AUTHOR FILE

Paul Bertone

Software can be like music, but an evaluation of software tools can involve some hard-to-play chords.

Originally, Paul Bertone set out to be a music composer. He switched to science in graduate school at Yale University. Writing music feels much like writing complex software, as if both tasks "use the same parts of your brain," he says. Bertone is a research group leader, bioinformatician and stem cell biologist at the European Bioinformatics Institute (EBI).

Paul Bertone

At Yale, he was in Mark Gerstein's bioinformatics lab and Michael Snyder's genomics lab. "That was amazing," he says. "I was a full-time member of both labs." He did experiments, explored analysis methods and was exposed to collaborative projects.

This experience primed him for one of his roles at the EBI, which he joined after completing his PhD. Together with his colleagues he ran an evaluation of software tools used to reconstruct RNA transcripts from high-throughput sequencing data: the RNA-sequencing (RNA-seq) Genome Annotation Assessment Project (RGASP). The project's goal is to keep tabs on how well RNA-seq analysis can be automated. Human and mouse genomes have been annotated by hand, but automated annotation of genomes is a challenge, Bertone says.

Scientists can choose from over two dozen open-source RNA-seq analysis tools that offer a range of approaches. Labs can test all the tools on their own, whereas community efforts can be cumbersome, even painful, with tempers flaring and personalities clashing. "You have to be very diplomatic, but the overriding thing is that you just know this is going to benefit everybody," Bertone says. He had to enforce rules such as prohibiting teams who submit data from also being involved in evaluating results. But the advantages of community-based evaluations are plentiful, he says, such as interaction with developers who help optimize tools.

RNA-seq data are not genomic data: introns have been removed. "When messenger RNA is spliced, you sometimes have an abbreviation, you could say, of the sequence," he says. The aligning software has to be aware of these abbreviations, detect where splits in the genome occur and place each read accurately. Transcript assembly from RNA-seq data remains a difficult problem. "The hope really is that we produce solid results that the community can trust."

RGASP was not a popularity contest; nor do the most popular programs do the best, he says. Overall, the analysis results on nematode and fruit fly data were "pretty good," but complex genomes with much more splicing yielded "surprisingly diverse" results.

Bertone hopes RGASP will help the community and his own work on stem cell pluripotency, too. Inaccurate RNA-seq analysis can have subtle effects; it might misidentify genes that are regulated in a given pathway. But effects can also be severe. "You could mischaracterize the state of a cell, for example," he says.

RNA-seq analysis challenges may start disappearing when it becomes possible to sequence whole transcripts at both high quality and high throughput, which would deliver unambiguous results in terms of exon linkage and read count, he says. Ideally, scientists want one read per transcript. "We're headed there," he says. "Third-generation instruments have the promise to do that, but we're not there yet."

John Rinn, a stem cell biologist at Harvard Medical School, worked with Bertone in the Snyder lab. "Paul inspired me early in my career to bridge experimental and computational science—and not to play video games with him," says Rinn. His favorite aspect of graduate school was going to the computer room "to jam out" and brainstorm with Bertone, with the added bonus that snacks were permitted there.

> "The hope really is that we produce solid results that the community can trust."

Bertone has always enjoyed jamming out to science. As the father of two young daughters, he says he would encourage them to choose a life of science, if they so desire. But he would not hide from them the demands of such a life, too.

He traces his bond with science to time spent with his cousin, now a nuclear astrophysicist. "We loved science together, and that was really important to us growing up," says Bertone. Although his day leaves too little time for music composition, he listens avidly. On any given day, baroque composers, fusion or world music flood his headphones.

Vivien Marx

Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **10**, 1177–1184 (2013).

Engström, P.G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10**, 1185–1191 (2013).