1. Tang, C. *et al. Nat. Biotechnol.* **29**, 829–834 (2011).
2. Boyse, E.A. & Old, L.J. *Annu. Rev. Genet.* **3**, 269–290 (1969).
3. Artzt, K. *et al. Proc. Natl. Acad. Sci. USA* **70**, 2988–2992 (1973).
4. Wright, A. & Andrews, P.W. *Stem Cell Res. (Amst.)* **3**, 3–11 (2009).
5. Solter, D. & Knowles, B.B. *Proc. Natl. Acad. Sci. USA* **75**, 5565–5569 (1978).
6. Adewumi, O. *et al. Nat. Biotechnol.* **25**, 803–816 (2007).
7. Enver, T., Pera, M., Peterson, P. & Andrews, P.W. *Cell Stem Cell* **4**, 387–397 (2009).
8. Fenderson, B.A., Eddy, E.M. & Hakomori, S. *Bioessays* **12**, 173–179 (1990).
9. Natunen, S. *et al. Glycobiology* **21**, 1125–1130 (2011).
10. Choo, A.B. *et al. Stem Cells* **26**, 1454–1463 (2008).

Semiconductors charge into sequencing

Keith Robison

The convergence of semiconductor chips and DNA sequencing begets a strong contender in the race to build the best sequencers.

After much secrecy and mystery, a new approach to sequencing DNA known as 'ion sequencing' was unveiled in early 2010 and launched commercially later that year. But few details of the detection methodology were made public—until now. Writing in *Nature*, Rothberg *et al.*[1] reveal the inner workings of their non-optical, electronic sequencing instrument and sketch possible routes to future improvements in performance through increasing the number of sequencing reads produced per run. Sequencing of bacterial and human genomes on this instrument shows the promise of the technology to provide fast and inexpensive sequencing, but issues of quality and fidelity remain to be solved.

All commercially available sequencers are based on optical detection of DNA extension from a known priming site, either by a polymerase (sequencing-by-synthesis) or by a ligase (sequencing-by-ligation)[2]. Most sequencers use fluorescently labeled nucleotides, but the sequencing-by-synthesis method known as pyrosequencing relies instead on an enzymatic cascade that converts pyrophosphate released during nucleotide incorporation into flashes of light[3]. Sequencing processes on various machines also operate at different time scales, resulting in faster or slower overall sequencing runs and imaging challenges. These can be divided into real-time observation of individual polymerases incorporating single nucleotides[4], imaging of light pulses coupled to nucleotide incorporation, and scanned imaging of stable patterns resulting from incorporation of reversible terminator nucleotides. Runs from instruments that use real-time imaging tend to be completed in a matter of hours, whereas the scanning-based instruments have runs of many days or even approaching two weeks. However, platforms that use scanning can read higher numbers of sequencing runs per read, resulting in greater overall production. Another difference between the systems is that most use spatially localized clonal DNA populations, whereas a few use single molecules[2]. But all sequencers require high-precision optics to resolve a densely packed array of sequence-producing elements.

In contrast, ion sequencing relies on electronic detection. In some ways it resembles high-throughput pyrosequencing, which Rothberg and five co-authors helped launch in 2005 (ref. 3). In both methods, beads bearing clonal populations of DNA are arrayed in wells and incubated serially with pure nucleotides. Incorporation of a nucleotide is detected by measuring ions ejected by polymerization—pyrophosphate in pyrosequencing and hydrogen ions in ion sequencing. But whereas pyrophosphate is detected optically in pyrosequencing, in ion sequencing each well lies above an ion-sensitive metallic oxide layer coupled to an electronic sensor that registers miniscule (0.02 pH unit) and transient (with a half-life <1 s) pH changes (**Fig. 1a**). Previous publications demonstrated the ability to detect such changes[5,6], but the new paper[1] parallelizes this to millions of sensors on a semiconductor chip housed in a compact instrument that delivers reagents, controls the process and collects data. Because the wells and sensors are produced as a single unit, they do not require alignment, in contrast to optics-based sequencing methods. An appealing aspect of the approach is that the entire sequencing chip is fabricated using well-established technology from the semiconductor industry. However, the scale of the electronics is far from cutting edge, with feature densities multiple orders of magnitude below those found in consumer electronics devices.

To convert electrical signals to DNA base calls, Rothberg *et al.*[1] use a computational model of the chemical processes that occur on the chip, such as diffusion of the released hydrogen ions and their neutralization by the buffer. The sequencing process also involves downstream filters that identify and eliminate polyclonal beads (as they produce signal too frequently) and beads whose signal poorly matches the physical model. The model corrects for dephasing, the tendency of clonal DNA populations to lose synchronization when nucleotides are not uniformly incorporated in all molecules in the population, and it is used to estimate quality scores that identify bases at the ends of reads that should be removed because they cannot be confidently called. These quality scores are reported as accuracies rather than as error rates. Although the two metrics are interchangeable (error rate is 1 – accuracy), individual error rates are more easily combined to estimate consensus error rates, which facilitate comparison across technologies and are typically used to assess the confidence of DNA assemblies or resequencing. Per base error rate in the first 50 bases is measured to be 0.4%, rising to 1.1% in the first 100 bases of a read.

A key theoretical advantage of ion sequencing and high-throughput pyrosequencing over many other approaches is the use of unmodified nucleotides, which reduces cost and eliminates the difficulties of engineering DNA polymerases to accommodate bases modified with fluorescent labels or chain-termination moieties[2]. Ion sequencing also does not require the downstream enzymatic cascade used in pyrosequencing, further lowering costs. However, unterminated polymerization can make it difficult to count incorporation events in homopolymeric DNA. Although the signal should be proportional to the number of emitted hydrogen ions, various sources of noise complicate this relationship. Rothberg *et al.*[1] report 3% error when measuring homopolymers 5 bases long, but do not give information for longer lengths. In contrast, Illumina sequencing does not encounter errors in homopolymers until these regions exceed 7 bases[7]. Very large genomes contain many such stretches, and so at present ion sequencing may find limited acceptance in applications where homopolymer measurement is critical, such as *de novo* genome sequencing. In most cases, however, Rothberg *et al.*[1] misestimate homopolymers by only a single base.

Keith Robison is at Infinity Pharmaceuticals, Cambridge, Massachusetts, USA.
e-mail: keith.robison@infi.com.

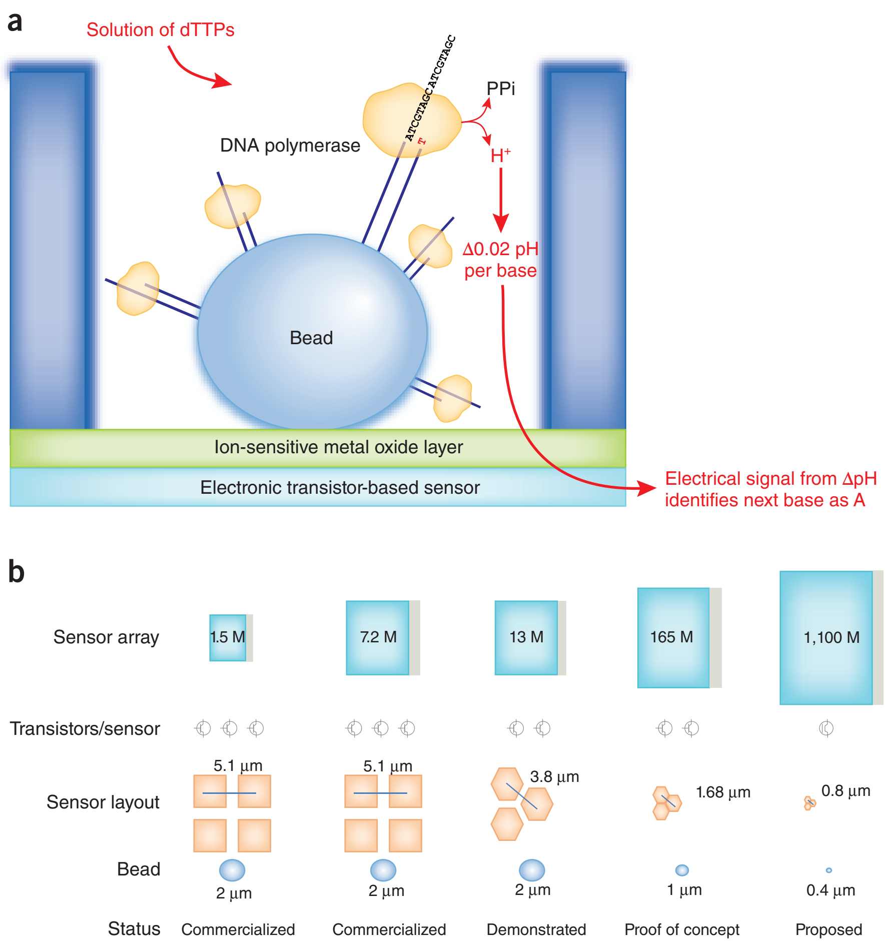

Figure 1 Ion sequencing. (**a**) Schematic of a well on an ion sequencing chip. Clonal DNA immobilized on a bead is extended by polymerase in the presence of a pure solution of one nucleotide (here 'T'). Nucleotide incorporation releases a pyrophosphate (PPi) and a hydrogen ion. The change in pH caused by release of the hydrogen ion alters the surface potential of the ion-sensitive metal oxide layer. This is converted to a voltage signal by transistors. The wells are washed and exposed sequentially to pure solutions of other nucleotides. For comparison, in high-throughput pyrosequencing, the pyrophosphate is converted to chemiluminescence by an enzymatic cascade and optically imaged. The size of the well relative to the bead has been exaggerated, although each well contains a single bead. (**b**) Evolution of ion sequencing chips. Increases in sensors per chip can be achieved by increasing the physical area of the sensor array, reducing the number of transistors per chip, arranging the sensors in a hexagonal rather than rectilinear geometry and reducing the well and bead size. Sensors are drawn to scale, and gray indicates sensor area not accessible to fluid. The 13-million (M) sensor design was used by Rothberg *et al.*[1] to sequence DNA from both *Escherichia coli* and human. Data for a fixed ('key') DNA sequence was shown for the 165-million sensor design. The 1,100-million sensor design was proposed but its feasibility was not shown.

Rothberg *et al.*[1] demonstrate the utility of their system by sequencing three bacterial genomes. Each was chosen to represent a range of guanine cytosine (GC)-dinucleotide content, as the accurate and consistent sequencing of DNA with high or low GC content has often proven difficult. For each genome, the expected frequency of reads is closely matched by the observed frequency, deviating only slightly in regions at the ends of the %GC distribution for that organism. The authors also present data from a low-coverage (10.6×) genome sequence of Gordon Moore, a pioneer in the semiconductor industry. Cross-validation of the Moore genome with more established sequencing-by-ligation technology at 15× coverage shows a high (>99.9%) concordance in single-nucleotide polymorphism calls, albeit with a high frequency of variants that were not called because of the low sequencing depth obtained from each instrument. Notably, the ion sequencing process seems to be less efficient at sequencing samples with low GC content, as human samples with very low GC content yielded about 20% less data than what would be expected.

Improvements to previous high-throughput sequencing technologies have been achieved in several ways, in particular, by increasing the density of DNA molecules in the flow cell and the amount of high-quality sequence obtained from each molecule[2]. Rothberg *et al.*[1] outline how the density of ion sensors could be increased (**Fig. 1b**). Their early-generation chip contained 1.5 million sensors (about 1.1 million of which are useable), each with three transistors. Increasing the physical size of the chip and the fraction of the chip accessible to fluid yielded a chip with five times as many sensors. The next jump required reducing the number of transistors per sensor to two, enabling a more compact hexagonal (rather than rectilinear) grid of wells that increased the sensor density by another 1.8-fold to 13 million sensors.

These advances show that although progress in ion sequencing technology may follow Moore's Law (a doubling of semiconductor performance to price every two years)[8], the strategies for improving ion sequencing and semiconductor technology are quite different. Gains in semiconductor performance have been driven largely by ever-smaller feature size; in contrast, the increase from 1.5 million to 13 million sensors came almost as much from enlarging the chip surface as from increasing sensor density, and sensor density was increased without altering the feature size.

Future increases in sensor density may require some radical changes. A newer generation of semiconductor technology allowed Rothberg *et al.*[1] to pack transistors at a threefold higher density, but required a halving of the bead size to support this and a corresponding fourfold reduction in the number of DNA molecules per bead. A consensus signal plot in the supplementary materials shows the ability to read the fixed 'key sequence' at the beginning of each read from a 1-µm bead, albeit with much narrower signal peaks. But whether this holds true for the rest of the read is unclear as such dense packing may trigger unusable levels of cross-talk between wells when every well is not reporting the same fixed sequence. In addition, data for this experiment are not provided to explore signal decay across a run, which could lead to lower qualities and shorter reads. Still, such chips could prove useful for applications in which a large numbers of short tags are sufficient, including expression profiling or chromatin immunoprecipitation. For even greater densities, the authors propose a reduction to single-transistor sensors, but this would necessitate more than halving the bead again to fit 0.5-µm wells. Such a chip, sporting approximately one billion sensors, might be able to sequence a human genome in a single run—in sharp contrast with the hoard of chips applied to Moore's genome: 1,601 of the 1.5-million sensor chips, 267 of the 7.2-million chips and 28 of the 13-million chips.

Rothberg *et al.*[1] do not provide a roadmap for improving read lengths and the efficiency and consistency of loading the sensor plate. In the current chip, only 20–40% of accessible sensors yield useable data. Only 55–80% of reads are >100 bp, although >90% reach 50 bp. Perhaps more concerning than raw numbers is high variability, which makes it difficult to plan experiments requiring set numbers of reads or reads of specific lengths. The authors present a single 212-bp perfect read to suggest future potential in this area, but this read lies far from the bulk distribution of read lengths.

Will ion sequencing become a major player? Scientists tend to pick sequencing instruments based on ease of use, cost, speed, read length and accuracy. The ion instrument excels in only some of these dimensions. In terms of cost, the list price of a complete instrument setup is about $70,000, substantially less than the $200,000–$700,000 for many other instruments. Cost per run is <$1,000, but yield in terms of bases is much lower than with other systems. For example, the top performance reported by Rothberg *et al.*[1] of ~10 million reads is greatly dwarfed by the more than 1 billion reads obtained from an Illumina HiSeq flow cell, although a HiSeq run costs about 10- to 20-fold more and takes over a week compared with a few hours for ion sequencing (library and template preparation for each system add additional time). Furthermore, the HiSeq routinely delivers 100 bases from each end of a DNA fragment rather than the single read of often <100 bases from ion sequencing. In short, ion sequencing is currently adept at delivering smaller amounts of sequence data very quickly at a low cost per experiment, whereas HiSeq can deliver enormous data volumes more slowly at a low cost per base or fragment. Hence, deciding between these two systems depends on an experimenter's needs, with ion sequencing perhaps at an advantage in fields that emphasize speed, such as biosurveillance[9].

Similarly, there are tradeoffs compared with other sequencers on the market. Roche's 454 pyrosequencing instrument has longer running times, similar numbers of reads as the second-generation ion sequencing chip, and a much higher cost (per base, per read and for the instrument), but offers longer reads of 400–800 bp. Pacific Bioscience's single-molecule sequencing instrument promises reads of a kilobase or more in short times but costs >10 times more than the ion sequencing instrument and delivers many fewer reads per sample per unit time.

DNA sequencing will continue to garner large public and private-sector investments. Optical sequencing-by-synthesis methods should continue to improve through the introduction of new chemistries and instruments. Many companies are also vying to produce fast and cheap benchtop machines for the scientific masses. Finally, all of these technologies may be displaced by completely new approaches, such as those using nanopores for single-molecule detection[10]. Whether Rothberg *et al.*[1] will be able to produce an ion sequencing instrument that can sequence a human genome at a high depth of coverage remains to be seen. Ion sequencing is a promising new entrant already delivering useful science[9], but one in a crowded field of equally impressive competitors.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Rothberg, J.M. *et al. Nature* **475**, 348–352 (2011).
2. Fuller, C.W. *et al. Nat. Biotechnol.* **27**, 1013–1023 (2009).
3. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
4. Eid, J. *et al. Science* **323**, 133–138 (2009).
5. Sakata, T. & Miyahara, Y. *Angew. Chem. Int. Edn. Engl.* **45**, 2225–2228 (2006).
6. Pourmand, N. *et al. Proc. Natl. Acad. Sci. USA* **103**, 6466–6470 (2006).
7. Albers, C.A. *et al. Genome Res.* **21**, 961–973 (2011).
8. Moore, G.E. *Electronics* **38**, 114–117 (1965).
9. Rohde, H. *et al. N. Engl. J. Med.* published online, doi: 10.1056/NEJMoa1107643 (27 July 2011).
10. Akeson, M., Branton, D., Kasianowicz, J.J., Brandin, E. & Deamer, D.W. *Biophys. J.* **77**, 3227–3233 (1999).

Genome remodeling

Yizhi Cai & Jef D Boeke

Alteration of a stop codon across an entire bacterial genome opens the way to exploring the biotech potential of synthetic genetic codes.

In much the same way as chemistry graduated from simply analyzing the structure of matter to synthesizing chemical compounds, biology is transitioning from an era centered on deciphering genome-sequence information to one focused on building and studying synthetic genomes[1]. Several technologies are being worked out for designing and fabricating genomes from scratch[2,3], but the technical challenges of *de novo* genome synthesis have spurred the development of alternative approaches. In a recent issue of *Science*, Isaacs *et al.*[4] present a method for editing DNA on a genome-wide scale *in vivo*. The authors first used a previously described technique called multiplex automated genome engineering (MAGE)[5] to modify a rare stop codon in many discrete sections of the *Escherichia coli* genome and then applied conjugative assembly genome engineering (CAGE) to hierarchically assemble the modified sections into larger pieces. MAGE/CAGE can be considered complementary to *de novo* genome synthesis, especially for cases in which the desired variant genome closely resembles the wild-type sequence.

The divide-and-conquer strategy used by Isaacs *et al.*[4] involved dividing the *E. coli* genome into 32 roughly equal-size fragments, with 31 fragments each containing 10 of 314 TAG stop codons and the last fragment having the remaining 4 TAG codons (**Fig. 1**). MAGE, which uses synthetic DNA oligonucleotides to direct modification by homologous recombination, was applied to each fragment to find and replace TAG codons, the rarest stop codon in the *E. coli* genome, with synonymous TAA stop codons. That this modification does not compromise viability of the organism is not unexpected as the specialized release factor that recognizes TAG, prfA, is known to be dispensable under specific genetic conditions[6]. Next, to knit the 32 fragments together, the authors exploited bacterial conjugation, a mechanism whereby bacterial cells exchange genetic material. Adjacent fragments were paired and merged, and the process was repeated until four strains lacking TAG codons in one quarter of the genome were generated (**Fig. 1**). Two CAGE steps remain to finish the job.

An obvious advantage of this method compared with *de novo* genome synthesis is that it allows one to closely monitor any aberrant phenotypes that may result from stop-codon swaps in each fragment. If a genome is viewed as a book, the chemical synthesis method developed at the Venter Institute[2] writes the book letter by letter and does not undertake proofreading until the entire book has been completed. In contrast, the MAGE/CAGE method of Isaacs *et al.*[4] starts with an existing book, divides it into paragraphs, replaces particular words in each paragraph with synonyms or even other types of other words, and proofreads each of the revised paragraphs before reassembling them. Thus, errors are read-

Yizhi Cai and Jef D. Boeke are in the Department of Molecular Biology and Genetics and the High Throughput Biology Center, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA.
e-mail: jboeke@jhmi.edu