Nature Reviews Genetics | AOP, published online 28 February 2012; doi:10.1038/nrg3206

GENOMICS

How pervasive are defective genes?

Substantial variation in protein-coding genes among human populations is being revealed through troves of genome sequence data. However, a new study presents a more thorough assessment of the number of variants in the human genome and of which of these variants detrimentally affects genes. The authors indicate that although the frequency of loss-of-function mutations may have been overestimated, they are still pervasive.

MacArthur *et al.* used the whole-genome sequences of 185 humans from the 1000 Genomes project, which collectively have been claimed to contain 2,951 putative loss-of-function sequence variants. Apparent loss-of-function variants can result from many factors — including sequencing artefacts, erroneous annotation of sequence reads or inappropriate functional interpretation — so the authors carried out rigorous quality-control procedures to filter these variants into a 'high-confidence' list. First, they filtered the variants informatically to remove those variants for which the sequence context suggested no major effect on gene function (for example, a location at the 3′ end of the open reading frame). Next, they carried out independent sequencing approaches and re-annotation to filter the variants further, retaining only those that passed this technical validation. Overall, 1,285 (43.5%) of variants survived filtering, suggesting that sequencing projects have overestimated the prevalence of loss-of-function variants by more than twofold. However, this still implies that there are ~80 heterozygous and, importantly, ~20 homozygous loss-of-function variants in a typical healthy individual. The numbers could be higher, as the authors acknowledge that additional loss-of-function variants, such as rare variants or large-scale rearrangements, may have been missed by the initial sequencing projects.

Further analyses indicated how these loss-of-function mutations might be tolerated in healthy individuals. Genes that were affected by loss-of-function variants were more likely to be a part of a gene family (suggesting buffering through redundancy) and had lower connectivity in gene and protein networks (implying a peripheral role in cellular processes). Also, variants in known disease-causal genes were almost exclusively heterozygous.

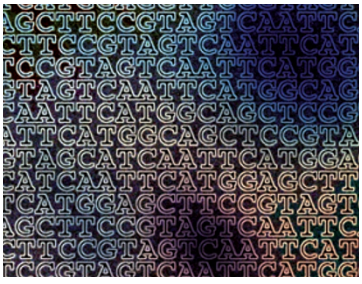
A thorough experimental characterization of the effects of these variants at the gene and organismal levels is difficult to expand to a genome scale. Instead, the authors looked for effects of the loss-of-function variants on mRNA expression on the basis that truncated proteins can induce nonsense-mediated decay (NMD) of their transcripts. Seven of the twenty-eight tested variants that were predicted to trigger NMD actually had lower-than-wild-type transcript expression; whether the other variants affect gene function in alternative ways remains unclear.

How will these findings inform clinical sequencing studies for finding disease genes? Knowing the background rate of loss-of-function variants and the accuracy of sequencing project data is crucial for predicting the importance of reported novel variants. In addition, MacArthur *et al.* used the properties of homozygously tolerated variants compared to known disease mutations to formulate a bioinformatic tool for predicting the severity of novel variants.

This study highlights the need for quality to keep pace with quantity in disease sequencing projects.

Darren J. Burgess

ORIGINAL RESEARCH PAPER MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012)



NPG/GETTY