Addendum: The mutational constraint spectrum quantified from variation in 141,456 humans

https://doi.org/10.1038/s41586-021-03758-y

Published online: 9 August 2021

Addendum to: *Nature* https://doi.org/10.1038/s41586-020-2308-7

Published online 27 May 2020

Open access

Check for updates

Sanna Gudmundsson, Konrad J. Karczewski, Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, Kristen M. Laricchia, Andrea Ganna, Daniel P. Birnbaum, Laura D. Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A. Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M. England, Eleanor G. Seaby, Jack A. Kosmicki, Raymond K. Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X. Chong, Kaitlin E. Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H. O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S. Ware, Christopher Vittal, Irina M. Armean, Louis Bergelson, Kristian Cibulskis, Kristen M. Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E. Talkowski, Genome Aggregation Database Consortium*, Benjamin M. Neale, Mark J. Daly & Daniel G. MacArthur

This analysis explores the extent of loss-of-function (LoF) tolerance in human disease genes.

Databases of human population genetic variation, such as the Genome Aggregation Database (gnomAD), are generally expected to be depleted for variation with severe effects on health. As such, it is expected that genes that carry highly disruptive changes, predicted (p)LoF variants, in these databases are less likely to be responsible for severe human disease. However, the precise relationship between pLoF tolerance and human disease causation is not well-characterized.

In our Article, we reported a total of 2,636 variants in 1,815 genes that were homozygous in at least one individual and annotated as pLoF after applying both automated filtering and manual curation of both sequencing quality and functional annotation. We labelled these genes as 'LoF-tolerant', indicating that total functional loss of these genes appears to be compatible with life. This does not exclude the involvement of these genes in diseases compatible with presence in individuals in gnomAD[1]. Neither the 'LoF Transcript Effect Estimator' (LOFTEE) nor manual curation took previous gene–phenotype associations into account, as this would create a bias that affects downstream analyses and also may result in the spurious exclusion of true LoF-tolerant genes owing to previous false-positive reported associations with disease. This unbiased approach is appropriate for permitting downstream analyses, but it means that the enrichment of pLoF artefacts will remain higher in genes for which genetic disruption is genuinely associated with severe disease.

Prompted by comments on our original Article, we explored the degree to which our LoF-tolerant list includes genes associated with disease by manually curating the 158 genes (with 217 pLoF variants) on the LoF-tolerant list associated with autosomal recessive and X-linked traits in 'Online Mendelian Inheritance in Man' (OMIM) by an additional biocurator[1].

Of these genes, 71% ($n = 112$) are associated with phenotypes that are likely to be found in gnomAD, on the basis of gnomAD inclusion criteria. These are phenotypes such as infertility, hearing or visual impairment, benign or mild metabolic or haematological phenotypes, expected at similar frequency as the general population (95 phenotypes) and, to a lesser extent, traits that are likely to be depleted from gnomAD, but for which someone with the condition may participate in a common disease study (17 phenotypes). We observed an overrepresentation of traits that are likely to be found (60% versus 33%) and an underrepresentation of traits that are not expected to be found (29% versus 53%) in gnomAD (early-onset severe or lethal rare disease that generally would restrict participation in genetic studies) versus a control set of 100 random selected autosomal recessive and X-linked OMIM traits ($P = 3.0 \times 10^{-5}$, Fisher's exact test) (Fig. 1a). We performed a thorough literature review of the 46 phenotypes that were initially not expected to be found in gnomAD, which revealed that 35% (16 out of 46) can be explained by evidence of mechanism of disease not being LoF ($n = 2$), variable expressivity ($n = 5$) or penetrance ($n = 3$), phenotype being responsive to treatment ($n = 4$) and onset after age of the individual in gnomAD ($n = 2$) (Fig. 1b, blue).

In contrast to what is expected to be found in gnomAD, 32 pLoF variants are in 30 genes for which homozygous LoF has been associated with severe or lethal phenotypes in OMIM. However, 10 of these 30 genes had a limited number of cases reported ($n = 7$) or no reported biallelic LoF variants in humans ($n = 3$) (Fig. 1b, light red) and only 5 genes meet current ClinGen standards for a known LoF mechanism[2]. We evaluated the 32 variants by applying more stringent criteria, and identified several cases in which a variety of mechanisms may result in an evasion of true loss of gene function. For 15 variants, we found evidence that disputed our previous prediction (Fig. 1c, purple), including variants that are suspected to escape nonsense-mediated decay but that did not meet the criteria for rescue applied in our original Article ($n = 12$), one variant that was within a small homopolymer and thus is more likely to represent a sequencing error, one alignment error, and one variant that is in an overprinted transcript and is more probably a synonymous variant in the most biologically relevant transcript. For the 17 variants for which we cannot identify conclusive ($n = 9$) (Fig. 1c, pink) or any ($n = 8$) (Fig. 1c, grey) evidence for evasion of pLoF, there are several explanations that even our stringent curation cannot confidently exclude: for example, sample swaps, a variety of residual sequencing and annotation artefact classes, the presence of an individual in gnomAD who does actually have the expected phenotype, or simply variable expressivity, late age of onset or reduced penetrance of the disease phenotype itself. Further details regarding variant curation are are available in Supplementary Table 1 and from https://gnomad.broadinstitute.org/downloads, or the curation data can be viewed at the respective gene page at https://gnomad.broadinstitute.org.

In summary, this result emphasizes the well-established need for extremely careful curation of any pLoF variant observed in a population database such as gnomAD, especially for genes for which such variants are expected to be deleterious. The variants curated here are found at low frequency and are enriched for both sequencing and annotation errors[3,4]. This enrichment is expected to be even larger in genes for

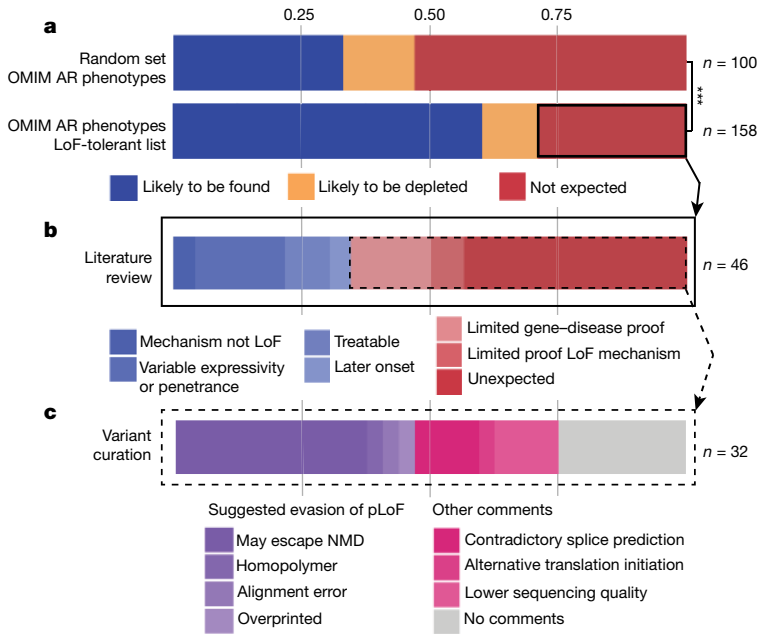

Fig. 1 | Assessment of pLoF variants in LoF-tolerant genes associated with autosomal recessive and X-linked phenotypes in OMIM. **a**, Autosomal recessive and X-linked (AR) OMIM phenotypes: likely to be found (blue), likely to be depleted (yellow) or not expected (red) to be found in gnomAD, for the 158 phenotypes associated with LoF-tolerant genes in gnomAD and a set of 100 randomly selected AR and X-linked OMIM traits. ***$P = 3.0 \times 10^{-5}$, Fisher's exact test. **b**, Extended literature review of the 46 out of 158 OMIM phenotypes not expected to be found in gnomAD. **c**, Extended variant curation of 32 pLoF variants in 30 LoF-tolerant genes beyond criteria presented in our original Article revealed pLoF with suggested evasion of pLoF (purple), and pLoF with no conclusive (pink) or no evidence (grey) contradicting pLoF in these genes. NMD, nonsense-mediated decay. Further details are provided in Supplementary Table 1.

which inactivation is associated with severe disease, because sequencing and annotation artefacts are distributed approximately uniformly across the genome, whereas true LoF variation is depleted in genes in which it results in a more detrimental effect. Although the pLoF variants found in the gnomAD dataset have been subjected to thorough quality control, any filtration process other than comprehensive experimental validation is insufficient to remove all artefacts.

In conclusion, population databases such as gnomAD are a powerful source of information when predicting human tolerance towards gene disruption. The list of LoF-tolerant genes identified in gnomAD is a useful class for downstream analysis that appears to largely comprise genes for which true homozygous disruption does not cause severe early-onset disease.

Authors S.G. and M.S.-B. carried out the analysis described in this Addendum. K.J.K., A.O.-L. and D.G.M. contributed to the experimental design, and A.O.-L. and D.G.M. supervised the work. S.G., M.S.-B., K.J.K., A.O.-L. and D.G.M. wrote the Addendum. A.O.-L. and D.G.M. contributed equally to this work.

We thank C. Arnoult, P. Ray and N. Thierry-Mieg for presenting the opportunity to further clarify the term LoF tolerance.

Supplementary Information is available in the online version of this Amendment.

1. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
2. Abou Tayoun, A. N. et al. Recommendations for interpreting the loss of function PVS1 ACMG/AMP variant criterion. *Hum. Mutat.* **39**, 1517–1524 (2018).
3. MacArthur, D. G. & Tyler-Smith, C. Loss-of-function variants in the genomes of healthy humans. *Hum. Mol. Genet.* **19**, R125–R130 (2010).
4. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).