

Executive Summary

First-generation DNA sequencing, dominated by the Sanger method, reached its peak with the introduction of automated capillary electrophoresis-based instruments, which facilitated the timely completion of the revolutionary Human Genome Project. Despite this victory, low throughput translated to high cost for projects needing extensive sequencing. At its peak, Sanger capillary systems could sequence a human genome for no less than about \$10 million. Shortly after the start of the new millennium, a few scientists began to envision a next generation of DNA sequencing technologies that might enable a new vision, a metaphor really, that came to be known as the “\$1,000 genome.” Even before the first of the new-generation sequencers became a reality, scientists were beginning to demonstrate the possibility of sequencing single DNA molecules.

This report is divided into seven chapters. Following a brief introduction, Chapter 2 provides historical background on first-generation sequencing. Chapter 3 provides a detailed overview of next-generation technologies, and the following chapter describes applications of these technologies. Chapter 5 covers market-related issues and includes results from a survey of next-generation sequencing users. The sixth chapter provides general observations and conclusions from the report. Chapter 7 contains transcripts of six in-depth interviews conducted for this report with people knowledgeable in the field.

Technologies

The first major breakthrough in commercially practical next-generation sequencing came from a company called 454, which developed technology that combines miniaturized target amplification on beads, immobilized sequencing reagents, and picotiter plates on which chemistry and imaging were conducted. The resulting GS20 sequencer, which entered commerce in April 2005 with delivery of the first unit

to the Broad Institute, permitted more than 1 million 400-base reads in a 10-hour run. This kind of throughput surpassed by far anything previously possible. The company was later acquired by Roche.

The second major player, Illumina, entered the next-generation race in late 2006 through acquisition of Solexa, a company which earlier that year had begun to field its own sequencer. In operation, DNA fragments are bridge-amplified on a flow cell surface to generate polony clusters. These then undergo a sequencing-by-synthesis process, which employs labeled chain-terminating nucleotides that are imaged and cleaved at the end of each chemistry cycle to permit further chain elongation. Read length, currently around 50 bases, is considerably less than for the 454 system, although other advantages have made Illumina the clear market leader in next-generation sequencing. Through acquisition of Avantome and investment in Nanopore, Illumina has signaled its intention to evolve its sequencing technology as necessary.

Applied Biosystems (ABI) acquired Agencourt Personal Genomics in mid-2006. They became the third commercial player in next-generation sequencing with the October 2007 launch of the SOLiD system. SOLiD employs technology developed originally in the Harvard laboratory of Dr. George Church, which employs polony production on beads followed by sequential ligation of multiple octamer oligonucleotides, each bearing combinations of two bases and six nonspecific nucleotides. Read length is similar to that of Illumina, and throughput measured 20 gigabases per run as of early 2009. ABI's parent, Life Technologies, acquired single-molecule sequencing technology developer, VisiGen, in October 2008.

Among these three market leaders, 454 provides the longest reads at the highest cost per megabase. Applied Biosystems offers shorter reads and the lowest cost per megabase, while Illumina provides short reads at costs intermediate between the other two. All three are working frenetically to improve performance. Instrument and chemistry improvements appear frequently, and these often translate to greater read length, more sequence per run, or new applications.

Helicos, founded in 2003 based on technology invented in the Stanford University lab of Dr. Stephen Quake, entered the market in late 2008 with the first single-molecule sequencer, the HeliScope. The system employs a special flow surface that permits use of total internal reflection fluorescence to detect labels on single DNA fragments contained on the surface at densities up to 100 million per square centimeter. The

sequencing-by-synthesis chemistry uses Virtual Terminators with labels that are cleaved in each cycle after the imaging step. Read lengths are relatively short, throughput is exceptionally high, and the HeliScope costs about three times as much as competing systems.

Complete Genomics has announced a mid-2009 introduction of whole-genome sequencing services using distinctive sequencing-by-hybridization technology developed by cofounder Dr. Radoje Drmanac. The cPAL (combinatorial probe-anchor) technology, which permits sequencing of diploid human genomes, employs billions of DNA “nanoballs” immobilized on a slide. Throughput is expected to be high, initially about 200 gigabases per run. Complete Genomics plans to sequence whole human genomes for as little as \$5,000 for large volume customers.

A number of other companies are in varying stages of development with innovative next-generation systems. Pacific Biosciences, considered a prime contender in the single-molecule sequencing area, plans to introduce its system in the second half of 2010. VisiGen, recently acquired by Life Technologies, is developing its own single-molecule sequencing-by-synthesis technology. Intelligent Bio-Systems uses reversible terminator chemistry and plans to begin beta testing before the end of 2009. Other contenders include ZS Genetics, Reveo, LightSpeed Genomics, and Dover Systems.

Several companies are attempting to develop nanopore-based sequencing systems, which in principle can read bases in long chains of DNA as they move rapidly through nanoscale pores, either biological or synthetic in nature. Oxford Nanopore is considered a leading contender in the field, although they have yet to demonstrate feasibility for actual sequencing. Another player, NABsys, is working to develop hybridization-assisted nanopore sequencing (HANS).

Regarding bioinformatics, next-generation systems generate huge image files generally considered too large for archiving, which is usually done after conversion to sequencing data. Sequence assembly presents challenges when read lengths are short, and mapping reads to known sequence for re-sequencing requires special programs as well. System manufacturers provide much of the necessary software and continually upgrade it. However, some of the most innovative software in the field comes from academic groups. Third-party software companies make valuable contributions, especially in the area of integrating various software tools into user-friendly systems.

Applications

Next-generation sequencers with their high throughput and low cost per gigabase are particularly useful for whole-genome sequencing, as evidenced by the aforementioned Complete Genomics' near-term target price of \$5,000 per genome. The international 1,000 Genomes Project, which began in early 2008, uses next-generation technology in order to provide a highly detailed collection of human genetic variants that goes well beyond anything available previously. Sequencing of 1,200 genomes is slated for completion by the end of 2009. Even simpler and less expensive, exome (all exons) sequencing is a subtopic of the 1,000 Genomes Project, which has no particular disease focus. The exome effort aims to correlate genetic variations with disease phenotypes, and hopes to reduce costs to \$1,000 per exome.

Next-generation technologies promise to be particularly useful for RNA re-sequencing, a field dominated to date by microarrays and qPCR technology. Next-generation systems identify transcripts or other RNA types by sequencing short segments that are long enough to identify their locus of origin in the genome. Paired-end read sequencing permits detection of RNA splice variants as well. Unlike microarrays, next-generation sequencing enables detecting and quantifying RNAs over large dynamic ranges with high sensitivity.

Another application, ChIP-seq, permits chromatin immunoprecipitation studies of protein-binding genomic sequences as an alternative to ChIP-chip microarray technology. The sequencing alternative is still the more expensive of the two, but could well dominate as costs come down. DNA methylation epigenetic studies are another important and viable target for ChIP-seq.

Disease prediction and diagnosis provides a particularly attractive prospect for next-generation sequencing. Current targets are few, but next-generation systems are currently enabling the genome-wide association studies needed to identify the genome variants central to deeper understanding of the molecular basis of genetically complex diseases and predicting individual risk of developing them. Illumina, for example, plans to open a CLIA-certified laboratory to offer proprietary assays derived from its licensing and discovery program. They also plan to conduct "platform partnering" to collaborate with customers in developing diagnostic applications. Additionally, the NHGRI-sponsored ClinSeq study is sequencing about 1,000 individuals with signs of impending coronary artery disease to associate sequence variants with phenotype. A number of other such studies are in progress.

The field of personal genomics emerged recently with the founding of several companies that offer to determine genetic variants in the DNA of individuals at affordable prices. Companies such as 23andMe, Navigenics, and deCODEme provide individuals with reports on variants found in their DNA using microarray technology, and the probability that they will develop particular diseases during their lifetime. The medical value of these services is still controversial, but should become less so as validation studies in progress begin to generate results. In any event, we can expect improvements as next-generation technologies uncover additional variants, likely to be of greater value than current ones. Next-generation sequencing should begin to impinge on microarrays for personal genomics as costs continue to fall.

Market Dynamics

Next-generation sequencing instruments and consumables are selling rapidly and contributing nicely to growth rates of the three market leaders, each of which were already profitable corporations. A second round of product introductions in the next-generation market, which began with the recent introduction of the Helicos single-molecule system, will continue over the next year or two. Second-round players must offer significant improvements over current market entrants in order to merit viable market share. Significant barriers to market entry exist, and candidates can expect difficulty in overcoming the advantages of market momentum and power of current leaders. Third-round technologies, particularly those involving nanopores, remain speculative at present, but could serve to accelerate market decentralization by virtue of extremely low costs and high speeds.

Deal flow has been relatively light in the still-young next-generation sequencing arena. Most of the activity to date has involved the three market leaders either acquiring the technologies needed for their current systems or those that might support a next-next generation of systems.

A survey of next-generation sequencing system users in both industrial and academic institutions generated a number of interesting observations and opinions, some of which are mentioned here:

- When asked which next-generation platform would dominate the market for the next two to three years, a majority of respondents selected Illumina's Genome Analyzer.
- A significant minority felt that Pacific Biosciences has the most exciting third-generation technology.

- Personal genomics and cancer genomics are thought to be the two leading application areas.
- Industry-based users feel they derive good value from their next-generation systems, but academic users are more divided in opinion.
- Opinion suggests strongly that the cost of a human genome sequence will fall to between \$1,000 and \$50,000, with a substantial minority betting on the \$1,000 to \$5,000 range.

Conclusions and Observations

Next-generation sequencing is characterized by great vitality, dynamism, and inventiveness. The market is not only growing nicely, but diversifying strongly into labs that have not previously been involved in sequencing. Next-generation is also starting to impinge on DNA microarray usage and should continue to do so as costs drop.

It is now feasible to sequence large numbers of genomes and exomes, which will reveal finally just how valuable is the notion of predicting disease risk at the genome level. Information gleaned from such studies may not prove useful *per se*, but will definitely provide major clues for investigations at the protein, metabolite, and systems biology levels.

The next several years are likely to prove a Golden Age for next-generation sequencing as the fruits of past innovation and continuing ingenuity are harvested to finally plum the depths of the human genome in all its complexity and variability. Each of the current market leaders are powerful companies with diversified product lines that extend well beyond their sequencing platforms. A number of smaller companies have technologies that promise some advantage over existing systems, but whether these advantages will be great enough to merit significant market share remains very much an open question. Meanwhile, existing next-generation sequencing users are striving to expand the case for personal genomics and indeed for personalized medicine in general.