GENETICS

RNA studies under fire

High-profile results challenged over statistical analysis of sequence data.

BY ERIKA CHECK HAYDEN

High-throughput RNA sequencing has yielded some unexpected results in the past few years — including some that seem to rewrite conventional wisdom in genetics. But a few of those findings are now being challenged, as computational biologists warn of the statistical pitfalls that can lurk in data-intensive studies.

The latest case centres on imprinted genes. Humans and most other animals inherit two copies of most genes, one from each parent. But in some cases, only one copy is expressed; the other copy is silenced. In such cases, the gene is described as being imprinted. In July 2010, a team led by Catherine Dulac and Christopher Gregg, both then at Harvard University in Cambridge, Massachusetts, published a study[1] in *Science* estimating that 1,300 mouse genes — an order of magnitude more than previously known — were imprinted.

Now, researchers are arguing that a flawed analysis led Dulac and Gregg to vastly overestimate imprinting in their paper. "The reason this paper was published in *Science* is that they made this big claim that they saw an order-of-magnitude more genes that are imprinted, and I don't think that's true," says Tomas Babak, a computational biologist at Stanford University in California, who challenged the study in a paper[2] published on 29 March.

Dulac counters that she and her team "absolutely stand by those data", adding that they have confirmed some of their findings by other means. The situation resembles an ongoing debate over another RNA-sequencing paper[3] published in 2011. In that study, Vivian Cheung of the University of Pennsylvania in Philadelphia and her colleagues reported evidence that RNA editing — which creates differences between a gene's DNA sequence and the RNA sequence it gives rise to — is "widespread" in the human genome. RNA editing had been seen before, but the finding that it was so frequent challenges the central dogma, which holds that an organism's genes are transcribed faithfully.

Other scientists have argued that Cheung's results arose largely from errors in data analysis and that the true extent of RNA editing is probably no greater than previously thought[4]. Cheung did not respond to *Nature*'s request for comment on this story, but she has stood by her results.

For their study, Dulac and Gregg used high-throughput RNA sequencing to search mouse RNA for single nucleotide polymorphisms (SNPs) — one-letter variations in genetic sequence. The researchers then asked whether the SNPs they found for each gene could be traced to one or to both parents. If the SNPs were encoded mainly by one parent's copy of the gene, the team concluded that the gene was imprinted (see 'The silence of the genes').

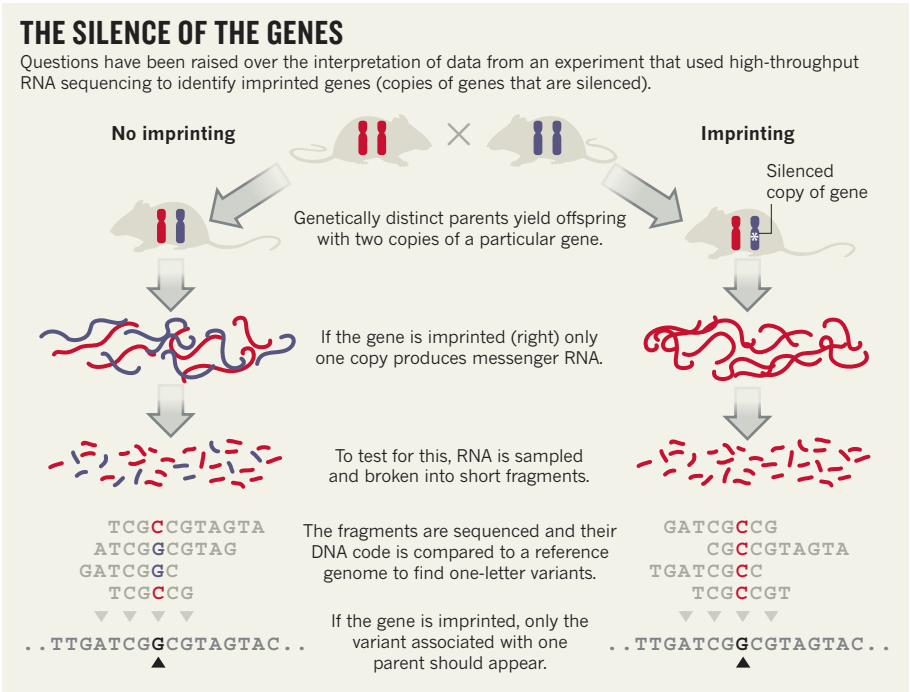
But Babak says that the statistical methods Dulac and Gregg used were not rigorous enough to rule out false positives. His team used multiple methods to estimate the false discovery rate — for instance, by applying stricter criteria for what could be considered instances of imprinting and by estimating how many spurious examples of imprinting would appear by chance if mice from identical genetic backgrounds were bred together. Babak's team then applied its false discovery rate to Dulac and Gregg's data and concluded that most of the instances of imprinting identified in the original paper were probably false positives. Dulac counters that Babak's analysis may be filtering out legitimate but complex instances of imprinting.

"What's happened in the first few papers on these problems is that the statistics and analysis in general have not been done very carefully," says Lior Pachter, a computational biologist at the University of California, Berkeley. "And that means you may get completely wrong answers." Researchers have had many years to develop standard methods to minimize errors and biases in DNA sequencing, but such methods are still being developed for high-throughput RNA sequencing.

Pachter says that another key problem is that high-profile papers in the field may be well reviewed for their biology but not their computational foundations. "The culture is not the same in biology as it is in statistics or math, where reviewers sit with a paper for months, check the statistics and the math, and run the programs and test them," he says.

The debate has implications for any sequencing-based study that requires statisticians to identify rare genetic phenomena using relatively new methods. "If you don't deal with the analytical details very carefully, you're going to get into trouble because of the low signal-to-noise ratio" in these types of experiments, says Jin Billy Li, a genomicist at Stanford University who was one of the critics of Cheung's RNA-editing paper.

Dulac says that she and her colleagues are now using different statistical methods to reanalyse the imprinting data, but adds, "I am quite confident that we will find things that are likely to be around the same order of magnitude" as originally reported. ■



1. Gregg, C. *et al. Science* **329,** 643–648 (2010).
2. DeVeale, B., van der Kooy, D. & Babak, T. *PLoS Genet.* **8,** e1002600 (2012).
3. Li, M. *et al. Science* **33,** 53–58 (2011).
4. Check Hayden, E. *Nature* http://10.1038/nature.2012.10217 (2012).