High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity

Nicholas J. Loman[1], Chrystala Constantinidou[1], Jacqueline Z. M. Chan[1], Mihail Halachev[1], Martin Sergeant[1], Charles W. Penn[1], Esther R. Robinson[2] and Mark J. Pallen[1]

Abstract | Here, we take a snapshot of the high-throughput sequencing platforms, together with the relevant analytical tools, that are available to microbiologists in 2012, and evaluate the strengths and weaknesses of these platforms in obtaining bacterial genome sequences. We also scan the horizon of future possibilities, speculating on how the availability of sequencing that is 'too cheap to metre' might change the face of microbiology forever.

In bacteriology, the genomic era began in 1995, when the first bacterial genome was sequenced using conventional Sanger sequencing[1]. Back then, sequencing projects required six-figure budgets and years of effort. A decade later, in 2005, the advent of the first high-throughput (or 'next-generation') sequencing technologies signalled a significant advance in the ease and cost of sequencing[2], delivering bacterial genome sequences in hours or days rather than months or years. High-throughput sequencing now delivers sequence data thousands of times more cheaply than is possible with Sanger sequencing. The availability of a growing abundance of platforms and instruments presents the user with an embarrassment of choice. Better still, vigorous competition between manufacturers has resulted in sustained technical improvements on almost all platforms. This means that in recent years our sequencing capability has been doubling every 6–9 months — much faster than Moore's law.

Here, we describe the sequencing technologies themselves, examine the practicalities of producing a sequence-ready template from bacterial cultures and clinical samples, and weigh up the costs of labour and kits. We look at the types of data that are delivered by each instrument, and describe the approaches, programs and pipelines that can be used to analyse these data and thus move from draft to complete genomes.

Several high-throughput sequencing platforms are now chasing the US$1,000 human genome[3]. Given that the average bacterial genome is less than one-thousandth the size of the human genome, a back-of-the-envelope calculation suggests that a $1 bacterial genome sequence is an imminent possibility. In closing, we assess how close to reality the $1 bacterial genome actually is and explore the ways in which high-throughput sequencing might change the way that all microbiologists work.

A variety of approaches

High-throughput sequencing platforms can be divided into two broad groups depending on the kind of template used for the sequencing reactions. The earliest, and currently most widely used, platforms depend on the production of libraries of clonally amplified templates. These are produced through amplification of immobilized libraries made from a single DNA molecule in the initial sample. More recently, we have seen the arrival of single-molecule sequencing platforms, which determine the sequence of single molecules without amplification. Within these broad categories, there is considerable variation in performance — including in throughput, read length and error rate — as well as in factors affecting usability, such as cost and run time.

Template amplification technologies. In general terms, all of the platforms that are currently on the market rely on a three-stage workflow of library preparation, template amplification and sequencing (FIG. 1). Library preparation begins with the extraction and purification of genomic DNA. Depending on the protocol, the amount of DNA required can vary from a few nanograms to tens of micrograms, meaning that success in this step depends on the ability to grow sufficient biomass. For some microorganisms, obtaining suitable DNA — in terms of quantity and quality — can prove difficult. Therefore, before using expensive reagents for library preparation and sequencing, it is advisable to confirm, by fluorometry, that DNA of sufficient quantity and quality has been obtained. However, purchasing a suitable instrument to do this adds to the costs of establishing a sequencing capability (BOX 1).

For shotgun sequencing, an initial fragmentation step is required to generate random, overlapping DNA fragments. Depending on the platform and application, these fragments can range from 150 bp to 800 bp in length; size selection either involves harvesting from agarose gels or exploits paramagnetic-bead-based technology. The selected fragments must also be sufficiently abundant to provide comprehensive and even coverage of the target genome. Two types of fragmentation are widely used: mechanical and enzymatic. Early protocols relied on mechanical methods such as nebulization or ultrasonication. Nebulization is an inexpensive method that can be easily adopted by any laboratory, but it results in large losses of input material and a broad range of fragment sizes, runs the risk of cross-contamination and cannot handle parallel processing. By contrast, ultrasonication instruments such as systems from Covaris or the Bioruptor systems from Diagenode allow parallel sample processing and minimize hands-on time and sample loss but come at a price that could be prohibitive for small laboratories. Mechanically generated fragments require repair and end-polishing before platform-specific adaptors can be ligated to

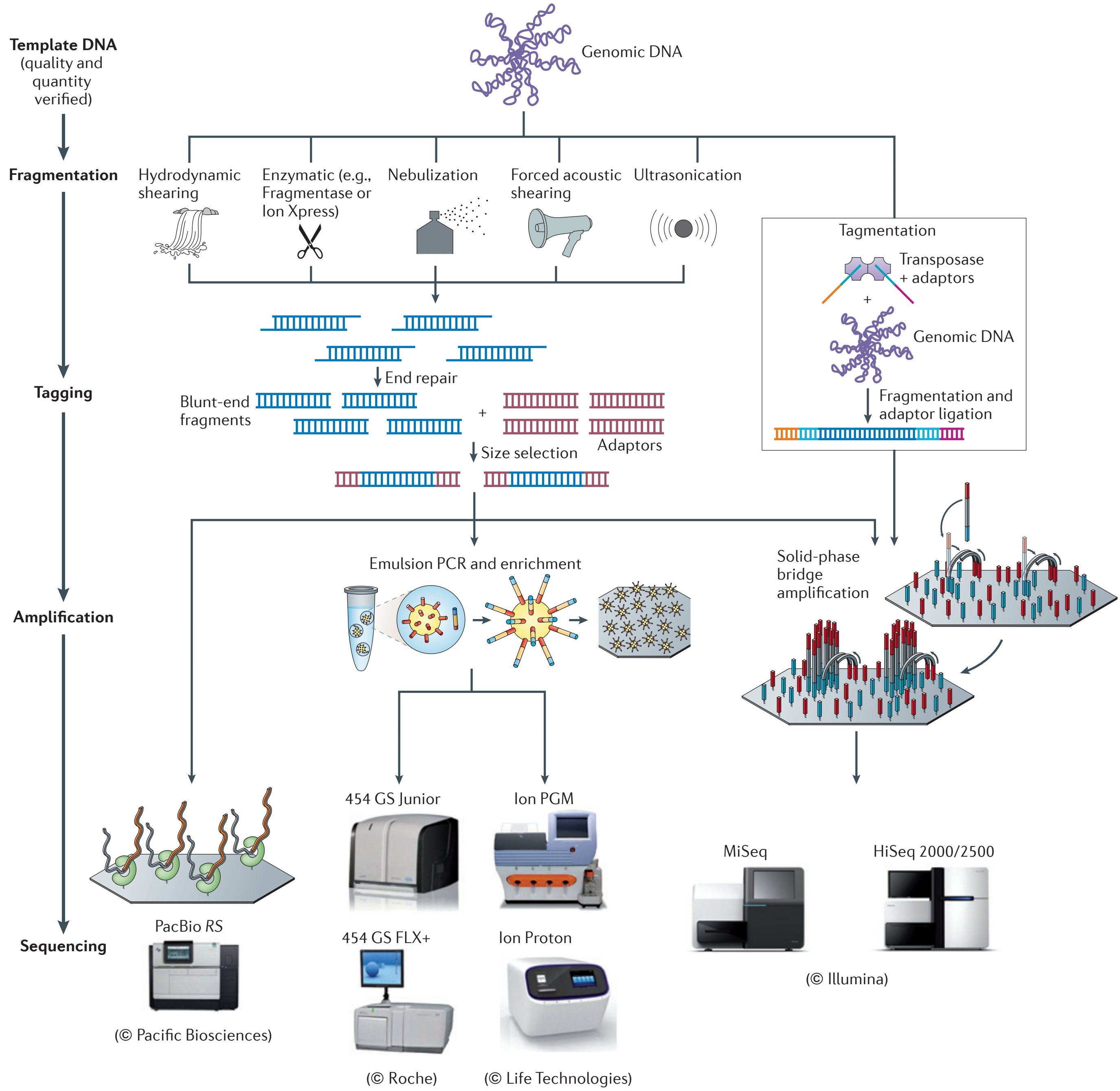

Figure 1 | **High-throughput sequencing platforms.** The schematic shows the main high-throughput sequencing platforms available to microbiologists today, and the associated sample preparation and template amplification procedures. For full details, see main text. PGM, Personal Genome Machine. The tagmentation schematic is modified, with permission, from REF. 57 © (2010) BioMed Central.

the ends of the target molecules. These adaptors act as primer-binding sites for the subsequent template amplification reaction.

More recently, enzymatic methods have provided an alternative approach to producing random fragments of the desired length. These require less input DNA and offer easier, faster sample processing. Fragmentase (from New England Biolabs) is a mixture of a nuclease, which randomly nicks double-stranded DNA, and a T7 endonuclease, which cleaves the DNA. Together, these enzymes generate random double-strand DNA breaks in a time-dependent manner, allowing the user to tailor protocols in order to obtain products of the required length. Adaptors can then be ligated to these fragments in the usual way. Tagmentation[4] is a promising transposase-based approach that, in a single step, fragments DNA and incorporates sequence tags, which then take the place of adaptors. Currently, the only available implementation of tagmentation is within the Nextera system, which is only available for the Illumina platform. Several companies have produced automated liquid-handling machines that greatly reduce the hands-on time required for fragmentation approaches but significantly increase costs (BOX 1).

In addition to supporting fragment-based sequencing, all template amplification platforms support mate pair sequencing, in which the ends of DNA fragments of a certain size (typical sizes are 3 kb, 6 kb, 8 kb or 20 kb) are joined together to form circular molecules. These molecules are then fragmented a second time. Fragments flanking the joins are then selected and end adaptors added. Sequencing through the joins provides valuable information about the location of sequences dispersed across the genome, facilitating assembly.

Paired-end sequencing has similarities to mate pair sequencing, but DNA fragments are sequenced from each end without the need for additional library preparation steps. The Illumina platform has direct support for paired-end sequencing. Short fragments that are less than the read length from the forward and reverse ends (for example, 180 bp fragments combined with 2 × 100 base sequencing) permits overlapping pseudo long reads to be generated. Alternatively, fragments of up to ~800 bp can be used. Longer fragments may result in a loss of amplification efficiency. The Ion Personal Genome Machine (PGM) (using the Ion Torrent platform, from Life Technologies) also has a bidirectional sequencing protocol that requires the removal of the chip after the initial run, a digestion step and a second sequencing run using a different sequencing primer. All platforms can handle PCR products, allowing adaptor sequences to be incorporated into the 5′ ends of primers.

For all platforms, it is highly advisable to assess the quality and quantity of the sequence library before subjecting it to amplification. Different instruments for quality assessment are recommended by different manufacturers. Examples include the 2100 Bioanalyzer (from Agilent Technologies), fluorometers such as the NanoDrop 3300 (from Thermo Scientific) or the Qubit (from Life Technologies), and quantitative PCR using any of a number of available quantitative PCR machines along with either own-design or commercially available assays. Purchasing a suitable instrument for this step can add several thousand dollars to the costs of establishing a sequencing capability (BOX 1).

In preparation for amplification, template molecules are immobilized on a solid surface, which is a flow cell for sequencing with the Illumina platform and solid beads or ion sphere particles for other approaches. Simultaneous solid-phase amplification of millions or billions of spatially separated template fragments prepares the way for massively parallel sequencing. For the Illumina platform, template amplification is automated and is performed either directly on the instrument (for the MiSeq, and the HiSeq 2500 sequencer in rapid-run mode) or using the cBot, a separate instrument that is dedicated to this task (used in conjunction with the Genome Analyzer IIx and the HiSeq 2000 machine). Clusters are generated by bridge amplification on the surface of the flow cell. For platforms that use bead-based immobilization (the SOLiD (from Life Technologies), 454 and Ion Torrent platforms), amplified template sequence libraries are prepared off-instrument, relying on an emulsion PCR, in which the beads are enclosed in aqueous-phase microreactors and are kept separated from each other in a water-in-oil emulsion.

Box 1 | The add-on costs of sequencing

The costs of sequencing instruments and reagents are not the only issues that need to be taken into account when setting up a sequencing facility for microbial applications. So, what else do you need? Well, first you have to buy a high-end fluorometer such as a Life Technologies Qubit (around US$2,000) and/or an Agilent Technologies 2100 Bioanalyzer (around $18,000). Then, if you want to save time by parallel processing, you should consider investing in an ultrasonicator (for example, from Covaris, at around $45,000) and a liquid-handling robot (for example, the Biomek FX^{P}, at around $310,000, or one of the SPRIworks systems, at around $45,000; both from Beckman Coulter). To carry out sequencing on the 454 GS FLX+ instrument from Roche, you need a bead counter for emulsion PCR (up to $20,000), and for the Genome Analyzer IIx or HiSeq machines from Illumina, you need to buy an Illumina cBot (~$55,000). For some platforms, you may have to buy additional centrifuges and/or rotors; for example, the ULTRA-TURRAX Tube Drive system from IKA ($1,000) is required by the Ion Torrent platform (from Life Technologies) if the OneTouch system is not used. You also need to buy a server to take receipt of the data coming off your instrument (for example, a $5,000 desktop), and then a cluster of servers for analysing and storing the data (ranging from $20,000 upwards). In addition, you may have to update your laboratory infrastructure by investing in a dedicated electrical connection and appropriate air-conditioning units for your sequencing instrument, and uninterruptible power supplies for your sequencer and servers. Most laboratories also want to invest in a backup solution that is both fast and available. This may be a mirrored set of hard drives, or even a shelf full of disconnected USB drives. Illumina offers a cloud-based backup and basic-analysis solution called BaseSpace which can store sequence results as they are generated on the Illumina MiSeq. Currently, this is a free solution, but users are likely to have to pay a subscription in the future.

Sequencing chemistry. Although these platforms rely on a sequencing-by-synthesis design, they differ in the details of the sequencing chemistry and the approach used to read the sequence. The Illumina sequencing platform depends on Solexa chemistry[5], which includes reversible termination of sequencing products. In each sequencing cycle, a mixture of fluorescently labelled 'reversible terminator' nucleotides with protected 3′-OH groups (and a different emission wavelength for each nucleotide) is perfused across the flow cell. Wherever a complementary nucleotide is present on the template strand, the terminator is incorporated and imaged, and then the signal is quenched and the terminator nucleotide is chemically deprotected at the 3′-OH group.

The 454 and Ion Torrent sequencing platforms avoid the use of terminators. Instead, in each cycle a single kind of dNTP is flowed across the template. When there is base complementarity between the dNTP and the next available position in the template, the DNA polymerase incorporates the base onto the extending strand, liberating pyrophosphate and hydrogen ions. When there is no complementarity, DNA synthesis is halted temporarily; each type of dNTP is flowed across the template in turn according to the dispensing cycle, and DNA synthesis is thus re-initiated when the next complementary dNTP is added. The 454 platform exploits a pyrosequencing approach[6,7] whereby the presence of pyrophosphate is signalled by visible light as the result of an enzyme cascade. The order and intensity of the light peaks are recorded as 'flowgrams'. The Ion Torrent platform relies on a modified silicon chip to detect hydrogen ions that are released during base incorporation; the resulting lack of reliance on imaging makes this platform the first 'post-light' sequencing instrument[8].

The SOLiD platform[9] and the platform from Complete Genomics[10] depend on sequencing by ligation. In this approach, fluorescent probes undergo iterative steps of hybridization and ligation to complementary positions in the template strand at the 5′ end of the extending strand, followed by fluorescence imaging to identify the ligated probe.

Single-molecule sequencing. Single-molecule sequencing brings the promise of freedom from amplification artefacts as well as from onerous sample and library preparations. The HeliScope Single-Molecule Sequencer (from Helicos BioSciences) was the first platform for single-molecule sequencing to hit the market place in 2009 (REF. 11). This technology applies one-colour reversible-terminator sequencing to unamplified single-molecule templates. However, this platform has been hampered by its high price and poor instrument sales and, following the delisting of the company from the stock market, there are significant doubts over the future of the platform.

More recently, Pacific Biosciences has delivered 'real-time sequencing', in which dye-labelled nucleotides are continuously incorporated into a growing DNA strand by a highly processive, strand-displacing φ29-derived DNA polymerase[12]. Each DNA polymerase molecule is tethered within a zero-mode waveguide detector, which allows continuous imaging of the labelled nucleotides as they enter the strand[13].

Choosing a platform

High-end instruments. The high-throughput sequencing market presents the user with a challenging choice between bulky, expensive high-end instruments and the new generation of bench-top instruments (TABLES 1,2). The high-end machines include PacBio *RS* (from Pacific Bioseciences), the HiSeq instruments, Genome Analyzer IIx, the SOLiD 5500 series and the 454 GS FLX+ system. These deliver a high throughput and/or long read lengths but come with set-up costs of hundreds of thousands of dollars, placing them beyond the reach of the average research laboratory or even department. These machines are thus only suitable for large sequencing centres or core facilities. This raises the important question of where an 'average' microbiologist should source sequencing from.

These instruments can deliver dozens to thousands of bacterial genomes per run, as illustrated by several high-impact publications on bacterial genomes and metagenomes[14–17]. However, to achieve efficiencies in time and cost, optimum sequencing of microbial samples on such instruments requires onerous and expensive bar-coding and multiplexing of samples and/or subdivision of runs (for example, through gaskets or the use of single channels on the Illumina platform), as well as a sophisticated scheduling system. Compare sequencing a single human genome with the equivalent sequencing throughput for 1,000 average-sized bacterial genomes: although the sequencing run itself may be comparable in both scenarios, >1,000 samples and libraries need to be prepared for the bacterial run, compared with just one for the human genome. The costs and effort involved in sequencing 1,000 bacterial genomes therefore vastly outweigh the requirements for sequencing a single human genome, so the hasty calculation that one human genome-sequencing project equates to 1,000 bacterial genome-sequencing projects starts to look rather optimistic.

Bench-top instruments. Three modestly priced bench-top instruments with throughputs and workflows that are well suited to microbial applications have recently hit the market. The 454 GS Junior was released in early 2010 and is a smaller, lower-throughput version of the 454 GS FLX+ machine, exploiting similar emulsion PCR and pyrosequencing approaches but with lower set-up and running costs[18]. The Ion PGM was launched in early 2011 and saw almost immediate use in the crowd-sourced analysis of the Shiga toxin-producing *Escherichia coli* (STEC) outbreak in Germany[19,20]. This platform has also shown the greatest improvement in performance in recent months: an assembly for the STEC outbreak strain was generated in May 2011 using data from five Ion Torrent 314 chips and consisted of more than 3,000 contigs, whereas comparable data from a single newer 316 chip assembled into fewer than 400 contigs. The MiSeq, which began to ship to customers in late 2011, is based on the existing Solexa chemistry but has dramatically reduced run times compared with the HiSeq (hours rather than days). This is made possible by the use of a smaller flow cell, leading to a reduced imaging time and faster microfluidics.

Each of these bench-top instruments is capable of sequencing a whole bacterial genome in days. The performance of all three instruments was recently compared by sequencing a British isolate from the German STEC outbreak of 2011 (REF. 18). In this evaluation, all three bench-top sequencing platforms generated useful draft genome sequences with assemblies that mapped to ≥95% of the reference genome, so by these criteria all could be judged fit for purpose. However, no instrument was able to generate accurate one-contig-per-replicon assemblies that might equate to a finished genome.

The MiSeq was found to have the highest throughput per run, lowest error rate and most user-friendly workflow of the three instruments: hands-on time is low because template amplification is carried out directly on the instrument without manual intervention. However, a paired-end 150-base sequencing run took more than 27 hours. The MiSeq is notable for being able to sequence fragments from both ends (paired-end mode) without changes to the library preparation stage or additional intervention during sequencing.

The 454 GS Junior produced the longest reads (mean 522 bases) and generated the least fragmented assemblies but had the lowest throughput and a cost-per-base that was at least one order of magnitude higher than the cost for the other two platforms. The Ion PGM delivered the fastest throughput per hour (80–100 Mb) and had the shortest run time (around 3 hours) but also had the shortest reads (mean 121 bases), although kits producing 200 bases have since been made available for this instrument. The Ion PGM and 454 GS Junior were both prone to making mistakes in homopolymeric tracts, and these mistakes caused assembly errors that resulted in frame-shifts in coding regions, even when data were assembled at high read coverage.

Coping with the data

The high-end sequencing platforms make considerable demands on the local information technology infrastructure in terms of data tracking and analysis, short-term storage and long-term archiving. Bench-top instruments have more modest information technology requirements. However, each platform delivers data in a slightly different format, and saying that one has sequenced a bacterial genome means different things on different platforms and can create difficulties when comparing or combining data generated on different platforms (TABLE 2).

There are two main analytical approaches to the exploitation of high-throughput sequencing data: reads can be aligned — that is, mapped — to a known reference sequence or subjected to *de novo* assembly. The choice of strategy depends on the read length obtained (short reads are better mapped to a reference), the availability of a good reference sequence and the intended biological application (for example, genomic epidemiology versus pathogen biology).

To document genetic variation in the genomes of multiple highly related strains, a mapping approach is efficient and often sufficient. In this situation, sequence variants can be called by aligning reads to a reference genome using short-read-mapping tools (see Supplementary information S1 (table)). A mapping approach is problematic when dealing with reads from repetitive regions or from parts of the genome that are absent

Table 1 | **Comparison of next-generation sequencing platforms**

Machine (manufacturer)	Chemistry	Modal read length* (bases)	Run time	Gb per run	Current, approximate cost (US$)‡	Advantages	Disadvantages
High-end instruments							
454 GS FLX+ (Roche)	Pyrosequencing	700–800	23 hours	0.7	500,000	• Long read lengths	• Appreciable hands-on time • High reagent costs • High error rate in homopolymers
HiSeq 2000/2500 (Illumina)	Reversible terminator	2 × 100	11 days (regular mode) or 2 days (rapid run mode)§	600 (regular mode) or 120 (rapid run mode)§	750,000	• Cost-effectiveness • Steadily improving read lengths • Massive throughput • Minimal hands-on time	• Long run time • Short read lengths • HiSeq 2500 instrument upgrade not available at time of writing (available end 2012)
5500xl SOLiD (Life Technologies)	Ligation	75 + 35	8 days	150	350,000	• Low error rate • Massive throughput	• Very short read lengths • Long run times
PacBio *RS* (Pacific Biosciences)	Real-time sequencing	3,000 (maximum 15,000)	20 minutes	3 per day	750,000	• Simple sample preparation • Low reagent costs • Very long read lengths	• High error rate • Expensive system • Difficult installation
Bench-top instruments							
454 GS Junior (Roche)	Pyrosequencing	500	8 hours	0.035	100,000	• Long read lengths	• Appreciable hands-on time • High reagent costs • High error rate in homopolymers
Ion Personal Genome Machine (Life Technologies)	Proton detection	100 or 200	3 hours	0.01–0.1 (314 chip), 0.1–0.5 (316 chip) or up to 1 (318 chip)	80,000 (including OneTouch and server)	• Short run times • Appropriate throughput for microbial applications	• Appreciable hands-on time • High error rate in homopolymers
Ion Proton (Life Technologies)	Proton detection	Up to 200	2 hours	Up to 10 (Proton I chip) or up to 100 (Proton II chip)	145,000 + 75,000 for compulsory server	• Short run times • Flexible chip reagents	• Instrument not available at time of writing
MiSeq (Illumina)	Reversible terminator	2 × 150	27 hours	1.5	125,000	• Cost-effectiveness • Short run times • Appropriate throughput for microbial applications • Minimal hands-on time	• Read lengths too short for efficient assembly

*Average read length for a fragment-based run. ‡Approximate cost per machine plus additional instrumentation and service contract. See REF. 58. §Available only on the HiSeq 2500.

from the reference genome, or when a closely related reference genome is unavailable.

De novo assembly is more informative when dealing with a new pathogen or a new strain of a well-known pathogen. Sequencing errors can have a significant impact on assembly. When platforms produce random errors, the effect of these errors on assembly can be overcome by increasing the depth of coverage. However, when errors are systematic and occur in predictable contexts (for example, in homopolymers), increasing the depth of coverage is unlikely to help, and it may be necessary to sequence the troublesome regions using an alternative technology. Very high-quality, near complete references may be obtained by a hybrid approach, such as in recent studies combining Pacific Biosciences and Illumina data[21,22].

A variety of commonly used assemblers is now available (see Supplementary information S1 (table)), ranging from the platform specific (for example, Newbler from Roche) to the more generally applicable (for example, MIRA[23], Velvet[24], and the CLC Genomics Workbench from CLC Bio).

Table 2 | **The applicability of the major high-throughput sequencing platforms**

Example application in bacteriology	Desirable characteristics	Machine*						
		454 GS Junior‡	454 GS FLX+‡	Ion Personal Genome Machine (318 chip)§	MiSeq‖	HiSeq 2000‖	5500xl SOLiD§	PacBio *RS*¶
De novo sequencing of novel strains to generate a single-scaffold reference genome	• Long reads • Paired-end protocol and/or long mate-pair protocol • Even coverage of genome	✓	✓✓	✓	✓	✓	X	✓✓
Rapid characterization of a novel pathogen (draft *de novo* assembly of a genome for a single strain)	• Total run time (library preparation plus sequencing) of under 48 hours • Sufficient coverage of a bacterial genome in a single run	✓	✓✓	✓✓	✓✓	X	X	✓✓
Rough-draft *de novo* sequencing of small numbers of strains (<20) for comparative analysis of gene content	• Long or paired-end reads • High throughput • Ease of library and sequencing workflow • Cost-effective	X	✓	✓	✓✓	✓✓	✓	✓
Re-sequencing of many similar strains (>50) for the discovery of single nucleotide polymorphisms and for phylogenetics	• Very high throughput • Low-cost, high-throughput sequence library construction • High accuracy	X	X	✓	✓	✓✓	✓	✓
Small-scale transcriptomics-by-sequencing experiments (for example, two strains under four growth conditions with two biological replicates, so 16 strains)	• High per-isolate coverage	X	✓	✓	✓	✓✓	✓✓	✓✓
Phylogenetic profiling to genus-level using partial 16S rRNA gene amplicon sequencing	• High coverage • Long amplicon input (≥500 bp) • Long reads • High single-read accuracy (error rate <1%)	✓	✓✓	✓	✓✓	✓	✓	X
Whole-genome metagenomics for the reconstruction of multiple genomes in a single sample	• Long reads or paired-end reads • Very high throughput • Low error rate	X	✓	✓	✓	✓✓	✓	✓

*✓✓, particularly well suited; ✓, suitable; X, not suitable. ‡From Roche. §From Life Technologies. ‖From Illumina. ¶From Pacific Biosciences.

De novo assemblies can be compared using Mauve[25] or Mugsy[26], and the assemblies can be manually examined using the Tablet viewer[27]. For annotation of assemblies, Glimmer[28] works well for coding-sequence prediction, while tRNAScan-SE[29] and RNAmmer[30] work well for stable-RNA prediction. There are numerous pipelines for automatic annotation of *de novo* assemblies, including RAST[31], IMG/ER[32] and the IGS Annotation Engine (developed by the Institute for Genome Sciences, University of Maryland School of Medicine, USA), although care must be taken when interpreting results from such services, as the public databases used contain annotation errors that are then propagated to newly sequenced genomes[33].

For microbial applications, all of the above programs run quickly (in minutes or hours) and are not particularly processor intensive. Some workflows combine a series of programs and provide an accessible interface for microbiologists who are not bioinformatics specialists. For example, xBASE-NG provides a 'one-stop shop' for assembly, annotation and comparison of bacterial genome sequences[34]. Sophisticated phylogenetic analyses are more demanding and may be beyond the capability of the average research group. One particular issue when constructing bacterial whole-genome phylogenies is the clouding of phylogenetic signal by recombination events and homoplasy[35]. Algorithms such as ClonalFrame[36] and ClonalOrigin[37] take multiple whole-genome alignments as input and attempt to identify blocks of recombination. These approaches are computationally very expensive, and there is no 'off the shelf' solution to comparing hundreds or thousands of bacterial genomes. There is a growing interest in alignment-free approaches for constructing bacterial phylogenies, as it is thought that these approaches may help address the computational challenges of these analyses[38].

A recurring problem with data from high-throughput sequencing is meeting the requirement, as stipulated by journals and funders, that data be lodged in the public domain. Unannotated assembled sequences can be uploaded to conventional sequence databases, such as GenBank, fairly easily. However, submission of annotated sequences can be onerous, slowing down the process of publication even further. Submission of sequence reads to short-read archives may be hampered by slow data transfer rates, and it remains uncertain how sustainable such archives will prove to be in the future. There may come a time when the easiest way to obtain such data will be to re-sequence the sample, rather than upload, archive and retrieve large data sets.

Current applications and future prospects

High-throughput sequencing has already transformed microbiology. Rapid, low-cost genome sequencing has helped make genomic epidemiology a reality, allowing us to track the spread of pathogens through hospitals[39,40], communities[19,20,41] and across the globe[16,42,43]. High-throughput sequencing has already had a huge impact on our understanding of microbial evolution, whether within a single patient over years or decades (for example, *Pseudomonas aeruginosa* in a patient with cystic fibrosis[44]) or globally across the centuries (for example, influenza virus in the 1918 influenza pandemic[45] or mediaeval *Yersinia pestis* in the Black Death[46]). Genome sequences have even been obtained from single microbial cells[47].

There are many applications beyond mere genome sequencing. High-throughput sequencing has opened up new avenues for sequence-based profiling and metagenomics of complex microbial communities, including those associated with human health and disease[14,15]. Particularly exciting is the promise of culture-independent approaches to pathogen discovery and detection[48]. In the research laboratory, sequencing is taking over from microarrays as the method of choice for studying gene expression (using RNA sequencing (RNA-seq))[49–51], mutant libraries (using Tn-seq and transposon-directed insertion site sequencing (TraDIS))[52,53] and protein–DNA interactions (using chromatin immunoprecipitation followed by sequencing (ChIP–Seq))[54].

So, what does the future hold? For current platforms, we can expect to see cheaper, easier library preparation methods and ever-higher sequencing throughputs. However, with the arrival of transformative new technologies[55] (BOX 2), this might be seen as tinkering around the edges. The tipping point has already been reached such that the staff and infrastructure costs of handling and analysing sequence data outweigh the costs of generating that data. If the promise of portable, single-molecule, long-read-length sequencing bears fruit and these technologies show the same steady increase in functionality and cost-effectiveness that we have seen with earlier high-throughput sequencing platforms, we could be just a few years away from user-friendly, '$1-a-pop' bacterial genome sequencing.

As we have argued elsewhere[56], high-throughput sequencing may well be poised to make a decisive impact on clinical microbiology, but there are still many difficulties to be overcome — for example, in presenting complex information to clinicians, in agreeing common formats for data sharing, in integrating genomics with clinical informatics and clinical practice, in benchmarking novel technologies and in gaining regulatory approval (from the US FDA and other bodies) for clinical applications of these technologies. One thing is certain: thanks to the expected relentless progress in sequencing technology, microbiology in the next 20 years will look nothing like it does now.

Box 2 | Oxford Nanopore: the game changer?

In February 2012, at a conference in the United States, the British company Oxford Nanopore Technologies announced a new, near-market "strand sequencing" technology that exploits protein nanopores embedded in an industrially fabricated polymer membrane. As a DNA strand is fed through a nanopore by a processive enzyme, the trinucleotides in contact with the pore are detected through electrochemistry.

The manufacturers have already claimed that they can sequence the 50 kb phage λ genome on both strands, and they claim that there is no theoretical read length limit. They also claim that sequencing can be paused, the sample recovered and replaced, and sequencing then started again. Plus, there is no need for onerous sample preparations: sequences can be read directly from blood (and probably also bacterial lysates).

Oxford Nanopore Technologies has announced two products, both scheduled to ship in late 2012. The MinION is a disposable US$900 sequencer housed in a USB stick, with 512 nanopores, each capable of running 120–1,000 bases per minute per pore for up to 6 hours. The MinION can generate 150 Mb of sequence per hour, all without fluidics or imaging, and bases are streamed live to a laptop through the USB connection. The GridION is a rack-mountable sequencer with 2,000 nanopores and is capable of generating tens of gigabases over 24 hours. Both machines promise astonishing read lengths at low cost and with minimal sample preparation. However, this technology currently suffers from a high error rate (~4%) that is chiefly due to deletion errors but, according to their February 2012 press conference, the manufacturers are confident that they can fix this.

How will access to a disposable sequencer change the way we do microbiology? With no capital costs or cumbersome set-up and installation, this technology certainly has the power to democratize sequencing even further. Will prices fall enough for it to be worth sequencing one bacterial genome per MinION, or will the long read lengths mean that we can mix samples and then disaggregate the genomes with little effort? If read lengths really can be obtained in the ≥100 kb range, then all the existing problems of short-read assembly in genomics and metagenomics will be rendered obsolete. Furthermore, we can now take the sequencer to the patient's bedside or out into the field. Microbial ecologists need no longer depend on molecular barcodes such as the 16S rRNA gene when they can have whole genomes instead, and latter-day John Snows can use disposable sequencing, not just to detect cholera, but also to track its evolution and spread.

Of course, the reality may not match the hype, and we eagerly await the first independent evaluation of this technology. But if the dream comes true, most of the rest of this article will soon be redundant.

Nicholas J. Loman, Chrystala Constantinidou, Jacqueline Z. M. Chan, Mihail Halachev, Martin Sergeant, Charles W. Penn and Mark J. Pallen are at the Institute of Microbiology and Infection, University of Birmingham, Birmingham B15 2TT, UK.

Esther R. Robinson is at the Nuffield Department of Clinical Laboratory Sciences, University of Oxford, Oxford OX3 9DU, UK.

Correspondence to M.J.P.
e-mail: m.pallen@bham.ac.uk

doi:10.1038/nrmicro2850
Published online 6 August 2012

1. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
2. Metzker, M. L. Emerging technologies in DNA sequencing. *Genome Res.* **15**, 1767–1776 (2005).
3. Venter, J. C. Multiple personal genomes await. *Nature* **464**, 676–677 (2010).
4. Caruccio, N. in *High-Throughput Next Generation Sequencing: Methods and Applications. Methods in Molecular Biology* Vol. 733 (eds Kwon, Y. M. & Ricke, S. C.) 241–255 (Humana Press, 2011).
5. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
6. Ronaghi, M., Uhlen, M. & Nyren, P. A sequencing method based on real-time pyrophosphate. *Science* **281** 363–365 (1998).
7. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
8. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
9. Valouev, A. *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).
10. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
11. Bowers, J. *et al.* Virtual terminator nucleotides for next-generation DNA sequencing. *Nature Methods* **6**, 593–595 (2009).
12. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
13. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).
14. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
15. Hess, M. *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463–467 (2011).
16. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462–465 (2011).

17. Harris, S. R. *et al.* Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. *Nature Genet.* **44**, 413–419 (2012).
18. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotech.* **30**, 434–439 (2012).
19. Rohde, H. *et al.* Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N. Engl. J. Med.* **365**, 718–724 (2011).
20. Mellmann, A. *et al.* Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS ONE* **6**, e22751 (2011).
21. Bashir, A. *et al.* A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotech.* 1 Jul 2012 (doi:10.1038/nbt.2288).
22. Koren, S. *et al.* Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotech.* 1 Jul 2012 (doi:10.1038/nbt.2280).
23. Chevreux, B. *et al.* Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14**, 1147–1159 (2004).
24. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
25. Darling, A. C., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
26. Angiuoli, S. V. & Salzberg, S. L. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics* **27**, 334–342 (2011).
27. Milne, I. *et al.* Tablet—next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010).
28. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).
29. Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **33**, W686–W689 (2005).
30. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
31. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
32. Markowitz, V. M. *et al.* IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* **25**, 2271–2278 (2009).
33. Richardson, E. J. & Watson, M. The automatic annotation of bacterial genomes. *Brief. Bioinform.* 9 Mar 2012 (doi:10.1093/bib/bbs007).
34. Chaudhuri, R. R. *et al.* xBASE2: a comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res.* **36**, D543–D546 (2008).
35. Marttinen, P. *et al.* Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* **40**, e6 (2012).
36. Didelot, X. & Falush, D. Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266 (2007).
37. Didelot, X., Lawson, D., Darling, A. & Falush, D. Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* **186**, 1435–1449 (2010).
38. Domazet-Lošo, M. & Haubold, B. Alignment-free detection of local similarity among viral and bacterial genomes. *Bioinformatics* **27**, 1466–1472 (2011).
39. Köser, C. U. *et al.* Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N. Engl. J. Med.* **366**, 2267–2275 (2012).
40. Lewis, T. *et al.* High-throughput whole-genome sequencing to dissect the epidemiology of *Acinetobacter baumannii* isolates from a hospital outbreak. *J. Hosp. Infect.* **75**, 37–41 (2010).
41. Gardy, J. L. *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N. Engl. J. Med.* **364**, 730–739 (2011).
42. Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2011).
43. Beres, S. B. *et al.* Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. *Proc. Natl Acad. Sci. USA* **107**, 4371–4376 (2010).
44. Cramer, N. *et al.* Microevolution of the major common *Pseudomonas aeruginosa* clones C and PA14 in cystic fibrosis lungs. *Environ. Microbiol.* **13**, 1690–1704 (2011).
45. Dunham, E. J. *et al.* Different evolutionary trajectories of European avian-like and classical swine H1N1 influenza A viruses. *J. Virol.* **83**, 5485–5494 (2009).
46. Bos, K. I. *et al.* A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* **478**, 506–510 (2011).
47. Woyke, T. *et al.* Assembling the marine metagenome, one cell at a time. *PLoS ONE* **4**, e5299 (2009).
48. Lipkin, W. I. Microbe hunting. *Microbiol. Mol. Biol. Rev.* **74**, 363–377 (2010).
49. Sorek, R. & Cossart, P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Rev. Genet.* **11**, 9–16 (2010).
50. Passalacqua, K. D. *et al.* Structure and complexity of a bacterial transcriptome. *J. Bacteriol.* **191**, 3203–3211 (2009).
51. Sharma, C. M. *et al.* The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**, 250–255 (2010).
52. van Opijnen, T., Bodi, K. L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature Methods* **6**, 767–772 (2009).
53. Langridge, G. C. *et al.* Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. *Genome Res.* **19**, 2308–2316 (2009).
54. Grainger, D. *et al.* Direct methods for studying transcription regulatory proteins and RNA polymerase in bacteria. *Curr. Opin. Microbiol.* **12**, 531–535 (2009).
55. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nature Biotech.* **26**, 1146–1153 (2008).
56. Pallen, M. J. & Loman, N. J. Are diagnostic and public health bacteriology ready to become branches of genomic medicine? *Genome Med.* **3**, 53 (2011).
57. Adey, A. *et al*. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Res.* **11**, R119 (2010).
58. Glenn, T. C. A field guide to next generation DNA sequencers. *Mol. Ecol. Res.* **11**, 759–769 (2011).

Acknowledgements

The authors thank the anonymous reviewers for their help and suggestions.

Competing interests statement

The authors declare competing financial interests; see Web version for details.

FURTHER INFORMATION

Mark J. Pallen's homepage: http://pathogenomics.bham.ac.uk/index.html

The *Pathogens: Genes and Genomes* blog: http://pathogenomics.bham.ac.uk/blog

SUPPLEMENTARY INFORMATION

See online article: S1 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF