Tests for the role of NMD in selection on germline PTC variants

We compared the observed frequencies of NMD-triggering and NMD-evading rare population PTCs to their expected frequencies, which we simulated from trinucleotide spectra of rare population variants (Methods; we note that trinucleotide sequence composition of different NMD-evading regions does not vary appreciably, Extended Data Fig. 2a-b). The fraction of NMD-triggering PTCs expected at random (60%) is significantly higher than in observed rare variants (52% NMD-triggering PTCs; p<2.2e-16, Chi-square goodness-of-fit test). This difference from a simulated random baseline grows more pronounced for higher-frequency variants (60% simulated versus 25% observed at MAF>10%; Extended Data Fig. 2c), suggesting a genomic signature of selection on NMD-inducing PTCs in the human germline.

We wanted to further substantiate our observation of the enrichments of NMD-detected/evaded PTC variants in individual disease genes (Extended Data Fig. 1b), which was seen after normalizing to the lengths of gene regions seen or not seen by NMD (illustrated in Extended Data Fig. 1c). To this end, we therefore also identified enrichments for PTCs that do or do not trigger NMD, after normalizing the PTC density to the local density of rare missense ClinVar variants, which cannot trigger NMD. In other words, if a gene has a high pathogenic PTC-to-missense ratio in NMD-triggering regions, this suggests that those PTCs seen by NMD tend to be pathogenic and therefore that NMD aggravates the disease phenotype due to PTCs. Conversely, if the pathogenic PTC-to-missense ratio is low in NMD-triggering regions, this suggests PTCs seen by NMD tend to be less pathogenic and thus NMD alleviates the phenotype. This analysis yielded overall similar results as the analysis without normalization for missense variants, with 22 of 30 genes remaining significant at FDR<5% (total 30 significant genes, Extended Data Fig. 3a-b), and further underscoring the predominance of genes enriched for NMD-detected pathogenic PTCs (19 vs 11 significant in this test).

Examples of disease genes where NMD ameliorates phenotypes

Even though we predict NMD often aggravates disease phenotypes, there are, however, multiple diseases where we predict that activity of NMD alleviates the phenotype (n=13 and n=40 disease-associated genes, at 5% and 25% FDR, respectively; Extended Data Fig. 3d). One example
concerns the nucleophosmin (\textit{NPM1}) gene where somatic mutations cause acute myeloid leukemia, commonly via a 4 nt insertion that introduces a cytoplasmic localization signal, resulting in an oncogenic gain-of-function (GoF). These insertions normally happen in the last exon of NPM1, where they are invisible to NMD\textsuperscript{1}. A related example is \textit{NOTCH3}, where GoF truncating variants in the last exon that escape NMD have been suggested to prolong protein half-life by removing a PEST sequence\textsuperscript{2}. In these cases inhibiting NMD would not have a favorable impact on the disease phenotype. Some of the other genes where we predict NMD inhibition would often not be helpful include CFH and TGFBR2 (Extended Data Fig. 3d).

\textbf{Predicting the pathogenicity of PTCs using NMD rules}

The ability to escape NMD is surprisingly prevalent in standing genetic variation with \textasciitilde50\% of rare PTCs evading NMD to some extent\textsuperscript{3} (Fig. 2a). This general result is also evident in disease genes: in a majority of the genes we examined (59\% of genes with \textasciitilde20 PTC variants in ClinVar), we predict that at least a quarter of the pathogenic PTCs do not trigger NMD (Extended Data Fig. 1b). In many cases, this becomes evident only after incorporating the two new non-canonical NMD evasion rules. For example, the long-exon rule applies to \textasciitilde40\% of PTC variants observed in the important cancer predisposition genes MSH6 and BRCA2, among others (Extended Data Fig. 3). Additionally, while the start-proximal NMD evasion rule covers a smaller part of each gene, it nonetheless applies to PTCs observed in a variety of disease genes. The presence of many NMD-evading PTCs in disease genes is not unexpected, given the overall abundance of NMD-evading germline variants in human populations (Fig. 2a). This data suggests that PTC-bearing disease genes should, therefore, quite widely be expected to produce truncated proteins, which are known to often retain partial functionality\textsuperscript{4}, with implications for disease severity and therapeutic prospects.

To investigate whether the NMD activity rules can systematically improve pathogenicity predictions, we trained a logistic regression model to distinguish known benign from known pathogenic PTC variants based on a set of sequence features in ALoFT, a state-of-the-art method to predict the impact of putative loss-of-function variants\textsuperscript{5}. Including our NMD predictions significantly increases the accuracy of the model (area under precision-recall curve (AUPRC) increases from 0.989 to 0.991, \textit{P}=4\textit{e}−19 by ANOVA; Fig. 2d; crossvalidation results are reported in Supplementary Table 1). At a fixed FDR of 1\% (equivalent to precision=99\%), the recall of the model increases from 58\% to 66\% of the pathogenic PTC variants recovered when including the
NMD predictions. This illustrates how even a modest increase in AUPRC can translate to meaningful increases in coverage with predictions at high-stringency thresholds. We provide examples of 23 ClinVar PTCs which were re-classified from pathogenic to putatively non-pathogenic by the NMD-aware model (Supplementary Table 2), of which only 9 are subject to the last-exon NMD evasion rule while 14 are subject to the new NMD rules.

**Gene 5’ ends are not optimal CRISPR-Cas9 target sites because of NMD evasion**

The non-canonical start-proximal and long exon rules included in NMDetective are important because they apply to a substantial proportion of gene sequence (Fig. 2a) including in disease genes (Extended Data Fig. 1b). Both rules cover ~12% of human protein-coding sequence, compared to ~3% and ~18% covered by the canonical 50nt and last exon rules, respectively (note that the last exon rule also encompasses intronless genes).

The start-proximal NMD rule is also extremely relevant for gene editing applications using CRISPR-Cas9, where sgRNAs are often designed to target 5’ ends of the coding region to induce gene inactivation via frameshifting indels. This 5’ bias is indeed used as an important criterion in some popular sgRNA design software such as E-CRISP6. Analyzing the genomic distribution of sgRNAs used in genome-wide CRISPR screens and sgRNAs designed by two popular sgRNA design tools reveals that on average 36% of the sgRNAs are targeted to the first 150 nt of the coding sequence (Extended Data Fig. 5b-e), regions that do not trigger NMD. For example, only 14% of the best ranking sgRNAs designed by E-CRISP tool against the top 100 most cited genes (from http://doi.org/10.5281/zenodo.1066066) target regions that will trigger NMD, with 69% targeting the first 150 nt of the coding sequence (Extended Data Fig. 5e).

We therefore evaluated the start-proximal rule at high resolution in another experimental dataset, where ‘saturation genome editing’ using CRISPR-Cas9 was applied to the initial and terminal exons of the BRCA1 gene and cell survival and mRNA levels were measured7. PTCs covered by the canonical last exon and 50nt rules indeed cause smaller reductions in mRNA levels (Fig. 3c, 2.5-fold and 1.5-fold, respectively compared to a 4.7-fold reduction in regions predicted to trigger NMD, p=2.0e−5 and p=1.1e−3 by Mann-Whitney U test). As a control, missense mutations (that do not trigger NMD) in the same regions had no effect on mRNA levels (Fig. 3c).
Strikingly, there is an even stronger reduction of NMD efficacy in the 5’ end of the gene coding region: PTCs in the first 150nt of the BRCA1 coding sequence hardly change mRNA levels (1.2-fold) compared to the change in mRNA levels caused by PTCs in the immediate downstream region of 151 nt - 300 nt (2.3-fold, p=1.7e−8) and in all regions predicted to trigger NMD (4.7-fold, p=3.0e−12). Taken together, these data show that the start-proximal rule of NMD evasion is an important determinant of the efficiency of gene inactivation by CRISPR-Cas9. We therefore strongly recommend that sgRNAs be designed such that they target a locus past the initial 150 nt of the coding region.

In addition to this mRNA level-based analyses, experiments measuring changes in protein levels also supports the importance of start-proximal and the long-exon NMD rules. In six of the nine genes, we observe that loss of protein levels is also attenuated when using sgRNAs that target the 5’ gene end, consistent with the start-proximal rule of NMD evasion (Fig. 3a; median 1.2-fold difference, IQR 0.8 to 1.7-fold, Extended Data Fig. 4a,c). In four genes, this data suggests a near-absence of NMD activity on PTCs in the start-proximal region. One of the nine genes (encoding the human CD13 marker) had an exon subject to the long-exon rule, providing an opportunity to validate also this rule. Indeed, sgRNAs targeting this long exon also exhibited a diminished reduction in protein levels (Fig. 3a, 1.2-fold), consistent with NMD evasion. Overall, the NMDetective predictions based on sgRNA-target locations were significantly correlated with a decrease in protein levels (range: r=0.21 to 0.81 across genes, all genes significant at p<2e−16). This analysis of how protein levels are affected by sgRNAs targeted to various loci in nine genes supports the relevance of the rules in NMDetective for predicting the efficiency of gene inactivation experiments using CRISPR-Cas9.

**NMD evasion is a biomarker of tumor immune reactivity**

We first tested for associations between the burden of frameshifting indels predicted to trigger or not trigger NMD, and markers of immune cell infiltration into a tumor, using gene expression data in the TCGA pan-cancer cohort.

We found that infiltrated lymphocytes were significantly (FDR=2.5%) higher in tumors with many NMD-evading (n>10) but few NMD-detected frameshifting indels (0<n≤10), when compared to tumors many NMD-detected (n>10) but few NMD-evading (0<n≤10) frameshifts (Extended Data Fig 7a). We found similar trends for cytotoxic CD8+ T-cell infiltration, T-cell receptor richness and
entropy scores, as well as the RNA-Seq expression levels of genes encoding PD-1, PD-L1 and CTLA-4 proteins (all FDR<25%; Extended Data Fig. 7b). Overall, this suggests that a high number of frameshifting indels that do not trigger NMD - but not a high number of frameshifting indels that trigger NMD - results in a higher immune reactivity of a tumor.

Additional evidence that the NMD pathway underlies this difference is provided by tumors bearing somatic mutations in UPF1, a gene required for NMD activity. Tumors with damaging mutations in UPF1 show significantly higher infiltration by CD8+ T-cells and by NK cells than tumors bearing low-impact UPF1 mutations in UCEC samples (uterine corpus endometrial carcinoma, the cancer type in TCGA which has the highest prevalence of UPF1 mutations, with 8.5% tumors harboring a mutation; Extended Data Fig. 8b). This suggests that inhibiting the NMD pathway can increase the immune reactivity of a tumor by increasing the number of neoantigens produced from transcripts containing frameshifting indels. This enhanced immunoreactivity when NMD is inhibited may have therapeutic relevance (see below).

In order to ascertain whether NMD evasion is also associated with better clinical outcomes, we examined survival in a large cohort of renal cell tumors, a cancer type proposed to be highly immunogenic because indels dominate its mutational landscape. Indeed, overall survival was more favorable for patients with more frameshifts in regions that evade than that trigger NMD, in contrast to patients with more frameshifts in regions that trigger than evade NMD (75% of patients survive 8 years versus 4.4 years, respectively; p=0.011). This also holds true after stratifying for the overall frameshift burden (at p<0.079; Extended Data Fig. 8a). Interestingly, even in the group of patients with only a single frameshifting indel in protein-coding regions, whether this single mutation triggers NMD or not is predictive of survival (p=0.050), underscoring the strong antigenic potential of individual frameshifted peptides and also of NMD in controlling their expression.

**Contribution of NMD features to a model for predicting immunotherapy response**

We estimated the contribution of the NMD rules to a joint predictive model that classifies response versus no-response based on (a) the tumor mutation burden (of single-nucleotide variants), an established marker for immunotherapy and (b) the burden of frameshifting indels, separated into those triggering NMD or not (Extended Data Fig. 9b-d, crossvalidation results are reported in Supplementary Table 1). In the joint model, total mutation burden has the highest contribution (p=1.6e−5 by Chi-Square Test; McFadden's pseudo-$R^2$=5.5%; Extended Data Fig. 9c). NMD-
triggering frameshifts are not significant and contribute very little to prediction \( (\text{pseudo-} R^2 = 0.16\%) \). In contrast, the number of NMD-evading frameshifts is significant \( (p=0.0021) \) and the second-most predictive variable \( (\text{pseudo-} R^2 = 2.8\%) \). The number of exonic frameshifts per tumor sample is usually low (Q1-Q3: 1-5 frameshifting indels in protein coding regions) and the number of NMD-evading and NMD-triggering frameshifts are not well correlated \( (R^2 = 0.33) \). As a result, the total number of frameshifts is not an ideal proxy for the number of the immunogenic, NMD-escaping frameshifts, underscoring the importance of considering predicted NMD effects on individual frameshifting indels.

The accuracy of the logistic regression model to classify responders from non-responders increases the fraction of patients correctly classified to 71.1\%, in a model which considers TMB and NMD-evading frameshifts, compared to 65.2\% correctly classified in a baseline model that considers only TMB. The specificity of the model increases to 92.3\% (TMB+NMD) from 89.3\% (TMB only), while the sensitivity increases to 37.1\% (TMB+NMD) from 26.7\% (TMB only). (For the purposes of calculating sensitivity and specificity, the responders \([n=105]\) were considered as positive examples while the non-responders \([n=168]\) were considered as negative examples).

**Supplementary Note references:**

1. Falini, B. *et al.* Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N Engl J Med* **352**, 254-66 (2005).
2. Gripp, K.W. *et al.* Truncating mutations in the last exon of NOTCH3 cause lateral meningocele syndrome. *Am J Med Genet A* **167A**, 271-81 (2015).
3. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-11 (2013).
4. Guo, Y. *et al.* Dissecting disease inheritance modes in a three-dimensional protein network challenges the "guilt-by-association" principle. *Am J Hum Genet* **93**, 78-89 (2013).
5. Balasubramanian, S. *et al.* Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. *Nat Commun* **8**, 382 (2017).
6. Heigwer, F., Kerr, G. & Boutros, M. E-CRISP: fast CRISPR target site identification. *Nat Methods* **11**, 122-3 (2014).
7. Findlay, G.M. *et al.* Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217-222 (2018).
8. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830 e14 (2018).
9. Turajlic, S. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol* **18**, 1009-1021 (2017).