Faster human genome sequencing

Yingrui Li & Jun Wang

Advances in parallelization allow a human genome to be sequenced using single-molecule technology.

Genome sequencing has reached another milestone: the sequencing of a human genome using single-molecule technology (Fig. 1a). This feat, reported by Pushkarev et al. in this issue, was achieved by three researchers in a matter of weeks for an estimated reagent cost of $48,000, highlighting the rapid pace of technology development in the sequencing field and bringing us a step closer to the day when genome analysis is routinely applied in personalized medicine.

Pushkarev et al. used a new sequencer from Helicos BioSciences, one of a growing number of cutting-edge sequencing instruments that includes Roche 454’s GS FLX Titanium, Illumina’s Genome Analyzer II (ref. 4) and Applied Biosystems’ SOLiD 3.0 (ref. 5). All of these machines are based on similar cyclic-array sequencing principles—‘sequencing-by-synthesis’ or ‘sequencing-by-ligation’—but each uses distinct chemistries, leading to different performance characteristics (Table 1).

Rapid advances in the speed, efficiency and cost of sequencing have led to a recent flurry of published human genomes, including those of Craig Venter, Jim Watson, a Han Chinese, a Yoruban and two Koreans, as well as the sequence of a cancer cell (Fig. 1b), and this number is expected to increase rapidly in coming years. In addition to strings of bases, these reports have revealed millions of single-base differences between individuals, or single-nucleotide polymorphisms (SNPs), and hundreds to thousands of instances of ‘structural variation’, a term that includes insertions, deletions, rearrangements and copy-number differences. Cataloging this variation should lead to better understanding of the genetic differences between populations, which might be

Yingrui Li and Jun Wang are at the Beijing Genomics Institute at Shenzhen, Shenzhen, China, and Jun Wang is in the Department of Biology, University of Copenhagen, Copenhagen, Denmark. e-mail: wangj@genomics.org.cn

Figure 1 Single-molecule genome sequencing. (a) Single-molecule sequencing as implemented by the Helicos sequencer. Unamplified DNA is fragmented and denatured to form sequencing templates, tailed with poly(dA) and hybridized to the flow cell surface with tethered poly(dT) anchors (green), which also serve as primers. The sequencing process involves cyclically incorporating fluorescently labeled nucleotides to nascent complementary strands, imaging the labeled nucleotides and, finally, cleaving off the fluorescent labels. (b) A survey of published human genome sequences. The size of each circle represents the estimated cost of sequencing reagents (data from Supplementary Table 1 of Pushkarev et al.). *Cost and runs for the first human genome sequences are from the public sequencing project. **Cost and runs of the two Korean genome sequences are estimates based on their use of technology similar to that used for the Yoruban and Chinese genomes.
accuracy, the results reported by Pushkarev et al. should be noted. First, although single-molecule sequencing is expected to outperform other sequencing platforms in throughput and molecule sequencing is expected to outperform single-molecule sequencing should surpass variants with an accuracy in line with the best

Several issues with the approach of Pushkarev et al.1 should be noted. First, although single-molecule sequencing is expected to outperform other sequencing platforms in throughput and accuracy, the results reported by Pushkarev et al.1 still fall short. The Helicos sequencer produced ~2.5 billion bases of mappable (that is, usable) data per day, which is about equal to the throughput of the Genome Analyzer II or the SOLiD 3.0 in research laboratories. In addition, the SNP error rate (at ~30-fold sequencing passes) was 0.2% when evaluated against array genotyping—similar to what was achieved last year by the predecessor to the Genome Analyzer II (Illumina’s Genome Analyzer I).

Second, the read length reported by Pushkarev et al.1 is relatively short, 32 bp on average, and the read insertion-deletion (or indel) error rate—that is, the rate at which extra nucleotides are introduced or missed in the read sequences—is ~4%. This limits the usefulness of this instrument for large-scale de novo genome assembly, which requires either long reads or low error rates that have very few indel errors. Even for sequencing applications, genomic indel identification would be a difficult challenge, although performance in detecting this type of polymorphism was not discussed in the paper.

Third, all sequencing technologies are undergoing rapid development, reducing the costs of human genome sequencing. Both the Genome Analyzer II and the SOLiD 3.0 now have 30 Gbp production per instrument run, and are expected to achieve ~90 Gbp in about 6 months. Recently, Illumina announced a personal genome sequencing service for $48,000.

Taken together, these points suggest that the current Helicos instrument may not be more competitive than other approaches. Regardless, the scientific community has long been looking forward to seeing single-molecule technology fulfill its potential. In support of the technology’s promise, Pushkarev et al.2 demonstrate evenness of coverage substantially better than that of other platforms at 30-fold sequencing passes. In addition, accuracy of the relatively early form of the commercialized instrument is already on par with that of other technologies that have been commercially available for a longer time with many upgrades.

What should be the future goals of sequencing technology development? If the purpose of genome sequencing is to discover genetic variation, then a useful technology must be able to identify the different types of variation, including SNPs, insertions and deletions, structural variation and novel sequences of any length, on both sets of chromosomes. Such comprehensive analysis can be achieved only by de novo diploid genome assembly—a process that is sensitive to read length, base accuracy and error type. Thus, it will be important to improve these performance characteristics in future sequencing instruments.

Moreover, as Pushkarev et al.1 note, every cell in an individual may contain genetic variation that affects cellular function, and such genomic heterogeneity could be relevant to cancer and other complex disorders. As single-molecule technologies are best suited to measuring the genetic variation in single cells, technology development should aim for efficient single-molecule DNA extraction in a single cell and sequencing methods capable of reading multiple diploid genomes affordably for personalized medicine.

### Silencing prostate cancer

Ivanka Toudjarska & Antonin de Fougerolles

Systemic delivery of an siRNA–aptamer chimera leads to prostate cancer regression in mice.

Harnessing RNA interference therapy has the potential to modulate targets that are undruggable by conventional agents. Small interfering RNA (siRNA) can be generated to specifically silence any endogenous or viral mRNA, but efficient delivery of siRNA to a particular cell type or organ remains a key challenge.1 In this issue, Giangrande and colleagues2 improve the delivery and activity of a previously described RNA aptamer–siRNA conjugate directed against prostate cancer2. Whereas the earlier conjugate required intratumoral injection for efficacy, the new, optimized conjugates achieve growth inhibition of prostate cancer xenografts after systemic injection.

Any approach to systemic delivery of siRNA must address two fundamental issues: (i) delivery of intact siRNA to a particular target organ or cell type and (ii) transfer of the siRNA to the

### Table 1 Comparison of single-molecule and other commercialized second-generation sequencing platforms

<table>
<thead>
<tr>
<th>Sequencing platform</th>
<th>Sequencing biochemistry</th>
<th>Performance characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roche 454 TiTi</td>
<td>Emulsion PCR and polymerase pyrosequencing</td>
<td>400° 0.8–1.2 0.5–1.5 Insertion &amp; deletion &gt;0.5% (ref. 7)</td>
</tr>
<tr>
<td>Illumina Genome Analyzer II</td>
<td>Cluster PCR and reversible terminators polymerase sequencing</td>
<td>75–125 1.5–2.5 0.2–2 Substitution &lt;0.1% (refs. 8 and 9)</td>
</tr>
<tr>
<td>AB SOLiD 3.0</td>
<td>Emulsion PCR and sequencing by ligation</td>
<td>50 0.8–3.0 &lt;0.1 (2 bp coding) Substitution NA</td>
</tr>
<tr>
<td>Helicos SMS Heliscope</td>
<td>Amplification-free and single-molecule sequencing</td>
<td>32 2.5 Stanley &amp; deletion 0.2% (ref. 1)</td>
</tr>
</tbody>
</table>

NA, not available. Sequencing of a single human genome with SOLiD technology has not been published.

*Based on the experience of Y.L. and J.W. and quality controlled to exclude failed runs. **Approximately 200 bp for paired reads. ‡Judging from the color-space mismatches and mappable proportion of data (40–80% typically), the raw color-space error rate would be >2%.