Published in final edited form as:
Nat Genet. 2016 October ; 48(10): 1112–1118. doi:10.1038/ng.3664.

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

Rik G.H. Lindeboom[1,2,5], **Fran Supek**[1,2,3], and **Ben Lehner**[1,2,4]

[1]EMBL-CRG Systems Biology Unit, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 08003 Barcelona, Spain

[2]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain

[3]Division of Electronics, Rudjer Boskovic Institute, 10000 Zagreb, Croatia

[4]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

Abstract

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-quarters of the non-random variance in NMD efficiency across thousands of PTCs. We also show that dosage compensation may 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 of tumor suppressor genes.

Introduction

The nonsense-mediated mRNA decay (NMD) pathway is an mRNA surveillance system that protects eukaryotic cells by reducing the production of harmful truncated proteins translated from premature termination codon (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.

Correspondence should be addressed to B.L. (ben.lehner@crg.eu) and F.S. (fran.supek@crg.eu).
[5]Present address: Department of Molecular Biology, Radboud Institute for Molecular Life Sciences, 6525GA Nijmegen, the Netherlands

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 paper; B.L. conceived the project.

Competing financial interests
The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts

Therefore, a complete understanding about how the NMD machinery selects which transcripts to degrade is crucial for predicting the phenotypic consequences of nonsense mutations in humans.

The canonical model for how PTC-containing transcripts are recognized in mammalian cells is the exon junction complex (EJC) model, which was defined by introducing PTCs into reporter genes6,7. EJC proteins are deposited at exon-exon junctions by the splicing machinery but, after nuclear export, are stripped off mRNAs by the translating ribosome. If a PTC is present at least 50 nucleotides (nt) upstream of the last exon junction, an EJC will remain bound after the pioneering round of translation has terminated8. The interaction between termination factors and the downstream EJC subsequently triggers NMD9.

While the EJC model is well supported, exceptions have been reported10,11,12, highlighting the possible relevance of an alternative "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* and *Caenorhabditis elegans*15,16. One further exception to the canonical EJC model has been observed with individual examples of human genes bearing PTCs close to the start codon, rendering them insensitive to NMD17,18, even with downstream EJCs present. It was hypothesized that 5′ proximal PTCs circumvent NMD by re-initiating translation downstream, or, alternatively, that the proximity to poly(A) binding protein in a closed mRNA loop inhibits NMD17,19. It is not clear how widely these exceptions from 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 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 the 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 ~3/4 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 of human tumor suppressor genes.

Results

To systematically examine determinants of NMD efficiency we compiled a dataset of somatic nonsense mutations from 9,769 cancer patients with available exome sequences, copy number alteration (CNA) and mRNA expression data 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

(Methods). The NMD efficiency of each nonsense mutation was estimated as the fold-change in mRNA expression level compared to the median expression of 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 frequency of the somatic variants in each tumor, high expression levels and low variation of the wild-type gene across samples (Methods). This resulted in 2,840 high-confidence nonsense mutations with 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 the TCGA tumor samples (3,151 indel-induced PTCs in 1,957 genes from 1,156 patients after filtering) and heterozygous germline nonsense variants in 452 lymphoblastoid cell lines from the Geuvadis study20 where the NMD efficiency was quantified as allele-specific mRNA expression of the reference over the variant allele21 (1,784 high-confidence PTCs in 487 unique loci after filtering on 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 *versus* PTCs inside the last exon (Fig. 2b,e; Supplementary Fig. 1c,d) shows that NMD is, overall, indeed inefficient in the latter case (1.7% *vs.* 84.7% NMD efficiency – calculated as observed NMD efficiency over maximum observable NMD efficiency using our NMD model (Methods), $p < 2.2 \times 10^{-16}$). We also observe a sharp decrease towards a complete NMD insensitivity over the 50 nt boundary upstream of the last exon, demarcating the physical location of the EJC (2.2% *vs.* 84.7% NMD efficiency, $p = 5.3 \times 10^{-5}$). Thus, a downstream EJC is a widespread signal for efficient NMD in many human genes. Consistently, intronless transcripts that bear PTCs are not degraded by NMD (Supplementary Fig. 1f), as expected from their lack of deposited EJCs.

Next, we examined whether the yeast/*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 to PTC proximity to the transcript 3′ end (Supplementary Fig. 1e,f, p = 0.92, t-test for significance of Pearson R). This conclusion is also upheld when using frameshift-inducing indels in TCGA or germline truncation variants (Supplementary Fig. 1g,h, p > 0.54). Thus, while faux 3′ UTR-related mechanisms may be important in individual cases23,24, they do not appear to be a prevalent influence on NMD efficiency in human cells. Consistently, a study in HeLa cells did not observe a relationship between transcript 3′ UTR length and mRNA turnover rates in transcripts targeted by the key NMD factor UPF125.

However, transcripts harboring an intron within the 3′ UTR show a different trend, where PTCs in the second-last exon do not trigger NMD despite the presence of a downstream exon junction (7% efficient NMD for PTCs in penultimate exon *vs.* 81% in the upstream exons, p = 0.045; 12% *vs.* 81% for frameshift indels, p = 0.006; Fig. 2c,e, 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 second-last splice site.

Start-proximal PTCs avoid NMD by reinitiating translation

Previously, PTCs introduced into the first half of exon 1 of the β-globin gene were shown not to reduce mRNA abundance18, suggesting that the NMD machinery did not recognize these transcripts. Our cancer genome data support the generality of this result, with a widespread decrease in NMD efficiency when PTCs are located in close proximity to the start codon (Fig. 2d,e, 35% *vs.* 93% NMD efficiency for PTCs within 200 nt of the start codon *vs.* >200 nt from the start codon, $p < 2.2 \times 10^{-16}$; Supplementary Fig. 2b,c). PTCs located in the first 100 coding nts are rarely targeted for NMD, followed by a gradual increase in NMD efficiency up to 200 nt from the start codon (Fig. 2d,e). This start-proximal NMD insensitivity is not explained by the canonical EJC model of NMD because the downstream EJCs fail to trigger NMD. The same conclusion is reached analyzing somatic frameshift data (Supplementary Fig. 2b, $p = 5.5 \times 10^{-12}$) and allele-specific expression of germline variants (Supplementary Fig. 2c, $p = 0.0001$). Two mechanisms have been proposed to account for the examples of NMD insensitivity in transcripts with start-proximal PTCs: transcript stabilization by an interaction between the poly(A) binding protein C1 (PABPC1) and the translation termination machinery19 or a bypass of NMD by the re-initiation of translation at a downstream AUG codon17.

Comparing the reading frame of the first AUG downstream from the PTCs shows NMD insensitivity in transcripts with an in-frame AUG downstream of the PTC (Fig. 2a,e, 1.9% *vs.* 57% NMD efficiency, $p = 3.4 \times 10^{-4}$; Supplementary Fig. 2d,e), suggesting that translation re-initiation is a widespread mechanism for evading NMD in human genes. The distance between the original start codon and the PTC is a strong predictor for the start-proximal NMD insensitivity, while the distance to the downstream AUG or the Kozak sequence strength of the downstream AUG does not affect the NMD efficiency (Supplementary Fig. 2f,g). However, we observe that efficiency of NMD is still reduced for start-proximal PTCs even in those cases when there are no in-frame downstream AUG codons (57% *vs.* 93% NMD efficiency), suggesting that other mechanisms can result in NMD bypass. Shorter distances between the best match of the PABPC1 motifs and the PTCs in a hypothetical looped mRNA conformation did not associate with lower NMD efficiency (Supplementary Fig. 2h-k).

Long exons and the distance to stop codon attenuate NMD

While the start codon proximity and the presence of downstream EJCs explain a substantial part of the variance in NMD efficiency, we hypothesized that there could be additional rules governing whether a nonsense mRNA is targeted by NMD. To elucidate these, we first factored out the known effects of the start proximity and the presence of downstream EJCs on the NMD efficiency (Supplementary Fig. 3f,g). Moreover, we filtered out possible sources of systematic bias in gene expression by a procedure based on principal component analysis26 (Methods).

Next, we used Random Forest (RF) regression to predict NMD efficiency using a large set of gene features of plausible importance for NMD. Overall, eight different features contributed to NMD ($p < 0.05$, RF permutation test; features listed in Supplementary Table 1 and 2, significant features are listed in Supplementary Table 3). As a trivial example that serves as a validation of our methodology, the estimate of NMD efficiency is higher when the variant allele frequencies in the tumor are higher (Supplementary Fig. 3a); our further Random Forest analyses control for this covariate, while searching for rules that govern NMD.

We found a significant reduction of NMD efficiency when PTCs are located in exceptionally long exons (Fig. 3a, 61% *vs.* 98% NMD efficiency, $p < 4.8 \times 10^{-9}$, for exons >400 nt). In these long exons, we found that the NMD efficiency is 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 the 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 (Fig. 3c, 99% *vs.* 67% NMD efficiency in the top 20% longest PTC-to-stop distances, $p = 1.3 \times 10^{-7}$). This is observed after controlling for the start-proximal NMD evasion and the presence of downstream EJCs or long exons. Prior to mRNA decay, the activated UPF1 helicase translocates to the 3′ end of the mRNA to remodel the messenger ribonucleoprotein (mRNP), while scanning and unwinding the RNA at slow speeds of <1bp/s27. It is conceivable that UPF1 translocation may become rate-limiting for NMD when the distance between the PTC and the normal termination site is large. These new rules validated for frameshift indels and by using Geuvadis allele-specific expression of 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 have a reduced NMD efficiency (74% *vs.* 1.2% NMD efficiency when mRNA half-life <1 hour, $p = 6.3 \times 10^{-8}$, Fig. 3d; Supplementary Fig. 3h,i) and this effect is also observed using mRNA half-life measurements from a different cell type (HeLa *vs.* B-cells; Supplementary Fig. 3k). These transcripts are already rapidly degraded and enhanced degradation by NMD is likely to only have a small effect on the steady-state mRNA level. In addition, genes with higher wild-type expression levels tend to exhibit more efficient NMD when containing a PTC (Supplementary Fig. 3j); of note, our analyses exclude the lowly expressed genes (≤5 transcripts-per-million, TPM) where the ability to detect NMD may be overwhelmed by technical noise (Methods). As mRNA half-life and gene expression level correlate, it is likely that the reduced NMD efficiency in lowly expressed genes is confounded by shorter mRNA half-lives to some extent. We thus explicitly factored out the expression levels from the NMD efficiency measure prior to 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 of lowly-expressed, rapid-turnover genes are noisier, thus increasing the observed variability of NMD

efficiency after adjusting for expression levels. However after pooling the points into five expression level bins a robust association is evident (Fig. 3d; Supplementary Fig. 3k).

We next tested whether known28 and *de novo* inferred RNA-binding protein motifs (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 – the binding sites for SRSF1, PABPN1, SNRPB2 and ACO1 (Fig. 3e; $p < 0.10$ in Geuvadis and in frameshifts; pooled p-value < 0.005; see URLs, Methods). Existence of these motifs in proximity to a PTC or in the wild-type 3′ UTR is associated with substantially altered NMD efficiency (15%–48%) in either direction, and we find they may also regulate mRNA degradation in the absence of PTCs (Supplementary Fig. 5a-c; individual examples are discussed in Supplementary Note).

As remodeling of mRNA-protein complexes is necessary for NMD, we also tested whether sequence composition (dinucleotide frequencies) and mRNA secondary structure (RNA-RNA interaction probability per nt; Methods) impact NMD efficiency. However, we found neither factor to be associated with NMD efficiency (Supplementary Table 2), consistent with a high *in vivo* efficiency of the UPF1 helicase in translocating through structured RNA27. In addition, transcript features such as sequence conservation or the stop codon identity of either the PTC or wild-type stop codon did not affect NMD after controlling for other features (not shown). Examined independently, codon usage biases associated with efficient translation show a marginally significant association with lower NMD efficiency (79% *vs.* 62% NMD efficiency in the top 20% most biased genes, $p = 0.04$; Supplementary Fig. 5d,e), in agreement with previous yeast data29. This effect is 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 open reading frame in their 5′ UTR31. We found that such transcripts exhibit normal NMD efficiencies when additionally targeted for PTC-induced NMD (not shown).

NMD model explains ~3/4 of the variance in efficiency

To quantify how much variance in the efficiency of NMD can be predicted using the rules and features 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; Methods). The regression model explains 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 the prediction to uncover the impact of each NMD rule. Adding the most significantly contributing features one-by-one shows that the canonical EJC model is the most important predictor of NMD, accounting for nearly 1/2 of the observed variance in NMD efficiency. Start-proximal NMD evasion and inefficient NMD in long exons further explain 17% and 4% of the remaining variance, respectively. The mRNA half-life and the distance between the PTC and the *wild-type* stop codon each explain 1.6%, while the identified motifs account for 0.1% of the variance not explained by all previous features.

The amount of explained variance (R^2) is a compound measure of the accuracy of each predictor and also of its global coverage of individual examples (PTCs). This can manifest as a low R^2 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 PTC that fall into this portion of the gene indeed show a 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 they have only a subtle contribution to genome-wide prediction accuracy.

We reasoned that one explanation for 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 level of a gene containing an NMD-triggering PTC could be rescued by compensatory up-regulation of the non-mutated allele. In *Drosophila*, it has been estimated that 50% of genes exhibit some degree of dosage compensation (DC) when heterozygously deleted32, while in humans this effect still remains to be assessed in a systematic manner.

Cancer genomics data present an opportunity to globally assess the DC of human genes. In the absence of DC, somatic copy number alterations (CNA) are expected to have a strong effect on gene expression levels33. We thus compared the estimates of copy number gains and losses in tumors (from SNP arrays) to changes in gene expression (by RNA-Seq) across 9,769 cancer patients from the TCGA study. Next, we contrasted the 20% of genes with the strongest evidence of dosage-sensitivity (highest correlation of CNA to expression; putatively non-compensated), with the 20% of genes with least evidence of dosage-sensitivity (low correlation; Methods). We can predict NMD efficiency four-fold more accurately for dosage-sensitive genes (Fig. 4b; R^2 = 85% *vs.* 29%, p = 0.0051). This suggests that feedback regulation of a gene 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 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 genomes22. However, negative (purifying) selection against detrimental mutations has proven difficult to detect34. We reasoned that accounting for the variability of NMD-efficiency amongst 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 to NMD-insensitive regions of each gene, after normalizing for the local synonymous mutation rate (Fig. 5a,b). This showed that nonsense mutations are depleted in NMD-sensitive regions compared to the NMD-insensitive last exon and first 250 nt (2.2-fold depletion, p = 0.01, Fig. 5c) in oncogenes and also in a set of human essential genes (1.8-fold depletion in the NMD-sensitive regions, p = 0.04, Fig. 5d). Next, we examined which other gene functional categories show 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 (Fig. 5e, odds ratio > 2, $p < 3.2 \times 10^{-4}$), 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 and tumor suppressor genes

In contrast to oncogenes, the inactivation of tumor suppressor genes promotes tumor proliferation and survival. NMD of PTC-containing transcripts should therefore be positively selected during tumor development35. Indeed, PTCs are enriched in regions of tumor suppressor genes predicted to trigger NMD (2.1-fold increase, $p = 0.003$, 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 is accompanied by the deletion of the other allele, leading to a complete inactivation (Supplementary Table 4). A principal component analysis on the NMD and somatic deletion frequencies shows three broad clusters of tumor suppressors (Fig. 5g), which suggest a classification of mechanisms of inactivation.

First, ~1/3 of the examined tumor suppressors exhibit NMD combined with frequent heterozygous deletion, leading to biallelic inactivation (Fig. 5g-h, bold circles) as expected for haplosufficient tumor suppressors inactivated by a classical 'two-hit' mechanism. A second cluster containing ~35% of the analyzed tumor suppressor genes undergo deletion less frequently in patients carrying a PTC (Fig. 5g, asterisks), suggesting that a heterozygous inactivation already has functional consequences in some cases (haploinsufficiency). Examples include *NF1* (Fig. 5h), where haploinsufficiency is known to contribute to cancer36. Lastly, there is a smaller cluster of tumor suppressors that are 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, triangles). Here the truncated protein could be partially functionally inactive, degraded, or 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 based on an excess of predicted loss-of-function mutations37. For nine such genes with sufficient PTCs in our data ($n \geq 50$), we do not find evidence for frequent two-hit inactivation mechanisms via simultaneous NMD and gene deletion (Supplementary Fig. 6a,b) consistent with these novel cancer genes acting as haploinsufficient tumor suppressors37.

Discussion

In this study we have used the data from close to ten thousand 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 or do not trigger changes in mRNA levels can have very different functional consequences.

Europe PMC Funders Author Manuscripts

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 genome-wide 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 level of an mRNA may remain unchanged. The additional features important for predicting NMD genome-wide effects included distance 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 lines20. 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 as contributing to cancer. Quantifying such positive selection on NMD-triggering *vs* 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 genome34, while we have shown widespread purifying selection against nonsense mutations that trigger NMD in both oncogenes and the general set of essential genes in human (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 mutations38.

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.

URLs

Method for pooling p-values, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.4686.

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 20-04-2015 from The Cancer Genome Atlas (TCGA) data portal. Copy number alteration (CNA) and nonnegative 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

Europe PMC Funders Author Manuscripts

when the translation would prematurely stop (median distance frameshift to PTC = 47 nt) and used these PTCs as a separate validation dataset. Germline nonsense mutations with corresponding allele-specific expression values for 462 lymphoblastoid cell lines from the Geuvadis RNA-sequencing project were downloaded from EBI ArrayExpress under accession E-GEUV-1.

Dominant isoforms

The mutations were re-annotated 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 is at least two times higher than the second-highest expressed isoform42, or if the gene had a single annotated isoform. Dominant isoforms were selected per 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 the TCGA datasets was quantified by comparing the mRNA expression level in transcripts-per-million (TPM) of a PTC bearing transcript to the median mRNA expression level of the same transcript in tumor samples where the PTC was absent. To minimize interference from inter-tumor 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 PTC-bearing transcript was then divided by the median mRNA expression level of that transcript, which was $-\log_2$ transformed to get NMD efficiency values where 0 indicates no mRNA degradation and 1 complete heterozygous mRNA degradation by NMD.

NMD efficiency in the Geuvadis dataset was quantified by dividing the allele-specific expression of a PTC bearing allele by the reference allele. Allele-specific expression values were taken from the Geuvadis study20, which were determined as described in Pirinen *et al.* 43. In short, for every SNV the RNA-Seq reads were aligned to both the reference genome and the mutated genome, allowing for no mismatches. The $-\log_2$ transformed ratio of the reads mapping to the variant allele divided over the reference allele reads is here defined as NMD efficiency, where 0 indicates no mRNA degradation and complete heterozygous mRNA degradation approaches infinity. Compared to TCGA, the Geuvadis dataset contains six times fewer PTCs at unique loci, which is 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).

While 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 SNV. Therefore we investigated the rules of NMD with NMD efficiency measures derived from mRNA expression values (TGCA nonsense mutations, method 1) and validated the findings using an independent set of PTCs with mRNA expression values (TCGA frameshift induced PTCs, 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 PTCs affected by the respective NMD rule. Here, the maximum observable NMD efficiency is defined as the median NMD efficiency of NMD inducing PTCs when taking the known rules into account and used as 100% NMD efficiency. Since 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 the NMD efficiency. PTC 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 all transcripts to compare the expression of a PTC bearing transcript with, 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 principal component analysis (PCA). Several studies have shown that PCA can successfully increase power to associate gene expression changes with SNVs by removing non-random sources of noise26,44. 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 based on 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 ‘rfPermute’ 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 (see URLs) was used. Briefly, p_1, p_2 and p_3 are the three p-values to combine, ρ is their product $p_1p_2p_3$, and the combined p-value is found thusly:

$$p_c=\rho\left\{1+\log\frac{1}{\rho}+\frac{1}{2}\left(\log\frac{1}{\rho}\right)^2\right\}$$

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 95% confidence interval of the fit.

In all figures, box plots are drawn using the defaults of the R package ggplot: the center line is the median of the data distribution, the notch around this line is the approximate 95% C.I. of the median, 1.58 * IQR / sqrt(n), the hinges are the 1st and the 3rd quartile, and the whiskers extend to the lowest/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, 50 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 (‘rfPermute’ 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 the translation efficiency and was computed by dividing the amount of optimal in-frame codons over the codon count. Optimal codons were defined as the ones corresponding to the tRNA genes with the highest copy number in the human genome45. Weeder 2.0 was used to discover *de novo* RNA 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 be significantly contributing to the Random Forest were manually investigated to determine a threshold for the binding score. Mono- and di-nucleotide frequencies were computed by counting the occurrence of the four nucleotides and all 64 dinucleotide combinations, respectively. The above described features were computed for the following regions of interest of each transcript: +/– 100 nt of the PTC, +/– 100 nt of the exon junction downstream the PTC, the 3′ UTR region, the 5′ UTR region, 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.*46 (Supplementary Table 2: RNA half-lives in human B-cells) and is validated with mRNA half-life measures in HeLa cells47. A list of translated upstream open reading frames from Andreev *et al.*48 was kindly provided by Patrick BF 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 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 the NMD-sensitive region. The nonsense mutation density was normalized to the synonymous mutation density, to correct for differences in the mutational landscape. Two standard deviations of 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 Census39 (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 done on the UniProt-GOA release 15449 by calculating

the $\log_2$ odds ratio of NMD-sensitive and insensitive somatic nonsense mutations normalized for the synonymous mutation densities for every gene set bigger than 30 genes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

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, Agencia de Gestio d'Ajuts Universitaris i de Recerca (AGAUR), FP7 project 4DCellFate (277899), and the EMBL-CRG Systems Biology Program. FS was also supported by FP7 grants MAESTRA [ICT-2013-612944] and InnoMol [FP7-REGPOT-2012-2013-1-316289].

References

1. Maquat LE. When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells. RNA. 1995; 1:453–65. [PubMed: 7489507]
2. Mendell JT, Sharifi NA, Meyers JL, Martinez-Murillo F, Dietz HC. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. Nat Genet. 2004; 36:1073–8. [PubMed: 15448691]
3. Frischmeyer PA, Dietz HC. Nonsense-mediated mRNA decay in health and disease. Hum Mol Genet. 1999; 8:1893–900. [PubMed: 10469842]
4. Hall GW, Thein S. Nonsense codon mutations in the terminal exon of the beta-globin gene are not associated with a reduction in beta-mRNA accumulation: a mechanism for the phenotype of dominant beta-thalassemia. Blood. 1994; 83:2031–7. [PubMed: 8161774]
5. Kerr TP, Sewry CA, Robb SA, Roberts RG. Long mutant dystrophins and variable phenotypes: evasion of nonsense-mediated decay? Hum Genet. 2001; 109:402–7. [PubMed: 11702221]
6. Zhang J, Sun X, Qian Y, LaDuca JP, Maquat LE. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. Mol Cell Biol. 1998; 18:5272–83. [PubMed: 9710612]
7. Thermann R, et al. Binary specification of nonsense codons by splicing and cytoplasmic translation. EMBO J. 1998; 17:3484–94. [PubMed: 9628884]
8. Le Hir H, Gatfield D, Izaurralde E, Moore MJ. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. EMBO J. 2001; 20:4987–97. [PubMed: 11532962]
9. Chamieh H, Ballut L, Bonneau F, Le Hir H. NMD factors UPF2 and UPF3 bridge UPF1 to the exon junction complex and stimulate its RNA helicase activity. Nat Struct Mol Biol. 2008; 15:85–93. [PubMed: 18066079]
10. Brogna S, Wen J. Nonsense-mediated mRNA decay (NMD) mechanisms. Nat Struct Mol Biol. 2009; 16:107–13. [PubMed: 19190664]
11. Wang J, Gudikote JP, Olivas OR, Wilkinson MF. Boundary-independent polar nonsense-mediated decay. EMBO Rep. 2002; 3:274–9. [PubMed: 11850396]
12. Buhler M, Paillusson A, Muhlemann O. Efficient downregulation of immunoglobulin mu mRNA with premature translation-termination codons requires the 5'-half of the VDJ exon. Nucleic Acids Res. 2004; 32:3304–15. [PubMed: 15210863]
13. Eberle AB, Stalder L, Mathys H, Orozco RZ, Muhlemann O. Posttranscriptional gene regulation by spatial rearrangement of the 3' untranslated region. PLoS Biol. 2008; 6:e92. [PubMed: 18447580]
14. Mangus DA, Evans MC, Jacobson A. Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. Genome Biol. 2003; 4:223. [PubMed: 12844354]

15. Gatfield D, Unterholzner L, Ciccarelli FD, Bork P, Izaurralde E. Nonsense-mediated mRNA decay in Drosophila: at the intersection of the yeast and mammalian pathways. EMBO J. 2003; 22:3960–70. [PubMed: 12881430]
16. Longman D, Plasterk RH, Johnstone IL, Caceres JF. Mechanistic insights and identification of two novel factors in the C. elegans NMD pathway. Genes Dev. 2007; 21:1075–85. [PubMed: 17437990]
17. Zhang J, Maquat LE. Evidence that translation reinitiation abrogates nonsense-mediated mRNA decay in mammalian cells. EMBO J. 1997; 16:826–33. [PubMed: 9049311]
18. Romao L, et al. Nonsense mutations in the human beta-globin gene lead to unexpected levels of cytoplasmic mRNA accumulation. Blood. 2000; 96:2895–901. [PubMed: 11023527]
19. Silva AL, Ribeiro P, Inacio A, Liebhaber SA, Romao L. Proximity of the poly(A)-binding protein to a premature termination codon inhibits mammalian nonsense-mediated mRNA decay. RNA. 2008; 14:563–76. [PubMed: 18230761]
20. Lappalainen T, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501:506–11. [PubMed: 24037378]
21. Rivas MA, et al. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science. 2015; 348:666–9. [PubMed: 25954003]
22. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science. 2015; 349:1483–9. [PubMed: 26404825]
23. Kurosaki T, et al. A post-translational regulatory switch on UPF1 controls targeted mRNA degradation. Genes Dev. 2014; 28:1900–16. [PubMed: 25184677]
24. Singh G, Rebbapragada I, Lykke-Andersen J. A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay. PLoS Biol. 2008; 6:e111. [PubMed: 18447585]
25. Tani H, et al. Identification of hundreds of novel UPF1 target transcripts by direct determination of whole transcriptome stability. RNA Biol. 2012; 9:1370–9. [PubMed: 23064114]
26. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007; 3:1724–35. [PubMed: 17907809]
27. 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. 2015; 6:7581. [PubMed: 26138914]
28. Ray D, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013; 499:172–7. [PubMed: 23846655]
29. Zhang Z, et al. Nonsense-mediated decay targets have multiple sequence-related features that can inhibit translation. Mol Syst Biol. 2010; 6:442. [PubMed: 21179015]
30. Supek F, Skunca N, Repar J, Vlahovicek K, Smuc T. Translational selection is ubiquitous in prokaryotes. PLoS Genet. 2010; 6:e1001004. [PubMed: 20585573]
31. Rehwinkel J, Raes J, Izaurralde E. Nonsense-mediated mRNA decay: Target genes and functional diversification of effectors. Trends Biochem Sci. 2006; 31:639–46. [PubMed: 17010613]
32. Malone JH, et al. Mediation of Drosophila autosomal dosage effects and compensation by network interactions. Genome Biol. 2012; 13:r28. [PubMed: 22531030]
33. Fehrmann RS, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat Genet. 2015; 47:115–25. [PubMed: 25581432]
34. Ju YS, et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. Elife. 2014; 3
35. Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE. Nonsense-mediated decay approaches the clinic. Nat Genet. 2004; 36:801–8. [PubMed: 15284851]
36. Gutmann DH, et al. Haploinsufficiency for the neurofibromatosis 1 (NF1) tumor suppressor results in increased astrocyte proliferation. Oncogene. 1999; 18:4450–9. [PubMed: 10442636]
37. Davoli T, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell. 2013; 155:948–62. [PubMed: 24183448]
38. McFarland CD, Mirny LA, Korolev KS. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. Proc Natl Acad Sci U S A. 2014; 111:15138–43. [PubMed: 25277973]

39. Forbes SA, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015; 43:D805–11. [PubMed: 25355519]
40. Hart T, Brown KR, Sircoulomb F, Rottapel R, Moffat J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol Syst Biol. 2014; 10:733. [PubMed: 24987113]
41. Solimini NL, et al. Recurrent hemizygous deletions in cancers may optimize proliferative potential. Science. 2012; 337:104–9. [PubMed: 22628553]
42. Gonzalez-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. 2013; 14:R70. [PubMed: 23815980]
43. Pirinen M, et al. Assessing allele-specific expression across multiple tissues from RNA-seq read data. Bioinformatics. 2015; 31:2497–504. [PubMed: 25819081]
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. 2010; 6:e1000770. [PubMed: 20463871]
45. Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. Cell. 2014; 156:1324–35. [PubMed: 24630730]
46. Friedel CC, Dolken L, Ruzsics Z, Koszinowski UH, Zimmer R. Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. Nucleic Acids Res. 2009; 37:e115. [PubMed: 19561200]
47. Tani H, et al. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res. 2012; 22:947–56. [PubMed: 22369889]
48. Andreev DE, et al. Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. Elife. 2015; 4:e03971. [PubMed: 25621764]
49. Huntley RP, et al. The GOA database: gene Ontology annotation updates for 2015. Nucleic Acids Res. 2015; 43:D1057–63. [PubMed: 25378336]

Europe PMC Funders Author Manuscripts

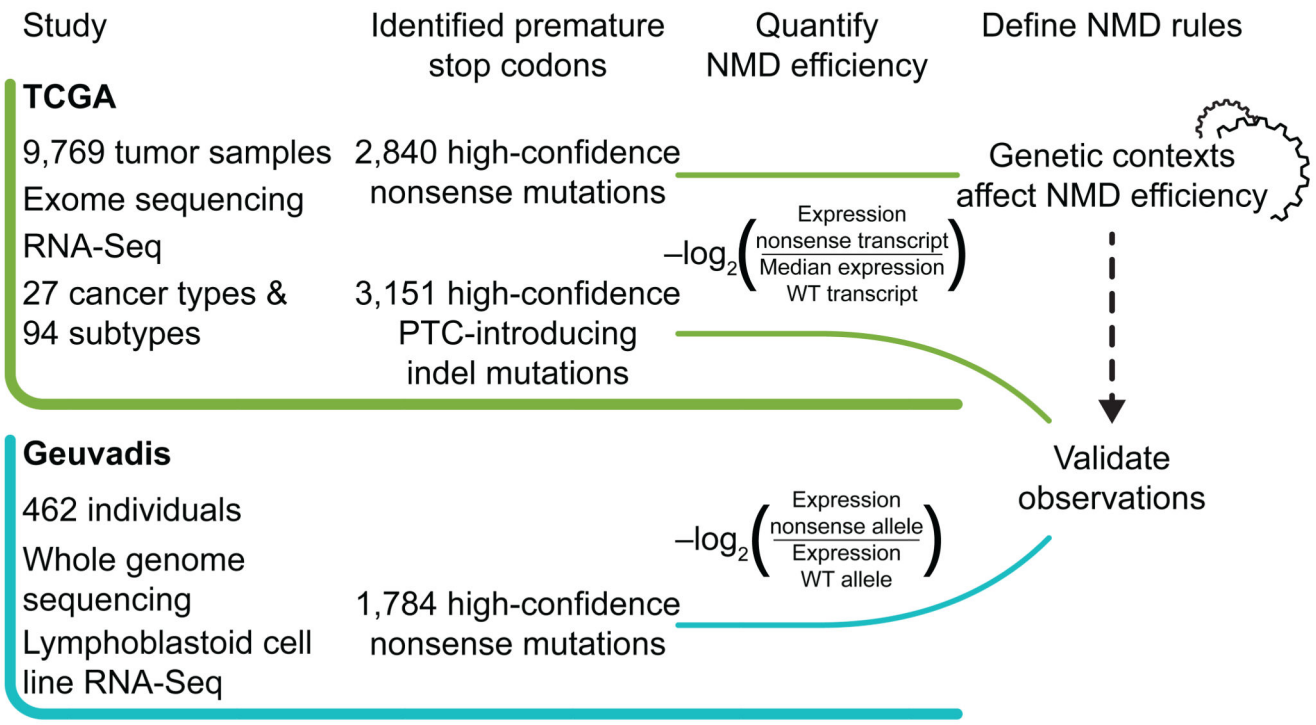

Figure 1. Study overview.

The rules of NMD were inferred from nonsense somatic mutations (n=2,840) in TCGA tumor genome sequences, and then validated using somatic frameshift changes (n=3,151) in TCGA and germline nonsense variants (n=1,784) in Geuvadis. NMD efficiency was quantified via mRNA expression levels for the TCGA (after control for somatic copy-number changes), and via allele-specific expression for the Geuvadis cohort.

Europe PMC Funders Author Manuscripts

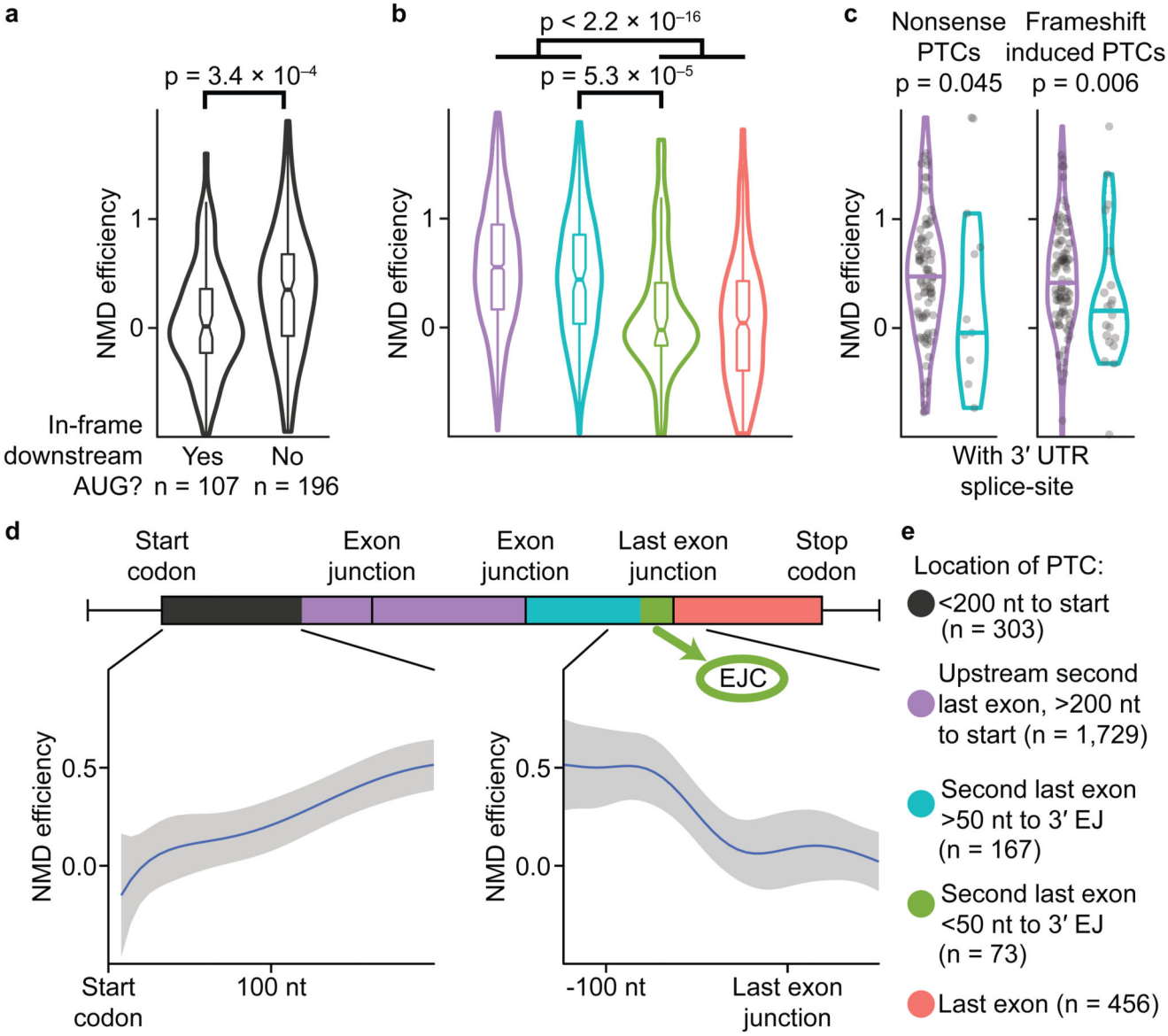

Figure 2. A downstream EJC and the proximity to the start codon are widespread signals for NMD.

NMD efficiency (–$\log_2$ nonsense mRNA level / wild-type mRNA level) in different gene regions. (**a**) NMD efficiency for PTCs in the first 200 coding nt with and without a downstream AUG codon. (**b**) NMD evasion in the last exon (red) and 50 nt upstream of the last exon (green), in comparison to efficient NMD in the second-last (blue) and further upstream (purple) exons. (**c**) PTCs in the second-last exon do not induce NMD when an intron is present in the 3′ UTR, shown for the nonsense mutations (left) and PTCs resulting from frameshift mutations (right). (**a-c**) P-values are by Mann-Whitney U-test, two-tailed, not adjusted for multiple testing. (**d**) Above: a schematic representation with color-coded gene regions that show different NMD efficiencies; black boxes represent exons. Below: a loess fit and its 95% C.I. shows NMD efficiency trends (individual points not shown). Bottom-left: increase in NMD efficiency across the first 200 coding nt. Bottom-right:

variation in NMD efficiency around the last exon junction, where the assumed location of the EJC is demarcated in the top of the plot. (**e**) Color-coding of gene regions.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

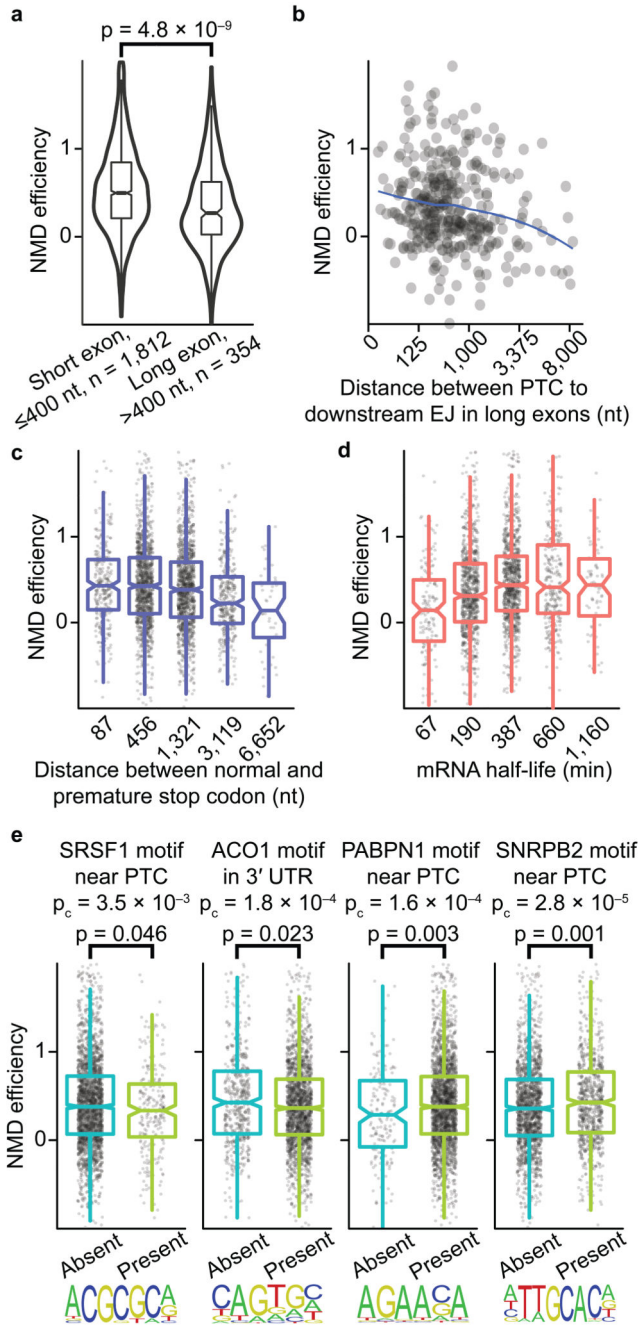

Figure 3. Exon length, distance to stop codon, mRNA decay rates and RBPs influence NMD efficiency.

(**a**) Reduced NMD efficiency in long exons. (**b**) Effects of distance between the PTC and the downstream exon junction. The line is a local polynomial regression fit. (**c**) Reduced NMD efficiency at PTCs very distant from the normal stop codon (after controlling for the EJC model and start-proximal NMD insensitivity; Methods). X-axis shows the median distance between the PTC and the normal stop codon of boxplots. (**d**) Reduced NMD efficiency in transcripts with short half-lives. X-axis shows the median mRNA half-life in minutes in each

boxplot. Bins in (c) and (d) are equal-width on the square root-transformed scale. (**e**) RNA binding protein motifs associated to NMD efficiency when located within +/– 100 nt from the PTC (SRSF1, PABPN1 and SNRPB2) or in the normal 3′UTR (ACO1). NMD efficiency distribution in left or the right box corresponds to absent or present motif, respectively. The position weight matrices used to detect motifs are shown below. In all boxplots, the central line and the notch are median and its approximate 95% C.I., the box shows the interquartile range, and the whiskers are extreme values upon removing outliers. (**a,e**) P-values are from Mann-Whitney U-test, two-tailed.

Europe PMC Funders Author Manuscripts

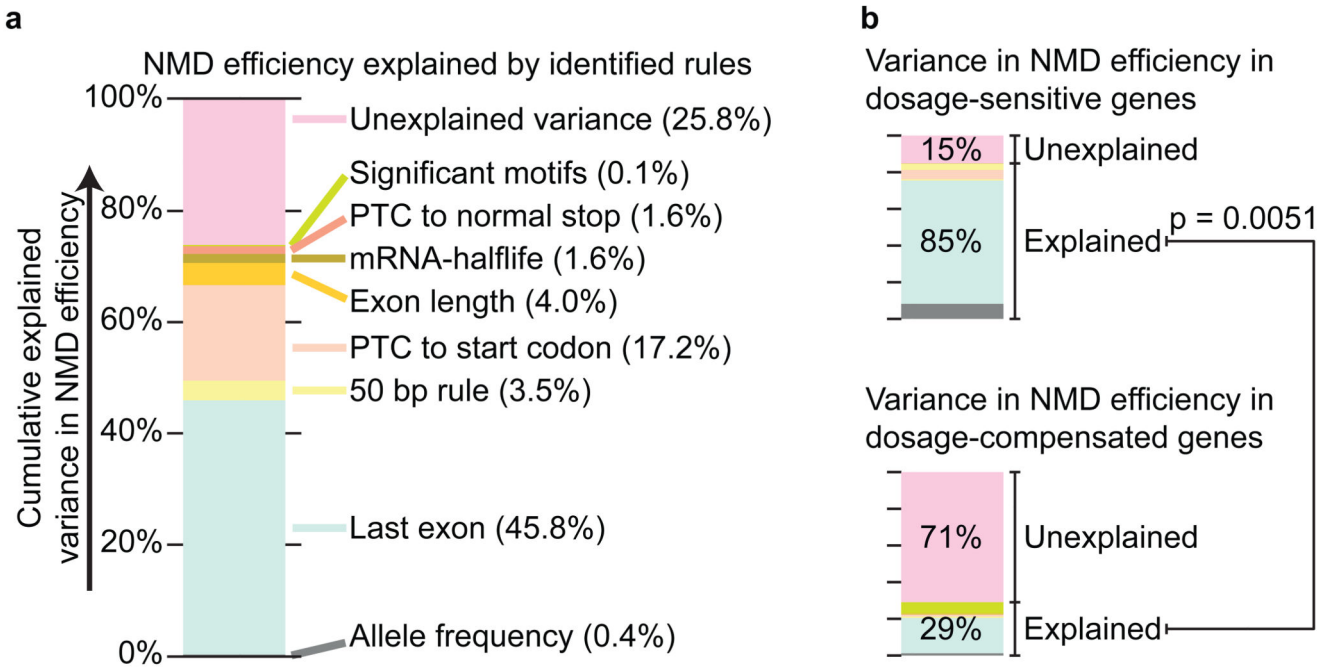

Figure 4. The identified NMD rules explain a large part of the NMD efficiency but NMD can have little effect when there is dosage compensation.

(**a**) The identified rules for NMD explain different amounts of variance in NMD efficiency between PTCs, up to a total of R^2=74.2%. The predictive power of every feature was determined by sequentially introducing the features to RF models, based on the order of the features suggested by an initial analysis of the entire feature set (Supplementary Table 3); in the bar chart, the bottommost features were added first, and the topmost features added last. The gain in R^2 upon adding each feature is shown after normalizing by the maximum attainable R^2 given the noise in the data (estimated from repeated occurrences of the same PTC in different tumors; Methods). (**b**) A significant decrease in explained NMD efficiency is observed in dosage-compensated genes. Explained NMD efficiency in the top 20% least and most-dosage compensated genes (definition in Methods) are shown in the top and bottom bars, respectively. Colors in the bars match the features in (a). Significance by a t-test on the Fisher-transformed Pearson correlations of the RF model.

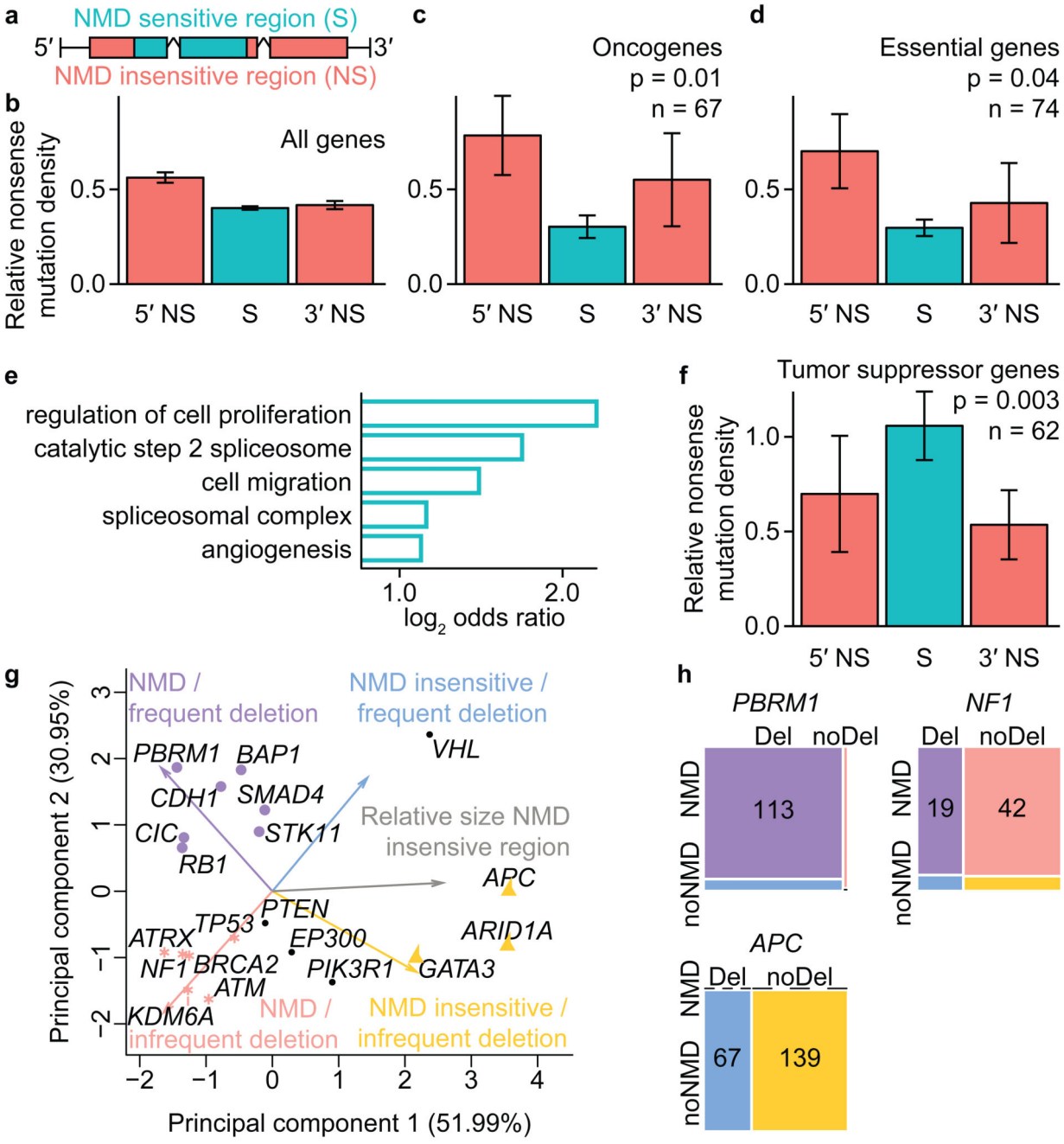

Figure 5. Signatures of negative/positive selection on somatic nonsense mutations.

(**a**) Genes were divided into NMD-insensitive regions: (i) last exon plus 50 nt upstream (3′NS), and (ii) first 250 nt after start codon (5′NS), and (iii) the NMD-sensitive remainder (S). (**b-d**) Density of nonsense mutations normalized by synonymous mutations within each gene region, for all genes (b), oncogenes (c) from the Cancer Gene Census39 and essential genes (d) from40. (**e**) Gene Ontology enrichment analysis of genes under negative selection, based on NMD-sensitive *vs.* insensitive proportions (Methods). Gene sets ≥30 genes and odds ratio > 2 are shown. (**f**) As (b-d), but for the 200 highest scoring ‘STOP’ tumor

Europe PMC Funders Author Manuscripts

suppressor (TS) genes from41. Panels (b-d, f), error bars are 95% C.I. obtained by bootstrapping; *n* = number of examined PTCs; p-values are from randomization tests, comparing against the baseline in (b). (**g**) A principal components analysis of occurrence of nonsense mutations in NMD-sensitive or NMD-insensitive gene regions and copy-number alteration data in the same patients (Methods). Clusters with different putative mechanisms of inactivation shown with distinct shapes and colors. Arrows show correlations of principal components and original features; colors as in (h). TS genes with ≥50 nonsense mutations in the TCGA cohort are shown. (**h**) Example TS genes from each cluster: *PBRM1*- two-hit genes; *NF1*- haploinsufficient TS genes; *APC*- genes that rarely undergo NMD. Columns show relative frequencies of tumors harboring a deletion ("Del") or not ("noDel") of a gene; rows are tumors with nonsense mutations in NMD-sensitive ("NMD") or NMD-insensitive regions ("noNMD").

Europe PMC Funders Author Manuscripts

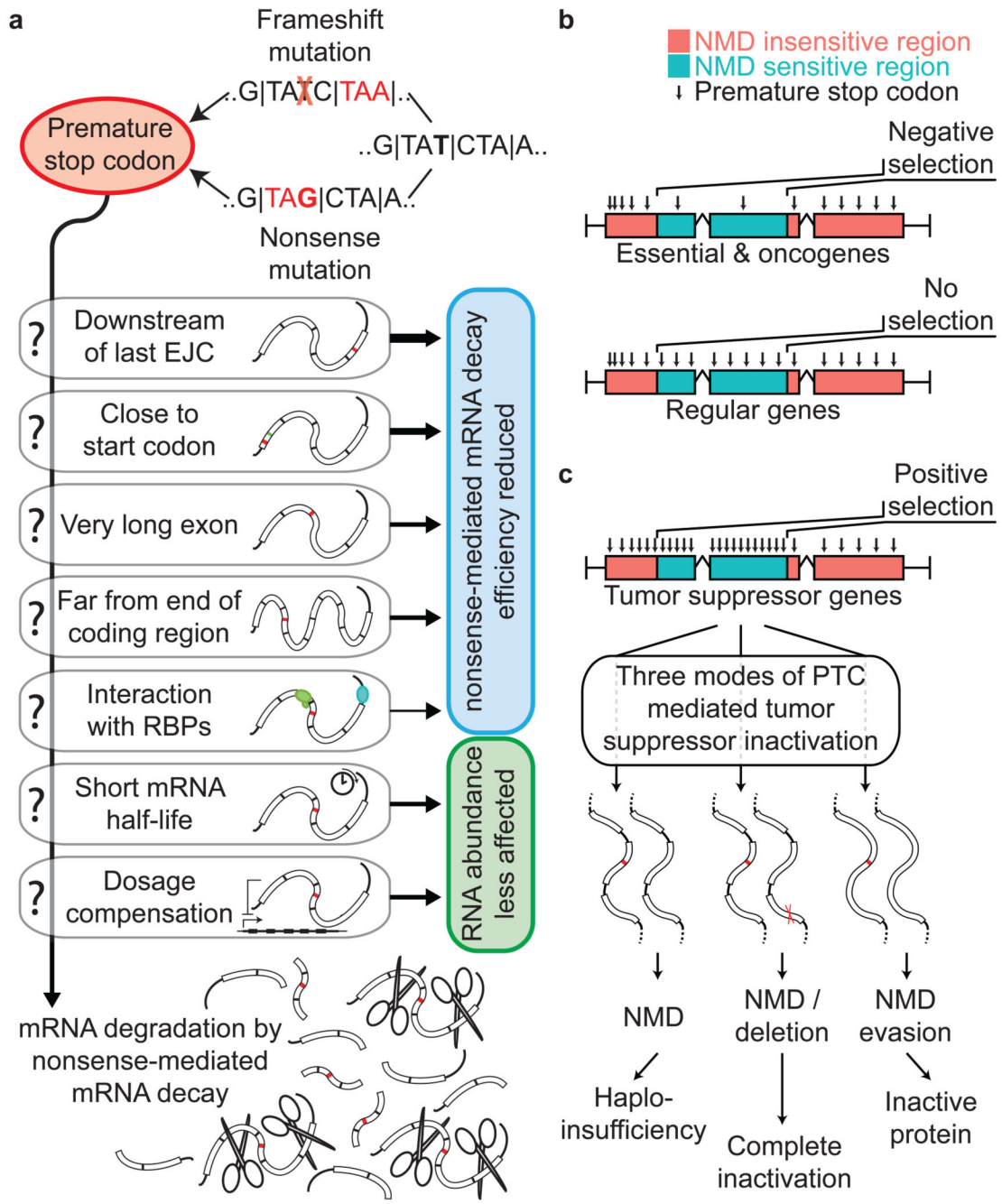

Figure 6. Model summarizing the rules governing NMD in human cells.

(**a**) Overview of the determinants of NMD efficiency. (**b**) Negative selection against NMD-triggering somatic mutations. (**c**) Relevance of NMD for tumor suppressor inactivation.