NAR Genomics and Bioinformatics, 2024, **6**, lqae111
https://doi.org/10.1093/nargab/lqae111
Advance access publication date: 29 August 2024
Application Notes

Context-adjusted proportion of singletons (CAPS): a novel metric for assessing negative selection in the human genome

Mikhail Gudkov [1,2,†], Loïc Thibaut [3,4,†] and Eleni Giannoulatou [1,2,*]

[1]Victor Chang Cardiac Research Institute, NSW 2010, Australia
[2]School of Clinical Medicine, St Vincent's Healthcare Clinical Campus, Faculty of Medicine and Health, UNSW Sydney, Australia
[3]Centre for Population Genomics, Garvan Institute of Medical Research, NSW 2010, Australia
[4]School of Mathematics and Statistics, UNSW Sydney, NSW 2052, Australia

[*]To whom correspondence should be addressed. Tel: +61 2 92958669; Email: E.Giannoulatou@victorchang.edu.au
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Interpretation of genetic variants remains challenging, partly due to the lack of well-established ways of determining the potential pathogenicity of genetic variation, especially for understudied classes of variants. Addressing this, population genetics methods offer a practical solution by evaluating variant effects through human population distributions. Negative selection influences the ratio of singleton variants and can serve as a proxy for deleteriousness, as exemplified by the Mutability-Adjusted Proportion of Singletons (MAPS) metric. However, MAPS is sensitive to the calibration of the singletons-by-mutability linear model, which results in biased estimates for certain variant classes. Building up on the methodology used in MAPS, we introduce the Context-Adjusted Proportion of Singletons (CAPS) metric for assessing negative selection in the human genome. CAPS produces corrected estimates with more accurate confidence intervals by eliminating the mutability layer in the model. Retaining the advantageous features of MAPS, CAPS emerges as a robust and reliable tool. We believe that CAPS has the potential to enhance the identification of new disease-variant associations in clinical and research settings, offering improved accuracy in assessing negative selection for diverse SNV classes.

Introduction

Most disease-variant association studies these days are hindered by the lack of well-established ways of variant prioritization, which remains one of the key challenges in modern genomic analyses. The problem of variant prioritization can be reduced to comparing variants based on their potential deleteriousness. However, for many classes of genetic variation their overall deleteriousness remains poorly quantified.

With the ever-increasing size of open databases of human genetic variation, population genetics methods have the potential to provide the means of variant prioritization that are based on patterns of genetic variation that occur naturally in human populations. One of the key considerations when analyzing variants through the population genetics approach is the variability in mutation rates, which is known to affect variant analysis (1). Indeed, as some variants have higher mutability—or susceptibility to change—than others, variants with high mutability are less likely to be rare and can reach saturation in large genomic databases. This creates a need for the use of the finite sites model and mutability correction (2).

Transversions, CpG transitions and non-CpG transitions constitute the three main types of single-nucleotide variants. Transversions (for example, AAG to ATG) have much lower mutation rates than CpG transitions, because the mutations that they create are more complex biochemically. It has been shown that trinucleotide sequence context is sufficient to account for a large proportion of variability in mutation rates and that the baseline mutability level can be obtained from the mutation rates of trinucleotide sequences in noncoding regions (2,3). Using this approach, each variant falls into one of 104 possible groups, as there are 32 trinucleotide contexts with each having three possible mutations of the middle nucleotide, with the additional 8 groups coming from the 4 CpG contexts' medium and high levels of methylation.

A particular example of a population genetics method that utilizes this approach is the Mutability-Adjusted Proportion of Singletons (MAPS) metric (4,5), which can be used as a tool for estimating negative selection and deleteriousness. The general assumption in MAPS is that if a particular variant is damaging, it will be rare in the population, because purifying selection will be trying to remove it. MAPS scores are calculated as the scaled excess or deficit of singletons (variants with allele count of 1), where the expected number is derived based on context sequence from a singletons-by-mutability linear model calibrated on a relatively neutral class of variants, using data from the genome aggregation database (gnomAD).

Synonymous variants have some unique properties in gnomAD, making them a good choice for the calibration of MAPS (4,5). In particular, these variants have good coverage in both exomes and genomes and, despite having low overall

Received: February 8, 2024. **Revised:** July 24, 2024. **Editorial Decision:** August 6, 2024. **Accepted:** August 14, 2024

selective pressure, exhibit a background selection component (6). Importantly, however, in synonymous variants only 100 out of the total of 104 context groups can be observed, due to the specifics of how the genetic code works. In MAPS, these 4 missing contexts are ignored, as the singletons-by-mutability model can only be fitted on the 100 contexts that are present in synonymous variants.

Here, we show that the mutability layer in MAPS makes its estimates biased, with the severity of the bias varying depending on the variant type composition in the dataset of interest. We present CAPS, the Context-Adjusted Proportion of Singletons metric, as a drop-in replacement for MAPS with better performance.

Materials and methods

Score calculation procedure

CAPS is a new metric of negative selection with the expected level of rare variation derived on a per-context basis from synonymous variants.

Even though it has been shown that intergenic variants may be used as an alternative class of variants for the estimation of the baseline level of rare variation (7), we had insufficient evidence that intergenic variants could be used interchangeably with synonymous variants for the purpose of model calibration without any significant effect on the results.

Given that in synonymous variants 4 out of 104 contexts are never observed, we were initially unable to use synonymous variants to calculate CAPS scores for some classes of genetic variation where those contexts were observable. We, therefore, approximated the expected proportions of singletons for those missing contexts from intronic variants using a probit regression model. This enabled us to apply CAPS to all classes of variants, including those with all 104 contexts present (for example, missense). Importantly, CAPS can be calibrated for both genomes and exomes, just like MAPS.

One of the key advantages of using per-context estimates of the expected proportion of singletons over the singletons-by-mutability approach is that each context can be seen as a binomial random variable, which allows the calculation of per-context variance in a mathematically sound way. This, in turn, results in wider and more realistic confidence intervals compared with MAPS' binomial confidence intervals for the mean with the assumption of normality.

CAPS scores can be calculated using either a simple method, which is based on the total variance in the observed number of singletons, or using the posterior predictive distribution (PPD) of CAPS. The actual values of the scores are identical between the two methods; however, the produced confidence intervals differ: the intervals estimated using the PPD method are wider, as they take into account the additional uncertainty around the probabilities that are used to calculate the expected level of singletons for each context (see Supplementary Table S1). Details of how CAPS scores are calculated can be found in Supplemental Note 1.

Data

For model calibration and all analyses we used 125 748 exomes (WES) and 15 708 genomes (WGS) from gnomAD v2.1.1, filtered based on the QC criteria from the original 2020 gnomAD flagship paper (4). The total number of QC-compliant variants per class is shown in Supplementary Tables S2 and S3, with additional per-context statistics shown in Supplementary Tables S4 and S5.

Results

We developed CAPS, a novel metric of negative selection where the expected level of rare variation is derived on a per-context basis instead of using a mutability-based model.

Improved estimates of negative selection

The estimates of MAPS carry a strong bias coming from the singletons-by-mutability model (Figure 1A), which is an integral part of MAPS' design. Specifically, in MAPS, transversion variants are more likely to be assigned a missense-level negative selection score merely due to their low mutation rates (Figure 1B). It is important to note, however, that this bias may not be noticeable when MAPS is calculated over large variant sets. CAPS eliminates this mutability bias completely, as its design does not include the mutability correction layer (Figure 2). As a result, unlike MAPS, CAPS can be used safely to study a wide range of different classes of genetic variation, regardless of the variant type composition in the variant set of interest. To validate the improvement in the estimates after the elimination of the mutability bias, we compared a total of eight pairs of contexts with highest differences in the residuals from the MAPS model. As a reference, we used averaged values from the state-of-the-art pathogenicity predictor AlphaMissense (8). The results demonstrate that in CAPS the estimates are always either in agreement with AlphaMissense values or the difference is not statistically significant, which is not the case in MAPS (Figure 1C and Supplementary Figure S1).

As shown in Figure 3, CAPS' scores of negative selection are highly consistent with those of MAPS, with the estimates of the two metrics agreeing for all categories of variants when either the exome or genome frequency data from the gnomAD database are used. Specifically, both MAPS and CAPS capture the same upward trend in the deleteriousness of synonymous, missense and predicted loss-of-function (pLoF) variants. The major difference in the results observed is in the confidence intervals of the estimates, which are more accurate in CAPS (Supplementary Table S1).

CAPS as a drop-in replacement for MAPS

To demonstrate the applicability of CAPS we sought to apply it to those classes of variants which had been previously studied using MAPS. Supplementary Figures S2 and S3 show how CAPS can be used as a drop-in replacement for MAPS, using reproduced scores from previously reported analyses in upstream open reading frames (uORFs) and near-splice regions, respectively (9–11). As evident from the figures, the CAPS' negative selection scores show good concordance with those estimated using MAPS. These results confirm that uORF uAUG-creating and some uORF stop-removing variants are subject to strong negative selection and that this selection is dependent on the effects that these variants induce and contextual information. Our results also confirm that variants affecting intron-exon junctions are particularly deleterious.

CAPS captures background selection in constrained genes

To check the validity of CAPS' corrected estimates of selection, we compared CAPS and MAPS scores in sets of variants with

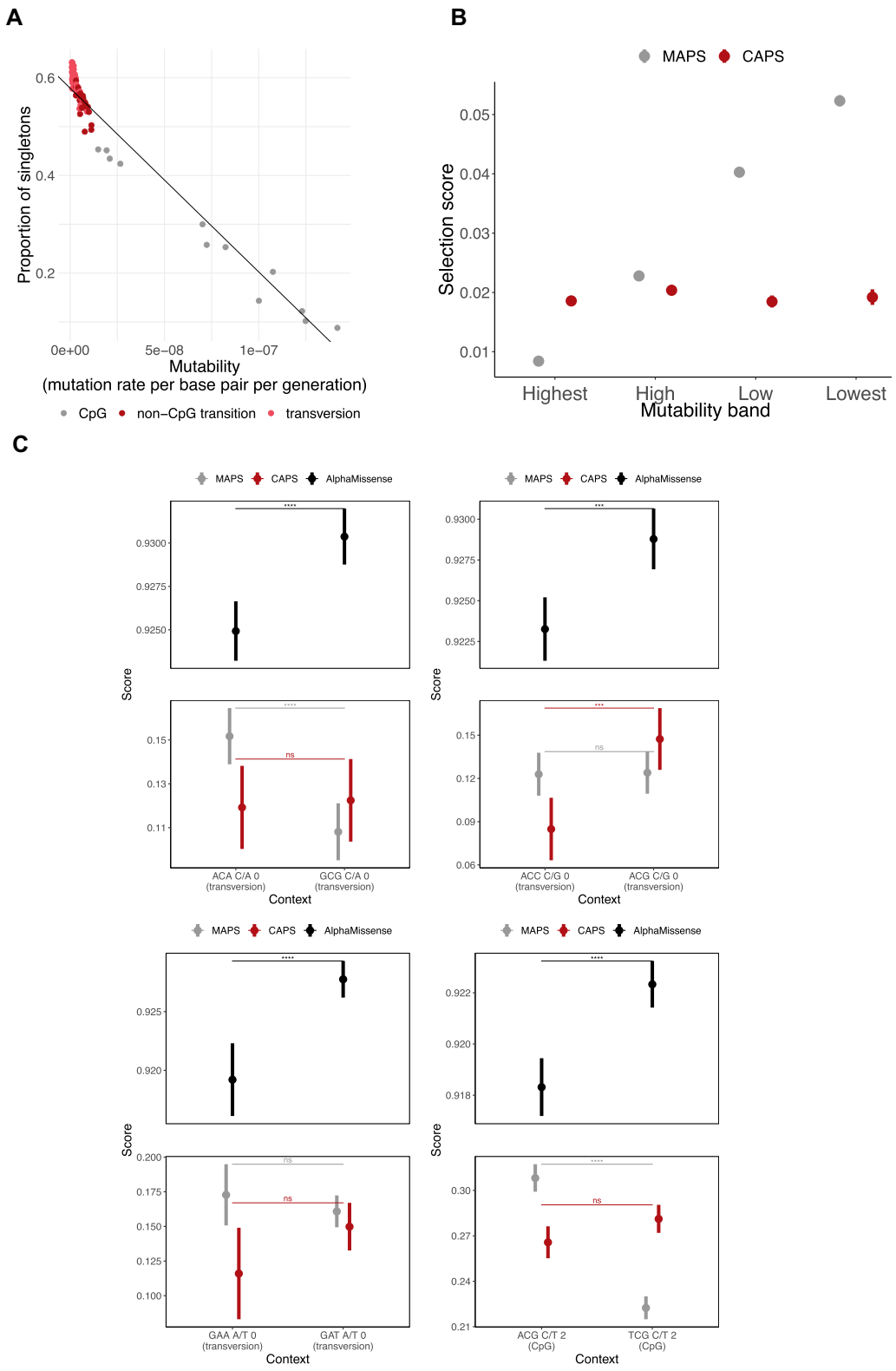

Figure 1. MAPS' singletons-by-mutability model (**A**), the resulting mutability bias in MAPS (**B**) and examples of the bias affecting MAPS estimates (**C**). Error bars are 95% confidence intervals. (A) The singletons-by-mutability model was calibrated on 100 unique contexts observed in QC-compliant WES synonymous variants. The model underestimates the real proportion of singletons for transversions, while overestimating it for CpGs, though this may not be critical when averaging over large numbers of variants. (B) The mutability bias was assessed based on CAPS and MAPS estimates in 'Lowest' (0–25%), 'Low' (25–50%), 'High' (50–75%) and 'Highest' (75–100%) mutability bands on all QC-compliant WES variants. (C) CAPS and MAPS estimates in pairs of contexts with highest difference in the residuals from the MAPS model, with CAPS estimates showing agreement with the corresponding average AlphaMissense values, or the difference being not statistically significant. Labels are made up of the trinucleotide context, reference/alternate alleles, methylation level (for CpGs) and variant type. In each pair, only the trinucleotide context sequences are different. Only QC-compliant missense WES variants with the AlphaMissense class 'pathogenic' and an AlphaMissense score of at least 0.8 were included. Bonferroni-adjusted *P*-values are shown for each comparison (Welch modified two-sample *t*-test): '*' (<0.05), '**' (<0.01), '***' (<0.001), '****' (<0.0001), 'ns' (not significant).

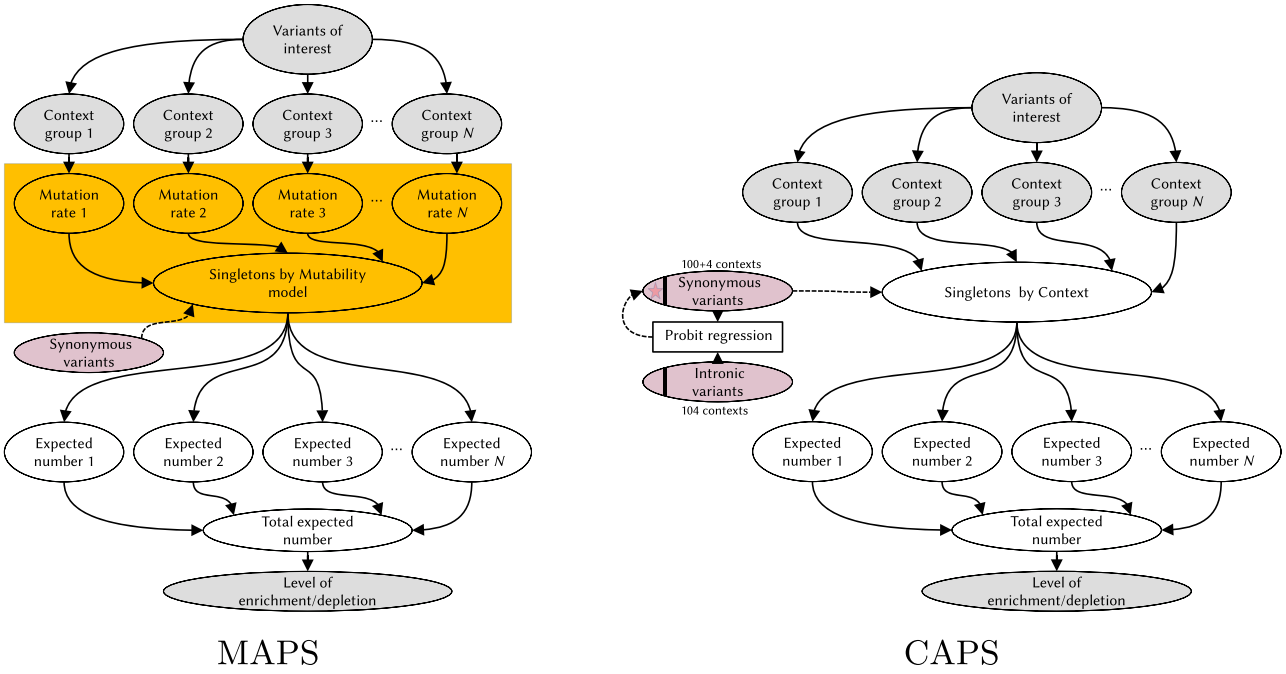

Figure 2. Differences between CAPS and MAPS in the derivation of the scores. In CAPS, the complexity of the model is reduced via complete elimination of the mutability layer.

increasing levels of gene constraint using LOEUF, a metric of intolerance to variation based on the deficit of pLoF variants in a gene (4). To minimize the effect of targeted selection, we limited our dataset to synonymous variants and stratified all variants by gene intolerance level.

In line with our previous findings (6), Supplementary Figure S4 shows that both metrics are sensitive enough to capture the background selection that affects synonymous variants in constrained genes and have striking similarity in the scores. Wider confidence intervals can be seen for CAPS.

Discussion

In this work, we introduced CAPS, a novel metric of selection-based deleteriousness and a drop-in replacement for MAPS. CAPS should be treated as a tool for quantifying the differential selective pressure between synonymous variants and other groups of variants. CAPS eliminates the mutability bias that was present in its predecessor and performs reliably even on small variant sets. Besides, CAPS' confidence intervals are calculated in a more accurate way compared to the simplified intervals used in MAPS.

Even though it is possible to reduce the bias of MAPS towards transversion variants by modifying the fit of the model, we argue that this approach can be seen as *ad hoc* and not future-proof, considering that every new release of gnomAD would require refitting the model. Given the complexity of mutability estimation, there is no straightforward way of modelling mutation rates; however, attempts have been made to incorporate additional mutability-related annotations in the singletons-by-mutability model and use intergenic variants instead of synonymous ones for calibration (7). Overall, we believe that our proposed context-based approach is superior compared with using mutability as a proxy for the estimation of the expected level of rare variation, as mutation rates modelling proves a difficult task.

However, a number of factors should be taken into account when using CAPS (and MAPS, for that matter). First, with CAPS being calibrated on gnomAD v2 data, the scores derived are comparable only when subsets of variants from gnomAD v2 are contrasted. Besides, with CAPS, only a fraction of possible pathogenic variation can be analyzed. This is because in common complex diseases (such as Alzheimer's disease or Type 2 diabetes) the associated variants are likely to exhibit attenuated patterns of purifying selection due to the complex evolutionary impact of these variants, rendering such variation hard to capture with selection-based metrics. Finally, and perhaps most importantly, CAPS—being an instance of the population-level family of metrics, which includes MAPS, LOEUF and pLI (4)—can only serve as an imperfect proxy for selection (12–14). Indeed, all such metrics are only able to estimate the mutation's heterozygous deleteriousness effect (referred to as hs), but cannot disentangle the dominance coefficient h and the homozygous selection coefficient s from it. Hence, being a crude metric based exclusively on the concept of singleton proportions, CAPS has limitations in the analysis of haploinsufficiency, being unable to distinguish between a dominant gene with low selection levels and a recessive gene with higher selection levels, although, again, this also holds true for other population-level metrics.

Nevertheless, we believe that CAPS holds promise to reveal new insights into understudied classes of variation with unknown or poorly quantified deleteriousness, especially for smaller classes of variants. Compared with complex pathogenicity predictors, the design of CAPS is simpler and more interpretable. We believe that this transparency will help

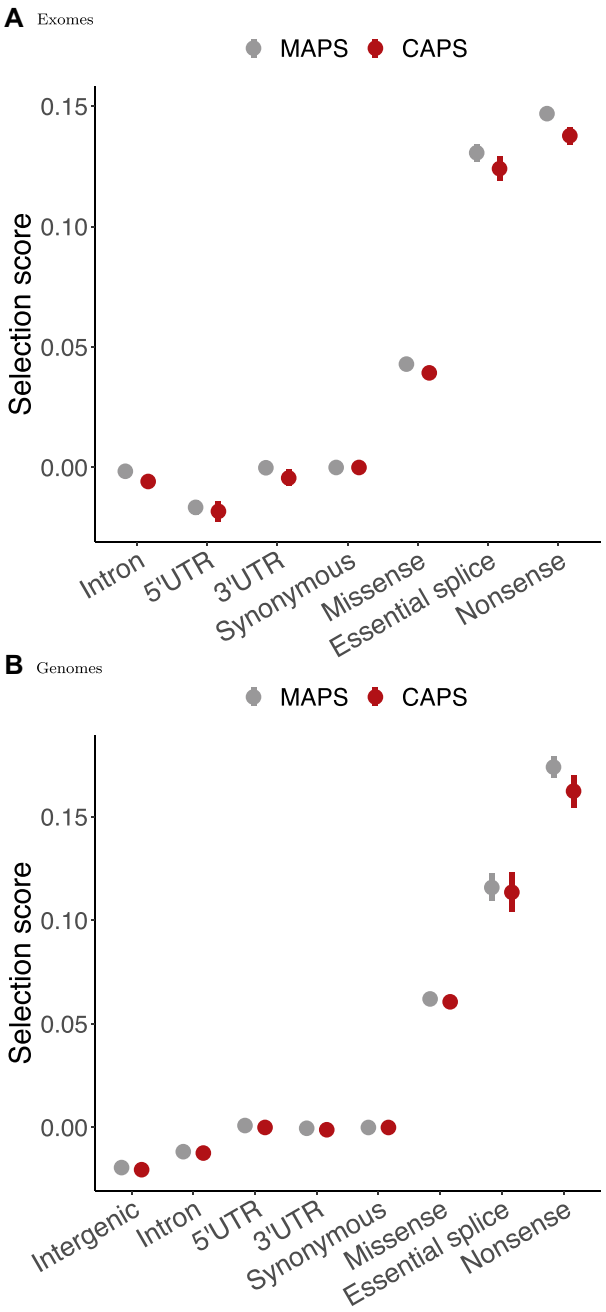

Figure 3. Original (MAPS) and corrected (CAPS) estimates of negative selection in SNVs by variant class for exomes (**A**) and genomes (**B**). Error bars are 95% confidence intervals. All QC-compliant variants.

researchers to better understand the potential impact of genetic variants on human health.

Web resources

- Variants
  - gs://gcp-public-data–gnomad/release/2.1.1/ht/exomes/gnomad.exomes.r2.1.1.sites.ht
  - gs://gcp-public-data–gnomad/release/2.1.1/ht/genomes/gnomad.genomes.r2.1.1.sites.ht
- Coverage
  - gs://gcp-public-data–gnomad/release/2.1/coverage/exomes/gnomad.exomes.r2.1.coverage.ht
  - gs://gcp-public-data–gnomad/release/2.1/coverage/genomes/gnomad.genomes.r2.1.coverage.ht
- Mutability
  - gs://gcp-public-data–gnomad/papers/2019-flagship-lof/v1.0/model/ mutation_rate_methylation_bins.ht
- Methylation and trinucleotide context data
  - gs://gcp-public-data–gnomad/papers/2019-flagship-lof/v1.0/context/ Homo_sapiens_assembly19.fasta.snps_only.vep_20181129.ht
- LoF metrics (including LOEUF) by gene
  - https://storage.googleapis.com/gcp-public-data--gnomad/release/2.1.1/constraint/gnomad.v2.1.1.lof_metrics.by_gene.txt.bgz
- Whiffin et al., Characterising the loss-of-function impact of 5′ untranslated region variants in 15 708 individuals (2020)
  - https://github.com/ImperialCardioGenetics/uORFs/tree/master/data_files
- AlphaMissense scores
  - https://zenodo.org/records/8208688

Data availability

The code generated during this study is available at https://doi.org/10.5281/zenodo.13282923.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We thank Professor Daniel MacArthur, Director of the Centre for Population Genomics, for his valuable feedback on this work.

Funding

NSW Health Early-Mid Career Fellowship (to E.G.); National Health and Medical Research Council [Investigator Grant 2018360 to E.G.).

Conflict of interest statement

None declared.

References

1. Karczewski,K.J. and Martin,A.R. (2020) Analytic and translational genetics. *Annu. Rev. Biomed. Data Sci.*, **3**, 217–241.
2. Harpak,A., Bhaskar,A. and Pritchard,J.K. (2016) Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. *PLoS Genet.*, **12**, e1006489.
3. Samocha,K.E., Robinson,E.B., Sanders,S.J., Stevens,C., Sabo,A., McGrath,L.M., Kosmicki,J.A., Rehnström,K., Mallick,S., Kirby,A., *et al.* (2014) A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.*, **46**, 944–950.
4. Karczewski,K.J., Francioli,L.C., Tiao,G., Cummings,B.B., Alföldi,J., Wang,Q., Collins,R.L., Laricchia,K.M., Ganna,A., Birnbaum,D.P., *et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.
5. Lek,M., Karczewski,K.J., Minikel,E.V., Samocha,K.E., Banks,E., Fennell,T., O'Donnell-Luria,A.H., Ware,J.S., Hill,A.J.,

Cummings,B.B., *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
6. Gudkov,M., Thibaut,L. and Giannoulatou,E. (2024) Quantifying negative selection on synonymous variants. *HGG Adv.*, **5**, 100262.
7. Findlay,S.D., Romo,L. and Burge,C.B. (2024) Quantifying negative selection in human 3' UTRs uncovers constrained targets of RNA-binding proteins. *Nat. Commun.*, **15**, 85.
8. Cheng,J., Novati,G., Pan,J., Bycroft,C., Žemgulytė,A., Applebaum,T., Pritzel,A., Wong,L.H., Zielinski,M., Sargeant,T., *et al.* (2023) Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, **381**, eadg7492.
9. Whiffin,N., Karczewski,K.J., Zhang,X., Chothani,S., Smith,M.J., Evans,D.G., Roberts,A.M., Quaife,N.M., Schafer,S., Rackham,O., *et al.* (2020) Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *Nat. Commun.*, **11**, 2523.
10. Lord,J., Gallone,G., Short,P.J., McRae,J.F., Ironfield,H., Wynn,E.H., Gerety,S.S., He,L., Kerr,B., Johnson,D.S., *et al.* (2019) Pathogenicity and selective constraint on variation near splice sites. *Genome Res.*, **29**, 159–170.
11. Blakes,A.J.M., Wai,H.A., Davies,I., Moledina,H.E., Ruiz,A., Thomas,T., Bunyan,D., Thomas,N.S., Burren,C.P., Greenhalgh,L., *et al.* (2022) A systematic analysis of splicing variants identifies new diagnoses in the 100,000 genomes project. *Genome Med.*, **14**, 79.
12. Fuller,Z.L., Berg,J.J., Mostafavi,H., Sella,G. and Przeworski,M. (2019) Measuring intolerance to mutation in human genetics. *Nat. Genet.*, **51**, 772–776.
13. Agarwal,I., Fuller,Z.L., Myers,S.R. and Przeworski,M. (2023) Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs. *eLife*, **12**, e83172.
14. Zeng,T., Spence,J.P., Mostafavi,H. and Pritchard,J.K. (2023) Bayesian estimation of gene constraint from an evolutionary model with gene features. *Nat. Genet.*, **56**, 1632–1643.

Received: February 8, 2024. **Revised:** July 24, 2024. **Editorial Decision:** August 6, 2024. **Accepted:** August 14, 2024