GENOMICS

Searching for mismatches in a vast genomic landscape

Raw data of millions of sequences used to assemble the reference genomes of ten organisms are analyzed in search of mismatches indicative of editing events. Findings include candidate sites for *in vivo* DNA and RNA editing, and a common sequencing error.

It is generally thought that an organism contains identical genomic information in all of its cells and that this genome remains unaltered throughout the organism's life, with the exception of rare and random somatic mutations that might occur. However, certain enzymes present in the cell can change or 'edit' particular sequences in the DNA as well as in the RNA. The occurrence of RNA editing has been well documented in several model organisms as well as in humans, and it is thought to be important in the regulation of gene expression. DNA editing, however, although reported in viruses and mice retrotransposons, has yet to be seen in the human genome.

In a recent study, co–first authors Alexander Wait Zaranek and Erez Levanon from George Church's laboratory at Harvard University set out to look for evidence of editing in the vast, publicly available databases of genomic information. They turned to the National Center for Biotechnology Information Trace Archive, a repository of 'raw' sequencing data obtained from traditional capillary electrophoresis sequencing methods. The two billion sequences contained in this archive serve primarily for the assembly of consensus reference genomes of multiple organisms. But this collection of data can also be a source of answers to biological questions, these researchers thought—if, of course, one can make sense of it.

Church and collaborators took 600 million sequences from this archive from ten different organisms and aligned them to their corresponding reference genomes. Then they looked for mismatches, locations with single–base-pair differences compared to the reference. They first focused on hallmarks of editing by APOBEC3, a member of a family of cytidine deaminases that produces characteristic changes in the DNA and RNA. "This task turned out to be more difficult than initially thought," says Levanon, now at Bar-Ilan University, Israel, because the majority of these mismatches were the product of a systematic sequencing error. This error has likely been incorporated into some of the resources and reference genomes that are used by the scientific community such as the HapMap database for common human genomic variations.

Once the researchers identified the mismatch motifs caused by this sequencing error, they eliminated the errors from the dataset using stringent quality thresholds. The researchers were then confident that the remaining thousands of mismatches represented genuine editing events. Among these, they found the first evidence of candidate DNA editing sites in the human genome.

In this study, they also discovered thousands of new RNA editing sites in both human and mouse genomes as well as extensive RNA editing in the frog *Xenopus tropicalis*. "These findings are probably an underrepresentation of the true number of editing events occurring in these genomes," says Zaranek. Applying this approach to survey other genomes will help determine the magnitude of DNA and RNA editing in different organisms. Although the importance and scope of these phenomena needs to be explored in future work, this study shows the utility of using raw genomic datasets for the discovery of candidate editing events.

The next logical step will be to adapt the technique to data obtained from next-generation sequencing methods. Extracting meaningful information from the shorter reads that these technologies generate will undoubtedly be a challenge, but it will open up an even greater landscape of genomic data to dive into.

Erika Pastrana

RESEARCH PAPERS

Zaranek, A.W. *et al.* A survey of genomic traces reveals a common sequencing error, RNA editing, and DNA editing. *PLoS Genet.* **6**, e1000954 (2010).