The new paradigm of flow cell sequencing

Robert A. Holt and Steven J.M. Jones

Genome Res. 2008 18: 839-846
Access the most recent version at doi:10.1101/gr.073262.107

The new paradigm of flow cell sequencing

Robert A. Holt[1] and Steven J.M. Jones

British Columbia Cancer Agency, Genome Sciences Centre, Vancouver, British Columbia V5Z 4E6, Canada

DNA sequencing is in a period of rapid change, in which capillary sequencing is no longer the technology of choice for most ultra-high-throughput applications. A new generation of instruments that utilize primed synthesis in flow cells to obtain, simultaneously, the sequence of millions of different DNA templates has changed the field. We compare and contrast these new sequencing platforms in terms of stage of development, instrument configuration, template format, sequencing chemistry, throughput capability, operating cost, data handling issues, and error models. While these platforms outperform capillary instruments in terms of bases per day and cost per base, the short length of sequence reads obtained from most instruments and the limited number of samples that can be run simultaneously imposes some practical constraints on sequencing applications. However, recently developed methods for paired-end sequencing and for array-based direct selection of desired templates from complex mixtures extend the utility of these platforms for genome analysis. Given the ever increasing demand for DNA sequence information, we can expect continuous improvement of this new generation of instruments and their eventual replacement by even more powerful technology.

Since the establishment of DNA as hereditary material and the elucidation of its structure, there has been insatiable demand for sequence information and remarkable innovation in the methods used to obtain it. Like many technologies, DNA sequencing has advanced by punctuated equilibrium, where a new approach to sequencing is introduced, adopted, and improved upon incrementally for some period of time, then replaced by the next wave. The very earliest sequencing techniques involved variations on the theme of cleavage of short polynucleotides and subsequent identification by their migration characteristics using two-dimensional paper chromatography. Using this approach it was possible to infer short sequences, such as that of the *Escherichia coli* lac operon (Gilbert and Maxam 1973), and it was feasible at the time to report the data from an entire sequencing project in a paper's abstract. A transition of major significance was spearheaded by the Sanger group in the mid 1970s, when they introduced the notion of using primed template replication by polymerase and separation of the extension products by gel electrophoresis (Sanger and Coulson 1975) to obtain DNA sequence information. Modifying this approach to allow base-specific chain termination by di-deoxy nucleotides (Sanger et al. 1977) laid the foundation for sequencing for the next 30 yr. Further incremental improvements during this time included using fluorescent rather than radiolabeled terminators, separation on acrylamide matrices in capillaries rather than slab gels, and, ultimately, the deployment of mechanized production lines for template preparation and devices for automated generation and reading of sequence ladders. This industrial approach to sequencing spawned the modern era of genomics and has provided an archive of complete reference genome sequences. Yet demand for DNA sequence is undiminished and we find ourselves in a new period of rapid change. If the hallmark of the past paradigm was electrophoretic separation of terminated DNA chains, then the hallmark of new paradigm is flow cell sequencing, with stepwise determination of DNA sequence by iterative cycles of nucleotide extensions done in parallel on massive numbers of clonally amplified template molecules. If one takes the broad view of a flow cell as a reaction chamber that contains template tethered to a solid support, to which nucleotides and ancillary reagents are iteratively applied and washed away, then the new instruments on the market (the Roche GS-FLX, the Illumina 1G analyzer, and the Applied Biosystems SOLiD) are all flow cell sequencers (as are instruments anticipated in the near future such as the Helicos HeliScope and the Danaher Polonator). Massively parallel approaches using flow cells allow DNA to be sequenced markedly faster and cheaper than ever before. This means that lines of scientific inquiry that once were prohibitively expensive are now feasible, and this is good because there is much to explore. For example, human genome sequences have been compiled but represent a miniscule proportion of the ~100 million kilograms of human DNA that is on the planet on any given day. It is certain that novel template from the biosphere will continue to drive consecutive waves of innovation in sequencing technology for some time to come.

[1]Corresponding author.
E-mail rholt@bcgsc.ca; fax (604) 877-6085.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.073262.107.

The technology

Templates and sequencing chemistries

While all of the latest commercial sequencing instruments use flow cells and massive parallelization to increase sequencing capacity, the specifics of template preparation, sequencing chemistry, and flow cell configuration differ among the platforms. There is often a misconception that the new generation of sequencers perform sequencing on single molecules. In fact, all currently available platforms (the Roche GS-FLX, Illumina 1G analyzer, and the Applied Biosystems SOLiD) require PCR-based amplification of fragmented template DNA to obtain sufficient signal for base calling. However, these methods utilize a single DNA molecule as the initial substrate for amplification allowing each sequenced molecule to represent a single haplotype. This has proven to be useful for robust polymorphism detection particularly in cancer-derived material, where associated normal tissue may obscure heterozygote calls using traditional Sanger sequencing of PCR products. As discussed further below, the instrument being developed by Helicos stays with the single molecule throughout analysis.

The premiere of flow cell sequencing was the GS20 (454 Life Sciences), which in a single machine run provided shotgun sequence data for de-novo assembly of the *Mycoplasma genitalium* genome with 96% coverage at 99.96% accuracy (Margulies et al. 2005). The current model of this instrument is the GS-FLX marketed by Roche Applied Science, but the core technology is the same, and the method is still referred to as 454 sequencing. The 454 flow cell supports a "picotiter" plate, a fiber optic slide with ~1.6 million 75-picoliter wells. For 454 sequencing, template is amplified, as follows, by emulsion PCR. Using limiting dilution, each individual molecule of sheared template DNA is captured on a separate bead, and each bead is compartmentalized in a private droplet of aqueous PCR reaction mixture within an oil emulsion. Template is clonally amplified on the bead surface by thermocycling, and the template-loaded beads are then distributed into the wells of the picotiter plate, ideally with one or fewer beads per well. Sequence is obtained by iterative pyrosequencing (Nyren et al. 1993; Ronaghi et al. 1996, 1998), whereby wells are loaded once with bead-tethered sequencing enzymes (polymerase, sulfurylase, and luciferase), and buffer containing one of four dNTPs is passed horizontally over the wells. If there is a match to the primed template, polymerase incorporates the nucleotide and releases a pyrophosphate molecule which, when converted to ATP by sulfurylase, generates a luciferase-catalyzed luminometric signal. After washout of residual nucleotides, the cycle is repeated with the next dNTP. Because dNTPs are used, homopolymers in the template DNA present a potential problem for this sequencing chemistry, as they allow the incorporation of multiple nucleotides in a single flow. The increase in observed luminescence allows the length of a homopolymer to be estimated, but the ability to discriminate decreases with the length of the homopolymer. Discrimination is reliable for up to four bases, and homopolymers approaching eight bases are typically irresolvable (Margulies et al. 2005; Huse et al. 2007), although this is continuously improving through better software and decreased cross-talk among wells. Further issues with the 454 approach are failed wells due to incomplete extension of a homopolymer, the misincorporation of excess nucleotides that are not completely washed away after a previous cycle, beads with mixed templates, and multiple copies of the same template on different beads. While 454's pyrosequencing approach appears at first to be somewhat complex and indirect, and although homopolymers, carry-forward, and phasing effects can be problematic, over 100 studies using this technology have been published and are a testimonial to its utility and robustness.

After the 454 system, the next platform on the market was the 1G Analyzer developed by Solexa, and now owned and marketed by Illumina (www.illumina.com). This is the first of the massively parallel short-read platforms. The Illumina flow cell is a planar optically transparent surface similar to a microscope slide, which contains a lawn of oligonucleotide anchors bound to its surface. To prepare template DNA, adapters complimentary to oligos on the flow cell surface are ligated to the ends of size-selected DNA. Adapted single-stranded DNAs are bound to the flow cell and amplified, as follows, by solid-phase "bridge" PCR. In each PCR cycle, priming occurs by arching of the template molecule such that the adapter at its untethered end hybridizes to and is primed by a free oligo in the near vicinity on the flow cell surface. This process results in a raindrop pattern of clonally amplified templates. Sequencing proceeds by synthesis using reversible four-color fluorescence (e.g., a mix of the four bases each labeled with a different cleavable fluorophore, such that they can be used simultaneously rather than sequentially to interrogate a given nucleotide position in the template). Labeled terminators, primer, and polymerase are applied to the flow cell. After base extension and recording of the fluorescent signal at each cluster, the sequencing reagents are washed away, labels are cleaved, and the 3′ end of the incorporated base is unblocked in preparation for the next nucleotide addition. The key innovations of this system are in-situ template amplification and four-color Sanger-like, but reversible, terminators. A new version of the Illumina instrument (the GAII), with improved optics capable of handling higher cluster densities, is anticipated.

Of the platforms that are presently commercially available, the latest addition is the SOLiD (Supported Oligonucleotide Ligation and Detection) instrument from Applied Biosystems. Certain elements of the platform are directly analogous to features of both the 454 and Illumina systems. As with the 454 system, template amplification is by emulsion PCR, and as with the Illumina system, template is applied at high density to a flow cell. There is no separation of the bead from template, but rather, after a step to cull failed beads, template is deposited on the flow cell, bead and all. The key distinguishing feature of the SOLiD platform is, as the name suggests, the ligation-based sequencing chemistry. The ligation approach is based on early work characterizing thermostable ligase (Barany and Gelfand 1991; Housby and Southern 1998) and subsequent implementation of high-throughput sequencing protocol by the Church group (Shendure et al. 2005). In this seminal study, emulsion PCR-amplified DNA from *E. coli* MG1655 was subjected to 26 cycles of sequencing by ligation to generate 1.16 million mappable reads. Combining these reads with data from a separate run provided coverage of 91.4% of the *E. coli* genome. Of interest, this effort by the Church laboratory used off-the-shelf instrumentation and reagents, and the package is being developed for commercial distribution by Danaher Motion as the "Polonator" sequencing platform.

In contrast to polymerase-based sequencing, in sequencing by ligation, bases are inferred indirectly based on successful ligation events. As implemented on the SOLiD platform, a primer complementary to the adapter sequence at the template/bead junction provides a 5′ phosphate group, to which four-color dye-labeled oligos compete for ligation. Each oligo has three universal bases then two fixed bases, such that the actual bases being interrogated are offset by three nucleotides from the point of ligation. Further, the oligos are ligation terminators, and only after cleaving the fluorescent label, clearing the flow cell, and ligating with fresh reagents can additional bases be inferred. In this second round of ligation, the bases interrogated are not consecutive but rather offset from the initially interrogated pair of bases by the three universal bases at 5′ end of the second incorporated oligo. In this manner a template is first sequenced at every fourth and fifth base for, typically, seven cycles. Then, to obtain the missed bases, the ligated oligos are stripped, a new $n - 1$ primer is applied, and the process is repeated. Four complete ligation series are done, each offset by one nucleotide, such that each base is interrogated twice. This scheme is referred to as "two-base encoding," and the double interrogation provides an error-checking mechanism and thus greater overall accuracy. However, the two-base encoding method introduces a level of abstraction beyond other sequencing chemistries. Two-base encoding does not provide sequence DNA directly but rather adjacency information between base pairs where each color represents any one of four possible dinucleotides. The result is a degenerate four-color alphabet termed "color space." In color space,

single nucleotide polymorphisms are identifiable as two adjacent color changes. Sequence errors, however, will introduce a "frame-shift" during translation of the color-space sequence into text (e.g., A, G, C, and T) and, as a result, the inferred DNA sequence will be completely erroneous. To avoid such frame-shifts, the manufacturer recommends that all downstream analysis be carried out in color space. This requires that reference genome sequences in text format be converted into color space prior to sequence alignment and subsequent analysis. Some alignment algorithms, such as Maq (http://maq.sourceforge.net/maq-man.shtml), are already providing some tools to facilitate conversion into and out of color space. It remains to be seen whether color space will be readily embraced by the community as a new syntax for nucleotide sequence.

There are not yet any true single molecule sequencing platforms commercially available, but it is anticipated that this will change with introduction of the HeliScope instrument by Helicos. This platform has its beginnings in seminal work performed in the Quake laboratory (Braslavsky et al. 2003), where read lengths of up to five bases were obtained by primed synthesis of single templates immobilized on a quartz flow cell. Primers were labeled with Cy3 to register the location of each template molecule, and detection of incorporated bases by Cy3/Cy5 single-pair fluorescence resonance energy transfer reduced the effect of background fluorescence to a point where the single molecules could be reliably imaged. After imaging, photobleaching was required to detect the next incorporated nucleotide. The Helicos platform has industrialized this approach and introduced some key modifications. Principally, fluorescent labels are cleaved between cycles rather than photobleached, and nucleotide analogs called "virtual terminators" have been developed that reduce the processivity of polymerase, allowing it to step through a series of identical bases one at a time.

Using single DNA molecules as sequencing template makes signal detection difficult and presents a significant challenge for this approach. However, single molecule detection has the following key advantage. Theoretically, because of the lack of dephasing (signal decay due to misextensions/nonextensions in a subset of clonally amplified template molecules), each base extension should be as detectable as the first, providing read lengths limited only by the extent of the template molecules. In practice, read lengths from single molecule templates appear not to give such read extensions. The reasons for this are not clear but may be related to secondary conformations taken up by the DNA that prevent efficient sequencing. Reagent quality may also have an impact. Since the addition of a single nucleotide molecule needs to be detected as the template is sequenced, a nucleotide that has failed to be covalently labeled with fluorophore will be unable to provide a signal for that base. Therefore, in contrast to sequencing methodologies that use multiple template molecules that can tolerate some fraction being unlabelled, it is likely that more rigorous and more expensive manufacturing processes will be required to provide the reagents for these single molecule platforms. To circumvent this issue and possibly other stochastic problems in detecting a single labeled molecule, it is possible to strip the synthesized strand from the template and resequence one or more times. This approach is supported by the Helicos platform and provides a means of improving sequence quality but at the cost of increasing the consumables used and the machine run time. In principle, prior to resequencing, incorporating controlled amounts of unlabelled bases provides the means to sample the template at various positions, resulting in a number of linked sequences or indeed the complete shotgun sequencing of the short template sequence itself, given enough iterations.

Paired-end methods

Because of the restricted length of any sequence read, obtaining information from opposite ends of long templates has been recognized for some time as a means of deriving positional information to facilitate sequence assembly (Roach et al. 1995). The ultra-short reads produced by the majority of the current flow cell sequencers make paired-end approaches even more compelling. All platforms have devised strategies for paired-end sequencing. DNA templates for flow cell sequencing are typically size selected to be a few hundred bases in length, but the method to obtain sequence from either end differs between platforms. The Illumina approach is to resynthesize template, and this requires a modified paired-end enabled flow cell and a different hardware module for bridge PCR. The first end sequence is obtained by amplification, linearization, and dehybridization of DNA to obtain single-stranded template covalently attached to the flow cell, followed by primed synthesis. To obtain sequence at the opposite end, double-stranded template is regenerated, and the opposite strand is dehybridized and sequenced by primed synthesis. This approach is very similar to that of Helicos, where polyA tailed template is annealed to and primed by flow cell bound polyT oligos. After numerous sequencing cycles, the read is extended to the opposite end of the template with unlabelled bases, the original template is melted off, and sequence at the far end of the molecule is obtained by primed synthesis back toward the flow cell using the previously synthesized strand as template. Likewise, the ABI approach to obtaining mate-pair sequence appears to be priming a ligation series from the 5′ end of the first and then second stand of a template. The superior read lengths offered by the 454 system are a distinct advantage in paired-end sequencing in that both ends can be obtained simply by reading the entire template.

All platforms require that the template directly used in the flow cell be no longer than a few hundred base pairs. To escape this constraint and obtain the sequence of mate pairs separated by longer distances, the same basic approach of template circularization and cleavage is used that was originally devised for rapidly obtaining distant mate pair information in shotgun sequencing projects (Venter et al. 2001; Holt et al. 2002; Mural et al. 2002; Gibbs et al. 2004). Briefly, DNA is sheared, accurately size selected, and recircularized in the presence of a stuffer carrying sites for type IIS restriction enzymes such as MmeI or EcoP15I that cleave downstream from their recognition sequence. The restriction sites are in opposite orientation such that after digestion short (~18–25 bp) fragments comprised of the stuffer flanked by end sequence tags are isolated and used as sequencing template. To address the difficulty in uniquely mapping, in complex genomes, the short sequence tags produced by type IIS enzymes, a random shearing method has been developed that generates much longer end segments (Korbel et al. 2007). Here, biotinylated linkers are added to starting linear DNA fragments. These end-modified DNAs are then circularized by ligation, as above. The DNA circles are then randomly sheared by nebulization, and the biotinylated adapter sequences, now flanked by end sequences from the original DNA, are captured on streptavidin beads for subsequent sequencing. To date, this approach has been used for 454 sequencing, but in principle it is

applicable to all platforms and its utility should increase as read length capabilities improve.

While all paired-end approaches have enormous utility for acquiring position information that can facilitate assembly and enable the detection of structural variants, it is important to note that for all but the 454 platform, the cost and run times for paired-end protocols are essentially twice the requirements of single read approaches.

Targeted sequencing

Flow cell sequencing approaches are best suited for generating a lot of data (Mbp to Gbp) on one or a small number of samples. Millions of Sanger reads have been generated from individual whole-genome shotgun libraries to assemble reference genomes, but the Sanger one-read-per-sample format has also lent itself well to large-scale resequencing of sets of PCR amplified genes for variant detection (Bustamante et al. 2005; Greenman et al. 2007; Wood et al. 2007). Since a flow cell can be partitioned into only a small number of lanes and each partition results in some degree of loss of valuable flow cell surface area, other approaches have been sought for targeted sequencing. In principle, a high degree of multiplexing can be achieved by tagging individual DNA samples with alternative or extended adapters during library preparation, then pooling them for sequencing. The extra sequence provides a barcode that allows the sorting of reads from mixed data according to sample of origin. This approach will become more appealing as read lengths increase such that the proportion of each read dedicated to tag sequence is reduced.

An approach to targeted sequencing that is showing considerable promise is the direct selection of desired sequences from a complex DNA pool. Using a programmable microarray, Porreca et al. (2007) synthesized 55,000 molecular inversion probes (Hardenbol et al. 2003; Dahl et al. 2007) that targeted selected human exons. A library was generated by solution hybridization of released probes with sheared human gDNA, and then sequenced on the 1G analyzer. Approximately 10,000 (28%) of the 55,000 anticipated targets were identified at least once, and the degree of redundancy ranged widely. Using an even more direct approach, Albert et al. (2007) hybridized sheared human gDNA to microarrays containing oligonucleotide probes complementary to either dispersed short targets (6726 human exons) or single long genomic segments up to 5 Mbp in length. Following capture, gDNA was eluted and sequenced on the GS-FLX. Depending on the particular experiment, between 65% and 77% of reads were from targeted regions, and between 93% and 95% of targets were hit by at least one sequence read (with median coverage of five- to sevenfold). A method for capture of all annotated human exons on a 2.1-million feature Nimblegen array is under development by this group. Finally, employing a similar strategy, Hodges et al. (2007) used a series of seven oligonucleotide microarrays to attempt to capture ~200,000 human exons. After elution of captured DNA and sequencing on the 1G analyzer this group reported that up to 98% of target exons were sequenced, and the average coverage was 1.2-fold. Depending on the specific chip, between 55% and 85% of captured and sequenced fragments were from target regions. Thus, collectively, these studies show that rapid enrichment by oligonucleotide hybridization is simple and reasonably effective for constructing sequencing libraries enriched in sequences from desired contiguous or noncontiguous segments of gDNA and go a long way toward impedance matching of sample preparation and massively parallel flow cell sequencing.

Throughput and accuracy

The manufacturer's specifications for instrument configuration, throughput, and operating costs (based on vendor supplied list prices), current as of this writing, are presented in Table 1. These figures are independent of requirements for informatics infrastructure. Interestingly, because of relatively rapid run times, the GS-FLX still has several fold higher sequencing capacity than the other platforms, measured in terms of raw bases per week. However, this remains the most expensive platform at ~$85,000 per Gbp, a cost more than thirty times that of the other platforms. It is important to note, however, that for all platforms there is still substantial headroom for increasing throughput and decreasing cost per Gbp. In principle, the greatest advances may come from single molecule platforms such as Helicos, where read lengths are not inhibited by dephasing and single molecule templates can be introduced onto the flow cell at very high density. It is expected that in the future there will also be considerable improvements in the speed and accuracy of flow cell imaging, which is currently a major rate-limiting step for the non-pyrosequencing platforms. Since instrument runs can currently take on the order of a week to complete, machine amortization needs to be considered seriously in terms of a hidden cost of DNA sequencing, particularly in this rapidly changing technology landscape.

Sequence data, no matter how rapidly and cheaply it is produced, is only useful if it is accurate. Key to the utility of high-throughput Sanger sequencing has been an automated method and universal standard for defining accuracy: the *phred* score (Ewing and Green 1998; Ewing et al. 1998). The *phred* software package assigns a log-transformed error probability informed by peak characteristics to each base in an electropherogram and stores these results in a dedicated file. *phred* quality values have been essential for trimming low confidence data from sequence reads and have greatly facilitated the sharing of capillary sequence data. An analogous quality metric that would allow comparison of data from the new flow cell sequencing platforms, where per base error rates are higher than Sanger sequencing, remains a key unmet need. While a universal quality standard is desirable, it may be lacking for some time to come because sequencing chemistries have diverged and each has its own issues affecting accuracy. While the different flow cell sequencing methods may report quality values on the same numerical scale as *phred*, these quality values are not comparable across platforms. For now, the most reliable approach is to determine error rates empirically, by sequencing a known standard. Pyrosequencing data from the GS-FLX is typically preprocessed by removing reads that lack the primer sequence, have more than 5% ambiguous bases, have more than 3% borderline positive calls, and have more than four bases that fall outside normal signal ranges. It appears that while over- or undercalls of homopolymer length are the principal error source, a substantial proportion of pyrosequencing errors are due to miscalls of mixed-template beads, and data sets can be easily and substantially improved by recognizing and removing these problematic mixed-template reads (Huse et al. 2007). A comprehensive treatment of error probability scoring for pyrosequencing has recently been presented by Brockman et al. (2008). Using training sets, they determined, empirically, error predictors for individual bases (e.g., miscalls and over- or undercalling of homopolymer length) and also the influence on per base error

Table 1. Manufacturer's specifications for instrument configuration and production of single end sequences from a single flow cell

Platform	Method	Template prep	Starting DNA (μg)	Instrument configuration	Throughput statistic	Data per run (Gbp)	Reagent cost per run ($)[a]	Run time
454 GS-FLX	Pyrosequencing	Emulsion PCR	3–5	Single picotiter plate, partitionable into 8 lanes	238-bp read[b]	0.1	8500	7.5h
Illumina 1G	Four-color SBS with reversible terminators[c]	Bridge PCR	0.1–1	Single flow cell, partitionable into 8 lanes	35-bp read	1.3	3000	3 d
ABI SOLiD	Oligonucleotide ligation with two-base, four-color encoding	Emulsion PCR	0.1–20	Independently controlled dual-flow cells, each partitionable into 8 lanes	35-bp reads, mapped to reference sequence allowing up to three mismatches	4	3400	7 d
Helicos HeliScope	Single-color SBS with virtual terminators	Not applicable	Not available	Single 25-lane flow cell	30-bp read	7.5	18,000	14 d

[a]Reagent costs are list prices.
[b]Average read length for a typical whole-genome library, using long read kit.
[c](SBS) Sequencing by synthesis.

of properties of the read as a whole (e.g., homopolymer count for the whole read, observed noise for the whole read, and position of the base in the read). The *phred* algorithm was then used to integrate these sources of error into an error probability for each base. This software is publicly available (Brockman et al. 2008) and should prove to be of considerable utility in processing pyrosequencing data.

The Illumina platform has implemented a four-value-per-base quality calling scheme to report the relative probabilities that a given base is an A, G, C, or T. The highest value indicates the most probable base. As with *phred*, calls are made according to read characteristics and values are reported on a log-transformed scale. For practical purposes, when the top Illumina base score is greater than about 15, it is essentially equivalent to a *phred* score. An important consideration is the current need to generate an error model for each run using a training set. The training set can be the actual run data, if the target is known, or can be some other known sequencing target run in one of the flow cell lanes specifically for the purpose of calibration. Because the Illumina platform uses four-color fluorescence, matrix calibration to correct for spectral overlap is also required, and this can be done using the same sample that is used to generate the error model. At our center, we have found it useful to follow the recommended procedure of including a single lane of bacteriophage phiX174 on each flow cell. Based on 56 lanes of 27-cycle phiX174 runs, we observe the per-base error rate of the Illumina platform to be 1.3 ± 0.9%, which is slightly better than the manufacturer's specification of 98.5% per-base accuracy. It is clear that error rates begin to increase sharply toward the end of reads and, therefore, subjecting reads to quality trimming is desirable for many applications. However, depending on stringency, quality trimming can result in discarding a large proportion of the data. Helicos self-reports accuracy of >99% for the HeliScope and ABI >99.9% for the SOLiD, but because these platforms have not yet entered widespread use, their accuracy and calibration requirements await critical evaluation. Given that both of these platforms' sequencing protocols involve error checking (template resequencing and two-base encoding, respectively) expectations should be high. The Illumina instrument could also be run twice on the same clusters to improve error rates.

Data handling

With the ability to sequence more DNA in a week than many of the larger sequencing centers generated in a year using capillary sequencing, many laboratories may quickly find themselves lacking appropriate computational hardware and expertise to handle and make any sense of the data they are generating. Indeed, any of these new machines running at full capacity for a year will generate, in raw DNA, more sequence than existed in the whole of the National Center for Biotechnology Information (NCBI)–GenBank database at the beginning of 2008. Analysis of the sequence data has rapidly become the limiting step and will likely become the most expensive part. The sheer volume of data will provide challenges in processing, networking, storage, and analysis of the flow-cell images just to provide the initial base calling. While some manufacturers, such as Illumina, currently rely on the existence of laboratory computer resources to provide the downstream processing, other instruments, such as the SOLiD and HeliScope, provide substantial dedicated disk and computer resources for this purpose. Historically, genomic sequencing centers have chosen to archive at least the chromatogram data derived from capillary sequencing machines. Storing the 0.5–1 terabytes of raw image data from each of the next-generation DNA sequencing instrument runs is unlikely to be useful or indeed practical where the cost of long-term physical storage will be close to the cost of regenerating the sequence data itself.

To efficiently archive and exchange the large data sets generated by the latest sequencers it will be important to standardize data formats. NCBI has already established a provisional Short Read Archive or SRA (http://www.ncbi.nlm.nih.gov/Traces/sra) that aims to provide a central repository for submission, storage, and retrieval of short sequences. As of this writing, the archive contains 235 submissions, all but two of which are 454 data. Extending beyond the notion of capillary sequencing reads, secondary analyses such as assemblies and alignments will be handled by the SRA, as will extensive metadata related to, for example, the particulars of the researcher group, sequencing platform, specimens, libraries, and experiments that generated the data. At the present time, the SRA is accepting data in ZTR format, a format originally developed to store ABI trace files. However, a new unifying DNA sequence format called SSR (Short Sequence Read) that is compatible with the principles of the SRA is being developed by the bioinformatics community (http://srf.sourceforge.net/). SSR files are independent of sequencing technology and will support individual reads and sets of multiple reads without the need for associated image data.

Using the data

Once generated, sequence data will need to be either compared with a reference genome or used for de-novo sequence assembly. Software approaches to achieve this efficiently are still in their infancy, and the alignment of millions of short sequences to a reference genome poses a computational challenge. The popular DNA alignment program BLAST, for example, using default parameters would take ~12 yr to align the 40 million 36 base pair reads typical of a complete flow cell run to the human genome if computed on a single CPU. Clearly, more efficient alignment algorithms and multiple computers need to be utilized. Aligners suitable for this mapping task include Maq, Eland, and Exonerate (Illumina) (Slater and Birney 2005). Most of the approaches gain the required increases in speed by determining ungapped alignments, and therefore alignment will be blind to the presence of small base pair deletions and insertions. Both Exonerate and the commercial alignment algorithm SXOligoSearch (Synamatix) do provide the potential to identify insertion and deletions. Utilizing gapped alignments with such short sequences without conservative gap opening and extension penalties against a mammalian-sized genome would likely generate too many false positive and ambiguous mappings to be useful. However, paired-end sequencing approaches where one end can be unambiguously anchored allowing a gapped alignment of the other read may prove useful. It should be stressed that the complete resequencing of a mammalian genome still represents a challenging and expensive task, remembering that for this purpose the effective size of the human genome is 6 Gbp, not the 3 Gbp of the haploid reference sequence. For the current generation of machines the 15- to 20-fold coverage required of a human genome would sequester a machine for the best part of a year, a dedicated use that few sequencing laboratories would probably consider regardless of reagent costs.

A somewhat unexpected initial application of new sequencing technology has been in the characterization of chromatin immunoprecipitated DNA. The success of this application with the technology has been partially due to the fact that it is far less dependent on sequence quality or sophisticated sequence alignment. As long as an unambiguous mapping to the genome can be achieved, a sequence read can contribute to a genome-wide DNA–protein interaction profile. Successful studies have been carried out for both transcription factors (Johnson et al. 2007; Robertson et al. 2007) and histone modifications (Barski et al. 2007; Mikkelsen et al. 2007). It is also anticipated that the simple ability to map sequence reads unambiguously will provide means to characterize transcriptomes as well as genomic rearrangements such as inversions and deletions, particularly where paired-end sequencing methods are deployed.

De-novo assembly is an appealing application for flow cell sequencing but a challenging one for the short-read platforms. It is recognized that a contig accurately assembled from short reads can be a useful and cost-effective surrogate for single Sanger reads in many applications. Six different algorithms for short-read assembly have already been published, and more are under development. The first programs, SSAKE (Warren et al. 2007), VCAKE (Jeck et al. 2007), SHRAP (Sundquist et al. 2007), and SHARCGS (Dohm et al. 2007), all use conservative seed and extension approaches, with modifications to reduce errors in contig elongation either through the use of sequence depth information (Jeck et al. 2007) or by identification of ambiguities between overlapping reads that are candidates to extend contigs (Dohm et al. 2007). The most recent programs, Edena (Hernandez et al. 2008) and ALLPATHS (Butler et al. 2008), take a different approach and assemble short reads by computing an overlap graph. All of these assemblers can already build accurate contigs for bacterial genomes that are on the order of 10 kbp in length, and the development of paired-end protocols promises to drastically improve the utility of such sequencing approaches for de-novo assemblies. One current unknown is the fidelity of end-paring from the Illumina, ABI, and Helicos approaches where the sequence substrates are randomly distributed on the flow cell. Even small errors rates in the assignment of sequence pairs may be sufficiently obfuscating for the current generation of assemblers, requiring new algorithms capable of handling softer mate-pair assignments. The longer-read 454 pyrosequencing system has proven to be particularly successful for de-novo assembly (Goldberg et al. 2006; Hofreuter et al. 2006; Hiller et al. 2007; Pearson et al. 2007; Smith et al. 2007). It is clear that the ability of the 454 platform is enhanced further with the recent introduction of longer (>200 bp) reads and paired-end read protocols, the latter being effective in detecting structural variation within the human genome (Korbel et al. 2007). Familiarization with the data has also shown that stringent filtering for sequence quality can provide significant improvements in assembly quality (Huse et al. 2007), indicating that eliminating erroneous reads, even at the expense of removing significant numbers of good reads, can be beneficial in assembly applications.

Conclusions

Flow cell-based sequencers are now revitalizing the field of DNA sequencing, providing capacities to drive a host of new applications including the deep characterization of immunoprecipitated DNA and transcribed sequences. While the complete derivation of bacteria and fungal genomes now becomes extremely tractable, at the current sequencing throughput and read length, the generation, de novo, of a high-quality mammalian-sized genome remains far from trivial. We will inevitably see more individual human genomes being sequenced (Levy et al. 2007; Qiu and Hayden 2008) but the associated costs in both reagents and machine time will ensure that such sequencing will remain in the research domain for some time, and the promise of genomes for personalized medicine will remain a distant goal. While we wait for capacities to improve and costs to come down, technologies that allow the selection and enrichment of specific sequence targets, such as exons, will be of the utmost importance and will represent a very active research area in both the academic and commercial domains.

Acknowledgments

R.A.H. and S.J.M.J. are Michael Smith Foundation for Health Research scholars. We thank representatives of 454, ABI, Illumina, and Helicos for readily providing current and official instrument specifications. We also thank Martin Hirst and Gabor Marth for helpful discussion of the manuscript.

References

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* **4:** 903–905.

Barany, F. and Gelfand, D.H. 1991. Cloning, overexpression and nucleotide sequence of a thermostable DNA ligase-encoding gene. *Gene* **109:** 1–11.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

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

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W.L., Russ, C., Lander, E.S., Nusbaum, C., and Jaffe, D.B. 2008. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.* **18:** 763–770.

Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., et al. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437:** 1153–1157.

Butler, J., Maccallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C., and Jaffe, D.B. 2008. ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res.* **18:** 810–820.

Dahl, F., Stenberg, J., Fredriksson, S., Welch, K., Zhang, M., Nilsson, M., Bicknell, D., Bodmer, W.F., Davis, R.W., and Ji, H. 2007. Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci.* **104:** 9387–9392.

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. 2007. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res.* **17:** 1697–1706.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8:** 186–194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.

Gilbert, W. and Maxam, A. 1973. The nucleotide sequence of the lac operator. *Proc. Natl. Acad. Sci.* **70:** 3581–3584.

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

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C.,

Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. 2007. Patterns of somatic mutation in human cancer genomes. *Nature* **446:** 153–158.

Hardenbol, P., Baner, J., Jain, M., Nilsson, M., Namsaraev, E.A., Karlin-Neumann, G.A., Fakhrai-Rad, H., Ronaghi, M., Willis, T.D., Landegren, U., et al. 2003. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* **21:** 673–678.

Hernandez, D., Francois, P., Farinelli, L., Osteras, M., and Schrenzel, J. 2008. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Res.* **18:** 802–809.

Hiller, N.L., Janto, B., Hogg, J.S., Boissy, R., Yu, S., Powell, E., Keefe, R., Ehrlich, N.E., Shen, K., Hayes, J., et al. 2007. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: Insights into the pneumococcal supragenome. *J. Bacteriol.* **189:** 8186–8195.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39:** 1522–1527.

Hofreuter, D., Tsai, J., Watson, R.O., Novik, V., Altman, B., Benitez, M., Clark, C., Perbost, C., Jarvie, T., Du, L., et al. 2006. Unique features of a highly pathogenic *Campylobacter jejuni* strain. *Infect. Immun.* **74:** 4694–4707.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298:** 129–149.

Housby, J.N. and Southern, E.M. 1998. Fidelity of DNA ligation: A novel experimental approach based on the polymerisation of libraries of oligonucleotides. *Nucleic Acids Res.* **26:** 4259–4266.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8:** R143. doi: 10.1186/gb-2007-8-7-r143.

Jeck, W.R., Reinhardt, J.A., Baltrus, D.A., Hickenbotham, M.T., Magrini, V., Mardis, E.R., Dangl, J.L., and Jones, C.D. 2007. Extending assembly of short DNA sequences to handle error. *Bioinformatics* **23:** 2942–2944.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420–426.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol.* **5:** e254. doi: 10.1371/journal.pbio.0050254.

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

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448:** 553–560.

Mural, R.J., Adams, M.D., Myers, E.W., Smith, H.O., Miklos, G.L., Wides, R., Halpern, A., Li, P.W., Sutton, G.G., Nadeau, J., et al. 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296:** 1661–1671.

Nyren, P., Pettersson, B., and Uhlen, M. 1993. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal. Biochem.* **208:** 171–175.

Pearson, B.M., Gaskin, D.J., Segers, R.P., Wells, J.M., Nuijten, P.J., and van Vliet, A.H. 2007. The complete genome sequence of *Campylobacter jejuni* strain 81116 (NCTC11828). *J. Bacteriol.* **189:** 8402–8403.

Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F., et al. 2007. Multiplex amplification of large sets of human exons. *Nat. Methods* **4:** 931–936.

Qiu, J. and Hayden, E.C. 2008. Genomics sizes up. *Nature* **451:** 234. doi: 10.1038/45123a.

Roach, J.C., Boysen, C., Wang, K., and Hood, L. 1995. Pairwise end sequencing: A unified approach to genomic mapping and sequencing. *Genomics* **26:** 345–353.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4:** 651–657.

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

Ronaghi, M., Uhlen, M., and Nyren, P. 1998. A sequencing method based on real-time pyrophosphate. *Science* **281:** 363–365.

Sanger, F. and Coulson, A.R. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94:** 441–448.

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

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

Slater, G.S. and Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31. doi: 10.1186/1471-2105-6-31.

Smith, M.G., Gianoulis, T.A., Pukatzki, S., Mekalanos, J.J., Ornston, L.N., Gerstein, M., and Snyder, M. 2007. New insights into *Acinetobacter baumannii* pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. *Genes & Dev.* **21:** 601–614.

Sundquist, A., Ronaghi, M., Tang, H., Pevzner, P., and Batzoglou, S. 2007. Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE* **2:** e484. doi: 10.1371/journal.pone.0000484.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Warren, R.L., Sutton, G.G., Jones, S.J., and Holt, R.A. 2007. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23:** 500–501.

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjoblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., et al. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* **318:** 1108–1113.