. & Salzberg, S. L. Bioinformatics challenges of new sequencing technology. *Trends Genet.* **24**, 142–149 (2008).
An accessible overview of the computational challenges presented by new sequencing technologies.
2. Trombetti, G. A., Bonnal, R. J., Rizzi, E., De Bellis, G. & Milanesi, L. Data handling strategies for high throughput pyrosequencers. *BMC Bioinformatics* **8**, S22 (2007).
3. Hall, N. Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* **210**, 1518–1525 (2007).
4. Holt, R. A. & Jones, S. J. The new paradigm of flow cell sequencing. *Genome Res.* **18**, 839–846 (2008).
A comprehensive description of sequencing technologies and their applications.
5. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
6. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
7. Marguerat, S., Wilhelm, B. T. & Bähler, J. Next-generation sequencing: applications beyond genomes. *Biochem. Soc. Trans.* **36**, 1091–1096 (2008).
8. Medini, D. *et al.* Microbiology in the post-genomic era. *Nature Rev. Microbiol.* **6**, 419–430 (2008).
9. Rusk, N. & Kiermer, V. Primer: sequencing — the next generation. *Nature Methods* **5**, 15 (2008).

10. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nature Methods* **5**, 16–18 (2008).
11. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nature Biotechnol.* **26**, 1135–1145 (2008).
Contains detailed descriptions of sequencing technologies and their applications, and a useful survey of available software.
12. Snyder, L. A., Loman, N., Pallen, M. J. & Penn, C. W. Next-generation sequencing — the promise and perils of charting the great microbial unknown. *Microb. Ecol.* **57**, 1–3 (2009).
13. Steinberg, K. M., Okou, D. T. & Zwick, M. E. Applying rapid genome sequencing technologies to characterize pathogen genomes. *Anal. Chem.* **80**, 520–528 (2008).
14. Wold, B. & Myers, R. M. Sequence census methods for functional genomics. *Nature Methods* **5**, 19–21 (2008).
15. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
16. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S, R. Sequence information can be obtained from single DNA molecules. *Proc. Natl Acad. Sci. USA* **100**, 3960–3964 (2003).
17. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–109 (2008).
18. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15**, 589–594 (2005).
19. Velicer, G. J. Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor. *Proc. Natl Acad. Sci. USA* **103**, 8107–8112 (2006).
20. Mardis, E., McPherson, J., Martienssen, R., Wilson, R. K. & McCombie, W. R. What is finished, and why does it matter. *Genome Res.* **12**, 669–671 (2002).
21. Stiens, M. *et al.* Comparative genomic hybridisation and ultrafast pyrosequencing revealed remarkable differences between the *Sinorhizobium meliloti* genomes of the model strain Rm1021 and the field isolate SM11. *J. Biotechnol.* **136**, 31–37 (2008).
22. La Scola, B. *et al.* Rapid comparative genomic analysis for clinical microbiology: the *Francisella tularensis* paradigm. *Genome Res.* **18**, 742–750 (2008).
23. Dinsdale, E. A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **455**, 830 (2008).
The 454 GS20 technology developed by Roche enabled the authors to find that metagenomes from different biomes encode distinctly different metabolic profiles.
24. Ossowski, S. *et al.* Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.* **18**, 2024–2033 (2008).
The authors tackle genome-wide polymorphism by integrating 'resequencing' approaches with *de novo* assembly.
25. Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376 (2008).
26. Holt, K. E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nature Genet.* **40**, 987–993 (2008).
27. Liu, Z. *et al.* Patterns of diversifying selection in the phytotoxin-like *scr74* gene family of *Phytophthora infestans*. *Mol. Biol. Evol.* **22**, 659–672 (2004).
28. Kamoun, S. A catalogue of the effector secretome of plant pathogenic oomycetes. *Annu. Rev. Phytopathol.* **44**, 41–60 (2006).
29. Srivatsan, A. *et al.* High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS Genet.* **4**, e1000139 (2008).
30. Loman, N. J. & Pallen, M. J. XDR-TB genome sequencing: a glimpse of the microbiology of the future. *Future Microbiol.* **3**, 111–113 (2008).
31. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
32. Cheung, F. *et al.* Analysis of the *Pythium ultimum* transcriptome using Sanger and pyrosequencing approaches. *BMC Genomics* **9**, 542 (2008).
33. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* **5**, 613–619 (2008).
34. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA–Seq. *Nature Methods* **5**, 621–628 (2008).
35. Nagalakshmi, U. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
36. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
This ambitious and comprehensive survey of the epigenome was enabled by sequencing technology developed by Illumina.
37. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA–Seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
38. Shendure, J. The beginning of the end for microarrays? *Nature Methods* **5**, 585–587 (2008).
39. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
40. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold B. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
41. Taylor, K. H. Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res.* **67**, 8511–8518 (2007).
42. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
43. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
44. Hakimi, M. A. & Deitsch, K. W. Epigenetics in Apicomplexa: control of gene expression during cell cycle progression, differentiation and antigenic variation. *Curr. Opin. Microbiol.* **10**, 357–362 (2007).
45. Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. & Bushman, F. D. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* **17**, 1186–1194 (2007).
46. Molnár, A., Schwach, F., Studholme, D. J., Thuenemann, E. C. & Baulcombe, D. C. miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature* **447**, 1126–1129 (2007).
47. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. *Genome Res.* **17**, 1697–1706 (2007).
48. Warren, R. L., Sutton, G. G., Jones, S. J. & Holt, R. A. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23**, 500–501 (2007).
49. Jeck, W. R. *et al.* Extending assembly of short DNA sequences to handle error. *Bioinformatics* **23**, 2942–2944 (2007).
50. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA* **98**, 9748–9753 (2001).
51. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
52. Chaisson, M. J. & Pevzner, P. A. Short read fragment assembly of bacterial genomes. *Genome Res.* **18**, 324–330 (2008).
53. Hernandez, D., François, P., Farinelli, L., Osterås, M. & Schrenzel, J. *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.* **18**, 802–809 (2008).
54. Butler, J. *et al.* ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.* **18**, 810–820 (2008).
55. Phillippy, A. M., Schatz, M. C. & Pop, M. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.* **9**, R55 (2008).
56. Huang, W. & Marth, G. EagleView: a genome assembly viewer for next-generation sequencing technologies. *Genome Res.* **18**, 1538–1543 (2008).
57. Farrer, R. A., Kemen, E., Jones, J. D. G. & Studholme, D. J. *De novo* assembly of the *Pseudomonas syringae* pv. *syringae* B728a genome using Illumina/Solexa short sequence reads. *FEMS Microbiol. Lett.* **291**, 103–111 (2009).
58. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
59. Kent, W. J. BLAT — the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
60. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
61. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**,1851–1858 (2008).
62. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
63. Smith, A. D., Xuan, Z. & Zhang, M. Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* **9**, 128 (2008).
64. Prüfer, K. *et al.* PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics* **24**, 1530–1531 (2008).
65. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
66. Jiang, H. & Wong, W. H. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24**, 2395–2396 (2008).
67. Coarfa, C. & Milosavljevic, A. Pash 2.0: scaleable sequence anchoring for next-generation sequencing technologies. *Pac. Symp. Biocomput.* 102–113 (2008).
68. Fejes, A. P. *et al.* FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* **24**, 1729–1730 (2008).
69. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP–Seq data. *Nature Methods* **5**, 829–834 (2008).
70. Stein, L. D. The generic genome browser: a building block for a model organism system database. *Genome Res.* **12**, 1599–1610 (2002).
71. Barton, G. *et al.* EMAAS: an extensible grid-based rich internet application for microarray data analysis and management. *BMC Bioinformatics* **9**, 493 (2008).
72. Huntley, D., Tang, Y. A., Nesterova, T. B., Butcher, S. & Brockdorff, N. Genome Environment Browser (GEB): a dynamic browser for visualising high-throughput experimental data in the context of genome features. *BMC Bioinformatics* **9**, 501 (2008).
73. Field, D. *et al.* The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnol.* **26**, 541–547 (2008).
74. Aury, J. M. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. *BMC Genomics* **9**, 603 (2008).
75. Reinhardt, J. A. *et al. De novo* assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. *oryzae*. *Genome Res.* **19**, 294–305 (2009).

Acknowledgements

We are grateful to S. Kamoun, E. Kemen, S. Foster and M. Pallen for useful discussions and suggestions on the manuscript. This work was supported by Gatsby Foundation core funding to The Sainsbury Laboratory.

DATABASES

Entrez Genome Project: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
Arabidopsis thaliana | *Bacillus subtilis* | *Chlamydomonas reinhardtii* | *Escherichia coli* | *Francisella tularensis* subsp. *holartica* | *Mycobacterium tuberculosis* | *Myxococcus xanthus* | *Neurospora crassa* | *Phytophthora infestans* | *Salmonella enterica* subsp. enteric serovar Typhi | *Sinorhizobium meliloti*

FURTHER INFORMATION

David J Studholme's homepage: http://www.tsl.ac.uk/Bioinformatics.htm
1001 Genomes Project: http://1001genomes.org/
454 sequencing: http://www.454.com
***Chlamydomonas reinhardtii* silencing RNA database:** http://cresirna.cmp.uea.ac.uk/
EMBL–EBI: http://www.ebi.ac.uk/
Ensembl: http://www.ensembl.org/index.html
Genomes OnLine Database: http://www.genomesonline.org/
Helicos BioSciences: http://www.helicosbio.com
Pacific Biosciences: http://compbio.cs.toronto.edu/shrimp/
SEQanswers: http://seqanswers.com/forums/showthread.php?t=43
SHRiMP: http://compbio.cs.toronto.edu/shrimp/
Solexa: http://www.solexa.com
SOLiD Systems Sequencing: http://solid.appliedbiosystems.com
SourceForge.net (sequence read format description): http://srf.sourceforge.net/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF