Translational plasticity facilitates the accumulation of nonsense genetic variants in the human population

Sujatha Jagannathan1,2,3 and Robert K. Bradley1,2

1Computational Biology Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA; 2Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA; 3Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA

Genetic variants that disrupt protein-coding DNA are ubiquitous in the human population, with about 100 such loss-of-function variants per individual. While most loss-of-function variants are rare, a subset have risen to high frequency and occur in a homozygous state in healthy individuals. It is unknown why these common variants are well tolerated, even though some affect essential genes implicated in Mendelian disease. Here, we combine genomic, proteomic, and biochemical data to demonstrate that many common nonsense variants do not ablate protein production from their host genes. We provide computational and experimental evidence for diverse mechanisms of gene rescue, including alternative splicing, stop codon readthrough, alternative translation initiation, and C-terminal truncation. Our results suggest a molecular explanation for the mild fitness costs of many common nonsense variants and indicate that translational plasticity plays a prominent role in shaping human genetic diversity.

[Supplemental material is available for this article.]
production using a combination of genome-wide data and directed biochemical experiments.

**Results**

**Many common nonsense variants do not disrupt protein production**

We set out to test the hypothesis that many common loss-of-function variants do not ablate protein production from their host genes, thereby explaining the apparently mild phenotypic consequences of these variants even when found in a homozygous state. To test this hypothesis, we took advantage of several recently published data sets representing successive stages of gene expression. These data sets consist of genome-wide measurements of mRNA levels by RNA-seq (Lappalainen et al. 2013), mRNA:ribosome association (Battle et al. 2015), and protein abundance (Battle et al. 2015) in lymphoblastoid cell lines (LCLs) that were established by the HapMap and 1000 Genomes Projects (International HapMap 3 Consortium 2010; 1000 Genomes Project Consortium 2012). For each data set, we used published sample genotypes to identify SNVs that induced synonymous or nonsense codon changes with respect to the reference genome in one or more RefSeq coding transcripts.
We specifically focused our analysis on common nonsense variants that were present in a homozygous state and thereby predicted to induce complete gene knockouts. We focused on nonsense variants rather than the broader spectrum of loss-of-function variation because nonsense variants can be identified in a straightforward manner from both DNA and RNA sequencing, their impact upon the mRNA is unambiguous, they are relevant to many Mendelian diseases, and their allelic expression can be reliably quantified from RNA-seq data. Of all nonsense variants that we identified, a total of 194 (RNA-seq), 118 (ribosome profiling), and 106 (mass spectrometry) nonsense variants were present in a homozygous state in the genome of at least one assayed sample in each respective data set (Fig. 1B).

By studying individuals with either two (genotype 0/0; 0 = reference allele) or zero (genotype 1/1; 1 = alternate allele) intact gene copies, we were able to unambiguously determine how each nonsense variant affected protein production. Restricting to those variants, we found that levels of the parent mRNAs containing each nonsense variant were similar in 0/0 and 1/1 individuals (Fig. 1C). The similarity of mRNA levels in 0/0 and 1/1 individuals suggested that many of the mRNAs containing the nonsense variants studied here escaped from nonsense-mediated decay (NMD). To determine the effect of NMD on these variants, we quantified allele-specific expression (ASE) of the alternate alleles illustrated in Figure 1C in heterozygous (genotype 0/1) individuals. We restricted our analysis to nonsense variants covered by at least 20 RNA-seq reads in at least one heterozygous sample (so that we could accurately measure ASE) and computed the median ASE across all such heterozygous samples for each variant. The 36 nonsense variants meeting this coverage requirement exhibited a median ASE of 45%, only modestly lower than the 50% expected if sense variants meeting this coverage requirement exhibited a median ASE < 10%, consistent with the observed allelic expression, removing those inconsistent data points did not affect our conclusion that neither mRNA expression, mRNA:ribosome association, nor protein expression was abolished by the analyzed homozygous nonsense variants (Supplemental Fig. S1B–D). Therefore, genotyping errors do not explain why these nonsense variants do not abolish protein production.

Permissive RNA processing may enable protein production from genes containing nonsense variants

The extensive regulation that occurs between the initiation of gene transcription and the completion of mRNA translation offers ample opportunities for protein production from genes containing nonsense variants (although gene function may or may not be preserved). Protein can be produced from genes containing nonsense variants by one of two means (Fig. 1F). First, a nonsense variant may be isoform specific, such that at least one coding isoform of the parent gene does not contain the variant. In this case, the nonsense variant may prevent translation of one or more, but not all, isoforms. Second, the stop codon introduced by a nonsense variant may not be sufficient to abolish productive translation. This can occur if the ribosome can initiate translation downstream from the stop codon, if the stop codon is subject to readthrough, or if the nonsense variant is near the normal stop codon.

Sequence analyses reported in previous studies suggest that both alternative splicing and protein truncation may contribute to the robust protein production from genes containing nonsense variants that we observed (Fig. 1C–E). Approximately one-third of previously studied nonsense variants are isoform specific (MacArthur et al. 2012; Kaiser et al. 2015); minor coding isoforms are enriched for nonsense variants relative to major isoforms (Liu and Lin 2015); and nonsense variants preferentially occur at the 3′ ends of their host CDSs (MacArthur et al. 2012; Sulem et al. 2015). We therefore set out to systematically identify molecular mechanisms that enabled protein production from genes containing nonsense variants. Our general strategy was to combine RNA-seq, ribosome profiling, and SILAC data to identify the mechanisms that enabled protein production from genes containing the common nonsense variants studied in Figure 1, and then computationally predict the likely relevance of each mechanism to lower-frequency nonsense variants using sequence analysis of the approximately 10,000 nonsense variants identified by the 1000 Genomes Project.

Nonsense variants are frequently isoform specific and removed by alternative splicing

We first tested whether nonsense variants are more frequently isoform specific than their synonymous counterparts. For each nonsense variant within a multiexon gene, we determined whether the gene contained at least one RefSeq coding transcript for which
the alternate allele did not introduce an in-frame stop codon or whether the nonsense variant was contained within alternatively spliced sequence of the mRNA. We similarly measured isoform specificity for synonymous variants, restricted to the subset of synonymous variants lying within genes that contain nonsense variants in order to control for the fact that nonsense variants preferentially occur in specific gene classes (MacArthur et al. 2012; Sulem et al. 2015). We eliminated variants lying within genes encoding olfactory receptors from this and subsequent analyses, as this large gene family has been subject to frequent and likely largely neutral gene inactivation during human evolution (Gilad et al. 2003).

Approximately 28% of synonymous variants in multiexon genes are isoform specific. In order to test for a relationship between isoform specificity and the prevalence of variants, we classified each variant as rare, low frequency, or common, corresponding to allele frequencies of ≤0.5%, 0.5%–5%, and ≥5% in the 1000 Genomes data. We observed no relationship between isoform specificity of synonymous variants and their frequency in the human population, consistent with our expectation that most synonymous codon substitutions are functionally neutral. In contrast, ∼31% of rare nonsense variants are isoform specific, rising to 39% and 43% of low-frequency and common nonsense variants, suggesting that isoform-specific nonsense variants are subject to relaxed purifying selection (Fig. 2A). When we considered all synonymous variants rather than restricting to those within genes containing nonsense variants, we observed slightly greater disparities between the isoform specificity of synonymous versus nonsense variants (Supplemental Fig. S2A). Our estimates of isoform specificity for nonsense variants are consistent with a previous estimate of 32.3% based on a smaller cohort of 185 individuals (MacArthur et al. 2012).

The functional consequences of an isoform-specific variant depend upon the frequency with which those isoforms are produced as mature mRNAs. To test whether nonsense variants preferentially fall within isoforms that are used infrequently, we quantified the inclusion of alternatively spliced sequence containing synonymous or nonsense variants across 16 human tissues. Isoform-specific synonymous variants exhibited a median inclusion of ∼90% irrespective of allele frequency, while low-frequency and common isoform-specific nonsense variants exhibited median inclusions of 64% and 53% (Fig. 2B; Supplemental Fig. S2B). We conclude that even among the isoform-specific nonsense variants, those variants that are most frequently spliced out of mature mRNA are subject to further relaxed selection.

We next measured the empirical isoform specificity of the specific set of homozygous nonsense variants that we analyzed in LCLs (Fig. 1C–E). For each SNV within an expressed gene, we quantified the expected versus observed number of ribosome footprints overlapping the SNV as a simple empirical measure of isoform specificity (e.g., an isoform-specific SNV is expected to exhibit an observed:expected ratio of less than one). We estimated the expected footprint coverage from the total number of footprints aligned to the parent transcript of each SNV and averaged this footprint coverage over all samples with a 0/0 genotype for that SNV. (We restricted the coverage analysis to 0/0 samples in order to measure position-specific mRNA:ribosome association in the absence of premature stop codons.) The resulting distributions were similar for most synonymous and nonsense variants, consistent with our genome-wide prediction that the majority of variants are not isoform specific. Nonetheless, a subset of nonsense variants was strongly depleted for ribosome footprint coverage beyond the background distribution estimated from synonymous variants, indicating that they are frequently excluded from mature mRNAs engaged with ribosomes in LCLs (Fig. 2C). Low-frequency and common nonsense variants exhibited greater depletion for footprint coverage relative to rare nonsense variants, again consistent with relaxed purifying selection (Fig. 2C).

Each of the mechanisms proposed in Figure 1F contributed to the isoform specificity of nonsense variants. Specific examples include nonsense variants within a frame-preserving cassette exon of LGALS8, an alternate promoter of MOB3C, and an alternatively spliced sequence of TMEM218 that is either coding or non-coding, depending upon which start codon is used (Fig. 2D–I; Supplemental Fig. S2D–G). SILAC data were available for LGALS8 and MOB3C, allowing us to confirm that protein levels were sustained in LCLs with homozygous copies of the corresponding nonsense variants. In both cases, protein levels were higher for 0/1 and 1/1 samples relative to 0/0 samples, perhaps due to compensatory translational up-regulation in individuals carrying the alternate alleles or differential stability of the relevant protein isoforms (Fig. 2F,I).

As alternative splicing is frequently tissue specific (Wang et al. 2008), the degree to which isoform specificity mitigates the consequences of a nonsense variant may differ substantially between different cell types. For example, the LGALS8 cassette exon is highly tissue specific, with inclusion ranging from 5% in colon to 87% in testes of individuals with genotype 0/0 (inclusion is ∼55% in LCLs with genotype 0/0) (Supplemental Fig. S2E). Therefore, both the levels of LGALS8 protein in 1/1 individuals and the consequences of specifically ablating the inclusion isoform may differ between cell types.

Nonsense variants may be subject to stop codon readthrough

Genes containing nonsense variants within constitutively included sequence cannot produce mRNAs lacking the variants but could potentially produce protein nonetheless through permissive mRNA translation (Fig. 1F). We therefore searched for read coverage patterns indicative of potential stop codon readthrough, wherein the translating ribosome does not efficiently terminate at the stop codon but rather decodes the stop codon as a sense codon and continues elongating. A nonsense variant that is subject to readthrough might suppress levels of the encoded protein due to imperfectly efficient readthrough but would induce only a single amino acid change at the nonsense variant.

To identify potential examples of readthrough, we compared ribosome footprint coverage of genes containing nonsense variants in samples with genotypes 0/0 or 1/1 for the variants. Of the 118 nonsense variants present as 1/1 in at least one sample within the Yoruba cohort (Fig. 1B), 15 variants had both 0/0 and 1/1 samples available and lay within genes with sufficient footprint coverage in LCLs to identify potential readthrough (median coverage of relevant transcripts of about one footprint per million mapped reads). Of those 15, two nonsense variants within PVRIG and SLFN13 exhibited similar patterns of mRNA:ribosome association in 0/0, 0/1, and 1/1 samples, including in the immediately vicinity of the premature termination codon (Fig. 3; Supplemental Fig. S3). This pattern of ribosome footprints is consistent with stop codon readthrough, in the sense that there is similar quantitative read coverage before and after the nonsense variant. However, it is important to note that this pattern of read coverage could potentially arise from ribosome scanning and translation initiation downstream from the nonsense variant as well.
We tested whether our analysis of PVRIG and SLFN13 was confounded by mismapping of ribosome footprints from homologous transcribed genomic loci. The first two coding exons of PVRIG, which contain the nonsense variant, have 96.8% homology to the transcribed pseudogene PVRIG2P. However, as the density of mismatches between PVRIG and PVRIG2P is high enough to uniquely map ribosome profiling reads and as PVRIG2P exhibits no ribosome footprint coverage, we conclude that the ribosome association illustrated in Figure 3A likely arises from the PVRIG mRNA and not a homologous locus. Similarly, while the first

Figure 2. Alternative splicing, promoter usage, and polyadenylation remove nonsense genetic variants from mature mRNA. (A) Percentages of variants that are isoform specific. Isoform-specific variants are defined as those that do not induce the indicated coding change within at least one RefSeq CDS of the parent gene or that lie within known alternatively spliced mRNA sequence. Plot restricted to variants that lie within genes containing nonsense variants. (N) Number of analyzed variants. Error bars, 95% confidence intervals as estimated by the binomial proportion test. (B) Median inclusion of variants lying within known alternatively spliced mRNA sequence across 16 human tissues. An inclusion level of 75% indicates that 75% of mRNAs transcribed from the parent gene contain the variant, while 25% do not. Inclusion was computed using the Body Map 2.0 data. (Notches) Approximate 95% confidence interval for the median. Plot restricted to variants that lie within genes containing nonsense variants. (C) Histogram of observed versus expected coverage of variants by ribosome footprints. Observed coverage was computed as the number of footprints overlapping each variant; expected coverage was computed as the total number of footprints overlapping each variant’s host CDS, normalized such that the median of the ratio observed:expected over all variants was equal to one. All footprint coverage calculations were restricted to 0/0 samples to avoid potentially confounding effects of nonsense variants, and the plotted values indicate medians over those samples. (Black line) Best-fit normal distribution estimated from synonymous variants. Plot restricted to variants that lie within genes containing nonsense variants and that are present in samples for which ribosome profiling data were available. RNA-seq (D) and ribosome profiling (E) read coverage of a nonsense variant lying within a cassette exon of LGALS8, stratified by sample genotype. Coverage was normalized per sample to control for sequencing depth and then averaged over all samples with the indicated genotypes. Units are reads per million. (Red triangle) Location of nonsense variant. (F) LGALS8 protein levels relative to a sample standard with genotype 0/1. RNA-seq (G) and ribosome profiling (H) read coverage of a nonsense variant lying within an alternate 5′ exon of MOB3C. (I) MOB3C protein levels relative to a sample standard with genotype 0/1.
coding exon of SLFN13, which contains the nonsense variant, has 82.9% homology to the first coding exon of the expressed gene SLFN11, the density of mismatches is sufficient to uniquely map ribosome footprints. Finally, both RNA-seq and ribosome profiling reads mapping to the PVRIG and SLFN13 nonsense variants supported 100% expression of the alternate allele in the 1/1 samples, indicating that genotyping errors are unlikely.

We conclude that stop codon readthrough may enable translation of some mRNAs containing nonsense variants. However, we were unable to estimate the fraction of nonsense variants that might be so rescued given the infeasibility of predicting readthrough from mRNA sequence alone.

Nonsense variants frequently induce N- or C-terminal protein truncation

In addition to stop codon readthrough, stable expression of a truncated protein might enable protein production from a nonsense variant–containing mRNA. Protein truncation can occur at the N or C terminus through different mechanisms. An N-terminally truncated protein results from translation initiation at a start site downstream from the nonsense variant. In contrast, C-terminal truncation results if the nonsense variant is sufficiently close to the normal stop codon and the host transcript is not efficiently degraded via NMD.

We took a sequence-based approach to identify nonsense variants that might induce N- or C-terminal truncation. Simply plotting the relative positions of nonsense variants within their host CDSs revealed that these variants are enriched at the 3′ ends of their host CDSs and are otherwise uniformly distributed, while synonymous variants exhibited no peak at the 3′ end (Fig. 4A; Supplemental Fig. S4A). This pattern is consistent with previous studies of different cohorts (MacArthur et al. 2012; Sulem et al. 2015) and suggests that C-terminal truncation occurs frequently.

We did not observe any enrichment for nonsense variants at the 3′ ends of their host CDSs, suggesting that start codon misannotation—which can cause a variant within a 3′ untranslated region (UTR) to be incorrectly labeled as a nonsense variant (MacArthur and Tyler-Smith 2010; Sulem et al. 2015)—is not a frequent confounding factor in our analyses.

To test whether N-terminal truncation via translation initiation at downstream start sites might also occur, we restricted to SNVs lying within the first 10% of their host CDSs and tested whether those variants were followed by an in-frame methionine within 50 amino acids downstream from the SNV. Approximately 52% of such synonymous variants had a downstream methionine within this distance. In contrast, both rare and common nonsense variants exhibited statistically significant enrichment for downstream methionines, with 57% and 85% of such variants categorized as potential candidates for downstream initiation (Fig. 4B). We next performed two controls to confirm that methionines were specifically enriched downstream from nonsense variants that were downstream translation initiation candidates. First, we measured the occurrence of out-of-frame methionines downstream from SNVs and found no enrichment for out-of-frame methionines downstream from nonsense variants relative to synonymous variants (Supplemental Fig. 5B–D). Second, we searched for in-frame methionines downstream from SNVs but restricted to variants within the middle of the CDS, where downstream translation initiation would produce a severely truncated protein. We observed no statistically significant differences in occurrence of downstream methionines for synonymous versus nonsense variants (Fig. 4C). We conclude that in-frame methionines are specifically enriched downstream from nonsense variants that lie near the 3′ ends of CDSs.

Finally, we tested whether nonsense variants that induced C-terminal truncation might be subject to relaxed selection. Approximately 10% of synonymous variants fell within the last 10% of their host CDSs, independent of allele frequency. In contrast, 13% of rare nonsense variants fell within this region, increasing to 21% for common nonsense variants (Fig. 4D). Therefore, we conclude that the positional distribution of nonsense variants within CDSs is consistent with frequent induction of relatively small N- or C-terminal truncations of the encoded protein.

Ribosome profiling data similarly indicated that downstream translation initiation contributes to translation of mRNAs containing nonsense variants. We observed likely translation initiation downstream from nonsense variants at both canonical (ATG; CCHCR1) and noncanonical (CTG or GTG; ABHD14B) start codons (Fig. 4E,F; Supplemental Fig. S4E,F). Translation initiation at noncanonical start codons has been previously observed in mammalian genomes, although it is uncommon relative to initiation at canonical start codons (Ingolia et al. 2011). We confirmed that there are no regions in the reference human genome assembly with high homology to the exon containing the ABHD14B
nonsense variant and also observed 100% expression of the alternate allele in the 1/1 sample, indicating that genotyping errors are unlikely. Together, our data provide both sequence-based and direct evidence of translation initiation at canonical and noncanonical start codons downstream from common nonsense variants.

Truncated proteins encoded by nonsense variant–containing mRNAs are produced in a heterologous system

Of the nonsense variants highlighted above (within LGALS8, MOB3C, TMEM218, PVRI, SLFN13, CCHCR1, and ABHD14B), we could only confirm LGALS8 and MOB3C protein levels due to mass spectrometry’s incomplete coverage of the proteome. SILAC data were not available for proteins affected by the other highlighted nonsense variants. We therefore used a reporter system to experimentally confirm the conclusions of our genomic analyses, as well as to test whether translation of nonsense variant–containing mRNAs could be recapitulated in a heterologous system (vs. the endogenous context in which the genomic assays were performed). We focused on nonsense variants that we predicted to be subject to stop codon readthrough or to induce N- or C-terminal protein truncation, as those involve incompletely understood biochemical mechanisms or the production of novel proteins not present in cells lacking the nonsense variant.

![Figure 4](image-url)

**Figure 4.** N- and C-terminal protein truncation enable translation of nonsense variant–containing mRNAs. (A) Positional distribution of variants within their host transcripts. Plot indicates the percentages of variants lying within the indicated deciles of their host CDS’s lengths. Error bars, 95% confidence intervals as estimated by the binomial proportion test. Percentages of variants lying within the first 10% (B) or middle 10% (C) of their host CDS’s length that have a downstream methionine (M) within 50 amino acids. Plot restricted to variants that lie within genes containing nonsense variants. Error bars, 95% confidence intervals as estimated by the binomial proportion test. (D) Percentages of variants lying within the last 10% of their host CDS’s length. Plot restricted to variants that lie within genes containing nonsense variants. Ribosome profiling read coverage of nonsense variants lying within CCHCR1 (E) and ABHD14B (F). Units, data normalization, and notation are as in Figure 3. (Blue bars) CTG or GTG noncanonical start codons (no downstream ATG codons are present within the plot region). (Full/half-height blue bars) CTG or GTG codons that are/are not within a Kozak consensus context, defined as RnnTGG (R = A/G, S = C/G).
carrying both tags. In the case of N- or C-terminal truncation, however, only a partial protein product carrying a HA or FLAG tag, respectively, is expected (Fig. 5B).

We introduced each construct containing the reference or variant allele of each candidate gene individually into HEK293 cells via transient transfection and assayed their protein products 48 h post-transfection by detecting the FLAG and HA tags. Constructs containing the reference alleles of CCHCR1, ABHD14B, NDUFV3, and TRIM38 all produced a protein product of the expected size carrying both N- and C-terminal tags (Fig. 5C; Supplemental Fig. S5A). The other reference constructs produced either no protein or protein products that were not of the expected size and so were excluded from further analysis. We additionally tested constructs for PVRIG and SLFN13 that contained their endogenous 5′ and 3′ UTRs, but found that those constructs also did not produce the expected protein products (data not shown).

Table 1. Characteristics of nonsense variants selected for reporter assays

<table>
<thead>
<tr>
<th>RefSeq ID</th>
<th>Locus</th>
<th>Alleles</th>
<th>q (%)</th>
<th>ASE (%)</th>
<th>Isoform specific</th>
<th>Relative pos. (%)</th>
<th>Ref. validated</th>
<th>Origin of protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABHD14B</td>
<td>NM_001146314</td>
<td>G&gt;A</td>
<td>1.3</td>
<td>37</td>
<td>Yes</td>
<td>7.6</td>
<td>Yes</td>
<td>No protein</td>
</tr>
<tr>
<td>CCHCR1</td>
<td>NM_001105564</td>
<td>C&gt;T</td>
<td>47.0</td>
<td>43</td>
<td>Yes</td>
<td>8.9</td>
<td>Yes</td>
<td>Downstream initiation</td>
</tr>
<tr>
<td>CCHCR1</td>
<td>NM_001105564</td>
<td>C&gt;A</td>
<td>10.0</td>
<td>40</td>
<td>Yes</td>
<td>4.6</td>
<td>Yes</td>
<td>Downstream initiation</td>
</tr>
<tr>
<td>FCGR2A</td>
<td>NM_001136219</td>
<td>C&gt;T</td>
<td>9.6</td>
<td>—</td>
<td>Yes</td>
<td>19.5</td>
<td>No</td>
<td>—</td>
</tr>
<tr>
<td>NDUFV3</td>
<td>NM_021075</td>
<td>C&gt;T</td>
<td>6.4</td>
<td>16</td>
<td>Yes</td>
<td>42.0</td>
<td>Yes</td>
<td>C-terminal truncation</td>
</tr>
<tr>
<td>PVRIG</td>
<td>NM_024070</td>
<td>G&gt;T</td>
<td>24.0</td>
<td>35</td>
<td>No</td>
<td>5.2</td>
<td>No</td>
<td>—</td>
</tr>
<tr>
<td>SDR42E1</td>
<td>NM_143168</td>
<td>G&gt;A</td>
<td>13.0</td>
<td>50</td>
<td>No</td>
<td>7.4</td>
<td>No</td>
<td>—</td>
</tr>
<tr>
<td>SUN13</td>
<td>NM_144682</td>
<td>G&gt;T</td>
<td>10.0</td>
<td>53</td>
<td>Yes</td>
<td>1.6</td>
<td>No</td>
<td>—</td>
</tr>
<tr>
<td>TLR10</td>
<td>NM_030956</td>
<td>G&gt;A</td>
<td>1.4</td>
<td>45</td>
<td>No</td>
<td>45.4</td>
<td>No</td>
<td>—</td>
</tr>
<tr>
<td>TRIM38</td>
<td>NM_006355</td>
<td>C&gt;T</td>
<td>3.0</td>
<td>15</td>
<td>No</td>
<td>35.0</td>
<td>Yes</td>
<td>C-terminal truncation</td>
</tr>
</tbody>
</table>

(q) Allele frequency of the alternate allele in the 1000 Genomes data. (ASE) Allele-specific expression based on RNA-seq. ASE was computed for each variant by restricting to samples that were heterozygous (0/1) for the variant and had at least 10 reads covering the variant. No 0/1 samples had sufficient reads to compute ASE for the variant in FCGR2A. “Isoform specific” was defined as in Figure 2A. [Relative pos. (%)] Relative position of the codon containing the SNV within the host CDS. (Ref. validated) Construct containing the reference allele produced a protein product of the expected length with both N- and C-terminal tags. (RefSeq ID) ID of the RefSeq parent transcript for which each variant introduces an in-frame stop codon. (Locus) Genomic coordinates for the GRCh37/hg19 genome assembly. (Alleles) Nucleotides associated with the reference>alternate alleles.

The absence of detectable protein from some constructs may be due to requirements for cell-type–specific translational machinery that is absent from HEK293 cells, problems caused by our introduction of N- and/or C-terminal tags, or possibly incorrect coding sequence annotations. The encoded proteins might also be unstable in HEK293 cells.

Constructs containing either of the two nonsense variants of CCHR1 produced a N-terminally truncated protein fragment indicative of initiation downstream from the premature termination codon (Fig. 5D; Supplemental Fig. S5B). This fragment was produced from the construct containing the reference allele as well, suggesting that CCHCR1 is subject to downstream initiation.

Figure 5. A reporter system recapitulates translation of N- and C-terminally truncated proteins from mRNAs containing nonsense variants. (A) Design of reporter constructs carrying the reference (sense) or alternate (nonsense) alleles of selected genes. (B) Cartoon depicting how permissive translation may enable productive translation of mRNAs containing nonsense variants, as well as the expected protein products from our reporter constructs for each mechanism of permissive translation. (C) Western blot of total protein from HEK293 cells transfected with constructs containing the reference alleles of all candidate genes. The FLAG and HA tags are illustrated in green and red. (White arrows) Constructs that produced the expected protein with both FLAG and HA tags. (D) As in C, but for constructs containing either the reference (Ref) or alternate (Alt) alleles of the genes marked with white arrows in C. For CCHCR1, the “Alt1” and “Alt2” variants correspond to SNVs at genomic positions 31,124,649 and 31,125,257 on Chromosome 6.
alternative translation initiation even in the absence of a nonsense genetic variant. The construct containing the nonsense variant of ABHD14B did not make a detectable protein in HEK293 cells, despite our expectation given uninterrupted ribosome footprints throughout the open reading frame in LCLs, suggesting that an N-terminally truncated protein is likely translated but unstable. Constructs containing the NDUFV3 and TRIM38 nonsense variants produced C-terminally truncated protein fragments that were abundant despite their small size relative to the full-length proteins. These variants in NDUFV3 and TRIM38 were expressed in RNA-seq from 0/1 samples at allelic levels of 16% and 15% (Table 1), indicative of incomplete degradation by NMD that could potentially enable these short proteins to be similarly produced from their endogenous loci. In total, we detected stable expression of N- and C-terminally truncated protein products from constructs encoding four of five unambiguously tested nonsense variants.

Stop codon readthrough rescues protein production from an endogenous gene containing a nonsense variant

As technical limitations of our heterologous reporter system prevented us from experimentally confirming that readthrough occurs for PVRIG and SLFN13 mRNAs, as predicted from ribosome profiling data, we turned to measurements of endogenous protein levels. We identified LCLs that were homozygous for either the reference (0/0) or nonsense (1/1) variants in PVRIG and SLFN13 and obtained five such cell lines from the Coriell Biorepository (Fig. 6A). We used an enrichment strategy to detect endogenous protein, wherein we immunoprecipitated using antibodies directed against PVRIG or SLFN13 and then probed the immunoprecipitate using the same antibodies (Fig. 6B). We detected PVRIG protein at the expected molecular weight in LCLs that were homozygous for either the reference (0/0) or nonsense (1/1) variants, supporting our computational prediction that stop codon readthrough rescues protein production from the nonsense variant–containing allele of PVRIG. PVRIG protein was detectable in the LCLs after enrichment of the endogenous protein using two distinct antibodies, as well as in a positive control consisting of lysate from PVRIG-overexpressing cells, confirming the specificity of our assay (Fig. 6C,D). SLFN13 was not detectable in any LCL, suggesting either that the two tested antibodies against SLFN13 are not effective probes for immunoblotting or that SLFN13 is an unstable protein. Together with data from our heterologous reporter system, these measurements of endogenous protein levels support our hypothesis that diverse mechanisms of permissive translation enable protein production from genes containing nonsense variants.

Discussion

Our data demonstrate that many common nonsense variants have only modest impacts upon the levels of total protein produced from their seemingly disabled parent genes. Translation of mRNAs transcribed from genes containing nonsense variants...
arises from diverse mechanisms during the gene expression process, including transcription start and stop site choice, alternative splicing, translation start site choice, downstream translation initiation, and stop codon readthrough. Given the severe fitness costs of many complete gene knockouts in murine models (Ayadi et al. 2012; de Angelis et al. 2015), our data therefore suggest that permissive RNA processing and translation in human cells facilitates the accumulation of otherwise deleterious genetic variation in the human population.

What fraction of nonsense variants disable gene function? We were able to directly answer this question for common variants by relying on individuals that are homozygous for the variants of interest. However, a direct answer is inaccessible for rare variants, which almost never occur in a homozygous state. (The low coverage of ribosome profiling and SILAC data prevents reliable analysis of allele-specific translation or protein levels in heterozygotes.) We therefore used sequence analyses to estimate that 43%–49% of rare and low-frequency variants identified by the 1000 Genomes Project are isoform specific or may induce N- or C-terminal truncation (Table 2). These predictions do not take into account the incompleteness of current isoform annotations, the possibility of initiation at noncanonical start codons, or the possibility of stop codon readthrough and so may be underestimates. We therefore speculate that permissive RNA processing frequently rescues protein production from genes containing rare nonsense variants, although at lower rates than their common counterparts.

It is important to note that maintenance of protein production may not be equivalent to rescue of gene function in many cases. Different isoforms of a gene may encode proteins with distinct or even opposing functions, so that an isoform-specific nonsense variant may have unpredictable consequences. Even modest N- or C-terminal truncation could alter or abolish protein function by removing peptide signals that direct subcellular localization, prenylation, or other activities or modifications, or induce unexpected gain of function by placing a normally internal peptide signal at the N or C terminus. For example, a recent study identified a novel C-terminal truncating variant in CD164 that segregated with nonsyndromic hearing impairment in an affected family. Only the last six amino acids of CD164 were removed, but without the encoded canonical sorting motif, the truncated protein was trapped on the cell membrane (Nyegaard et al. 2015).

The most unexpected finding of our study was the prominent role played by mRNA translational plasticity in enabling protein production from genes containing nonpathogenic nonsense variants. The regulation of mRNA translation initiation and termination is incompletely understood, yet relevant for interpreting genetic data. Our data indicate that downstream translation initiation is associated with relaxed purifying selection and gives rise to N-terminally truncated proteins for some nonsense variants. Our data also show that stop codon readthrough enables translation of at least one mRNA containing a common nonsense variant. Stop codon readthrough is thought to be relatively rare in mammals, with only a few known examples (Eswarappa et al. 2014; Loughran et al. 2014; Schueren et al. 2014; Stiebler et al. 2014). Nonetheless, efficient readthrough of an otherwise fatal nonsense variant has been observed clinically (Pacho et al. 2011), highlighting the importance of translational plasticity when interpreting genetic data.

Together, our data demonstrate that permissive RNA processing has the potential to convert loss-of-function variants from genetic nulls into hypomorphic, silent, or even neomorphic alleles. Our results may therefore help to explain recent observations that some apparently healthy individuals carry homozygous putative loss-of-function variants within genes implicated in Mendelian disease (Kaiser et al. 2015). An improved mechanistic understanding of RNA processing and translation is essential for ongoing attempts to predict the pathogenicity of genetic variation (Kircher et al. 2014; MacArthur et al. 2014).

### Methods

#### Genotypes

Genotypes from the phase 3 variant set called by the 1000 Genomes Project (1000 Genomes Project Consortium 2012) were downloaded from ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502.

We restricted our analysis of ribosome profiling data from Battle et al. (2015) to the 62 samples with genotypes called in this variant set, and used the 1000 Genomes genotyping for all analyses. For the RNA-seq data from Lappalainen et al. (2013), we used the genotyping information from the original study, available at http://www.ebi.ac.uk/arrayexpress/files/GEUV-1/genotypes/ and restricted to the 421 samples that were genotyped by sequencing. For global SNV analyses based on the 1000 Genomes variant set, we restricted to SNVs for which the alternate allele was present in a subsampled fraction of the 1000 Genomes data, wherein we restricted each of the 26 populations to no more than 90 samples to help to even the genotyping coverage among the available populations. For all data sets, we restricted genotyping calls to SNVs.

#### SNV functional annotation

We used SnpEff (Cingolani et al. 2012) to identify variants that affected coding DNA sequences and annotate their functional effects. We ran snpEff v4_1c with the options “-v -lof -no-downstream -no-intragenic -no-intron -no-upstream -no-utr –formatEff –download hg19” and then restricted to SNVs with predicted effects of “synonymous variant” or “stop gained” for at least one RefSeq transcript. For all transcript-level analyses, such as computing the amino acid affected by a SNV, we relied upon the AnnotationHub and BSGenome packages within Bioconductor (Huber et al. 2015).

#### Genome annotations

A merged genome annotation was created by combining the UCSC knownGene annotations (Meyer et al. 2013), Ensembl 71 (Flicek et al. 2013), and isoform annotations from MISO v2.0 (Katz et al. 2010) as previously described (Dwinge et al. 2014). A file of all

---

**Table 2. Percentages of nonsense variants that are predicted to potentially not abrogate protein production**

<table>
<thead>
<tr>
<th></th>
<th>Rare</th>
<th>Low</th>
<th>Common</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isoform specific</td>
<td>30.2%</td>
<td>36.9%</td>
<td>40.2%</td>
</tr>
<tr>
<td>Downstream initiation</td>
<td>5.5%</td>
<td>5.7%</td>
<td>9.8%</td>
</tr>
<tr>
<td>C-terminal truncation</td>
<td>13.4%</td>
<td>18.4%</td>
<td>21.4%</td>
</tr>
<tr>
<td>Total</td>
<td>42.8%</td>
<td>48.6%</td>
<td>52.7%</td>
</tr>
</tbody>
</table>

Isoform-specific variants are defined as in Figure 2A. Candidates for downstream initiation and C-terminal truncation are defined as in Figure 4. The last row (total) specifies the percentage of variants that are isoform specific, that are candidates for downstream initiation, or that cause C-terminal truncation. As described in the Discussion, these numbers are likely underestimates and do not include the possibility of stop codon readthrough.
possible splice junctions, consisting of all splice junctions present in the merged genome annotation as well as all possible combinations of 5′ and 3′ splice sites within each parent gene, was created for subsequent read mapping.

**RNA-seq and ribosome profiling read mapping**

FASTQ files were downloaded for RNA-seq data from LCLs from EBI’s ArrayExpress under accession number E-GEUV-1 (Lappalainen et al. 2013) for ribosome profiling data from LCLs from the NCBI Gene Expression Omnibus under accession number GSE61742 (Battle et al. 2015) and for the Illumina Body Map 2.0 data set from ArrayExpress under accession number E-MTAB-513. RNA-seq and ribosome profiling reads were mapped to the UCSC hg19 (NCBI GRCh37) genome assembly as previously described (Dvinge et al. 2014). (Because our analyses consisted of comparisons between the relative expression of the reference versus alternate alleles, which are unaffected by the specific genome assembly, using the NCBI GRCh38 genome assembly instead of GRCh37 would not significantly affect our conclusions.) Briefly, RSEM (Li and Dewey 2011) was used to map reads to annotated transcripts, and the resulting unaligned reads were then aligned to the genome and the set of all possible splice junctions using TopHat v2.0.8b (Trapnell et al. 2009). We required that reads map with a maximum of three mismatches for the RNA-seq data, which consisted of 2x75 bp reads, and one mismatch for the ribosome profiling data, which consisted of shorter mixed-length reads. Reads mapped by RSEM and TopHat were then merged to create final BAM files of aligned reads.

**SILAC mass spectrometry data**

Gene-level quantification of relative protein abundance based on SILAC was taken from Supplemental Table 4 of the work by Battle et al. (2015).

**Transcript expression, ASE, and isofrom ratio measurements**

Transcript expression (e.g., as illustrated in Fig. 1C,D) was quantified for RNA-seq or ribosome profiling data in units of fragments per kilobase per million mapped reads (FPKM) as follows. For each coding gene annotated in Ensembl 71, we counted the numbers of reads (or fragments, in the case of paired-end reads) that aligned to the longest CDS of that gene based on the RefSeq transcripts and then divided by the CDS length in kilobases. We then normalized these counts by dividing with a sample-specific normalization factor, defined as \((10^6 \times \text{total number of reads mapping to any Ensembl coding gene})\). Samples were then classified according to their genotype, and the average coverage was computed for each genotype by averaging over the normalized coverage values for each sample. RNA-seq read coverage plots were generated via an identical procedure, but the reads were not trimmed (as all reads were of equal length). These plots relied upon the AnnotationHub and BSgenome packages within Bioconductor (Huber et al. 2015).

**Plotting and graphics**

All plots and figures were generated with the dplyr (http://CRAN.R-project.org/package=dplyr) and ggplot2 packages (Wickham 2009).

**Reporter constructs**

The reporter constructs for reference alleles were obtained as full-length cDNAs in the pcDNA3.1 backbone carrying FLAG and HA tags at the N and C termini, respectively (GenScript). Point mutations were introduced into constructs containing the reference alleles to generate the variants (GenScript).

**Immunoprecipitation of PVRIG and SLFN13 from lymphoblastoid cell lysates**

LCLs were obtained from the Coriell Biorepository and cultured under recommended conditions in RPMI with 15% FBS. Twenty-five million cells were pelleted and lysed in NP-40 lysis buffer (50 mM Tris at pH 8.0, 150 mM NaCl, 1% NP-40) containing EDTA-free protease inhibitor mini tablets (Pierce; catalog no. 868666). The lysate was incubated overnight at 4°C with 25 µL of protein G Dynabeads (Invitrogen) prebound to 10–50 µg of antibody (PVRIG: ab168153 [Abcam], NBP2-13832 [Novus Biologicals]; SLFN13: ab173951 [Abcam], NBP1-93879 [Novus Biologicals]). The beads were washed thrice with the Dynabead wash buffer and the bound proteins eluted in the nondenaturing elution buffer. The eluate was then subjected to Western blotting as described below using the same antibodies used for the immunoprecipitation. A transient overexpression lysate of PVRIG (LY411386; Origene Technologies) served as a positive control to determine the specificity of the antibodies used for Western blotting.

**Western blotting**

HEK293 cells were grown in DMEM containing 10% FBS and transfected with the reporter constructs using Lipofectamine 3000 transfection reagent (Invitrogen). After 48 h, the cells were lysed in TRIzol reagent (Invitrogen). Total protein was extracted following the manufacturer’s instructions, and the protein pellet resuspended in sample buffer containing 5% SDS and 0.5 M unbuffered Tris base to ensure efficient solubilization, aided by gentle sonication in an ultrasonic water bath. Protein concentrations were determined using the BCA protein assay (ThermoScientific), 5 µg of total protein was resolved on 12% NuPAGE Bis-Tris acrylamide gels (Novex), and Western blotting was performed using the LICOR system. The primary antibodies used were mouse monoclonal anti-FLAG M2 antibody (Sigma-Aldrich; catalog no. F1804) and rabbit monoclonal anti-HA antibody (EMD Millipore; catalog no. 05-9028R). IRDye-conjugated secondary antibodies (LICOR) were used for quantitative detection of the tags. Rough estimates of the protein molecular weights were made by comparison with the Precision Plus Protein Kaleidoscope standards (Bio-Rad).
Acknowledgments
We thank Jesse Bloom and Guo-Liang Chew for comments on the manuscript. This research was supported by the Ellison Medical Foundation AG-NS-1030-13 (R.K.B.) and the FSH Society FSFS-22014-01 (S.J.).

Author contributions: S.J. and R.K.B. performed the experiments, analyzed the data, and wrote the paper.

References


Received February 2016; accepted in revised form September 16, 2016.

1650 Genome Research www.genome.org
Translational plasticity facilitates the accumulation of nonsense genetic variants in the human population

Sujatha Jagannathan and Robert K. Bradley

*Genome Res.* 2016 26: 1639-1650 originally published online September 19, 2016
Access the most recent version at doi:10.1101/gr.205070.116

Supplemental Material

http://genome.cshlp.org/content/suppl/2016/11/09/gr.205070.116.DC1.html

References

This article cites 41 articles, 14 of which can be accessed free at:
http://genome.cshlp.org/content/26/12/1639.full.html#ref-list-1

Open Access

Freely available online through the *Genome Research* Open Access option.

Creative Commons License

This article, published in *Genome Research*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

© 2016 Jagannathan and Bradley; Published by Cold Spring Harbor Laboratory Press