

NEWS AND COMMENTARY

Technology

A genome sequencing center in every lab

Michael E Zwick

European Journal of Human Genetics (2005) 13, 1167–1168.
doi:10.1038/sj.ejhg.5201504

Two recent papers suggest that a revolution in DNA sequencing is at hand^{1,2} and explore compelling alternative DNA sequencing technologies that show promise to change the way human genome resequencing occurs.

Scientific revolutions, nearly by definition, often change the research landscape in unanticipated ways. Completing the human genome project-required multiple innovations,³ not the least of which was the large-scale application of gel electrophoresis and Sanger sequencing chemistry⁴ in a small number of highly automated industrial genome sequencing centers. One upshot of the human genome project is the future vision of 'individualized medicine.' Genomics technologies that could resequence individual genomes might provide genetic explanations for phenotypic variation in disease susceptibility and drug response, leading to improved patient care.

In the course of sequencing a human reference genome, the traditional industrial model has consistently produced ever-greater quantities of data at a reduced cost. However, it seems unlikely that current approaches are sufficiently scalable to fulfill the promise of individual genomic medicine.⁵ Realizing this future possibility will seemingly require another revolution in DNA sequencing technology: a revolution that these new papers indicate might be upon us.

Genome sequencing can be divided into four steps: (1) break a large DNA polymer into smaller fragments, (2) isolate and amplify single fragments, (3) determine the fragment sequence and (4) perform automated data quality assessment and

sequence assembly to reconstruct the original DNA polymer sequence. Traditional DNA sequencing protocols use libraries of cloned fragments and Sanger sequencing chemistry to accomplish the first three steps. Alternative approaches accomplishing these same tasks are presented in these two papers.^{1,2}

Shendure *et al*¹ employed a multiplex polymerase colony, or polony, protocol to generate approximately 1.6 million fragments that are each 135 basepairs (bp) in length (step 1). Each fragment has 100 bp in common, and contains two mate-pair tags of 17 and 18 bp from the genome being sequenced. The tags consist of random genome sequences selected to be approximately 1000 bp apart. Each fragment is attached to a separate 1 μ m bead, amplified using a water-in-oil emulsion PCR protocol (step 2), and immobilized in a 1.5 cm² acrylamide gel. The fragments are sequenced in parallel via a ligation protocol that uses four dyes to identify each possible base (step 3). In all, 13 basepairs from each tag, or a total of 26 bp, are determined for each fragment. It is striking that the protocols described were implemented with off-the-shelf instrumentation and reagents suggesting

that in principle, it is possible for single laboratories to perform these assays.

Margulies *et al*² also avoid traditional library construction. They shear an entire genome to generate 300 bp long DNA fragments (step 1), add specialized common adaptors, capture individual fragments on beads, and clonally amplify each fragment within an emulsion (step 2). The beads are then distributed across open wells of a fiber-optic slide and pyrosequencing chemistries are used to determine the sequence of each fragment (step 3). Average read lengths of 100 bp were reported. The authors suggest that mate pair reads are possible by sequencing the same fragment on a bead from different directions. A commercial system using this approach, that requires limited laboratory space and personnel to operate, is currently available.

High-throughput methods of data generation require quantitative measures of data quality (step 4). Traditional DNA sequencing uses Phred scores to determine the probability of a basecalling error.^{6,7} A similar approach can be used to evaluate these two technologies. The Shendure/ Porreca group estimated data quality by resequencing an *Escherichia coli* MG1655 genome expected to differ relative to the reference sequence at a number of known and unknown sites. In all, 70% of the 3.3 Mb genome had 4 \times or greater coverage. No substitution errors were observed, implying an error rate <1 per 3.3 million bases sequenced, or a Phred score of 65 (Table 1). Margulies *et al* sequenced a *Mycoplasma genitalium* (580 kb) genome to estimate data quality. At high coverage sites (98.2% of the genome), the error rate was 3.0E-6, which corresponded to a Phred score of 55 (Table 1). Remarkably, Margulies *et al* were able to resequence the genome eight times for 40-fold coverage in only 243 min of instrument run time.

Table 1 Representative error rates and their quantification

| Error rate | Phred score | Expected number of errors (5 Mb region) | Comment |
|------------|-------------|---|-------------------------------------|
| 1.0E-03 | 30 | 5000 | — |
| 1.0E-04 | 40 | 500 | Bermuda standard |
| 1.0E-05 | 50 | 50 | Finished sequence |
| 3.0E-06 | 55 | 15 | Margulies <i>et al</i> ² |
| 3.0E-07 | 65 | 2 | Shendure <i>et al</i> ¹ |

Phred scores are calculated as Phred = $-10 \log_{10}$ (error rate).

So it seems both methods can produce very high-quality data.

What do these approaches have in common? They perform sample preparations directly on entire genomes avoiding slower and more expensive clone-base methodologies. They both use similar emulsion PCR to clonally amplify single fragments. While both papers report raw accuracies and read lengths (26 bp,¹ 100 bp²) significantly lower than Sanger sequencing (~700 bp), they compensate by generating many more sequences (~1 600 000,¹ ~300 000²) as compared to Sanger sequencing (96), per single system run. Perhaps most impressive, the cost per high-quality base with either technology is roughly an order of magnitude lower than that of conventional sequencing. If one considers the lower costs associated with limited infrastructure and personnel, these approaches become even more attractive. Future improvements in the library density and read length for both technologies will further reduce cost while increasing throughput.

These studies focused on resequencing relatively small bacterial genomes. However, the methods of library construction and sequencing are general, so they are relevant to human genetics. Human genomic regions containing putative disease-causing alleles are typically identified through family-based linkage or case-

control whole genome association studies. These regions are often roughly the size of bacterial genomes. In the near term, approaches enabling specific DNA isolation from localized regions in the human genome, such as that under a 5 Mb linkage peak, could be sequenced efficiently and accurately in single laboratories using these technologies (see expected number of errors in Table 1). Since variation detection in human genetics is often rate limiting, these advances have the very great potential to significantly speed the identification of human disease-causing variants.

Meeting the longer-term goal of a \$1000 genome will require further improvements in scaling these or other technologies whose cost per high-quality base is far lower than traditional sequencing technologies. The notion that thousands of laboratories could generate genomic sequence at rates meeting or exceeding that of a conventional sequencing center will surely cause a revolution itself. For example, many traditional software/statistical packages used to map human disease traits already struggle with genomic data sets, and will face similar problems with large genome sequencing data sets. Developing efficient algorithms and computing infrastructure that can meet the challenges of handling, storing, exploring and analyzing such enormous data sets will prove formidable. Clearly, even after

the completion of the Human Genome Project, the genomics revolution continues to advance rapidly, changing the perception and practice of human genetics and the potential role of genomics technologies in medical practice ■

Dr ME Zwick is at the Department of Human Genetics, Emory University School of Medicine, 615 Michael Street, Suite 301, Atlanta, GA 30322, USA.

E-mail: mzwick@genetics.emory.edu

References

- 1 Shendure J, Porreca GJ, Reppas NB *et al*: Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005.
- 2 Margulies M, Egholm M, Altman WE *et al*: Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005.
- 3 Collins FS, Morgan M, Patrinos A: The Human Genome Project: lessons from large-scale biology. *Science* 2003; **300**: 286–290.
- 4 Sanger F, Air GM, Barrell BG *et al*: Nucleotide sequence of bacteriophage phiX174 DNA. *Nature* 1977; **265**: 687–695.
- 5 Shendure J, Mitra RD, Varma C *et al*: Advanced sequencing technologies: methods and goals. *Nat Rev Genet* 2004; **5**: 335–344.
- 6 Ewing B, Green P: Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 1998; **8**: 186–194.
- 7 Ewing B, Hillier L, Wendl MC, Green P: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 1998; **8**: 175–185.