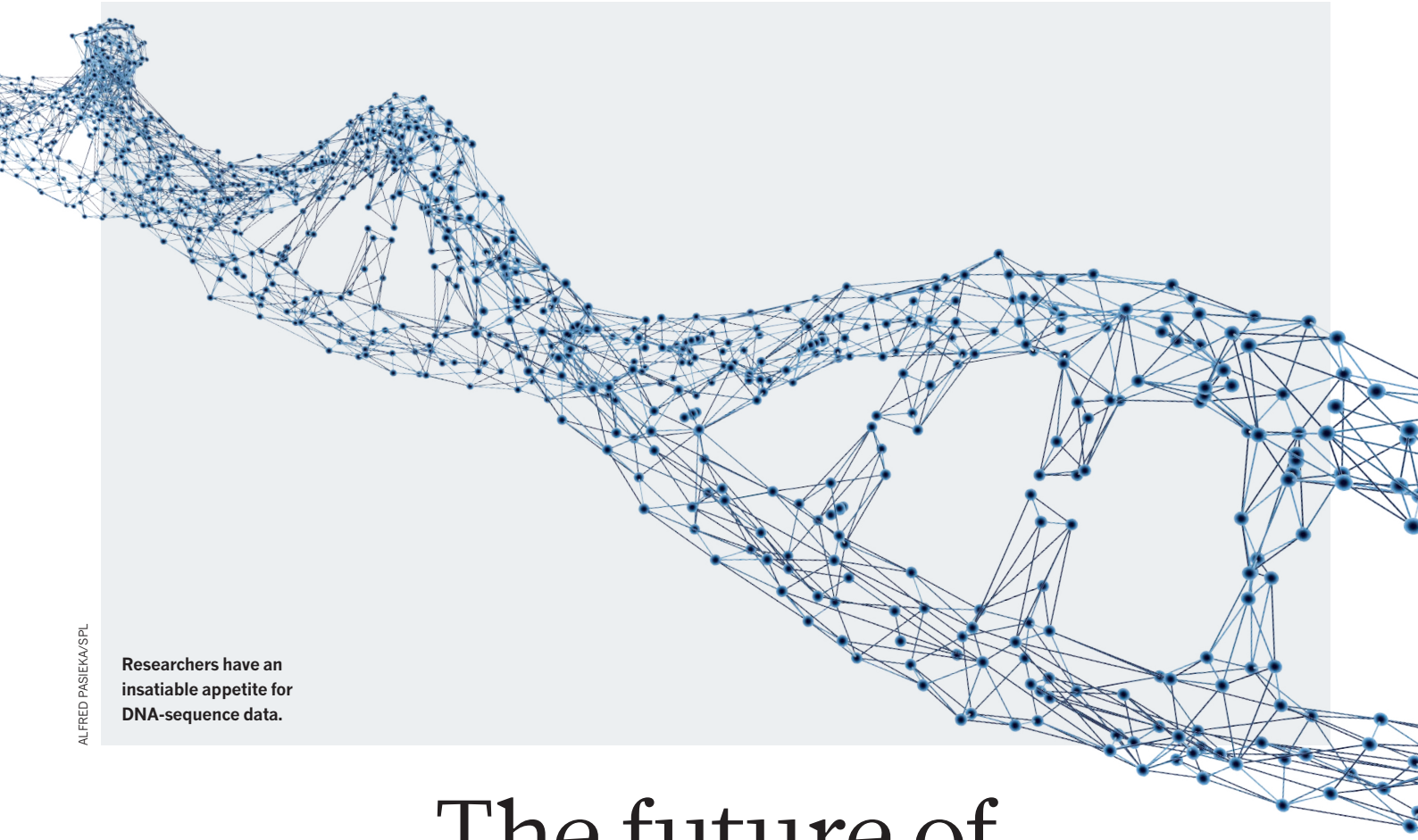
COMMENT



ALFRED PASIEKA/SPL

Researchers have an insatiable appetite for DNA-sequence data.

The future of DNA sequencing

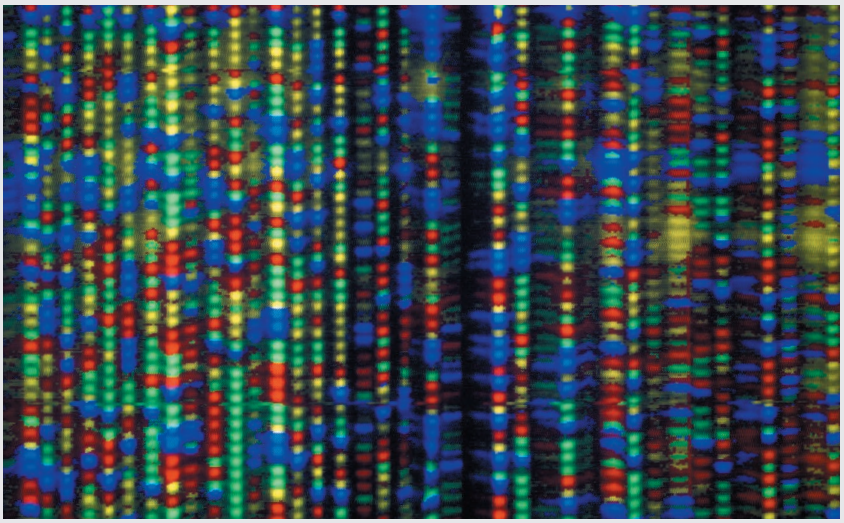
Eric D. Green, **Edward M. Rubin** and **Maynard V. Olson** speculate on the next forty years of the applications, from policing to data storage.

Forty years ago, two papers[1,2] described the first tractable methods for determining the order of the chemical bases in stretches of DNA. Before these 1977 publications, molecular biologists had been able to sequence only snippets.

The evolution of DNA sequencing from these nascent protocols to today's high-throughput technologies has occurred at a breathtaking pace[3]. Nearly 30 years of exponential growth in data generation have given way, in the past decade, to super-exponential growth. And the resultant data have spawned transformative applications in basic biology and beyond — from archaeology and criminal investigation to prenatal diagnostics.

What will the next 40 years bring? Prognosticators are typically wrong about which technologies — or, more importantly, which applications — will be the most disruptive. In the early days of the Internet, few predicted that e-mail that would achieve staggering popularity. Similarly, traders on Wall Street and investors in Silicon Valley failed to foresee that games, online video streaming and social media would come to dominate the use of today's available processing power and network bandwidth.

We would probably fare no better in ▶



KAREN KASMAUSKI/NGC

Coloured DNA bands.

PLATFORM PROGRESS

Many ways to sequence DNA

Over the past 40 years, the platforms for DNA sequencing have repeatedly been replaced.

By 1985, almost all DNA sequencing was performed with the Sanger or dideoxy chain-termination method[2]; reaction products were labelled with radionucleotides, separated on acrylamide slab gels, and detected with autoradiography (the use of X-ray or photographic film to detect radioactively labelled samples). By 2000, the four-colour-fluorescence method reigned supreme; reaction products were labelled with chain-terminating nucleotide analogues, separated electrophoretically in capillaries filled with a jelly-like media, and detected with energy-transfer fluorescent dyes. By 2010, the techniques had diversified. The dominant instruments were based on massively parallel analyses of DNA 'polonies' (clonal amplifications of a single DNA molecule) and on sequencing-by-synthesis chemistries (these rely on reversible chain-terminators).

From now on, the requirements for each DNA-sequencing platform will depend on what it is to be used for. In oncology and medical genetics, the goal will often be to identify every base correctly and to define every variant of genomic segments that exist in multiple copies. By contrast, when a yes or no 'match' is required — for instance, in species identification — the ability to run tests quickly and easily in the field may be more important than accuracy.

Another factor that will probably change is the relative need for centralized versus decentralized DNA sequencing. An epidemiologist trying to assess in real time what virus has affected a particular village in Sierra Leone might need cheap, portable devices. But for those generating massive data sets, it might be more efficient and cost effective to ship samples to centralized commercial operations, especially when the laboratories are required to meet exacting standards for quality control and sample tracking, as in clinical applications.

▶ predicting the future of DNA sequencing. So instead, we offer a framework for thinking about it. Our central message is that trends in DNA sequencing will be driven by killer applications, not by killer technologies.

IN DEMAND

Improvements in a technology can either increase or decrease demand. Microsoft co-founder Bill Gates famously cited radial tyres as an example of the latter: because they were more durable than earlier designs, the need for tyres dropped and the tyre industry shrank.

We think that DNA sequencing will follow the pattern of computing and photography, not of tyres. As it becomes cheaper and more convenient, applications will proliferate, and demand will rise (see 'Better, cheaper, faster'). As DNA sequencing breaks out of the research market and into clinical, consumer and other domains, the rule of 'more supply means more demand' will hold ever more strongly.

Researchers have an insatiable appetite for DNA-sequence data. In the 1990s, the idea of sequencing a human genome seemed daunting. Now, geneticists would like to have DNA sequences for everyone on Earth, and from every cell in every tissue at every developmental stage (including epigenetic modifications), in health and in disease. They would also like to get comprehensive gene-expression patterns by sequencing the complementary DNA copies of messenger RNA molecules. Meanwhile, archaeologists are beginning to reconstruct the flow of genes through ancestral populations, just as they previously deduced the flow of languages, cultural practices and material objects. And taxonomists, ecologists, microbiologists and evolutionary biologists are seeking to analyse the genomes of all living (and extinct) species — and even whole ecosystems.

Obviously, a sustained demand for data would require that the vast cataloguing efforts proffer actual understanding. At present, the bottleneck is analysing and interpreting all the DNA-sequence data. But just as new informatics approaches and massive data sets have dramatically improved language translation and image recognition, we predict that massive DNA-sequence data sets coupled with phenotypic information will enable researchers to deduce the biological functions encoded within genome sequences.

What's more, much of the basic science needed to interpret the data is already in place for a growing repertoire of practical applications (such as high-quality reference sequences of bacterial genomes, or the rules by which certain gene networks operate in healthy people). These range from recognizing microbial DNA sequences in unbiased surveys of environmental or clinical samples to identifying genome changes associated with known biological consequences.

KILLER APPLICATIONS

Over the years, the platforms for DNA sequencing have changed dramatically (see 'Many ways to sequence DNA'). Yet the trajectories of other technologies for which there is a seemingly insatiable demand — smartphones, the Internet, digital photography — suggest that the real disrupters will be the resulting applications, not the new technologies.

One domain where we are confident that DNA sequencing will be truly transformative is medicine.

Today's 'breakout' clinical application of DNA sequencing — in terms of the sheer number of tests conducted — is prenatal testing for the presence of an abnormal number of chromosomes, such as trisomy 21, which causes Down's syndrome. This test now relies on detecting the small amount of cell-free fetal DNA that circulates in maternal blood. Not even imagined at the end of the Human

Genome Project, it has been described as "the fastest growing genetic test in medical history"[4]. In fact, experts in the field estimate that some 4 million to 6 million pregnant women are now receiving this test each year worldwide, and that the number will surpass 15 million within a decade (D. Bianchi, D. Lo and D. Zhou, personal communication). Some of the hallmarks of the test seem likely to characterize many future applications of DNA sequencing in primary care: it is non-invasive, easy to perform and has low requirements for nucleotide-level accuracy (chromosomes can be counted without assessing sequence variation).

In high-income countries, genome sequencing is already used routinely to evaluate children with ill-defined congenital conditions. Analyses of the resulting sequences can reveal the disease-causing mutations in around 30% of such cases[5,6] — a figure that will only rise as the ability to interpret the data matures. In some instances, the resulting diagnoses have led to dramatic improvements in clinical management[7,8]. More typically, they benefit both families and physicians by ending a diagnostic odyssey and providing clinical clarity.

In oncology, considerable investments are being poured into the development of liquid biopsies[9]. It is easy to imagine such a sequence-based cancer test becoming a routine screening tool, used much like Pap smears and colonoscopies. With the advent of cancer treatments that target specific mutations, rather than tumour types[10], liquid biopsies could ultimately guide therapeutic interventions even when tumours are known to exist only from DNA-sequence signatures present in blood samples.

Various applications can be envisioned outside the clinic, too, particularly for handheld DNA sequencers. Epidemiologists and even caregivers working in rural areas could use such devices to test air, water, food, and animal and insect vectors, not to mention human throat swabs and body fluids. In fact, easy access to DNA-sequencing technologies in low- and middle-income countries is already facilitating projects such as the Global Virome Project. This aims to sequence numerous samples of wildlife DNA to identify a significant fraction of the viruses that can be transmitted into humans and cause disease.

"DNA-sequencing appliances could become the next 'smart' or 'connected' devices."

Meanwhile, public-health specialists are starting to discuss how they might sequence the DNA of all the microorganisms in the waste-water outlets of entire cities to speed up the recognition of disease outbreaks. And marine biologists are exploring ways to monitor the health of the oceans through systematic metagenomic studies.

On the street, portable instruments could bring DNA analysis out of the crime lab and make it a front-line policing tool. Police might be able to 'read' people's DNA, much as they currently check car number plates or identification documents. In fact, the degree to which cheap and easy DNA sequencing opens up possibilities for mass surveillance has recently sparked concern among human-rights groups.

In the home, DNA-sequencing appliances could become the next 'smart' or 'connected' devices, after smoke alarms and thermostats. One commentator even identified the toilet as the ideal place to monitor family health through real-time DNA sequencing[11].

HITTING LIMITS

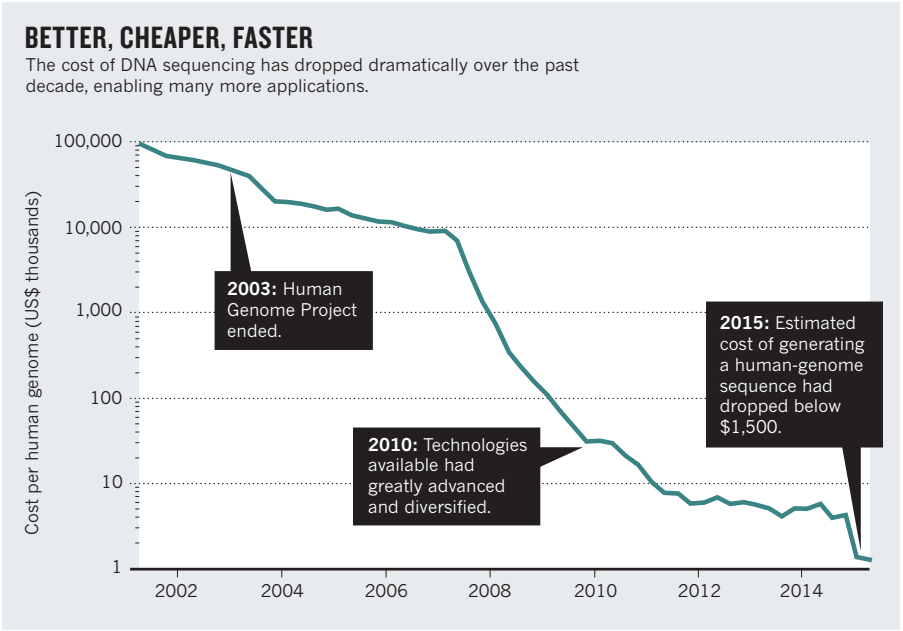
What are the stumbling blocks?

In a mere 40 years, the central goal of putting molecular data about cells to practical use has changed from an informational challenge to a meta-informational one.

Take clinical applications of genome-sequence data. It may soon be possible to use DNA sequencing routinely to analyse body fluids obtained for any clinical purpose. But only a vast amount of well-organized data about the multi-year medical histories of millions of people will provide the meta-information needed to establish when to ignore such data and when to act on them.

With respect to medicine, we echo the recommendations of advisory groups such as the US National Research Council's Precision Medicine Committee[12] on the need to create a vast "information commons". This would overlay molecular and clinical data onto the germ-line genome sequences of millions of individuals. Several such population-scale efforts are under way, including the UK Biobank resource and the US All of Us Research Program.

Here we have laid out our best guesses. Surprises are a certainty. In fact, it is possible that decades from now, much of the world's data (now residing on hard drives or in the cloud) will be stored in DNA, and that the main driver of DNA sequencing will be not our quest to tackle disease, but our insatiable appetite for data storage. ■



SOURCE: NATIONAL HUMAN GENOME RESEARCH INSTITUTE

Eric D. Green *is director of the US National Human Genome Research Institute at the US National Institutes of Health, Bethesda, Maryland, USA.* **Edward M. Rubin** *is chief scientific officer at Metabiota, San Francisco, California, USA.* **Maynard V. Olson** *is professor emeritus of medicine and genome sciences at the University of Washington, Seattle, Washington, USA.*
e-mail: egreen@nhgri.nih.gov

1. Maxam, A. M. & Gilbert, W. *Proc. Natl Acad. Sci. USA* **74,** 560–564 (1977).
2. Sanger, F., Nicklen, S. & Coulson, A. R. *Proc. Natl Acad. Sci. USA* **74,** 5463–5467 (1977).
3. Shendure, J. *et al. Nature* http://dx.doi.org/10.1038/nature24286 (2017).
4. Paxton, A. *CAP Today* (March 2017); available at go.nature.com/2hoipsp.
5. Bick, D. *et al. J. Pediatr. Genet.* **6,** 61–76 (2017).
6. Eldomery, M. K. *et al. Genome Med.* **9,** 26 (2017).
7. Worthey, E. A. *et al. Genet. Med.* **13,** 255–262 (2011).
8. Bainbridge, M. N. *et al. Sci. Transl. Med.* **3,** 87re3 (2011).
9. Alix-Panabières, C. & Pantel, K. *Cancer Discov.* **6,** 479–491 (2016).
10. Garber, K. *Science* **356,** 1111–1112 (2017).
11. Erlich, Y. *Genome Res.* **25,** 1411–1416 (2015).
12. National Research Council. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease* (National Academies Press, 2011); available at go.nature.com/2fmz99