Tech News

Finding the true \$1000 genome

This April, the field of molecular biology will celebrate two remarkable anniversaries: sixty years since James Watson and Francis Crick introduced the world to the structure of DNA, and ten years since the completion of the Human Genome Project (HGP). The former taught us what genetic material looks like; the latter taught us what it says.

Today, the HGP seems almost quaint—what took 15 years and nearly \$3 billion to complete in 2001 can now be accomplished in about a day for less than \$10,000, and soon that cost will likely drop below \$1000. (Indeed, that line may already be crossed, if only the sequencing process itself is considered.) The NHGRI Genome Sequencing Program estimates its per-genome cost is \$7666. Illumina, whose HiSeq DNA sequencing systems produce the bulk of human genome sequence reads, offers its sequencing services for as little as \$4000 in bulk, while Complete Genomics offers its genomes for \$5000 or less.

“2013 is likely to be the year where we see the \$1000 genome. It’s going to become that cheap to sequence the human genome,” predicted Daniel Franklin, executive editor of *The Economist* (and nephew of DNA pioneer Rosalind Franklin), in an online video (1) accompanying the magazine’s “The World in 2013” issue. If nothing else, the barrier could fall this September when the \$10 million Archon X Prize competition takes place. To claim the prize, a team must produce 100 human genome sequences in 30 days or less, at a maximum cost of \$1000 per genome sequence (with an error rate of at least 10^{-6} and at least 98% completeness).

The concept of a \$1000 genome was first articulated at the start of the millennium. It’s a price point that a sequenced genome on par with a laptop computer or a clinical test—a psychologically pleasing, if not completely arbitrary, value. But just what does it mean?

“That is the first and best question on this topic,” says Chad Nusbaum, co-director of the Genome Sequencing and Analysis Program at the Broad Institute in Cambridge, MA. The term itself is nebulous: Is \$1000 a cost or a price? Does it include only the cost of reagents or does it fold in sample preparation and labor? And what about the bioinformatics analyses, the genome interpretation, and long-term data storage?

Is \$1000 a cost, or a price? Chad Nusbaum—co-director of the Genome Sequencing and Analysis Program at the Broad Institute in Cambridge, MA—wants to know the true cost of sequencing, not just the outer pricetag. **Photo credit: Len Rubenstein.**

The time has come to ask just what that \$1000 will buy.

The \$5000 genome

As a general rule, research and development (R&D) is most expensive at the start, when producing that first success comes at the cost of many initial failures. Subsequent attempts cost far less as failure rates fall, technology prices drop, and economies of scale kick in.

Genomics is no different. And if there is one sequencing facility that can count on economies of scale, it is the BGI in Shenzhen, China. BGI houses 128 HiSeq 2000 instruments, each capable of producing 600 gigabases (Gb) of sequence data per run. As the human genome is 3 billion base pairs long, achieving an average 30x coverage means researchers must collect around 90 Gb of sequence per genome. Thus, during each 10-day HiSeq 2000 run, scientists can, in theory, interrogate up to six human genomes, depending on coverage.

BGI declines to quote exact prices, but citing an online quote from Illumina for a New York-based medical school, Deputy

Director Xun Xu estimates a single HiSeq 2000 run costs about \$12,000 in consumables; assuming five samples per run, reagent costs run about \$2400 per genome. That represents the bulk of genome sequencing cost today, Xu says; sample preparation and library construction are relatively inexpensive and largely automated, working out to \$500 —1,000 per genome. In total, that adds up to around \$3000 for reagents and labor, plus another \$1000 or so for instrument depreciation.

Noting the apparent differences in genome pricing at different sequencing centers, Nusbaum—who also declines to quote a per-genome cost for the Broad Institute—says it is important to consider the “all-in” or “true” cost of sequencing: that is, not just the reagents, but factoring in everything from blood tubes to labor to instrument amortization. “You want to be very clear about how you are determining the cost,” Nusbaum explains. “There is a difference between the cost of doing the experiment, versus the true cost...for genomes, I think it’s important to do the ‘true cost.’”

Still, he says, even that cost excludes the upfront R&D expense required to develop the sequencers and sequencing technologies—costs that are typically included in the HGP accounting. But according to Nusbaum, the genome he can provide today is a pale shadow of that first HGP genome sequence, and lacks most long range structural information. The HGP data were assembled *de novo* (rather than merely aligned to a reference) and then each base was manually “finished”, subjected to extensive quality control, manual cross-checking and error correction, none of which occurs routinely today. “It’s not the same genome,” he says. In contrast, today’s genomes are generated by the far less complicated and expensive process of “resequencing” and alignment to a reference.

Still, Nusbaum figures he and others in the field could probably squeeze costs by a “couple factors of two” from current processes—improving automation of DNA extraction and library preparation steps, for instance. This, plus anticipated instrument technology improvements in read lengths and read densities, could get his per-genome cost down to around \$2000 over the next two years or so.

\$1000 genome; \$100,000 ‘interpretome’

When it comes to genomics, sequencing is the easy part; all that data must then be analyzed and interpreted. Quality metrics on raw sequence needs to be evaluated, data assembled and compared to a reference, and genetic variants identified. Many mature computational pipelines are now available, including the GATK package developed by the Broad Institute and the SOAP package from BGI. Both are largely automated and cost less than \$500 per genome, Xu says (though a full accounting of cost would also have to include the processing time required by computers to run these jobs).

If the goal is simply to collect a bunch of genomes to assess human genetic variability, then the process is done at this stage. But if the effort is discovery-driven in order to identify variants underlying phenotype or human disease—say, for a clinical application—those automated steps are not sufficient.

“It cannot be done by automation pipelines on a computer; you have to have a lot of range of knowledge and certain kinds of data-interpretation technical skills,” says Xu. “Those kind of people could be very expensive, and it’s not very efficient to do that, because you have to manually interpret each data point you care about.” \$1000 genomes may be in sight, but what are they worth alongside so-called million-dollar “interpretomes”? Indeed, in 2010, Washington University

geneticist and Genome Institute codirector Elaine Mardis penned an article in *Genome Medicine* entitled “The \$1000 genome, the \$100,000 analysis,” a reference to a comment she once made to then-NHGRI head Francis Collins, when NHGRI was actively funding technology development towards the \$1000 genome.

“My admonishment to Francis at the time was that he really needed to start thinking about funding bioinformatics,” recalls Mardis, “or he was going to have his \$1000 genome but

Bioinformatics bottleneck. Collecting a genome is one thing, but analyzing it, warns Elaine Mardis—codirector at the Genome Institute at Washington University—can be much more costly. **Photo credit:** Robert Boston/Washington University Public Affairs.

he would still have a million-dollar analysis.”

That isn’t hyperbole. When the institute sequenced its first tumor-normal genome pair in 2008 using next-generation technology, it cost about a million dollars to collect all those As, Cs, Gs, and Ts, and another \$600,000 to analyze the data once the reads were done. “The analyses had never been done before,” says Mardis. “We had to figure out how to do it as we went.”

Today, the Genome Institute at Washington University has more than 30 Illumina HiSeq 2000s, each costing about \$750,000. Each sample, Mardis reckons, costs about \$8000 “fully-loaded” to generate the sequencing data. But the downstream analysis costs still add up, especially for clinical applications.

On average, each tumor genome contains about 3.5 million single-base variants from a

matched normal. For each one, the variant must be logged and analyzed. Does this variant change an amino acid? Is that residue conserved? Is the variant seen in the matched normal sample, and what is its frequency in the human population? Much of that work can be automated in informatics pipelines, but the analytical interpretation of the final list is not yet easily automated.

Ultimately, however, somebody—perhaps a medical oncologist or geneticist—needs to sit down with the data and consider each novel variant to determine which, if any, might either cause a disease or represent an actionable target. This may involve researching the genes and digging into the literature, all of which takes time. And money.

Distinguished Scientist Gary Schroth, at Illumina, says bioinformatics costs largely accrue at this step. “That’s where you have to consider the vastness of the human genome. Three billion letters is a lot to sift through.” Even after multiple rounds of filtration, the list of variants can number in the tens or hundreds. “You can’t have a bioinformaticist or geneticist staring at that list for very long and keep the total cost under \$1000,” he says.

To get a sense of the complexity of the problem, consider the Laboratory for Molecular Medicine (LMM) at the Partners Center for Personalized Genetic Medicine in Cambridge, MA. Director of Operations Lisa Mahanta has a single HiSeq 2000 in her lab, which she uses to run panels of 11–71 genes from as many as 80 patients at a time, searching for causes of deafness and cardiac problems, among other ailments. This is targeted sequencing and, according to Mahanta, “ridiculously small” in light of current genomics capabilities.

Still, it takes Mahanta’s lab six weeks to return a 71-gene otolaryngology panel, and not because of backlog. The next-gen sequencing is done in about two weeks, and then it takes four to five days to pass the data through the lab’s automated bioinformatics pipeline. (Data transfer alone—it’s one terabyte of information—takes 24 hours.) The pipeline churns through the raw data, eliminating common variants known to be harmless or irrelevant. What’s left, anywhere from 5 to 20 previously undocumented variants, now has to be scrutinized: first through independent validation by Sanger sequencing, and then with literature searches to determine potential variant impact.

It should come as no surprise, then, that the lab’s recently launched genome sequencing and analysis service requires some time for turnaround. For \$9000, the lab will outsource the genome sequencing to Illumina after which LMM personnel will interpret the resulting data, focusing specifically on “medical anomalies” with unclear genetic

NEXTflex™ NGS Kits

Bioo Scientific has the kit to match all of your NGS needs. Our line of innovative kits offer sensitivity, flexibility and speed to NGS library prep for the Illumina, Ion Torrent and SOLiD sequencing platforms. Take a look:

DNA-Seq

Best Ligation Efficiency Available

ChIP-Seq

As Low As 1 ng

Bisulfite-Seq

Complete Kit

Amplicon-Seq

PCR-based Target Selection

Methyl-Seq

MeDIP & MeCAP

RNA-Seq

10 ng of RNA

Small RNA-Seq

Bias Reducing Technology

Directional RNA-Seq

Strand Specific

PCR-Free DNA-Seq

No GC/AT Bias

Pre-Target Capture

Library Prep, Barcodes & Blockers

Visit **BiooNGS.com** to discover and fulfill all of you Next-Gen Sequencing needs.

The NGS Experts™

Illumina® Compatible • Ion Torrent Compatible
SOLiD Compatible

causes. All total, Mahanta anticipates taking four to six months to process each sample.

Of course, time and knowledge will speed the process: once a troublesome genomic variant has been identified and characterized, it never need be characterized again, notes Harvard geneticist George Church. In other words, once a mutation and its consequence (and potential treatment) have been defined, those details can be codified into an algorithm that can scan variant lists automatically.

Church's lab has developed a piece of software to go along with his Personal Genome Project for this very purpose. Called Get-Evidence, the software is a collaborative research tool in which researchers and physicians can enter information about novel genetic variants and their potential clinical impact. Once a variant has been added, "then no one has to do that again—at least until the next breakthrough in that gene," Church notes. "The second person is basically free. And the third person and the fourth person; the downstream cases are free."

Church notes that one participant in his PGP project found out they carry a previously unrecognized mutation in the JAK2 gene, which causes clotting problems and dark spots on the retina. The solution: baby aspirin. "That did involve some research. But the next time somebody comes in with a JAK2 mutation, getting to the point of clinician involvement will be easier and more algorithmic."

Genome sequencing, STAT

When it comes to the \$1000 genome, another critical question is whether the price point really matters at all. A desperate patient or family may not mind paying \$3000 for a genome, as opposed to \$1000. Neither, perhaps, will insurance companies. In some sense, then, the notion of a \$1000 genome is arbitrary.

But for researchers like Nusbaum, the cost difference does matter. "The difference between \$1000 and \$5000 is five times as many samples, and that's five times more [statistical] power," he says. "Even a two-fold difference in the number of samples is a real difference in how quickly you can get to your answer."

In early 2012, Illumina released an update of its HiSeq 2000, the HiSeq 2500. The HiSeq 2500 can generate about enough data to produce a 30x genome (~120 Gb) in 24 hours. Recently, Stephen Kingsmore of the Center for Pediatric Genomic Medicine at Children's Mercy Hospital in Kansas City, MO, published a paper demonstrating that it was possible to couple this instrument with some clever bioinformatics processing to perform pediatric genomic diagnostics in just over two days.

The technique, called STAT-Seq, costs \$13,500 and returns results in 50 hours. That includes 4.5 hours for sample preparation, 26 hours for sequencing, and 15 hours for base calling, sequencing alignment, and variant calling. The final few hours comprise data interpretation.

Following sequencing, a home-built software pipeline filters the ~4 million genetic variants found in each genome by their location in protein-coding genes, allele frequency, protein class, disease likelihood, and so on. These data are then passed to a tool called SSAGA.

SSAGA "is a new clinicopathological correlation tool that maps the clinical features of 591 well established recessive genetic diseases with pediatric presentations to corresponding phenotypes and genes known to cause the symptoms," according to the STAT-Seq publication (2). In short, says Emily Farrow, a research scientist at Children's Mercy Hospital and coauthor of the paper, clinicians input a patient's clinical symptoms so the software can narrow its search field. "We're filtering out the noise, so to speak, and really narrowing down the area that we have to look at," she explains.

Farrow says her lab has now processed seven STAT-Seq families (22 samples total). By comparison, in 2012 her lab sequenced more than 1,200 exomes and some 500 targeted gene panels.

Researchers hoping to take advantage of processes like STAT-Seq can send their samples to Children's Mercy, or pass them to a more local core facility. In many cases, basic bioinformatics analyses no longer represent a financial or even intellectual bottleneck: cloud-based, automated pipelines like DNAnexus and Illumina's BaseSpace service have, in some sense, democratized the field. All of which means the power of the human genome will finally make its way to the people who paid for it. Not bad for 10 year's work, or even 60.

References

1. "Another layer on reality," *The Economist*, January 4, 2013. www.economist.com/blogs/newsbook/2013/01/year-ahead-landmarks-edition.
2. **Saunders, C.J. et al.** 2012. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. *Sci. Transl. Med.* 4:154ra135.

Written by Jeffrey Perkel, Ph.D.

BioTechniques 54:71-74 (February 2013)
doi 10.2144/000113988

To purchase reprints of this article, contact:
biotechniques@fosterprinting.com