

Straws in a haystack

With new ways to examine the effects of mutations on gene expression within the 3D genome and increased emphasis on finding these variants by sequencing whole genomes, we would really like to know more about the rules that govern noncoding and regulatory sequences.

Whole-genome sequencing captures a larger proportion of the variation in the readily interpretable coding regions than whole-exome sequencing, and its price is becoming competitive. As thousands of whole genomes become available, the noncoding regulatory and functional structural elements of the genome are emerging from the background because they are relatively constrained in their variation within the human population (*Nat. Genet.* **50**, 333–337, 2018) or conserved in comparison to other animals. In this issue, Michael Talkowski and colleagues ([doi:10.1038/s41588-018-0107-y](https://doi.org/10.1038/s41588-018-0107-y)) show that it is possible to design a case-control association study to identify de novo mutations in a defined catalog of noncoding elements and genome features. The drawback, emphasized in the accompanying News & Views by Naomi Wray and Jacob Gratten ([doi:10.1038/s41588-018-0113-0](https://doi.org/10.1038/s41588-018-0113-0)), is that tens of thousands of families will need to be recruited to gain statistical support for associations to variants outside protein-coding exons. In some respects, this study deals with a set of relatively tractable early-onset traits (the autism spectrum disorders) where family quartets are

available for analysis and about 15% of cases are already accounted for by newly acquired rare coding mutations in a restricted set of genes. Other complex diseases will be, well, more complex. Consequently, we need to be careful to continue to correct for multiple hypotheses when associating function with noncoding sequence variation in whole genomes.

For heritable structural variants in human and mouse, it is now possible not only to identify their tissue-specific perturbations of the structure of the entire genome in the nucleus experimentally via proximity ligation methods such as that developed by Gang Cao and colleagues ([doi:10.1038/s41588-018-0111-2](https://doi.org/10.1038/s41588-018-0111-2)) but also to study in silico the topological consequences of chromosomal rearrangements. Stefan Mundlos and colleagues ([doi:10.1038/s41588-018-0098-8](https://doi.org/10.1038/s41588-018-0098-8)) use a polymer model to predict how such rewiring causes genes to become misexpressed (see the News & Views by Ralph Stadhouers; [doi:10.1038/s41588-018-0112-1](https://doi.org/10.1038/s41588-018-0112-1)).

What types of regulatory perturbation might we expect mutations to cause by disrupting chromatin loops or enhancer-gene interactions? Perhaps some variants

produce graded gene expression changes like a rheostat, whereas others act in an all-or-nothing manner. In this respect, genetic variants might be interpreted by comparing their effects to the graded response to pharmacological disruption of the protein contacts of a single enhancer-promoter pair. At the other extreme, we might sometimes see large stepwise deregulation by disturbance of the multivalent protein contacts of enhancer clusters (for example, *Cell* **169**, 13–23, 2017).

On our wish list, of course, is high-throughput prediction of the effect of common and recurrent structural variants on the long-distance interactions within the genome. This could lead to a database that intersects genome-wide sequence constraint data, gene annotations and allelic topological configurations in relevant cell types. Until then, rare noncoding variants will need to be collected in well-powered meta-analyses of tens or hundreds of thousands of cases and controls, all with whole-genome sequencing and detailed phenotypes. □

Published online: 26 April 2018
<https://doi.org/10.1038/s41588-018-0125-9>