GENOME WATCH

Genome annotation: man versus machine

Nicola K. Petty

Following on from last month's discussion of sequence assembly and correction, this month's Genome Watch examines genome annotation in the context of advances in second-generation sequencing.

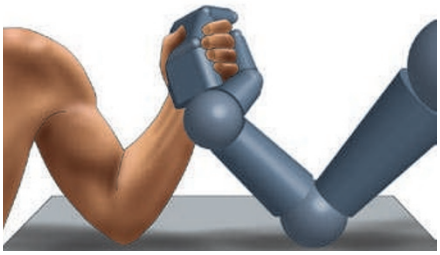
Genome sequencing has been brought within reach of many researchers owing to huge advances in second-generation sequencing technologies. The cost of sequencing a microbial genome can now be covered by funding from an average grant, and microbiologists can even sequence the genome of their favourite organism themselves, with the arrival of new bench-top genome sequencers.

The ability to produce a reasonable assembly of a draft genome sequence in a matter of weeks has led to the increased need for rapid genome annotation. Automated annotations are increasingly performed owing to the time- and cost-saving advantage that they have over a manual annotation. Perhaps the quickest way of producing the most accurate annotation for new microbial genome sequences, and certainly the most accessible without sophisticated bioinformatic resources, is to transfer the annotation from a reference strain using tools such as the Rapid Annotation Transfer Tool (RATT; http://sourceforge.net/projects/ratt/). RATT is particularly useful for transferring annotation between different assemblies of a genome sequence so that it can be annotated and analysed concurrently with improvement to the assembly of the underlying reads.

No annotation transfers will be 100% accurate, and so they will still require manual curation, although RATT output files will highlight regions of difference with the reference genome to facilitate a faster manual annotation of problem regions. However, the use of such tools is limited to situations in which closely related genomes, annotated to a high standard, are available.

Where a *de novo* annotation is more appropriate, for example, in the genome annotation of a new species or genus, there are several automated annotation tools available; however, these tools vary greatly in their ability to predict and annotate genes. For example, Bakke *et al.*[1] recently compared the performance of three of the most popular automated annotation tools — Integrated Microbial Genomes (IMG; http://img.jgi.doe.gov/cgi-bin/pub/main.cgi), Rapid Annotation using Subsystems Technology (RAST; http://rast.nmpdr.org/) and the JCVI Annotation Service (http://www.jcvi.org/cms/research/projects/annotation-service/overview/) — in the genome annotation of *Halorhabdus utahensis*, a halophilic archaeon. The study highlighted numerous discrepancies between the IMG, RAST and JCVI annotations, including variations in start sites and predicted products, as well as differences in the number of unique genes (79, 39 and 254, respectively). The authors concluded that a manual comparison and curation of the output from multiple annotation services is required for the most accurate annotation.

Most genome sequences currently deposited in the public databases have been annotated using automated methods. However, accurate genome annotation — frequently updated to include new biological and genome knowledge, and to reflect advances such as the recent development of directional transcriptome sequencing, which was shown to aid gene prediction[2] — is important, not just for the understanding of the particular organism, but also for the accuracy and usefulness of the public databases. This was the rationale behind the recent complete re-annotation of the genome of uropathogenic *Escherichia coli* str. CFT073 (REF. 3). The re-annotation used improvements in annotation techniques and bioinformatics methods, together with new biological discoveries and data from the many *E. coli* genomes sequenced since the CFT073 genome was first published[4] (there are currently 37 complete *E. coli* genomes in the EMBL database). The re-annotation resulted in the removal of 608 previously predicted coding sequences (CDSs) and the addition of 299 CDSs, as well as the correction of the start sites of 341 CDSs. The update provided new insights into the virulence of this pathogen, including the identification of 19 new potential virulence genes.



Automated annotations can go some way to relieving the bottleneck in the ability to translate the vast quantities of sequencing data into meaningful results, and they may be sufficient for certain studies. However, there is still a clear need for manual curation and continual updating, at least of key reference genomes, to reduce the propagation of inadequate annotation, although the commitment and funding required for such endeavours, and perhaps the inability to translate these efforts into outputs that are measurable by funding bodies, makes this an unattractive proposition.

Nicola K. Petty is at the Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.
e-mail: microbes@sanger.ac.uk

doi:10.1038/nrmicro2462

1. Bakke, P. *et al.* Evaluation of three automated genome annotations for *Halorhabdus utahensis*. *PLoS ONE* **4**, e6291 (2009).
2. Croucher, N. J. *et al.* A simple method for directional transcriptome sequencing using Illumina technology. *Nucleic Acids Res.* **37**, e148 (2009).
3. Luo, C., Hu, G. Q. & Zhu, H. Genome reannotation of *Escherichia coli* CFT073 with new insights into virulence. *BMC Genomics* **10**, 552 (2009).
4. Welch, R. A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **99**, 17020–17024 (2002).

Competing interests statement
The author declares no competing financial interests.