first exploits mass spectrum signal intensities, the accuracy of which has greatly improved owing to recent advances in chromatography and ionization (for example, nanoflow electrospray ionization) and in mass spectrometers themselves (for example, the Thermo Electron Corporation LTQ/Orbitrap, which has an innovative mass analyzer[6]). As a consequence, Silva *et al.*[7] found that a protein's abundance could be well estimated from the average mass spectrum peak intensity of its three best-detected peptides. A second approach, spectral counting, analyzes the observed counts of MS/MS spectra attributable to each protein. In a recent development for large-scale absolute protein expression measurements (APEX), Lu *et al.*[8] improved the accuracy of spectral counting by incorporating differential peptide ionization propensities into the computation.

Malmström *et al.*[1] combine these three approaches—SRM measurements of a limited set of internal reference standards, the average mass spectrum signal intensities of the top three peptides selected per protein, and weighted MS/MS spectral counts—to more completely quantify the proteome (**Fig. 1**). By using the SRM measurements of reference standards to calibrate the two computational abundance calculations, they achieve abundances accurate to ~2-fold on average for 769 proteins using the approach of Silva *et al.*[7] and to ~3-fold for 1,095 more proteins with the technique of Lu *et al.*[8]. This enables them to measure abundances for >1,800 proteins, or 83% of the proteome detectable by mass spectrometry under these conditions and 51% of the *L. interrogans* proteome (based on predicted open reading frames). Combining the high accuracy of SRM with the high coverage of the two computational approaches minimizes the costs of isotopic labeling while maximizing coverage and accuracy (**Fig. 1**). The abundance estimates are validated with molecule concentrations measured by single-cell cryo-electron tomography for flagellar proteins, flagellar motors and periplasmic methyl-accepting chemotaxis protein receptors.

As with any mass spectrometry method, the techniques used by Malmström *et al.*[1] are limited by the peptides' amenability to ionization and by the mass spectrometer's ability to detect low abundance molecules. Although >200 of the ~1,000 proteins monitored after exposure of *L. interrogans* to the antibiotic ciprofloxacin changed their abundance more than twofold, the limitations of sensitivity for differentially expressed proteins may be even lower[8], depending on whether the observed quantification errors are consistent across samples and systematic in nature, which is unknown at present.

Although there is no theoretical upper limit to the size of the proteome for which this approach should be effective, current mass spectrometers and practices restrict it to a few thousand proteins; this covers the majority of proteins for simple organisms but typically represents only a fraction of the expressed proteome for higher organisms. Fractionation of samples before analysis can substantially increase the proteome coverage, but further work remains to determine how fractionation affects these quantification methods. For example, the SRM calibrants might have to be chosen appropriately to sample the different fractions. Perhaps more importantly, resolving the differential expression of splice variants, which are common in proteomes of higher organisms, is still a challenging problem in shotgun proteomics. Nonetheless, given that these approaches offer protein quantification without the need for genetic modification or extensive isotopic labeling, the combination of approaches presented by Malmström *et al.* should be widely applicable to many systems.

The availability of absolute protein concentration data will be indispensable to fulfilling the promise of systems biology. Owing to extensive post-transcriptional regulation, protein abundances are only partially correlated with the abundances of the corresponding mRNAs[8–10]. This has led many to argue that direct assessment of protein levels is often more informative of the cellular state than analysis of mRNA levels. Indeed, protein abundances seem more conserved across evolution than mRNA transcript abundances[10]. Quantitative mass spectrometry is now poised to routinely provide such data at large scale and with high accuracy—a testament to the rapid progress in quantitative shotgun proteomics over the last few years.

1. Malmström, J. *et al. Nature* **460**, 762–765 (2009).
2. Han, X., Aslanian, A. & Yates, J.R. III. *Curr. Opin. Chem. Biol.* **12**, 483–490 (2008).
3. Lange, V. *et al. Mol. Syst. Biol.* **4**, 222 (2008).
4. Addona, T.A. *et al. Nat. Biotechnol.* **27**, 633–641 (2009).
5. Kito, K. & Ito, T. *Curr. Genomics* **9**, 263–274 (2008).
6. Hu, Q. *et al. J. Mass Spectrom.* **40**, 430–443 (2005).
7. Silva, J.C. *et al. Mol. Cell. Proteomics* **5**, 144–156 (2006).
8. Lu, P. *et al. Nat. Biotechnol.* **25**, 117–124 (2007).
9. Anderson, L. & Seilhamer, J. *Electrophoresis* **18**, 533–537 (1997).
10. Schrimpf, S.P. *et al. PLoS Biol.* **7**, e48 (2009).

Combinatorics and next-generation sequencing

Nick Patterson & Stacey Gabriel

The massive capacity of today's sequencing machines can be harnessed efficiently by sequencing pooled samples and decoding the results.

In the last year alone, the average yields of a single DNA sequencing instrument have increased by at least tenfold, and ten billion bases can now be obtained routinely in a single run. Indeed, for many applications, current sequencing throughput is vastly greater than what is needed to process a single sample—a situation that brings not only new opportunities but also new challenges. Two recent papers in *Genome Research*, by Erlich *et al.*[1] and Prabhu and Pe'er[2], present improved methods for exploiting this technological capability. Using ideas from a branch of mathematics called combinatorics, they show that thousands of pooled samples can be sequenced *en masse* and the results decoded.

The new sequencing technologies will have many applications[3], but here we concentrate on methods for discovery of rare mutations, which are likely to account for much of the genetic basis of disease. The high yields of the latest instruments allow us to deeply sequence genes of medical interest for thousands of individuals[3]. As only tens to hundreds of kilobases are of interest in such studies, and as even the smallest functional unit of a sequencer—a single 'lane'—generates data amounting to many thousand–fold coverage of such targets, the challenge is how to use a sequencer efficiently on samples requiring only a fraction of its capacity. An equally daunting challenge is the need to individually amplify and create sequencing templates for thousands of samples. The cost of the amplification and the difficulties of sample tracking and automation are substantial.

Pooling DNA samples promises to solve both of these challenges. Grouping many samples together in each run makes the most effective use of the high depth of sequencing coverage and alleviates the problem of handling many individual samples. Simply mixing all of the samples together, however, makes it impossible to determine which individual contributed

Nick Patterson and Stacey Gabriel are at the Broad Institute, Cambridge, Massachusetts, USA.
e-mail: nickp@broadinstitute.org

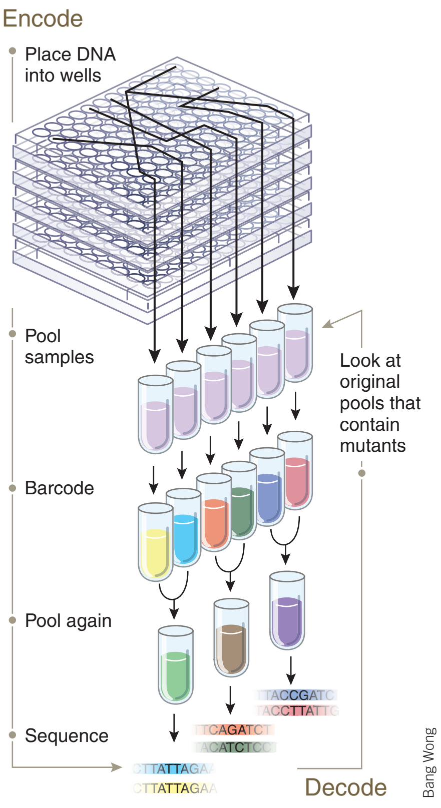

Figure 1 Pooling samples using a combinatorial design. Also shown is optional barcoding and further pooling. The details of the decoding strategy will depend on the design used to combine samples, but a general method is to assume the pattern of the mutants (that is, which samples contain the mutant allele) and then work through the pooling steps to calculate the likelihood of observations given the assumption. This 'brute force' method is computationally intensive. More efficient decoding methods are discussed by Erlich *et al.*[1] and Prabhu and Pe'er[2].

any rare DNA variants that are detected. Furthermore, a pool may 'fail'—that is, no useful sequence may be recovered owing to experimental difficulties—and we would like a strategy that is robust to small numbers of such failures.

A conceptually simple workaround is to cut the DNA in each sample into short fragments suitable for sequencing and ligate the fragments to a short, sample-specific DNA sequence, or 'barcode'[4]. This approach has been used to sequence pooled samples, although, at present, barcoding every sample is costly.

An alternative to barcoding is to design a pooling strategy that allows the sequence data to be decoded—for instance, to identify rare alleles and the samples that contain them. In practice, this may be achieved by arraying a large number of samples (denoted by N) into distinct wells in microtiter plates. Then, by machine or by hand, the N source wells are pooled into T destination wells in another set of microtiter plates. An effective pooling strategy allows T to be much smaller than N. With a suitable design and a small number of mutant alleles, there are effective decoding strategies to reconstruct which samples contain the mutants. If the pool is still too small for next-generation sequencing to be economical, then barcoding and further pooling can be applied (**Fig. 1**).

What should be the rule for assigning source wells to destination wells? It turns out that this kind of problem has been extensively studied in combinatorics. Simple designs in which each sample is assigned to a unique pool offer no hope of reconstructing which pools contain the mutant and are also vulnerable to failure of a single pool. There are advantages to mapping one source well to several destination wells—a so-called 'overlapping' design. With such a design we can achieve 'robustness' (resistance to a small number of pool failures), and we may also be able to work out which samples contain the mutant allele.

In designing a pooling strategy, we would like to keep small the number of pools (which reduces costs), the maximum number of samples in any pool, called the 'compression level' (which reduces experimental problems), and the maximum number of pools that two samples can share, called the 'intersection number'. A large intersection number makes the design redundant: for instance, if two samples are each assigned to the same set of pools, we have no chance of reconstructing which contains a mutant. Thus, we would like a low intersection number to avoid duplication of information, and we would like robustness, so that the chance of success is good if coverage is not very high or some pools fail.

Erlich *et al.*[1] and Prabhu and Pe'er[2] discuss these considerations and use techniques from combinatorics to construct the design rule. Related techniques are being applied to protein arrays[5]. The two papers have subtly different aims. Erlich *et al.*[1] propose methods that allow rare alleles to be discovered and that determine with high probability which samples contain the allele. Prabhu and Pe'er[2] are concerned primarily with discovering rare alleles. However, their designs are able to detect and identify 'singletons'—a mutation occurring in just one sample—with very good reliability.

Erlich *et al.*[1] suggest a design based on elementary number theory. Using the Chinese remainder theorem[6], they show that the intersection number of their design is equal to one, meaning that no two samples will be jointly placed in more than one pool. This is essentially the best possible solution, as a design with an intersection number of zero would simply group samples into pools without overlap. Such a design cannot identify the sample carrying a mutation and is not robust if a pool fails. The number of pools in the Erlich *et al.*[1] design will be a small multiple of the square root of the number of starting samples ($\sqrt{N}$), and the compression will always be smaller than $\sqrt{N}$.

Prabhu and Pe'er[2] suggest several designs that use a smaller number of pools than Erlich *et al.*[1], but their methods are less able to identify which samples carry the rare allele, at least when there is more than one such sample. Their most interesting design uses code words from the Golay code[7], a remarkable combinatorial configuration that has many consequences in pure mathematics. Using these theoretical foundations, one can combine 759 source wells into 24 pools with a relatively high compression of 253 (that is, their decoding method must be able to handle data from 253 pooled samples). For a sample size of 759, which is at present a moderate number of samples, their method shows extremely good robustness for detecting a rare allele. Even if three pools fail (which is not uncommon in practice), a sample with a rare allele will still be sequenced in five other pools.

Erlich *et al.*[1] implemented their strategy in an experiment using 40,000 bacterial clones and 1,900 pools. Several practical difficulties emerged, notably low-level contamination, with some sequences appearing in several hundred pools. Nevertheless, after some modifications to their decoder, they were able to achieve good results. Prabhu and Pe'er[2] simulated an experiment by downloading sequences from the 1000 Genomes project (http://www.1000genomes.org) and simulating pooling *in silico*. This is not entirely satisfactory as unmodeled difficulties would likely have arisen in practice, but it is certainly a reasonable proof of principle.

It is exciting to see combinatorial designs being used in biology in this way. Their usefulness will ultimately depend on how the technologies develop. Pooling certainly introduces new technical issues (such as achieving equimolar pooling of samples) but drastically reduces some of the process costs. If barcoding becomes very simple and inexpensive, the way forward will be to barcode every sequence fragment. If, by contrast, sequencing costs fall rapidly compared with those of barcoding, combinatorial methods should prove increasingly valuable.

1. Erlich, Y. *et al. Genome Res.* **19**, 1243–1253 (2009).
2. Prabhu, S. & Pe'er, I. *Genome Res.* **19**, 1254–1261 (2009).
3. Mardis, E.R. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
4. Craig, D. *et al. Nat. Methods* **5**, 887–893 (2008).
5. Xin, X. *et al. Genome Res.* **19**, 1262–1269 (2009).
6. Andrews, G.E. *Number Theory* (Dover, Mineola, NY; 1994).
7. Conway, J.H. & Sloane, N.J.A. *Sphere Packings, Lattices and Groups* 3rd edn. (Springer, New York; 1998).