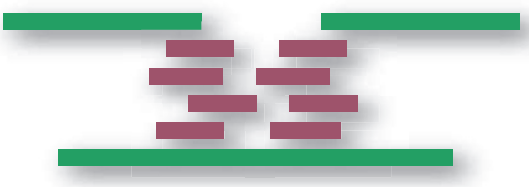
GENOME WATCH

Seeking perfection

Thomas D. Otto

This month's Genome Watch discusses ways to automatically produce 'base-perfect' genome sequences.

Recent advances in sequencing technologies have enabled a host of new applications for whole-genome sequencing, which have been reviewed in this column in recent months. As many laboratories now have access to sequencing machines that can generate high-depth data, one might think that the quality of the assembled genome sequences would also increase. However, the only easily determined measure of quality for genome sequencing is the number of gaps in the sequence. Other quality control criteria, such as the number of mis-assemblies, the number of base errors in 100 kb and whether the assembly captures all replicons (including kinetoplast DNA, chromosomes or plasmids), can only be determined after additional microanalysis. This explains why 1,284 of the 2,453 submitted bacterial genome sequences in the NCBI Entrez Genome database (http://www.ncbi.nlm.nih.gov/sites/genome) were submitted as draft sequences — that is, they contain gaps and might have low-quality regions. One reason why so many genomes are submitted with gaps is the smaller reads (25–450 bp) that second-generation sequencing technologies give compared with the older capillary sequencing technologies, which could produce up to 800 bp of good-quality sequence. It is difficult to correctly assemble repeats that are longer than the read length. The quality of the library — in terms of coverage, representation and sequencing errors — also influences the outcome of the assembly.



One reason for producing 'only' a draft genome might be that the cost and effort of manual finishing, during which gaps are closed by designing and sequencing PCR products, is not justified by its scientific value. An elegant algorithm to overcome this issue was recently developed by Tsai *et al.*[1], exploiting the fact that each DNA template is sequenced from both ends to yield two reads that are called a mate pair. If one of these reads flanks a gap, then one can infer that its mate falls inside the gap; local assemblies can then be used to close the gap or at least reduce its size. The success rate of this technique is high: up to 80% for two *Salmonella enterica* genomes. However, this method has its limits; for example, gaps that are caused by repeats cannot be closed, contigs must be scaffolded and, clearly, a paired short-read library is required. To close the remaining gaps, tools such as ABACAS (algorithm-based automatic contiguation of assembled sequences)[2] can be used to align contigs against a reference and then automatically design PCR primers.

Some groups are simply lucky. For example, Unemo *et al.*[3] obtained just three contigs out of a 454 sequencing assembly for a 1 Mb *Chlamydia trachomatis* genome. The gaps were due to the two ribosomal RNA operons. After the draft sequence had been assembled manually, the next step was to find base errors in the assembly using iCORN (iterative correction of reference nucleotides)[4]. This tool iteratively maps short reads against completed genome sequences of the same isolate to find discrepancies and correct them. In the iCORN paper, several errors were found in finished genomes, including 2,000 errors in the *Plasmodium falciparum* reference genome, and many 454 sequencing homopolymer tracts were corrected. Systematic errors (involving long homopolymer tracts of >12 bp) were also observed in Illumina reads, indicating that the dream of obtaining more 'base-perfect' genome sequences using this technology could be some way off. iCORN allows only those corrections that do not decrease the coverage of perfect mapping reads, as determined using the read-mapping tool SNP-o-matic[5]. Obviously, one discrepancy in the sequence will make it impossible to map reads over this position, and the perfect mapping coverage plot would therefore drop to nearly zero. The limitations of this approach involve, again, short tandem repeats and larger repeats. Most of the bases of the *C. trachomatis* genome were covered with perfect mapping reads of at least 40×. The exceptions were around the origin of replication, as SNP-o-matic does not map onto circular genomes.

Some people might argue that enough has been done, as long as all gene models are assembled and they are covered by perfect mapping reads. However, if a genome sequence will be used as a reference sequence for further projects for population genetics, phylogenetics or RNA-Seq experiments, gaps and small errors can hinder the downstream analysis. At the very least, regions with lower confidence should be tagged, which would make life easier for downstream users of the sequence.

Thomas D. Otto is at the Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.
e-mail: tdo@sanger.ac.uk

1. Tsai, I. J., Otto, T. D. & Berriman, M. Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* **11**, R41 (2010).
2. Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**, 1968–1969 (2009).
3. Unemo, M. *et al.* The Swedish new variant of *Chlamydia trachomatis*: genome sequence, morphology, cell tropism and phenotypic characterization. *Microbiology* **156**, 1394–1404 (2010).
4. Otto, T. D., Sanders, M., Berriman, M. & Newbold, C. Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics* **26**, 1704–1707 (2010).
5. Manske, H. M. & Kwiatkowski, D. P. SNP-o-matic. *Bioinformatics* **25**, 2434–2435 (2009).

Competing interests statement
The author declares no competing financial interests.