

GENETIC VARIATION

# ExAC boosts clinical variant interpretation in rare diseases

The largest catalogue of protein-coding genetic variation to date is reported by the Exome Aggregation Consortium (ExAC). This project includes aggregation, harmonization and joint analysis of exome sequence data for 60,706 individuals from more than 20 research studies. ExAC's openly accessible genetic variation database has already proved to be a crucial resource for research as well as clinical studies, in particular for rare disorders.

Analysis of the ExAC data set brings our first views into the landscape of very rare protein-coding genetic variation. Lek *et al.* identify more than 7.4 million high-confidence genetic variants, on average 1 every 8 bases, the majority of which are novel (72% are not found in any existing database) and extremely rare (99% with a frequency of <1%; 54% are seen only once in ExAC). From the data set, they are able to document recurrent rare mutations emerging independently, providing an estimate of the frequency of recurrence, which has never been observed systematically before owing to the need for such large sample sizes. Furthermore, the authors examine the level of selective constraint against protein-truncating variation, identifying 3,230 genes that appear highly loss-of-function-intolerant. Reassuringly, this includes most known human haploinsufficient disease genes; however, 72% do not yet have an established human disease phenotype. Although some of these genes may be associated with weaker phenotypes or embryonic lethality, this points to how much more we have yet to understand about the phenotypic consequences of loss of function in human genes.

In coordinated work, Ruderfer *et al.* analyse the rates and properties of rare copy number variation (CNV) within the ExAC data set. They find that ~70% of individuals

carry at least one rare genic CNV, with an average of 0.81 deleted genes and 1.75 duplicated genes per individual. The authors also estimate relative intolerance to CNVs for each gene and show that this statistic is correlated with single-nucleotide variation (SNV) and evolutionary measures of genic constraint.

The authors also demonstrate the use of ExAC to improve variant interpretation in rare diseases. Lek *et al.* find that ExAC participants have on average ~54 genetic variants previously classified as causal for a disease, and suggest that most may be attributable to misclassified variants. Lek *et al.* also review the evidence for pathogenicity of 192 previously reported pathogenic variants for rare Mendelian disorders. Only nine of these variants had sufficient support for disease association, with a high proportion of these variants present at an implausibly high frequency in the ExAC data set, suggesting that many have been incorrectly classified as pathogenic.

“ value of large reference panels such as ExAC for filtering variants ”

In two additional publications from the ExAC project, the authors analyse large patient case series in efforts to move towards resolution of prior disease associations. Walsh *et al.* systematically re-examine evidence for genes implicated in inherited cardiomyopathies, which are collectively one of the most common and severe rare disorders. For this, the authors analyse sequence data for selected cardiac genes from 7,855 individuals with a clinical

diagnosis of cardiomyopathy. Although they validate some cardiomyopathy genes, they also find a sizeable fraction of purported cardiomyopathy genes and variants that do not show support for pathogenicity, including some that are included in gene panels used clinically. Similarly, Minikel *et al.* collected data from 16,025 individuals with confirmed prion disease, the largest case series ever available for prion disease, for which ~10–15% of cases are estimated to be caused by mutations in the prion protein (*PRNP*) gene. They find numerous variants in *PRNP* that are thought to be pathogenic and highly penetrant, but actually appear to be likely benign.

These findings highlight the necessity to carefully evaluate the literature for rare genetic disorders and reinforce the value of large reference panels such as ExAC for filtering variants seen in patient sequence data. The ExAC project continues to expand in size, hoping to increase to more than 120,000 exome sequences over the next year, as well as 20,000 whole-genome sequences, bringing additional sample size, diversity and exploration of non-coding regions that will further aid these efforts.

Orli G. Bahcall, Senior Editor, Nature

**ORIGINAL ARTICLES** Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016) | Ruderfer, D. M. *et al.* Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat. Genet.* <http://dx.doi.org/10.1038/ng.3638> (2016) | Walsh, R. *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet. Med.* <http://dx.doi.org/10.1038/GIM.2016.90> (2016) | Minikel, E. V. *et al.* Quantifying prion disease penetrance using large population control cohorts. *Sci. Transl. Med.* **8**, 322ra9 (2016)

**FURTHER READING** MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014)

**WEB SITE**  
ExAC database: <http://exac.broadinstitute.org>