Opinion

Anticipating the $1,000 genome

Elaine R Mardis

Address: Genome Sequencing Center, Washington University School of Medicine, 4444 Forest Park Boulevard, St. Louis, MO 63108, USA.
Email: emardis@wustl.edu

Published: 27 July 2006

Genome ***Biology*** 2006, **7:**112 (doi:10.1186/gb-2006-7-7-112)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/7/7/112

Abstract

A new generation of DNA-sequencing platforms will become commercially available over the next few years. These instruments will enable re-sequencing of human genomes at a previously unimagined throughput and low cost. Here, I examine why the $1,000 human genome is an important goal for research and clinical diagnostics, and what will be required to achieve it.

In April 2003, 50 years after Watson and Crick first described the chemical structure of DNA [1], the DNA sequence that makes up the human genome was proclaimed "essentially complete" [2]. Following on from this, in October 2005, the project of the HapMap consortium to identify the locations of one million common single-nucleotide polymorphisms (SNPs) in the context of this reference human genome sequence were completed [3]. Accomplishing these two genomic milestones required the development, testing and implementation of technology platforms that could produce data at previously unprecedented throughputs, as well as of the bioinformatics tools and computational capabilities to analyze the resulting data and to interpret it in meaningful ways. It is this critical interplay of technology and bioinformatics that will usher in the next era of genome sequencing technology, commonly referred to as 'the $1,000 genome' on the basis of its targeted price per genome in US dollars; today, we find ourselves poised at the brink of this era. In this paradigm, the cost of determining an individual genome sequence would fall to a price of around $1,000, placing it firmly in the realm of advanced clinical diagnostic tests. As a result, determining a person's genome sequence might ultimately become an important first step upon entering a health insurance network or a health care provider's practice, akin to determining their height, weight and blood type, for example.

Why aim for a $1,000 genome?

Given this paradigm, one might ask why a $1,000 genome is an important or necessary goal to achieve. Fundamentally, even with the significant achievements of the HapMap Project [3], we have little context for comprehending the breadth of human genomic diversity, encompassing all types of variation beyond common single-nucleotide variants. Capturing this range of diversity, at the current cost of around $10-20 million per genome sequence, places it firmly outside the bounds of fiscal reality. Yet without this 'baseline', genome scientists and statisticians lack a contextual framework within which to evaluate the genome-characterization projects that are presently under consideration. Some such projects are described here.

As recently outlined by Hartwell and Lander [4], the comprehensive characterization of all differences in DNA sequence and chromosome organization between cancer genomes and their corresponding normal genomes should have a significant impact on our understanding of the spectrum of genomic alterations that underlie malignancy. Similar projects are currently being championed by the National Human Genome Research Institute [5] to provide sequencing and analysis of focused regions in the genomes of individuals with mapped, uncloned, autosomal Mendelian disorders, X-linked disorders and specific common diseases. These and similar projects require a comprehensive combination of focused genome re-sequencing that targets specific 'suspect' genes, a characterization of chromosomal amplifications and/or deletions, comprehensive gene-expression profiling, and karyotyping, when possible. Taken together, this broad-brush approach will potentially further our understanding of a particular disease, with genome- and transcriptome-level characterizations that identify the shared somatic changes associated with the phenotype.

Projects such as these are 'discovery' efforts: they aim to characterize a relatively small (typically not statistically significant) number of affected individuals or samples, first evaluating genome alterations for each individual and then establishing shared, statistically significant somatic alterations for the group. A subsequent phase would follow these discovery efforts that would aim to evaluate the genome alterations ascertained from the smaller group in thousands of similar samples, using high-throughput, inexpensive assays that can provide the necessary statistical power to establish (or refute) the contribution of each mutation. Ultimately, this approach will yield genome-wide sequence-based biomarkers; the mutations, copy-number changes, rearrangements and other alterations that are diagnostic for the disease in question. With knowledge of these biomarkers in hand, the availability of rapid and inexpensive human genome re-sequencing (so-called as the reference sequence is already known) heralds an era in which re-sequencing becomes a clinical diagnostic or prognostic tool.

The technologies currently available

Placing the task of developing a $1,000 genome technology in context requires a quick overview of the current state of the art in genome re-sequencing technology. There are already several commercial platforms that can evaluate known human SNPs in a rapid and massively parallel manner, including those offered by Illumina [6] and Affymetrix [7]. These technologies predominately use DNA:DNA hybridization and are ideal for genotyping known SNPs and identifying copy-number differences, but are unsuitable for discovery of novel SNPs or other polymorphisms (insertions or deletions - 'indels' - or rearrangements). For novel polymorphisms, DNA sequencing of products obtained by PCR from genomic DNAs, using primers designed to match selected regions of the reference human genome, represents the best technology to date. A PCR-based re-sequencing approach has limitations but has been implemented by brute force for several large-scale projects [8-13].

In many ways, this situation is somewhat reminiscent of the early days of large-scale genome sequencing, in that many of the components for automation, methodology and bioinformatics are, in my opinion, being developed in a 'just-in-time' fashion. We should hope, however, to be saved from this path because PCR-based re-sequencing of the human genome is ultimately limited by several factors. Selecting unique primers for every region of interest in the human genome is frankly not possible, because of SNPs and/or repetitive content that reduce the stability of primer-annealing sites near exons, and because of gene families and pseudogenic regions. Even when unique primers can be designed, data-quality issues frequently arise in PCR and/or sequencing as a result of structural features (high GC content or homopolymer and dinucleotide runs). Furthermore, the overall cost of PCR-based re-sequencing is about 2.5 times that of clone-based sequencing, primarily because of the expense of PCR compared with high-throughput subclone isolation. This cost, coupled with a higher inherent failure rate for PCR from genomic DNA templates, further increases re-sequencing costs and timelines, as either more patients must be sequenced to achieve mutation discovery across samples, or additional attempts at PCR and sequencing of fewer samples must be successfully completed to obtain the necessary data.

The technology needed to reduce the cost of genome sequencing

What, then, are the general features of a technology platform that can overcome the inherent limitations of our PCR-based re-sequencing paradigm and deliver a genome for significantly less cost and within a much faster time frame than at present? We can examine a generic *de novo* sequencing pipeline for clues (Figure 1a). One of the least automated and most error-prone steps at present is preparation of genomic subclone libraries. An ideal re-sequencing platform would therefore remove the sub-cloning step and sequence directly from a relatively small input quantity (several micrograms, for example) of genomic DNA. A side benefit of skipping conventional sub-cloning is that cloning bias, which can skew representation of a genome, is avoided (which is not to say that other types of bias might not be introduced). Another significant benefit is that subsequent clone-specific steps, such as picking clones from agar plates, isolation of individual subclone DNAs and sequencing reactions in microtiter plates, are eliminated along with much of the automation required to perform them. Contrast the more complex workflow in Figure 1a to a generic massively parallel instrument workflow in Figure 1b, for example.

Massively parallel data production in a single duty cycle of an instrument is another critical component for success, and the amount of data required to be produced depends solely on the size of the genome. Currently, a capillary sequencing instrument produces around 1.35 megabases (Mb) of sequence per 24-hour duty cycle (20 x 96 samples daily at an average read length of 700 bases), which means that about 4,500 duty cycles are needed to produce the raw base equivalent of a diploid human genome (I estimate roughly 12 years). By contrast, each '$1,000 genome' instrument should require about a month per diploid genome of raw data, in order to provide a suitable 'discovery' research platform (more realistically 1-5 days per genome would be required, once whole-genome re-sequencing moves into the clinical laboratory).

Specifications for the accuracy of base-calling, the length of sequence reads and the ability to produce sequence from the paired ends of each DNA fragment are equally critical. For re-sequencing, unequivocal placement of a read pair onto the reference genome and determining whether the fragment ends, so aligned, encompass a region of 'difference' (for example a

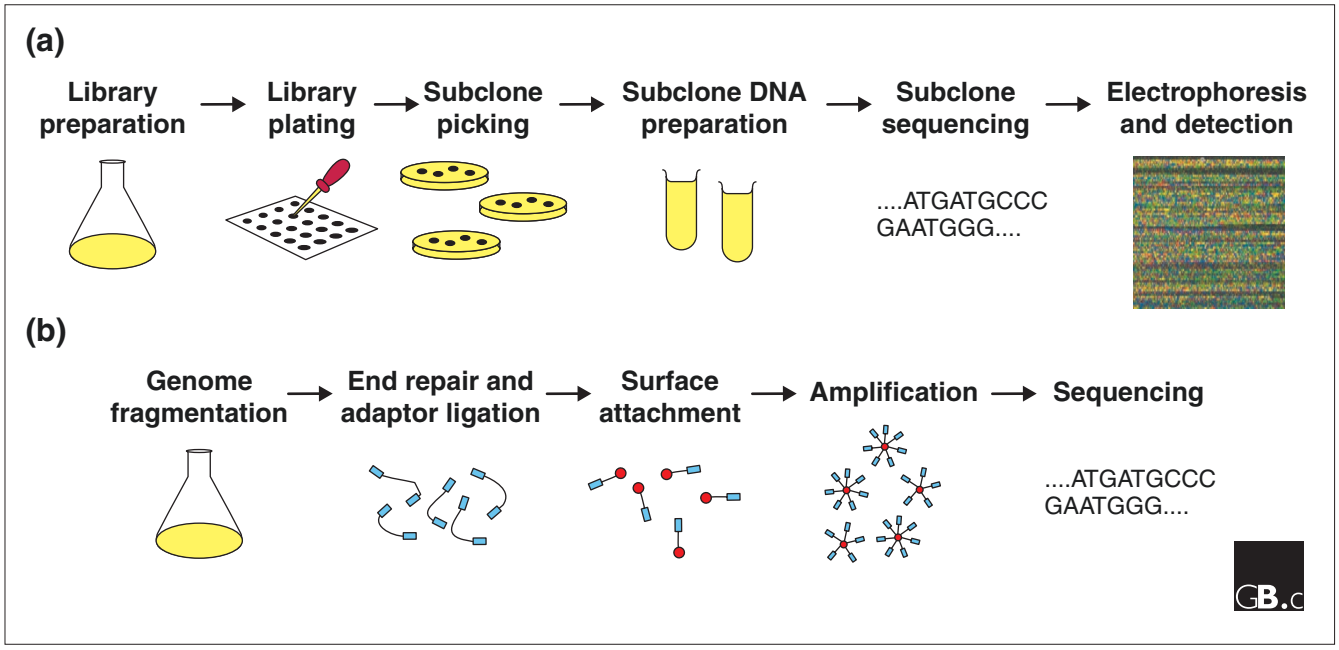

Figure 1
Comparison of conventional and massively parallel sequence pipelines. Both pipelines begin with a DNA fragmentation step. **(a)** The steps in a conventional genome-sequencing pipeline, most of which require dedicated automation and processing in a 384-well format. DNA fragments are subcloned into bacterial vectors and introduced into bacterial cells to prepare a library covering the whole genome. The transformed cells containing subclones are plated and grown and then harvested by robotic picking, and the DNA from each one is isolated and sequenced. The sequence is visualized by loading onto a capillary sequencing instrument. **(b)** The steps in a generic massively parallel genome-sequencing pipeline. Genomic DNA fragments first undergo end repair to provide blunt ends for adaptor ligation and then have specific adaptors ligated to their ends that contain priming sites for PCR and sequencing. The adaptor-ligated fragments are then hybridized to complementary adaptors that are fixed to a surface (a slide or bead), and then *in situ* PCR amplification is used instead of bacterial amplification *in vivo*. Sequencing reactions of the surface-amplified fragments take place on the surface. The sequence is visualized using either luciferase (pyrosequencing) or fluorescence reporting that is detected by a CCD camera.

mutation, indel, rearrangement, translocation, or other difference) relative to the reference are pivotal requirements. Alignment and determination of differences in read pair placements are directly affected by the accuracy of base-calling, by read length, and by the capability to obtain sequence from both ends of a genomic fragment. Maximizing read length and obtaining paired end reads could ideally converge to provide long reads across an entire fragment (and, by inference, to provide haplotype information); but even so, the accuracy of base-calling would determine the efficiency of re-sequencing (including the required coverage of the genome needed), which directly determines its cost. Base-calling accuracy is influenced by the reaction chemistry and by the algorithms developed to extract base-calls from raw data, among other factors. In addition to these considerations, it is also necessary to consider the ability of reaction chemistry and reaction conditions to overcome any secondary-structure effects in the templates that might truncate reads and affect coverage.

Current massively parallel sequencing technologies

So, are we anywhere near this lofty goal of the $1,000 genome? At present, the short answer is in fact 'no'. Given the level of interest in the goal, however, a significant amount of activity and several innovative, interdisciplinary technologies are now being pursued. I do not aim to describe these new technologies in detail here; a recent review [14] analyzes many of them comprehensively and comparatively. Of such technologies, only one massively parallel sequencing instrument, the GS-20 from 454 Life Sciences has so far achieved commercial availability [15]. This instrument uses pyrosequencing (the enzymatic sequencing method that reports nucleotide incorporations using the reporter firefly luciferase [16-18]), of genomic fragments that have been captured and amplified on agarose beads to produce up to 20 Mb of sequencing data (in 100 base-pair read lengths) per 4 hour instrument run.

At its present base-calling accuracy, read length and cost per run, a human genome cannot be sequenced for even $100,000 using the 454 instrument, but it nevertheless represents an important first step toward the $1,000 genome goal. Realistically, the 454 platform will continue to improve, providing longer read lengths and higher base-calling accuracy, and the commercial pressures of other massively parallel instruments will drive down the costs per run. The imminent entry of another massively parallel

platform from Solexa Ltd, due this summer (2006), will help to fuel this trend [14]. Others will follow in the ensuing months and years. It is therefore conceivable that we are quite close (within 1-2 years) to having instruments suitable for the research laboratories that will provide the 'discovery' setting described earlier.

The challenges of having so much data

If, for the sake of argument, we assume that novel, massively parallel platforms will be developed and implemented to rapidly and inexpensively re-sequence human genomes, there are related concerns to point out. Namely, are the challenges posed by the enormous data-generation capabilities and by the analysis of these data also being anticipated? The tracking, storage and submission of the data from such platforms will certainly pose significant challenges, even for large sequencing centers. Similarly, intelligent algorithms that use computing resources efficiently must be developed for aligning and mapping re-sequencing data onto the reference genome, evaluating the aligned sequences for mutations, indels, rearrangements, and so on, and reporting genome-wide alterations in an annotated and organized fashion. Most challenging of all, it will be of critical importance to develop meta-analyses and statistical analysis tools that integrate across disparate data types, such as whole-genome re-sequencing data, gene-expression data, copy-number alterations, biochemical pathway information, clinical parameters (age, sex, diagnosis and treatment), outcomes, and so on, and thereby enable researchers to collectively interpret these data for all samples in a study and to form testable hypotheses from this discovery phase. These bioinformatic challenges may well be as daunting as the development of the instruments themselves, and ultimately they will determine whether, once ready, the instruments can be utilized immediately and effectively.

Incorporating the new and ever-increasing functional knowledge of the non-genic portions of the genome, which often comes from incongruent sources, into meta-data analyses will also determine whether we can make sense of sequence changes that are found outside exons and known regulatory regions. After all, the primary reason that current PCR-based re-sequencing approaches typically focus on exons is that the impact of a sequence change on an encoded protein can be readily deciphered. Follow-on functional studies of the altered protein can characterize the impact of a mutation on function, can suggest other pathway-related effects of the mutation, and ultimately may identify treatments to counteract the mutation. It is at this interface - where genome-scale sequence information and its consequences begin to have an impact on clinical practice, directing treatment, indicating genetic predisposition to disease and predicting outcomes - that applications of the $1,000 genome concept in a clinical context begin to take shape. But first, much of the aforementioned discovery phase has to take place. Without this, the contextual framework for understanding an altered genome is not there.

Finally, once these discovery phases are completed for specific diseases, confirmed in thousands of affected individuals and subjected to the rigor of approved clinical tests, they must enter into the collective practice of medicine. Given that this paradigm shift will require changes both in medical education and in acceptance by health insurance providers and ethicists [19], we have a long way to go. Even if we can never comprehensively interpret the entirety of each re-sequenced genome, the efforts under way to revolutionize DNA sequencing, and to dramatically decrease its cost so that multitudes of human genomes can be sequenced for discovery and ultimately for clinical means, are well worth it. Simply put, having this capability not only facilitates our efforts to understand the genomic basis of disease, but also opens our minds to questions not yet imagined.

References

1. Watson J, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.** *Nature* 1953, **171:**737-738.
2. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome.** *Nature* 2004, **431:**931-945.
3. International Human Genome Sequencing Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437:**1299-1320.
4. **Hartwell LH, Lander ES: Report to National Cancer Advisory Board: NCAB Working Group on Biomedical Technology** [http://deainfo.nci.nih.gov/Advisory/ncab/sub-bt/NCABReport_Feb05.pdf]
5. **The Medical Sequencing Program at the National Human Genome Research Institute** [http://www.genome.gov/15014882]
6. **Illumina** [www.illumina.com]
7. **Affymetrix** [www.affymetrix.com]
8. Lalani S, Safiullah AM, Fernbach SD, Harutyunyan KG, Thaller C, Peterson LE, McPherson JD, Gibbs RA, White LD, Hefner M, *et al.*: **Spectrum of CHD7 mutations in 110 individuals with CHARGE Syndrome and genotype-phenotype correlation.** *Am J Hum Genet* 2006, **78:**303-314.
9. Jiang J, Paez JG, Lee JC, Bo R, Stone RM, DeAngelo DJ, Galinsky I, Wolpin BM, Jonasova A, Herman P, *et al.*: **Identifying and characterizing a novel activating mutation of the FLT3 tyrosine kinase in AML.** *Blood* 2004, **104:**1855-1858.
10. Paez J, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, *et al.*: **EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy.** *Science* 2004, **304:**1497-1500.
11. Wilson RK, Ley TJ, Cole FS, Milbrandt JD, Clifton S, Fulton L, Fewell G, Minx P, Sun H, McLellan M, *et al.*: **Mutational profiling in the human genome.** *Cold Spring Harb Symp Quant Biol* 2003, **68:**23-29.
12. Woloszynek JR, Rothbaum RJ, Rawls AS, Minx PJ, Wilson RK, Mason PJ, Bessler M, Link DC: **Mutations of the SBDS gene are present in most patients with Shwachman-Diamond syndrome.** *Blood* 2004, **104:**3588-3590.
13. Ley TJ, Minx P, Walter MJ, Ries RE, Sun H, McLellan M, DiPersio JF, Link DC, Tomasson MH, Graubert TA, *et al.*: **A pilot study of high-throughput, sequence-based mutational profiling of primary human acute myeloid leukemmia cell genomes.** *Proc Natl Acad Sci USA* 2003, **100:**14275-14280.
14. Metzker ML: **Emerging technologies in DNA sequencing.** *Genome Res* 2005, **15:**1767-1776.
15. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.
16. Hyman ED: **A new method of sequencing DNA.** *Anal Biochem* 1988, **174:**423-436.

17. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P: **Real-time DNA sequencing using detection of pyrophosphate release.** *Anal Biochem* 1996, **242:**84-89.
18. Ronaghi M, Uhlen M, Nyren P: **A sequencing method based on real-time pyrophosphate.** *Science* 1998, **281:**363-365.
19. Robertson JA: **The $1000 genome: ethical and legal issues in whole genome sequencing of individuals.** *Am J Bioeth* 2003, **3:**W-IF1. doi: 10.1162/152651603322874762.