DNA SEQUENCING: NUMBERS, NUMBERS, NUMBERS...

Peter Milburn*
Australian Cancer Research Foundation Biomolecular Resource Facility,
John Curtin School of Medical Research, Australian National University, ACT 0200
*Corresponding author: peter.milburn@anu.edu.au

Introduction

The first generation of techniques for reading DNA sequence by the electrophoretic separation of dideoxynucleotide terminated products from DNA-dependent DNA polymerase synthesis (Sanger sequencing) (1,2) remains the gold standard for both accuracy and read length. *But for how much longer?* Sequencing eukaryote genomes with this technique alone required near-industrial scale effort and, largely, has been superceded by more recently developed strategies that have orders of magnitude higher output. The quantum increase in sequencing power has brought about a paradigm shift, such that we now ask the hitherto inconceivable question *'How many individual genomes shall we sequence from how many species?'* Beyond accelerating the investigation of biological variation at the genome level, this massive sequencing power has enabled new ways of exploring the mechanisms by which the expression of genes is regulated. Not only the underlying control mechanisms may be explored, but also, the genomic elements that are translated into the transcriptome may be determined in quantitative fashion. It is also possible to interrogate whole populations of microorganisms in spatial samples such as air, soil, water or food in metagenomic applications. The emergence of new sequencing platforms and improvements to those extant occur on a regular basis. The relationship between biological questions and the burgeoning choice of commercially available sequencing platforms is complex and continually evolving.

Numbers Past

In September 2005, the details of a sequencing-by-synthesis technology that had been developed by 454 Life Sciences were published (3). The first commercially available version of the instrument provided more than an order-of-magnitude increase in productivity and nearly an equivalent reduction in cost compared to the best Sanger technology platforms. However, the median read length of around 100bp seemed meagre and was uninspiring to a scientific community used to the >750bp high-quality reads of the incumbent sequencing engines. *How useful could these short sequence reads be?* Additionally, base-calling accuracy was below that luxurious standard to which we had become accustomed. The fact that several hundred thousand reads were obtained in a single run was somehow overlooked. Perhaps there were just too many for an era when 96 sequences per run seemed like a lot and people liked to print out their data! The median read length increased to ~250bp within two years and then to almost 400bp with the release of the Titanium chemistry in late 2008. Importantly, the latter development brought both higher density and reduced reactor size in the picotitre plate, which increased the number of reads obtainable to over one million in a single run. The chemistry is fast and 200 cycles of chemistry are accomplished readily in a day.

In late 2005, Solexa announced the release of its 1G Genome Analyzer, in which the sequencing-by-synthesis took place on a glass surface rather than in microfabricated reactors. This apparently simple difference, along with spectral resolution of base addition, enabled the sequencing chemistry to occur at much higher density than in the picotitre plate format. Tens of millions of sequences were obtained in a single run, over a period of days rather than hours, but read lengths were only up to 35bp. *How useful could these even shorter sequence reads be?* Just as 454 Life Sciences had combined with Roche, Solexa combined with Illumina and technical improvements flowed steadily from the union. From the birth of the Illumina GA*I* in 2007, developments in chemistry, software and hardware yielded progressive increases in read length and data output through to the GA*IIx* in 2009, which delivers up to 100bp read length, >400 million reads from a paired end run and >40Gbases of DNA sequence!

Also in 2005, a publication from Church and coworkers (4) described a method for multiplex sequencing of paired genomic tags that involved the ligation of fluorescently labelled sequence probes to produce discontinuous templated sequence reads. This was commercialised and launched as the SOLiD System in mid-2007. This technology was well outside the comfort zone of those of us used to reading DNA sequence by identifying the templated sequential addition of a nucleotide by a polymerase. Importantly, this method allows each locus to be interrogated more than once, which yields the potential for very high primary base-calling accuracy. The by now familiar rapid developments in chemistry, software and hardware yielded progressive increases in read length and data output through SOLiD 3 that produces *billions* of reads to 50bp.

Each of these three sequencing technologies requires monoclonal DNA polymerase amplification of DNA fragments to produce reads that are the canonical sequence of the initial fragments in a massively parallelised process, somewhat similar to the Sanger sequencing of individual PCR amplicons. The HeliScope, launched in early 2008, was the first commercially available instrument that read DNA sequence from *single molecules*. An obvious advantage of this fact was that it removed cost and labour associated with amplification of the sample libraries prior to sequencing. More subtle was the point that any biases in the amplification process are obviated, making it ideal for quantitative applications that call for counting

sequence tags, e.g., chromatin profiling (5). Billions of tags are obtained at read lengths of around 35bp and upwards across 50 sample lanes. It should be noted that the primary base-calling accuracy for single molecule reads is inherently lower than for canonical reads, making them less desirable for *de novo* sequencing applications.

In their initial publication in 2005, Margulies *et al.* (3) dealt with the perceived problem of lower base-calling accuracy by stressing that the "high" volume of data output allowed one to generate consensus sequence that was of a similar accuracy to that obtained from Sanger sequence data. Thus, even at the much lower coverage levels attainable in the past, *the short and very short reads obtained from these new sequencing platforms were useful indeed!*

It soon became apparent that data obtained from the Roche 454 and Illumina GA platforms were complementary, the former being more prone to small indel errors and the latter to base substitution errors. Aury *et al.* (6), for example, demonstrated the production of high-quality draft sequences for prokaryotic genomes *without using Sanger sequencing*; the long Roche 454 reads producing high continuity and paired-end Illumina GA reads polishing the consensus sequence by reducing the number of indel errors. This study is an example of how combining data with different attributes proves to be more powerful than increasing coverage alone, the benefit of which plateaus rather quickly.

Numbers Present

The present is a very small space when it comes to technical discussion of DNA sequencing platforms! In just a few years, the scientific community has become comfortable with less-than-perfect base calling and shorter read lengths. However, Roche 454 have recently presented data with median read lengths well in excess of 700bp, *approaching that of contemporary Sanger methodology*. These longer read lengths are particularly useful for *de novo* sequencing, sequencing through repeat regions, co-locating exons and linking associated variants as phased haplotypes. We await the release of this extended chemistry with bated breath! This is now a very well established sequencing technology, and at the time of writing, there were 748 peer-reviewed publications collated at http://454.com/publications-and-resources/publications.asp

Early this year, Illumina announced the advent the HiSeq 2000, and at 200Gbases from 1 billion sequences per run, this has a 5-fold improvement in data output over the existing technology at launch. Once the initial libraries have been made, the technology is now largely 'plug-and-play' and has been widely adopted. At the time of writing, there were 363 peer-reviewed publications collated at http://www.illumina.com/publications

One week after the launch of the HiSeq 2000, Life Technologies launched the SOLiD 4 System that generates up to 100Gbases of mappable sequence from greater than 1.4 billion reads per run and scales to 300Gbases of mappable sequence from greater than 2.4 billion reads per run with the SOLiD 4*hq* System upgrade later in the year. Significantly, the SOLiD 4*hq* System offers a change in sequencing fundamentals by interrogating each base five times before the sequence is called, generating base-call accuracy in excess of 99.98%. The SOLiD technology was later to market and less familiar to the scientific community so has remained somewhat in its own niche. At the time of writing, there were 52 peer-reviewed publications collated at http://www5.appliedbiosystems.com/tools/citations/

The benefit of increasing coverage to improve consensus accuracy only applies to fixed problems, e.g., describing the sequence of a genome. While random effects disappear at high coverage, overcoverage per se is not a good thing because systematic errors start to appear at low but realistic frequency. Currently, 30-fold coverage is considered to be the gold standard for optimal consensus accuracy.

We might reasonably expect the demonstrated output from both HiSeq and SOLiD to leapfrog each other in the race to sequence a complex >3Gb genome at 30-fold coverage for under the benchmark US$1000. While it is truly remarkable that so much sequencing data may be generated in a matter of days at such low cost, one has to remember that these data must be stored somewhere and are of little use without bioinformatics analysis. Both storage and analysis have an associated cost! A new challenge for DNA sequencing technologies is to increase the base:byte ratio in the data output.

Increased output is particularly useful for quantitative applications that call for counting sequence tags, e.g., digital transcriptomics (7). The key number here is the number of mappable sequence tags obtained per run, rather than the total number of bases sequenced. The ability to both map and assemble DNA sequence reads increases dramatically if paired reads are obtained that are separated by a predictable interval. This has been recognised for a long time, but it is only recently that reliable methods for attaining this goal have been made available. It is more productive to achieve modest coverage with large insert paired reads than overly high coverage with short insert paired end reads.

We have entered an era wherein not only several mammalian genomes may be re-sequenced at 30-fold coverage in a single sequencing run, but also where this increased sequence output also allows one to mine ever deeper into open systems such as environmental samples. In this type of application, there is no theoretical limit to the number of different sequences that might be sampled, so we require as many of the longest possible reads at the highest possible accuracy. It seems hard to imagine, but there remains room for significant improvement in DNA sequencing.

Numbers Future

The future will soon be upon us! Process speed from sample to sequence is the new number to watch. Early this year, Pacific Biosciences announced that it would soon be shipping its Single Molecule Real Time (SMRT) Sequence Analysis System (8). In its highly engineered SMRT Cell, up to 80,000 individual polymerase molecules capture primed DNA templates and report the incorporation of one to three fluorescently labeled nucleotides per

second. Very long reads, in excess of 10,000bp, have been demonstrated, but we do not know if this is the exception or the rule because no data on median read length are available. *Such read lengths are far beyond those available from Sanger sequencing* (1) and would admit a range of exciting possibilities, from sequencing complete transcripts to the unambiguous identification of viral quasispecies. Base-calling accuracy may be improved by producing consensus sequence from multi-pass reads of a circularised template, but this requires sacrifice of overall read length. The process requires neither template amplification nor cycles of chemistry, which clearly reduces cost and *saves a lot of time*. The SMRT cells are loaded in strips and run sequentially with essentially no downtime between runs. Moreover, the sequencing process itself is fast; the polymerase adds bases directly from a pool in a fashion unfettered by the need for cycles of chemistry. The kinetics of addition are sensitive not only to sequence context but also to base modification in the template such that, for example, *methylated bases are detected directly*. This is something fundamentally new and is unique to the realm of single molecule DNA sequencing.

Ion Torrent also announced that they intend to ship sequencing platforms later this year. Their technology is based on the electrical detection of DNA synthesis on a surface (9). DNA molecules are tethered in wells on complementary metal oxide semiconductor (CMOS) chips. Each of the four nucleotides is delivered sequentially in cyclic fashion and voltage changes proportional to the number of bases added reveals the sequence of the template. The process is reminiscent of that employed in the Roche 454 sequencing method, but individual protons are detected rather than bursts of photons. Details of the process are scant, but the method does not require optical imaging or molecular tagging, which brings significant savings in chemistry, hardware requirements and data space. It is claimed that, in a run that takes about an hour, the instrument will generate 'hundreds of millions of bases' and 'millions' of highly accurate reads per run, each several hundred bases in length each.

There is no doubt that further sequencing platforms will be announced in the not too distant future.

Conclusions

Base-calling accuracy, read length, speed per gigabase, mappable reads per run, cost per gigabase and sample-to-sequence speed are the numbers that best describe performance. Every technology available commercially has a different sweet spot and that is great news for nucleomics.

References

1. Sanger, F., Nicklen, S., and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467
2. Sanger, F., (1998) *Ann. Rev. Biochem.* **57**, 1-29
3. Margulies M., Egholm, M., Altman, W.E., *et al.* (2005) *Nature* **437**, 376-380
4. Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. (2005) *Science*, **309**, 1728-1732
5. Goren, A., Ozsolak, F., Shoresh, N., Ku, M., Adli, M., Hart, C., Gymrek, M., Zuk, O., Regev, A., Milos, P.M., and Bernstein, B.E. (2010) *Nature Methods* **7**, 47-49
6. Aury, J.-M., Cruaud, C., Barbe, V., Rogier, O., Mangenot, S., Samson, G., Poulain, J., Anthouard, V., Scarpelli, C., Artiguenave, F., and Wincker, P. (2008) *BMC Genomics* **9**, 603
7. Blow, N., (2009) *Nature* **458**, 239-242
8. Eid, J., Fehr, A., Gray, J., *et al.* (2009) *Science* **323**, 133-138
9. Pourmand N., Karanek, M., Persson, H.H.J., Webb, C.D., Lee, T.H., Zahradnikova, A., and Davis, R.W. (2006) *Proc. Natl. Acad. Sci. USA* **103**, 6466-6470