The rules and impact of nonsense-mediated mRNA decay in human cancers

Rik G H Lindeboom1,2,5, Fran Supek1–3 & Ben Lehner1,2,4

Premature termination codons (PTCs) cause a large proportion of inherited human genetic diseases. PTC-containing transcripts can be degraded by an mRNA surveillance pathway termed nonsense-mediated mRNA decay (NMD). However, the efficiency of NMD varies; it is inefficient when a PTC is located downstream of the last exon junction complex (EJC). We used matched exome and transcriptome data from 9,769 human tumors to systematically elucidate the rules of NMD targeting in human cells. An integrated model incorporating multiple rules beyond the canonical EJC model explains approximately three-fourths of the non-random variance in NMD efficiency across thousands of PTCs. We also show that dosage compensation may sometimes mask the effects of NMD. Applying the NMD model identifies signatures of both positive and negative selection on NMD-triggering mutations in human tumors and provides a classification for tumor-suppressor genes.

The NMD pathway is an mRNA surveillance system that protects eukaryotic cells by reducing the production of harmful truncated proteins translated from PTC-bearing transcripts1. In addition, NMD is involved in post-transcriptional regulation of global gene expression2. PTC-introducing variants cause human genetic diseases3, making NMD an important modulator of disease outcome. In particular, NMD can both protect against disease4 and aggravate disease phenotypes5. Therefore, a complete understanding of how the NMD machinery selects transcripts to degrade is crucial for predicting the phenotypic consequences of nonsense mutations in humans.

The canonical model of how PTC-containing transcripts are recognized in mammalian cells is the EJC model, which was defined in yeast, where a very long 3′ UTR triggers NMD13,14, and also applies to Drosophila melanogaster and Caenorhabditis elegans15,16. One further exception to the canonical EJC model has been observed in Drosophila, where a long 3′ UTR13,14 triggers NMD downstream of an EJC13,14. This is consistent with the ‘faux 3′ UTR’ model for mammalian cells. This model was originally defined in yeast, where a very long 3′ UTR triggers NMD13,14, and also applies to Drosophila melanogaster and Caenorhabditis elegans15,16.

We hypothesized that 5′-proximal PTCs circumvent NMD by reinitiating translation downstream or, alternatively, that proximity to poly(A)-binding protein in a closed mRNA loop inhibits NMD17,19. It is not clear how widely these exceptions to the canonical EJC model apply beyond the few transcripts in which they were discovered.

To systematically test the general validity of these proposed rules for NMD efficiency and to discover new NMD rules requires a global and unbiased approach. A recent study examined the effects of germline protein-truncating variants from 635 individuals on allele-specific mRNA expression20, providing support for the canonical EJC model but suggesting that additional rules are also likely to be important21. We hypothesized that human cancer genomes and their matched transcriptomes are a useful resource to discover and test these additional rules. Cancer exomes can, in some instances, harbor hundreds of somatic single-nucleotide variants (SNVs), including new PTCs, most of which are passenger mutations with little consequence for tumor cells22. Here we systematically elucidate the rules governing NMD in human cells using the nonsense variants from nearly 10,000 human tumors and are able to explain approximately three-fourths of the non-random variance in NMD efficiency. Applying the model identifies both positive and negative selection on NMD-inducing somatic mutations in human tumors and provides a classification for human tumor-suppressor genes.

RESULTS

To systematically examine determinants of NMD efficiency, we compiled a data set of somatic nonsense mutations from 9,769 patients with cancer for whom exome sequence, copy number alteration (CNA) and mRNA expression data were available from The Cancer Genome Atlas (TCGA) (Fig. 1). We considered 27 cancer types separately (Supplementary Fig. 1b), further subdividing them into 94 subtypes by clustering based on global gene expression patterns (Online Methods). The NMD efficiency of each nonsense mutation was estimated as the accuracy of nonsense-mediated mRNA decay (NMD).

The NMD efficiency of each nonsense mutation was estimated as the accuracy of nonsense-mediated mRNA decay (NMD).

1EMBL-CRG Systems Biology Unit, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain. 2Universitat Pompeu Fabra (UPF), Barcelona, Spain. 3Division of Electronics, Rudjer Boskovic Institute, Zagreb, Croatia. 4Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. 5Present address: Department of Molecular Biology, Radboud Institute for Molecular Life Sciences, Nijmegen, the Netherlands.

Received 17 March; accepted 11 August; published online 12 September 2016; doi:10.1038/ng.3664
fold change in the mRNA expression level for the mutant transcript in comparison to the median expression level for wild-type transcripts of the same gene in the same cancer subtype (Supplementary Fig. 1a). We applied stringent filters to ensure absence of CNA, high allelic frequency of the somatic variants in each tumor and high level but low variability of wild-type gene expression across samples (Online Methods). This filtering yielded 2,840 nonsense mutations with high-confidence NMD efficiency estimates in 1,900 different genes originating from 1,271 patients (median = 1 nonsense mutation per patient, 25 patients with >10 nonsense mutations).

We validated our findings in two additional data sets (Fig. 1): somatic small indels that result in a downstream out-of-frame PTC in TCGA tumor samples (3,151 indel-induced PTCs in 1,957 genes from 1,156 patients after filtering) and heterozygous germline nonsense variants in 462 lymphoblastoid cell lines from the Geuvadis study20 where NMD efficiency was quantified as allele-specific mRNA expression of the reference over the variant allele 21 (1,784 high-confidence PTCs in 487 unique loci after filtering on the basis of read counts).

The standard EJC model applies widely to human genes

The EJC model postulates that translation termination at least 50 nt upstream of an exon junction triggers NMD6–7. Comparing the NMD efficiency of PTCs located upstream of the last EJC and PTCs inside the last exon (Fig. 2b and Supplementary Fig. 1c,d) showed that, overall, NMD was indeed inefficient in the latter case (84.7% versus 1.7% NMD efficiency; calculated as observed NMD efficiency over maximum observable NMD efficiency using our NMD model (Online Methods); P < 2.2 × 10^-16; two-tailed Mann–Whitney U test). We also observed a sharp decrease in efficiency approaching complete NMD insensitivity over the 50-nt boundary upstream of the last exon, demarcating the physical location of the EJC (2.2% versus 84.7% NMD efficiency, P = 5.3 × 10^-5). Thus, a downstream EJC is a widespread signal for efficient NMD in many human genes. Consistently, intronless transcripts that bore PTCs were not degraded by NMD (Supplementary Fig. 1f), as expected from their lack of deposited EJCs.

Next, we examined whether the yeast and Drosophila faux 3’ UTR model11–14 also applies to human cells, meaning that a long 3’ UTR would be sufficient to trigger NMD, even in the absence of EJCs. When examining PTCs located downstream of the last exon junction or in intronless genes, we did not observe a correlation of NMD efficiency with PTC proximity to the 3’ end of the transcript (P = 0.92, t test for significance of the Pearson R value; Supplementary Fig. 1c,f). This conclusion was upheld when examining frameshift-inducing indels in TCGA or germline truncation variants (P > 0.54; Supplementary Fig. 1g,h). Thus, although faux 3’ UTR–related mechanisms may be important in individual cases23,24, they do not seem to be a prevalent influence on NMD efficiency in human cells. Consistently, a relationship was not observed between transcript 3’ UTR length and mRNA turnover rates for transcripts targeted by the key NMD factor UPF1 in a study in HeLa cells25.

However, transcripts harboring an intron in the 3’ UTR showed a different trend, in which PTCs in the penultimate exon did not trigger NMD, despite the presence of a downstream exon junction (7% NMD efficiency for PTCs in penultimate exons versus 81% efficiency for PTCs in upstream exons, P = 0.045; 12% versus 81% NMD efficiency for frameshift-inducing indels, P = 0.006; Fig. 2c and Supplementary Fig. 2a). This suggests that EJCs in 3’UTRs are generally less capable of interacting with the termination machinery to initiate NMD, and the NMD-triggering EJC appears to be the one located at the penultimate splice site.

Start-proximal PTCs avoid NMD by reinitiating translation

Previously, PTCs introduced into the first half of the first exon of the β-globin gene were shown not to reduce mRNA abundance18, suggesting that the NMD machinery did not recognize transcripts with these PTCs. Our cancer genome data support the generality of this result, showing a widespread decrease in NMD efficiency when PTCs are located in close proximity to the start codon (35% versus 93% NMD efficiency for PTCs within 200 nt of the start codon versus >200 nt from the start codon, P < 2.2 × 10^-16; Fig. 2d and Supplementary Fig. 2b,c). PTCs located in the first 100 coding nucleotides were rarely targeted for NMD, and there was a gradual increase in NMD efficiency up to 200 nt from the start codon (Fig. 2d). This start-proximal NMD insensitivity is not explained by the canonical EJC model of NMD because downstream EJCs fail to trigger NMD: The same conclusion was reached in analyzing somatic frameshift data (P = 5.5 × 10^-12; Supplementary Fig. 2b) and allele-specific expression of germline variants (P = 0.0001; Supplementary Fig. 2c). Two mechanisms have been proposed to account for the examples of NMD insensitivity for transcripts with start-proximal PTCs: (i) transcript stabilization by interaction between poly(A)-binding protein C1 (PABPC1) and the translation termination machinery19 and (ii) bypass of NMD by the reinitiation of translation at a downstream AUG codon20.

Comparing NMD efficiency on the basis of the reading frame of the first AUG codon downstream of start-proximal PTCs showed NMD insensitivity for transcripts with an in-frame AUG downstream of the PTC (1.9% versus 57% NMD efficiency for transcripts in which the downstream AUG was in a different reading frame, P = 3.4 × 10^-4; Fig. 2a and Supplementary Fig. 2d,e), suggesting that translation reinitiation is a widespread mechanism of evading NMD for human genes. The distance between the original start codon and the PTC was a strong predictor of start-proximal NMD insensitivity, whereas the distance to the downstream AUG or the Kozak sequence strength of the downstream AUG did not affect NMD efficiency (Supplementary Fig. 2f,g). However, we observed that the efficiency of NMD was still reduced for start-proximal PTCs even in cases where the downstream AUG was in a different reading frame (57% versus 93% NMD efficiency for PTCs located >200 nt downstream of the start codon), suggesting that other mechanisms can result in NMD bypass. Shorter distance between the best match to PABPC1 motifs and the PTC in a hypothetical looped mRNA conformation did not associate with lower NMD efficiency (Supplementary Fig. 2h–k).
Figure 2 A downstream EJC and proximity to the start codon are widespread signals for NMD. NMD efficiency (−log2(mutant mRNA levels/wild-type mRNA levels)) is shown for PTCs in different gene regions. (a) NMD efficiency for PTCs in the first 200 coding nucleotides with an in-frame or out-of-frame AUG codon downstream. (b) NMD evasion for PTCs in the last exon (red) and 50 nt upstream of the last exon (green), in comparison to efficient NMD for PTCs in the penultimate exon (blue) and exons further upstream (purple). (c) PTCs in the penultimate exon do not induce NMD when an intron is present in the 3′ UTR, as shown for nonsense mutations (left) and PTCs resulting from frameshift mutations (right). In a–c, P values were determined by two-tailed Mann–Whitney U tests not adjusted for multiple testing. Color-coding for gene regions is as in d. (d) Top, schematic with color-coded gene regions showing different NMD efficiencies; boxes represent exons. Bottom, loess fits with 95% confidence intervals (shaded regions) showing NMD efficiency trends (individual data points not shown). Bottom left, increase in NMD efficiency across the first 200 coding nucleotides. Bottom right, variation in NMD efficiency around the last exon junction; the presumed location of the EJC is demarcated at the top of the plot. Exons >400 nt vs. exons ≤400 nt, P < 4.8 × 10−9; Fig. 3a). In these long exons, we found that NMD efficiency was best explained by the distance between the PTC and the downstream exon junction (Fig. 3b). It is conceivable that the large distance between the stalled ribosome and the EJC results in reduced physical contact with ribosome-bound UPF1, leading to lower degradation rates.

In addition to PTCs in long exons, we also found an unexpected association between reduced NMD efficiency and the distance from the PTC to the normal translation termination site (67% NMD efficiency for the 20% longest PTC-to-stop distances versus 99% efficiency for shorter distances, P = 1.3 × 10−7; Fig. 3c). This association was observed after controlling for start-proximal NMD evasion and the presence of downstream EJCs or long exons. Before mRNA decay, activated UPF1 helicase translocates to the 3′ end of mRNA to remodel the messenger ribonucleoprotein (mRNP), while scanning and unwinding the RNA at slow speeds of <1 bp/s (ref. 27). It is conceivable that UPF1 translocation may become rate limiting for NMD when the distance between the PTC and the normal translation termination site is large. These new rules were validated for frameshift indels and by using Geuvadis allele-specific expression data for germline variants (Supplementary Fig. 3).

mRNA turnover and sequence motifs modulate NMD efficiency

An additional feature used by the model was mRNA half-life. Transcripts with a short mRNA half-life had reduced NMD efficiency (1.2% NMD efficiency for mRNA with a half-life <1 h versus 74% efficiency for mRNA with a half-life ≥1 h, P = 6.3 × 10−4; Fig. 3d and Supplementary Fig. 3h,i), and this effect was also observed using mRNA half-life measurements from a different cell type (HeLa versus B cells; Supplementary Fig. 3k). These transcripts are already rapidly degraded, and enhanced degradation by NMD is likely to have only a small effect on steady-state mRNA levels. In addition, genes with higher wild-type expression levels tended to exhibit more efficient NMD when containing a PTC (Supplementary Fig. 3j); of note, our analyses excluded weakly expressed genes (≤5 transcripts per million, TPM) for which the ability to detect NMD may be overwhelmed by technical noise (Online Methods). As mRNA half-life and gene expression level correlate, it is likely that the reduced NMD efficiency for weakly expressed genes is confounded to some extent by shorter mRNA half-lives. We thus explicitly factored out expression levels from the NMD efficiency measure before further analysis and found a similar trend (Supplementary Fig. 3l.m), implying that short mRNA half-life is independently associated with inefficient NMD. Expression level estimates for weakly expressed genes with rapid turnover were noisier, thus increasing the observed variability in NMD efficiency after adjusting for expression levels. However, after pooling the data points into five expression level bins, a robust association was evident (Fig. 3d and Supplementary Fig. 3k).

We next tested whether known29 and de novo-inferred RNA-binding protein motifs (Online Methods) explain a significant portion of the remaining variance in NMD efficiency. We found nine significant motifs (Supplementary Fig. 4), of which four validated in both independent data sets, including binding sites for SRSF1, PABPN1, SNRPB2 and ACO1 (P < 0.10 in Geuvadis and in frameshifts; pooled P value < 0.005; see URLs; Fig. 3e and Online Methods). Existence of these motifs in proximity to a PTC or in the wild-type 3′ UTR sequence was associated with substantially altered NMD efficiency (15–48%) in either direction, and we found that these motifs might also regulate mRNA degradation in the absence of PTCs (Supplementary Fig. 5a–c; individual examples are discussed in the Supplementary Note).
As remodeling of mRNA–protein complexes is necessary for NMD, we also tested whether sequence composition (dinucleotide frequency) and mRNA secondary structure (RNA–RNA interaction probability per nucleotide; Online Methods) influence NMD efficiency. However, we found neither factor to be associated with NMD efficiency (Supplementary Table 2), consistent with high in vivo efficiency of the UPF1 helicase in translocating through structured RNA27. In addition, transcript features such as sequence conservation and the stop codon identity of either the PTC or wild-type stop codon did not affect NMD after controlling for other features (data not shown). Examined independently, codon usage biases associated with efficient translation showed a marginally significant association with lower NMD efficiency (62% NMD efficiency for the top 20% most biased genes versus 79% efficiency for other genes, $P = 0.04$; Supplementary Fig. 5d,e), in agreement with previous yeast data29. This effect was subtle and restricted to the most biased genes, consistent with previous estimates of the prevalence of selected codon usage30. Finally, gene expression is known to be potentially regulated by NMD in wild-type genes when transcripts contain a 3′ UTR intron or a translated upstream ORF in the 5′ UTR11. We found that such transcripts exhibited the expected NMD efficiency when additionally targeted for PTC-induced NMD (data not shown).

NMD model explains approximately three-fourths of variance in efficiency

To quantify how much variance in the efficiency of NMD can be predicted using the rules described above, we used random forest regression to compare the explained variance in NMD efficiency to the maximum observable variance, given the noise in the data (correction for attenuation; Online Methods). The regression model explained 74% of the variance in NMD efficiency by drawing on the general NMD features we have defined above (Fig. 4a). We next investigated how much every feature contributes to model fit to uncover the impact of each NMD rule. Adding the most significantly contributing features one by one showed that the canonical
EJC model was the most important predictor of NMD, accounting for nearly half of the observed variance in NMD efficiency. Start-proximal NMD evasion and inefficient NMD in long exons further explained 17% and 4% of the remaining variance, respectively. mRNA half-life and distance between the PTC and the wild-type stop codon each explained 1.6%, while the identified RNA-binding protein motifs accounted for 0.1% of the variance not explained by all previous features.

The amount of variance explained ($R^2$) is a compound measure of the accuracy of each predictor and its global coverage of individual examples (PTCs). This can manifest as a low $R^2$ value for very strong predictors with limited genomic coverage. For example, the 50-nt boundary rule is highly accurate in our data, meaning that all examples of PTCs that fall into this portion of the gene indeed show near-complete NMD insensitivity. However, the rule covers only a small part of the gene sequence and is thus not pertinent for many occurrences of PTCs in actual transcripts. Similarly, individual RNA-binding protein motifs may have a large effect on NMD efficiency (Supplementary Note), but their relatively rare occurrence means that they have only a subtle contribution to genome-wide prediction accuracy.

We reasoned that one explanation of why we observe residual variance in NMD efficiency could be compensatory changes in gene expression (feedback regulation) masking the effects of increased mRNA decay rates. For example, the expression levels of a gene containing an NMD-triggering PTC could be rescued by compensatory upregulation of the non-mutated allele. In Drosophila, it has been estimated that 50% of genes exhibit some degree of dosage compensation when heterozygously deleted, whereas in humans this effect still remains to be assessed in a systematic manner.

Cancer genomics data present an opportunity to globally assess the dosage compensation of human genes. In the absence of dosage compensation, somatic CNAs are expected to have a strong effect on gene expression levels. We thus compared the estimates of copy number gains and losses in tumors (from SNP arrays) to changes in gene expression (from RNA-seq) across 9,769 patients with cancer from the TCGA study. Next, we contrasted the 20% of genes with the strongest evidence of dosage sensitivity (highest correlation of gene copy number with expression; putatively non-compensated) with the 20% of genes with the least evidence of dosage sensitivity (low correlation). We could predict NMD efficiency threefold more accurately for dosage-sensitive genes ($R^2 = 85\%$ versus 29%, $P = 0.0051$; Fig. 4b). This suggests that feedback regulation of genes may account for a certain part of the unexplained variance in NMD efficiency (Fig. 4b).

### Negative selection acts on NMD-inducing somatic mutations

Having systematically determined the influences on NMD efficiency in human cells, we next examined possible roles in carcinogenesis. Positive selection on cancer-promoting driver mutations is well established in cancer genomes. However, negative (purifying) selection against detrimental mutations has proven difficult to detect. We reasoned that examining the variability in NMD efficiency among pooled PTCs should provide more power to detect purifying selection in cancer genomes.

To test for purifying selection on classes of genes, we quantified the ratio of nonsense mutations in NMD-sensitive regions to NMD-insensitive regions for each gene, after normalizing for the local synonymous mutation rate (Fig. 5a,b). This analysis showed that nonsense mutations were depleted in NMD-sensitive regions in comparison to the NMD-insensitive last exon and the first 250 nucleotides in oncogenes ($2.2\text{-fold}$ depletion, $P = 0.01$; Fig. 5c) and also in a set of human essential genes ($1.8\text{-fold}$ depletion in NMD-sensitive regions, $P = 0.04$; Fig. 5d). Next, we examined which other gene functional
categories showed depletion of NMD-triggering nonsense mutations. The top scoring gene ontology term was ‘regulation of cell proliferation’ (Fig. 5e). Other significant categories implicate the spliceosome, cell migration, angiogenesis and the endosome (odds ratio > 2, P < 3.2 × 10−4; Fig 5e), suggesting their importance for the fitness of cancer cells. Thus, accounting for variation in the ability of PTCs to trigger NMD suggests widespread purifying selection against detrimental mutations during the evolution of human tumors.

Positive selection for NMD targeting tumor-suppressor genes

In contrast to oncogenes, the inactivation of tumor-suppressor genes promotes tumor proliferation and survival. NMD of PTC-containing tumor suppressors should therefore be positively selected during tumor development. Indeed, PTCs were enriched in regions of tumor-suppressor genes predicted to trigger NMD (2.1-fold enrichment, \( P = 0.003; \) Fig. 5f). To systematically investigate how NMD contributes to the somatic inactivation of individual tumor-suppressor genes, we quantified how often an NMD-triggering PTC was accompanied by deletion in the same tumor sample, leading to complete gene inactivation (Supplementary Table 4). PCA of NMD and somatic deletion frequencies showed three broad clusters of tumor suppressors (Fig. 5g), which suggest a classification of mechanisms of inactivation.

First, approximately one-third of the examined tumor suppressors exhibited NMD combined with frequent heterozygous deletion, leading to biallelic inactivation (Fig. 5g,h), as expected for haplosufficient tumor suppressors inactivated by a classical ‘two-hit’ mechanism. A second cluster containing ~35% of the analyzed tumor-suppressor genes underwent deletion less frequently in patients carrying an allele with a PTC (Fig. 5g), suggesting that heterozygous inactivation by itself has functional consequences (haploinsufficiency). This category included NF1 (Fig. 5h), where haploinsufficiency is known to contribute to cancer. Lastly, there was a smaller cluster of tumor suppressors that were frequently mutated by PTCs but whose transcripts are likely ineffectively degraded by NMD, as most of their sequence is predicted to be NMD insensitive (Fig. 5g,h). Here the truncated protein could be partially functionally inactive or degraded, or could act as a dominant negative.

Finally, we examined the prevalence of NMD-mediated inactivation events in an extended list of putative tumor suppressors proposed previously on the basis of an excess of predicted loss-of-function mutations. For nine such genes with sufficient PTCs in our data \((n \geq 50)\), we did not find evidence for frequent two-hit inactivation mechanisms via simultaneous NMD and gene deletion (Supplementary Fig. 6a,b), consistent with these new cancer-related genes acting as haploinsufficient tumor suppressors.

**DISCUSSION**

In this study, we have used the data from close to 10,000 tumors to provide a systematic and unbiased evaluation of the rules that govern the efficiency of premature stop codons in triggering NMD in human cells. Elucidating these rules is important for the interpretation of clinical genetic data because, as we have illustrated for human tumors, nonsense mutations that do and do not trigger changes in mRNA levels can have very different functional consequences.

Our analyses confirmed the canonical EJC model as the most important individual determinant of NMD efficiency. However, they also identified other features important for predicting NMD efficiency globally or in individual transcripts. Moreover, we suggest rapid mRNA turnover and dosage compensation as possible reasons why, even if NMD is triggered, the overall expression levels of an mRNA may remain unchanged. Additional features important for predicting NMD genome-wide effects include distance of the PTC to the start codon, exon length and the presence of individual protein-binding motifs both close to a PTC and elsewhere in an mRNA.

The rules that we determined using somatic nonsense mutations from human tumors (summarized in Fig. 6a) were validated using an independent set of tumor frameshift mutations and also an independent set of germline variants in human lymphoblastoid cell lines. These rules are thus likely to apply widely across human tissues and diseases.

Finally, we showed that by accounting for variation in NMD we could detect both widespread positive and negative selection during the evolution of human tumors. Positive selection on cancer-promoting driver mutations is widely appreciated to contribute to cancer. Quantifying such positive selection on NMD-triggering versus NMD-evading PTCs allowed us to classify known and putative tumor-suppressor genes (Fig. 6c). Purifying selection against detrimental somatic mutations has, to our knowledge, previously only been detected for the mitochondrial genome, whereas we have shown widespread purifying selection against nonsense mutations that trigger NMD in both oncogenes and the general set of essential genes in humans (Fig. 6b). This has important implications for understanding tumor evolution because it supports the notion that subclones are eliminated during tumor progression because they carry detrimental mutations.

Taken together, this study provides important mechanistic insight into NMD and tumor evolution, as well as a broader framework for predicting the effects of nonsense variants in human disease.
METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by a European Research Council (ERC) Consolidator grant (616434), the Spanish Ministry of Economy and Competitiveness (BFU2011-26206 and ‘Centro de Excelencia Severo Ochoa 2013–2017’ SEV-2012-0208), the AXA Research Fund, the Agencia de Gestió d’Ajuts Universitaris i de Recerca (AGAUR), FP7 project 4DCell Fate (277899) and the EMBL-ERGG Systems Biology Program. F.S. was also supported by FP7 grants MAESTRA (ICT-2013-612944) and InnoMol (FP7-REGPOT-2012-2013-1-316289).

AUTHOR CONTRIBUTIONS

R.G.H.L. performed all analyses. R.G.H.L., F.S. and B.L. designed analyses, interpreted the data and wrote the manuscript. B.L. conceived the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Maquat, L.E. When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells. RNA 1, 453–465 (1995).
2. Mendell, J.T., Sharifi, N.A., Meyers, J.L., Martinez-Murillo, F. & Dietz, H.C. Nonsense-mediated mRNA decay: a post-translational regulatory switch on UPF1 controls targeted translation of mammalian mRNAs. EMBO J. 17, 3484–3494 (1998).
3. Leir, H., Hafid, D., Izaurralde, E. & Moore, M.J. The exon–exon junction complex and Upf complex recruitment governs human nonsense-mediated mRNA decay. Sci. Signal. 6, ra111 (2013).
4. Eberle, A.B., Stalder, L., Mathys, H., Orozco, R.Z. & Möhlemann, O. Posttranscriptional gene regulation by spatial rearrangement of the 3′ untranslated region. PLoS Biol. 6, e92 (2008).
5. Zhang, J. & Maquat, L.E. Evidence that translation reinitiation abrogates nonsense-mediated mRNA decay in mammalian cells. EMBO J. 16, 826–833 (1997).
6. Romão, L. et al. Nonsense mutations in the human β-globin gene lead to unexpected levels of cytoplasmic mRNA accumulation. Blood 90, 2895–2901 (2002).
7. Silva, A.L., Ribeiro, P., Inácio, A., Liebhaber, S.A. & Romão, L. Proximity of the poly(A)-binding protein to a premature termination codon inhibits mammalian nonsense-mediated mRNA decay. RNA 14, 563–576 (2008).
8. Lopalaíñen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–511 (2013).
9. rivas, M.A. et al. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science 348, 666–669 (2015).
10. martincroce, I. & campbell, P.J. Somatic mutation in cancer and normal cells. Science 349, 1483–1489 (2015).
11. Kuróski, T. et al. A post-translational regulatory switch on UPF1 controls targeted mRNA degradation. Genes Dev. 28, 1900–1916 (2014).
12. Singh, G., Rebbapragada, I. & Lysk-Reanders, J.A. A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay. PLoS Biol. 6, e111 (2008).
13. Tani, H. et al. Identification of hundreds of novel UPF1 target transcripts by deep sequencing of whole transcriptome stability. RNA Biol. 9, 1370–1379 (2012).
14. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 3, 1724–1735 (2007).
15. Fiorini, F., Bagchi, D., Le Hir, H. & Croquette, V. Human Upf1 is a highly processive RNA helicase and translocase with RNP remodelling activities. Nat. Commun. 6, 7581 (2015).
16. Ray, D. et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature 499, 172–177 (2013).
17. Zhang, Z. et al. Nonsense-mediated decay targets have multiple sequence-related features that can inhibit translation. Mol. Syst. Biol. 6, 442 (2010).
18. Supé, F., Sunčka, N., repar, J., vlaho, V., koc, J. & Smuc, T. Translational selection is ubiquitous in prokaryotes. PLoS Genet. 6, e1001004 (2010).
19. Rehwinkel, J., raes, J. & Izaurralde, E. Nonsense-mediated mRNA decay: target genes and functional diversification of effectors. Trends Biochem. Sci. 31, 639–646 (2006).
20. Malone, J.H. et al. Mediation of Drosophila autosomal dosage effects and compensation by network interactions. Genome Biol. 13, r28 (2012).
21. fehrmann, R.S. et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat. Genet. 47, 115–125 (2015).
22. Kurosaki, T. Somatic mutation in cancer and normal cells. Science 349, 1483–1489 (2015).
23. tani, H. et al. Identification of hundreds of novel UPF1 target transcripts by deep sequencing of whole transcriptome stability. RNA Biol. 9, 1370–1379 (2012).
ONLINE METHODS

Data acquisition. Somatic nonsense and frameshift mutations aligned to hg19/GRCh37 for available exome sequences of tumors (n = 9,769) together with matching mRNA sequencing data were downloaded on 20 April 2015 from the TCGA data portal. CNA and non-negative matrix factorization (NMF) clustering data (based on mRNA levels) for these tumor samples were downloaded from the Broad Institute ‘TCGA Genome Data Analysis Center Firehose pipeline ‘analyses __2014_10_17’. A frameshift mutation results in the translation of an out-of-frame peptide, which will often lead to the incorporation of a premature stop codon. We took the frameshifted sequence to calculate when the translation would prematurely stop (median distance frameshift to PTC = 47 nt) and used these PTGs as a separate validation data set. Germline nonsense mutations with corresponding allele-specific expression values for 462 lymphoblastoid cell lines from the Geuvadis RNA-seq project were downloaded from EBI ArrayExpress under accession E-GEUV-1.

Dominant isoforms. The mutations were reannotated to dominant isoforms in the ‘knownGene’ database provided in the ‘June2011’ bundle at the TCGA data portal. A gene had a dominant isoform if the expression level of the isoform with the highest expression was at least two times higher than that of the second most highly expressed isoform or if the gene had a single annotated isoform. Dominant isoforms were selected for each tumor type to ensure that the RNA-seq signal of other isoforms of the same gene would not mask changes in expression caused by premature stop codons leaving 15,890–16,506 dominant transcripts, depending on the tumor type.

Quantifying NMD efficiency. NMD efficiency in TCGA data sets was quantified by comparing the mRNA expression level in TPM of a PTG-bearing transcript to the median mRNA expression level of the same transcript in tumor samples where the PTC was absent. To minimize interference from intertumoral heterogeneity when calculating the median mRNA expression level for a transcript, we first separated the samples by tumor type and then used the most robust consensus NMF clustering (Broad GDAC Firehose) to further separate tumor samples. Within these clusters, a median mRNA expression value was calculated for every transcript from the samples that had no nonsense mutations, indel mutations or copy number variations overlapping this transcript. The mRNA expression level of every PTG-bearing transcript was then divided by the median mRNA expression level of that transcript, which was –log2 transformed to get NMD efficiency values where 0 indicates no mRNA degradation and 1 indicates complete heterozygous mRNA degradation by NMD.

NMD efficiency in the Geuvadis data set was quantified by dividing the allele-specific expression of a PTG-bearing allele by that of the reference allele. Allele-specific expression values were taken from the Geuvadis study, which were determined as described in Pirinen et al. In short, for every SNP, the RNA-seq reads were aligned to both the reference genome and the mutated genome, allowing for no mismatches. The –log2-transformed ratio of the reads mapping to the variant allele divided by the reference allele reads is here defined as NMD efficiency, where 0 indicates no mRNA degradation and complete heterozygous mRNA degradation approaches infinity. In comparison to TCGA, the Geuvadis data set contains six times fewer PTGs at unique loci, which are not enough data points for rigorous validation for NMD rules that only apply to a very small part of the gene sequence (Supplementary Figs. 2a and 3e).

Although quantifying NMD efficiency with allele-specific expression has the major advantage that there is no need to compare different samples, it comes at the cost of having fewer data points to accurately approximate the expression level. Normal mRNA expression values are derived from all RNA-seq reads aligning to any part of the transcript sequence, which can be thousands of sequence reads. In contrast, allele-specific expression is based on only the sequence reads that overlap with the SNP. Therefore, we investigated the rules of NMD with NMD efficiency measures derived from mRNA expression values (TCGA nonsense mutations, method 1) and validated the findings using an independent set of PTGs with mRNA expression values (TCGA frameshift-induced PTGs, method 1) and also the allele-specific expression–derived NMD efficiency measures (Geuvadis nonsense mutations, method 2) to evaluate any method-related biases. The effect sizes of the characterized NMD rules supplied in the text are calculated by comparing the maximum observable NMD efficiency to the median NMD efficiency of the PTGs affected by the respective NMD rule. Here the maximum observable NMD efficiency is defined as the median NMD efficiency of NMD-inducing PTGs when taking the known rules into account and used as 100% NMD efficiency. Because we characterize NMD rules in a nested fashion throughout this manuscript, wherein we control for all previously examined effects while measuring the next one, the maximum observable NMD efficiency increases throughout this manuscript as new rules are added.

Variant filtering and noise reduction. Stringent filtering was applied on the TCGA data when calculating NMD efficiency. PTG-bearing transcripts were removed if they contained additional indel, nonsense or splice-site-disrupting (±3 nt from an exon junction) variants or had overlap with a CNA. Furthermore, nonsense and frameshift mutations needed to be present at an allele frequency of 20% or higher. When calculating the median expression level of wild-type transcripts to which to compare the expression of a PTG-bearing transcript, we excluded transcripts with a median expression level lower than five and transcripts with large variation in expression among the tumor samples (coefficient of variation > 0.5). A threshold of at least 10 wild-type measurements (samples) per transcript per NMF cluster was used to calculate an accurate median expression value. Lastly, to increase the power of the random forest regression to detect new NMD rules, we reduced variability in the expression data by removing hidden covariates with PCA. Several studies have shown that PCA can successfully increase power to associate gene expression changes with SNVs by removing non-random sources of noise.

We applied PCA to the mRNA expression data and took out the first four principal components for every tumor type. These four components comprise sources of unwanted technical of biological variation but not variation introduced by individual SNVs.

To accurately determine NMD efficiency on the basis of allele-specific expression, we only included variants that had at least one variant read and at least ten mRNA-seq reads in total. To prevent overfitting on highly prevalent nonsense mutations, we collapsed SNVs that were found in more than one individual and used the median NMD efficiency for further analyses.

Data analysis. All analyses were performed in Python 2.7.2 and R 3.1.2 (R Core Team). PCA was done with the FactoMineR package in R, where we modified the reconst function to take out principal components. We used the randomForest package for random forest regression and the rIPermute package to compute P values for feature contribution. Unless stated otherwise, a two-tailed Mann–Whitney test was used to compute P values. When P values were pooled across multiple tests, a formula described previously was used. Briefly, P1, P2 and P3 are the three P values to combine, ρ is their product, P2/P1 and the combined P value is found using the formula

\[
P_c = \rho^{1 + \log \left( 1 + \frac{1}{2} \left( \log \frac{P_1}{P_2} \right)^2 \right)}
\]

In Supplementary Figures 1–3, the blue lines show the fits to the data point using the R package ggplot, function geom_smooth. In cases with ≤1,000 points, this uses the R function loess with default parameters (linear fit, using the smoothing parameter width of 0.75). In cases with ≥1,000 points, the R function gam is used, invoking a generalized additive model with smoothing terms represented by penalized regression splines. The shaded areas are the 95% confidence intervals of the fit.

In all figures, box plots are drawn using the defaults of the R package ggplot2, the center line is the median of the data distribution, the notch around this line is the approximate 95% confidence interval of the median, 1.5 × IQR or quartile (IQR), the hinges are the first and third quartiles, and the whiskers extend to the lowest and highest non-outlying values (those that are within 1.5 × IQR of the upper or the lower hinge).

Modeling NMD. Random Forest regression was used to identify gene features that influence NMD efficiency. Random Forest is a robust and easily interpretable machine learning approach that accepts mixed data types and internally
controls for overfitting. To learn about new features that influence NMD efficiency, we first factored out the major rules of NMD that we found manually. In particular, we predicted NMD efficiency with a Random Forest regression having the knowledge whether the PTC was located on the last exon, 30 nt before the last exon, in proximity to the start codon and the allele frequency of the nonsense mutation (RandomForest package, 1,000,000 trees, mtry = 1). We subtracted the predicted NMD efficiencies from the original values to remove the confounding effects of these known NMD rules. Next, we used Random Forest regression again to predict these adjusted NMD efficiencies from 2,390 gene features hypothesized to influence NMD efficiency (RandomForest package, 100,000 trees, mtry = 13). Here, the 100 most important features (highest percentage increase in mean squared error) were used to determine which features contribute significantly at \( P < 0.05 \) to the NMD efficiency prediction (rPermute package, 10,000 trees, 1,000 permutations), resulting in eight significantly contributing features that were selected for further manual inspection.

To build our random forest model, we made a comprehensive table of different features that we hypothesized to influence NMD efficiency (Supplementary Tables 1 and 2). Sequence conservation of all vertebrates in phyloP46way downloaded from the UCSC Genome Browser website was used to create the sequence conservation features. A basewise probability score for the mRNA secondary structure was computed with Rfold 0.1-2. Optimal codon usage was used as a proxy for translation efficiency and was computed by dividing the amount of optimal in-frame codons by the codon count. Optimal codons were defined as the ones corresponding to the tRNA genes with the highest amount of optimal in-frame codons by the codon count. Optimal codon usage with known RNA-binding protein motifs from Ray de novo motifs that influence NMD efficiency, by ranking the PTC-bearing transcripts by NMD efficiency and comparing the top 10% to the bottom 10% of the transcripts for motif enrichment. De novo motifs were pooled with known RNA-binding protein motifs from Ray et al.28 and used to scan mRNA sequences. Motifs that were found to significantly contribute to the random forest were manually investigated to determine a threshold for the binding score. Mono- and dinucleotide frequencies were computed by counting the occurrence of the 4 nucleotides and all 64 dinucleotide combinations, respectively. The above described features were computed for the following regions of interest for each transcript: ±100 nt of the PTC, ±100 nt of the exon junction downstream of the PTC, the 3′ UTR, the 5′ UTR, the region between the PTC and the normal stop codon and the whole transcript sequence. In addition, RNA motif scanning was also performed in the first and last 200 nt of each transcript. For sequence conservation, RNA folding score and codon optimality, the median score for each region was used as input for the random forest regression. The mRNA half-life feature is based on Friedel et al.44 RNA half-lives in human B cells (Supplementary Table 2) and is validated with mRNA half-life measures in HeLa cells.47 A list of translated upstream ORFs from Andreev et al.48 was kindly provided by P.B. O’Connor.

A correction for attenuation procedure was used to remove the measurement error to investigate how much of the variation in NMD efficiency could be explained with the rules we defined. A Spearman correlation on the NMD efficiency of replicated data points (same nonsense mutation in several tumor samples) showed 22.5% explainable variance in the data. Next, we used random forest regression (randomForest package, 100,000 trees, mtry = 3) to determine the added explained variance for each feature, starting with the most important one. Including all validated rules in the random forest explained 16.6% of the variance, which equals 74% of the explainable variance in NMD efficiency.

**Purifying selection.** To test whether there is purifying selection in essential and cancer-related genes, we defined the first 250 nt, the 50 nt before the last exon and the whole last exon of a transcript as NMD-insensitive regions and compared the nonsense mutation density to that of the NMD-sensitive region. The nonsense mutation density was normalized to the synonymous mutation density, to correct for differences in the mutational landscape. 2 s.d. from 10,000 bootstrap samples are shown in the error bars, while empirical \( P \) values were computed with 10,000 permutations. The 200 most essential genes published by Hart et al.40 were used as essential genes, and 220 oncogenes were taken from the Cancer Gene Census49 (Molecular Genetics = ‘Dom’), after manual removal of leukemia-specific cancer genes. To test whether genes that promote carcinogenesis when successfully silenced are enriched for nonsense mutations in NMD-sensitive regions, we took the 200 highest scoring STOP genes from Solimini et al.41. Gene Ontology enrichment analysis for gene sets under negative selection of somatic nonsense mutations was performed on UniProt-GOA release 154 (ref. 49) by calculating the log2-transformed odds ratio of NMD-sensitive and NMD-insensitive somatic nonsense mutations normalized for the synonymous mutation densities for every gene set bigger than 30 genes.

42. González-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 14, R70 (2013).
43. Pirinen, M., et al. Assessing allele-specific expression across multiple tissues from RNA-seq read data. Bioinformatics 31, 2497–2504 (2015).
44. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput. Biol. 6, e1000770 (2010).
45. Supek, F., Milošević, B., Valcarcel, J., Gabaldón, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. Cell 156, 1324–1335 (2014).
46. Friedel, C.C., Dölken, L., Ruzsics, Z., Koszinowski, U.H. & Zimmer, R. Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. Nucleic Acids Res. 37, e115 (2009).
47. Tani, H. et al. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res. 22, 947–956 (2012).
48. Andreev, D.E. et al. Translation of 5′ leaders is pervasive in genes resistant to eIF2 repressor. eLife 4, e03971 (2015).
49. Huntley, R.P. et al. The GOA database: gene ontology annotation updates for 2015. Nucleic Acids Res. 43, D1057–D1063 (2015).