**Response to Ramos, *et al*.**

Eleanor G. Seaby1,2,3,4, Diana Baralle1, Heidi L. Rehm2,3, Anne O’Donnell-Luria2,4\*, Sarah Ennis1\*

1. Faculty of Medicine, University of Southampton, Southampton, UK
2. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
3. Center for Genomic Medicine, Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
4. Division of Genetics and Genomics, Boston Children’s Hospital, Boston, MA, USA

\*Authors contributed equally

Corresponding author:

Dr Eleanor G. Seaby

Duthie Building, MP 808, University Hospital Southampton, Tremona Road, Southampton, SO16 6YD, UK.

E.Seaby@soton.ac.uk

**Ethics approval and consent to participate**All patients included in this study consented to participate in the 100,000 Genomes Project - ethics approval REC: 14/EE/1112; IRAS: 166046.

**Competing interests**
No competing interests or conflicts to declare.

**Availability of data and materials**
The anonymised phenotype and genotype data that support the findings of this study are only available as a registered GeCIP member in the Genomics England Research Environment, but restrictions apply to the availability of these data due to data protection and are not publicly available. Information regarding how to apply for data access is available at the following url: <https://www.genomicsengland.co.uk/about-gecip/for-gecip-members/data-and-data-access/>. No individual’s variant and phenotype data can be shared or published due to data sharing restrictions.

We thank Ramos *et al*. for their correspondence1 and interest in our recent article that applied constraint metrics to uplift novel disease gene discovery.2 The authors provide valuable insight by applying an approach similar to ours to identify novel genes using single sample (as opposed to trio) exome sequencing in Brazil and in narrowing the gap of health inequality for genetic testing.

Between 2017 and 2021, Ramos *et al*. searched for predicted loss-of-function (pLoF) variants in novel genes (unassociated with disease at the time) and absent from gnomAD3 with a gene pLI score4 >0.99 in a cohort of Brazilian patients with neurodevelopmental phenotypes, all of whom had proband-only exome sequencing performed. After manual curation, which included review of exon expression in the central nervous system, the authors identified 55 candidate variants. Following Sanger sequencing or parental exome sequencing, 25/55 (40%) of the variants they identified were *de novo* or assumed *de novo*. 15/25 (60%) of these *de novo* candidates have now been published as novel gene discoveries, which is in keeping with our observations that genes constrained for loss-of-function are predicted to be disease-causing.2

Ramos *et al*. were surprised by the high proportion of inherited pLoF variants in their cohort. They postulated that the 30 inherited variants (60% of the total they identified) could be due to underrepresentation of their cohort’s diverse ancestries in gnomAD. When replicating their filtering criteria on 13,494 trios in the 100,000 Genomes Project, we identified 3,359 pLoF variants, absent from all gnomAD populations, in novel genes (absent from OMIM) with a pLI score >0.99. To further enrich for high quality variants, we removed 1199 variants that failed QC and/or were not on the Matched Annotation from NCBI and EMBL-EBI (MANE) transcript5, leaving a total of 2160 pLoF variants, of which 131/2160 (6%) were *de novo*. We observe a much smaller proportion of *de novo* variants when using Ramos*’* criteria, compared with their observations; 6% versus 40% respectively. This suggests that ancestral differences do not fully explain the high number of inherited pLoF variants in constrained genes, which is perhaps in keeping with 10% of gnomAD v2.1.1 representing Latino/Admixed American populations.3 Furthermore, we would expect that a less well represented population in gnomAD would have a lower *de novo* rate. While we are unclear of the cohort size used by Ramos *et al.*, we expect that the Brazilian cohort is considerably smaller, which may explain some of the disparity in results. Additionally, the phenotypic composition (and percentage of the cohort with a *de novo* pathogenic variant) may differ between cohorts.

The study by Ramos *et al*. is important as it raises an excellent point regarding genetic diagnostic inequity across the globe. In the UK, genome sequencing is available as a test on the National Health Service, whereas in Brazil, both exome and genome sequencing are not publicly funded. Despite the lower diagnosis rate compared to trio exome sequencing6, Ramos *et al.* explain that proband-only exome sequencing is covered by some Brazilian health insurance companies, but only in select circumstances and segregation analysis is at the discretion of reporting laboratories. There is clear demand to narrow global inequality of diagnostic testing and to not discriminate against families where parental data are unavailable nor covered by public or private healthcare.

Whilst we agree that trio analysis is economically impracticable in low-middle income countries, we also raise concern about interpretability of pLoF variants in constrained genes without segregation data available, particularly if up to 94% of variants are inherited from unaffected individuals. Ramos *et al*. manually curated their variants to exclude false positive candidates, including removing variants that escaped nonsense mediated decay and that were on transcripts poorly expressed in neurodevelopmental tissues. Manual curation of variants is a worthwhile but time-intensive process3,7, and while tools such as LOFTEE3 and pext8 can automate aspects of pLoF curation, there is no single automated tool that can reliably identify commonly-observed false positive pLoF variants. If trio data were unavailable in the 100,000 Genomes Project, this would involve curation of >2100 variants, of which many would pass curation yet still be present in an unaffected parent.3 That said, it is unlikely a clinical scientist would be curating all variants at once, therefore if resources allow, this remains a viable and important step in the analysis of pLoF variants observed in non-trio families. Additionally, to confirm candidates, Ramos *et al*. undertook Sanger or parental exome sequencing. While for the Brazilian group, this additional testing yielded a *de novo* or assumed *de novo* variant 40% of the time, for the 100,000 Genomes Project this would be less cost-effective, returning a putative *de novo* variant ~6% of the time.

Ramos *at al*. demonstrate the importance of global equity of access to genomic testing. They show that novel gene discovery is possible in the absence of trio analysis, which is an essential finding for countries where trio exome or genome sequencing are not routinely available or where parental DNA are unavailable. Additionally, some diseases present later in life, whereby parental testing would not be as informative. However, querying a large amount of data from the 100,000 Genomes Project showed that only ~6% of pLoF variants meeting criteria used to identify candidate variants in novel genes were in fact *de novo*. Therefore, we urge caution in the interpretation of novel (i.e. absent from gnomAD) pLoF variants in genes constrained for loss-of-function e.g. with a pLI >0.99 or LOEUF <0.2. We recommend that, where possible, trio data remains a powerful tool to increase confidence in pLoF variant pathogenicity but that in circumstances where segregation is unavailable, more caution in the interpretation of pLoF variants is recommended.

References:

1. Ramos LLP, Kitajima JP, de Athayde Costa LS, Monteiro FP, Kok F. Correspondence on “A gene-to-patient approach uplifts novel disease gene discovery and identifies 18 putative novel disease genes” by Seaby et al. Genetics in Medicine, in press*.* 2022.

2. Seaby EG, Smedley D, Taylor Tavares AL, et al. A gene-to-patient approach uplifts novel disease gene discovery and identifies 18 putative novel disease genes. Genetics in Medicine*.* 2022;24(8):1697-1707.

3. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature*.* 2020;581(7809):434-443.

4. Kosmicki JA, Samocha KE, Howrigan DP, et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. Nature genetics*.* 2017;49(4):504.

5. Morales J, Pujar S, Loveland JE, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature*.* 2022:1-6.

6. Farwell KD, Shahmirzadi L, El-Khechen D, et al. Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. Genet Med*.* 2015;17(7):578-586.

7. Gudmundsson S, Karczewski KJ, Francioli LC, et al. Addendum: The mutational constraint spectrum quantified from variation in 141,456 humans. Nature*.* 2021;597(7874):E3-E4.

8. Cummings BB, Karczewski KJ, Kosmicki JA, et al. Transcript expression-aware annotation improves rare variant interpretation. Nature*.* 2020;581(7809):452-458.