MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) doi: 10.1111/j.1755-0998.2011.03024.x

INVITED TECHNICAL REVIEW

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

© 2011 Blackwell Publishing Ltd

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

© 2011 Blackwell Publishing Ltd

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).

© 2011 Blackwell Publishing Ltd

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).

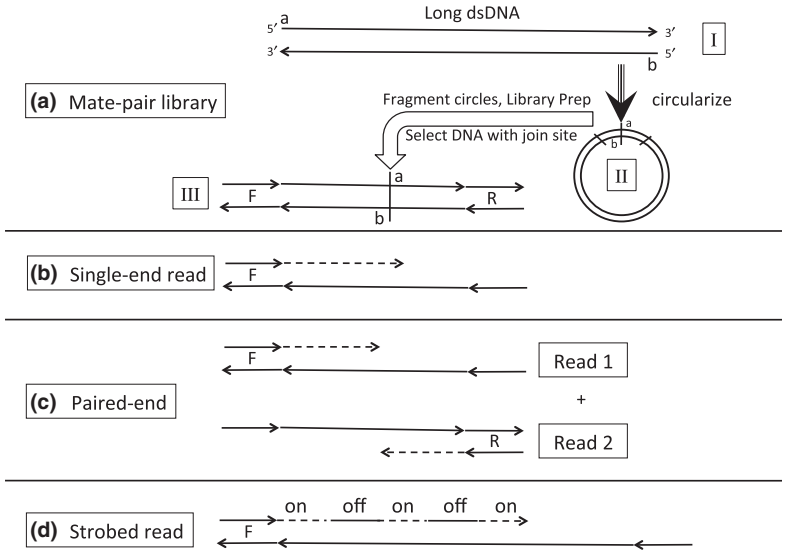


Fig. 1 Illustration of the methods used for the four types of reads. Arrowheads indicate 3′ ends of DNA. F, forward primer; R, reverse primer. Double-stranded adapters of F plus its complement, and R plus its complement are added during the library construction phase for NGS. (a) Mate-pair libraries 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.

© 2011 Blackwell Publishing Ltd

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 3rdgeneration 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

© 2011 Blackwell Publishing Ltd

Table 2 Comparison of sequencing instruments, sorted by cost/Mb, with expected performance by mid 2011

Instrument	Run time[a]	Millions of reads/run	Bases/read[b]	Yield Mb/run	Reagent cost/run[c]	Reagent cost/Mb	Minimum unit cost (% run)[d]
3730xl (capillary)	2 h	0.000096	650	0.06	$96	$1500	$6 (1%)
Ion Torrent – '314'chip	2 h	0.10	100	>10	$500	<$50	~$750 (100%)
454 GS Jr. Titanium	10 h	0.10	400	50	$1100	$22	$1500 (100%)
Starlight*	†	~0.01	>1000	†	†	†	†
PacBio RS	0.5–2 h	0.01	860–1100	5–10	$110–900	$11–180	†
454 FLX Titanium	10 h	1	400	500	$6200	$12.4	$2000 (10%)
454 FLX+[e]	18–20 h	1	700	900	$6200	$7	$2000 (10%)
Ion Torrent – '316'chip*	2 h	1	>100	>100	$750	<$7.5	~$1000 (100%)
Helicos[f]	N/A	800	35	28 000	N/A	NA	$1100 (2%)
Ion Torrent – '318'chip*	2 h	4–8	>100	>1000	~$925	~$0.93	~$1200 (100%)
Illumina MiSeq*	26 h	3.4	150 + 150	1020	$750	$0.74	~$1000 (100%)
Illumina iScanSQ	8 days	250	100 + 100	50 000	$10 220	$0.20	$3000 (14%)
Illumina GAIIx	14 days	320	150 + 150	96 000	$11 524	$0.12	$3200 (14%)
SOLiD – 4	12 days	>840[g]	50 + 35	71 400	$8128	<$0.11	$2500 (12%)
Illumina HiSeq 1000	8 days	500	100 + 100	100 000	$10 220	$0.10	$3000 (12%)
Illumina HiSeq 2000	8 days	1000	100 + 100	200 000	$20 120[h]	$0.10	$3000 (6%)
SOLiD – 5500 (PI)*	8 days	>700[g]	75 + 35	77 000	$6101	<$0.08	$2000 (12%)
SOLiD – 5500xl (4hq)*	8 days	>1410[g]	75 + 35	155 100	$10 503[h]	<$0.07	$2000 (12%)
Illumina HiSeq 2000 – v3[i]*	10 days	≤3000	100 + 100	≤600 000	$23 470[h]	≥$0.04	~$3500 (6%)

[a]Instrument time for maximum read length.
[b]Average length for high-quality reads >200 bases (mode is higher); typical maximum for reads ≤150 bases (most reads reach this length).
[c]Includes all stages of sample preparation for a single sample (i.e. library preparation through sequencing; capillary = sequencing only).
[d]Typical full cost (i.e. including labour, service contract, etc.) of the smallest generally available unit of purchase at an academic core laboratory provider for the longest available read (and percentage of reads relative to a full run, rounded to the nearest whole percentage).
[e]Upgrade of the FLX instrument, due out summer 2011.
[f]Instruments and reagents are no longer sold; services are available for any organism.
[g]Mappable reads [number of raw high-quality reads (as reported for all other platforms) is higher].
[h]More reads are obtained than is needed from any single sample within most experiments, but the value illustrates the costs.
[i]AnnouncedTruSeq v3 reagents & software, reads and yield are half for HiSeq1000.
*Information based on company sources alone (independent data not yet available).
†Detail not yet available.
'~' Indicates a likely value based on unpublished information available in March 2011 (i.e. author speculation).

and sample tagging has been clear for several years, during which time a variety of schemes have been developed (e.g. Binladen *et al.* 2007; Hoffmann *et al.* 2007; Meyer *et al.* 2007; Craig *et al.* 2008). Molecular ecologists (among others) will benefit from further development of low-cost library preparations in which many different sample tags are employed, thereby facilitating many samples being pooled together and thus dividing sequencing costs among many samples.

To understand the importance of library construction costs, consider that when the Illumina HiSeq was introduced in early 2010, standard Illumina RNA-Seq libraries cost about $400 each to construct. Researchers could pool 12 indexed samples per lane and yield about 5 million reads per sample (a sufficient number of reads for gene expression studies seeking modest sensitivity). Sequencing reagents for 192 samples were estimated at <$7000, but library preparations were estimated at 192 × $400 = $76 800 (i.e. library preparation was ten times more expensive than sequencing the libraries). Thus, library preparation and sample tagging continues to be an active area of research with important ongoing developments.

Purchase costs

NGS platforms currently range from $49 500 to $695 000 (Table 3). Additional ancillary equipment, extended service contracts and required computers extend the costs of most systems from about $75 000 to more than $1 000 000. Costs in Table 3 assume that equipment will be housed in an equipped, fully functioning laboratory. If the new sequencer will be placed into a new facility, then the costs will increase considerably.

There are now three instruments that are <$150 000 (Table 3), making them within the reach of many indi-

© 2011 Blackwell Publishing Ltd

Table 3 Instrument purchase cost, additional instrument costs, service agreement costs, computational resources needed, size of data files, primary errors and error rates for commercially available DNA sequencing platforms in 2011. All costs are list price in thousands of US dollars

Instrument	Purchase cost	Additional instruments[a]	Service contract[b]	Computational resources[c]	Data file sizes (GB)[d]	Primary errors	Error rate (%)[e]
3730xl (capillary)	$376	–	$19.8	Desktop	0.03	Substitution	0.1–1
454 GS Jr. Titanium	$108	$16	$12.6	$5 (desktop)	<3 images, <1 sff	Indel	1
454 FLX Titanium	$500	$30	$50.0	$5 (desktop)	20 images, 4 sff	Indel	1
454 FLX+[f]	$29.5	$30	$50.0	$5 (desktop)	~40 images, 8 sff	Indel	1*
PacBio RS	$695	–	$85	$65 cluster	20 pulsed, 2 Fastq	CG deletions	16
Ion Torrent – 314 chip	$49.5	$18[g]	$7.5	Desktop – $35	0.1Fastq	Indel	~1
Ion Torrent – 316 chip	$49.5	$18[g]	$7.5	Desktop – $35	0.6Fastq	Indel	~1*
Ion Torrent – 318 chip	$49.5	$18[g]	$7.5	Desktop – $35	TBD	Indel	~1*
SOLiD – 4	$475	$54[h]	$38.4	$35 cluster[i]	680[j]	A-T bias	>0.06*
SOLiD – 5500	$349	$54[h]	$29.0	$35 cluster[i]	74[k*]	A-T bias	>0.01*
SOLiD – 5500xl	$595	$54[h]	$38.4	$35 cluster[i]	148[k*]	A-T bias	>0.01*
Illumina MiSeq	$125	–	$12.5	Desktop	1[k*]	~Substitution	>0.1*
Illumina HiScanSQ	$405	$55[l]	$41.5	$222 cluster[m]	50[k*]	Substitution	≥0.1
Illumina GAIIx	$250	$100[n]	$44.5	$222 cluster[m]	600	Substitution	≥0.1
Illumina HiSeq1000	$560[o]	$55[l]	$62.0	$222 cluster[m]	≤300[k*]	Substitution	≥0.1
Illumina HiSeq2000	$690	$55[l]	$75.9	$222 cluster[m]	≤600[k*]	Substitution	≥0.1

[a]Does not include general purpose and library preparation equipment (e.g. Covaris [$45k], Agilent bioanalyzer [$18k], thermal cyclers, general purpose centrifuges, MilliQ water, etc.), but includes bead counters for emPCR (up to $20k), TissueLysers or similar for emPCR (up to $10k), specialty centrifuges, etc. when required by the instrument manufacturer. Many laboratories will need additional general purpose instruments.
[b]Annual maintenance agreements include on-site service, but do not include extra premiums for the fastest available service.
[c]Desktops assume higher-end models with multiple processors, ≥8 GB RAM, ≥1 TB HD, etc. (up to $5k; except capillary = $2k desktop).
[d]Data file size transferred from instrument server to offline cluster.
[e]Percentage of errors per base within single reads of the maximum length given in Table 2; rates among platforms are not exactly comparable; reported Ion Torrent rates range from 0.46% to 2.4%; SOLiD rates are from reads with bases consistent on double or triple sequencing only; for Illumina, the 0.1% rate applies to > 85% of reads (not all reads); see text for additional details.
[f]Upgrade to the 454 FLX instrument; FLX+ new purchases = $500k.
[g]Required $16.5k IonTorrent server for conversion of raw signals to basecalling; $1k ULTRA-TURRAX® Tube Drive; Argon gas tank (<$0.5k).
[h]Includes EZ Bead ePCR automation and required UPS; a Covaris ($45K) is also required but is not included to facilitate comparisons (because one is usually bought with any of the other sequencing systems).
[i]Compute cluster available from Life Technologies.
[j]85Gb run is a 2 × 50 run (the highest throughput on a SOLiD 4).
[k]New compressed binary data format saves base and quality-value data in a 1byte:1base ratio.
[l]Cost of cBot (required). Additional instruments for library preparation needed (Covaris, etc., similar to 454 FLX).
[m]Illumina Compute – Tier 1 system: 3 cluster nodes with 8 cores and 48 GB RAM per node (i.e. 24 cores and 144 GB RAM) and ~24 TB usable data storage; commodity equivalent systems are available for much less, but will require technical support (see text).
[n]Cost of cBot and Paired-end module (required for GAIIx).
[o]HiSeq 1000 is upgradable to HiSeq 2000 for $175k.
'~' Indicates a likely value based on unpublished information available in February 2011 (i.e. author speculation).
*Information based on company sources alone (independent data not yet available); also applies to Illumina TruSeq v3 chemistry.

vidual researchers. These instruments have significantly reduced total costs per run and/or experiment. The smaller footprint of these instruments and potential portability of at least some are also attractive features. Although these features will facilitate new research opportunities, researchers should be careful to weigh the significantly increased cost per read and cost per Mb of these instruments relative to other instruments and to the costs of outsourcing. It is obvious that these lower-cost instruments and low-cost runs will be invaluable in small-scale experiments, gathering pilot data, quality control and validation. Researchers with these instruments will, however, still often find it economically most advantageous to send out fully processed and validated samples to be run on other lower-cost per read/Mb instruments (e.g. library pools with many indexes could be validated on the MiSeq/Ion Torrent 314/GS Junior, but then

© 2011 Blackwell Publishing Ltd

Table 4 Primary advantages and disadvantages of each next-generation sequencing instrument

Instrument	Primary advantages	Primary disadvantages
3730xl (capillary)	Low cost for very small studies	Very high cost for large amounts of data
454 GS Jr. Titanium	Long-read length; low capital cost; low cost per experiment	High cost per Mb
454 FLX Titanium	Long-read length	High capital cost and high cost per Mb
454 FLX+	Double the maximum read length of Titanium	High cost per Mb
Helicos	Large numbers of reads directly from single molecules	Length of reads and questionable longevity of company
PacBio	Single molecule real-time sequencing, longest available read length, strobed reads, each instrument run = min, low cost per sample and many methods being developed	Error rates, low total number of reads per run, high cost per Mb, high capital cost, and many methods still in development
Ion Torrent	Low-cost instrument upgraded through dispos able chips (the chip is the machine), very simple machine with few moving parts and clear trajectory to improved performance	New platform with a variety of unknowns, and some known issues at the time of release
Ion Torrent – 314 chip	Low cost per sample for small studies, short time needed on instrument, suitable for microbial sequencing and targeted sequencing, and easily upgraded with new chips	Highest cost per Mb of all NextGen platforms and sample preparation takes longer time than on the instrument
Ion Torrent – 316 chip	Same as above, upgraded because of higher density chip	Sample preparation time and similar cost per Mb to 454
Ion Torrent – 318 chip	Same as above, upgraded because of higher density chip, lower cost per read and Mb allows more applications	Sample preparation time and similar cost to MiSeq
SOLiD – 4	EZ Bead simplifies emPCR, low-cost per Gb, throughput = 5–6 Gb/day	Unusual informatics with 2-base colour space encoding, relatively short reads and chip runs all at once
SOLiD – 5500	Each lane of Flow-Chip can be run independently, highest accuracy*, output in bases (not colour space); ability to rescue failed sequencing cycles, 96 validated barcodes per lane and throughput of 10–15 Gb/day	Not available until spring 2011, relatively short reads, more gaps in assemblies than Illumina data and less even data distribution than Illumina
SOLiD – 5500xl	Same as 5500, but with double the throughput	Same as SOLiD 5500 and high capital cost
Illumina MiSeq	Low-cost instrument and runs, lowest cost/Mb for small platforms and fastest Illumina run times	Relatively few reads and higher cost/Mb compared to other Illumina platforms
Illumina HiScanSQ	Versatile instrument for full catalogue of Illumina arrays and sequencing, and scalable in future	Higher cost/Mb than HiSeq for large amounts of data
Illumina GAIIx	Lower capital cost than HiSeqs	Slightly higher cost per Mb than HiSeq and not as scalable in the future
Illumina HiSeq 1000	Lower instrument cost than HiSeq 2000, same number of reads/lane and cost/lane as HiSeq 2000, field upgradable to HiSeq 2000 and future scalability	Not as flexible as HiSeq2000 because of having only 1 flow cell
Illumina HiSeq 2000	Same as HiSeq 1000, but runs two flow cells simultaneously; Most reads, Gb per day and Gb per run, lowest cost per Mb of all platforms*	High capital cost and high computation needs

Mb, megabase; Gb, gigabase.
*Information based on company sources alone (independent data not yet available).

sequenced to greater depth on the HiSeq/Ion Torrent 318/454 FLX).

Computational resources for analysis

The computational resources needed to process and analyse data from each platform vary tremendously (Table 3). Because short reads require more intensive analysis, and many platforms deliver very large numbers of short reads, the computational resources needed to use short-read data can be considerable. Several platforms (454, Ion Torrent, MiSeq), however, deliver relatively little sequence data (i.e. ≤5 Gb of sequence data), which can be analysed on higher-end desktop

© 2011 Blackwell Publishing Ltd

computers. Some analysis software runs only on Linux-based operating systems (OS), which are less common than Windows or Mac OS. Rather than dedicating an entire high-end desktop to DNA analyses in Linux, virtual machine software available commercially (e.g. Fusion or Parallels) or for free (e.g. http://www.virtualbox.org/) allows researchers to install Linux on Windows or Mac OS machines.

Computer resources can be modest for many analyses. For example, assembly of an *E. coli* genome with >600 000 reads from a half 454 titanium run requires less than 13 min on a dual quad-core MacPro purchased in early 2008, using the Roche gsAssembler software in a Linux virtual machine. The MacPro assembly time is about 30 s longer than the same analyses computed on a $30 000 computer cluster purchased with the 454 in mid-2009. A MacBook Pro purchased in early 2010 runs the same analysis in about twice the time. The virtual machines require only about 2 GB of RAM to perform these analyses (thus a total of 6 GB RAM is sufficient). This illustrates why Roche no longer sells an off-machine analysis cluster with the 454 and how many users can conduct analyses on high-quality desktops or laptops (see also Illumina, Inc. 2010).

For larger amounts of data, it is necessary to use high-performance computational clusters (HPCCs). The HPCCs from Illumina (Table 3) are much more extensive than the computational resources recommended by other companies. Illumina is recommending resources that enable full analysis of their data at the location of data generation, whereas other companies often assume that data are shared with users at remote locations who have access to shared HPCCs that could be used for large-scale analysis. Institutions lacking HPCCs can purchase commodity (i.e. generic) HPCCs at much lower cost than those sold by or through most sequencing companies. The downside to commodity resources is that they are not plug-and-play (i.e. someone has to put them together from their component parts, install the software, etc.). Thus, there can be considerable investments needed in information technology (IT) support for commodity HPCCs.

It is possible to reduce IT headaches and avoid purchasing HPCCs by renting computation time and data storage space on commercially available HPCCs (i.e. clouds such as Amazon EC2, Google AppEngine, or Rackspace; Pennisi 2011). However, the files are often so large that they take excruciatingly long to transfer (hours to days). Indeed, file transfers can take longer than shipping hard drives so that the drives can be physically plugged into local or cloud resources (i.e. sneakerNet can be faster than the internet). Storage costs can also be considerable (often costing much more than the CPUtime needed for analysis). Cloud resource providers clearly recognize these limitations and are working to overcome these constraints. Much work remains, however, in modifying existing workflows for cloud computation (Pennisi 2011). One possible development will be to move analyses to hubs of data rather than downloading data to where analyses are performed (Kahn 2011).

Error rates

Direct comparisons of error rates are particularly problematic. For example, most companies report errors based on sequence reads of particular templates that are favourable for their platform, generally the same template used for quality control (e.g. *E. coli*, Phi X, etc.). Errors increase near the end of each platform's maximum read length, except for PacBio. Indeed, maximum read length is limited by error tolerance (i.e. if one is willing to accept more errors, then longer reads are possible on all platforms). As noted earlier, NIST standards would be helpful for future comparisons.

Although the error rates of different platforms are not exactly comparable for a variety of reasons, and thus should not be taken too literally, Table 3 shows reasonable approximations that can be used to compare the platforms. The average error rate per base of the platforms varies by more than three orders of magnitude, ranging from 0.01% to 16% (Table 3). The SOLiD system has the lowest error rate of data accessible to users, whereas PacBio has the highest. The SOLiD actually has the second highest raw error rate, but uses double-or triple-encoding of each base (i.e. each base is sequenced independently two or three times, with inconsistent data becoming inaccessible to users) to achieve its low error rate. Similarly, PacBio suggests reading each template multiple times to overcome high per-read error rate to achieve a consensus sequence with low error (<0.1%), but PacBio gives users the option of obtaining single-pass data. Like the SOLiD, Illumina also touts error rates based on less than all of the data. Error rates of the Ion Torrent have minimal independent verification at this time and appear to be similar to the 454. In addition to raw error rates of sequencing reads, each platform has biases, such as evenness of coverage and percentage of perfect reads. Platform bias during library construction, amplification and sequencing strongly affects the utility of sequence data obtained for specific downstream applications. Thus, users need to consider a variety of variables in addition to reported error rates.

Platform choice – pick your poison

Each platform has significant advantages and disadvantages (Table 4). As mentioned previously, many platforms can be used for many tasks. No single platform,

© 2011 Blackwell Publishing Ltd

however, can do everything that users will want and do it well or economically. Overall, Illumina has the broadest utility and lowest cost per read and Mb, whereas the longer reads, more limited computational needs, and mature software from Roche/454 will continue to make 454 attractive. Ion Torrent is seeking to displace 454, but 454 is working on their own postlight pyrosequencing instrument (the DNAe). As read length grows and the new 5500s become available, the SOLiD will become attractive for more ecological work.

In addition to the information given in Tables 1 and 4, there are significant advantages and disadvantages of approaches underlying several of the platforms. For example, the 454, Ion Torrent and SOLiD systems use emulsion PCR (emPCR) to amplify templates. An advantage of emPCR is that it is a separate process from sequencing, so that if the emPCR fails to yield enough beads, it can be repeated before loading the beads and doing the sequencing. However, emPCR reagents are expensive, costing 10s–100s of dollars per reaction, depending on the scale. Additionally, the process of emPCR has consistency issues similar to those of Randomly Amplified Polymorphic DNA (RAPDs), which are notorious for problems of replicability within the molecular ecology community. Thus, emPCR suffers from high training costs, high template quantification costs, and high failure rates, all of which significantly increase the time and cost of processing samples. Alternatively, bridge PCR, used by Illumina, happens on the sequencing chip, which is more efficient, but it is a major portion of overall sequencing cost. So, if bridge PCR (also known as cluster formation) goes poorly for one or a few lanes, a major part of the sequencing cost is also lost: you either waste expensive sequencing reagents on poorly performing lanes or you toss out the expensive flow cell and start again. Thus, emPCR and bridgePCR both have significant downsides.

There is an ever-growing body of software that can be used for analyses. Illumina, Inc. (2010) compiled a very readable technical bulletin on practical impacts of read depth, read length, software and hardware on *de novo* assembly. Several recent reviews on alignment and assembly examine a large array of software tools (Flick & Birney 2009; Horner *et al*. 2009).

Because the instruments are changing so rapidly, expert opinion is generally the best option for choosing a platform in which to conduct an experiment. Unfortunately, most experts have biases because of their experiences, knowledge of specific instruments and analysis techniques, as well as financial conflicts of interest. Much like seeing a specialist physician, it is advisable to ask questions, read widely and seek a second opinion.

In general, all of the massively parallel platforms have some pain that purchasers and users will feel. This pain can be expressed in many forms, including reagent costs, wait time, instrument failure, failed library preparations, suboptimal numbers of reads or read length, or difficulty in analyses. The trick is to find systems for which the pain delivered best matches your tolerance, problem solving and coping skills.

Summary

It is an exciting time to be a molecular ecologist. We are gaining access to tools that are opening new avenues of investigation and developing techniques that promise to yield answers to long-standing ecological and evolutionary questions. In addition to focusing on our own research questions, we should also invest in developing knowledge and techniques that allow us to more wisely use these incredible tools – so that we can maximize our insight into ecological and evolutionary systems while most efficiently using the funding resources entrusted to us. I hope that the information contained in this guide will serve as a starting place for increased clarity about the similarities and differences among NGS platforms. Updates to the tables presented here will be posted at http://tomato.biol.trinity.edu/blog/next-gen-reagent-costs/, and additional information will be at http://dna.uga.edu/

Acknowledgements

I thank Brant Faircloth, Kevin Winker, Nick Crawford, all of my colleagues at the Georgia Genomics Facility, and the excellent sales force and scientific support staff of the companies who supplied much of the raw data used herein: James Hudson and Scott Taylor, Illumina; Jeffrey Smith, Ion Torrent; Jeff Robinson, Pacific Biosciences; Chad Alexander and Dale Yuzuki, Life Technologies; and Carl Woodham, Roche. Three anonymous reviewers facilitated tremendous improvements to the manuscript by making excellent suggestions – thank you. Partially supported by DEB-0614208.

Conflicts of interest

The author serves as faculty director of the Georgia Genomics Facility, a facility that owns a 454 and facilitates NGS via outsourcing for researchers at the University of Georgia.

References

Ansorge WJ (2009) Next-generation DNA sequencing techniques. *New Biotechnology*, **25**, 195–203.

Binladen J, Gilbert MTP, Bollback JP *et al*. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE*, **2**, e197.

© 2011 Blackwell Publishing Ltd

Bräutigam A, Gowik U (2010) What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. *Plant Biology*, **12**, 831–841.

Craig DW, Pearson JV, Szelinger S *et al.* (2008) Identification of genetic variants using bar-coded multiplex sequencing. *Nature Methods*, **5**, 887–893.

Dalloul RA, Long JA, Zimin AV *et al.* (2010) Multi-platform next-generation sequencing of the domestic turkey (*Meleagrisgallopavo*): genome assembly and analysis. *PLoS Biology*, **8**, e1000475.

Flick P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, **6**(11 Suppl.), S6–S12.

Gnerre S, MacCallum I, Przybylski D *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences USA*, **108**, 1513–1518.

Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, Bushman FD (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Research*, **35**, e91.

Horner DS, Pavesi G, Castrignano T *et al.* (2009) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics*, **11**, 181–197.

Illumina, Inc. (2010) *De Novo Assembly Using Illumina Reads. Technical Note: Sequencing*. Available at: http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly_ecoli.pdf, last accessed 28 Feb 2011.

Kahn SD (2011) On the future of genomic data. *Science*, **331**, 728–729.

Karrow J (2010) *Life Tech Details Real-Time Single-Molecule Tech at AGBT; Combines Qdots with FRET-Based Detection*. Genome Web, March 2, 2010. http://www.genomeweb.com/sequencing/life-tech-details-real-time-single-molecule-tech-agbt-combines-qdots-fret-based, last accessed 1 March 2011.

Karrow J, Toner B (2011) *In Sequence Annual Survey: Illumina Leads Market but Most Users Believe PacBio will Provide the Next Big Leap*. Genome Web, January 25, 2011. Available at: http://www.genomeweb.com/sequencing/sequence-annual-survey-illumina-leads-market-most-users-believe-pacbio-will-prov, last accessed 1 March 2011.

Li R, Fan W, Tian G et al. (2009) The sequence and *de novo* assembly of the giant panda genome. *Nature*, **463**, 311–317.

Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, **9**, 387–402.

Meyer M, Stenzel U, Myles S, Prüfer K, Hofreiter M (2007) Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Research*, **35**, e97.

Paszkiewicz K, Studholme DJ (2010) De novo assembly of short sequence reads. *Briefings in Bioinformatics*, **11**, 457–472.

Pennisi E (2011) Will computers crash genomics? *Science*, **331**, 666–668.

Richardson P (2010) Special issue: next generation DNA sequencing. *Genes*, **2010**, 385–387.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.

Tautz D, Ellegren H, Weigel D (2010) Next generation molecular ecology. *Molecular Ecology*, **19**(Suppl. 1), 1–3.

Worley KC, Gibbs RA (2010) Genetics: decoding a national treasure. *Nature*, **463**, 303–304.

Travis Glenn develops DNA tools, conducts research in environmental genomics and environmental health (often using novel biomedical model organisms), teaches, collaborates widely and advises researchers on use of genetic and genomic tools.

© 2011 Blackwell Publishing Ltd