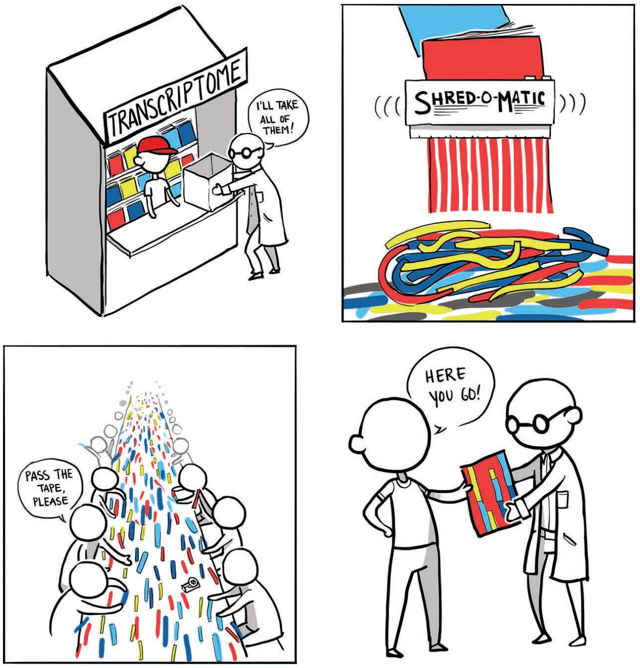
Genomics: the state of the art in RNA-seq analysis

Ian Korf

RNA-seq is a recent and immensely popular technology for cataloging and comparing gene expression. Two papers from the international RGASP consortium report on large-scale competitions to identify the best algorithms for RNA-seq analysis, with surprising variability in the results.

As a general rule, all cells of an organism contain the exact same DNA, and it is the expression of different RNAs (or different amounts of RNAs) that determines the identity of a cell. Important questions such as "Why are cancer cells different from normal cells?" can be answered by investigating all of the RNAs of a cell (its transcriptome) using a technique called RNA-seq. In this issue of *Nature Methods*, two papers from the RNA-seq Genome Annotation Assessment Project (RGASP) consortium evaluate the state of the art in RNA-seq analysis[1,2].

We often talk about the genome as the 'book of life' for an organism. To appreciate the gene expression problem, we need a slightly different analogy. Let's consider the transcriptome to be the 'newsstand of life' (**Fig. 1**). Some magazines (RNAs) have many copies on the shelf, whereas others are sold out. In the 'old days', newsagents would carry only the most popular magazines; each magazine was expensive, and so you only bought a few. Today (because of changes in sequencing methods and associated technologies), you buy the whole newsstand for one bargain price. The only catch is that the newsagent puts them all through a paper shredder first. Fortunately, in this imaginary world, there are armies of tape-wielding do-gooders (algorithm developers) eager to put your magazines back together and a statistics junkie (RGASP) who organizes a competition to find out who does the best job. The process of reassembling the magazines is called transcript reconstruction, and it is the main emphasis of the paper from Steijger *et al.*[1]. One of the underlying technologies in transcript reconstruction involves mapping RNA-seq reads back to the parent genome, which is the subject of the Engström *et al.* paper[2].



Abigail Yu

Figure 1 | Transcriptome reconstruction—akin to reassembling magazine articles after they have been through a paper shredder.

To be blunt, the results of the competitions are a little depressing. Despite the number of researchers working in this area, accuracy (as evaluated by several measures and with respect to a variety of sequence features) is nowhere near 100%. In the human genome, no method achieved even 60% accuracy for transcript reconstruction (averaging sensitivity and precision). Results were better in the nematode and fruit fly genomes, but these are much smaller and simpler genomes. Note that the three genomes in this study are some of the best-studied genomes to date. Many of the recently sequenced genomes have not had years of tinkering to improve their assemblies; as a result, the ability to map and reconstruct transcripts will be even less accurate in such genomes.

Conceptually, RNA-seq is a straightforward process: you isolate RNA, sequence it with a high-throughput sequencer, and put it all back together. What is the problem? There are several challenges. (i) The RNA may be from a different source than the reference genome it is compared with. This is certainly the case with the human reference genome, which is derived from several people. (ii) RNA preparations may contain incompletely processed RNAs or transcriptional noise. The deeper one sequences, the more frequently these rare events will appear, and this may explain why greater coverage depth often degrades

Ian Korf is at the UC Davis Genome Center, University of California, Davis, Davis, California, USA.
e-mail: ifkorf@ucdavis.edu

reconstruction accuracy. (iii) The sequencing technology may bias which sequences are present. Most sequencing procedures involve PCR, which is notorious for suppressing sequences with high GC content[3].

Sorting out which RNAs are signal and which are noise is a difficult problem and probably one of the major reasons why there is so much variability in the performance of the algorithms. The methods that contain a model of gene structure learned from prior data (Augustus[4], mGene[5] and Transomics (http://linux5.softberry.com/cgi-bin/berry/programs/Transomics/)) perform better than those that do not because they know what genes are supposed to look like. Algorithms may therefore improve with better models of genome structure. However, they will probably improve more with changes in technology. Lower error rate will improve alignment accuracy, and longer sequencing reads, such as those from Pacific Bioscience's sequencers, allow the reconstruction step to be skipped altogether.

Although RNA-seq analysis is difficult, you can be certain that it is improving on both the molecular and algorithmic fronts. In fact, these two articles are already somewhat out of date. Genomics and bioinformatics are changing so rapidly that by the time it takes to perform, write, review and revise a study, its obsolescence shortly after publication is a near certainty. This is especially true of studies that involve a consortium with large data sets. Highly critical readers may wonder what the point of these studies is when they do not unequivocally answer their question of "Which software should I use on my project tomorrow?" But these kinds of studies serve three critical roles: (i) they provide a historical record of where the field was at a particular time; (ii) they provide a template for current studies to test the latest methods; and (iii) they create a community whose intent is to move the entire field forward. This last point is the most important. If you have the opportunity to observe or participate in a project such as RGASP or the Assemblathon (http://assemblathon.org/)[6], you will be impressed by the creativity, honesty and generosity of the scientists involved. These projects may not be specifically funded and may involve data sets that are not perfectly aligned to a specific scientific question. There is a lot of "if only we had these data" chatter. The groups make the best of what they have, and not because they need to but because they want to. The immediate product does not always meet with the standards of the people on the inside or the outside. "Le mieux est l'ennemi du bien" in genomics, too.

COMPETING FINANCIAL INTERESTS

The author declares no competing financial interests.

1. Steijger, T. *et al. Nat. Methods* **10**, 1177–1184 (2013).
2. Engström, P.G. *et al. Nat. Methods* **10**, 1185–1191 (2013).
3. Nakamura, K. *et al. Nucleic Acids Res.* **39**, e90 (2011).
4. Stanke, M. *et al. Nucleic Acids Res.* **34**, W435–W439 (2006).
5. Schweikert, G. *et al. Genome Res.* **19**, 2133–2143 (2009).
6. Bradnam, K.R. *et al. GigaScience* **2**, 10 (2013).