Fast DNA Sequencing

DOI: 10.1002/anie.201106707

Decoding Genomes at High Speed: Implications for Science and Medicine

*Shankar Balasubramanian**

DNA · genome sequencing · genomics · personalized medicine · Solexa

The elucidation of the genetic code 50 years back[1] brought about a quantum leap in our understanding of nucleic acids, following the determination of the rules for recognition by Watson–Crick base pairing.[2] The code is based on a genetic alphabet of four bases that are read as sets of three bases which make up a codon corresponding to an amino acid. Thus the coding regions of a genome define the proteins that are encrypted in DNA sequences in nature. This knowledge generated an inevitable need for methods to decode the nucleotide sequence of DNA in the laboratory. It is noteworthy that the need for a method to determine nucleotide sequence had been recognized by Brown and Todd even earlier.[3] Two practical chemical strategies for decoding DNA emerged in the 1970s in the form of a sequence-selective chemical cleavage approach developed by Maxam and Gilbert,[4] and also a chain-termination DNA synthesis approach developed by Sanger and co-workers.[5] The Sanger approach ultimately prevailed as the most widely employed method and has undergone considerable optimization and automation to enable the routine sequencing of nucleic acids and small genomes. The ultimate application of the Sanger approach was the huge achievement of decoding of first human genome sequence by the International Human Genome Project Consortium.[6] The determination of nucleic acid sequence has become an essential and integral part of modern biology and medicine. There has very recently been an abrupt change in our ability to accurately decode DNA (and RNA) sequence at a speed, scale, and cost that stands to change the way we derive and utilize information about living systems. In this Essay I will focus mainly on describing the invention of a methodology called Solexa sequencing (now Solexa–Illumina sequencing)[7–9] and the impact that it is having on science, together with a vision for how it may influence medicine. There are indeed a number of other sequencing concepts described elsewhere that have either been reduced to practice or are in the process of being developed.[10]

In 1994 I started my independent academic career with just a small team of four graduate students, each working on a completely separate project. One of the projects employed Förster resonance energy transfer to explore details of how a DNA polymerase extends DNA. During that work I found it necessary to carry out time-resolved fluorescence measurements. This led me to consult and collaborate with my physical chemistry colleague, David Klenerman, who had also started his independent laboratory around the same time. Besides enabling the completion of the study, my interactions with David led to stimulating discussions on what else we might learn by interrogating molecules with optical methods that were (at that time) relatively new. We went on to explore single-molecule room-temperature imaging of a DNA polymerase with a template-primer DNA substrate to observe the incorporation of nucleotides during DNA synthesis. The design of various experiments inspired us to consider how the differential placement of fluorophores on the polymerase, the DNA, and the monomers (deoxynucleotide triphosphates, dNTPs) would enable us to visualize various aspects of the structure, function, and mechanism of DNA synthesis. In one particular arrangement we immobilized the DNA template-primer construct and visualized the template- and polymerase-directed incorporation of nucleotides that were fluorescently labeled. At the single-molecule level, this offered the prospect of observing DNA synthesis in real time. It was evident that such a format had the potential to decode the immobilized template through the sequential detection of the fluorophores being incorporated. At that time (1997), we were aware that our colleagues at the nearby Wellcome Trust Sanger Institute were making good progress on decoding the sequence of the first whole human genome as part of the International Human Genome Project Consortium. We were also deeply inspired by an imaginative concept for decoding (or sequencing) DNA by first synthesizing a DNA strand in which every base was encoded by a fluorophore and then systematically degrading the DNA from one end such that each nucleotide (and label) could be detected in sequence.[11] During 1997, David and I continued to enjoy intense discussions on the subject with some of our co-workers. In contrast to the degradative approach, we envisaged accurately decoding DNA by controlled stepwise solid-phase synthesis by a DNA polymerase using labeled building blocks

[*] Prof. S. Balasubramanian
Department of Chemistry, University of Cambridge
Lensfield Road, Cambridge, CB2 1EW (UK)
and
Cancer Research UK Cambridge Research Institute
Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE (UK)
and
School of Clinical Medicine, University of Cambridge
Cambridge, CB2 0SP (UK)
E-mail: sb10031@cam.ac.uk

modified with 3′O protecting groups and detachable fluorophores, whereby a cycle of incorporation would lead to the templated insertion of the correct nucleotide, which can be then read by imaging its color code (Figure 1). Removal of the color code and the 3′O protecting group resets the system in the next nucleotide position ready for the subsequent decoding cycle. Key to this approach was possibility of adding all four color-coded dNTPs together such that natural competition would facilitate the accurate incorporation of the correct nucleotide by the polymerase. This in turn required absolute chemical control to limit the incorporation to a single monomer, through the protecting group. An alternative, potentially simpler, approach would have been to avoid the need for a 3′O protecting group and simply incorporate one labeled nucleotide at a time and detect the absence or presence of a signal, then repeat these two steps systematically for each of the four nucleotides in turn. We did not favor the latter approach since genomes tend to be rich in stretches of more than one consecutive base of the same identity, and we reasoned that accurately counting the number of guanine, cytosine, thymine, or adenine units in such stretches could be best achieved by imposing absolute control of each incorporation.

The principle of sequencing DNA on a solid phase was attractive, as it would enable the simultaneous decoding of potentially enormous numbers of arrayed DNA samples or fragments in parallel, thus achieving a very large throughput. Figure 2 show a very early estimate that David Klenerman and I made on the basis of initial ideas. Our projections for read length and the degree of parallelization proved to be some way off of what was to come, but this simple analysis allowed us to foresee the potential to sequence on a human-genome scale (i.e. roughly a billion bases per experiment; for comparison a human genome corresponds to 3 billion bases). A practical approach to generate a solid-phase DNA sample array that is ready for sequencing had to be a "one-pot" sample prep that obviated the laborious individual preparation of each sample fragment. Inspired by "single-molecule thinking", we envisaged achieving this by fragmenting genomic DNA then immobilizing all the fragments on the surface of a chip at high dilution, such that an optically resolvable spot would be occupied by only one DNA fragment. This single-molecule DNA array would serve as a means to prepare many resolvable stretches of DNA that could each be decoded simultaneously by solid-phase sequencing.

Shankar Balasubramanian is the Herchel Smith Professor of Medicinal Chemistry at the University of Cambridge. His research interests are focused on the chemical biology of nucleic acids and the genome. He is a principal inventor of the leading next-generation sequencing methodology, Solexa sequencing. His other work includes the identification, elucidation, and manipulation of noncoding genetic elements, particularly four-stranded structures called G-quadruplexes.

David and I used our limited resources to go beyond our fundamental studies and initiate proof-of-concept work in our laboratories directed towards the chemistry and enzymology of solid-phase DNA sequencing along with imaging the decoding reactions on the surface of a glass chip. We were convinced that the approach could be completely reduced to practice. At the same time we realized that to do so in a format that could be practically utilized by others would require substantially more resources than we had, along with expertise in other disciplines. To help achieve our goal, we approached venture capital investors in late 1997 and after having our ideas, science, and plans scrutinized and challenged from numerous angles, we founded a spin-out company in the summer of 1998. We named the approach Solexa, which in part was derived from a joining and contraction of the words "***sol***o" and "mol***e***cul***a***r", with the inclusion of an "***x***" as a necessary consonant (!), as it had been stimulated by "single-molecule thinking".

Much of the next two years were spent advancing the proof-of-concept in Cambridge University, after which the project was moved out of the university into commercial premises (Solexa Limited) just outside Cambridge in 2000. Over the next five to six years an outstanding team of scientists and management was assembled to fully develop the Solexa sequencing concept into a robust commercial system. Along the way, the team raised further funding and there was also a takeover of a U.S. company, called Lynx, to provide a facility for instrument engineering and production.

The development of a sequencing system required an amalgamation of many technical areas that included chemistry to fully develop the sequencing nucleotides and engineer the surface of the sequencing chip, protein engineering to generate a high-fidelity, compatible DNA polymerase, molecular biology to create a practical sample preparation method, engineering and physics to build the imaging system and hardware, and computational skills to create the software and generate the algorithms for data analysis. A more detailed technical description can be found elsewhere.[7,8] In the fully developed system the one significant refinement from our original vision was the inclusion of a step to amplify the single-molecule sample array to generate many copies of the original DNA fragment on the same location of the chip. This was achieved by adapting an elegant bridge amplification method that had been invented elsewhere[12] and was being developed independently by a company called Mantea Predictive Medicine. Amplification of the single-molecule DNA array provided the advantages of a stronger signal that would require a lower-cost imaging system, whilst also reducing the negative impact of single-molecule stochastic errors to improve the accuracy of sequencing.

The very first genome to be sequenced using the Solexa methodology was that of the bacteriophage phiX174 in 2005. By 2006 the first exportable Solexa sequencing system, called the *Genome Analyser*, was launched with the capability of sequencing about 1 billion bases (1 gigabase, or 1 Gb) of DNA in a single experimental run. This milestone matched the "1 Gb" capacity that we had set as a goal in 1997 (Figure 2). This system was able to generate tens of millions of individual fragment reads, each with a read length of roughly

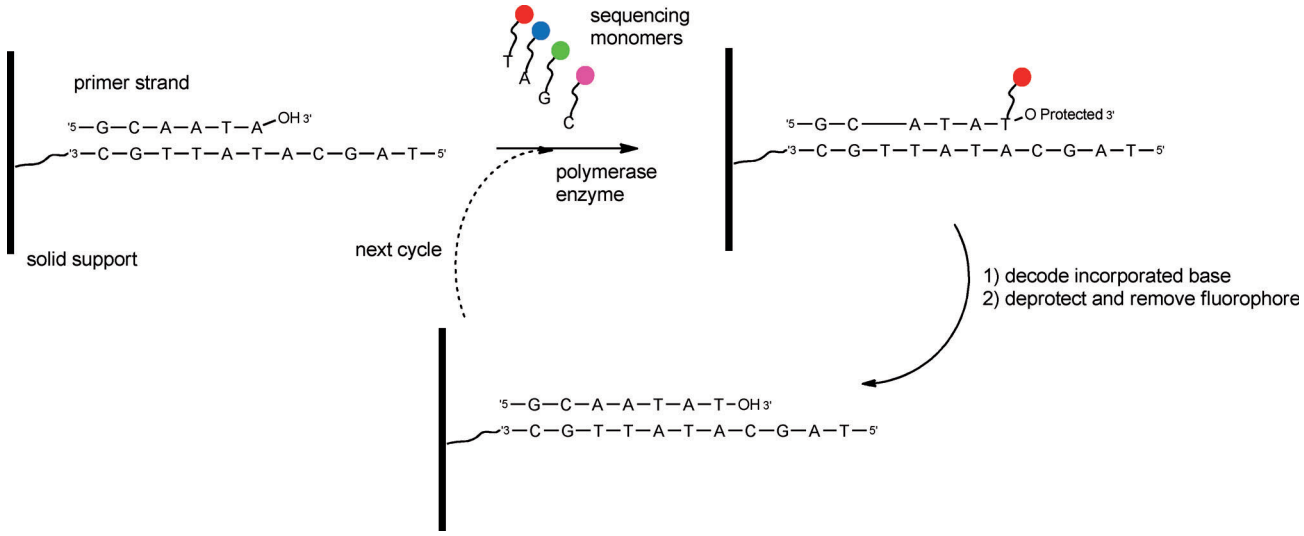

Figure 1. A schematic of solid-phase DNA sequencing. Each cycle permits decoding of a single base at each DNA sample by imaging the encoding fluorophore. This is now done at hundreds of millions to billions of DNA sample sites simultaneously on an integrated microfluidic system with flow cell(s).

Human genome 3 x 10^9 bases

Assume 300 X 300 array

10 s per cycle

$\therefore \frac{10^5}{10}$ bases per second = 10 kb s^{-1}

$\therefore$ one machine $\frac{3\times10^9}{10^4}$ s for human genome.

$\frac{3\times10^5}{60\times60}$ hours ≈ 100 hours ≈ 5 days

Figure 2. Our very first estimate (1997) of the potential sequencing capacity of the Solexa approach. In this approximate calculation, we had assumed only about 10^4 DNA fragments in a sample array. However, it was clear that sequencing roughly a billion bases of DNA would be possible in a matter of a few days. The *Genome Analyser* achieved this overall goal, but with a format that involved a much higher level of parallelization (ca. 10^7). Subsequently, the *HiSeq2000*, has exceeded these 1997 predictions by about 1000-fold.

35 bases. The *Genome Analyser* was used to generate the first human genomes to be sequenced by the Solexa approach, each of which was published in 2008; this included the genome of a Yoruban male,[7] the first genome of an Asian individual,[13] and the genome of a patient with acute myeloid leukemia along with a reference genome from normal skin cells from the same patient.[14] These genomes, published in a single journal edition, doubled the number of publicly available human genome data sets and put to rest any doubts that whole genomes could be sequenced by the massively parallel decoding of short DNA reads. What was then noteworthy was how rapidly the core technology was improved and optimized on many levels over the next few years within Illumina (who acquired Solexa and the sequencing technology in early 2007) to deliver a further three orders of magnitude of capacity. Today's incarnation of Solexa–Illumina sequencing, *HiSeq 2000*, routinely sequences about 600 Gb per run; this is about a million-fold higher than the routine capacity of the sequencing systems in common use in 1997. Overall, the past decade has seen the cost of sequencing a human genome fall from several hundred million U.S. dollars to below 5000 dollars—an approximately 100000-fold reduction.

The scientific passage from the core concepts to a high-speed sequencing system was not without doubts (or doubters). An early concern that sticks in my mind was whether any method for routine, low-cost sequencing of human genomes would be of interest to the scientific (and nonscientific) community. On reflection, that was not an unreasonable perspective given there was no human genome sequence (or market for whole-genome sequencing) back in 1997. We anticipated that the completion of the first human genome (to happen in 2003) would be followed by decades of further exploration in which routine human genome sequencing would be pivotal to the detailed elucidation of the nature of genome variation and its relevance to function and human disease. Our optimism, which was at least in part naive, was helpfully and enthusiastically endorsed by consultation with three eminent scientific leaders in human genome research[15] who were based at the Wellcome Trust Sanger Institute when we visited them in 1998.

Today, the sequencing of whole human genomes is starting to become routine and is also part of a number of prominent large-scale population studies that include the 1000 Genomes Project[16] and the International Cancer Genome Project.[17] We are also beginning to see the fruits of such studies. Comparative analysis of the growing number of accurate, complete human genome data sets is providing deep insights into the evolutionary history of human populations.[18] Whole-genome sequencing of cancer patients is providing detailed

knowledge of the somatic mutations that have contributed to the evolution of a cancer genome to the point where the cancer becomes symptomatic.[19] There are also early examples of how sequencing can track the evolution of a cancer genome of a patient undergoing standard cancer chemotherapy and subsequently inform clinicians of a more effective therapeutic that could stabilize the tumor.[20] A recent example has described how whole-exome (i.e. sequencing the protein-coding regions) sequencing of an infant with an unknown life-threatening disease revealed a mutation that led to a diagnosis and a subsequent life-saving bone-marrow transplant.[21] Such examples and others provide a glimpse of how accurate decoding of human genomes may one day transform the way we classify, diagnose, and treat many medical disorders. In the meantime, continued exploration of the applications of whole-genome sequencing will reveal the full potential of a vision for personalized, genomic medicine. There are clearly ethical and sociological considerations of great consequence that have been provoked by the advent of routine human genome sequencing; they include matters that relate to ownership, consent, privacy, and security of information and freedom of choice, to name but a few. Ethical frameworks are being carefully considered and in some cases implemented for research activity based on human genome sequencing.[22,23] This will remain an important, active area for engagement, debate, and formulation of policy that is essential if human genome sequencing is to constitute an integral part of daily life.

The ability to rapidly sequence genomes is also having a considerable impact on the study of organisms other than humans. The first complete genome sequences have been obtained for the giant panda,[24] numerous pathogens,[25] and plants.[26] The human microbiome project[27] seeks to apply deep sequencing of microbial genomes to characterize and understand the relationship between human health and changes in the human microbial flora. The benefits of high-capacity sequencing are also beginning to have a sizeable influence on strategically important areas of research and development that include agriculture and food security, environmental sciences, and the burgeoning area of bioenergy.

The ability to sequence "short" read lengths (35–100 bases) in a massively parallel fashion (on several billion fragments of DNA) has opened up a wide area of basic biology, besides genome sequencing, in a way that we had certainly not anticipated a decade back. Much of the credit for this rests with the creativity and imagination of users of the technology. The applications exploit the general principles that: 1) a short read from a DNA fragment can identify or "tag" a particular sequence context of say 40–100 bases; and that 2) the number of reads that comprise that sequence context allows one to digitally count the representation of that sequence context compared to others. A widely used application is the genome-wide high-resolution mapping of proteins (e.g. transcription factors or chromatin proteins) to genomic DNA (ChIPseq),[28] the determination of cytosine methylation sites to define the epi-genome (genome-wide epigenetics),[29] and high-resolution three-dimensional mapping of the structure of chromatin in cells.[30] A feature shared by all such endeavors is that they benefit from the capacity for such high-throughput sequencing methods to provide a means to generate outcomes without the need to presume any prior information about the sequences concerned. Such unbiased approaches are also well suited to the discovery and quantitative profiling of mRNA and noncoding RNAs such as micro-RNAs and other long noncoding RNAs. Given that much of the roughly 98% of the human genome that does not directly encode proteins (the noncoding sequences) is transcribed into RNA, there is clearly considerably more to be discovered and functionally elucidated in relation to the genome and the relevance of all that is encoded within the entirety of its sequence.

There has been a step change in the methodology, speed, cost, and capacity to sequence DNA that is helping science and medicine to be explored in a way that was not possible a decade ago. While the method I have described here is currently the most widely used approach for high-speed sequencing, there are other elegant approaches[10] that have been reduced to practice that include ligation-based sequencing on beads[31] or on DNA nano-arrays,[32] real-time single-molecule sequencing,[33] and a recently described proton-detection-based approach,[34] to name but a few. There are also creative sequencing concepts based on the use of nanopores currently being developed.[35] The speed and cost of genome-scale sequencing will continue to improve and accelerate the true democratization of large-scale sequencing; future applications are likely to be brought closer earlier than we might expect. The next technological step change (and challenge) for this field will be a method that enables a small hand-held, portable device to decode and interpret the genome, transcriptome, and epigenome from a spot of blood or saliva, in a matter of minutes; it is not inconceivable that such a device might be derived from one of the existing sequencing chemistries. I predict that whole-genome sequencing of individuals will become part of an integrated standard of care offered by the healthcare sector during the next ten years. The main hurdles to be overcome will likely be cultural, ethical, and economic, rather than technological.

In closing, I wish to draw attention to the fact that the basic concepts and experiments that underpinned Solexa–Illumina sequencing were born out of basic curiosity-driven research[36] that led to unexpected outcomes, rather than ideas that were strategically driven. I would also like to acknowledge that the reduction to practice of the ideas and methods into a variety of commercially available systems required the contributions of a large number of exceptional, talented, and committed individuals from the University of Cambridge, Solexa, and Illumina and their associates.

Received: September 21, 2011
Published online: December 5, 2011

[1] V. A. Erdmann, J. Barciszewski, *Angew. Chem.* **2011**, *123*, 9718–9724; *Angew. Chem. Int. Ed.* **2011**, *50*, 9546–9552.
[2] J. D. Watson, F. H. Crick, *Nature* **1953**, *171*, 737.
[3] D. M. Brown, A. R. Todd, *J. Chem. Soc.* **1952**, 52–58.
[4] A. M. Maxam, W. Gilbert, *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 560.

[5] F. Sanger, F. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5463.
[6] International Human Genome Project Consortium, *Nature* **2004**, *431*, 931.
[7] D. Bentley, S. Balasubramanian, H. Swerdlow et al., *Nature* **2008**, *456*, 53.
[8] S. Balasubramanian, *Chem. Commun.* **2011**, *47*, 7281 – 7286.
[9] There are currently over 1900 publications that have used the approach since its broad release. These are tracked and updated on the Illumina website: http://www.illumina.com/.
[10] M. L. Metzker, *Nat. Rev. Genet.* **2010**, *11*, 31 – 46.
[11] J. H. Jett, R. A. Keller, J. C. Martin, B. L. Marrone, R. K. Moyzis, R. L. Ratliff, N. K. Seitzinger, E. B. Shera, C. C. Stewart, *J. Biomol. Struct. Dyn.* **1989**, *7*, 301 – 309.
[12] C. Adessi, G. Matton, G. Ayala, G. Turcatti, J. J. Mermod, P. Mayer, E. Kawashima, *Nucleic Acids Res.* **2000**, *28*, 87e.
[13] J. Wang, W. Wang, R. Li, *Nature* **2008**, *456*, 60 – 65.
[14] T. J. Ley, E. R. Mardis, L. Ding, *Nature* **2008**, *456*, 66 – 72.
[15] They were: Dr. David Bentley, Dr. Richard Durbin, and Dr. Jane Rogers.
[16] The International 1000 Genomes Project Consortium, *Nature* **2010**, *467*, 1061.
[17] The International Cancer Genome Consortium, *Nature* **2010**, *464*, 993.
[18] H. Li, R. Durbin, *Nature* **2011**, *475*, 493 – 496.
[19] E. D. Pleasance et al., *Nature* **2010**, *463*, 191 – 196.
[20] S. M. Jones, J. Laskin, Y. Y. Li, *Genome Biol.* **2010**, *11*, R82.
[21] E. A. Worthey, *Genet. Med.* **2011**, *13*, 255 – 262.
[22] A. L. McGuire, T. Caulfield, M. K. Cho, *Nat. Rev. Genet.* **2008**, *9*, 152 – 156.
[23] For information on social, ethical, and legal issues provided by the U.S. Department of Energy and National Institutes of Health see: http://www.ornl.gov/sci/techresources/Human_Genome/elsi/elsi.shtml.
[24] R. Li, W. Fan, G. Tian, *Nature* **2010**, *463*, 311 – 317.
[25] C. Cunningham, D. Gatherer, B. Hilfrich, K. Baluchova, D. J. Dargan, M. Thomson, P. D. Griffiths, G. W. Wilkinson, T. F. Schulz, A. J. Davison, *J. Gen. Virol.* **2010**, *91*, 605 – 615.
[26] V. Shulaev, D. J. Sargent, R. N. Crowhurst, *Nat. Genet.* **2011**, *43*, 109 – 116.
[27] https://commonfund.nih.gov/hmp/.
[28] A. Barski, *Cell* **2007**, *129*, 823; D. S. Johnson, A. Mortazavi, R. M. Myers, B. Wold, *Science* **2007**, *316*, 1497.
[29] R. Lister, R. C. O'Malley, B. D. Gregory, C. C. Berry, A. H. Millar, J. R. Ecker, *Cell* **2008**, *133*, 523.
[30] E. Lieberman-Aiden, *Science* **2009**, *326*, 289.
[31] K. J. McKernan et al., *Genome Res.* **2009**, *19*, 1527 – 1541.
[32] R. Drmanac et al., *Science* **2010**, *327*, 78 – 81.
[33] J. Eid et al., *Science* **2009**, *323*, 133 – 138.
[34] J. Rothberg et al., *Nature* **2011**, *475*, 348 – 352.
[35] A. Meller, L. Nivon, E. Brandin, J. Golovchenko, D. Branton, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 1079 – 1084; J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, H. Bayley, *Nat. Nanotechnol.* **2009**, *4*, 265 – 270.
[36] I acknowledge the Biotechnology and Biological Sciences Research Council of the UK for funding the basic science that underpinned Solexa-Illumina sequencing.

