The American Journal of Human Genetics, Volume *103*

Supplemental Data

Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles

Zeynep Coban-Akdemir, Janson J. White, Xiaofei Song, Shalini N. Jhangiani, Jawid M. Fatih, Tomasz Gambin, Yavuz Bayram, Ivan K. Chinn, Ender Karaca, Jaya Punetha, Cecilia Poli, Baylor-Hopkins Center for Mendelian Genomics, Eric Boerwinkle, Chad A. Shaw, Jordan S. Orange, Richard A. Gibbs, Tuuli Lappalainen, James R. Lupski, and Claudia M.B. Carvalho

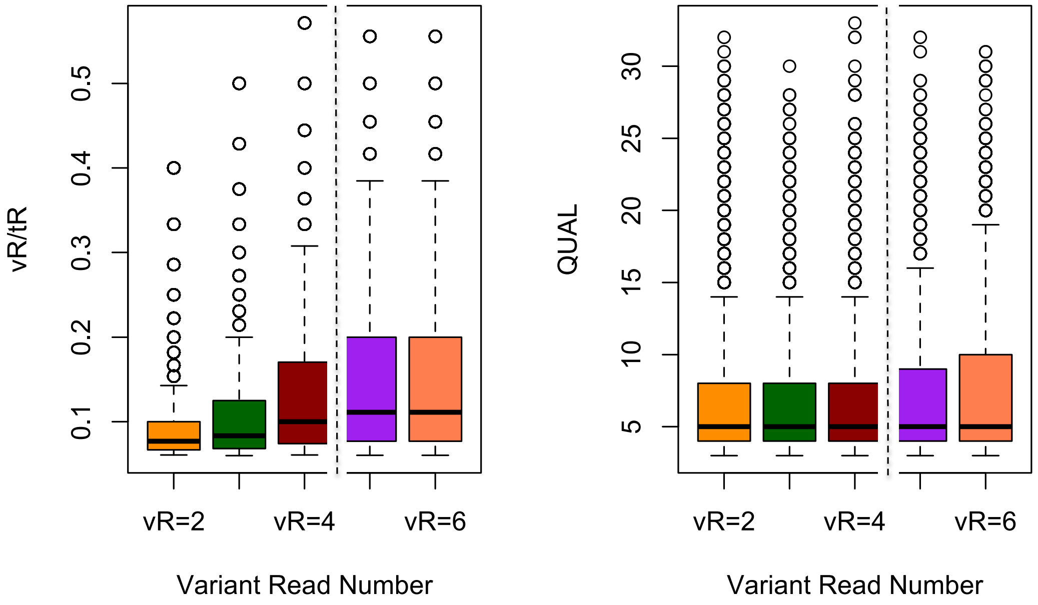


Figure S1: Dissection of quality control features for frameshifting indels and stopgain variants in the Baylor-CMG database

Box plots display variant read (vR) to total read (tR) ratio (vR/tR) and quality score values across variants called in the Baylor-CMG database with vR=2,3,4,5 and 6. vR/tR plateaus when vR reaches 5. Therefore, in the extraction of high-quality frameshifting indels and stopgain variants, the criteria that vR should be at least 5 reads was used.

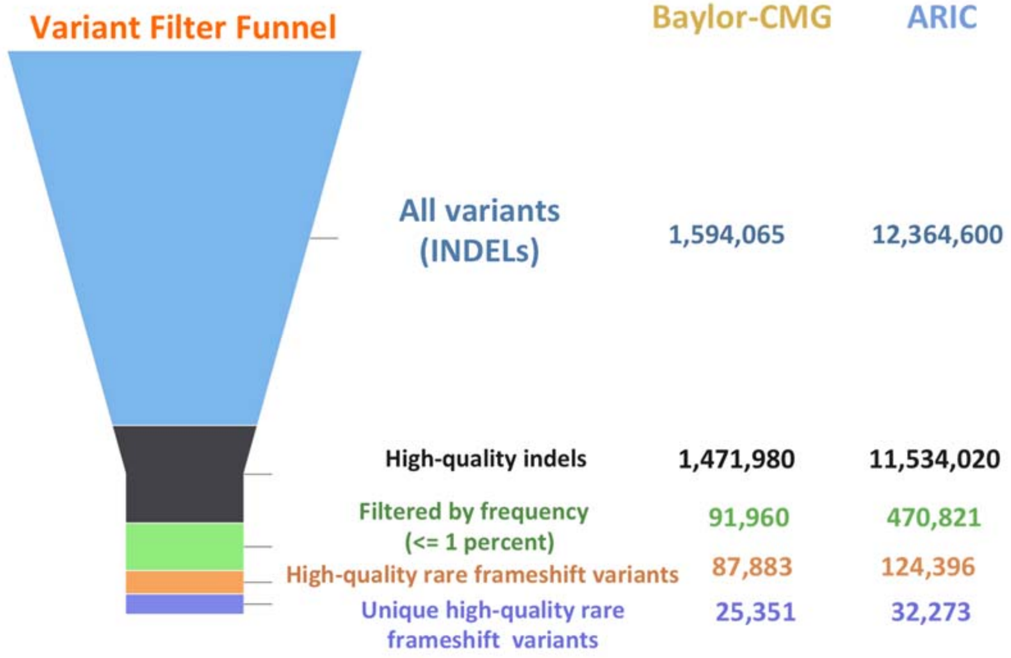


Figure S2: Variant filtering criteria for indels in Baylor-CMG and ARIC database

Variant prioritization workflow for frameshifting indels in the Baylor-CMG and the ARIC database was applied as follows. If an on-target indel has a variant read number (vR) >=5, it was included for further analysis. Then the indels were filtered based on the minor allele frequencies in our internal database (MAF <= 0.01). At this step, in-frame indels were also removed from the analysis.

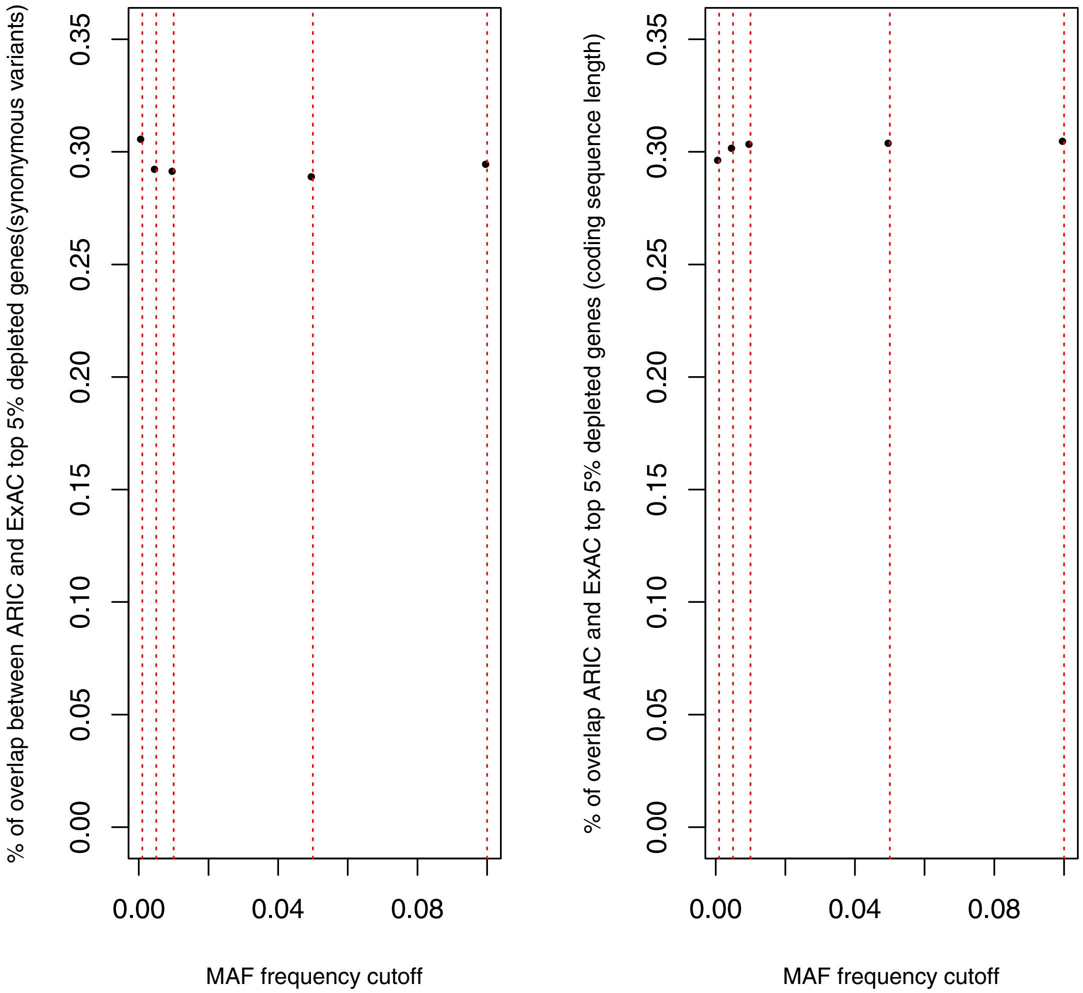


Figure S3: Sensitivity analysis in terms of the overlap of top 5% depleted genes for NMD$^-$ variants in ARIC vs. ExAC databases using: synonymous variant count normalization and coding sequence length normalization at different MAF cutoffs.

The sensitivity analysis was performed at different MAF cutoffs (0.1, 0.05, 0.01, 0.005 and 0.001) using synonymous variant count normalization (Left panel) and coding sequence length normalization (Right panel). The overlap between top 5% depleted genes for NMD$^-$ variants in ARIC vs. ExAC databases is similar using synonymous variant count normalization (29.2%) and coding sequence length normalization (30.3%) at MAF <= 0.01.

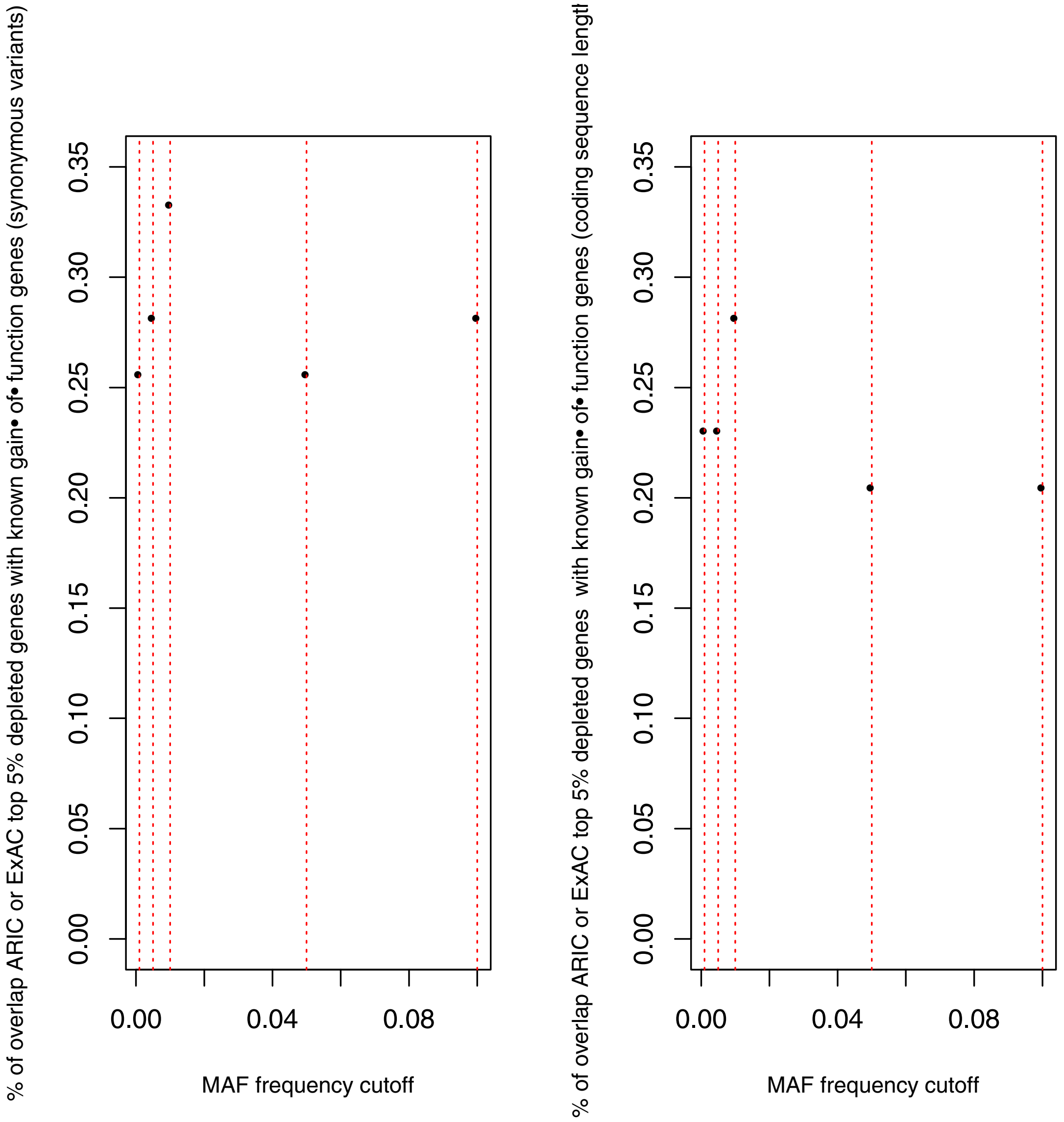
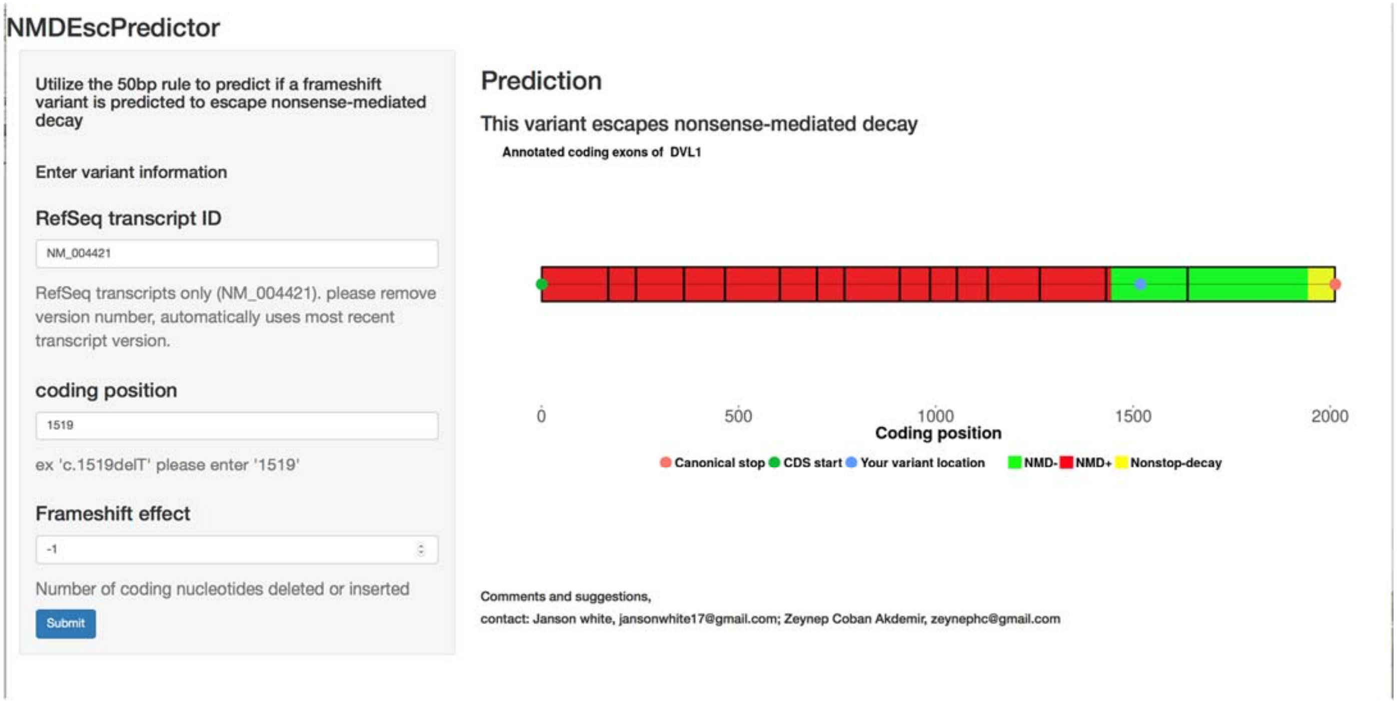


Figure S4: Sensitivity analysis in terms of the overlap of top 5% depleted genes for NMD$^-$ variants in either control database (N=1,996) with known gain-of-function genes extracted from OMIM database using synonymous variant count normalization and coding sequence length normalization at different MAF cutoffs.

The sensitivity analysis was performed at different MAF cutoffs (0.1, 0.05, 0.01, 0.005 and 0.001) using synonymous variant count normalization (Left panel) and coding sequence length normalization (Right panel). The overlap between top 5% depleted genes for NMD$^-$ variants in either control database and our control OMIM list of genes that cause disease via potential gain-of-function (N=39) drops significantly using coding-sequence length normalization (28.2%) compared to synonymous variant count normalization (33.33%) at MAF <= 0.01.



NMDEscPredictor

Utilize the 50bp rule to predict if a frameshift variant is predicted to escape nonsense-mediated decay

Enter variant information

RefSeq transcript ID

NM_004421

RefSeq transcripts only (NM_004421). please remove version number, automatically uses most recent transcript version.

coding position

1519

ex 'c.1519delT' please enter '1519'

Frameshift effect

-1

Number of coding nucleotides deleted or inserted

Prediction

This variant escapes nonsense-mediated decay

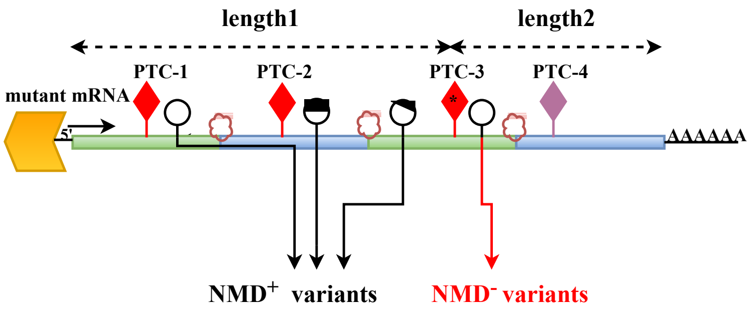
Annotated coding exons of DVL1

0
500
1000
1500
2000
Coding position
● Canonical stop ● CDS start ● Your variant location 🟩 NMD- 🟥 NMD+ 🟨 Nonstop-decay

Comments and suggestions,

contact: Janson white, jansonwhite17@gmail.com; Zeynep Coban Akdemir, zeynephc@gmail.com

Figure S5: An example of NMD prediction of a frameshifting indel in NMDescPredictor web-based tool (https://nmdpredictions.shinyapps.io/shiny/)



$$\boxed{\begin{array}{c}\text{\# of expected}\\ \text{NMD}^{-}\text{ variants}\end{array}} = \frac{\text{\# of NMD}^{-}\text{ variants + NMD}^{+}\text{ variants}}{\text{the total \# of rare synonymous variants / \# of rare synonymous variants in NMD}^{-}\text{ region}}$$

Figure S6: The normalization by the number of synonymous variants to calculate the expected number of NMD$^-$ variants

Number of expected NMD$^-$ variants per each canonical transcript was calculated as follows: The total # of variants was multiplied by the ratio of # of rare synonymous variants in NMD$^-$ region to the total # of rare synonymous variants.

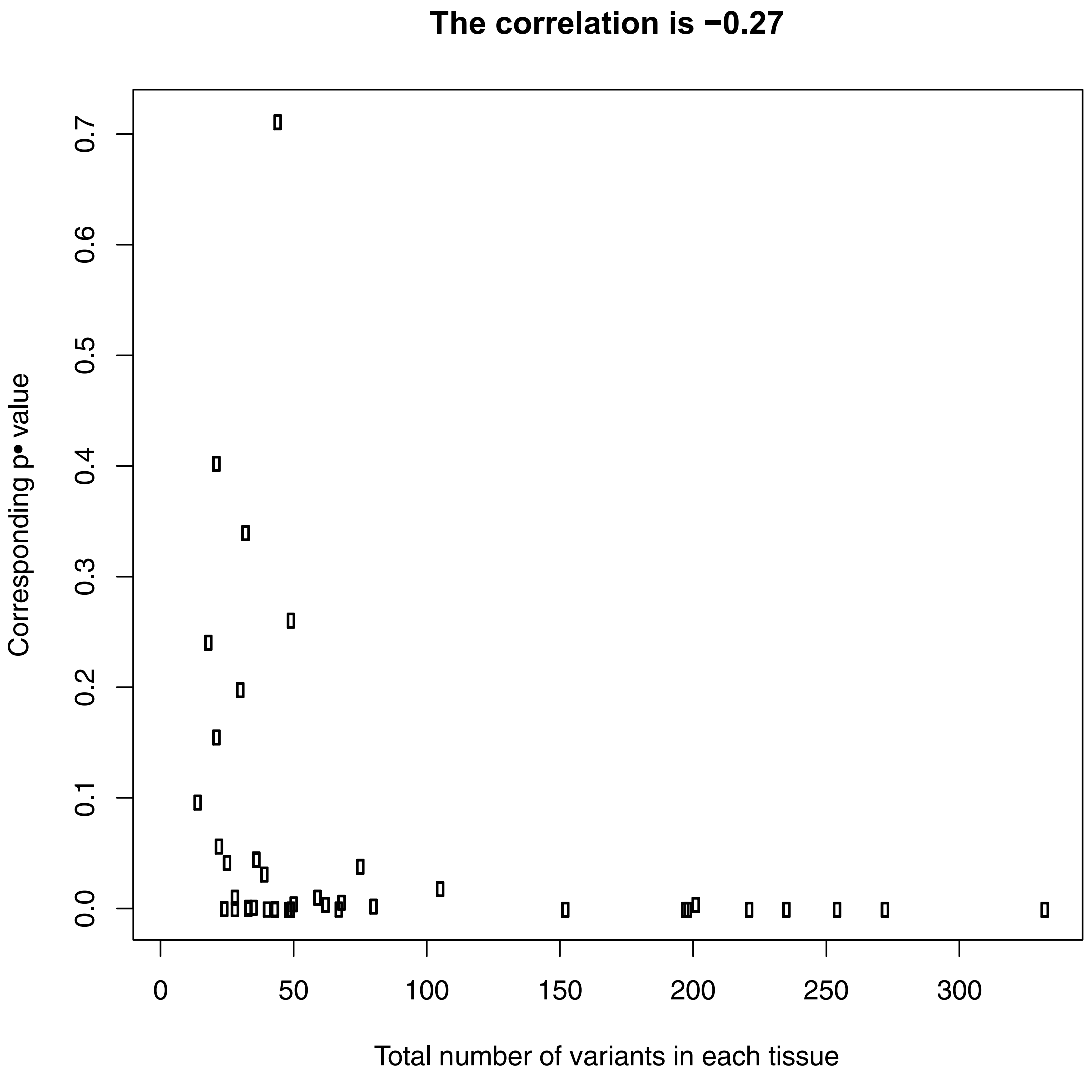


Figure S7: Correlation between the number of variants in each GTEx tissue and corresponding p-values.

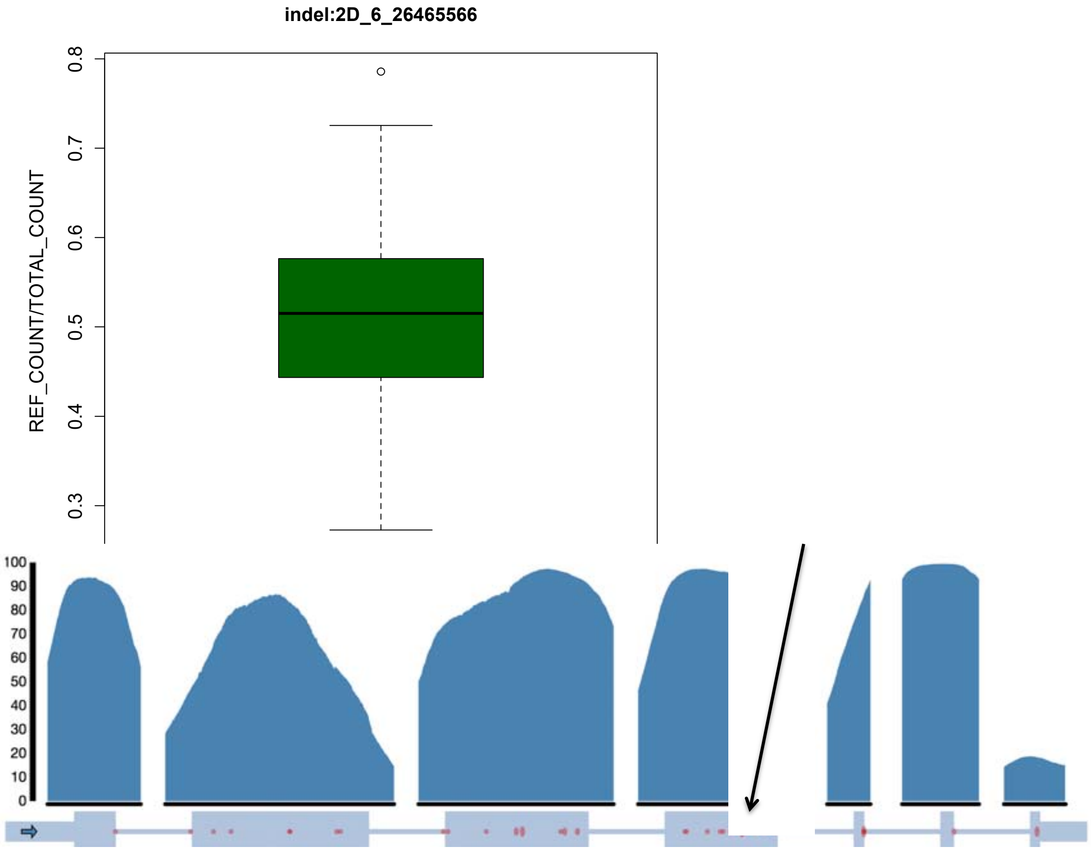


Figure S8: An example of a frameshifting indel in the fourth exon (nearly in the middle of the *BTN2A1* gene) was predicted to lead to NMD$^-$ by our tool. Allele-specific expression data in GTEx is concordant with this prediction.

The variant chr6:26,465,566_CAA>C that is located in exon 4 of transcript ENST0000042938 (7 coding exons in total) belonging to the *BTN2A1* gene was predicted to be NMD$^-$ by NMDEscPredictor based on the location of the boundary PTC. The average ratio of variant read count to total read count for this variant was extracted from GTEx RNA-Seq data and quantified as 0.478. This experimental finding supported the computational prediction of this variant as NMD$^-$ by NMDEscPredictor.

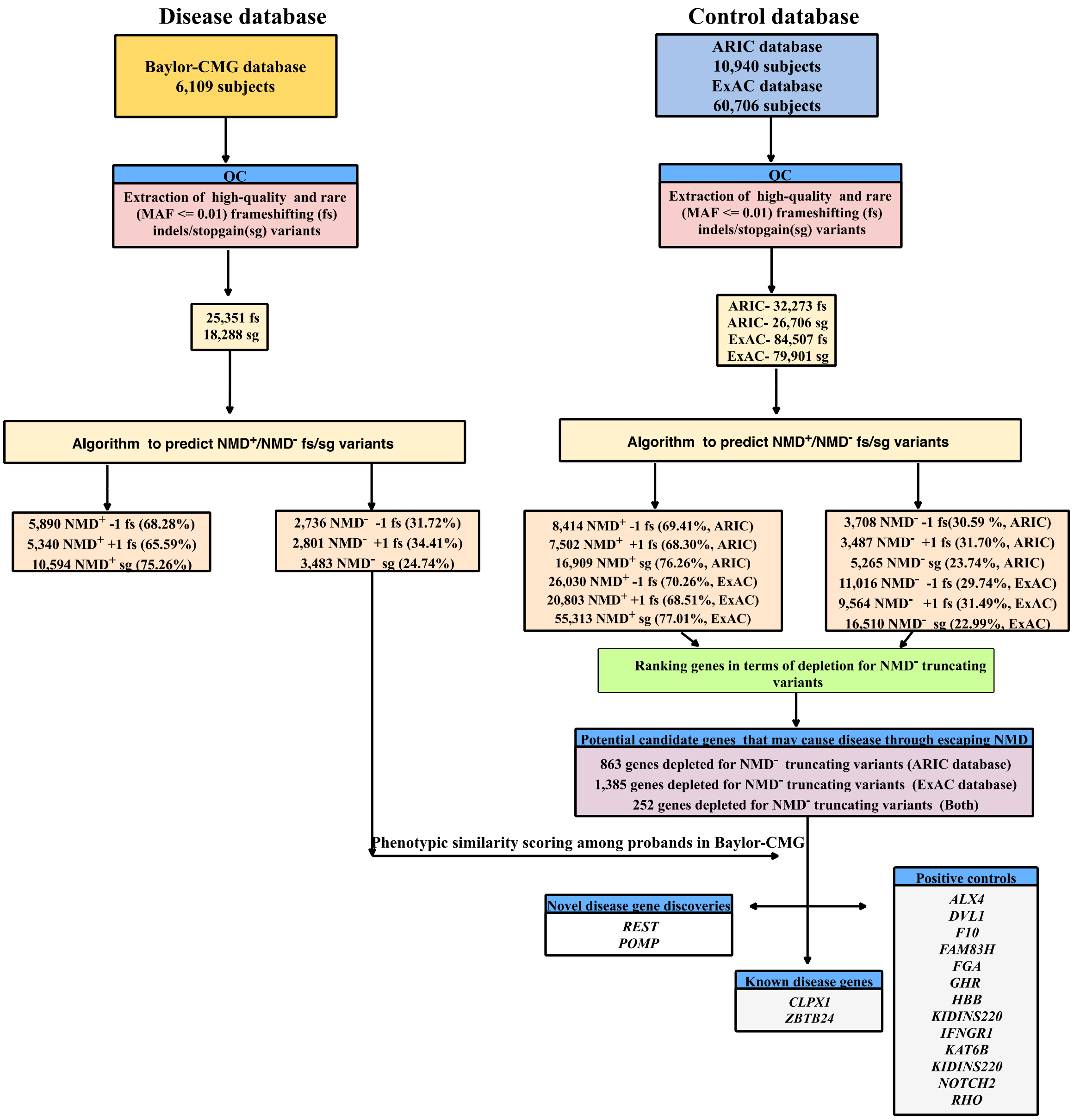

Disease database
Baylor-CMG database
6,109 subjects
OC
Extraction of high-quality and rare (MAF <= 0.01) frameshifting (fs) indels/stopgain(sg) variants
25,351 fs
18,288 sg
Algorithm to predict NMD^{+}/NMD^{-} fs/sg variants
5,890 NMD^{+} -1 fs (68.28%)
5,340 NMD^{+} +1 fs (65.59%)
10,594 NMD^{+} sg (75.26%)
2,736 NMD^{-} -1 fs (31.72%)
2,801 NMD^{-} +1 fs (34.41%)
3,483 NMD^{-} sg (24.74%)
Control database
ARIC database
10,940 subjects
ExAC database
60,706 subjects
OC
Extraction of high-quality and rare (MAF <= 0.01) frameshifting (fs) indels/stopgain(sg) variants
ARIC- 32,273 fs
ARIC- 26,706 sg
ExAC- 84,507 fs
ExAC- 79,901 sg
Algorithm to predict NMD^{+}/NMD^{-} fs/sg variants
8,414 NMD^{+} -1 fs (69.41%, ARIC)
7,502 NMD^{+} +1 fs (68.30%, ARIC)
16,909 NMD^{+} sg (76.26%, ARIC)
26,030 NMD^{+} -1 fs (70.26%, ExAC)
20,803 NMD^{+} +1 fs (68.51%, ExAC)
55,313 NMD^{+} sg (77.01%, ExAC)
3,708 NMD^{-} -1 fs(30.59 %, ARIC)
3,487 NMD^{-} +1 fs (31.70%, ARIC)
5,265 NMD^{-} sg (23.74%, ARIC)
11,016 NMD^{-} -1 fs (29.74%, ExAC)
9,564 NMD^{-} +1 fs (31.49%, ExAC)
16,510 NMD^{-} sg (22.99%, ExAC)
Ranking genes in terms of depletion for NMD^{-} truncating variants
Potential candidate genes that may cause disease through escaping NMD
863 genes depleted for NMD^{-} truncating variants (ARIC database)
1,385 genes depleted for NMD^{-} truncating variants (ExAC database)
252 genes depleted for NMD^{-} truncating variants (Both)
Phenotypic similarity scoring among probands in Baylor-CMG
Novel disease gene discoveries
REST
POMP
Known disease genes
CLPX1
ZBTB24
Positive controls
ALX4
DVL1
F10
FAM83H
FGA
GHR
HBB
KIDINS220
IFNGR1
KAT6B
KIDINS220
NOTCH2
RHO

Figure S9: Pipeline workflow for our algorithm

At the first step of the algorithm, as a quality control (QC) step, high-quality and rare (MAF <=0.01) frameshifting indels and stopgain variants were extracted from the Baylor-CMG (disease database, 6,109 exomes) and the ExAC and the ARIC control databases (60,706 and 10,940 exomes, respectively). Then, using the NMDEscPredictor algorithm, frameshifting indels and stopgain variants in each database were categorized into three categories as NMD escaping (NMD^{-}), NMD triggering (NMD^{+}). We then removed the variants that could not be annotated to any canonical transcript in Ensembl version 19 as well as variants mapped to transcripts without a predicted PTC, without a boundary PTC or mapped to single-

exon canonical transcripts. Next, each gene in the genome is ranked based on the depletion of NMD$^-$ relative to NMD$^+$ variants in control databases (NMD escape intolerance score metric). This analysis revealed a total of 1, 996 genes as the most depleted in either database (i.e. ranked in the top 5%). Those genes were further investigated for NMD$^-$ variants in the Baylor-CMG database (disease database). A subset of significantly depleted genes has NMD$^-$ variants in multiple unrelated individuals with similar clinical phenotypes (based on the phenotypic similarity scoring) in the Baylor-CMG database. Some of those genes were found to be causative for human disease through escape from NMD and include novel and known disease genes.

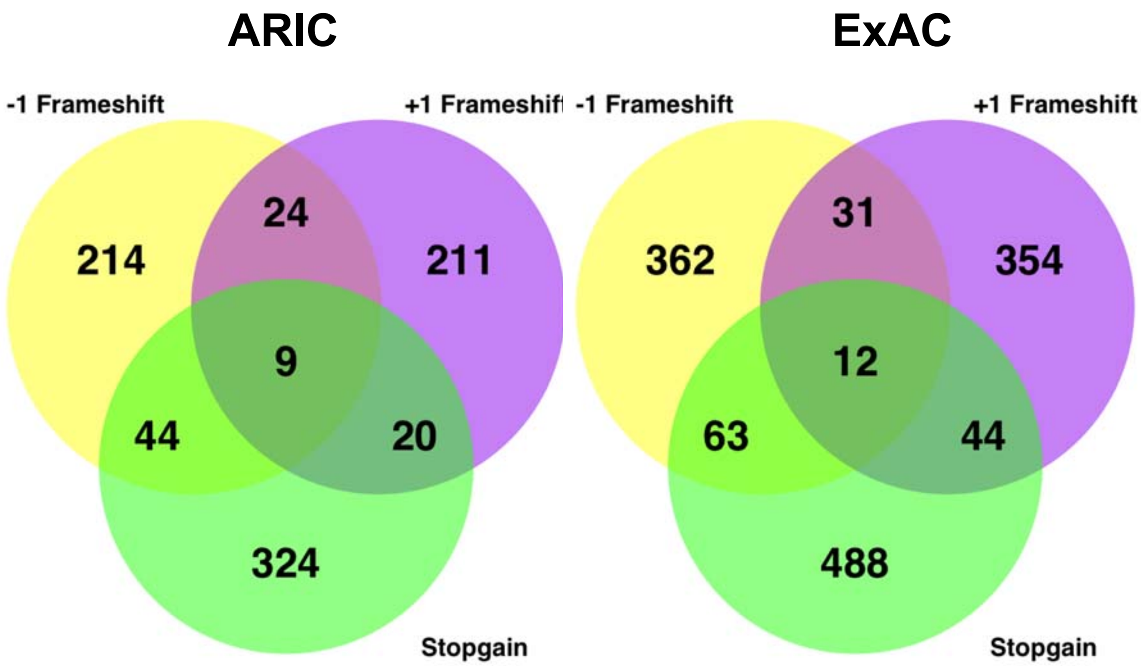


Figure S10: The overlapping number of genes depleted for NMD$^-$ variants (i.e. ranked in the top 5%) in each category of truncating variants within ARIC and ExAC database

Venn Diagrams display the number of genes depleted for NMD$^-$ variants (i.e. ranked in the top 5%) in -1 frame, +1 frame and stopgain categories in the ARIC and ExAC control databases.

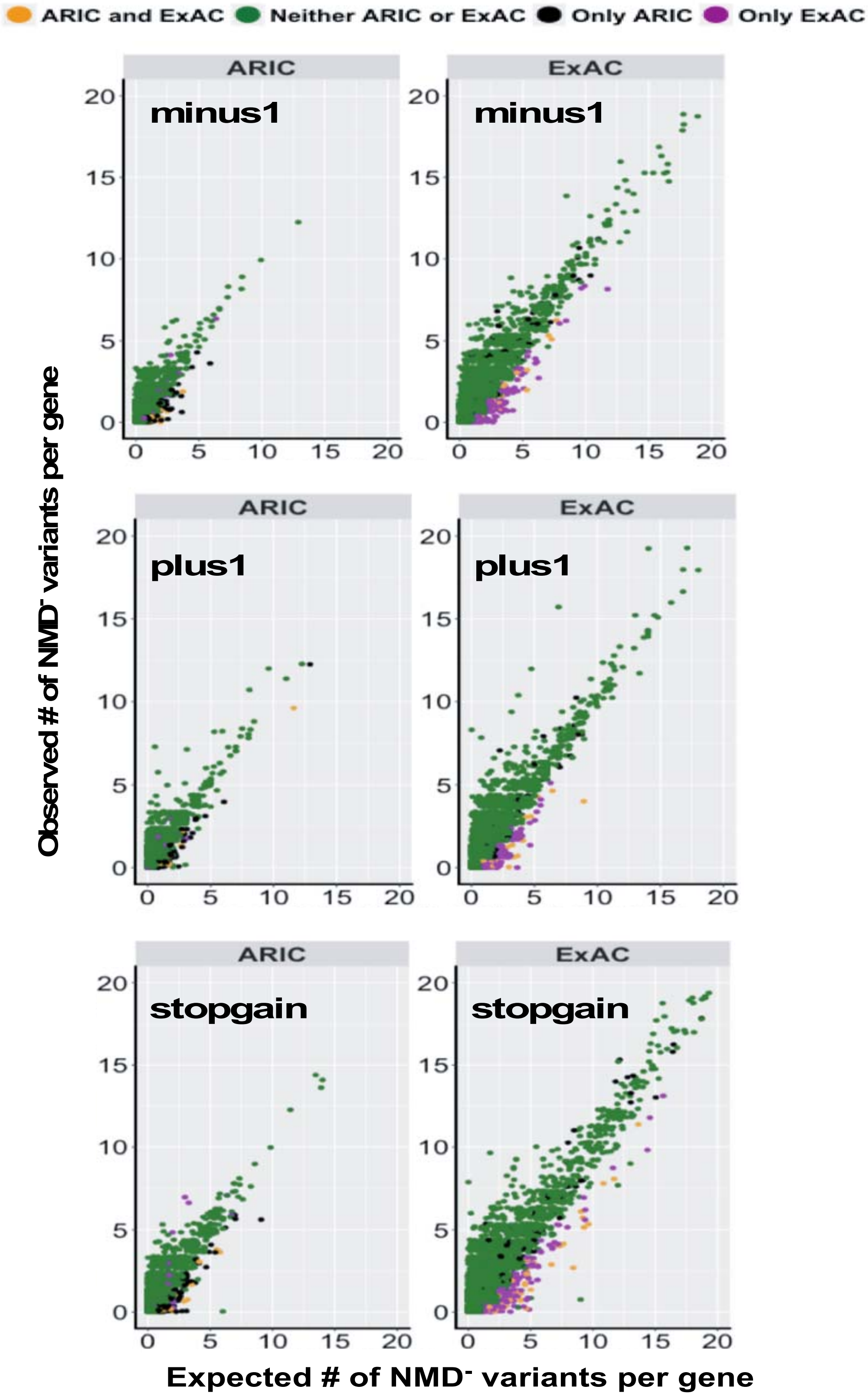

ARIC and ExAC
Neither ARIC or ExAC
Only ARIC
Only ExAC
ARIC
ExAC
minus1
minus1
plus1
plus1
stopgain
stopgain
0
5
10
15
20
Observed # of NMD- variants per gene
Expected # of NMD- variants per gene

Figure S11: Expected and observed number of NMD$^-$ variants in each category of truncating variants in the ARIC and ExAC control databases

The plots show the expected # of escape (NMD$^-$) variants (x axis) compared to the observed # of escape (NMD$^-$) variants (y axis) per gene in the -1 frame, +1 frame and stopgain categories in the ARIC and ExAC control databases. The genes for which the observed # of variants/gene relative to expected # of variants/gene were depleted in both databases were colored in orange, and those depleted only in ARIC database were colored in black and those depleted only in the ExAC database were colored in purple and those not depleted in either control database were colored in dark green.

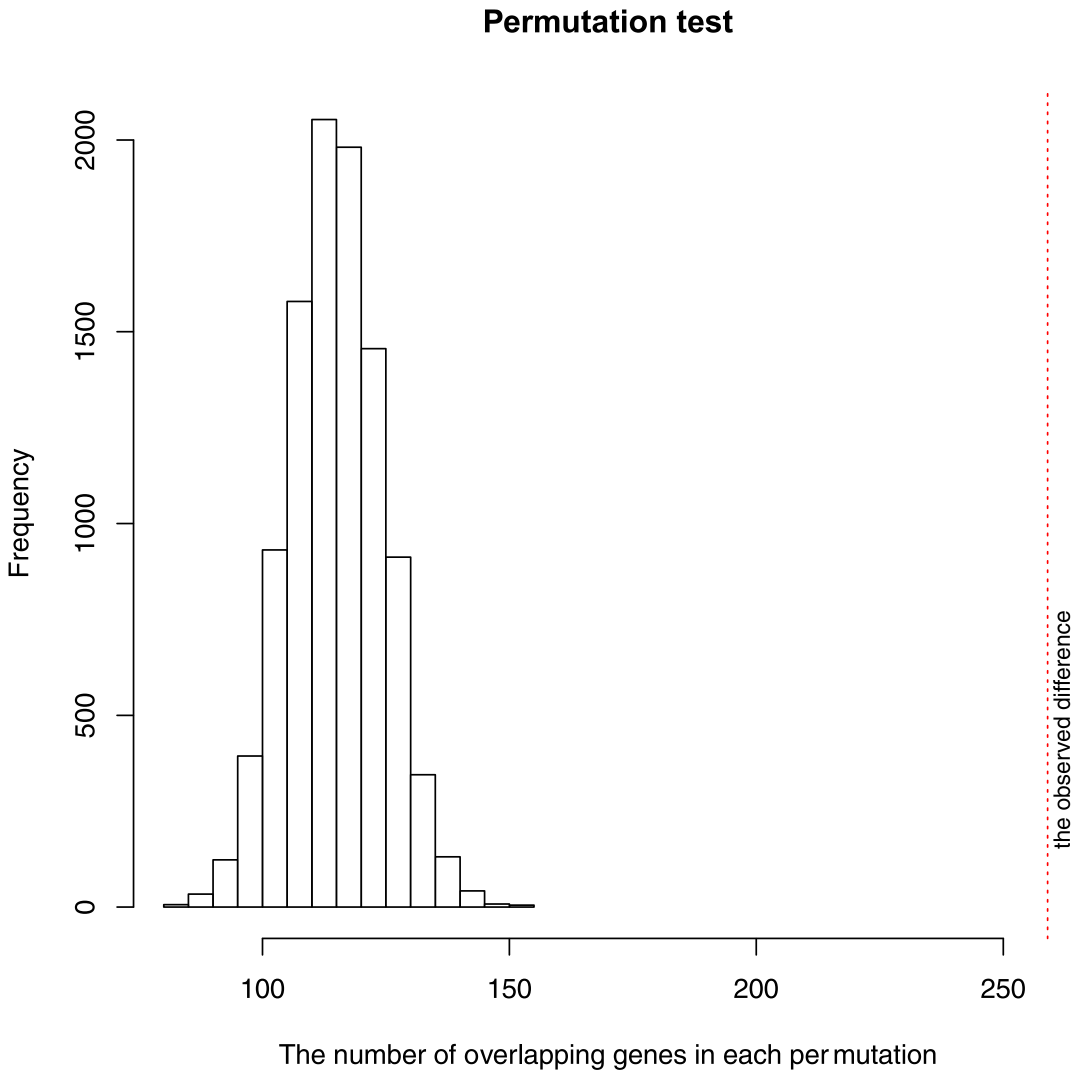


Figure S12: Permutation test results to quantify how often overlap between top 5% depleted genes for NMD$^-$ variants in ARIC vs. ExAC control database might occur by chance

10,000 permutations were done by generating random subsets of all of the genes considered in the analysis (N=16,411) at the size of ExAC gene set (N=1,385). In each permutation, the number of genes overlapping between ARIC and random set of genes were calculated. The red dashed line shows the observed value (N=252) of the number of genes overlapping between top 5% depleted genes for NMD$^-$ variants in ARIC (N=863) vs. ExAC database (N=1,385).

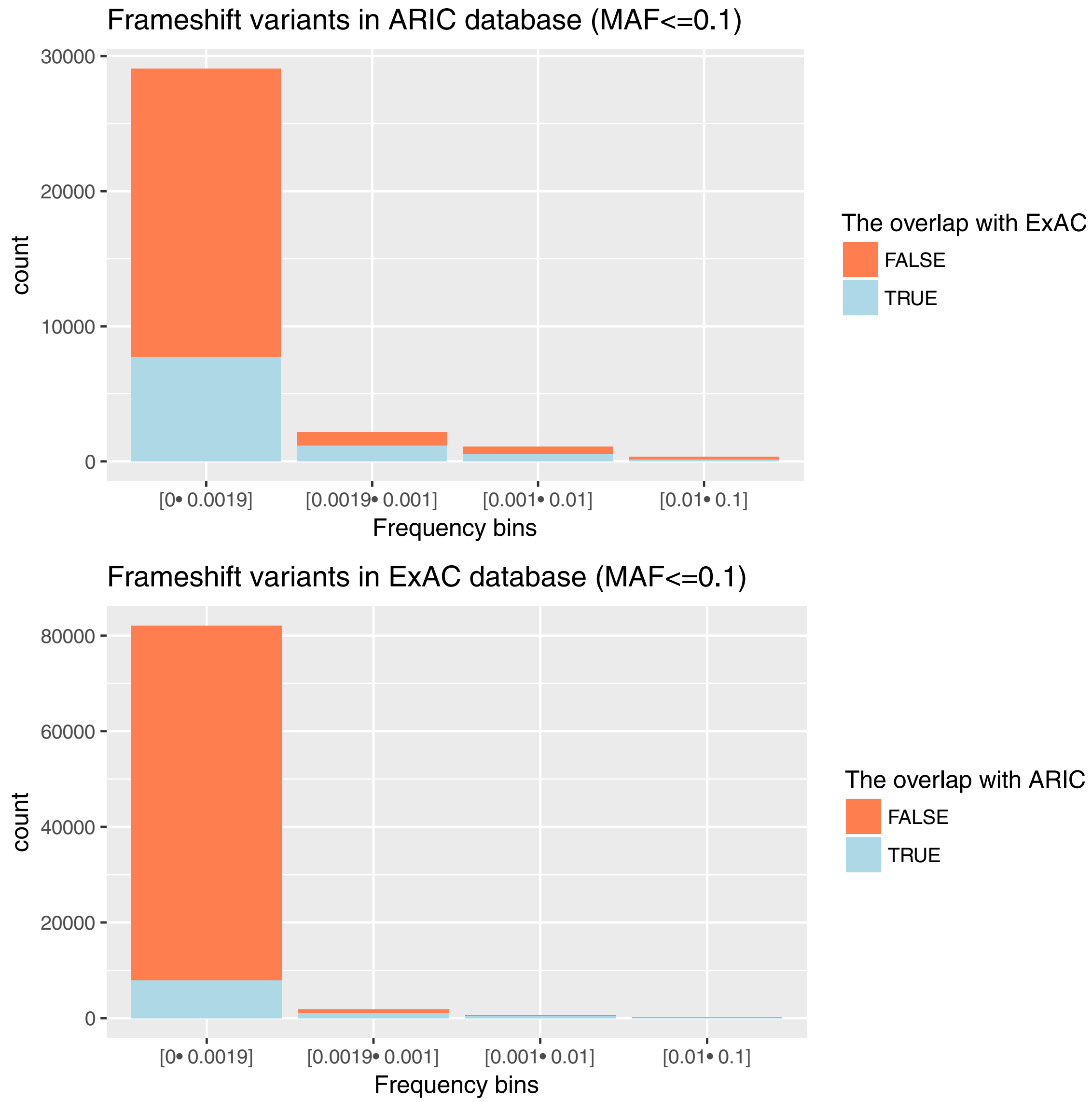
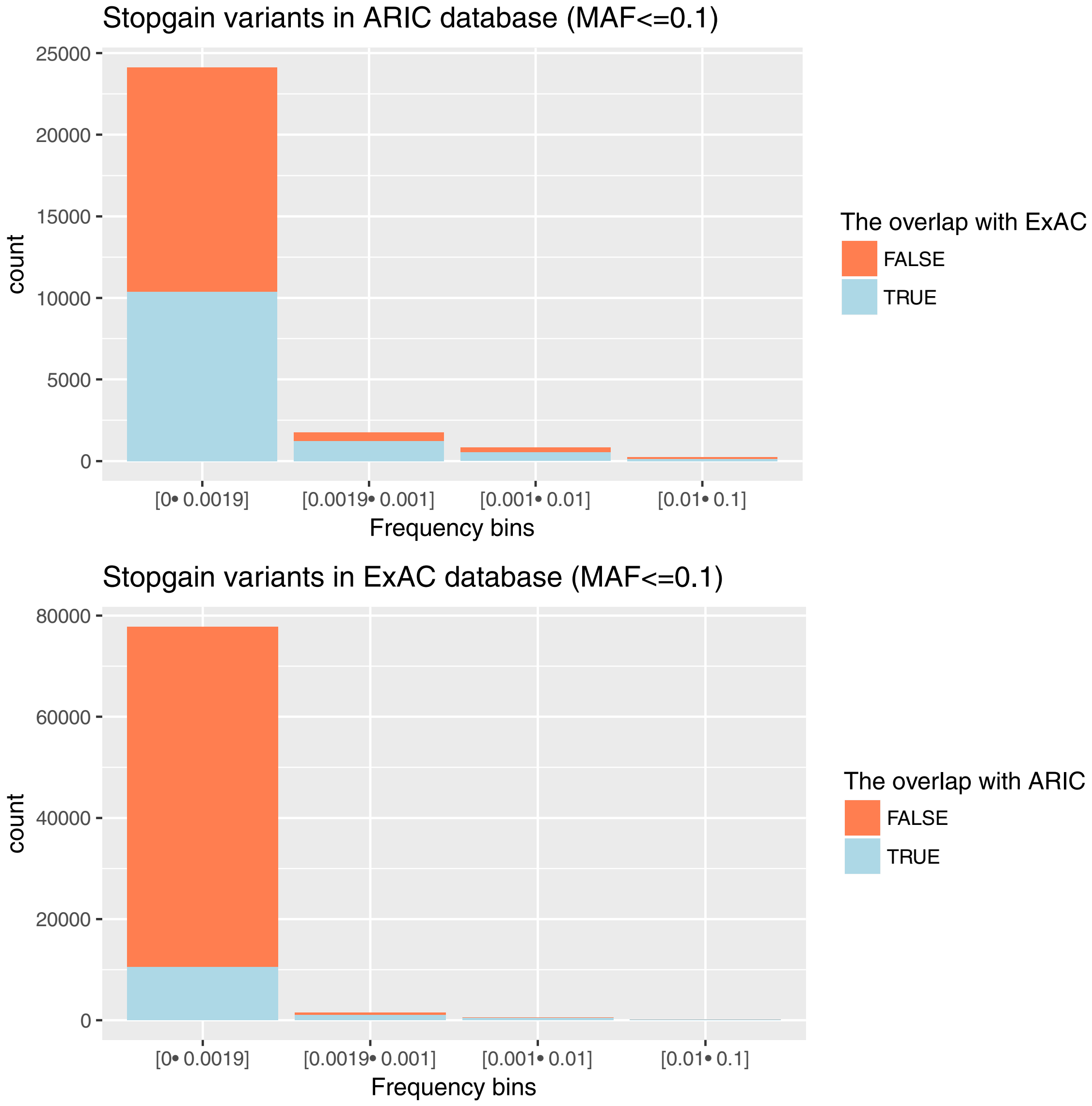


Figure S13: The percentage of overlapping frameshift variants between the ExAC and ARIC databases at different MAF cutoffs

The number of overlapping frameshift variants (light blue boxes) between the ExAC and ARIC databases were shown at different MAF intervals including 0-0.0019, 0.0019-0.001, 0.001-0.01 and 0.01-0.1.



Stopgain variants in ARIC database (MAF<=0.1)

Stopgain variants in ExAC database (MAF<=0.1)

Figure S14: The percentage of overlapping stopgain variants between the ExAC and ARIC databases at different MAF cutoffs

The number of overlapping stopgain variants (light blue boxes) between ExAC and ARIC databases were shown at different MAF intervals including 0-0.0019, 0.0019-0.001, 0.001-0.01 and 0.01-0.1.