bad clusters? It should be noted that in spectral library building, it is relatively easy to answer these questions, given that spectra are identified to peptides. Without that identification, however, it will be necessary to design other robust and automatic methods to verify clusters and remove errors in spectral archives, a process that will be especially challenging given the enormous scale. Frank *et al.* address some of these issues, including providing a rudimentary web interface for submitting data and querying the spectral archive, but further validation and improvement is probably needed. Nonetheless, these problems do not seem entirely intractable and should be fruitful areas of research in the near future.

At the same time, however, there are also hurdles that are not technical in nature. Proteomics researchers have so far been less forthcoming about sharing data than their counterparts in genomics. A change in mentality will likely be needed for such a radical approach to be embraced by the community. At the moment, inconvenience is a good excuse: sharing proteomics data is a troublesome task, and until recently the incentive to do so has not been very strong. Also somewhat understandable is the urge to protect hard-earned data from being exploited by others to make their own discoveries. Although a reasonable argument can be made that data directly contributing to publications should be made public, the case is much weaker for unpublished data or data not related to the premise of published papers. Yet in a context-blind approach such as spectral archives, even unpublished data can be of great value. Lastly, one might still question whether anyone in the community should be entrusted with such a demanding and critical role, although it is not hard to imagine governmental agencies taking on the responsibility, as has happened in the world of genomics[12,13].

If one takes an optimistic view, however, the idea of spectral archives may just be what is needed to encourage data sharing. Submitting proteomics data to repositories has so far been an unwelcome chore and an afterthought; there is not much to be gained by doing so, other than the good feeling of contributing to the scientific community. But with spectral archives, unknown spectra submitted will actually be analyzed free of charge. Peptide identifications, if any, and the corresponding high-quality consensus spectra will come back to the data submitter as a reward for sharing the data. The data submitter remains free to make new discoveries and can choose to analyze the returned consensus spectra as he or she sees fit, but with a higher chance of success due to better spectral quality. Additionally, with spectral archives, the data submitter will not be required to divulge the sample source or the experimental conditions (save perhaps the organism), maintaining some measure of confidentiality. By not insisting on full disclosure of metadata while providing a useful service, spectral archives will be better positioned to attract users.

Frank *et al.*'s clustering algorithm holds great promise as a game-changer in the field of computational proteomics. Today's static, dataset-centric data repositories are a necessary first step, but they are far from being truly useful. Turning them into spectral archives—dynamic, spectrum-centric, interactive and hence immensely useful—is a tantalizing possibility.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Aebersold, R. & Mann, M. *Nature* **422**, 198–207 (2003).
2. Steen, H. & Mann, M. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711 (2004).
3. MacCoss, M.J. *Curr. Opin. Chem. Biol.* **9**, 88–94 (2005).
4. Martens, L. *et al. Proteomics* **5**, 3537–3545 (2005).
5. Deutsch, E.W., Lam, H. & Aebersold, R. *EMBO Rep.* **9**, 429–434 (2008).
6. Hill, J.A., Smith, B.E., Papoulias, P.G. & Andrews, P.C. *J. Proteome Res.* **9**, 2809–2811 (2010).
7. Craig, R., Cortens, J.C., Fenyo, D. & Beavis, R.C. *J. Proteome Res.* **5**, 1843–1849 (2006).
8. Frewen, B.E., Merrihew, G.E., Wu, C.C., Noble, W.S. & MacCoss, M.J. *Anal. Chem.* **78**, 5678–5684 (2006).
9. Lam, H. *et al. Proteomics* **7**, 655–667 (2007).
10. Lam, H. *et al. Nat. Methods* **5**, 873–875 (2008).
11. Frank, A.M. *et al. Nat. Methods* **8**, 587–591 (2011).
12. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Sayers, E.W. *Nucleic Acids Res.* **39**, D32–D37 (2011).
13. Kulikova, T. *et al. Nucleic Acids Res.* **35**, D1–D5 (2006).

Fluorogenic pyrosequencing in microreactors

Jason A Steen & Matthew A Cooper

A technique that combines the speed of pyrosequencing with the sensitivity of fluorescent detection may lead to faster sequencing with smaller quantities of DNA.

High-throughput sequencing of nucleic acids has evolved over the last decade as a powerful new strategy for investigations into disease and disease-causing organisms, as evidenced by the number of commercial platforms and publications in this area. In this issue of *Nature Methods*, Sims *et al.*[1] describe a new variant of a contemporary sequencing method, pyrosequencing, in which the incorporation of fluorescently modified nucleotides followed by detection of nucleotide incorporation in an array format are used to increase throughput and sensitivity.

Current sequencing technologies can be broadly divided into two basic categories: sequencing by synthesis (Illumina, Roche, Ion Torrent, Helicos and Pacific Biosciences) and sequencing by ligation (Life Technology, Polonator and Complete Genomics). Sequencing by synthesis can be further divided into two distinct approaches: pyrosequencing, in which individual native nucleotides are added sequentially; and the reversible terminator approach, in which each nucleotide is labeled with a different fluorophore and a single base elongation with any base can be assayed concurrently. Pyrosequencing follows DNA polymerase progression along a DNA strand by allowing only a single dNTP to be available for incorporation at any time, and then takes advantage of the chemical reaction that occurs when the dNTP is incorporated by the polymerase (**Fig. 1a**). This reaction is detected either by inducing a bioluminescence cascade and detecting the emitted light (Roche 454)[2] (**Fig. 1b**) or by directly detecting protons released during incorporation as a change in pH (Ion Torrent)[3] (**Fig. 1c**). Pyrosequencing allows sequencing to proceed at a much faster rate than reversible terminator sequencing

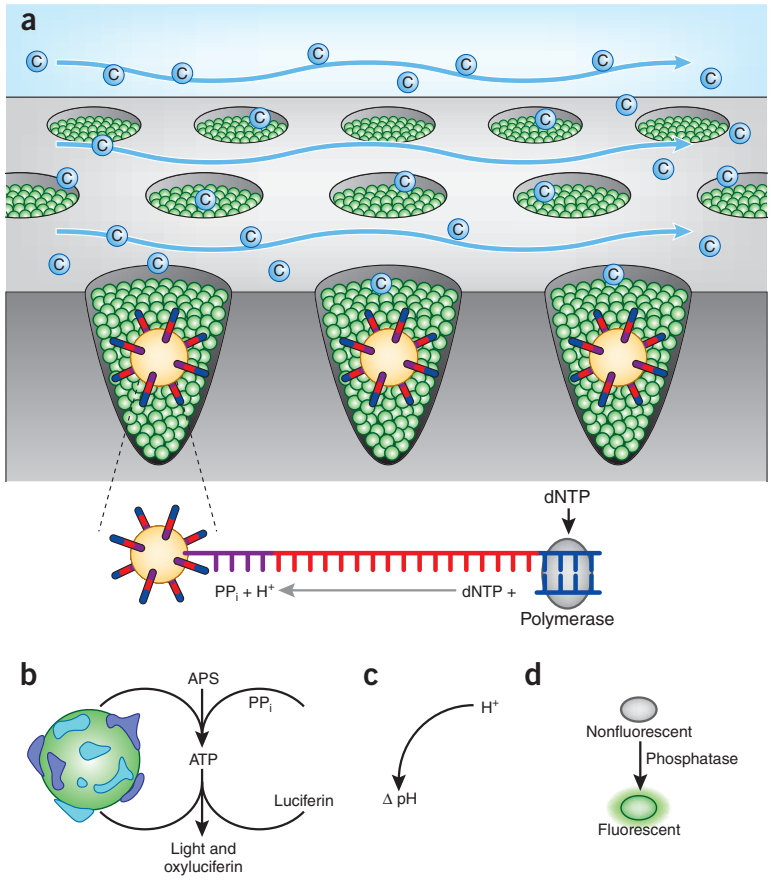
Jason A. Steen and Matthew A. Cooper are at the Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia.
e-mail: m.cooper@uq.edu.au

because fewer steps are required to detect a base and then to continue the extension of a template. As such, it is common to achieve 100-base-pair reads in under 3 h with pyrosequencing, compared to a week or more using other approaches.

The method described by Sims *et al.* provides a bridge between reversible terminator–based approaches and pyrosequencing. Here the dNTP is modified with a fluorescent reporter such that when a nucleotide is incorporated, an inactive polyphosphate-bound fluorophore is cleaved from the incorporated base. Additional phosphatases then cleave the polyphosphate chain, thereby activating the fluorophore and allowing detection of the event (**Fig. 1d**). The fluorophore needs to be spatially constrained near the molecule from which it was cleaved to tie the fluorescent signal back to a DNA base. This is achieved using microreactors composed of the versatile material polydimethylsiloxane (PDMS). PDMS microreactors are relatively easy to fabricate, can be sealed to surfaces in a reversible manner and possess low intrinsic autofluorescence. Other surface-based or solution-based microcompartment approaches could also be used with the method.

Like other pyrosequencing approaches, the technique relies on the initial immobilization and clonal amplification of the DNA molecules to be sequenced onto a bead, and then the physical separation of individual beads into microreactors. Clonal amplification generally uses emulsion PCR, in which the immobilizing bead, DNA fragment and DNA polymerase are emulsified in an oil-buffer mixture such that each droplet contains only one bead, one template DNA molecule and sufficient reagents for amplification. The amplification proceeds by conventional polymerase chain reaction, and final purification of the oil-water emulsion yields large numbers of clonally amplified molecules attached to immobilizing beads. This process can be slow and arduous, and one of the potential advantages of the approach presented in this study is that the lower concentration of DNA required may allow faster, simpler sample preparation for sequencing.

Although the ability to measure incorporation of a single dNTP in a reaction cycle is the backbone of pyrosequencing, it is also one of its major limitations, as dNTP incorporation will not necessarily stop at one nucleotide but rather continues for subsequent nucleotides in cases where there is a stretch of the same base. Multiple, discrete incorporations are thus detected as a single analog signal of varying magnitude. Unfortunately, the magnitude of the signal being detected does not always scale directly with the number of base incorporations. Hence, it is more difficult to determine the exact number of nucleotides in repeats longer than 7 or 8 bases. Although this new method is unlikely to achieve great advances in this area over either Roche's 454 or Life Technology's Ion Torrent, Sims *et al.*[1] show good separation of signal intensities for polynucleotide tracts up to 5 bases long, which could, with further optimization and development, lead to better performance in this area. Finally, in comparison to other fluorescence-based sequencing methods, the use of a single rather than multiple fluorophores could translate into lower costs for consumables.



Katie Vicari

Figure 1 | Overview of pyrosequencing. (**a**) Clonally amplified DNA beads are deposited into a microreactor, and dNTPs are sequentially washed over the wells. (**b**–**d**) When a nucleotide is incorporated, an enzymatic cascade produces light (**b**), H^+ ions are detected as a change in pH (**c**) or phosphatases cleave the phosphate from the fluorophore and the fluorescence is measured (**d**). APS, adenosine 5′ phosphosulfate.

The sequencing market continues to expand into science and medicine, with throughput/cost ratios outpacing the analogous Moore's Law metric for semiconductors. Pyrosequencing was first conceived by Pål Nyrén in 1987, culminating a seminal paper[4] and commercialization a decade later. With sequencing expanding into new markets and competition driving down costs per base, innovations such as this one may find a place in the sequencing race.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Sims, P.A. *et al. Nat. Methods* **8**, 575–580 (2011).
2. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
3. Rothberg, J.M. International Patent no. WO-2010/008480-A2 (2010).
4. Ronaghi, M., Uhlén, M. & Nyrén, P. *Science* **281**, 363–365 (1998).