Selection for or against Escape from Nonsense Mediated Decay is a Novel Signature for the Detection of Cancer Genes

Runjun D. Kumar[1], Briana A. Burns[1], Paul J. Vandeventer[1], Pamela N. Luna[1], Chad A. Shaw[1,2,3*]

[1] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA
[2] Baylor Genetics, Houston, TX, USA
[3] Department of Statistics, Rice University, Houston, TX, USA

Corresponding Author
Chad Shaw, PhD
One Baylor Plaza, MS225
Houston, TX, 77030
Email: cashaw@bcm.edu

Abstract

Escape from nonsense mediated decay (NMD-) can produce activated or inactivated gene products, and bias in rates of escape can identify functionally important genes in germline disease. We hypothesized that the same would be true of cancer genes, and tested for NMD- bias within The Cancer Genome Atlas pan-cancer somatic mutation dataset. We identify 29 genes that show significantly elevated or suppressed rates of NMD-. This novel approach to cancer gene discovery reveals genes not previously cataloged as potentially tumorigenic, and identifies many potential driver mutations in known cancer genes for functional characterization.

Frameshift insertions, deletions and stop-gain point mutations usually generate new predicted termination codons (PTC+) and result in nonsense-mediated decay of the mRNA prior to translation (NMD+)[1]. However, termination codons within 55bp of the final exon-exon junction or further downstream can escape NMD (NMD-) and produce a truncated or extended peptide. Somatic NMD- variants are known to affect immunogenicity in some tumor types[2], and the ratio of somatic PTC+ variants that are NMD+ or NMD- varies between tumor suppressors and oncogenes due to positive and negative selection on NMD+ variants[3].

Studies in germline datasets suggest that nearly 10% of protein-coding genes may demonstrate selection on NMD- variants as well, particularly c-terminal events which produce truncated or extended peptides[4]. In some cases these events are thought to activate the peptide or produce a dominant-negative effect, and these mechanisms are important in known cancer genes[5, 6]. Identifying activating mutations in cancer is crucial as these events may be more easily targeted with small molecule inhibitors; however we have previously shown that existing methods struggle to identify these events[7]. We hypothesized that potentially activating c-terminal, PTC+ NMD- events would be enriched or

depleted in some cancer genes in somatic cancer mutation data, and that new putative cancer genes could be identified with this novel mutation signature.

We first assessed whether PTC+ variants (stopgains and frameshift indels) have significantly higher or lower rates of NMD- versus PTC- variants (synonymous, missense and inframe indels) (Figure S1-2). NMD status is determined using the 55bp rule and the position of the predicted termination codon for PTC+ variants, and by the location of the variant itself for PTC- variants. Each gene was considered in the 0, +1 and –1 frames, with rates of NMD- compared between PTC- and PTC+ variants using Fisher exact tests in each frame. We then used Fisher's method to combine p-values into a single result for each gene. After false-discovery-rate (FDR) correction, we identified 29 genes with abnormal rates of NMD- among PTC+ variants from 14,762 genes assessed (Table 1, Table S1). In 26 of these genes there was a higher rate of NMD- than expected; the remainder had relative depletions (Figure S3).

Seven of 29 genes are also among 630 genes in the Cancer Gene Census (CGC)[8] (p=0.0001): four are identified as exclusively TSGs (*ACVR2A*, *CSMD3*, *PTCH1*, *RPL22*), one exclusively as an oncogene (*PPM1D*), and two are annotated as having a mixed role (*BCORL1, GATA3*). We ran our previously reported ensemble method for cancer gene discovery as well (Figure S4), and it identified 3 additional genes as predicted cancer genes (*DHX16*, *DOCK3*, and *ZBTB20* as TSGs); the model also would classify *GATA3* and *BCORL1* as TSGs and *ACVR2A* and *CSMD3* as oncogenes[9]. Four genes were potentially druggable based on their gene family (*ACVR2A*, *GPR161*, *STAMBPL1* and *TTK*)[10].

Interestingly, none of the 29 genes were identified in a prior analysis of NMD- depletion based on population germline datasets (Figure S5)[4]. However, *PPM1D* is one of 39 genes identified in the Online Mendelian Inheritance in Man (OMIM) as bearing activating c-terminal NMD- PTC+ variants, and these events are implicated in tumor development[6]. In general, we found evidence of NMD- bias in sets of cancer genes, but not in gene sets known to show NMD- bias in germline data sets (Figure S6).

Therefore, we assessed NMD- bias while limiting scope to sets of known cancer genes. We started with 87 high-confidence cancer genes identified prior to widespread tumor sequencing which we have previously used for benchmarking ("HiConf cancer genes")[9]. Only 5/87 were significantly biased at a FDR of 0.2 (*GATA3*, *PTCH1*, *VHL*, *FBXW7*, *APC*, Figure 1A). There is a general skew among p-values for these genes compared with others, suggesting we may not be sufficiently powered to detect cancer genes with significant NMD- bias (Figure 1B, Figure S6). The identified genes had a variety of effect sizes (Figure 1C-G), and extending this analysis to 431 CGC oncogenes and TSGs additionally identified *ACVR2A*, *ASXL1*, *ASXL2*, *CASP8*, *CSMD3* and *PPM1D* with significant deviations at FDR<0.2 (Figure S7). We also analyzed nonsense variants in the COSMIC dataset[8], and found that HiConf TSGs showed a relative depletion in NMD- compared with HiConf oncogenes (p=0.012, Figure S8), similar to the pattern in TCGA data, with the caveat that the COSMIC analysis contains only in-frame variants and combines different sequencing platforms.

Analysis of NMD- bias offers insight into specific cancer genes. *PPM1D* represents one pattern we expect to find for an activating variant, with enrichment of c-terminal PTC+ NMD- events in a known oncogene[6]. *GATA3* is another example; it has a complex role in breast cancer, but there is evidence that some c-terminal frameshifts result in an extended transcript that has gain-of-function properties, while others shorten the transcript and have a dominant negative effect[11]. The results of this analysis may have similar implications for the other known cancer genes identified in Table 1, several of which are usually categorized as tumor suppressors. PTC+ NMD- variants in these genes may cause loss-of-function despite expression of the peptide, which could potentially be rescued pharmacologically, or they may alter the function of the peptide in some other way, expanding the role of the gene in tumorigenesis.

Our results also implicate several genes in cancer development for the first time. We identify that *growth/differentiation factor 5 (GDF5)* is deficient in NMD- PTC+ variants, as would be expected of a tumor suppressor for the gain-of-function mechanism of NMD- PTC+ variants; however in this case the

absolute number of PTC+ variants is small, and the role of *GDF5* in tumorigenesis is not well established[12-14]. Many of the remaining genes in Table 1 have recurrent NMD- PTC+ variants that could produce activated and truncated peptides. For example, there is a recurrent event K405Rfs*21 event in *STAM-binding protein-like 1 (STAMBPL1)*. This gene is not well understood, though it appears to be positively associated with prostate and gastric cancers[15, 16], and mutations in this region have been implicated in canine polyp development[17].

The use of NMD- bias among PTC+ variants is a promising new strategy for identifying cancer genes which is substantially orthogonal to prior methods. While the effect on protein function may vary across NMD- variants, the inference that PTC+ variants cause loss-of-function does not hold in the NMD- case, and these variants present a novel class for oncological analysis. We developed a framework for identifying NMD- patterns of association within cancer somatic mutation datasets. We identified 29 genes that demonstrated significant bias for PTC+ in the NMD- region. The majority were enriched in NMD- events which could represent activating c-terminal truncations. Many of these genes are potential cancer genes. These variants also represent novel targets in several established cancer genes, many of which are currently classified as TSGs. We anticipate that this strategy could be even more fruitful in larger datasets, which would also allow further analysis within specific cancer sub-types.

Acknowledgements

We thank Dr. Zeynep Coban Akdemir for critically reading the manuscript and providing helpful feedback. We thank Dr. William Decker and his laboratory for wet lab mentorship of B.A.B., as well as funding her work through Alex's Lemonade Stand.

References

1. Lindeboom, R. G. H., Vermeulen, M., Lehner, B. & Supek, F. The impact of nonsense-mediated mRNA decay on genetic disease, gene editing and cancer immunotherapy. *Nat. Genet.* **51**, 1645-1651 (2019).

2. Litchfield, K. *et al*. Escape from nonsense-mediated decay associates with anti-tumor immunogenicity. *Nat. Commun.* **11**, 3800-5 (2020).

3. Lindeboom, R. G., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112-1118 (2016).

4. Coban-Akdemir, Z. *et al*. Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. *Am. J. Hum. Genet.* **103**, 171-187 (2018).

5. Toki, T. *et al*. De Novo Mutations Activating Germline TP53 in an Inherited Bone-Marrow-Failure Syndrome. *Am. J. Hum. Genet.* **103**, 440-447 (2018).

6. Martinikova, A. S., Burocziova, M., Stoyanov, M. & Macurek, L. Truncated PPM1D Prevents Apoptosis in the Murine Thymus and Promotes Ionizing Radiation-Induced Lymphoma. *Cells* **9**, 10.3390/cells9092068 (2020).

7. Kumar, R. D. & Bose, R. Analysis of somatic mutations across the kinome reveals loss-of-function mutations in multiple cancer types. *Sci. Rep.* **7**, 6418-x (2017).

8. Sondka, Z. *et al*. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer.* **18**, 696-705 (2018).

9. Kumar, R. D., Searleman, A. C., Swamidass, S. J., Griffith, O. L. & Bose, R. Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. *Bioinformatics* **31**, 3561-3568 (2015).

10. Kumar, R. D., Chang, L. W., Ellis, M. J. & Bose, R. Prioritizing Potentially Druggable Mutations with dGene: An Annotation Tool for Cancer Genome Sequencing Data. *PLoS One* **8**, e67980 (2013).

11. Mair, B. *et al*. Gain- and Loss-of-Function Mutations in the Breast Cancer Gene GATA3 Result in Differential Drug Sensitivity. *PLoS Genet.* **12**, e1006279 (2016).

12. Cook, M. B. *et al*. Genetic contributions to the association between adult height and testicular germ cell tumors. *Int. J. Epidemiol.* **40**, 731-739 (2011).

13. Margheri, F. *et al*. GDF5 regulates TGFss-dependent angiogenesis in breast carcinoma MCF-7 cells: in vitro and in vivo control by anti-TGFss peptides. *PLoS One* **7**, e50342 (2012).

14. Alshangiti, A. M. *et al*. Association of distinct type 1 bone morphogenetic protein receptors with different molecular pathways and survival outcomes in neuroblastoma. *Neuronal Signal.* **4**, NS20200006 (2020).

15. Yu, D. J. *et al*. STAMBPL1 knockdown has antitumour effects on gastric cancer biological activities. *Oncol. Lett.* **18**, 4421-4428 (2019).

16. Chen, X., Shi, H., Bi, X., Li, Y. & Huang, Z. Targeting the deubiquitinase STAMBPL1 triggers apoptosis in prostate cancer cells by promoting XIAP degradation. *Cancer Lett.* **456**, 49-58 (2019).

17. Wang, J. *et al*. Collaborating genomic, transcriptomic and microbiomic alterations lead to canine extreme intestinal polyposis. *Oncotarget* **9**, 29162-29179 (2018).

Methods

Datasets

Pan-cancer TCGA annotated mutations were downloaded from the public repository (https://gdc.cancer.gov/about-data/publications/mc3-2017, filename mc3.v0.2.8.PUBLIC.maf.gz, accessed 12/20/2020). Cancer Gene Census tier 1 and tier 2 genes were downloaded from COSMIC (https://cancer.sanger.ac.uk/census, accessed 12/20/2020). Other gene sets or annotations are drawn directly from the cited works unless specifically noted. We used the GenomOncology Precision Oncology API Suite (https://www.genomoncology.com/api-suite) to identify the distance of each variant from the exonic boundaries and determine NMD status.

Annotations and Filtering

Annotations are directly from the TCGA MAF file whenever possible. Only coding mutations (stopgains, frameshift indels, synonymous variants, missense variants, and inframe indels) were considered in the analysis. Variants were annotated with the frame they would produce. By definition, stopgains, synonymous, missense and inframe indels do not change the frame and are assigned a frame of 0. Frameshift indels that result in the frame moving forward by one position (e.g. deletion of one, four, seven base pairs, or the insertion of two, five, eight base pairs, *etc*) are assigned frame +1. Frameshift indels that result in the frame moving backward by one position are assigned frame -1.

We annotated mutations as being PTC+ (frameshift indel, stop-gains) or PTC- (synonymous, missense, inframe indel) and then annotated the PTC+ variants as being NMD+ or NMD- based the position of the predicted termination codon and the 55bp rule as previously described[4]. Each transcript was divided into NMD- and NMD+ regions for each potential reading frame. The boundary is defined as the most upstream position where a variant of that frame would produce a PTC within 55bp of the final exon-exon junction or further downstream. The transcript used was the canonical transcript used in the

original TCGA dataset. There were 1,306,510 mutations affecting 16,037 genes in 9,859 individuals with 34 cancer types at this point.

Analysis

We next confirmed the most appropriate control to assess rates of NMD- among PTC+ variants. We showed that synonymous mutations do not occur randomly between NMD+ and NMD- regions of the gene in at least some genes, and therefore chose observed mutation rates as a control (Figure S1). Synonymous mutations make up roughly one third of point mutations in the dataset; to maximize data usage, we assessed whether missense mutations were comparable to synonymous events with regards to occurrence within the NMD- region (Figure S2). We could find no genes where missense and synonymous variants significantly differed with regards to occurrence within the NMD- region. Therefore pooled synonymous and missense variants (PTC- variants) were compared to PTC+ variants to assess rates of NMD- (Figure S3). These control experiments did not test the +1 or -1 frames since missense and synonymous mutations occur exclusively in-frame.

We tested for NMD- bias by calculating a Fisher exact test for 2 by 2 tables for each frame in each gene (Figure S3). The p-values from these tests were combined using the Fisher Method, resulting in a single p-value for each gene. A pooled odds-ratio was also calculated by combining the counts across the three frames for each gene. These pooled odds-ratios and combined p-values are used throughout the manuscript unless otherwise noted.

Fisher Exact Tests were also used to assess the overlap between different gene sets throughout the analysis. False Discover Rates were controlled using Benjamini-Hochberg correction. To compare sets of genes, chi-square statistics were calculated for each frame and gene in the set and summed; this value was then compared to random draws of genes to create p-values. Methods and techniques for comparison analyses (e.g. druggability, cancer gene prediction) are found in the pertinent references.

Symbol	Transcript ID	Mutations	PTC+	NMD-PTC+	OR	p-value	CGC	Predicted Role	OMIM39	Druggable family
GDF5	ENST00000374372	104	13	0	0	2.11E-04				
PLAGL2	ENST00000246229	57	11	1	0	1.94E-06				
RPL22	ENST00000234875	80	56	6	0	1.17E-07	TSG	TSG		
CSMD3	ENST00000297405	1566	165	10	2.8	1.21E-04	TSG	Oncogene		
BCORL1	ENST00000540052	290	45	18	4.2	2.82E-04	Mixed	TSG		
GRIK2	ENST00000421544	251	27	9	4.9	1.43E-04				
RBM6	ENST00000266022	137	27	7	4.9	9.84E-05				
PTCH1	ENST00000331920	202	31	13	5.1	1.75E-04	TSG			
TTK	ENST00000369798	124	30	11	8	2.37E-04				Serine/threonine kinase
ADD3	ENST00000356080	100	19	10	8.2	1.61E-05				
ACVR2A	ENST00000241416	152	72	49	8.7	1.68E-08	TSG	Oncogene		Serine/threonine kinase
KMO	ENST00000366559	82	16	11	9.3	3.32E-04				
HFM1	ENST00000370425	246	39	7	9.5	2.04E-05				
DDC	ENST00000444124	113	15	11	11	2.85E-04				
DHX16	ENST00000376442	132	29	16	12.9	4.08E-07		TSG		
TCF20	ENST00000359486	220	22	6	13.5	3.23E-04				
GATA3	ENST00000379328	184	64	56	13.8	8.48E-09	Mixed	TSG		
FAHD2B	ENST00000414820	45	10	8	15.1	3.18E-04				
ZBTB20	ENST00000474710	196	56	48	16.7	2.06E-07		TSG		
STAMBPL1	ENST00000371926	55	22	15	19.1	9.30E-06				Protease
PPM1D	ENST00000305921	59	22	20	20.8	4.92E-05	Oncogene		+	
SMC4	ENST00000357388	163	21	9	22.1	1.88E-06				
PAX6	ENST00000419022	87	16	11	23.8	2.10E-05				
MYO1A	ENST00000442789	141	15	7	26.7	4.98E-06				
CAMSAP2	ENST00000358823	153	31	14	32.7	5.48E-08				
DOCK3	ENST00000266037	387	74	55	39.6	3.53E-11		TSG		
ALDH3A1	ENST00000457500	62	9	8	55.6	2.61E-04				
GPR161	ENST00000537209	83	8	7	64.6	1.08E-04				GPCR
NTAN1	ENST00000287706	39	11	11	100*	8.86E-05				

Table 1. Genes with enrichment or depletion for NMD- among variants causing premature termination codons based on FDR < 0.2. Abbreviations: CGC, cancer gene census; GPCR, g-protein coupled receptor; NMD-, escape from nonsense mediated decay; OMIM39, 39 genes with dominant NMD- events per Online Mendelian Inheritance in Man; OR, odds-ratio; PTC+, variant causing premature termination codon; TSG, tumor suppressor gene. Odds ratios with a zero count in the denominator are marked (*).

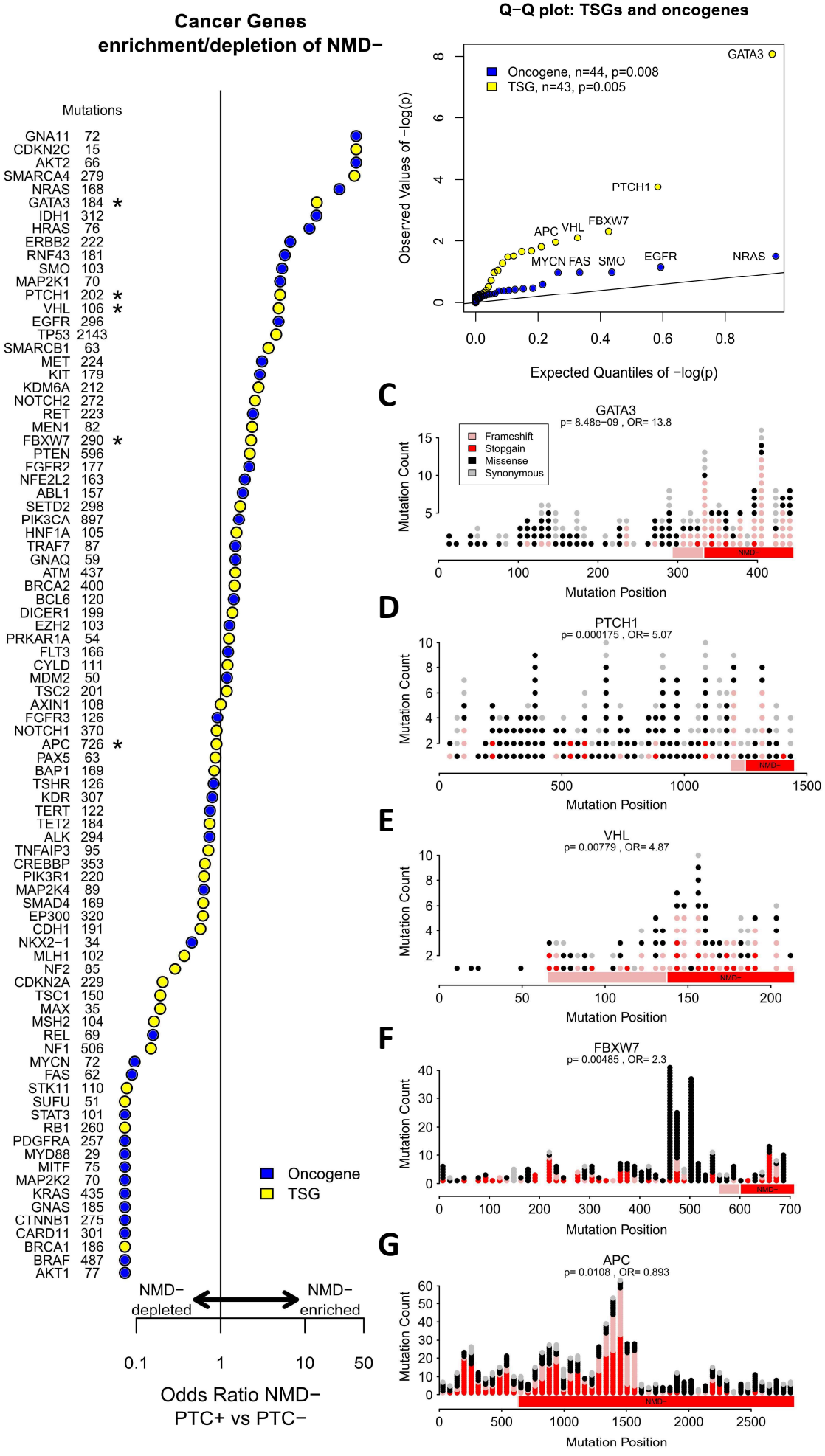


Figure 1. Enrichment or depletion of NMD- PTC+ events among oncogenes and tumor suppressors. A) Odds-ratios of NMD- for PTC+ compared with PTC- variants are shown for 87 high-confidence cancer genes. Mutation counts are displayed. Genes identified at FDR < 0.2 based on Fisher testing marked with (*). B) Q-Q plot of p-values for oncogenes and TSGs. C-G) Mutation diagrams for the 5 genes with FDR < 0.2. NMD- regions are labeled. The red area is NMD- in all frames. The pink area is NMD- for frameshift events. The boundary is drawn at the more extensive of -1 or +1 frames.

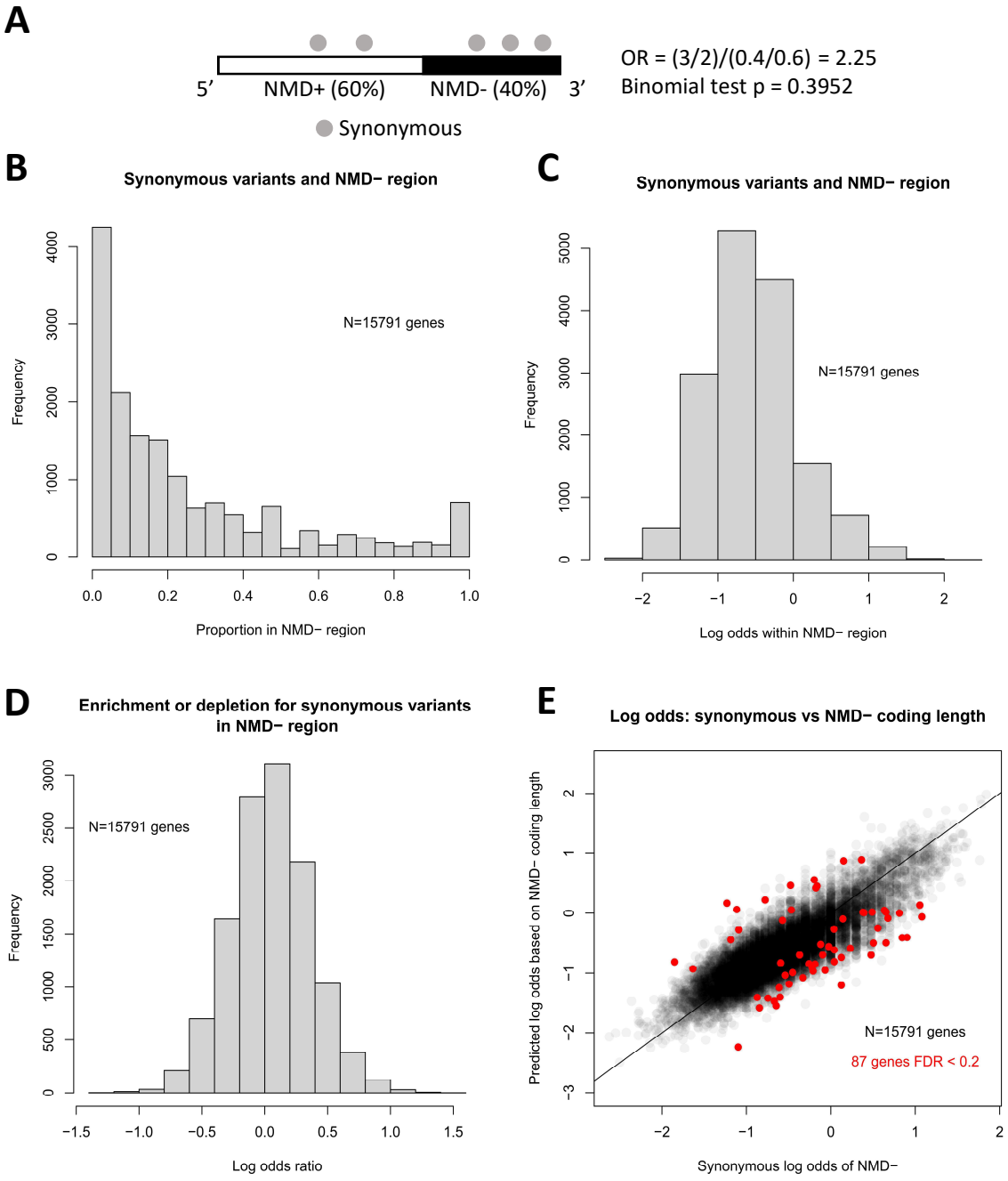


Figure S1. Synonymous variants do not occur randomly between NMD+ and NMD- regions. A) Schema for the analysis. A mature transcript is depicted with an NMD- region that is 40% of the coding length. B) Proportion of synonymous variants in the NMD- region for each gene. -1 and +1 frames are excluded. C) Log odds of synonymous variants occurring in the NMD- region. D) Log odds-ratios for synonymous variants occurring in the NMD- region, calculated as outlined in panel (A). E) Genes plotted by observed and expected odds of NMD- among synonymous variants, assuming variants occur randomly throughout the coding region. 87 genes with significant differences are highlighted.

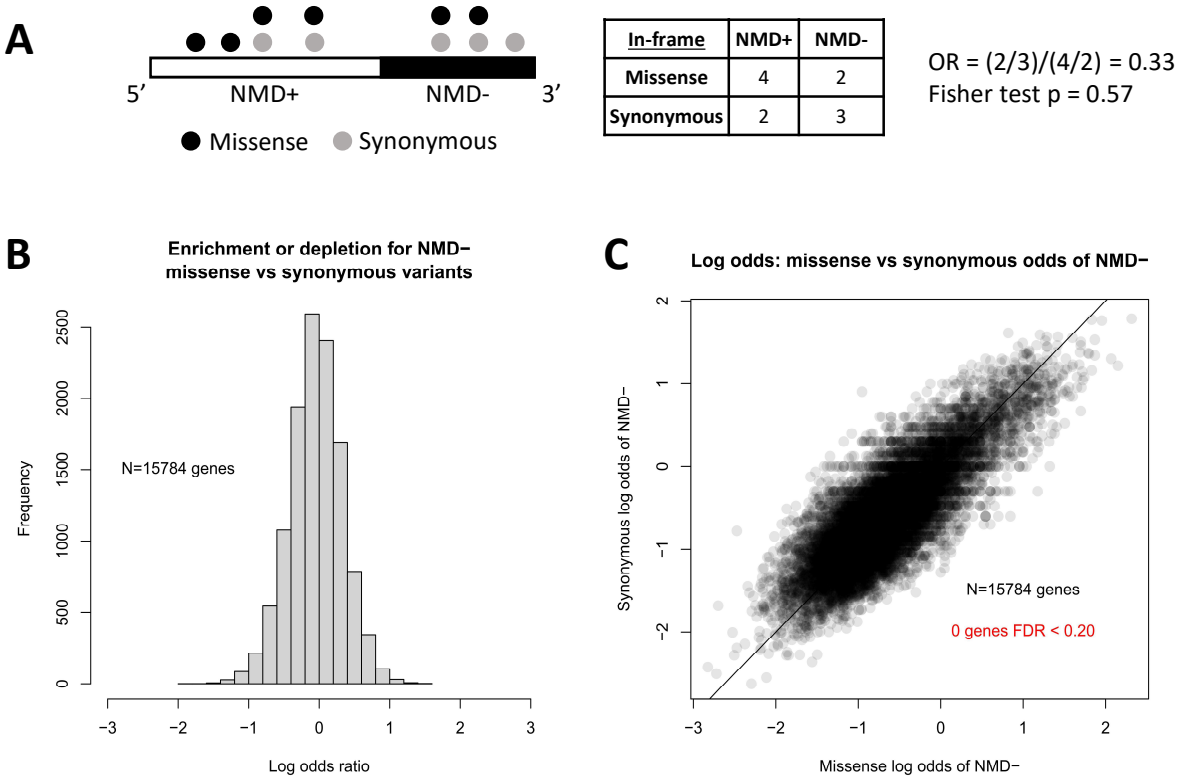
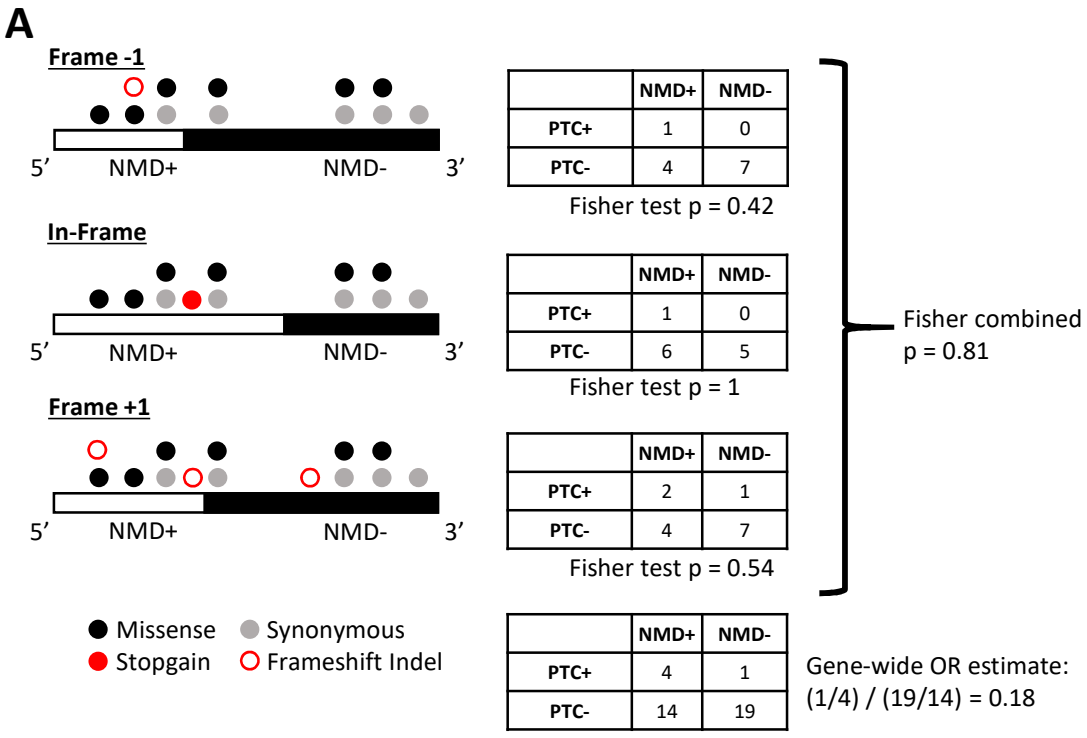


Figure S2. Missense and synonymous variants do not significantly differ with regards to presence in NMD- regions. A) Schema of analysis. A mature transcript is depicted. B) Log odds-ratios for rate of missense variants in NMD- region versus synonymous variants, as outlined in panel (A). -1 and +1 frames are excluded. C) Genes plotted by observed log odds of NMD- for missense and synonymous variants. The rates are compared with Fisher exact testing with no genes showing significant differences.



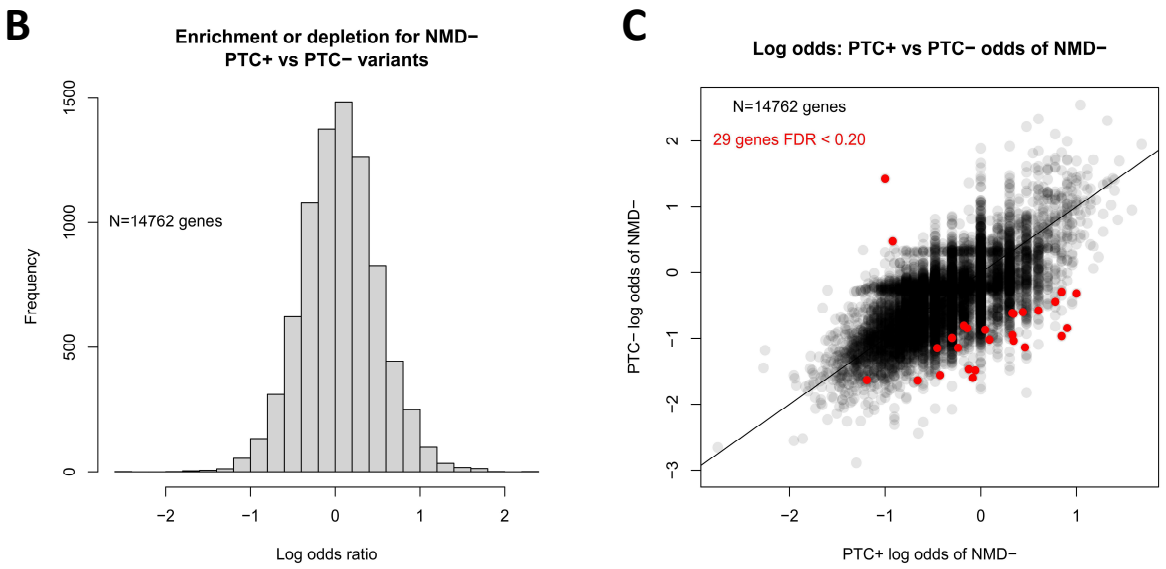


Figure S3. Analysis of PTC+ variants versus PTC- variants with regards to presence in NMD- region. A) Schema of analysis. A mature transcript is depicted, with NMD- regions defined for each reading frame using the 55bp rule. B) Log odds-ratios for rate of PTC+ variants in NMD- region versus synonymous and missense (PTC-) variants, as outlined in panel (A). C) Genes plotted by observed log odds of NMD- for PTC+ and PTC- variants. The 29 genes with significant NMD- bias are highlighted.

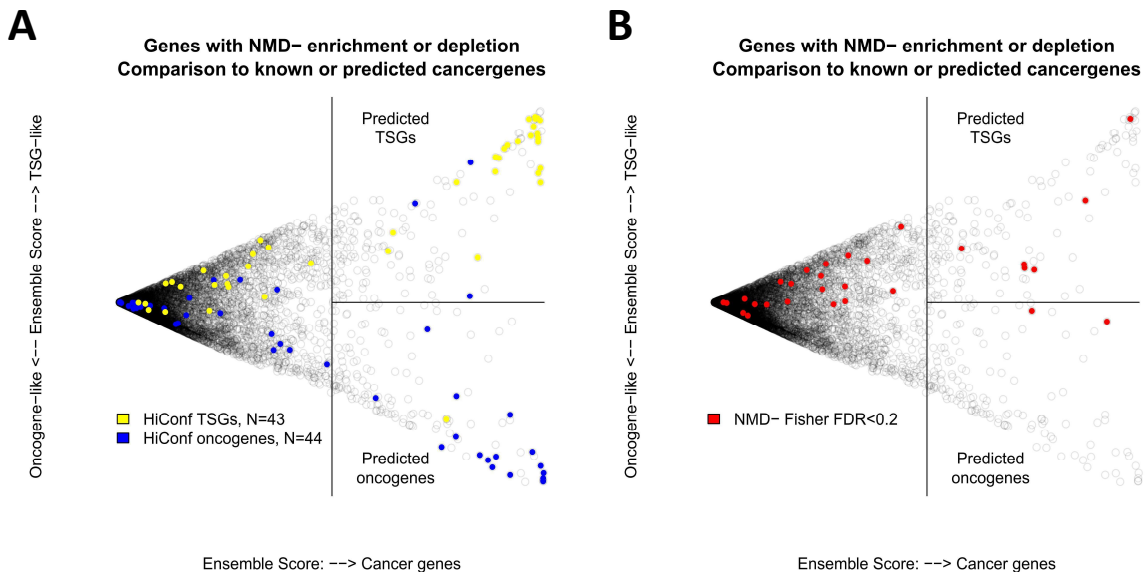


Figure S4. Prediction of tumor suppressors and oncogenes and comparison to genes enriched or depleted in NMD- PTC+ variants. A) An ensemble method identifies predicted oncogenes (lower right segment) and predicted TSGs (upper right segment). High confidence cancer genes identified prior to widespread tumor sequencing are highlighted ("HiConf cancer genes"). B) The same plot with the 29 genes showing NMD- bias highlighted.

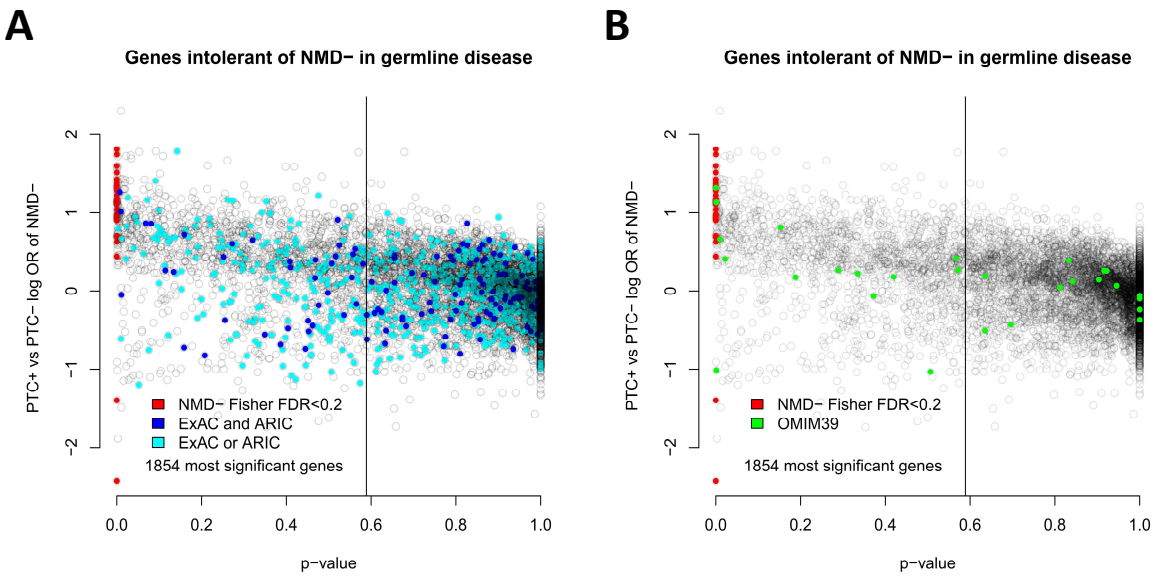


Figure S5. Comparison of current analysis in somatic cancer datasets to genes identified previously in germline datasets. "ExAC and ARIC" genes were identified in both of those two datasets as being depleted in NMD- PTC+ variants. "ExAC or ARIC" genes were identified in one dataset or the other. OMIM39 genes are annotated in OMIM as having NMD- PTC+ variants that cause dominant disease. To the left of the vertical line are as many genes as identified in the ExAC or ARIC set (n=1854).

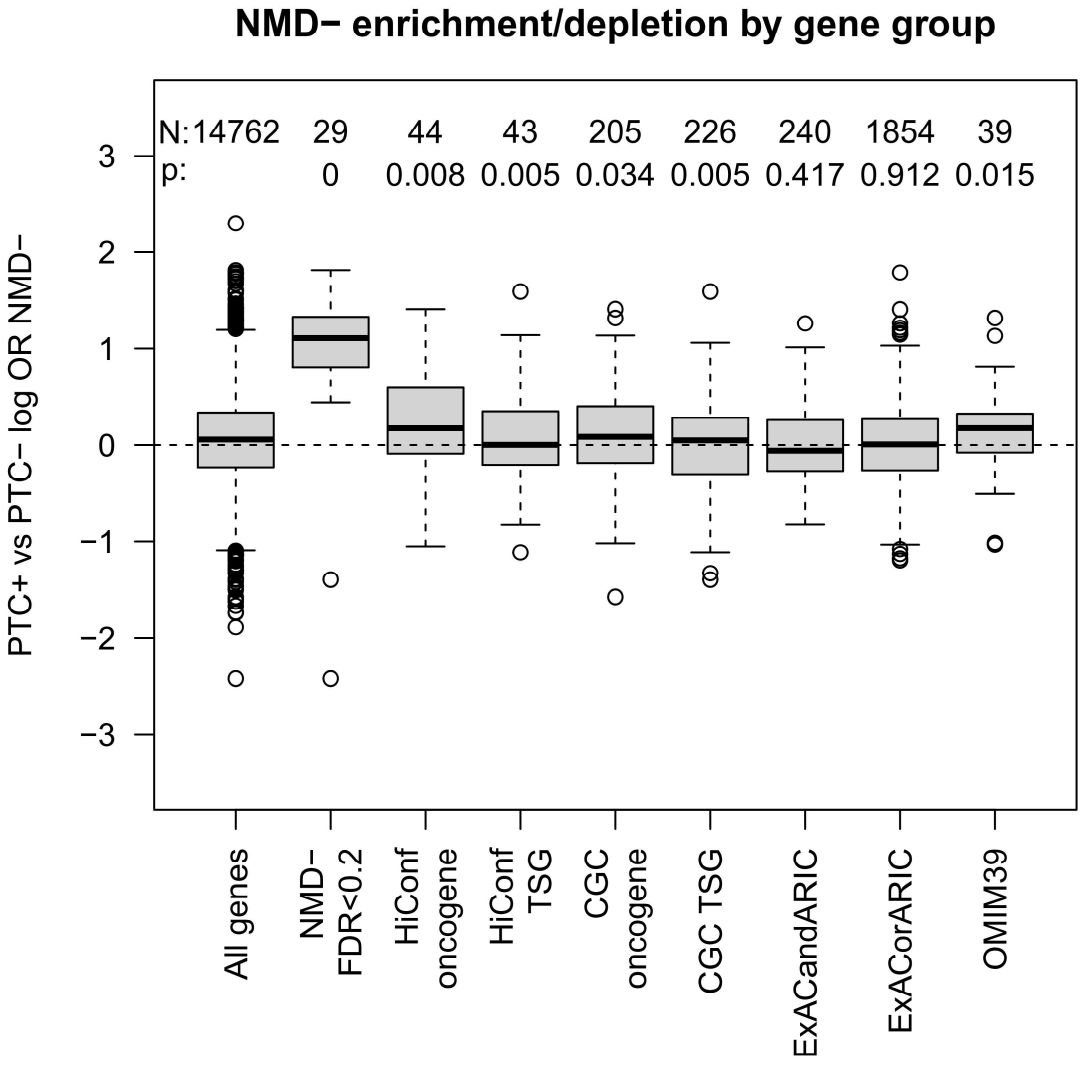


Figure S6. Log odds-ratios for NMD- enrichment or depletion among PTC+ variants by gene set. p-values indicate chi-square testing for the class as a whole, and test for abnormal enrichment or depletion among members of the class versus a randomly drawn set of equal size.

A

Symbol	Mutations	Gene Type	OR NMD-
ACVR2A	152	TSG	8.75
RPL22	80	TSG	0.04
PPM1D	59	TSG	20.83
CSMD3	1566	TSG	2.76
ASXL2	192	TSG	0.12
CASP8	167	TSG	3.33
ASXL1	183	TSG	6.26

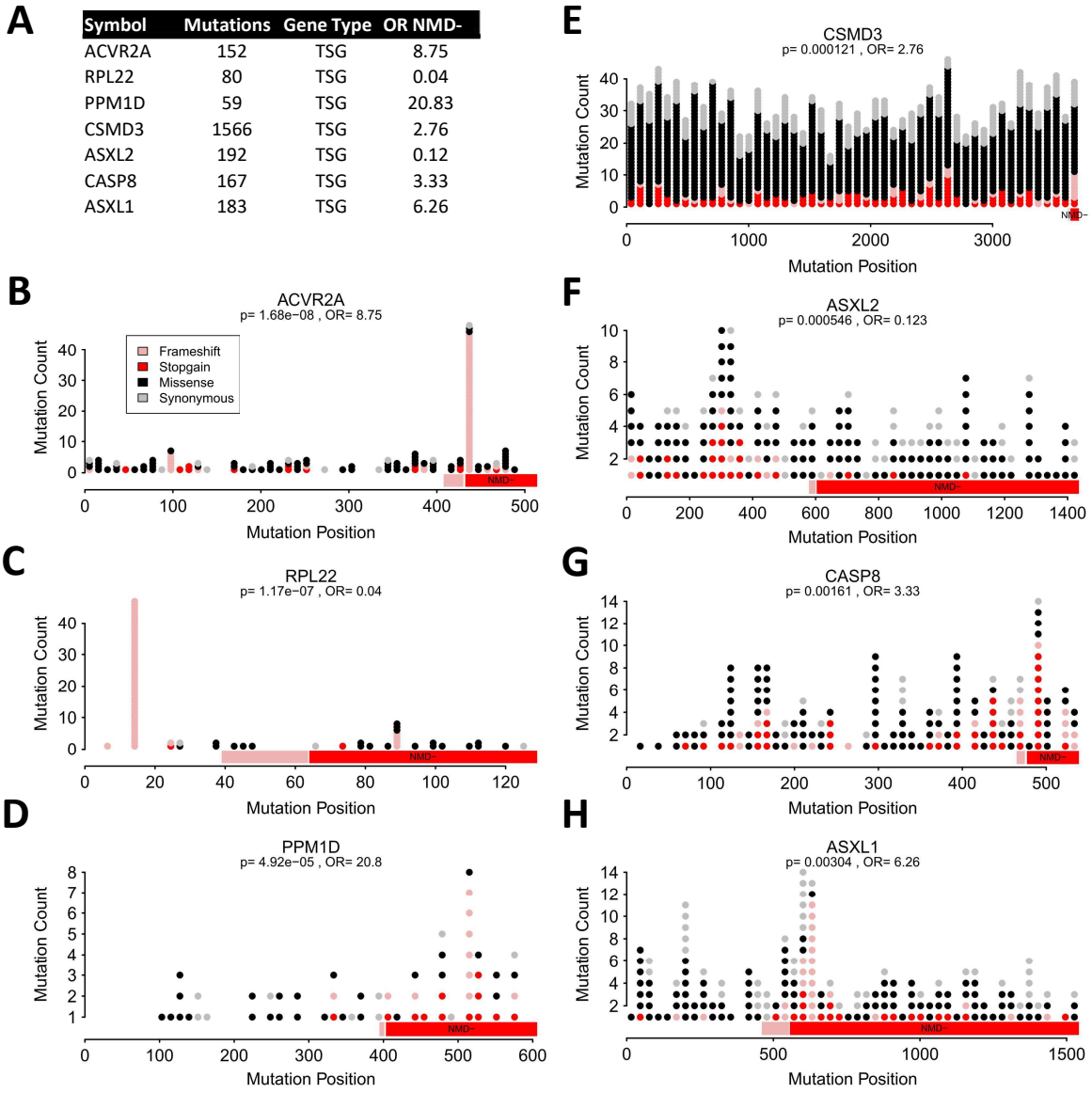


Figure S7. Additional oncogenes and tumor suppressor genes from the Cancer Gene Census with significant NMD- bias.

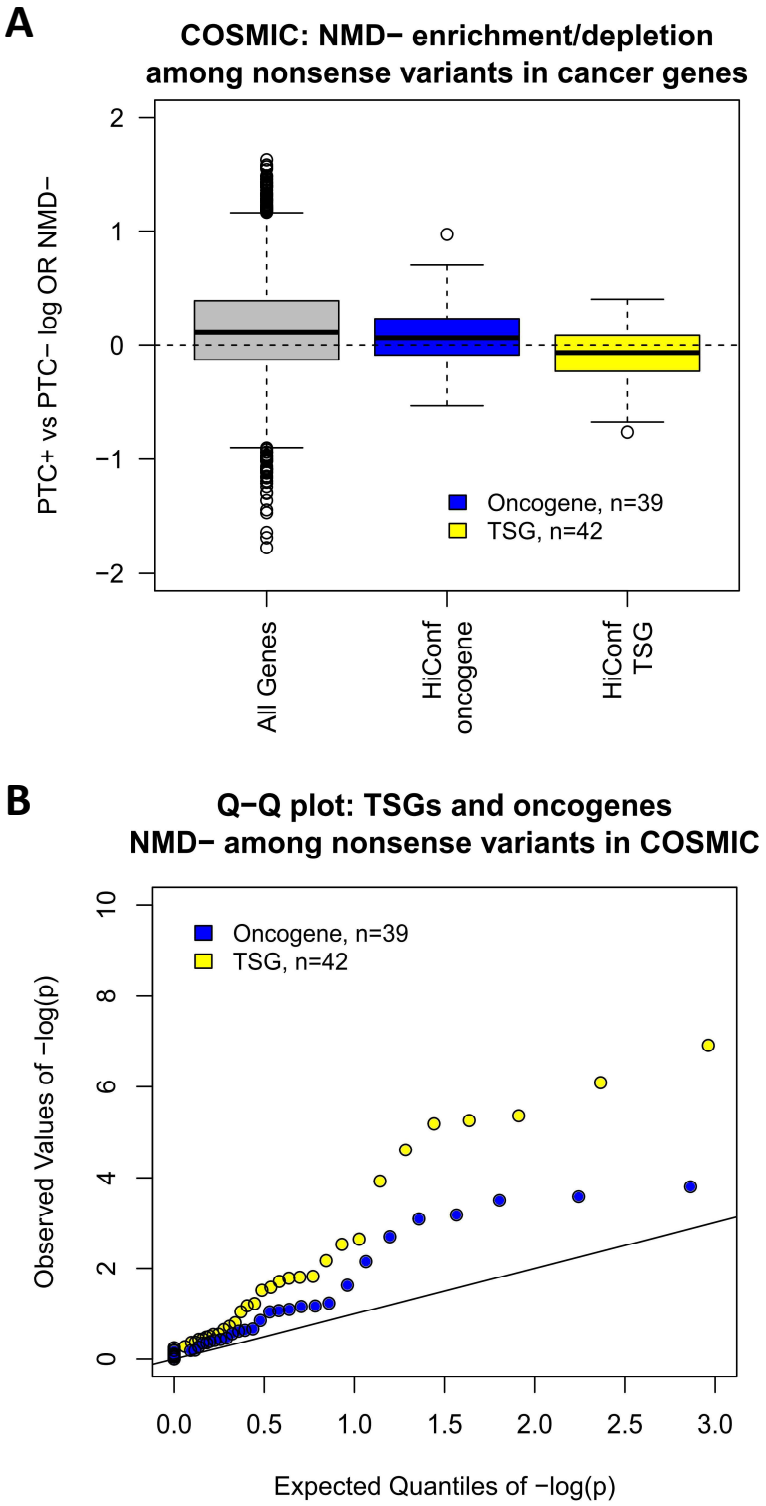


Figure S8. High-confidence cancer genes show similar patterns of NMD- among nonsense variants in the COSMIC dataset. A) Boxplot of log odds-ratios for all genes, HiConf oncogenes, and HiConf tumor suppressors based on nonsense variants only (frame=0) in the COSMIC dataset. Tumor suppressors are relatively NMD- depleted relative to the oncogenes (Student's t-test: p=0.012). B) QQ plot of p-values for the high confidence oncogenes and tumor suppressors. Two outliers with observed –log(p) > 15 are omitted.