MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11**, 759–769 doi: 10.1111/j.1755-0998.2011.03024.x

# Field guide to next-generation DNA sequencers

TRAVIS C. GLENN
*Department of Environmental Health Science and Georgia Genomics Facility, Environmental Health Science Building, University of Georgia, Athens, GA 30602, USA*

## Abstract

**The diversity of available 2nd and 3rd generation DNA sequencing platforms is increasing rapidly. Costs for these systems range from <$100 000 to more than $1 000 000, with instrument run times ranging from minutes to weeks. Extensive trade-offs exist among these platforms. I summarize the major characteristics of each commercially available platform to enable direct comparisons. In terms of cost per megabase (Mb) of sequence, the Illumina and SOLiD platforms are clearly superior (≤$0.10/Mb vs. >$10/Mb for 454 and some Ion Torrent chips). In terms of cost per nonmultiplexed sample and instrument run time, the Pacific Biosciences and Ion Torrent platforms excel, with the 454 GS Junior and Illumina MiSeq also notable in this regard. All platforms allow multiplexing of samples, but details of library preparation, experimental design and data analysis can constrain the options. The wide range of characteristics among available platforms provides opportunities both to conduct groundbreaking studies and to waste money on scales that were previously infeasible. Thus, careful thought about the desired characteristics of these systems is warranted before purchasing or using any of them. Updated information from this guide will be maintained at: http://dna.uga.edu/ and http://tomato.biol.trinity.edu/blog/.**

*Keywords*: 2nd and 3rd generation sequencing, 454, Helicos, Illumina, Ion Torrent, Life Technologies, massively parallel sequencing, Pacific Biosystems, Roche, SOLiD

*Received 17 March 2011; revision accepted 22 March 2011*

## Background

DNA sequencing technologies and platforms are being updated at a blistering pace, so much so that reviews of sequencing platforms resemble the work of Sisyphus. It is important, however, for molecular ecologists to keep pace with these technologies, because they are transforming what we can do, how we should do it, and how much it will cost. Institutions and researchers are committing up to a million dollars to purchase massively parallel sequencing instruments. Such purchases lock laboratories and institutions into specific paths for large annual expenditures in both consumable supplies and service contracts. Differences in instrument engineering, platform chemistry and economics related to design constrain what can be done with those instruments once they are purchased.

Several recent major announcements and acquisitions make this an opportune time to evaluate available platforms and what is likely to be available in the immediate future. In this brief guide, I summarize instruments currently available and those that have been announced by major companies. Although several of these platforms have very different strengths touted by the vendors, the weaknesses are often much less clear. I have therefore summarized available information in tables with categories of primary interest to purchasers and to users so that direct comparisons can be made. I will use the convention of 2nd generation to indicate a platform that requires amplification of the template molecules prior to sequencing, 3rd generation to indicate platforms that sequence directly individual DNA molecules, and next-generation sequencing (NGS) platforms to generically indicate 2nd or 3rd generation instruments.

This guide is intended to provide information for readers with little or advanced understanding of NGS platforms. I assume, however, that readers who are not familiar with these systems are learning details by: reading relevant publications (e.g. Mardis 2008; Shendure & Ji 2008; Ansorge 2009; Richardson 2010; Tautz *et al.* 2010), reading information at company and independent websites and talking with staff of the companies making NGS instruments.

My purpose is not to explain how these systems work in detail (that information is readily available from the sources noted above), but instead to focus on generally important traits of these systems and to provide relevant details for prospective buyers and users. In particular, my goal is to present information useful to researchers

Correspondence: Travis C. Glenn, Fax: 706 542 7472;
E-mail: travisg@uga.edu

who must determine what platform to use for their own experiments or who will recommend purchasing instruments so that they can make informed decisions and facilitate summaries of their decisions (e.g. for institutional purchasing support staff, administrators and in publications). I do not include information on Complete Genomics, deCode genetics, Knome or similar companies because they are focused solely on analysing human samples. I also will not cover the Polonator, Intelligent BioSystems, or other similar companies that have not yet been able to make significant commercial impact. I provide some information on Helicos because this company has only recently stopped selling instruments and reagents in favour of adopting a service-provider model, and their services are available for organisms of interest to molecular ecologists.

## Comparing the platforms

### *Caveats to the comparisons – need for standards*

All companies put out data and statements that cast their systems in the best possible light. I have generally accepted values from the companies to get at measures that can then be compared, but these comparisons have inherent flaws. There are no accepted standards for what measures the companies need to report, let alone particulars of how the data are analysed. The templates used, types of pre-analysis data filters used and number of runs used (e.g. best single run, average of many runs, etc.) can have significant impacts. Independent testing of NGS platforms to determine yield, error rates, etc. would be ideal, but is expensive and problematic because companies frequently update chemistry, software and other components of their systems. In several cases, available data give a broad range of values and I generally condense these data into a single number from the middle of the available data distribution. There are few places where I indicate dispersion of the values. For these reasons, many comparisons below are less than ideal. As in all field guides, the purpose here is to illustrate typical phenotypes.

Everyone using NGS data would benefit from the development of a standard set of conditions, analyses and a complex template (e.g. *Escherichia coli* genomic DNA) or set of templates (e.g. specific clones, *E. coli* genomic DNA, mouse cDNA, etc.) that could be adopted and used for testing of all platforms. Results from these templates could then be used to determine values that would allow direct comparison of NGS platforms, chemistry and software upgrades. Ideally, the standard template(s) would be similar to US National Institute of Standards and Technology (NIST) DNA standards for forensics and could be obtained from NIST or similar entities. Until such standards are developed and adopted, comparisons will remain difficult and inherently subjective, especially measures of error rate and mappable reads.

### *Basic characteristics*

Six 2nd and 3rd generation sequencing platforms are currently available, and a seventh is in advanced development (Table 1). Most platforms require that template DNA is short (200–1000 bp) and that each template contains a forward and reverse primer binding sites (i.e. a library of templates is needed). Libraries can be constructed in many different ways (see Cost per sample); an entire review on this subject alone is warranted. In the next section, I describe the most salient features of the platforms.

454 (http://www.454.com) was the 1st commercial NGS platform. 454 was acquired by Roche, but is still known as by the name 454. 454 uses beads that start with a single template molecule which is amplified via emPCR (Box 1). Millions of beads are loaded onto a picotitre plate designed so that each well can hold only a single bead. All beads are then sequenced in parallel by flowing pyrosequencing reagents across the plate.

Solexa (http://www.illumina.com) developed the 2nd commercial NGS platform. Solexa was subsequently acquired by Illumina and is now known by the name Illumina. Illumina uses a solid glass surface (similar to a microscope slide) to capture individual molecules and bridge PCR (Box 1) to amplify DNA into small clusters of identical molecules. These clusters are then sequenced with a strategy that is similar to Sanger sequencing, except only dye-labelled terminators are added, the sequence at that position is determined for all clusters, then the dye is cleaved and another round of dye-labelled terminators are added.

SOLiD (http://www.appliedbiosystems.com) was the 3rd commercial NGS platform. Invitrogen acquired Applied Biosystems, forming Life Technologies, but the name SOLiD has remained stable. SOLiD uses ligation to determine sequences and until the most recent release of Illumina's software and reagents, SOLiD has always had more reads (at lower cost) than Illumina.

Helicos (http://www.helicosbio.com) developed the HeliScope, which was the first commercial single-molecule sequencer. Unfortunately, the high cost of the instruments and short read lengths limited adoption of this platform. Helicos no longer sells instruments, but conducts sequencing via a service centre model.

Ion Torrent (http://www.iontorrent.com) uses a sequencing strategy similar to the 454, except that (i) hydrogen ions ($H^+$) are detected (instead of a pyrophosphatase cascade) and (ii) sequencing chips conform to common design and manufacturing standards used for

**Table 1** 2nd and 3rd Generation DNA sequencing platforms listed in the order of commercial availability

| Platform | Current company | Former company | Sequencing method | Amplification method | Claim to fame | Primary applications |
|---|---|---|---|---|---|---|
| 454 | Roche | 454 | Synthesis (pyrosequencing) | emPCR | First Next-Gen Sequencer, Long reads | **1***, **2**, 3*, 4, **7**, 8* |
| Illumina | Illumina | Solexa | Synthesis | BridgePCR | First short-read sequencer; current leader in advantages† | 1*, **2**, **3***, **4**, **5**, **6**, **7**, **8** |
| SOLiD | Life Technologies | Applied Biosystems | Ligation | emPCR | Second short-read sequencer; low error rates | 3*, **5**, **6**, **8** |
| HeliScope | Helicos | N/A | Synthesis | None | First single-molecule sequencer | **5**, 8 |
| Ion Torrent | Life Technologies | Ion Torrent | Synthesis ($H^+$ detection) | emPCR | First Post-light sequencer; first system <$100 000 | **1**, **2**, **3**, 4, **8** |
| PacBio | Pacific Biosciences | N/A | Synthesis | None | First real-time single-molecule sequencing | 1, 2, 3, 7, **8** |
| Starlight‡ | Life Technologies | N/A | Synthesis | None | Single-molecule sequencing with quantum dots | 1, 2, 7, **8** |

Bold indicates applications that are most often used, economical or growing.
1 = *de novo* BACs, plastids, microbial genomes.
2 = transcriptome characterization.
3 = targeted re-sequencing.
4 = *de novo* plant and animal genomes.
5 = re-sequencing and transcript counting.
6 = mutation detection.
7 = metagenomics.
8 = other (ChIP-Seq, μRNA-Seq, Methyl-Seq, etc.; see Brautigam & Gowik 2010, Shendure & Ji 2008).
*Pooling multiple samples with sequence tags (i.e. MIDs or indexes) is required for efficient use of this application
†Illumina currently leads in number and percentage of error-free reads, Illumina HiSeqs with v3 chemistry lead in reads per run, GB/run, and cost/GB.
‡A commercial launch date for the Starlight system is not yet known, but it is included here because it is in advanced development, and some information about its performance characteristics is known.

commercial microchips. Use of $H^+$ means that no lasers, cameras or fluorescent dyes are needed. Using common microchip design standards means that low-cost manufacturing can be used. Ion Torrent was purchased by Life Technologies in 2010, but is still known as Ion Torrent. The first early access instruments were deployed in late 2010.

PacBio (http://www.pacificbiosciences.com) has developed an instrument that sequences individual DNA molecules in real time. Individual DNA polymerases are attached to the surface of microscope slides. The sequence of individual DNA strands can be determined because each dNTP has a unique fluorescent label that is detected immediately prior to being cleaved off during synthesis. The first early access instruments were deployed in late 2010. The low cost per experiment, fast run times and cool factor have generated much enthusiasm for this platform, especially among investors.

Starlight uses quantum dots to achieve single-molecule sequencing. DNA is attached to the surface of a microscope slide where sequencing occurs in a manner similar to PacBio. A major advantage of Starlight relative to PacBio is that the DNA polymerase can be replaced after it has lost activity. Thus, sequencing can continue along the entire length of a template. Many characteristics of the Starlight technology are known (e.g. Karrow 2010), but timing of a commercial launch, target costs, etc. are unknown.

### *Broad characteristics*

The first three platforms (Table 1) are currently widely available through academic core laboratories and commercial service providers (see: http://pathogenomics.bham.ac.uk/hts/ for a hyperlinked global map of many NGS instruments; see http://seqanswers.com/forums/showthread.php?t=948/ for a list of NGS service providers; see Karrow & Toner 2011 for a recent survey). These three platforms have traditionally split their focus into fewer long reads (454) vs. more short reads (Illumina and SOLiD; see Box 1 for definitions). Long reads are optimal for initial genome and transcriptome characterization because longer pieces assemble more efficiently than shorter pieces. Alternatively, the lower costs and increased number of reads associated with shorter read-lengths are better suited for re-sequencing and for frequency-based applications (i.e. counting, such as in gene expression studies).

**Box 1** Glossary

**Barcode, index, MID or tag –** a short, unique sequence of DNA added to samples so they can be pooled, then processed and sequenced in parallel with each resulting sequence containing information to determine the source sample, used with some variance by all platforms.
**Bridge PCR –** PCR that occurs between primers bound to a surface, used by Illumina sequencers (see Shendure & Ji 2008, and references therein).
**cBot –** a required accessory instrument for many Illumina sequencers in which Bridge PCR is completed.
**de novo** – from the beginning (i.e. without prior information).
**emPCR or emulsion PCR –** PCR that occurs within aqueous microdroplets separated by oil so that up to thousands of independent reactions can occur per microlitre of volume; for NGS, one primer is usually covalently linked to a bead so PCR only occurs in microdroplets with beads, and a single template molecule per bead/microdroplet is needed, resulting in each bead having a homogeneous set of template molecules, used in 454, Ion Torrent, and SOLiD sequencers (see Shendure & Ji 2008, and references therein).
**Flow cell –** single-use sequencing chip/plate/slide used by Illumina sequencers (most use 8-channel flow cells; all channels must be used within a run); the SOLiD 5500 adopts a similar design, but channels may be run one at a time.

### Reads

**Mappable reads –** very short DNA sequences that can be determined to originate from a single location in the genome (~20–40 bases, length depends on genome complexity).
**Mate-paired reads –** DNA sequences from ends of DNA templates that have been circularized so that distant ends are physically ligated and read together (also known as Paired-end tags, PET or jump libraries; see Fig. 1a).
**Paired-end reads –** DNA sequences from each end of DNA templates (see Fig. 1c).
**Strobed reads –** DNA sequences determined at intermittent locations along the length of a single template; when illuminated the sequence is determined, when dark the polymerase continues at the same pace, but it is not degraded by the light. This is a way, for example, of spreading 900 bases of sequence data among three 300 base reads each separated by 300 bases (Fig. 1d).

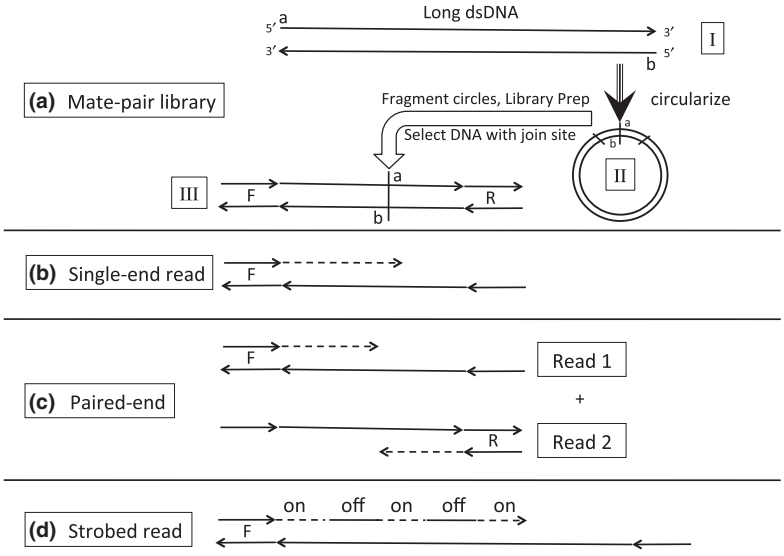


**Fig. 1** Illustration of the methods used for the four types of reads. Arrowheads indicate 3′ ends of DNA. F, forward primer; R, reverse primer. Double-stranded adapters of F plus its complement, and R plus its complement are added during the library construction phase for NGS. (a) Mate-pair libraries are constructed from fragments of double-stranded DNA (dsDNA) that are much longer than can be used directly for NGS libraries. In some embodiments, the join site may contain a linker that is used for selection purposes and to mark the join site. Following library construction, fragments are read using single- or paired-end reads. (b) Single-end reads yield data that are similar to Sanger sequences. (c) Paired-end reads allow both ends of a template to be sequenced. (d) Strobed reads spread the read length out along the template molecule by turning off the light source periodically, which allows synthesis to proceed at a known rate without photodegradation of the DNA polymerase. The data are used for the same purpose as mate-pair libraries.

No generally accepted standards exist for read length, but the following guidelines apply:
**Short reads –** sequences ≤50 consecutive bases.
**Mid-length reads –** sequences ≥51, but <400 consecutive bases.
**Long reads –** sequences≥400, but <1000 consecutive bases (i.e. similar to Sanger/capillary).
**Extended reads –** sequences >1000 bases; a small proportion of PacBio reads are up to a few kb; Starlight uses a replaceable polymerase allowing reads of indefinite length (up to the full length of the template).

**Computing**

**Cloud computing –** remote computational resources available (usually on a fee-for-use basis) via the internet [e.g. Amazon's Elastic Compute Cloud (http://aws.amazon.com/ec2)].
**Commodity alternatives/computing/resources –** computer parts and systems that conform to open standards and are thus available from many manufacturers and retailers (generally at low cost).
**Sneakernet –** transferring files by physically transporting hardware (i.e. carrying or shipping hard drives containing data).

The older NGS platforms have progressed significantly since they were first introduced. For example, 454 has progressed from reads of 100, to 250, to 400–500 bases, and is now on the verge of making 800-base reads available (mode = 800, average = 700). Illumina has progressed from reads of less than 36 bases to ≥100 bases on each end of templates, with SOLiD making slightly less striking increases. Thus, many of the platforms can be used for the same applications (Table 1) and such overlap is increasing.

Because it is possible to use most platforms for most applications, economics, length of time to data acquisition, length of time in the queue and downstream analysis constraints become important for selecting a platform. As the number and variety of instruments increase and costs continue to decrease, we will become constrained only by our knowledge of the systems and our creativity to develop and adapt techniques to obtain data efficiently. In particular, developments in sample multiplexing and sequence capture will drastically increase the amount of data available at affordable costs for molecular ecological studies.

### *Cost per run and cost per Mb*

Although all companies are continuously upgrading their platforms so that several fit into multiple read-length categories, the platforms can still be grouped into those that offer smaller numbers of middle-to-extended reads at relatively high cost per megabase (Mb) of sequence (i.e. 454, Ion Torrent, PacBio and Starlight) and those that offer larger numbers of short-to-middle-length reads at lower cost per Mb (i.e. Illumina, SOLiD, Helicose; Table 2). Technologies still in development (e.g. Oxford Nanopore, Roche+IBM, etc.) and expected updates to the current 3rd generation sequencing technologies (Karrow & Toner 2011) have the potential for many extended reads at low cost, but initial releases of the PacBio and Starlight platforms will not match the number of reads or cost per Mb of the short-read platforms (Table 2).

There is clearly a continuum of performance characteristics for massively parallel sequencers, with a reasonably strong dichotomy of these platforms in terms of the number of reads per run, cost per Mb and instrument time to conduct a run (Table 2). The variance in read lengths and supply costs per run are also important (Table 2). Because the read lengths of the Illumina sequencers can now equal or exceed 100 bases from each end of the template molecule, Illumina data can be used for *de novo* assemblies [e.g. Li *et al*. 2009 (but see Worley & Gibbs 2010); Paszkiewicz & Studholme 2010], especially when supplemented with mate-paired reads (Gnerre *et al*. 2011), and/or data from one of the longer-read platforms (e.g. Dalloul *et al*. 2010). Indeed, it is clear that the combination of Illumina or SOLiD data with-mate-paired reads on the 454 or Illumina, strobed reads from PacBio or extended reads from Starlight will facilitate many genome assemblies in the near future.

### *Cost per sample*

A major difference between the typical biomedical experiments targeted by NGS platforms and the uses for which molecular ecologists wish to employ these instruments is that the latter often want to process many samples (100s) at relatively modest numbers of loci (10s–1000s), and to do it with limited funds. A key to accomplishing low per-sample cost is to be able to attach an identifying tag (see Box 1) to each sample prior to expensive processing and sequencing. In this way, the cost of processing and sequencing can be divided among many samples.

All NGS platforms allow the use of sample tags. The importance of developing low-cost library preparations

DOCKET ALARM

# Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

## Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

## Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

## Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

## API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

**LAW FIRMS**
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

**FINANCIAL INSTITUTIONS**
Litigation and bankruptcy checks for companies and debtors.

**E-DISCOVERY AND LEGAL VENDORS**
Sync your system to PACER to automate legal marketing.