Next-generation gap

John D McPherson

There is a growing gap between the generation of massively parallel sequencing output and the ability to process and analyze the resulting data. New users are left to navigate a bewildering maze of base calling, alignment, assembly and analysis tools with often incomplete documentation and no idea how to compare and validate their outputs. Bridging this gap is essential, or the coveted $1,000 genome will come with a $20,000 analysis price tag.

© 2009 Nature America, Inc. All rights reserved.

DNA sequencing has undergone a remarkable evolution during the past 30 years. First published in 1977, dideoxynucleotide sequencing enabled DNA sequencing to become a mainstream laboratory protocol[1]. Initial methods used isotope-labeled dideoxynucleotides and vertical polyacrylamide slab gels, with DNA sequence read by hand from exposed film one base at a time. The throughput of the method overwhelmed neither the manual base-calling method nor the subsequent sequence analysis. It was not until the first fluorescence-based sequencers, with data collection largely automated, that large-scale DNA sequencing could be envisaged[2]. Although the data-collection phase of DNA sequencing was greatly simplified by the advent of capillary-based fluorescence sequencers, the massive scale of templates needed for production-scale sequencing limited high-throughput DNA sequencing to a relatively few, specialized laboratories.

Sequencing the human genome

Large-scale sequencing laboratories largely evolved during the Human Genome Project, with the publicly funded International Human Genome Sequencing Consortium effort using over 20,000 bacterial artificial chromosome (BAC) clones[3,4]. The availability of the clone-based maps assisted the sequencing of the human genome by making it possible to select clones for sequencing that would ensure comprehensive coverage and reduce sequencing redundancy. The use of BAC clones also mitigated the demands on sequence assembly software and computer hardware by reducing the project to manageable bites, restricting random shotgun sequencing and initial local assembly to individual clones. The final draft genome was produced by merging the clone contigs on the basis of overlaps, paired-end reads and known transcripts[5]. A parallel effort by Celera Genomics to shotgun-sequence entire genomes, including human, required considerable investment in proprietary software development to avoid the pitfalls of coassembly of regions that are similar in sequence but reside in distant regions of the genome[6–8]. By far the largest gap between DNA sequencing and analysis was seen in annotation and visualization, with several groups scrambling to package the new human genome in a usable format (UCSC Genome Browser[9], Ensembl Genome Browser[10] and NCBI Map Viewer[11]). The large-scale sequencing centers continued making incremental improvements to Sanger sequencing aimed at reducing overall cost of genome sequencing, with the research community benefiting from accessibility to a wide variety of genomes. Software for detection of single-nucleotide polymorphisms flourished, especially in the application of targeted resequencing projects using the sequenced genomes as a foundation and reference (for example, PolyPhred[12], PolyScan[13] and SNPDetector[14]).

Birth of a new generation

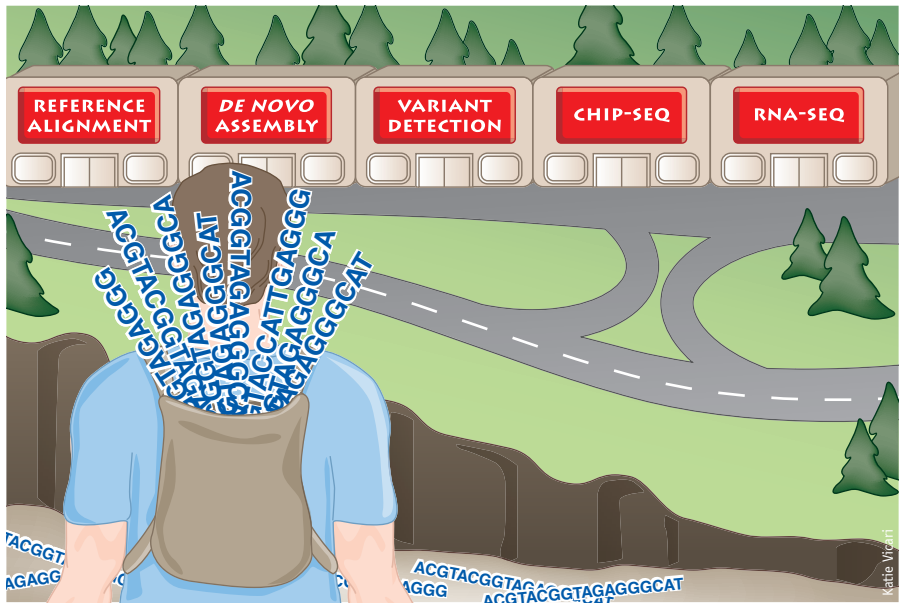
This landscape has dramatically changed in the past 5 years with the introduction of new, massively parallel sequencing platforms heralding the second generation of DNA sequencing.

Bursting on the scene in 2004, the Roche (454) Genome Sequencer (GS) began the revolution in DNA sequencing[15]. This instrument uses pyrosequencing to enable the simultaneous sequencing of several hundred thousand DNA fragments, with a read length greater than 100 base pairs (bp). The current GS FLX Titanium produces greater than 1 million reads in excess of 400 bp (http://www.454.com/products-solutions/system-features.asp). The GS was followed in 2006 by the Illumina (Solexa) Genome Analyzer (GA) which used sequencing-by-synthesis to generate then tens of millions of 32-bp reads. Today, the Illumina GAIIx produces ~200 million 75–100-bp reads (http://www.illumina.com/downloads/SQ_GAIIx_spec_sheet2_04_09LR.pdf). Applied Biosystems dominated the capillary sequencing hardware space and entered the next generation arena in 2007 with the SOLiD (Sequencing by Oligo Ligation and Detection) platform, capable now of producing 400 million 50-bp reads (combined total in two independent flow cells). As its name implies, it uses a ligation-based sequencing method with semidegenerate short oligonucleotides (http://www3.appliedbiosystems.com/AB_Home/applicationstechnologies/SOLiDSystemSequencing/overviewofsolidsystem/index.htm).

John D. McPherson is at the Ontario Institute for Cancer Research, Toronto, Ontario, Canada.
e-mail: john.mcpherson@oicr.on.ca

PUBLISHED ONLINE 15 OCTOBER 2009;
DOI:10.1038/NMETH.F.268

Each of the above sequencers uses template amplification to achieve the signal intensity needed for detection. A commercially available single-molecule sequencing platform, Helicos HeliScope, is also available, producing 400 million 25–35-bp reads (http://www.heliocosbio.com/Technology/TrueSingleMoleculeSequencing/tSMStradePerformance/tabid/151/Default.aspx). Detailed reviews of the chemistries behind these methods appear elsewhere[16,17], with the vendor websites providing up-to-date overviews of each of their platforms.



A gap exists between current sequence-generation and data-analysis capabilities.

Not all next-generation sequencers see a base the same way

Next-generation sequence throughput gains have been made not by scaling the capillary model but by often radical departures from these earlier platforms. Differences in chemistries and raw data collection require individualized data processing pipelines and hinder combining output from different next-generation platforms. The Illumina GAIIx and the Helicos HeliScope most approximate the capillary sequencers, generating base-specific signal intensities, with basic algorithms needed to determine the most likely template-directed base being incorporated. The output is readily obtained as simple base sequence. In contrast, the Roche GS FLX adds only one type of nucleotide at a time, allowing multiple base incorporations across mononucleotide stretches in a single cycle, resulting in a signal proportional to the number of bases incorporated. The resulting flowgram can be readily converted to bases, but with some uncertainty surrounding the length of long mononucleotide repeats. SOLiD uses dibase encoding, whereby two adjacent template bases at a time are interrogated by the incoming labeled oligonucleotide destined for ligation. This provides redundancy owing to overlap of the two bases read out at each position. The sequence output is encoded not as single bases but as the numbers 0, 1, 2 and 3, with each representing four possible dinucleotides. The SOLiD dibase sequences can also be decoded to simple base sequence if any portion of the sequence is known, but with loss of the dibase encoding data that provide discrimination between sequencing error and polymorphism (http://solidsoftwaretools.com/gf/).

Perhaps one of the most significant differences between capillary and next-generation sequencers is that only individual templates are sequenced on the latter. On the three main commercial next-generation platforms, each individual template is amplified in a clonal manner, with only the Helicos HeliScope providing direct sequencing of the unamplified DNA fragments. Past sequencing relied on simultaneous sequencing of multiple templates derived from the same PCR product or clone. This impairs variation detection in mixed pools of fragments: capillary-based sequencing struggled to distinguish multiple base calls at a single position without significant representation of each allele in the fragment population. With next-generation sequencing, each fragment within a mixture is analyzed independently, allowing deep single-nucleotide and small indel variant analysis within pooled samples and heterogeneous samples. Rapid and precise alignment of the sequence reads is essential to avoid misinterpretation of misaligned reads as positional variants.

These basic differences in how data are collected, as well as the sheer volume of data produced, have led to a gap between the traditional sequence detection and analysis tools on one side and the next-generation sequence data on the other. The initial drivers of this software gap are the variations in platform chemistries described above. Each platform has disparate output and unique error profiles, negating a Swiss-army-knife approach to universal base-calling and sequence analysis. At first, software support fell largely to the platform vendors, who produced base-calling software out of necessity. The proprietary nature of these early next-generation platforms did not allow the ready use of the software that had been generated during the Human Genome Project era. Simple conversion to strings of base and quality calls to enable use with the existing tools is insufficient as the new sequencing mechanisms generate data with unique properties, such as SOLiD dibase encoding, not appreciated by the traditional software packages. In addition, the extreme number of short reads generated readily cripple traditional sequence analysis software and hardware configurations. This affords enormous opportunity for third-party software development but complicates community utility as sequencing platforms continue to diversify.

Chartering new applications

To further complicate the next-generation sequencing landscape, these platforms have been quickly applied to many genomic methods not traditionally using sequence data. The first applications were obvious—small genome sequencing[18,19], small RNA discovery[20] and PCR amplicon analysis[21]—but the second wave of applications saw next-generation sequencing starting to replace traditional microarray-based output because the large number of sequences generated in a single run

© 2009 Nature America, Inc. All rights reserved.

rival or exceed the content and dynamic range of microarrays. This is not entirely new, as serial analysis of gene expression (SAGE[22]), requiring considerable sequence capacity, led a similar movement using capillary sequencing. Next-generation platforms now put the power of high-throughput sequencing within the grasp of a single investigator–led laboratory or core facility. Expression analysis is now possible through complete shotgun of all transcripts (RNA-seq[23]) or very deep analysis of sequence tags (DeepSAGE[24]). Chromatin immunoprecipitation (ChIP) has traditionally been analyzed using microarrays but is also readily replaced by direct sequencing of the captured material (ChIP-seq[25]). Likewise, copy-number variation (CNV) can also be detected by determining sequence depth of coverage across a genome (CNV-seq[26]). All these methods benefit from the transition to sequence-based output by many-fold increased dynamic range and freedom from the necessity for prior knowledge to determine the microarray content. For a discussion of software for these sequencing applications see refs. 27–29, in this issue. RNA-seq is perhaps the most complex next-generation application. Expression levels of specific genes can be obtained if sufficient sequence depth is obtained; however, there are many other subtleties in the data that are essential to the analysis. Differential splicing, allele-specific expression, RNA editing and fusion transcripts must be determined when comparing samples for disease related or mechanistic studies. These attributes are not readily obtained by microarray analysis[27]. CNV-seq can be combined with the analysis of paired sequence reads from the ends of contiguous DNA fragments to detect structural rearrangements. Inversions, translocations and large insertions and deletions are detected by analyzing the orientation of the paired reads with respect to a reference genome[28]. All of these methods are dependent on correct alignment of the sequence to a reference genome.

Alignment of sequence reads is at the root of all the above analyses, as well as sequence assembly and single nucleotide variant (SNV) detection. For RNA-seq, ChIP-seq and CNV-seq, the first step is to align sequences to a reference genome or set of reference transcripts. A growing number of software packages are appearing that, in simplest terms, count colocalized sequences to determine expression levels, binding locations or copy number. Specificity of alignment in large, complex genomes, such as the human, were initially hindered by the short length of next-generation reads, but this has become less of an issue as read lengths increase. Along with the increased read length, read numbers have dramatically increased and will continue to do so in the coming year. More than a dozen open-source or commercial short-read alignment tools are available[17,28].

In addition to reference alignment for local counting or analysis, the global assembly of sequence reads into a complete genome is an essential need for next-generation sequencing. One approach is a mapped assembly in which sequence reads are first aligned to a reference genome and a consensus sequence generated for the new genome. This type of assembly is potentially biased toward the reference genome used, possibly masking important structural differences. Matched read information from the same end of a contiguous DNA fragment (paired reads) or from captured distal ends of larger fragments (mate pairs) can be used to verify assemblies and for regions not contained in the reference. More important is a true unbiased *de novo* assembly using only the sequence reads alone. Longer read and longer mate-pair development on the next-generation platforms are providing much needed increased texture to the sequence data to drive *de novo* assembly, but with few gains in the ability to assemble large, complex genomes. Recently, a parallel short-read assembler, ABySS (Assembly by Short Sequences), was used to assemble 42-fold redundancy, whole-genome sequence reads of a Yoruban man[30]. Although this is an impressive first for next-generation sequence assemblers, there is room for much improvement, as the effort generated more than 2.7 million contigs longer than 100 bp and 680,000 contigs longer than 1,000 bp. These contigs were compared to the available reference genome for a more unbiased detection of sequence and structural variants. Other next-generation sequence assemblers have yet to break the vertebrate genome assembly barrier[17]. Second generation sequencing technology promises to deliver cost-effective genome coverage in the very near future, but a software and computational hardware gap for *de novo* assembly is likely to lag these developments.

The ever-increasing appreciation for the diversity of the human genome among individuals begs the question of the utility of a single, simple, linear reference genome for comparative alignments as described above. The 1000 Genomes project (http://www.1000genomes.org/) aims at capturing much of this diversity and may lead to an improved catalog of human genome variation, but an inadequate, linear representation of the variable genome is likely to remain. Unbiased assembly of individual genomes will undoubtedly produce more accurate results if they can be investigated relative to a genome knowledge base that captures structural variation of a higher order than can at present be represented by a linear reference.

Closing the gap

Genome centers around the world will continue to advance the definition of ultra-high-throughput sequencing as they embrace large fleets of second- and third-generation sequencers. As before, these centers will build the needed infrastructure to support these platforms, including sophisticated data analysis pipelines. But the next-generation sequencing platforms have also moved high-throughput sequencing into the hands of individual investigators, whether it be through acquiring an instrument or contracting through core facilities or third-party commercial sequencing operations. With a modest investment, data acquisition for any genomic project can become comprehensive, integrating multiple genome-wide data sets with great depth and resolution. Unfortunately, the software and computer hardware demands on these analyses are not much less than those of the large Genome Centers. From this perspective, the gap between large-scale genome centers and individual investigators may seem to be growing, not shrinking, as the next-generation platforms' apparent promise of a 'Genome Center in a box' may have only been half delivered, providing data without a full suite of tools. Fortunately, many software packages emerging are coming from smaller facilities and can be implemented without full data-center support. Nonetheless, navigating a plethora of software possibilities across multiple platforms will continue to be a daunting task. Cloud-computing solutions providing internet access to large clusters of computers offer some hope of accessing data pipelines in conjunction with significant hardware power at a reasonable cost. These

© 2009 Nature America, Inc. All rights reserved.

services provide infrastructure and software support, bringing a virtual data center to any laboratory. Next-generation sequencing software could be offered by such services as preconfigured pipelines while also supporting the uploading of custom virtual operating systems, allowing flexibility in the analyses performed. It is essential that such solutions combine cloud computing environments with local access to the growing sequence and biological databases. Owing to internet transfer speed limitations, it is already a formidable task just to deliver the newly generated sequence data set to the cloud for analysis, and is no longer tractable for each investigator to upload the rapidly growing genome data sets to query against. These decentralized computing solutions come with additional challenges, such as difficulty in documenting which data sets were available at the time of the query for future reference and, of course, a volume of processed sequence output data that may be as overwhelming to the average researcher as the initial input set of sequence data without cloud support for downstream analyses. Clear documentation of the analysis algorithms used is needed to ensure a full understanding of the processed data. Community adoption of output data standards and formats for reference alignments, assemblies and detected variants is also essential for easing the data management problem. Solving these issues may simply shift the software gap from sequence processing (base-calling, alignment or assembly, positional counting and variant detection) to sequence analysis (annotation and functional impact). Development of methods and tools to derive meaning from a genome sequence and put it in context with other genomes are also needed.

Democratization of next-generation sequencing

As next-generation sequencing becomes more commonplace, its full potential will only be realized when new users have a firm understanding of the data produced and have an understanding of the tools available for its analysis. There is not likely ever to be a one-size-fits-all solution, nor should there be. Each user must make informed decisions as to the appropriate analysis tool and have a confident understanding of the data set produced. Next-generation sequencing is a rapidly evolving field, and it is easy to spend more time evaluating software suites than analyzing the output data. Basic alignment and variant-detection tools are provided by the next-generation sequencing platform vendors and sequencing service providers and are a sensible place for new users to start. But they should also read current review articles that aim to demystify the software landscape, as well as learning from others who have bridged the gap through discussion forums such as SEQanswers (http://seqanswers.com/) that provide instant access to a next-generation sequencing user community. Whichever software is used, it is most important that its limitations be understood. Lastly, there is a tendency for next-generation sequencing to be seen as a hammer and every biological question as a nail. Next-generation sequencing in many cases will not provide the answer but rather is only one of many investigational tools needed.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

1. Sanger, F., Nicklen, S. & Coulson, A.R. *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467 (1977).
2. Smith, L.M. *et al. Nature* **321**, 674–679 (1986).
3. McPherson, J.D. *et al. Nature* **409**, 934–941 (2001).
4. Lander, E.S. *et al. Nature* **409**, 860–921 (2001).
5. Kent, W.J. & Haussler, D. *Genome Res.* **11**, 1541–1548 (2001).
6. Myers, E.W. *et al. Science* **287**, 2196–2104 (2000).
7. Huson, D.H. *et al. Bioinformatics* **17** (suppl. 1), S132–S139 (2001).
8. Venter, J.C. *et al. Science* **291**, 1304–1351 (2001).
9. Kent, W.J. *et al. Genome Res.* **12**, 996–1006 (2002).
10. Hubbard, T. *et al. Nucleic Acids Res.* **30**, 38–41 (2002).
11. Wheeler, D.L. *et al. Nucleic Acids Res.* **29**, 11–16 (2001).
12. Bhangale, T.R., Stephens, M. & Nickerson, D.A. *Nat. Genet.* **38**, 1457–1462 (2006).
13. Chen, K. *et al. Genome Res.* **17**, 659–666 (2007).
14. Zhang, J. *et al. PLoS Comput. Biol.* **1**, e53 395–404 (2005).
15. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
16. Mardis, E. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
17. Turner, D.J., Keane, T.M., Sudbery, I. & Adams, D.J. *Mamm. Genome* **20**, 327–338 (2009).
18. Edwards, R.A. *et al. BMC Genomics* **7**, 57 (2006).
19. Goldberg, S.M. *et al. Proc. Natl. Acad. Sci. USA* **103**, 11240–11245 (2006).
20. Girard, A., Sachidanandam, R., Hannon, G.J. & Carmell, M.A. *Nature* **442**, 199–202 (2006).
21. Binladen, J. *et al. PLoS One* **2**, e197 (2007).
22. Velculescu, V.E. *et al. Science* **270**, 484–487 (1995).
23. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. *Nat. Methods* **5**, 621–628 (2008).
24. Nielsen, K.L., Høgh, A.L. & Emmersen, J. *Nucleic Acids Res.* **34**, e133 (2006).
25. Robertson, G. *et al. Nat. Methods* **4**, 651–657 (2007).
26. Xie, C. & Tammi, M.T. *BMC Bioinformatics* **10**, 80 (2009).
27. Mortazavi, A. & Wold, B. *Nat. Methods* **6**, S21–S31 (2009).
28. Medvedev, P., Stanciu, M. & Brudno, M. *Nat. Methods* **6**, S13–S20 (2009).
29. Flicek, P. & Birney, E. *Nat. Methods* **6**, S6–S12 (2009).
30. Simpson, J.T. *et al. Genome Res.* **19**, 1117–1123 (2009).

© 2009 Nature America, Inc. All rights reserved.