BMC Genomics

RESEARCH

Open Access

# Mutational constraint analysis workflow for overlapping short open reading frames and genomic neighbors

Martin Danner[1,2], Matthias Begemann[1], Florian Kraft[1], Miriam Elbracht[1], Ingo Kurth[1] and Jeremias Krause[1*]

**Abstract**

Understanding the dark genome is a priority task following the complete sequencing of the human genome. Short open reading frames (sORFs) are a group of largely unexplored elements of the dark genome with the potential for being translated into microproteins. The definitive number of coding and regulatory sORFs is not known, however they could account for up to 1–2% of the human genome. This corresponds to an order of magnitude in the range of canonical coding genes. For a few sORFs a clinical relevance has already been demonstrated, but for the majority of potential sORFs the biological function remains unclear. A major limitation in predicting their disease relevance using large-scale genomic data is the fact that no population-level constraint metrics for genetic variants in sORFs are yet available. To overcome this, we used the recently released gnomAD 4.0 dataset and analyzed the constraint of a consensus set of sORFs and their genomic neighbors. We demonstrate that sORFs are mostly embedded into a moderately constrained genomic context, but within the gencode dataset we identified a subset of highly constrained sORFs comparable to highly constrained canonical genes.

**Keywords** Short open reading frames, Population genomics, Computational genomics

## Introduction

### Definition of short-open-reading-frames

Variants in the human genome can lead to a variety of pathologies and genome analysis is increasingly used as a basis for clinical decision making [1]. While sequencing technologies have improved rapidly in the past years, an extensive analysis of whole genome data lacks behind. The interpretation of genomic data is often limited to the protein coding parts of the genome, which makes up only 1–2% of the human genome [2]. For the far larger part of the genome, sometimes referred to as "Dark Genome" [3, 4], no adequate analysis strategies are available. Part of this "Dark Genome" are genomic elements called "short open reading frames" (sORFs). These sORFs are non-canonical reading frames shorter than the conventionally defined 100 codons, that may overlap with canonical coding regions and might encode for functional microproteins or fulfill regulatory functions [5]. sORFs are dispersed across the whole genome and several attempts have been made to classify and group them into different sub- categories [5–11].

### sORF datasets and nomenclature

Our analysis is based on a consensus paper published in 2022 by Mudge et al. [5] in which the used sORF consensus set has been presented and categorized by their

*Correspondence:
Jeremias Krause
jerkrause@ukaachen.de
[1]Institute for Human Genetics and Genomic Medicine Medical Faculty, RWTH Aachen University Hospital, Pauwelsstrasse 30, D-52074 Aachen, North-Rhine-Westphalia, Germany
[2]Scieneers GmbH, Kantstraße 1a, 76137 Karlsruhe, Baden-Wuerttemberg, Germany

genomic localization in comparison to canonical coding as well as non-coding elements. Six categories have been proposed and annotated with a terminology which we have adapted. These six categories are sORFs falling into canonical genes (intORFs), sORFs falling onto long noncoding RNAs (lncRNA-ORFs), sORFs falling into the 5' (uORF) or 3' (dORF) UTR of canonical genes and sORFs that start in the 5' (uoORF) or 3' (doORF) UTR of a canonical gene but reach into the canonical coding region of the gene [5]. For some of these sORFs a potential clinical relevance has been demonstrated [12, 13], although the biological role of the majority of sORFs remains elusive. The actual number of predicted sORFs varies within the literature. Conservative estimates used in the gencode sORF consensus set or in another small consensus set provided by Chen et al. [9] contain a few thousand sORFs. This contrasts with extremely large sORF datasets presented by Neville et al. [10] and Li et al. [11] which contain hundreds of thousands to millions of sORFs. Ultimately, experimental progress will result in the reduction of false positive annotations, although other strategies have been proposed. Of particular importance for our study is the strategy by Jain et al. [14]. They suggested using constraint metrics to reduce the number of predicted sORFs, by excluding sORFs which show close to no constraint. We selected the gencode dataset for our analysis due to several reasons: it contains sORFs which have been found in multiple ribo-seq studies, has redundant sORFs merged, includes sORFs with the canonical start codon ATG, covers both overlapping and non-overlapping sORFs and has undergone review by an international consensus working group.

### Quantifying the constraint of sORFs in the general population

We investigate the constraint of sORFs from different perspectives. At first, we investigate the suitability of existing genomic constraint scores (e.g. the Gnocchi score), secondly, we adapt constraint scores presented in the gnomAD constraint pipeline for a tailored constraint calculation, lastly, we compare the calculated sORF constraint to neighboring genomic elements.

### Proposing a constraint workflow for the analysis of overlapping sORFs

While constraint metrics of sORFs that do not overlap with canonical regions have been calculated before, to our knowledge, a constraint analysis for overlapping sORFs, such as those in the gencode dataset, has not been carried out before and provided to the public, particularly using the gnomAD 4.0 dataset. Analysis of overlapping genomic elements can be challenging, because it is necessary to differentiate between actual relevant constraints affecting the individual feature and out of frame effects from overlapping genomic elements. We approached the analysis of overlapping sORFs by calculating constraint metrics, which make different assumptions about the genomic elements in question. We hypothesize that by comparing these different scores, we might be able to partially separate the constraint of overlapping sORFs from their neighboring genomic background. We demonstrate that a subset of sORFs is highly constrained when analyzed for the intolerance for missense variants, while being less constrained when treated as a genomic region with unspecified single nucleotide variants.

## Results

### Sample size, statistical power and coverage

While sORFs have largely been investigated for evolutionary conservation [5, 11, 15–17], an extensive quantification for their selective pressure in a large-scale dataset has only partially been investigated. Reasons for that can be seen in their short length as sORF analysis can be hindered by low statistical power [18]. This is particularly noticeable when examining loss-of-function variants which are less likely to occur. The analysis of loss-of-function intolerance in short genomic elements can be complicated when sample size is limited. With the recent gnomAD [19] 4.0 release, these shortcomings can be overcome for some variant types. gnomAD 4.0 contains whole-genome data of 76,215 and whole-exome data from 730,947 individuals which brings up the total number of reference samples to 807,162 individuals. At first, we tested whether all gencode sORFs are contained in the whole-exome samples. This revealed that 4,274 out of the 7,264 gencode sORFs are included in regions present in gnomAD 4.0 exomes. However, not all of them have sufficient coverage for downstream analysis. Therefore, we limited some of our downstream analysis to the data from the 76,215 whole-genome samples contained in the gnomAD genomes.

### Analysis of mutational background

The descriptive analysis was conducted on the MANE select subset of coding variants present in genes and sORFs. In their current release the gnomAD genomes contain 4,320,631 unique missense variants, 2,188,653 synonymous variants and 376,271 high impact SNVs (start loss, stop loss, stop gain and frameshift variants) located in canonical genes. Considering the same MANE select transcripts, the gencode sORFs contain 101,445 missense variants, 44,118 synonymous variants and 19,441 high impact variants. The distribution is visualized in Fig. 1. Canonical genes and sORFs show a similar distribution of variants with the majority of variants being missense variants, followed by synonymous variants. Protein truncating variants like frameshift and stop-gain variants are a minority, although the gencode sORFs

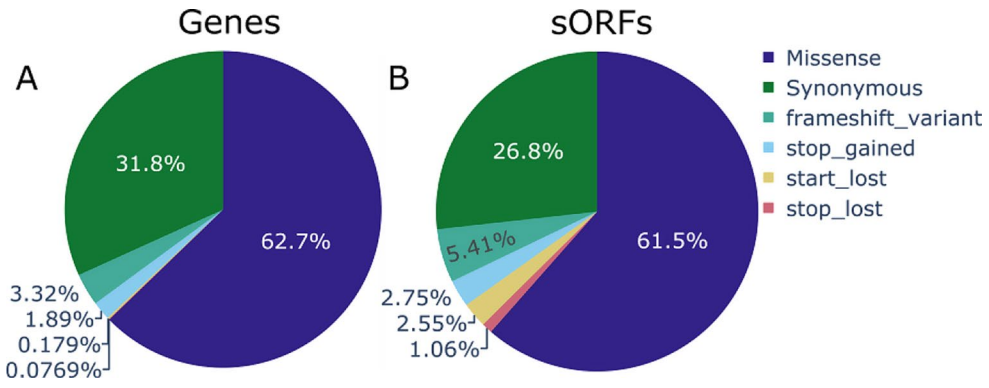


**Fig. 1** The distribution of variants present in the gnomAD 4.1.0 genomes **A**) Variant distribution for all MANE select transcripts (genes); **B**) Variant distribution for all Mane select transcripts (sORFs)

show a noticeably higher amount of high impact variants like frameshift variants and start loss variants.

#### Estimation of sORF constraint utilizing different genomic constraint metrics

As a next step we analyzed whether sORFs are generally intolerant against single nucleotide variants (SNVs) and whether this can be evaluated using established constraint scores. Therefore, we annotated the sORFs and for comparison three groups of non-coding genomic elements (lncRNAs, miRNAs and snoRNAs) with the recently published Gnocchi score [20]. The distribution of these four elements can be seen in Fig. 2. As demonstrated previously by Chen et al. [20] a noticeable portion of miRNAs (mean Gnocchi Score = 1.264) located on autosomes passes the cut-off value of 4 (15.4%), while only the minority of lncRNAs (mean Gnocchi Score = 0.605) show a highly constrained Gnocchi score (2%). snoRNAs (mean Gnocchi Score = 0.447) showed a Gnocchi score comparable with that of lncRNAs. Only 1.5% of snoRNAs fall into the highly constrained percentile of the genome. In comparison, some sORFs (mean Gnocchi Score = 0.152) show a higher constraint according to the Gnocchi score which places 7% into the most constrained percentile.

In the next step we wanted to analyze the sORFs with a more tailored approach, when compared to the larger Gnocchi score windows. Followingly, we calculated a constraint metric with a higher resolution, considering only the coding bases of the predicted sORFs. Using the genome data from gnomAD 3.0 / gnomAD 4.0, we calculated the SNV observed/expected upper bound fraction for SNVs falling into sORFs. We termed this value the SNVOEUF, and it is an adaption of the loss-of-function observed/expected upper bound fraction (LOEUF) score proposed by Karczewski et al. [18] and the missense observed/expected upper bound fraction (MOEUF) score put forward by Jain et al. [14].

Considering the gnomAD genome data, only 15 sORFs from the gencode dataset have less than 10 expected SNVs and were therefore discarded in the SNVOEUF analysis. The distribution of the SNVOEUF values can be seen in Fig. 3. Like the Gnocchi score, the SNVOUEF does not make any assumptions about the coding potential of the analyzed region. Instead, it provides an estimate of the overall constraint of the region for SNVs. For downstream analysis we compared the SNVOEUF score to the SNVOEUF score of MANE select transcript filtered coding genes and the SNVOEUF score of the UTR regions corresponding to these MANE select transcript filtered gene list. As visible in Fig. 3 only a small number of sORFs (53/7,249) passes the cut-off value for highly constrained regions, if the highest constraint decile of the SNVOEUF of the UTRs is taking as cut-off reference. When compared to genes, 700/7,249 sORFs fall into the highest constraint decile of the SNVOEUF of coding gene regions. The Gnocchi score and SNVOEUF do not correlate well (Kendall Rank Correlation Coefficient=-0.07, $p < 0.001$). This weak correlation was confirmed in a repeated experiment, in which only the SNVOEUF values of sORFs that fall exactly into on Gnocchi interval were considered (Kendall Rank Correlation Coefficient=-0.1, $p < 0.001$). The SNVOEUF distribution plots of the individual sORF subsets can be found in the appendix.

#### Most sORFs show a moderate constraint against missense variants

Because sORFs might encode for functional microproteins we additionally calculated a missense constraint score (MOEUF) and loss-of-function constraint score (LOEUF) to unravel the constraint for specific mutational categories, which might be missed when only the SNVOEUF is analyzed. Again, we computed the number of expected variants for both classes, to estimate whether reasonable assumptions can be made about a mutational constraint considering the sample size present in gnomAD. Utilizing the genomes from gnomAD 3.0 / 4.0

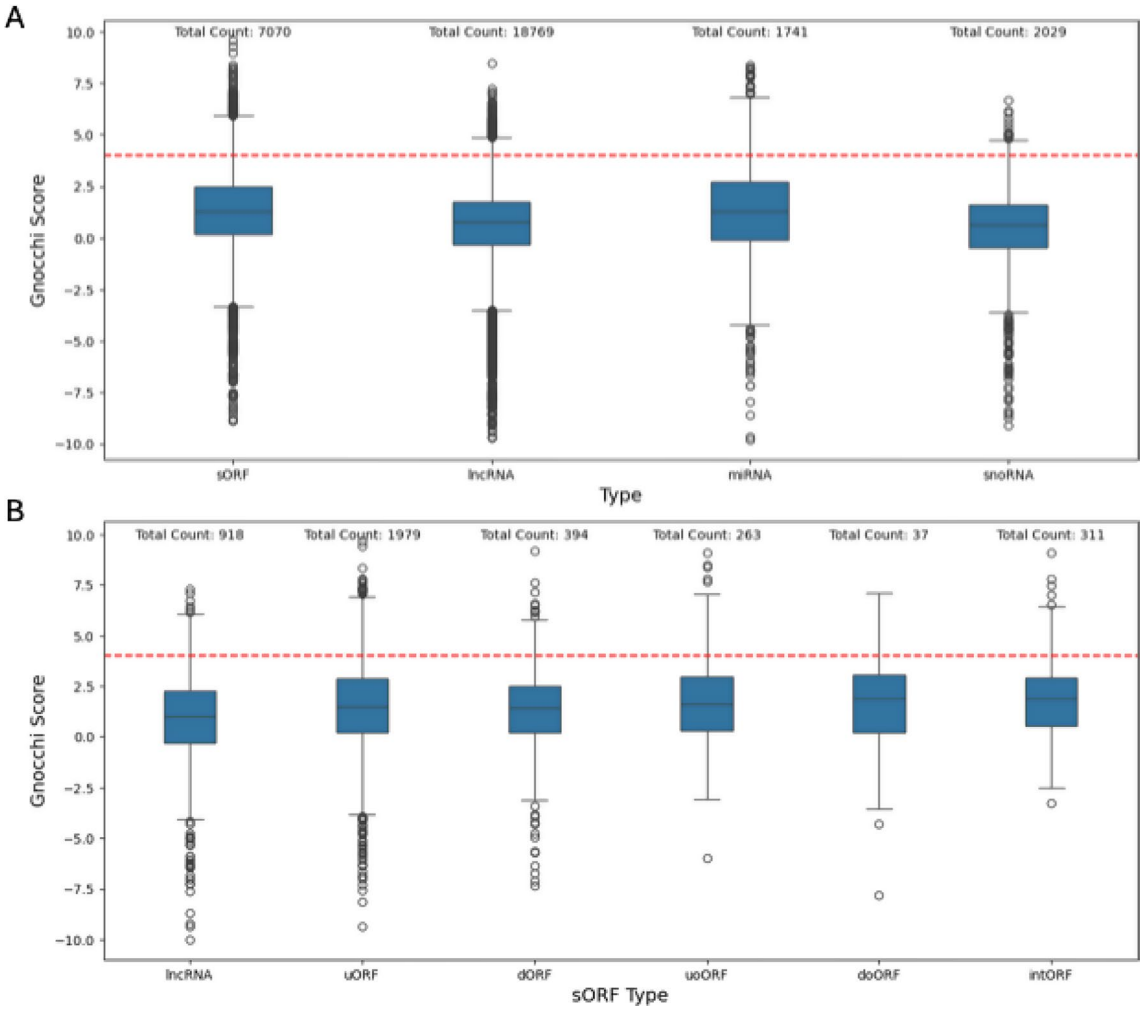


**Fig. 2** **A**) Boxplot showing the Gnocchi score distribution of lncRNAs, snoRNAs, miRNAs and sORFs. **B**) Boxplot showing the Gnocchi score distribution of the different classes of genecode sORFs. The cut-off value of 4 (here depicted by a vertical red line) marks the border of the most constrained percentile of the genome, as introduced by Chen et al. [20] The Gnocchi scores were taken from supplementary dataset 3 from the Gnocchi score paper [20]

5,573 out of 7,264 gencode sORFs had equal to or more than 10 expected missense variants. For the same dataset only 2 sORFs had equal to or more than 10 expected loss-of-functions variants, highlighting that despite its size gnomAD is not large enough to make estimates about loss-of-function intolerance using OEUF values for sORFs. Based on the results of the gnomAD flagship paper, which was published using data from 141,456 samples [18], the authors concluded that approximately 75% of genes had a power sufficient for constraint analyses and extrapolated their expected values to estimate a required sample size sufficient for constraint analyses. In the supplementary material of their article (supplementary Fig. 8 "The sample size required for well-powered constraint") the authors also included a figure for OEUF values for LOF variants. This figure highlights that for sufficient power for smaller genes the required sample size needs to be in the millions.

Since the average sORF is by orders of magnitude shorter than the average canonical gene, this limitation is followingly inherited by sORF encode microproteins. While LOEUF analyses therefore are currently out of scope of this paper we restricted our visualization to the distribution of MOEUF scores which can be seen in Fig. 4. Within the gnomAD genome dataset the sORF MOEUF distribution of (non-overlapping sORFs) differs significantly from the MOEUF distribution of MANE select transcripts of canonical genes when calculated from the genome dataset (Kolmogorov- Smirnov $p < 0.001$). Additionally, a significant difference between these filtered sORF MOEUF values ($M_{Rank}$ = 16897.79) and the MOEUF value of the MANE select transcripts of canonical genes ($M_{Rank}$ = 10857.72) was observed (Mann-Whitney-U-Test: $U = 77741709.5$, $p < 0.001$). This was further explored by calculating the Vargha and Delaney A measurement [21] which returned an estimate

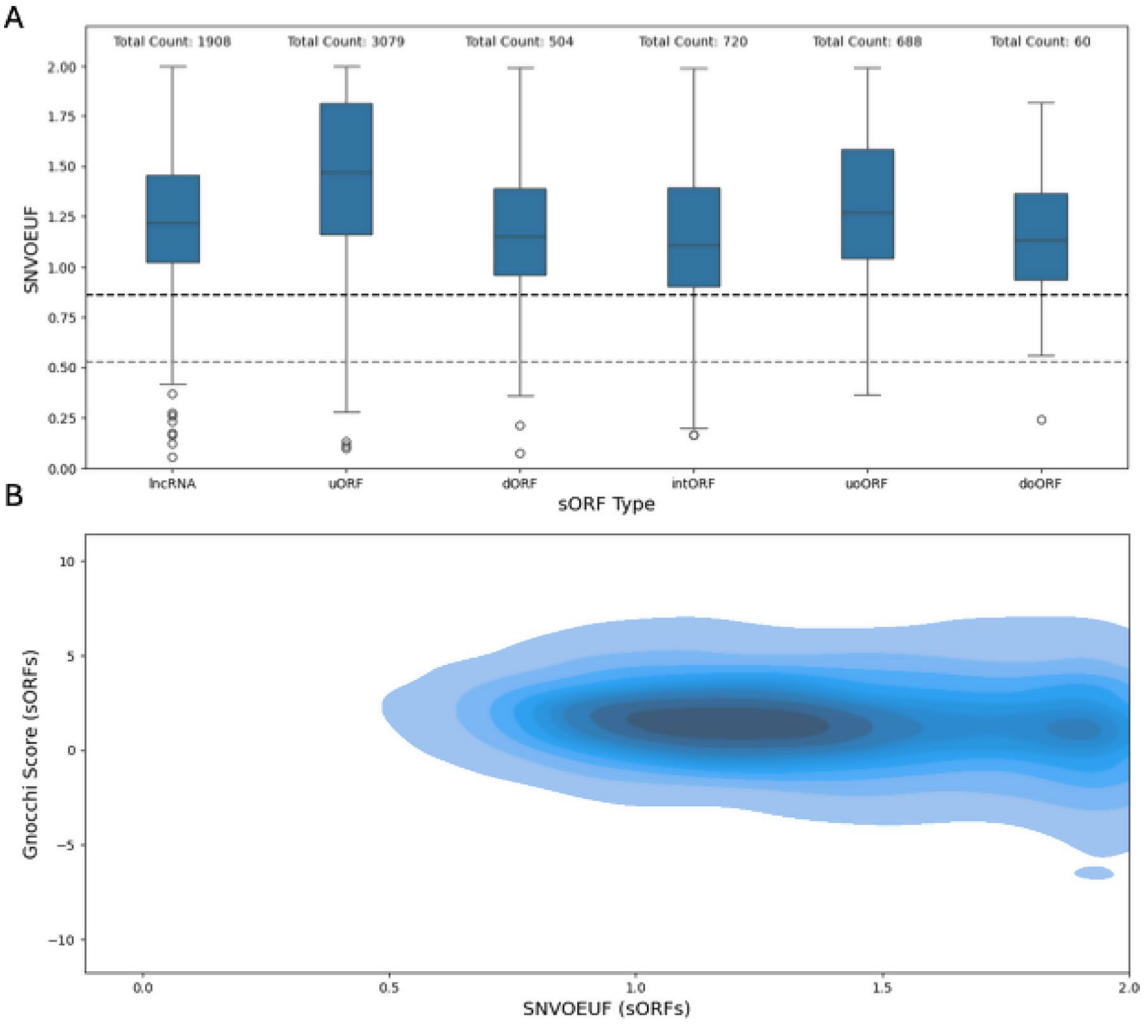


**Fig. 3** **A**) Boxplots of the SNV observed/expected upper bound fraction for the different classes of gencode sORFs. We used 0.86 and 0.53 as cut-off values for intolerance. OEUF scores are best interpreted in a continuous manner, but for downstream analysis decile-based filtering is suggested by gnomAD. 0.86 (black vertical line) marks the SNVOEUF cut-off value base on the most constrained decil with respect to the mane select transcripts of canonical genes. 0.53 (grey vertical line corresponds to the SNVOEUF cut-off value considering the UTRs from mane select transcripts. **B**) KDE (Kernel Density Estimate) plot depicting the relation between SNVOEUF values, and the Gnocchi score obtained from the supplementary dataset 3 from the Gnocchi score publication [20]

of 0.75. The comparison of the two complete distributions can be seen in Fig. 4B. As visible in Fig. 4, a small number of sORFs (111 / 5,573) fall into the highly constrained area, which is below the depicted cut-off values. To leverage the way larger sample size of the gnomAD 4.0 exome dataset and to estimate the effect of a sample size limitation, we also calculated the MOEUF scores of the previously mentioned 4,274 well covered sORFs included in the gnomAD exome regions. The cut-off value of the highest constrained decile for the MANE select transcripts of the canonical genes was used to study how many sORFs fall into this highly constrained decile. Of the 4,274 sORFs 3,578 had more than 10 expected missense variants and therefore have sufficient data for constraint analysis. Again, within the gnomAD exome dataset the sORF MOEUF distribution (of non-overlapping sORFs) differs significantly from the MOEUF distribution of MANE select transcripts of canonical genes when calculated from the exome dataset (Kolmogorov-Smirnov $p < 0.001$). Similar to the comparison in the genome dataset, a significant difference between these filtered sORF MOEUF values ($M_{Rank}$ = 14604.63) and the MOEUF value of the MANE select transcripts of canonical genes ($M_{Rank}$ = 9987.03) was observed (Mann-Whitney-U-Tests: $U = 31096690.5$, $p < 0.001$). Again, the Vargha and Delaney A measurement [21] was computed and returned an estimate of 0.72. 175 of these 3,578 sORFs fell into the highly constrained decile. Overall, our results are in line with the results presented by Jain et al. [14] who demonstrated that the distribution of MOEUF scores of sORFs tend to be similar to the MOEUF score of less constrained genes,

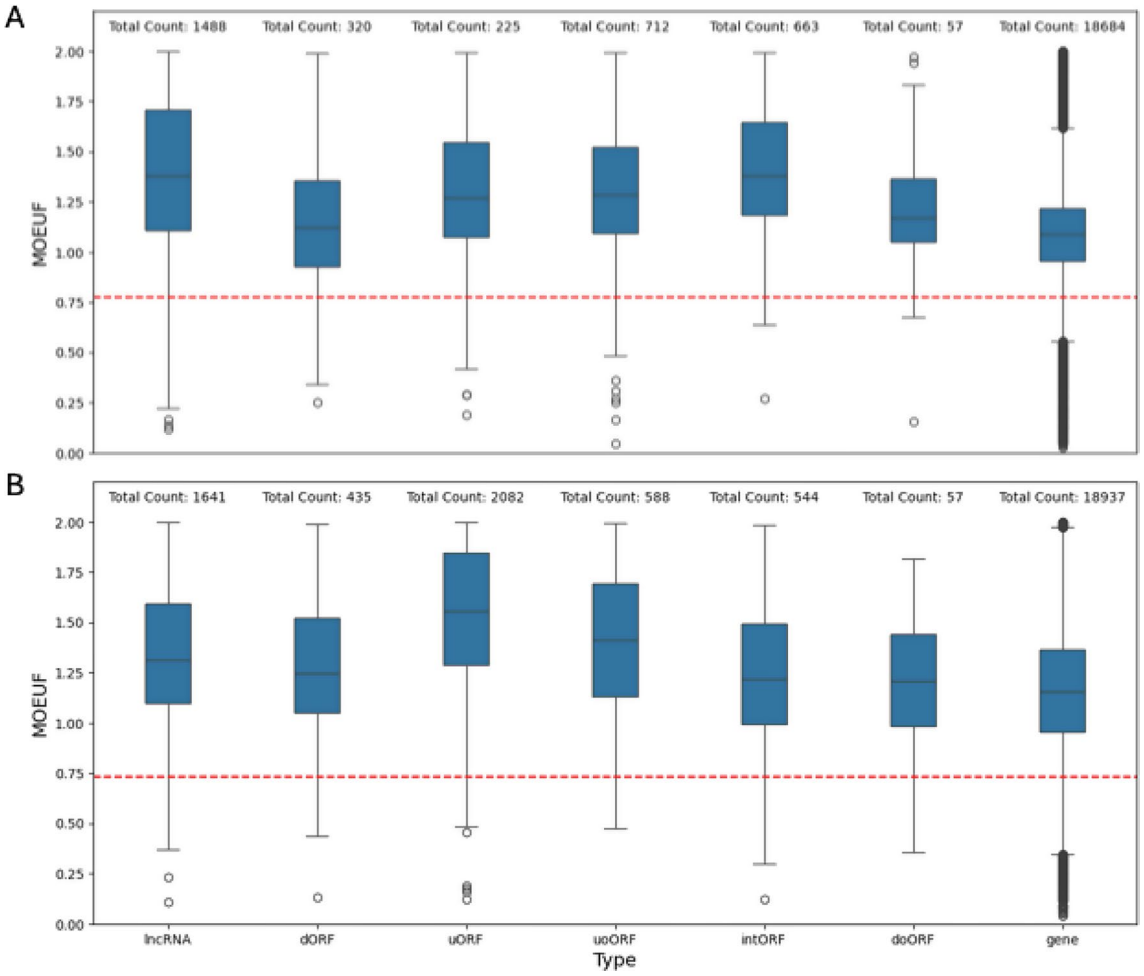


**Fig. 4** **A**) Boxplots showing the MOEUF score distribution of the mane-select transcripts of canonical genes and the different classes of gencode sORFs computed using the data of 730,947 exomes. The red vertical line (0.77) marks the most constrained decile of the canonical genes. **B**) Boxplots showing the MOEUF score distribution of the mane-select transcripts of canonical genes and the different classes of gencode sORFs computed using the data of 76,215 genomes. The red vertical line (0.73) marks the most constrained decile of the canonical genes. The plots share a common X-axis

as the majority of sORFs falls into the lesser constrained interval between a MOEUF value of 1 and 1.5 (1779/3578 sufficiently powered sORFs for the gnomAD exomes and 2560/4274 sufficiently powered sORFs for the gnomAD genomes). Plots of the MOEUF distributions of individual sORF categories can be found in the appendix.

### Identifying sORFs with OEUF values that respect their reading frame

To separate actual constrained overlapping sORFs from background out of frame effects from overlapping genomic elements, we compared the MOEUF values to the SNVOEUFs value of the sORFs. We demonstrate that 3084/3578 sORFs (of the sORFs sufficiently covered and sufficiently powered sORFs in the whole exome data) have a lower MOEUF value than SNVOEUF value (mean MOEUF value = 1.34 compared to a mean SNVOEUF value = 1.65). When the highly constrained subset of 185 sORFs from the whole exome data is analyzed, 160/185 sORFs show a lower MOEUF value, when compared to their corresponding SNVOEUF value (mean MOEUF value = 0.6 compared to a mean SNVOEUF value = 0.933). For the whole genome data 1506/5573 sufficiently powered sORFs have a lower MOEUF than SNVOEUF value (mean SNVOEUF value = 1.3 compared to a mean MOEUF value = 1.4). For the highly constrained sORFs in the whole genome dataset 62/129 sORFs have a lower MOEUF than SNVOEUF value (mean SNVOEUF value = 1.29 compared to a mean MOEUF value = 1.4).

### Most sORFs are neighbored by moderately constrained genes

While Jain et al. [14] compared the general distribution of sORFs to the distribution of RefSeq genes, to our knowledge the constraint of sORFs has not been previously compared to the directly neighbored genes or genes into which sORFs are embedded and therefore to the genomic background in which sORFs have evolved relatively

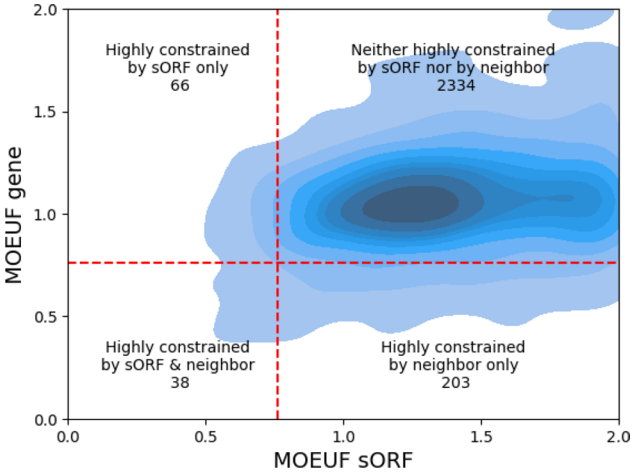


**Fig. 5** KDE (Kernel Density Estimate) plot illustrating the relationship between the MOEUF values of sORFs and their neighboring genes, calculated using gnomAD exome data. The quadrants interpret as follows: the lower left indicates sORFs highly constrained by both their own and neighboring genes' MOEUF values, the upper left shows sORFs highly constrained by their own MOEUF values alone, the lower right highlights sORFs highly constrained by the MOEUF values of their neighboring genes only, and the upper right represents sORFs which are neither highly constrained by their own nor neighboring genes

recently. Therefore, we used the gnomAD exome data to compare the MOEUF values of sORFs to the MOEUF values of their neighbored genes. Figure 5 highlights two points. Firstly, the gencode sORFs tend to be neighbored with genes that are subject to a moderate constraint. Secondly, with a correlation coefficient value of 0.12 (Kendall Rank Correlation Coefficient = 0.12, $p < 0.001$) there seems to be only a weak connection between the MOEUF values of sORFs and their neighbored genes. This is further supported if analyzed in a pairwise fashion. sORFs (median = 1.49) show a significantly higher MOEUF compared to their neighbored genes (median = 1.10) (Wilcoxon signed ranked test: W = 801505.5, $p < 0.001$). Plots of the individual subcategories of sORFs can be found in the appendix.

### Further analyzing the genomic neighbors of highly constrained sORFs

To get a hint for possible clinical relevance, we analyzed whether some of these highly constraint sORFs are neighbored by protein coding genes, which are involved in monogenic diseases, since it has been demonstrated by Mudge et al. that a subset of sORFs function as regulatory elements for their genomic neighbors. Of the 104 sORFs that are constrained by either both their own and neighboring genes' MOEUF values or solely their own, 101 have unique neighboring genes. Using Ensembl [22] Release 113, phenotype associations were identified for 30 genes. An overview of these genes and their respective phenotypes is provided as supplementary dataset 4. Figure 6 illustrates the distribution of sORF types among these 30 highly constrained genes linked to phenotypes. Further analysis of Ensembl Release 113 revealed Gene Ontology (GO) terms for 99 genes. Supplementary Figs. 30, 31, and 32 present bar plots of the top 20 GO terms by count for each gene term domain. The underlying data for the GO analysis is provided as supplementary dataset 5.

Additionally, we obtained phyloP values from the most recent Zoonomia [23] publication, which is a measurement for evolutionary conservation and mapped them to the subset of highly constrained sORFs and genes. Followingly, we calculated the average phyloP scores for the highly constrained sORFs from the exome data and subsequently also for the coding parts of the sORFs and genes and compared them towards each other. The results can be seen in Table 1.

### Constraint comparison between sORFs and UTRs

To further analyze the genomic context of sORFs we investigated the regional constraint of UTRs retrieved from the UTR 2.0 database [24]. For this we computed the SNVOEUF score for all UTRs within the UTR 2.0 database assigned to MANE select transcripts, using the data from the gnomAD genomes. Additionally,

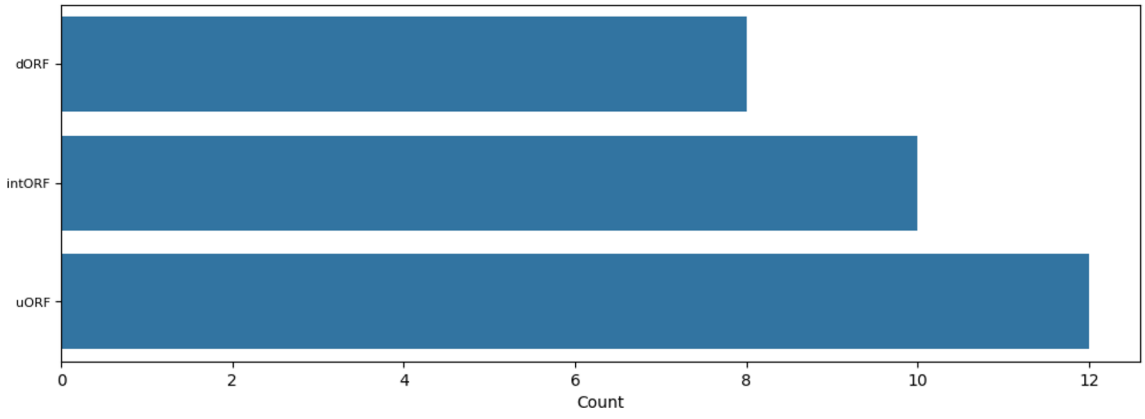


**Fig. 6** Distribution of sORF types among the 30 highly constrained genes linked to phenotypes

**Table 1** Evolutionary conservation of highly constrained sORFs neighbored by genes with associated phenotypes

| sORF | Gene | Mean phyloP sORF | Mean phyloP gene | sORF class | Conservation status established by Sandmann et al. [17] |
|---|---|---|---|---|---|
| C9NOREP73 | AOPEP | 0.11 | 2.62 | dORF | Primatomorpha |
| C18NOREP74 | BCL2 | 2.52 | 1.60 | uORF | Human |
| C20NOREP99 | EEF1A2 | 4.11 | 3.98 | dORF | Primatomorpha |
| C5RIBOSEQORF120 | ERGIC1 | 1.14 | 2.25 | dORF | Old work Monkeys |
| C8RIBOSEQORF34 | EXTL3 | 1.25 | 1.66 | uORF | Primatomorpha |
| CXRIBOSEQORF28 | FGD1 | 2.82 | 3.28 | uORF | Primatomorpha |
| CXNOREP99 | FGF13 | 2.92 | 0.38 | uORF | Primatomorpha |
| C19NOREP210 | FUT1 | -0.46 | 0.63 | uORF | Primatomorpha |
| C4RIBOSEQORF123 | HAND2 | 2.82 | 3.19 | uORF | Conserved |
| C1RIBOSEQORF39 | HNRNPR | 6.02 | 3.01 | intORF | Conserved |
| CXRIBOSEQORF37 | IGBP1 | -0.52 | 1.68 | uORF | Primatomorpha |
| CXNOREP24 | KDM6A | 4.85 | 3.71 | intORF | Conserved |
| C17RIBOSEQORF80 | KRT10 | 2.78 | 2.88 | dORF | Primatomorpha |
| C19NOREP118 | LSM14A | 5.17 | 3.55 | intORF | Conserved |
| C5NOREP142 | MATR3 | 5.66 | 3.86 | intORF | Conserved |
| CXRIBOSEQORF6 | MID1 | 3.71 | 1.68 | uORF | Conserved |
| C16RIBOSEQORF112 | MTSS2 | 2.96 | 2.59 | uORF | Primatomorpha |
| C1NOREP188 | NOTCH2 | 3.92 | 2.79 | intORF | Conserved |
| C13NOREP28 | NUDT15 | 0.56 | 1.11 | dORF | Primatomorpha |
| C5NOREP99 | PPIP5K2 | 2.12 | 1.38 | uORF | Primatomorpha |
| C7NOREP21 | RAC1 | 4.94 | 2.89 | intORF | Primatomorpha |
| CXRIBOSEQORF13 | SH3KBP1 | 2.96 | 2.76 | uORF | Conserved |
| C16NOREP74 | SIAH1 | 5.48 | 3.92 | intORF | Conserved |
| C8RIBOSEQORF45 | SLC20A2 | 4.18 | 2.71 | uORF | Primatomorpha |
| CXNOREP28 | SYN1 | 0.99 | 3.37 | dORF | Primatomorpha |
| C20NOREP24 | TASP1 | 5.06 | 3.81 | intORF | Old work Monkeys |
| C3NOREP214 | TBL1XR1 | 6.23 | 3.50 | intORF | Conserved |
| CXNOREP58 | TIMM8A | 2.24 | 2.08 | dORF | Primatomorpha |
| CXNOREP11 | TMSB4X | 2.29 | 2.95 | dORF | Primatomorpha |
| CXNOREP17 | ZFX | 3.98 | 1.98 | intORF | Conserved |

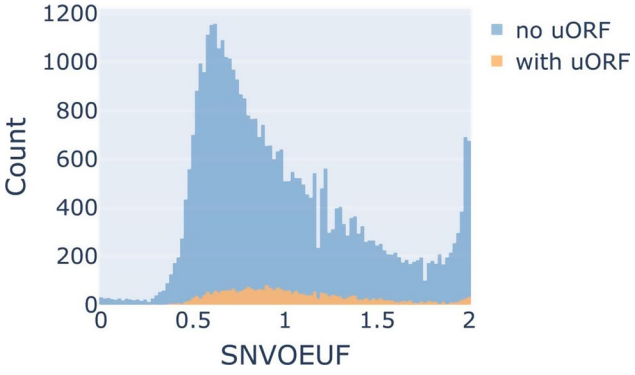


**Fig. 7** UTROEUF distribution of UTR regions containing uORFs (orange) and without uORFs (blue)

we differentiated between UTRs that contain gencode sORFs and those that do not. In a similar vein to the MOEUF comparisons, we calculated deciles and picked the maximum SNVOEUF value of the most constrained decile (0.53) as a cut-off value for SNV-intolerant UTR regions. As it can be seen in Fig. 7, UTRs which contain uORFs are dispersed across the SNVOEUF distribution of UTRs. Their distribution is significantly different to the SNVOEUF distribution of UTRs without uORFs (Kolmogorov-Smirnov $p < 0.001$). A significant difference between the sORF containing UTR SNVOEUF values ($M_{Rank}$=22516.53) and the SNVOEUF values of UTRs without sORF ($M_{Rank}$= 19926.71) was observed (Mann-Whitney-U-Test: $U = 46219468.0$, $p < 0.001$). The subsequently computed Vargha and Delaney A measurement returned an estimate of 0.56. To further analyze the UTR regions, we examined whether canonical genes with multiple UTR regions show regional constraint between their UTRs. As previously introduced, we filtered the annotated UTRs for UTRs with at least 10 expected variants. This reduced the original UTR dataset from 46,216 regions to 40,265 UTRs. Next, we filtered for genes with multiple UTRs, which revealed that out of a total of 18,546 genes, 16,217 genes contain

multiple UTRs. To analyze how many of these genes have regional UTR constraint we made use of the previously introduced decile method. We defined a gene as regionally UTR constrained if it contains multiple UTRs which are in separated deciles and delimited by at least two other deciles. Applying this definition 12,610 of the genes contained in the mane select filtered UTR 2.0 database fulfilled this criterion for regional UTR constrain. This is partially explainable by large base length differences observed across the UTRs, as OEUF values are influenced by base sequence length [18]. Thus, a correlation analysis showed a strong correlation between the SNVOEUF score and base sequence length (Kendall Rank Correlation Coefficient=-0.46, $p < 0.001$). Being limited by the sample size of gnomAD for small UTRs and to further adjust the analysis for sequence length, we repeated the previous analysis on a sequence length filtered UTR subset, by filtering the 40,265 UTRs for UTRs with a minimum length of 800 bases which corresponds to the mean UTR sequence length of the filtered 40,633 UTRs. Applying these filter criteria, 203 genes have multiple UTR sequences and 107 of these 203 genes have a regional UTR constraint following the definition above. Highlighting the potential relevance of constraint values that incorporate the genomic context of UTRs.

## Methods

### Annotation of sORF encoded variants

We utilized the custom annotations feature provided by the ensembl variant effect predictor (VEP) [25] to annotate the gnomAD chromosome reference VCF files for functional consequences in the sORF reading frame. To do so we created a gene transfer format (gtf) file in which we defined the sORF reading frame, by designating the sORF regions from the gencode ribo-seq orf bed file [5] as gene regions and the corresponding block regions as exons. Next to an annotation for functional consequence the general population frequency from gnomAD 4.0 exomes or genomes was added.

### Constraint evaluation of sORF encoded variants

For the constraint evaluation of sORF encoding variants, we calculate the observed/- expected upper bound fraction (OEUF) for different variant types and name them in correspondence with their variant type (MOEUF for missense variants, LOEUF for loss-of-function variants and SNVOEUF for not further divided SNVs). In brevity, the number of uniquely observed protein truncating variants is compared with the number of expected variants for a given mutation type calculated by a mutational model which assumes a neutral effect of these variants.

For the calculation of observed variants, we followed the recommendations used in the assembly of the gnomAD constraint scores [18]. This means that we annotated the gnomAD 4.0 release respecting the sORF context utilizing VEP, then filtered it for the number of unique variants for the mutation type of interest. We only included variants with a general gnomAD frequency less than 0.1% that passed all filters and had a median depth greater or equal to 1. For all constraint calculations we only considered single nucleotide variants, therefore for loss–of-function variants we only considered start-loss, stop-loss, and stop-gain variants. As a consequence of using VEP [25] for annotation, splice variants are predicted. Considering the limited known information of sORFs and therefore the uncertainty about the occurrence of possible splicing mechanisms in sORFs, we decided to discard variants that were only predicted as splicing variants. This exclusion applied to all analysis, including the loss-of-function analysis. To calculate the number of expected variants we at first calculated the number of possible variants for which we followed the protocol defined by Karczewski et al. [18]. In short, we estimate the number of expected SNVs, missense and synonymous variants by utilizing the mutation rates published in the paper by Chen et al. [20], where the Gnocchi score was proposed. Instead of relying on the Hail-framework previously introduced by Karczewski et al. [18] we reimplemented the workflow by means of Python and Spark. For this, we first extracted the base sequence of the gencode sORFs using Biopython [26]. Subsequently, we parsed the sequences and calculated the number of possible variants, for the class of interest, by iterating over each coding triplet and calculating the relevant context triplet for each base in the analyzed triplet. Adhering to the protocol set by Karczewski et al. [18] we reduced the number of possible variants. Variants were excluded if they originated from bases with a low-quality variant in the gnomAD data, had a high allele frequency (greater or equal to 0.1%) or fell into a region with insufficient coverage (mean depth less than 1). We additionally calculated a sORF mutation rate by summing the mutational rate of each individual base with its corresponding context triplet and the corresponding methylation status. Figure 8 illustrates the workflow schematically.

For the final OEUF calculations, we computed the 90% confidence interval (between the 5th and 95th percentiles) for an expected value of a Poisson distribution given an observed count and an expected count.

The mathematical formulation of the Poisson probability mass function (pmf) is:

$$G(X = k) = \frac{\lambda^{k} * e^{-\lambda}}{k!}$$

where:

- $\lambda$ is the average rate (expected value).

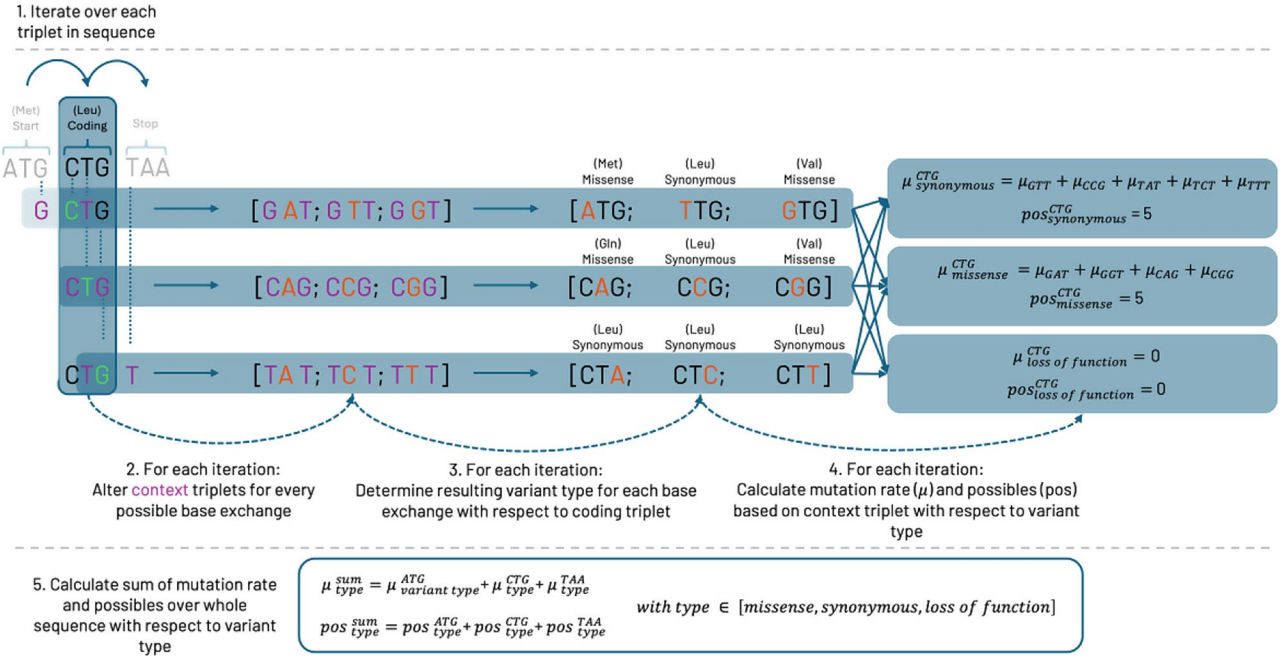


**Fig. 8** presents a detailed workflow for the calculation of possible variants and mutation rates for a given sequence, based on the type of variant. The iterative process is represented by a sample sequence composed of three codons: the start codon ATG, the triplet CTG, and the stop codon TAA (1). To further elucidate the process, the calculation of the possible variants and mutation rate for the CTG triplet is emphasized. The first step involves identifying the context triplets for each of the three bases (highlighted in green) of the coding triplet. This is done by considering their direct neighboring bases (shown in purple). Afterwards, both the coding and the context triplet are modified to create all possible single nucleotide variants. This is achieved by replacing the bases of the coding triplet (in green) with all other possible bases, resulting in a total of 9 altered triplets (2). Subsequently, the consequence, and thus the variant type, of each altered triplet is determined by evaluating its impact on amino acid translation (3). The mutation rate for a specific variant type of the whole codon is calculated by summing up the mutation rates of the corresponding context triplets (4). Complementary, the count of possible variants for a given variant type is simply the sum of triplets that align with that variant type (4). Finally, the mutation rate and total count of possible variants for the entire sequence are computed by summing up the mutation rates and possible variant counts, respectively, for each variant type across all codons in the sequence (5)

- k is the actual number of events (observed value).
- e is the base of the natural logarithm.
- $G(X = k)$ is the probability of k events occurring in an interval.

Here the Poisson pmf is calculated for a range of values from 0 to 2 with steps of 0.001 so that:

$$\lambda = \lambda\prime * x$$

where:

- x is the range value with:

$$\{x: x\varepsilon\ [0,\ 2],\ with\ x = n\ *\ 0.0001, where\ n\ \varepsilon\ N_0\ and\ 0 <= n <= 2000\}$$

- $\lambda$ is the average rate (expected value).

The lower and upper bounds of the expected value are then calculated as the first values in the cumulative distribution where the normalized values are greater than or equal to 0.05 and 0.95 respectively.

The following is a rough mathematical representation of the process:

1. Calculate Poisson pmf for each $\lambda\prime$ for a given expected and observed count
2. Calculate the cumulative sum of these Poisson pmf values.
3. Normalise these values.
4. Lower Bound = min( $\lambda\prime$ | Normalised Cumulative Poisson pmf > = 0.05)
5. Upper Bound = min( $\lambda\prime$ | Normalised Cumulative Poisson pmf > = 0.95)

### Comparison of OEUF values and defining constraint

OEUF values represent a continuous value which are best analyzed as a spectrum, however introducing cut-off values for downstream analysis can be helpful. Therefore, following the protocol by Karczewski et al. [18] we binned the resulting OEUF distributions into 10 equally sized bins and took the maximum OEUF value of the most constrained bin as a cut-off value for highly constrained items in this decile. Using this cut-off value we

further compared sORFs and their OEUF values to other genomic elements.

#### Matching with Gnocchi score

In brevity, the Gnocchi score is a constraint metric for which the calculation of the whole genome data from gnomAD 4.0 (76,215 individuals) was analyzed [20].For this the genome was fragmented into 1,000 base spanning intervals. Using a context dependent mutational model for each interval the number of expected variants was calculated and compared to the number of rare SNVs. The resulting ratio was transformed into a z-score and since this score was calculated for every arbitrary region, we paired it with the coordinates of the gencode sORFs located on autosomes [20]. Since not all sORFs might encode stable microproteins we wanted to start our analysis with a constraint score in comparison with three non-coding regions. For this we used the Supplemen- tary Dataset 3 from the recently published Gnocchi score publication [20]. This dataset provides genome wide scores at 1 kilobase resolution, calculated iteratively by sliding 100 base pairs. By means of Python the gencode sORFs, the snoDB 2.0 [27] snoRNAs, gencode [28] miRNAs and gencode [28] lncRNAs were paired with Gnocchi scores. To be more precise, we located all Gnocchi score Intervals that fell between the start and end coordinates of the genomic element of interest for all autosomes and then the averaged the Gnocchi scores over these intervals. If no proper interval was found containing the genomic element, the two closest Gnocchi score intervals were used to calculate an average Gnocchi score. To have a more direct comparison between the Gnocchi score and the SNVOEUF we subsequently repeated our analysis, by using the non-overlapping Gnocchi dataset (Supplementary Dataset 2 from the Gnocchi score publication [20]) to match the Gnocchi score with sORFs which fall into exactly one Gnocchi score interval. Not mappable sORFs where discarded in this repeated analysis.

#### Constraint analysis of UTR variants

UTR regions for each MANE select transcript were selected from the UTR 2.0 database [24]. We calculated the number of variants in gnomAD 4.0 within said UTR regions and compared them to the number of variants within sORFs located in said UTR region. This was carried out to estimate whether UTR regions containing the sORFs are subjected to a higher degree of selection. The underlying idea is that an UTR containing regulatory elements for the canonical CDS in addition to sORFs which might act as further regulatory elements or encode functional distinct micropeptides, should be subject to a stronger selection than an UTR only fulfilling regulatory purposes. To estimate this effect, we calculated the single nucleotide variant observed/- expected upper bound of these UTRs. To analyze a possible correlation between these two values we calculated the Kendall Rank Correlation Coefficient.

#### Constraint analysis of canonical genes and comparison to sORFs

For the analysis of canonical genes, we filtered the comprehensive VCFs for the current gnomAD 4.0 release for MANE select transcripts using Spark for observed variants with the above-mentioned criteria. We calculated the constraint values in the same fashion as described above and paired the sORFs with their neighbored genes. To analyze a possible correlation between these two values we calculated the Kendall Rank Correlation Coefficient.

#### Filtering gnomAD exomes for covered sORFs

Utilizing the allele sites published by gnomAD for the exomic regions and the depth summary, we removed possible variants which were insufficiently covered. For this we filtered out sORFs which were not completely contained in the exomic regions file. Additionally, following the gnomAD flagship paper we discarded regions in our analysis that showed a mean coverage of less than 1.

#### Creation of a whole genome methylation map

In a comparable approach to the gnomAD flagship paper, we obtained the bisulfite whole-genome data provided by the NIH Roadmap Epigenomics Consortium [29]. For each genomic position we averaged the methylation fraction across the provided 37 epigenomes from different tissue types and developmental time periods. Afterwards we performed a liftover to hg38 coordinates, using the UCSC chain files and the Python package liftover. Since we used the mutation rates published with the recent Gnocchi constraint score paper, we followingly binned the averaged methylation fractions into 16 bins between 0 and 1. This resulted in the following methylation fraction bins: [0-0.0625, 0.0625–0.125, 0.125–0.1875, 0.1875-0.25, 0.25-0.3125, 0.3125–0.375, 0.375–0.4375, 0.4375-0.5, 0.5-0.5625, 0.5625–0.625, 0.625–0.6875, 0.6875-0.75, 0.75-0.8125, 0.8125–0.875, 0.875–0.9375, 0.9375-1]. We used the corresponding bin of each genomic position to decide which mutation rate to select from the precalculated mutation rates [20] at potentially methylated regions.

#### Matching codings regions with the zoonomia phylop score

We obtained the latest phyloP values from the most recent Zoonomia [23] manuscript by Christmas et al. and paired them to the coding blocks of the whole gencode sORF dataset and their neighbored genes. For the genes we obtained the coding blocks from the most recent Ensembl Release GTF file. The paired values were

averaged using the groupby and mean function from Pandas.

### Statistics

Most statistical tests were performed using the Python library SciPy [30]. For the calculation of the Kendall Rank Correlation Coefficient [31] and the corresponding p-value the SciPy [30] function kendalltau was used. To test the equality of different OEUF distributions a two sample Kolmogorov-Smirnov [32] was performed using the ks 2samp function. To perform the comparison of the mean ranks of different OEUF distribution a Mann-Whitney-U test [33, 34] was performed using Scipys [30] mannwhitneyu function. The mannwhitneyu function from SciPy [30] does not return the calculated mean ranks of the two groups, therefore we utilised the rank function from the Python package pandas to rank the two groups in a concatenated dataframe. Subsequently we summed the ranks of the two groups and calculated an average. For the Mann-Whitney-U test we filtered the sORF dataset for uORFs, dORFs and lncRNA-ORFs to avoid dependencies which would result from using overlapping sORF and coding genomic regions (as present in intORFs, uoORFs and doORFs) for constraint calculation. For the comparison between the sORF MOEUF values and their neighbored gene MOEUF values a Wilcoxon signed ranked test [33, 34] was performed using SciPy [30] and the Wilcoxon function. Multiple hypothesis tests and correlation analysis resulted in p-values not distinguishable from 0 by scipy and R [35]. These values and p values below 0.001 were reported as p value < 0.001 as suggested by multiple scientific organizations (e.g. the APA). Following the suggestions by Lin, Lucas and Shmueli [36] for statistical inference in large and complex datasets, we accompanied most of our hypothesis tests with visualizations and analysis for effect sizes. We utilized the R package effsize [37] to compute the Vargha and Delaney A measurement [21], provided by the VD. A function. The Vargha and Delaney measurement can be described as the probability that a value from group one will be greater than a value from group two [21], thereby allowing a further quantification of the difference between the two groups.

### Data processing with databricks and spark

The gnomAD 4.0 VCFs underwent comprehensive processing, including the preparation of positional whitelists, the filtration of designated genomic regions such as MANE transcripts, and observed variant count. This processing was conducted using Databricks and Azure Cloud. The employed Databricks runtime was 14.3 LTS, which encompasses Apache Spark 3.5.0 and Scala 2.12. The computational infrastructure utilized was a multi-node E48d V4 Cluster, equipped with 384 GB memory and 48 cores per node.

## Discussion

Research on sORFs is an emerging field and the discussion of their function in the genome and their biological role is still ongoing. Here we tested different approaches to prioritize and evaluate variants that affect sORFs and neighbored regions, which is fundamental to understanding their role in health and disease.

### sORFs show a similar mutational background to coding regions

sORFs show a similar mutational background to canonical genes, yet they can contain a higher number of high impact variants. This can have multiple explanations. It might be that these regions are not intolerant against loss-of-function variants or that these non-constrained sORFs do not encode functional microproteins. This similarity in distribution, as seen in Fig. 1, on its own, does not bring sufficient evidence for a potential coding effect or conservation in sORFs, because the distribution might be fully explainable from a probabilistic standpoint, since synonymous and protein truncating variants have less opportunities to occur compared to missense variants. Followingly more complex workflows are required that respect the genomic context and the different effect of variants in differing reading frames.

### A streamlined context dependent genomic constraint workflow

We put one of these more complex workflows to the test, by analyzing OEUF values for the effect of sequence length, alternating reading frames in overlapping regions and different genomic contexts. We provide evidence for the context sensitivity of these scores, therefore highlighting the importance of carefully mapped genomic constraint maps and the need for clearly defined genomic regions. We further strengthen this point by showing the difference in genomic constraint between sORFs and their closest genomic neighbors and transferring mapped OEUF values onto other regulatory regions, such as UTRs. Thereby we present further evidence for regional constraint differences, even in related regions. This is a concept which is in the process of being adapted in the analysis of coding regions, where it already has been demonstrated that the interpretation of some genes benefits from regional constraint values in comparison to gene wide constraint values [38, 39]. For this, we sought to establish a more simplified method in comparison to the gnomAD Hail approach. This streamlined procedure enhances accessibility, particularly for Python-native developers who may not be as well-versed with the Hail framework. It empowers the wider community

to calculate constraint metrics for their specific genomic elements of interest. While this approach has been specifically adapted for the use case presented here, it is important to acknowledge that the Hail framework offers a far more extensive range of capabilities.

### Examining the explanatory power of OEUF scores from different perspectives

While OEUF values are a useful and versatile tool, their limitations need to be kept in mind, especially when comparing OEUF values from multiple regions to one another. OEUF scores are a one-sided conservative estimation. Low OEUF scores hint towards a higher level of constraint, while high OEUF scores can either depict a region which is only weakly constrained or signal a small sample size with limited explanatory power. The limited explanatory power is a present issue, even today with datasets like gnomAD. Our analysis of the mutational background in gnomAD and the calculation of the different constraint metrics demonstrate that the currently available sample size in gnomAD is insufficient to fully analyze the constraint level of sORFs. Especially the analysis of loss-of-function intolerance requires sample sizes which are by a magnitude larger than the current gnomAD release. As previously stated, small genomic regions suffer the most from the limited sample size since they already have a small number of mutational opportunities. Present metrics like the Gnocchi score try to reduce this problem by binning the genome in relatively large regions of 1,000 bases which are not mapped to the exact genomic architecture. We demonstrate that the OEUF value of SNVs in sORFs do not correlate well with the matched Gnocchi score of the sORFs. This might be explained by several possibilities. The first possibility is that the Gnocchi score bins relatively large regions of interest. sORFs only make up a small portion of these intervals and tend to fall within regions overlapping with canonical regions and regulatory elements. Consequently, sORFs might just be another factor under many more that influences the Gnocchi score. Highlighting the importance of context sensitive genomic constraint maps and highlighting that the Gnocchi score might benefit from a higher resolution, a suggestion also mentioned in the original Gnocchi score paper [20]. On the other hand, our SNVOEUF calculation might also suffer from the short sample size. Large sample sizes could reveal a correlation between the Gnocchi score and the SNVOEUF which might be masked by the fact that the sORF regions are just smaller and therefore tend to be more influenced by the small sample size, than the larger gnocchi score intervals. Thereby highlighting the interaction of OEUF values, genomic context and sequence length and the need for expanding the presented workflow to larger datasets than gnomAD 4.0.

### A subset of the sORF Gencode dataset is highly constrained

While the Gnocchi score and the SNVOEUF score can give a general overview of the constraint, additional metrics are required to analyze the coding potential of sORFs. Our analysis using the gnomAD genomes revealed that a portion of sORFs has a similar constraint to highly constrained genes with a similar OEUF value. While repeating this analysis with gnomAD exome data, we interestingly noticed a shift in the MOEUF distribution towards a smaller MOEUF value. We were surprised to see that the number of highly constrained sORFs increased. This could indicate that the current sample size for the gnomAD genomes might still be insufficient with respect to sORFs. As a result, this could potentially even lead to an underestimate of constraint in current datasets. These highly constrained sORFs, although a minority, might be of special interest in terms of potential clinical relevance since they show a similar constraint to the most constrained coding regions of the canonical genes. To our surprise intORFs were a minority in these highly constrained regions, which might highlight that these highly constrained regions are not purely ranked as highly constrained because of an overlap with canonical coding regions. To further explore the constraint of the sORFs, especially sORFs overlapping with other genomic regions, we suggest that an additional insight can be obtained by comparing the MOEUF values of genomic regions to their SNVOEUF value. Given a reading frame that is intolerant to missense variation it is to be expected that it has a lower number of total missense variants, than synonymous variants. If this reading frame is observed from an out of frame perspective this ratio might be different, because variants might be falsely interpreted as missense variants or vice versa. Therefore, we hypothesize that sORFs predicted to be more constrained against missense variants compared to the totality of SNVs are indeed more likely to be constrained against missense variants. This is because the predicted intolerance metrics accurately reflect the balance of missense and general SNVs observed in a genuine reading frame. After performing our analysis, we observe this pattern for the majority of sORFs from the consensus dataset, when analyzed using the gnomAD exome data. When the gnomAD genome data is used, the picture becomes less clear. We argument that this is due to the large sample size difference between those two datasets, since missense variants have less opportunities to occur, when compared to the totality of possibilities of SNVs, which is in turn influenced by the sample size. These highly constrained sORFs, which also show this pattern of MOEUF/SNVOEUF relationship, are prime candidates for experimental studies. To further explore their possible significance, we analyzed a recent preprint by

Deutsch et al. [40] in which the authors categorize this consensus dataset based on experimental evidence. 47 of the highly constrained sORFs from the exome dataset have an evidence class assigned to them. 13 of these are ranked as Tier 1B, one as Tier 2 A, 25 as Tier 2 B, 7 as Tier 3, 1 as Tier 2 A and one as Tier 4. For the genome date, 37 of the highly constrained sORFs are assigned an evidence class. Of these 37, 11 are assigned a Tier 1B ranking, 20 as Tier 2B, one as a Tier 2 A and 5 as Tier 3.

### Comparing the constraint of sORFs to the constraint of their genomic neighbors

When compared to the neighboring genes, which is used in the nomenclature of the gencode sORF set, we were not able to witness a correlation for both, gnomAD genome and exome data. This might be partially explained by the fact that the gnomAD constraint values are currently calculated gene wide. Some of these genes might have regional constraints, which will only be uncovered if a regional constraint calculation (exon wide or domain wide) is performed. To fully understand the connection between sORFs and neighboring genes it would be beneficial to analyze the sORFs and their closest neighboring gene element, annotated with regional constraint. This pattern of moderate constraint was continued when we compared uORFs with the constraint scores of UTR regions. This highlighted that uORFs evolved across a broad spectrum of constrained UTRs but were not enriched in highly constrained UTRs. Taken together we hypothesize that this moderately constrained background of genomic regions might have been a necessary condition for sORFs to evolve. This is supported by literature [5] as some sORFs are described as phylogenetically young and this background likely provides a balance between room for some genomic change and constraint.

### Comparing evolutionary conservation and constraint-based methods

There are multiple approaches for measuring sequence conservation. Common methods include evolutionary conservation metrics and population-based constraint approaches. Evolutionary conservation metrics involve aligning genomes from multiple specifies, followed by a sequence similarity comparison across the aligned genomes. The evolutionary background and developmental history of sORFs has been investigated at length previously [15–17]. Sandmann et al. [17] demonstrated that the majority (around 90%) of sORFs contained in the gencode catalogue can be classified as evolutionary young, while the remaining 10% was classified as evolutionary conserved across non-primate-mammals. They further identified a small subset (222/7264) of human specific sORFs. One limitation of evolutionary conservation-based approaches is the further classification of recently developed genomic regions. This limitation can be reduced by population-based constraint approaches, which are focused on the distribution of variants inside a given population. Therefore, to extend existing prioritization approaches, we investigated the constraint of sORFs in the general population, by implementing a workflow, which can be used standalone or can be combined with conservation-based approaches. To explore this further, we analyzed the phyloP score for the highly constrained sORFs from the exome data and subsequently in comparison also for their neighbored genes. The highly constrained sORFs have a mean phyloP value of 1.9 which is far from the median exome wide phyloP value of 4.9, as established by Christmas et al. [23]. However, this low phyloP values has to be interpreted under the previously introduced limitations. When those highly constrained sORFs are analyzed, which are neighbored by genes with already established phenotypes, as depicted in Table 1, we observed a mean phyloP value of 3.10. This is not really surprising, as we see a lot of associated GO-Terms in these genes regarding the highly conserved Notch-Pathway [41] and genes involved with neurodevelopmental disorders, which in the literature also have been described as more conserved when compared to other genes [42]. An interesting direction, which could be explored in further research is to analyze whether these sORFs neighbored to genes, already mapped to clinical relevance, might function as regulatory elements of these regions, which is increasingly investigated in some disease types, especially in cancer [43–45].

## Conclusion

We implemented a constraint calculation workflow in Python and Spark, which compares the number of expected genetic variants to the number of observed variants for different variant classes and genomic regions. Utilizing this workflow, we calculated constraint metrics for a consensus set of sORFs. We compared these computed metrics to the already established Gnocchi score and to constraints of the genomic neighbors of sORFs. We demonstrate that it might be beneficial to calculate tailored constraint value for the sORFs, instead of relying on larger binned intervals. Furthermore, we provide evidence that there is only a weak connection between the constraint of sORFs and their genomic neighbors, signifying the benefit of a targeted constraint approach. Finally, we highlight that a small subset of sORFs is constrained in a similar way to highly constrained canonical genes, highlighting the need for further research in this emerging field.

#### Abbreviations

| | |
|---|---|
| GO-Term | Gene-Ontology Term |
| GTF | Gene Transfer Format |
| KDE | Kernel Density Estimate |

| | |
|---|---|
| lncRNA | Long Non-Coding Ribonucleic Acid |
| LOEUF | Loss-of-function Observed/Expected Upper Bound Fraction |
| MOEUF | Missense Observed/Expected Upper Bound Fraction |
| OEUF | Observed/Expected Upper Bound Fraction |
| SNV | Single-Nucleotide-Variant |
| SNVOEUF | Single-Nucleotide-Variant Observed/Expected Upper Bound Fraction |
| sORF | Short Open Reading Frame |
| snoRNA | Small Nucleolar Ribonucleic Acid |
| UTR | Untranslated Region |
| VEP | Ensembl Variant Effect Predictor |

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11444-w.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Supplementary Material 7

**Acknowledgements**
We would like to thank Jan Höllmer and Lars Perchalla whose support in the early stages was crucial to the inception of this project and the research outcomes presented in this report.

**Author contributions**
JK, IK, ME, MB and FK conceived the Idea. JK and MD implemented the constraint workflow in Python and Spark. JK and MD parsed gnomAD, the gencode riboseq ORFs and the UTR database. JK and MD computed the OEUF metrics, the com- parison to the canonical OEUF values and created the plots. JK, MD and MB wrote the manuscript. JK and MD wrote tests and performed validations. IK, ME, MB, FK and JK provided supervision. All authors were involved in reviewing of the transcript and gave their consent for the submission of the final version.

**Funding**
Open Access funding enabled and organized by Projekt DEAL.
This research project was funded by the START-Program of the Faculty of Medicine, Uniklinik RWTH-Aachen. I.K. is supported by the Deutsche Forschungsgemeinschaft (DFG) (KU 1587/6 – 1, KU 1587/9 – 1, KU 1587/10 – 1, KU 1587/11 – 1). I.K. receives funding from the European Union's Horizon 2020 research and innovation programme under the EJP RD COFUND-EJP N° 825575.

**Data availability**
The predicted constraint values from the genome and exome data for the sORFs are available as supplementary dataset accompanying this article. The gnomAD constraint table and the gnomAD VCFs are publicly available on the gnomAD website https://gnomad.broadinstitute.org/downloads. The mutation rates were extracted from the supplementary datasets of gnomAD gnocchi Score paper [17]. The epigenomes are publicly available on the NIH epigenomics website https://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/. The gencode sORFs were taken from the corresponding project website https://www.gencodegenes.org/pages/riboseqorfs/. The snoRNAs were taken from the snoDB website https://bioinfo-scottgroup.med.usherbrooke.ca/snoDB/. The UTR data was obtained from the corresponding github page https://github.com/BioinfoUNIBA/UTRdb. The SNVOEUF and MOEUF values computed for the sufficiently powered genes is provided as supplementary dataset (1) The paired Gnocchi values with the different genomic elements is provided as supplementary dataset (2) The results from the UTR analysis are provided as supplementary dataset (3) The mapped phenotypes to genes neighbored by highly constrained sORFs is provided as dataset (4) The results from the GO Term analysis are provided as supplementary dataset (5) The Python code used to calculate the OEUF scores is available in the following public GitHub repository: https://github.com/IHGGM-Aachen/genomic-constraint-calculation.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare no competing interests.

Received: 5 August 2024 / Accepted: 4 March 2025
Published online: 14 March 2025

**References**

1. Shendure J, Findlay GM, Snyder MW. Genomic Medicine-Progress, pitfalls, and promise. Cell. 2019;177:45–57.
2. Hirsch N, Birnbaum RY. Dual function of DNA sequences: Protein-Coding sequences function as transcriptional enhancers. Perspect Biol Med. 2015;58:182–95.
3. Blanco-Melo D, et al. A novel approach to exploring the dark genome and its application to mapping of the vertebrate virus fossil record. Genome Biol. 2024;25:120.
4. Kafita D, Nkhoma P, Dzobo K, Sinkala M. Shedding light on the dark genome: insights into the genetic, CRISPR-based, and Pharmacological dependencies of human cancers and disease aggressiveness. PLoS ONE. 2023;18:e0296029.
5. Mudge JM, et al. Standardized annotation of translated open reading frames. Nat Biotechnol. 2022;40:994–9.
6. Leblanc S, et al. OpenProt 2.0 builds a path to the functional characterization of alternative proteins. Nucleic Acids Res. 2024;52:D522–8.
7. Cheng H, et al. Small open reading frames: current prediction techniques and future prospect. Curr Protein Pept Sci. 2011;12:503–7.
8. Choteau SA, Wagner A, Pierre P, Spinelli L, Brun C. MetamORF: a repository of unique short open reading frames identified by both experimental and computational approaches for gene and metagene analyses. Database J. Biol. Databases Curation. 2021:baab032.
9. Chen J, et al. Pervasive functional translation of noncanonical human open reading frames. Science. 2020;367:1140–6.
10. Neville MDC, et al. A platform for curated products from novel open reading frames prompts reinterpretation of disease variants. Genome Res. 2021;31:327–36.
11. Li Y, et al. SmProt: A reliable repository with comprehensive annotation of small proteins identified from ribosome profiling. Genomics Proteom Bioinf. 2021;19:602–10.
12. Hofman DA, et al. Translation of non-canonical open reading frames as a cancer cell survival mechanism in childhood Medulloblastoma. BioRxiv Prepr Serv Biol. 2023. https://doi.org/10.1101/2023.05.04.539399. 2023.05.04.539399.
13. Ge Q, et al. Micropeptide ASAP encoded by LINC00467 promotes colorectal cancer progression by directly modulating ATP synthase activity. J Clin Invest. 2021;131:e152911.
14. Jain N, Richter F, Adzhubei I, Sharp AJ, Gelb B. D. Small open reading frames: a comparative genetics approach to validation. BMC Genomics. 2023;24:226.
15. Bazzini AA, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. 2014;33:981–93.
16. Chothani SP, et al. A high-resolution map of human RNA translation. Mol Cell. 2022;82:2885–e28998.
17. Sandmann C-L et al. Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. Mol. Cell. 2023;83:994–1011.e18.
18. Karczewski KJ, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
19. Gudmundsson S, et al. Variant interpretation using population databases: lessons from GnomAD. Hum Mutat. 2022;43:1012–30.

20. Chen S, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2024;625:92–100.
21. Vargha A, Delaney HD. A critique and improvement of the 'CL' common Language effect size statistics of mcgraw and Wong. J Educ Behav Stat. 2000;25:101–32.
22. Harrison PW, et al. Ensembl 2024. Nucleic Acids Res. 2024;52:D891–9.
23. Christmas MJ, et al. Evolutionary constraint and innovation across hundreds of placental mammals. Science. 2023;380:eabn3943.
24. Lo Giudice C, et al. UTRdb 2.0: a comprehensive, expert curated catalog of eukaryotic mRNAs untranslated regions. Nucleic Acids Res. 2023;51:D337–44.
25. McLaren W, et al. The ensembl variant effect predictor. Genome Biol. 2016;17:122.
26. Cock PJA, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.
27. Bergeron D, et al. SnoDB 2.0: an enhanced interactive database, specializing in human snornas. Nucleic Acids Res. 2023;51:D291–6.
28. Frankish A, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47:D766–73.
29. Kundaje A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.
30. Virtanen P, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72.
31. Kendall MG. A new measure of rank correlation. Biometrika. 1938;30:81–93.
32. Massey FJ. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc. 1951;46:68–78.
33. Wilcoxon F. Individual comparisons by ranking methods. Biom Bull. 1945;1:80–3.
34. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat. 1947;18:50–60.
35. Core Team R. R and others. R: A language and environment for statistical computing. (2013).
36. Lin M, Lucas HC, Shmueli G. Research commentary: too big to fail: large samples and the p-Value problem. Inf Syst Res. 2013;24:906–17.
37. Torchiano M. Effsize - a package for efficient effect size computation. Zenodo. 2016. https://doi.org/10.5281/ZENODO.1480624.
38. Samocha KE et al. Regional missense constraint improves variant deleteriousness prediction. 148353 Preprint at https://doi.org/10.1101/148353 (2017).
39. Zhang X, et al. Genetic constraint at single amino acid resolution in protein domains improves missense variant prioritisation and gene discovery. Genome Med. 2024;16:88.
40. Deutsch EW et al. High-quality peptide evidence for annotating non-canonical open reading frames as human proteins. 2024.09.09.612016 Preprint at https://doi.org/10.1101/2024.09.09.612016
41. Bray SJ. Notch signalling: a simple pathway becomes complex. Nat Rev Mol Cell Biol. 2006;7:678–89.
42. Betancur C, Buxbaum JD. Gene constraint and genotype-phenotype correlations in neurodevelopmental disorders. Curr Opin Genet Dev. 2020;65:69–75.
43. Zhong Z, Li Y, Sun Q, Chen D. Tiny but mighty: diverse functions of uORFs that regulate gene expression. Comput Struct Biotechnol J. 2024;23:3771–9.
44. Dasgupta A, Prensner JR. Upstream open reading frames: new players in the landscape of cancer gene regulation. NAR Cancer. 2024;6:zcae023.
45. Silva J, Fernandes R, Romão L. Translational regulation by upstream open reading frames and human diseases. Adv Exp Med Biol. 2019;1157:99–116.


## Publisher's note