Published online: August 6, 2015

EMBO reports

Science & Society

Better tools, new problems

New technologies help to advance research in the life sciences, but the quantities of data generated are proving hard to manage and interpret

Philip Hunter

Technological advances have almost always led to accelerated progress in the life sciences. The invention of the light microscope paved the way for the discovery of cells as the building blocks of life, and bacteria as the causes of many infectious diseases. The electron microscope allowed biologists to peer into cells themselves and directly observe viruses. The unravelling of the structure of DNA or proteins was directly enabled by X-ray crystallography. DNA sequencing and PCR became the technological cornerstones of molecular biology, whereas NMR and mass spectrometry further improved and accelerated the analysis of proteins and their structures.

> *"As the sequencing of whole genomes becomes faster and cheaper, data production is running ahead of the ability to make sense of it..."*

Many of the inventions themselves have also seen rapid development and improvement. The combination of novel cryological techniques and image analysis algorithms has revived interest in the electron microscope as a tool to visualize the atomic structural details of proteins and cellular structures without the need to create crystals. DNA sequencing has made enormous advances since Fred Sanger's original method based on DNA polymerase followed by gel electrophoresis; today's next-generation sequencing (NGS) machines can perform millions of reactions in parallel. Light microscopy, one of the oldest tools of biologists, has broken the diffraction limit to let scientists observe the position and movement of individual protein molecules in living cells in real time in three dimensions. As much as all of these technologies help to understand how life works at the molecular level, they create a growing challenge for biologists to make sense of gigabytes or terabytes of data.

This is particularly true for next-generation sequencing, which has become a major challenge for bioinformatics. As the sequencing of whole genomes becomes faster and cheaper, data production is running ahead of the ability to make sense of it, according to Ron Zimmern, chairman of the PHG Foundation, a health policy think tank based in Cambridge, UK, that focuses on how genomics can deliver personalized and more effective health care. "Getting the raw sequence is the easy bit now," Zimmern said. "I'm in no doubt we will soon get to the $100 genome. But on its own it is totally useless to man or beast, and interpretation is by no means easy."

One of the problems is that although NGS yields more data at less cost and in a shorter time, it is actually a backwards step in terms of accuracy. The length of sequence obtained from each read operation is both shorter and less accurate than with older techniques. This sacrifice in accuracy has been considered worthwhile, given the huge increase in throughput, as the latest instruments are capable of yielding several hundreds of gigabases of DNA data in a single sequencing run. NGS thus requires a corresponding increase in computational power to reconstruct long sequences or even whole genomes from millions of code fragments. As a result, the data generated by NGS contains many errors, which can only be eliminated through multiple sequencing runs or comparison with other instances of a particular sequence to make sure that the detected variant is indeed real and not an artefact.

> *"The problems with data accuracy and data analysis are particularly relevant for the application of NGS in the clinic..."*

This is particularly important for using NGS to identify genetic variants that could indicate an elevated risk for certain diseases. One major challenge from NGS's accuracy problems is therefore making sure that sequence variants are not a mere sequencing artefact. "If it has been sequenced a number of times, then you can be more confident that a variant indicative of disease is really present," Zimmern said. After identifying and confirming a genetic variant, the next step is to assess how likely it is that the variant indicates disease risk. Zimmern noted that this is still a matter of human judgement for now, rather than quantitative analysis. Yet, more knowledge about genetic variants and their possible role in disease should eventually make diagnostics much easier, even without the need to resort to NGS. "So there is more than meets the eye to the whole issue of genome analysis," he said. "In the next few years at least, more conventional techniques will be used to identify variants, using microarrays to pick up those target variants, rather than looking across the whole genome."

The problems with data accuracy and data analysis are particularly relevant for the application of NGS in the clinic, which means that NGS will probably

Freelance journalist in London, UK. E-mail: ph@philiphunter.com
DOI 10.15252/embr.201541061 | Published online 7 August 2015

not arrive with a big bang, but will gradually extend its scope over years. The first major application will likely be for diagnosing inherited diseases associated with specific alleles. Such alleles—for diseases including cystic fibrosis, muscular dystrophy or Huntington's disease—can be identified without ambiguity from a blood-based genetic test, rather than a biopsy. As Zimmern noted, such inherited disorders are individually rare but collectively account for 5% of all human disease.

"SPIM now allows biologists to observe live specimens for up to a week or more with continuous observation, and is already yielding major insights..."

The second category of clinical application for NGS is the sequencing of whole genomes from tumour cells, rather than variants of individual genes. Most applications are still at an early stage, but there is great potential for personalizing treatment to both individual patients and specific tumours by matching tumour and normal tissues, according to Andrew Ludlow, Research Fellow at the University of Texas Southwestern Medical Centre. "This is the best example of clinical use of NGS against cancer, enabling genome-wide mutation profiling of matched tumour and normal tissues to personalize targeted therapeutic options for patients, with lots of examples now in the literature," he said.

Recent studies using NGS to match tumour genomes with normal tissue have made progress by probing the genetic origins of cancer and identifying whether the mutations were inherited or acquired [1]. Such information is decisive in determining whether a genomic aberration is somatic—that is found only in tumour cells and not in the matched normal tissue—or inherited through the germline. If the aberration is somatic, it may have arisen as a result of an environmental insult or a viral infection.

Identifying genomic aberrations—whether somatic or inherited—through NGS has great potential for early cancer diagnosis and screening, in particular by looking for copy number variation of individual genes or sequences of genes. Some of these variations are normal and heritable, while others are associated with diseases, including cancer. A key challenge that NGS is helping to address is identifying those somatic copy number alterations that are associated with cancer pathogenesis, as opposed to those that are the result of other genomic changes—in other words, separating cause from effect [2].

After cancer, the next important clinical application for NGS relates not directly to any specific disease type, but to the issue of drug selection and dose determination: NGS could be used to assess an individual's sensitivity to particular drugs and susceptibility to side effects. Its application can help by scanning for all the relevant enzyme-coding regions relevant for a particular drug and correlating them to assess the optimum dose. In a similar vein, NGS can also assess whether an individual is likely to suffer an adverse reaction against a particular compound.

The final frontier for NGS lies with complex diseases, such as asthma and rheumatoid arthritis, that have both genetic and environmental or infectious causes. "For these diseases, doing genetic tests on a single variant is totally useless," Zimmern said. "But if a software algorithm could combine 20 variants with five environmental variables, we might get a useful prediction." This application is some years away, however, and will involve a lot more research across multiple disciplines from genetics to epidemiology.

In addition to clinical applications, NGS also has great scope for use in epidemiology and disease monitoring. The ability to sequence whole bacterial genomes is valuable both for monitoring disease outbreaks and for analysing virulence and antibiotic resistance, according to W. Florian Fricke from the University of Hohenheim in Stuttgart, Germany. "Microbial whole-genome sequencing is likely to have a tremendous impact on public health and clinical microbiology," he said. "This will include increased accuracy, improved resolution, shorter reaction times and better integration with research from epidemiological studies and programs for surveillance and monitoring of microbial pathogens, as well as antimicrobial resistance and virulence traits across microbial populations. This knowledge could shift preventive measures from the clinic to the community and reduce over-prescription of antibiotics, resulting in decreased resistance development and dissemination."

Along with DNA sequencing, microscopy is another tool that has seen major advances in recent years, culminating in the 2014 Nobel Prize for Chemistry for Eric Betzig, Stefan Hell and William Moerner, who broke the diffraction limit in light microscopy, allowing biologists to now see individual molecules in the cell. Equally important was the invention of light sheet-based fluorescence microscopy (LSFM), which enables real-time observation of dynamical processes in living cells and tissues.

LSFM generates a light sheet by a laser that illuminates only a very thin slice of the sample, which minimizes photo damage. The sample is moved through the light sheet while a high-speed camera takes pictures from the successively illuminated planes, which are then combined to generate a three-dimensional image. The process has been improved in various ways, notably as selective/single plane illumination microscopy (SPIM), which enables sub-cellular resolution and greater real-time visualization [3]. SPIM now allows biologists to observe live specimens for up to a week or more with continuous observation and is already yielding major insights across both plant and animal biology, according to Ernst Stelzer, one of the field's pioneers, from Goethe University Frankfurt, Germany.

"The greatest challenges presented by new technologies in DNA sequencing, electron or light microscopy are [...] the enormous amount of data [...] that a single experiment can generate"

The application of SPIM in plant development has already yielded interesting insights into the processes underlying the formation of organs in plant embryos [4]. The key point in this study was that the development of living *Arabidopsis thaliana* samples could be observed over periods ranging from seconds to days with minimal damage. The study quantified the contribution of cell

elongation to the early morphogenesis of primordial lateral roots and observed diurnal variations in their subsequent growth.

This ability to observe live specimens over time in 3D presents some challenges, as Stelzer explained. First, the samples have to be prepared optimally for observation in 3D and have to be kept alive in conditions that are as close as possible to their natural environment. Second, the amount of data produced can be overwhelming. "The actual issue, in my opinion, is the fact that we look at specimens for more than a week; they survive and produce so much information and provide so many new insights that it simply takes a while to take it all in," Stelzer said. Notwithstanding, he believes that the enormous amount of data thus produced will encourage innovation and create a solid mathematical foundation for testing biological systems.

As much as LSFM and SPIM can benefit research in cell and developmental biology, the technology itself still presents challenges for scientists, according to Michael Knop, Group Leader at the DKFZ-ZMBH Alliance, in Heidelberg, Germany. The issue is the lack of commercial packages and the corresponding complexity in setting up a dedicated system for a given project. "You need computer specialists to set up, and all the solutions at present are experimental," Knop said. "Manufacturers have been relatively slow and have been reluctant to invest the resources given uncertain returns at present. Progress is also being held back by the old established companies trying to keep the technology to themselves. So things are moving very slowly." The benefits of SPIM are thus largely confined to wealthier, well-resourced institutions that have the money and knowledge to make the technology available to their research groups. But Knop is convinced that SPIM will become a major laboratory tool, by complementing the higher resolution of electron microscopy and being uniquely capable of addressing fundamental developmental questions. "We can start to connect things directly, to trace back and see, for example, how cells are organized and watch how nerves become active or fire. We can also see how the brain starts to work in a small embryo," Knop said. He added that SPIM could achieve even more if it became open to the same sort of innovation from start-up companies that has driven NGS.

The greatest challenges presented by new technologies in DNA sequencing, electron or light microscopy are thus not necessarily technological problems, but rather the enormous amount of data—in the gigabyte or terabyte range—that a single experiment can generate. The further success of these technologies and their application to medicine and to studying complex problems in biology will therefore depend on bioinformatics and the development of sophisticated algorithms for processing and analysing large data sets.

References

1. Li A, Liu Y, Zhao Q, Feng H, Harris L, Wang M (2014) Genome-wide identification of somatic aberrations from paired normal-tumour samples. *PLoS ONE* 9: e87212
2. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang C-Z, Wala J, Mermel CH *et al* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat Genet* 45: 1134 – 1140
3. Huisken J, Swoger J, Del Bene F, Wittbrodt J, Stelzer EHK (2004) Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science* 305: 1007 – 1009
4. Maizel A, von Wangenheim D, Federici F, Haseloff J, Stelzer EHK (2011) High-resolution live imaging of plant growth in near physiological bright conditions using light sheet fluorescence microscopy. *Plant J* 68: 377 – 385