Mind the gaps

Steven L Salzberg

By grouping short reads derived from the same long genomic fragment, the reads can easily be assembled into fragments that approach the length of capillary sequencing reads.

Amid all the excitement about next-generation sequencing, scientists often neglect to mention the problems that are caused by short read lengths. Genome assemblies produced from short reads are far more fragmented than those produced from long reads, with many more gaps and with relatively poor long-range linking information. Fortunately, technology to produce long reads is still being pursued vigorously in some laboratories. In this issue, a group led by Jay Shendure[1] describes a method to create 'reads' as long as 700 base pairs (bp) starting with pairs of reads that are just 20 bp and 76 bp.

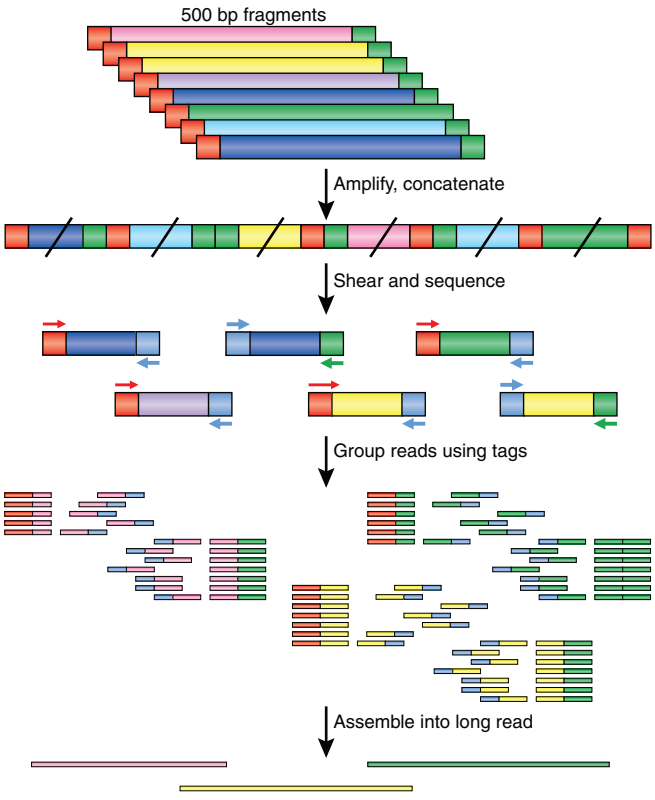

Figure 1 | Turning short reads into long reads using subassembly. The process begins with size-selected fragments of approximately 500 bp. The fragments get unique tags on both ends (red and green), and all fragments are then amplified using PCR. Random shearing breaks each fragment at many different places. The sheared fragments are sequenced from both ends, producing reads that originate all along the fragments. These reads can then be clustered together based on the unique tags and assembled to produce 'reads' that are nearly as long as the original DNA fragment.

Steven L. Salzberg is at the Center for Bioinformatics and Computational Biology and Department of Computer Science, University of Maryland, College Park, Maryland, USA.
e-mail: salzberg@umd.edu

For more than a decade after its introduction, automated sequencing steadily produced longer reads until, by the early 2000s, the ABI 3730 line of capillary sequencers routinely produced read lengths over 800 bp. As read lengths increased, costs came down, and genome assembly became easier. It was also easier to capture full-length transcripts for protein-coding genes with the longer reads, which in turn made gene discovery easier and more reliable.

Then along came next-generation sequencers, with read lengths of 25–100 bp, and although costs per megabase got dramatically cheaper, the quality of the assembled genome sequences got worse—in some cases much worse. Read length is once again increasing, but we still have not returned to the read lengths of the early 2000s. Genomes sequenced entirely with short reads, such as the recently published panda genome[2], are starting to appear. On the one hand, it is truly remarkable that we can assemble a large genome entirely from short reads; on the other hand, the number of gaps in the resulting assembly—over 200,000 in the panda genome—leaves much to be desired.

Similarly, the rapid decline in costs owing to next-generation sequencing has driven an explosion of metagenomics projects. Scientists begin with DNA extracted from an environment such as soil, water or the human gut and sequence the unknown mixture of species contained in it. Some of these environments are remarkably rich in species, containing thousands of unknown bacteria, archaea and viruses[3,4]. Unlike whole-genome shotgun sequencing projects, which begin with a clonal sample grown from a single bacterial cell, metagenomics samples contains complex populations, in which any one species might be represented by many slightly different variants.

For metagenomics projects, one of the main challenges is to identify the genes in the environment being sequenced. This becomes very difficult when the sequences

contain only fragments of those genes. Fortunately, bacterial genomes tend to be very gene-rich, and fragments of just a bit longer than a kilobase are likely to contain complete genes. If only there were a way to convert short, next-generation reads into longer fragments.

The method developed by members of Shendure's lab[1] can be used to create 'sub-assembled' (SA) reads averaging ~500 bp from a pair of short reads: a 20-bp 'tag' read and a 76-bp 'breakpoint' read. In this method (**Fig. 1**), first DNA is sheared and size-selected to pull out fragments of ~550 bp. These fragments are PCR-amplified, tagged, concatenated and sheared randomly to produce from each original fragment a set of shorter fragments. The shorter fragments are then sequenced from both ends using an Illumina Genome Analyzer II. The 20-bp sequences from tagged ends are used to group together the reads, and the breakpoint reads, all of which came from copies of the same original DNA fragment, are then assembled together. The resulting sub-assemblies, or SA reads, can then be treated as long reads and assembled using conventional genome assembly software.

To demonstrate that their method works, Hiatt *et al.*[1] applied it to two different problems: a conventional, whole-genome shotgun assembly of a bacterial genome and an assembly of a metagenomics sample. In the first case, they applied their method to a strain of *Pseudomonas aeruginosa*, a pathogenic bacterium associated primarily with lung infections. When combined with a modest amount of paired-end sequence data, the result was comparable to what one might expect from a conventional sequencing project: they assembled the 6.1 Mb genome into just 32 scaffolds covering >99% of the genome. For the metagenomics experiment, they used DNA collected from sediment at the bottom of a lake and produced SA reads with a median length over 400 bp. When compared to conventional Sanger sequencing of the same sample, the new technique produced a comparable amount of total sequence (after assembly), although somewhat shorter contig lengths, with far less sequencing effort.

A reasonable question is how well this method compares to pyrosequencing with the Roche 454 sequencer, which can currently produce read lengths of ~400 bp, similar in length to the SA reads. In supplementary data, Hiatt *et al.*[1] report that their cost (per megabase) is approximately half of that of 454 sequencing, and they also point out that their SA reads, because they are assembled from multiple overlapping short reads, have a much lower error rate than individual 454 reads. And although they focused on 'long' reads of 500 bp, in principle they could use any length for which next-generation sequencers can capture both ends.

With its very low costs, short-read sequencing will likely dominate the genome sequencing world until something better comes along. But short reads will always leave more gaps than long reads, and no one likes gaps. As the scientific community sequences an unending stream of genomes and metagenomes (and there seems to be an almost limitless number in our biosphere), techniques for producing long DNA sequences will always be needed, until the day arrives when we can grab onto one end of a chromosome and sequence the entire molecule in one go. That day is likely quite far off.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

1. Hiatt, J.B., Patwardhan, R.P., Turner, E.H., Lee, C. & Shendure, J. *Nat. Methods* **7**, 119–122 (2010).
2. Li, R. *et al. Nature* advance online publication, doi:10.1038/nature08696 (13 December 2009).
3. DeLong, E.F. *Nat. Rev. Microbiol.* **3**, 459–469 (2005).
4. Turnbaugh, P.J. *et al. Nature* **449**, 804–810 (2007).

Advancing neurochemical monitoring

Paul A Garris

Two new approaches to neurochemical monitoring *in vivo*—an improved real-time microsensor and genetically engineered cells that sense neurotransmitter levels—address the critical issue of brain reactivity to implanted devices.

Identifying the neural basis of behavior is a core focus of neuroscience. One prominent methodology in this pursuit is monitoring the neurotransmitters that underlie communication between neurons. Although technical improvements have advanced neurochemical measurements to the real-time domain, one critical limitation of present methods is the highly invasive nature of implanting a recording device and the subsequent reaction of brain tissue. Neuroinflammation not only alters the sampled microenvironment, but also results in a diffusion barrier that encapsulates the probe and therefore restricts access to released neurotransmitters. Taking radically different strategies, two new approaches address this key hurdle for achieving the long-standing goal of chronic, real-time neurochemical monitoring. In this issue of *Nature Methods*, Clark *et al.*[1] describe a microelectrode that retains the capability for subsecond dopamine measurements *in vivo* for months. In *Nature Neuroscience*, Nguyen *et al.*[2] report implantable genetically engineered cells for electrode-free acetylcholine sensing.

Microdialysis[3] and voltammetry[4] have dominated the modern era of neurochemical monitoring *in vivo*. With exquisite sensitivity and selectivity by virtue of removing brain analytes for *ex vivo* determination, microdialysis is better suited for measuring basal neurotransmitter levels. By using electrochemistry at the probe tip for *in situ* detection, the superior temporal resolution of voltammetry is more appropriate for capturing faster chemical signals.

Recent advances in voltammetry have overcome the historical criticisms of poor sensitivity and chemical specificity. Indeed, by providing nanomolar and subsecond measurements and a chemical signature in the form of a voltammogram, fast-scan cyclic voltammetry (FSCV) has met the demanding analytical criteria for monitoring phasic dopamine

Paul A. Garris is in the Department of Biological Science, Illinois State University, Normal, Illinois, USA.
e-mail: pagarri@ilstu.edu