SUPPLEMENTARY MATERIAL AND METHODS

Targeted gene sequencing

Custom RNA baits were designed complementary to all coding exons of 111 genes (genes listed in supplementary table 1) as per manufacturers' guidelines (SureSelect, Agilent, UK). Genomic DNA was extracted from peripheral blood granulocytes or bone marrow mononuclear cells for 738 patients with MDS, followed by whole genome amplification (WGA) with Phi 29 (Qiagen, UK). To control for the sensitivity of our analysis in a setting where a constitutional sample is not available for each sample in the study, as well as to evaluate the effect of WGA in the allelic representation of the variant calls, we used six independent control datasets. These include: exome sequencing of genomic and constitutional DNA for 10 samples that underwent WGA and targeted re-sequencing in the present study; technical replicates using genomic and WGA DNA for 18 samples; comparison of targeted re-sequencing results for *SF3B1* and *TET2*, to that obtained previously[1] using orthogonal sequencing platforms; 111 targeted gene screen of 22 normal DNA samples derived from peripheral blood from in house control DNA; analysis of 56 constitutional samples available from in house exome studies for the 111 genes using against the same unmatched reference sample and variant selection pipelines; exome or wholegenome sequencing data from 317 constitutional DNA samples.

Comparison of the variant allele fraction estimates was used by taking into account the number of reads reporting the variant as well as the total depth. and

calculating the 95% CI under a binomial distribution. If the two 95% CI estimates overlapped then the two estimates were considered to not be different.

A total of 125µl of 40ng/µl of WGA DNA was fragmented to an average insert size of 145bp (75-300) and subjected to Illumina DNA sequencing library preparation using the Bravo automated liquid handling platform. Individual samples were indexed using a unique DNA barcode via 6 cycles of PCR. Equimolar pools of 16 libraries were prepared and hybridized to custom RNA baits following the Agilent SureSelect protocol. Enriched pools of 96 cases were sequenced on two lanes of an Illumina HiSeq machine using the 75-base pair paired-end protocol.

Sequence variant detection and filtering criteria

Base substitutions and small insertions or deletions were identified by comparison of 738 MDS patients against an unmatched normal sample using established bioinformatic algorithms [2-4]. To account for the absence of matched control a bespoke variant selection pipeline was developed. Each putative variant was annotated using the following resources:

1. Known constitutional polymorphisms using known human variation databases, Ensembl GRCh37.5, 1000 genomes release 2.2.2 and ESP6500 [5-8];
2. Known somatic variation in myeloid and other common malignancies as reported in COSMIC v60 [9,10];

3. The presence of the same sequence change in exome or whole genome sequencing data derived from 317 constitutional DNA samples analysed in CGP (CGP normal panel); Specifically where the same base change was observed in at least two constitutional sample at allele fractions greater than 10% and the variant has not previously been confirmed as somatic in COSMIC or in two or more samples at < 10%.
4. Sequence context 5' and 3' to the reported sequence change highlighting regions of homopolymer sequence that are prone to PCR slippage and artefacts altering the last base of the homopolymer or inserting the same base as the homopolymer at +1, +2 of the track and often present in unidirectional reads and < 10% variant allele burden;
5. Variant specific metrics to include protein annotation, sequence depth and % of reads reporting the variant allele.

To enrich for high-confidence somatic variants that impact on protein function further filtering was conducted using the following criteria:

1. Removal of all variants with a predicted effect of a silent amino acid change on all transcripts corresponding to each gene;
2. Removal of known polymorphisms present in either of the human variation databases at a population frequency > 0.0014 (reflecting the population incidence of myeloid disease and potentially rare variants that could be associated with myeloid malignancies) and also present in the CGP normal panel;

3. Removal of known polymorphisms present in either of the human variation databases at a population frequency > 0.0014 (based on availabl incidence data of myeloid malignancies in the population and evaluation of known driver events i.e. V617F in public databases) unless variant is present as confirmed somatic mutation in COSMIC;
4. Removal of known polymorphisms present in human variation databases at a population frequency < 0.0014 and also represented in the extended normal CGP panel, available form in house exome and whole genome sequencing projects;
5. Retention of all variants present in human variation databases at a population frequency < 0.0014 and also present in COSMIC as confirmed somatic in Haematopoietic tissue;
6. Removal of all sequence variants that were represented in at least 2 normal individuals in the CGP normal panel with a minimum variant allele proportion of 10%;
7. Removal of variants present within regions prone to sequence context specific artifacts, including regions of high depth, enriched for reads of low mapping quality that harbor multiple mismatches;
8. Removal of all 1bp insertions or deletions present adjacent to regions of more than 5 homopolymer bases (i.e insA adjacent to AAAAA) and a variant allele proportion of < 12% and evidence of occurrence in CGP normal panel;
9. Removal of highly recurrent calls that presented with a narrow allele distribution in both MDS sample set and in the normal controls. These can

either be distributions of < 8-10% indicative of artifacts or distributions very close to 50% indicative of polymorphisms.

Sequence variant annotation

Once low confidence or likely polymorphisms were removed from the dataset, each high confidence variant was annotated as oncogenic, possible oncogenic or unknown in accordance to prior evidence in the literature in respect to the variants or genes association with myeloid disease. To reflect the confidence that one would use these as diagnostic biomarkers in the clinic variants were annotated conservatively. Broad variant annotation parameters are listed by variant type as follows:

a. Oncogenic

- Known oncogenic variants previously reported in the literature;
- Novel recurrent variants ($n \geq 2$) that cluster with known somatic variants in well characterised myeloid driver genes;
- Truncating variants (nonsense mutations, essential splice mutations or frameshift indels) in genes implicated in myeloid malignancies through acquisition of loss of function mutations.

b. Possible oncogenic

- Previously unreported variants that cluster (±3aa) with known oncogenic variants in COSMIC.

c. Unknown

- Variants identified outside the range of frequent driver variants in genes with known oncogenic variants;

- Variants (even if recurrent) in genes whose role in myeloid disease is not yet established.

Thus, for example, for tumor suppressor genes, nonsense mutations, frameshift indels and essential splice site mutations were classified as 'known oncogenic', as were sites of recurrent missense substitutions reported in the published literature. Amino acid substitutions or in-frame indels nearby previously reported hotspots were called 'possible oncogenic', whereas one-off missense mutations at some distance from recurrent variants were classified as of 'unknown significance'. Across previously well-characterized genes, such as *EZH2* and *CBL*, the pattern of mutations identified as 'known oncogenic' in our screen followed very similar distributions to those documented in the literature (Supplementary Figure 10), confirming that the manual curation of variants outlined here produced sensible predictions.

Karyotypic abnormalities were assigned with a minimum requirement of 2 metaphases and most often 5 and in accordance to the Mitelman rules. For monosomies; at least 3 metaphases with the same loss whilst for trisomies or translocations; at least 2 metaphases For cytogenetic studies, the following lesions were considered to be driver events: chromosome 3q rearrangements; deletions of 5q; monosomy 7 or del(7q); trisomy 8; del(11q); del(12p); abnormalities of chromosome 17; trisomy 19; del(20q); and complex karyotype. Neither del(Y) nor isolated non-recurrent cytogenetic lesions were considered as drivers, although presumably some of the latter category might indeed be relevant for

disease biology. Each karyotypic abnormality with a driver status accounted as a single event.

Collection of clinical data

For all participating patients, data relating to clinical pathology at presentation as well as long long-term follow up were collected. Minimal demographic information to include patients' sex and age of diagnosis was also obtained. Clinical data at diagnosis included date of diagnosis, bone marrow morphology, complete blood counts, karyotype and cytogenetic reports as well as computed WHO 2008 category and related IPSS score [11,12]. Information on the date of sample used in the study as well as the age of patient at sampling was also obtained. Follow up data included date of last follow up, outcome at last follow up, transformation to AML and where appropriate date of transformation to AML. The cohort had a variety of MDS treatments, reflecting the heterogeneous nature of the disease, treatment practices across the centers and the long period of time over which patients and samples were collected. We have therefore not factored therapy into the prognostic analyses.

Copy number analysis

Copy number evaluation of the targeted regions was performed using libraries from the ASCAT algorithm [13]. In brief, sequence data were interrogated for known polymorphic SNPs in the target capture region across all 738 MDS samples as well as constitutional DNA samples sequenced with the same capture array (Materials and methods). Both alleles of polymorphic SNPs were used to

compute B-allele frequencies and logR ratios. LogR levels were calculated as $\log_2$ of the sequencing yield, normalized with SNP data derived from sequencing data of constitutional DNA samples and also smoothed by GC correction. To calculate B-allele frequencies, the counts of sequencing reads showing the most commonly occurring allele were divided by the counts of sequencing reads showing the two most common alleles at a specific SNP locus for a specific sample. Areas of UPD were manually assigned by identification of a cluster of a minimum of 4 consecutive SNPs in the capture array, with B-allele frequencies for AA and BB genotypes deviating in intensity from those expected within the sample (when normal contamination is taken into account- hence the shift in positioning) and normal Log R .

Correcting variant allele fraction of indels

We observed a significantly lower variant allele fraction for all reported indels than for substitutions, a bias that was worse for insertions than deletions and for larger indels than shorter indels. To correct this for any given indel, we created a variant haplotype containing the indel and, using BLAT, realigned unmapped singleton reads anchored by a mapped read to the region of the indel. After PCR duplicate removal, we then added the newly mapped reads reporting the indel to the observed allele counts.

Statistical analysis

We analysed associations among genes and cytogenetic lesions to identify examples of pairs of genes that show a tendency to either co-occurrence (both genes mutated in more patients than expected) or mutually exclusive mutation

(one or other gene mutated, but rarely both together). Fisher tests were used to generate unadjusted p values from the 2x2 contingency table with counts for each pair of genomic alterations. The table entries were counts of patients with 'known oncogenic' or 'possible oncogenic' mutations in the particular gene versus unmutated and 'mutations of uncertain significance'. These were then adjusted for multiple hypothesis testing using the Benjamini and Hochberg approach.

Analysis of subclonal and clonal mutations

To test whether there was evidence for or against the null hypothesis that all driver mutations within a given patient were present in the same fraction of cells, we used Pearson goodness of fit tests, correcting for copy number of the locus. Where all mutations occur at the same copy number, we can use straightforward Fisher tests or chi-squared tests to evaluate heterogeneity. Where the copy number levels of the loci are different, this needs to be refined. For example, suppose we have a male patient with a mutation in *STAG2* (on the one copy of chromosome X) and a mutation in *SF3B1* (diploid chromosome 2, without LOH). Then, under the null hypothesis that both mutations are present in ρ cells, and observing r_{STAG2} reads reporting the mutation from n_{STAG2} reads etc, we can without much difficulty show that the maximum likelihood estimate of ρ is given by the solution to the quadratic:

$$(n_{STAG2} + n_{SF3B1})\hat{\rho}^2 - (r_{STAG2} + 2r_{SF3B1} + 2n_{STAG2} + n_{SF3B1})\hat{\rho} + 2(r_{STAG2} + r_{SF3B1}) = 0$$

Then, using this estimated value, we can calculate the expected number of reads in each cell of the 2x2 contingency table and calculate the Pearson chi-squared statistic in the conventional manner. We set a threshold of $p \leq 0.01$ for determining whether clonal heterogeneity existed in the patient. To reconstruct temporal precedences, we applied the pigeonhole principle, as described [14], only considering them informative when the phylogenetic relationship was unambiguous. From the set of genes in which 5 or more precedence were observed, we applied Bradley-Terry models using penalized maximum likelihood to the observed precedences.

Imputation of missing clinical data

Outcome data were available for 595 (79%) patients in the cohort, with most of the relevant prognostic variables complete. The missing clinical data were multiply imputed with the mice R package [15] using 5 chains and 100 iterations. Missing cytogenetic data were complemented prior to imputation by sequencing copy number data where available. Missing sequencing copy number data were imputed by 5 independent realizations of Bernoulli random variables with probability estimated by the mean of the complete cases. All categorical data were encoded as 0 or 1 and then re-centered by their mean. Non-negative continuous variables were log-transformed and logit-transformed, re-centered and rescaled by the range between the 0.05 and 0.95 quantiles to assert a similar magnitude as dichotomous covariates. For proportional data the logit transformation was used rather than the logarithm.

For the 595 patients with available outcome data, leukemia-free survival was the end-point, and log-rank tests were used for univariate hypothesis tests. Survival analyses were performed with the Kaplan-Meier method. OS was defined as the time (in months) between the date of diagnosis and the date of death (for patients who deceased) or last follow-up (for censored patients). Leukemia-free survival (LFS) was defined as the time (in months) between the date of sample (median difference between date of sample and date of diagnosis: 2 weeks) and the date of leukemic transformation, for patients who progressed to AML or last follow-up. Event-free survival (EFS) was defined as the time in months between the date of sampling and the date of first event (death or leukemic transformation for cases having at least 1 of these events) or last follow-up. Comparison between survival curves was carried out by Cox proportional hazards model. For multivariate survival analyses, missing data were estimated by multiple imputation and Cox proportional hazards models were built from three sets of predictor variables using stability selection [16,17]. To measure predictive accuracy under each scenario, we used a five-fold cross-validation scheme in which we randomly split the patient cohort into fifths and made outcome predictions for each fifth from models built on the remaining four fifths.

Variable selection

Variables were selected using the stability selection procedure [17] in a Cox proportional hazards regression context using the glmnet R package [16]. The endpoint for event-free survival regression was the earlier of AML

transformation or death of the patient or if received the date the patient received disease-modifying treatment (i.e. bone marrow transplant), each measured from the date of sample. Mutations per patient per gene were reduced to binary variables, separately for 'known oncogenic' or 'possible oncogenic' variants, with 1 referring to the mutated status and 0 denoting not mutated or a variant of unknown significance.

For the three scenarios considered, we identified the following sets of possible predictor variables:

- *IPSS:* Each IPSS category and the associated variables.
- Standard variables: Age; exp(age); sex; white cell count; platelet count; hemoglobin levels; bone marrow blast count; WHO classification; ring sideroblast percentage; IPSS score; cytogenetic abnormalities.
- *Standard and genetic variables:* Age; exp(age); sex; white cell count; platelet count; hemoglobin level; mutations in each of the 41 MDS genes; recurrent copy number changes identified by sequencing; interaction terms for all pairs of genetic and copy number variables; number of oncogenic point mutations per patient.

Center and date of sample were included in each set for stratification purposes.

Stability selection was performed with weakness parameter 0.5 and 1000 bootstrap samples, each drawn randomly from one of the three realizations of the imputed data. The region for variable selection was determined such that added binary noise variables at frequencies 1%, 5%, 10%, and 15% had a

selection probability smaller than 0.2. Variables were considered to be stable if their selection probability was greater than 0.6 (Supplementary Table 7). To assess the performance of different variable types, this procedure was repeated for different variable subsets to assess the performance of different variable types.

Model evaluation

The accuracy of Cox proportional hazards models based on stable variables was evaluated by five-fold cross-validation of the generalized R^2 value [18], as well as receiver-operating characteristic (ROC) curves and the area under the curve (AUC). ROC curves and AUC were each computed for event-free status after 24, 60 and 90 months with the nearest-neighbor smoothing algorithm with parameter span=$(\#data)^{-0.2}/4$ (using the survivalROC R package [19]). Data were split into five parts and cross-validated estimates were computed for each part, based on the risk predicted by Cox proportional hazards models trained on the remaining 4/5 of the data. Ridge penalization, equivalent to a Gaussian prior with mean zero and unit variance on the risk coefficients, was used for stabilizing the predictions. This procedure was repeated for all three imputations and the resulting model performance estimates were averaged over imputations and partitions. In addition to stable variable sets, models of lower and greater complexity were also evaluated by sequentially including terms with decreasing selection probability.

Prediction of clinical variables

We used generalized linear models provided in the glmnet R package to predict clinical variables using point mutation and cytogenetic data. We restricted the analysis to complete cases only and applied the transformations described above to clinical variables prior to regression. For each variable we used five-fold cross-validation to compute estimates of the coefficient of determination (R^2) as a function of the LASSO penalty parameter. We then plotted the cross-validated R^2 values against the terms included in a model trained on the entire data set with the given penalty, as depicted in Figure 6B-C.

References

1. Papaemmanuil E, Cazzola M, Boultwood J, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med*. 2011;365(15):1384-1395.
2. Papaemmanuil E, Cazzola M, Boultwood J, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med*. 2011;365(15):1384-1395.
3. Varela I, Tarpey P, Raine K, et al. Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature*. 2011.
4. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009.
5. Joober R. The 1000 Genomes Project: deep genomic sequencing waiting for deep psychiatric phenotyping. *J Psychiatry Neurosci*. 2011;36(3):147-149.
6. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. *Genome Res*. 2001;11(10):1725-1729.
7. Flicek P, Aken BL, Beal K, et al. Ensembl 2008. *Nucleic Acids Res*. 2008;36(Database issue):D707-714.
8. Andreasen C, Nielsen JB, Refsgaard L, et al. New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants. *Eur J Hum Genet*. 2013.
9. Forbes SA, Bindal N, Bamford S, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res*. 2011;39(Database issue):D945-950.
10. http://www.cancer.sanger.ac.uk.
11. Vardiman JW, Thiele J, Arber DA, et al. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood*. 2009;114(5):937-951.
12. Greenberg P, Cox C, LeBeau MM, et al. International scoring system for evaluating prognosis in myelodysplastic syndromes. *Blood*. 1997;89(6):2079-2088.

13. Van Loo P, Nordgard SH, Lingjaerde OC, et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A*. 2010;107(39):16910-16915.
14. Nik-Zainal S, Van Loo P, Wedge DC, et al. The life history of 21 breast cancers. *Cell*. 2012;149(5):994-1007.
15. van Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate Imputation by Chained Equations in R. . *Journal of Statistical Software*. 2011;45:1-67.
16. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*. 2011;39(5):1-13.
17. Meinshausen N, Buhlmann P. Stability selection. *Journal of the Royal Statistical Society (series B)*. 2010;72:417-473.
18. Nagelkerke NJD. A note on a general definition of the coefficient of determination. *Biometrika*. 1991;78:691-692.
19. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000;56:337-344.

Supplementary Figure 1. Targeted gene sequencing for MDS. Barcoded DNA samples were captured using RNA baits mapping to 111genes, in pools of 16 samples at a time. Pools of 96 samples were submitted for massively parallel sequencing.

Supplementary Figure 2. Average coverage of targeted genes. Histogram showing the average number of reads covering each coding base per patient per gene.

Supplementary Figure 3. Variant allele fraction estimates. (A) Variant allele fraction estimates of variants in the dataset annotated as oncogenic or possibly oncogenic. (B) Variant allele estimates of confirmed somatic variants identified by exome sequencing or gene resequencing using genomic DNA in the published literature. (C) Variant allele fraction estimates of variants identified by the present study with known rs IDs in population databases and documented minor allele frequencies. (D) Distribution of variants in study annotated to be of unknown signficiance.

Supplementary Figure 4. Frequency of driver mutations identified in the sequencing screen. (A) Barplots in patients with CMML (B) MDS-MPN including RARS-T and MDS-U (C) MDS-AML

Supplementary Figure 5. Copy number changes identified by targeted gene sequencing. (A) Copy number profile from a patient with deletions of chromosome 4q, 5, 7, 13 and 17p. (B), the upper panel depicts the normalized sequencing yields per exon; the lower panel depicts the variant allele fraction for heterozygous SNPs, highlighting a copy neutral region of UPD on chromosomes 21 and 22. (C) Copy-

number profile from a previous unidentified patient with del 5q and isochromosome X.

Supplementary Figure 6. Conditional patterns of co-mutation for each driver gene in the study. Mutational patterns were drawn conditioned on the mutational status of a given gene. Blue barplots indicate genes mutated amongst the samples with mutations in the gene (indicated by an asterix). In orange barplots distribution of mutations in the patients that are wild-type for the gene. For all figures, the same order of oncogenic genes was used and when conditioned on itself, the plot indicates it as an asterisk.

Supplementary Figure 7. Pairwise precedences established from patterns of clonal and subclonal driver mutations. Each square in the matrix is colored by the fraction of patients in which gene 1 (y axis) occurs before gene 2 (x axis), and sized by the number of patients in whom an informative precedence was observed.

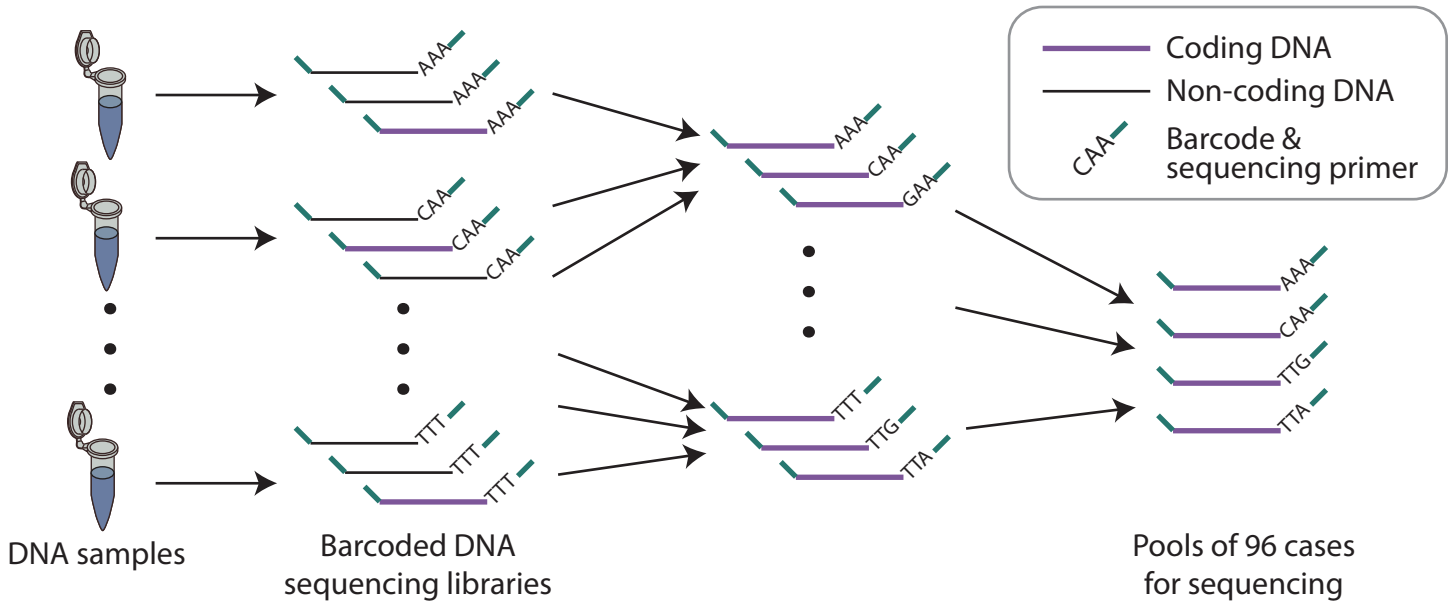
Supplementary Figure 8. Leukemia-free survival curves for 24 genes mutated in more than 5 patients. The p values refer to log-rank tests for an overall difference across groups.

Supplementary Figure 9. A. Leukemia free survival curves in patients grouped by number of oncogenic mutations identified in the gene sequencing data (excluding cytogenetic abnormalities). B Representation of gene mutations across the patient

categories with 1, 2, 3 or 4 oncogenic mutations in patients classified as IPSS LOW (B) , IPSS Int-1 (C), IPSS Int-2 (D).

Supplementary Figure 10. Comparison of mutation patterns between our screen and the published literature (as recorded in the COSMIC website). Mutations identified in the screen reported here are shown above the gene model, and those from the published literature are shown below for *EZH2* and *CBL*.

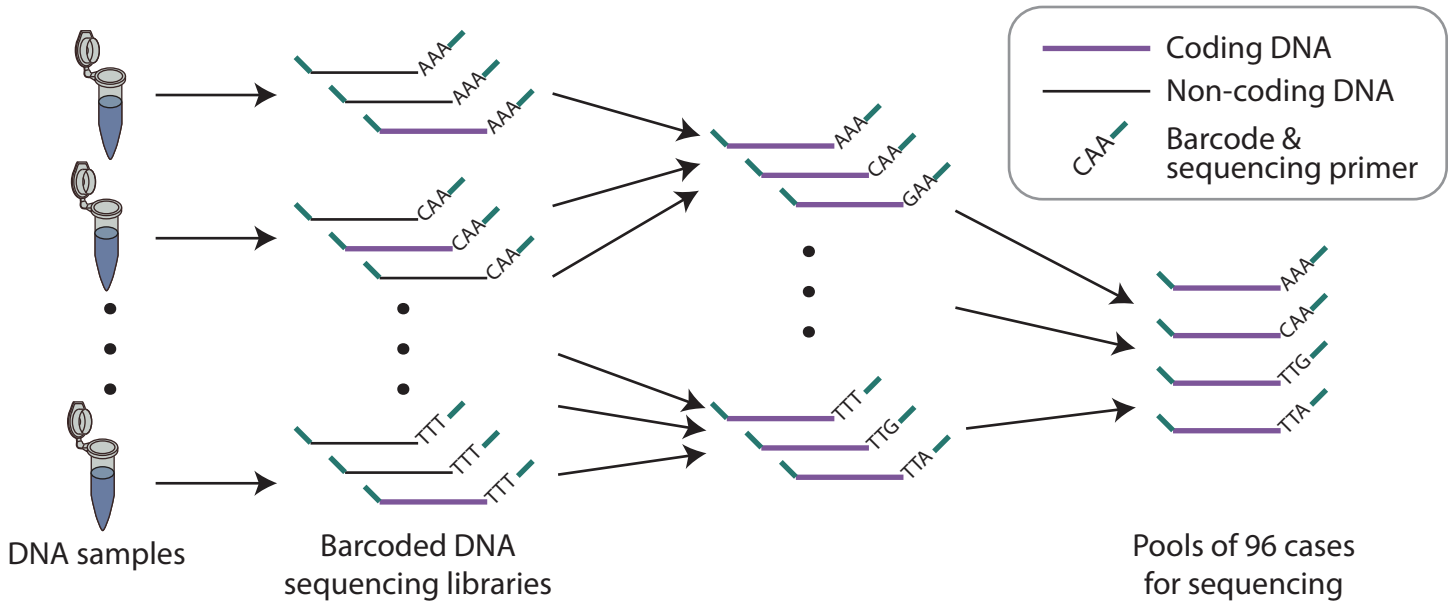
a



16 individually barcoded samples in each bait capture pool

Supplementary figure 1

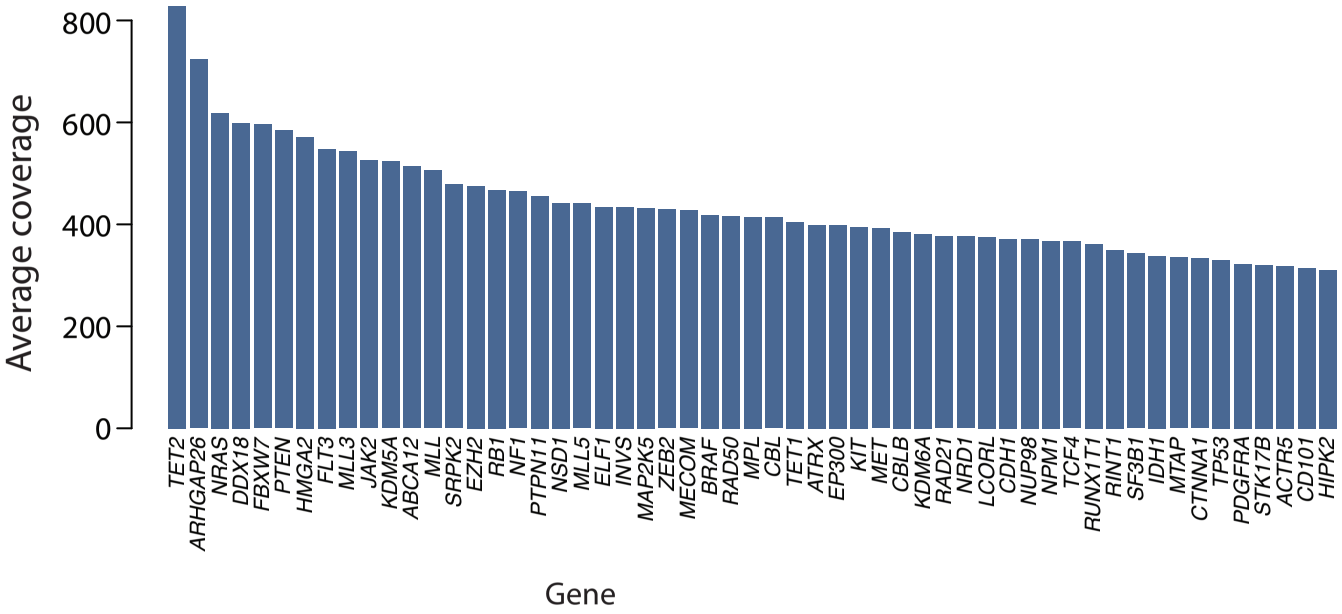
a

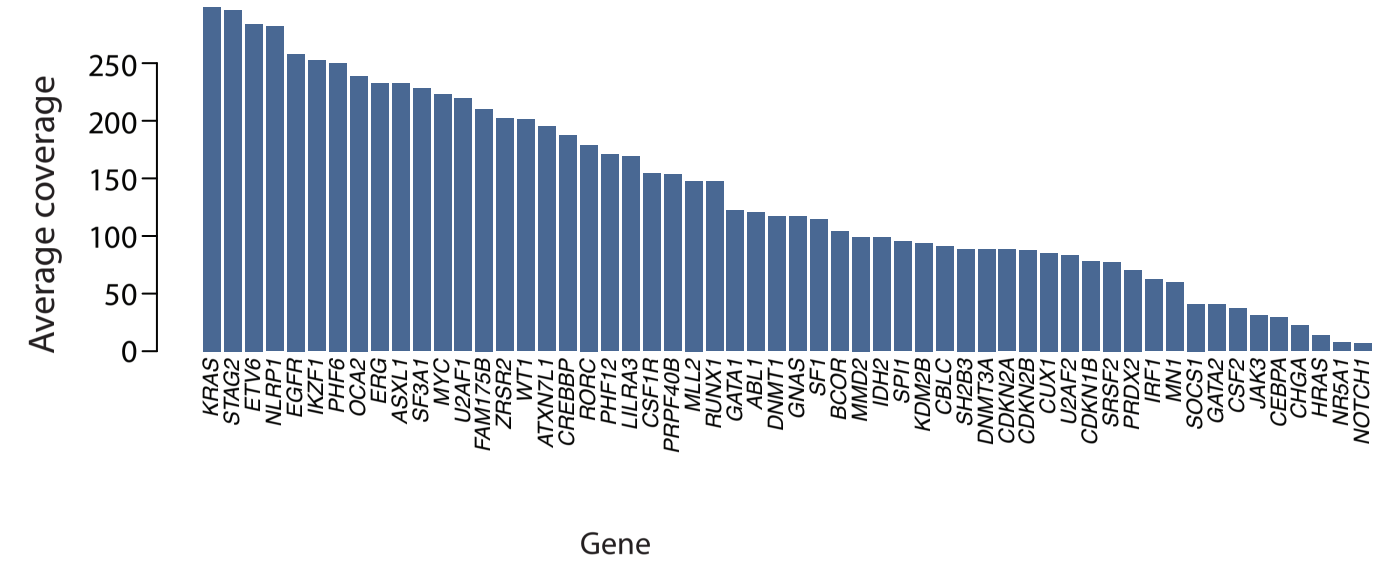


16 individually barcoded samples in each bait capture pool

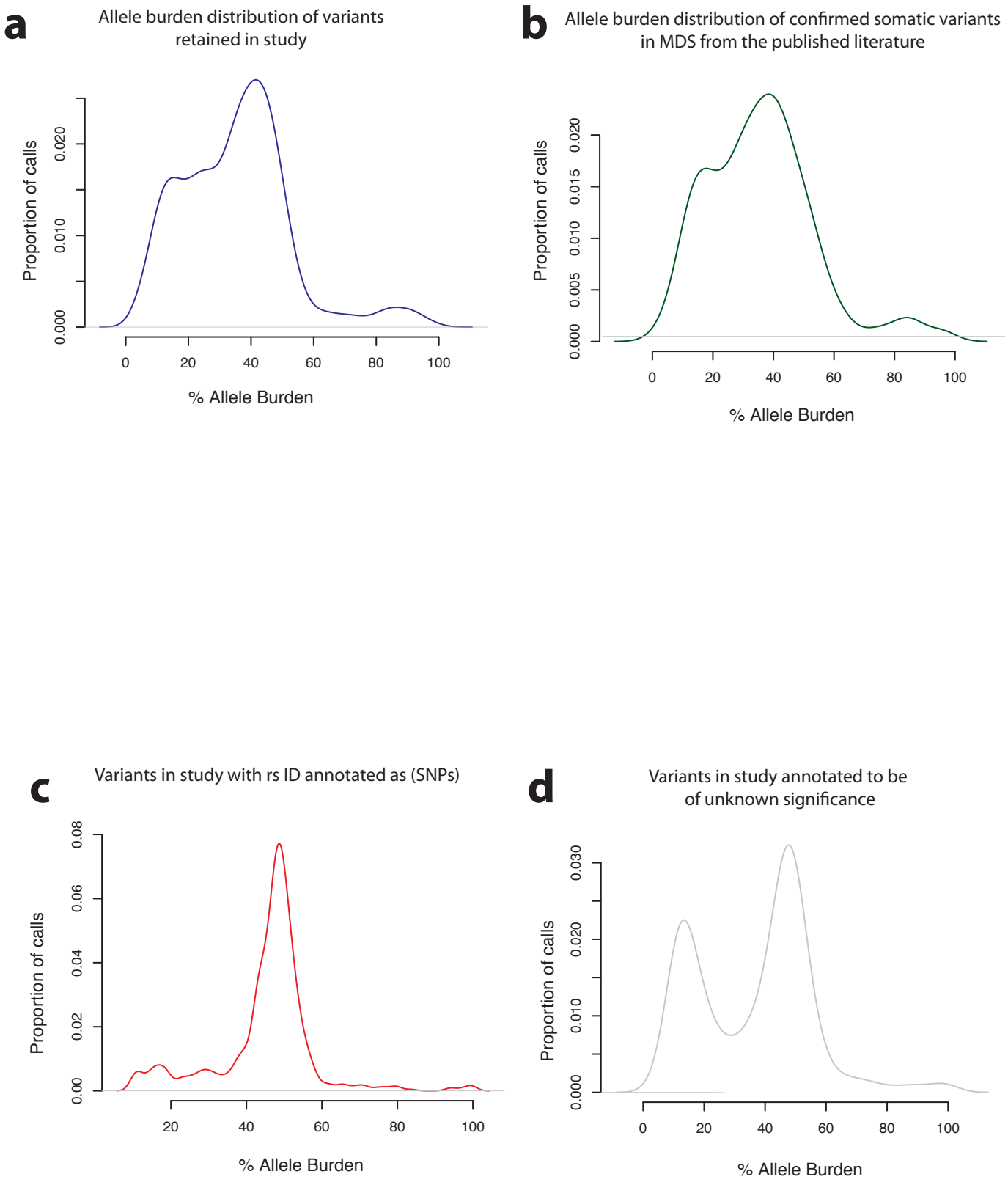
Supplementary figure 1

a

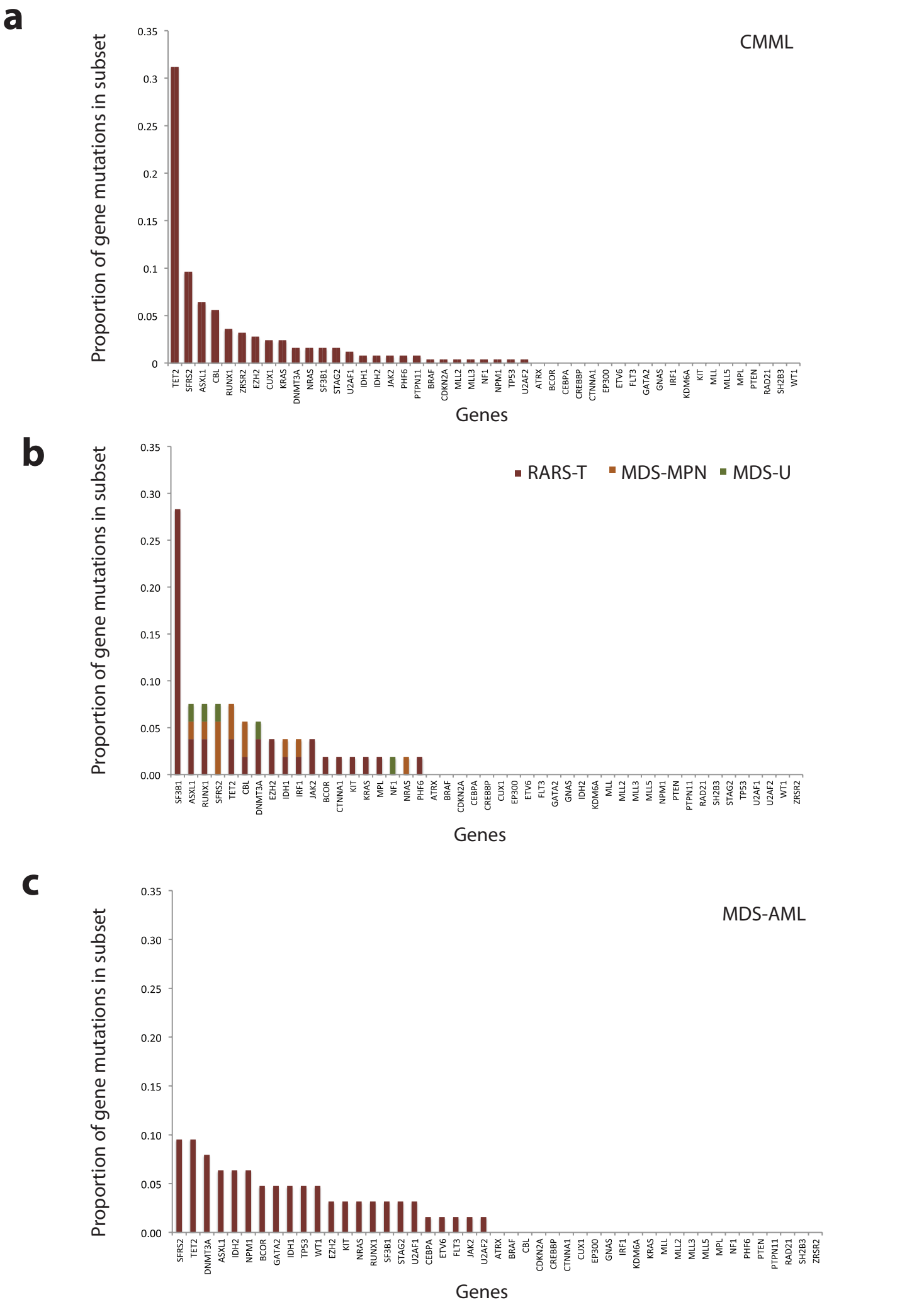




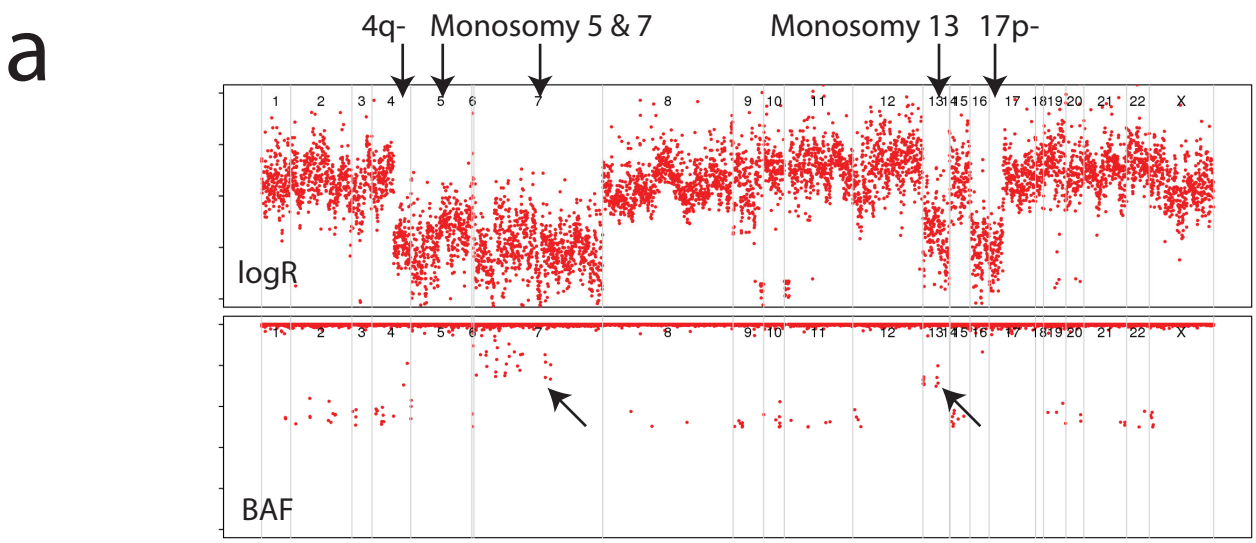
Supplementary figure 2

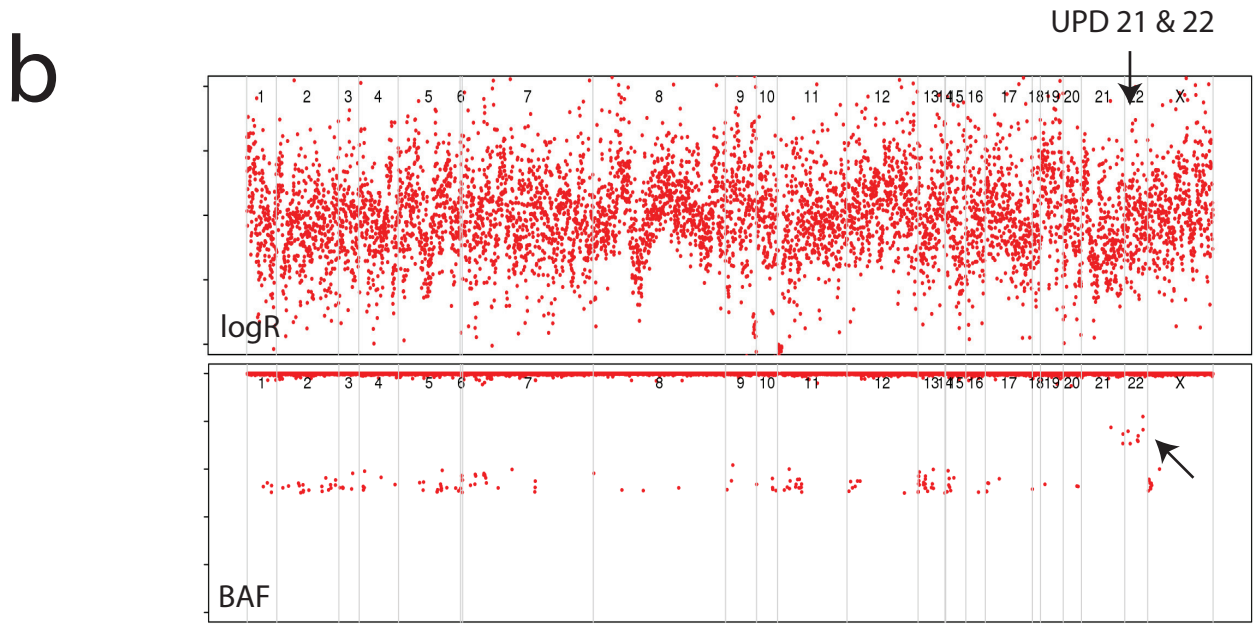


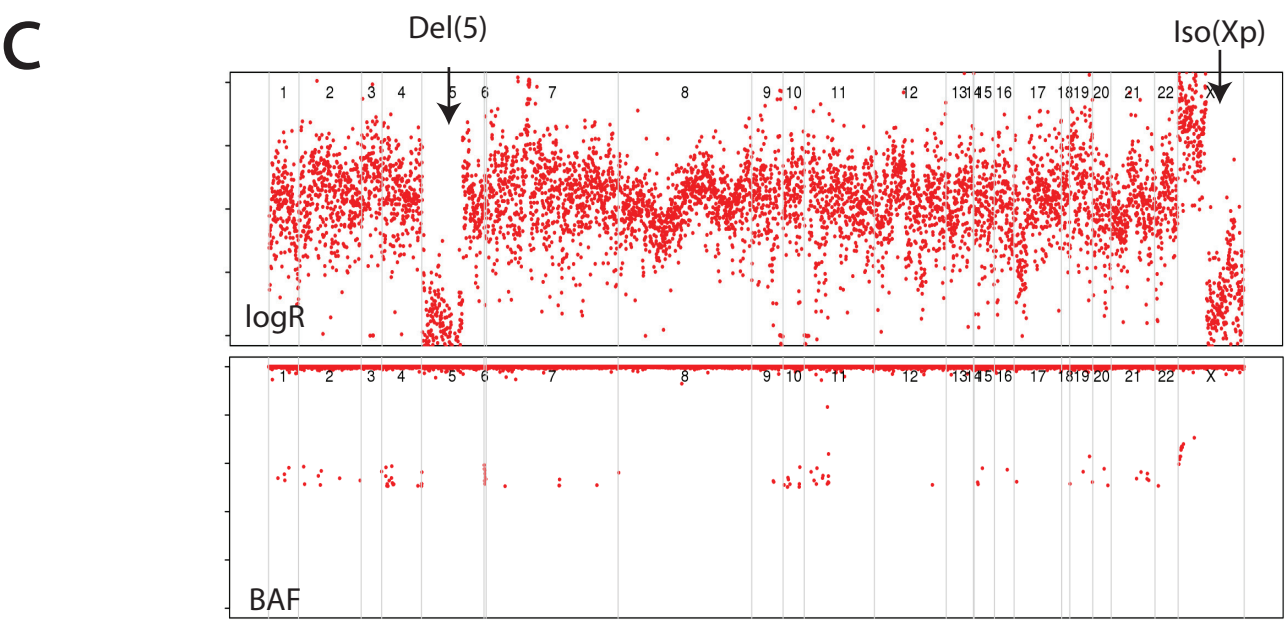
Supplementary figure 3



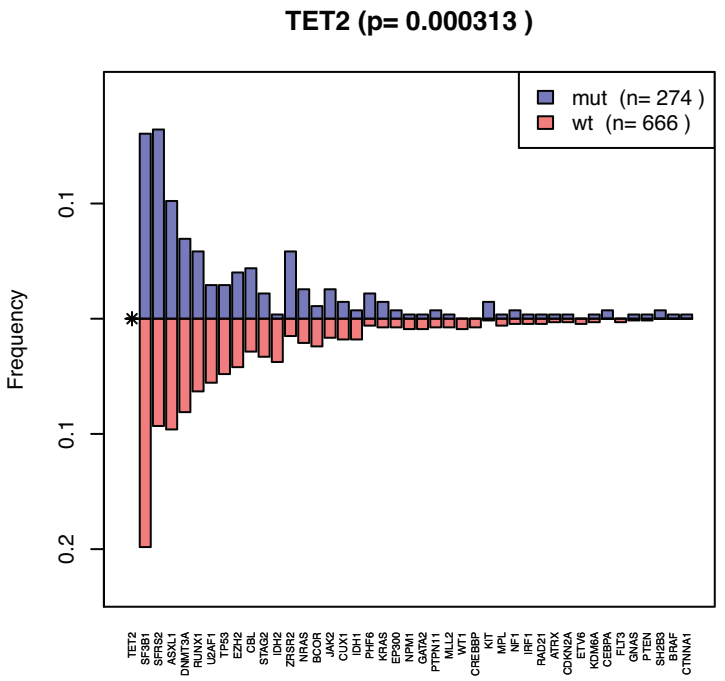
Supplementary figure 4

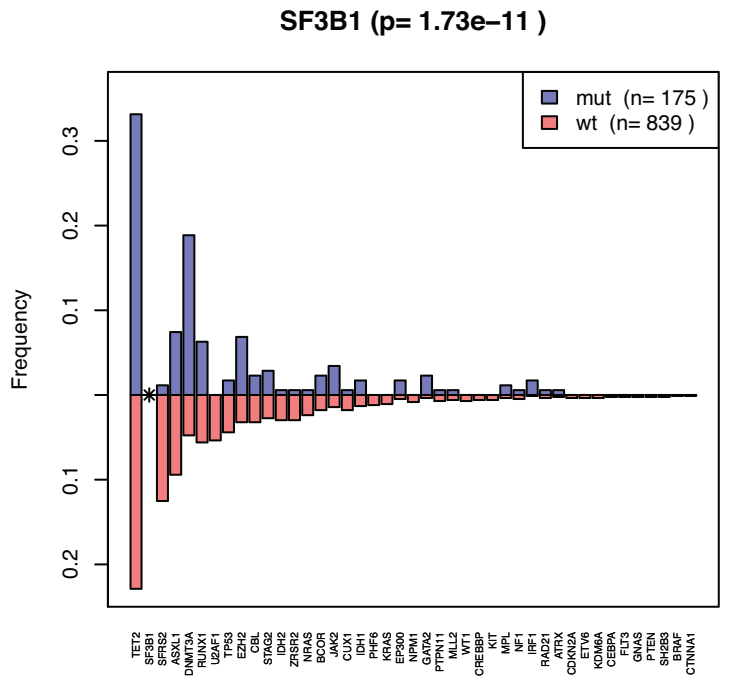


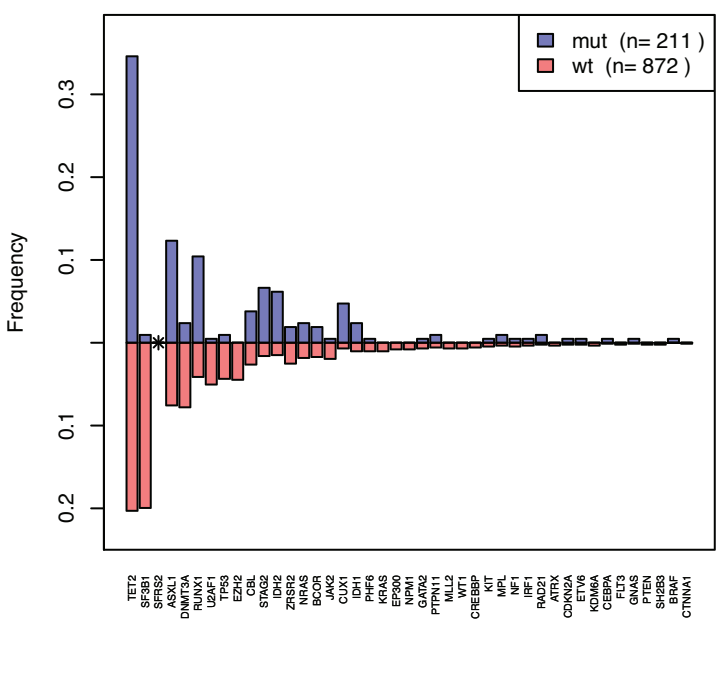


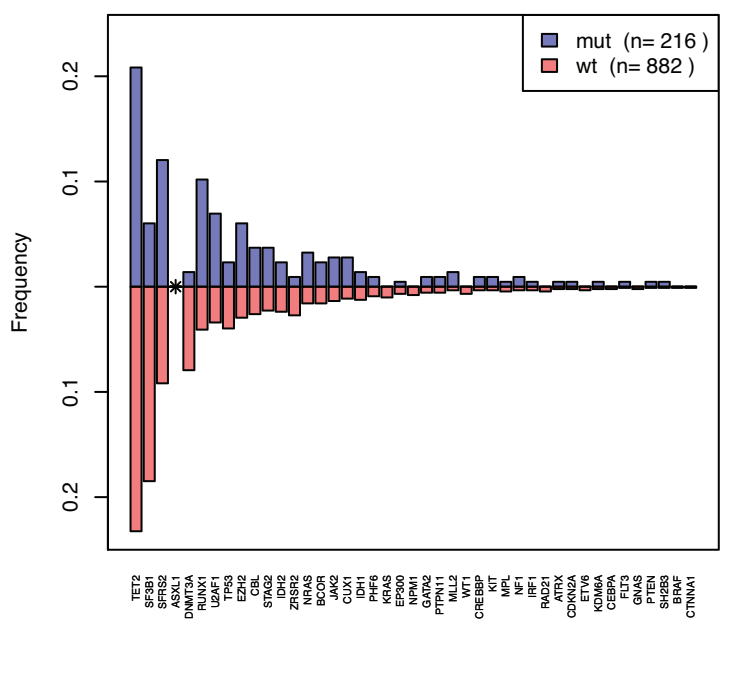


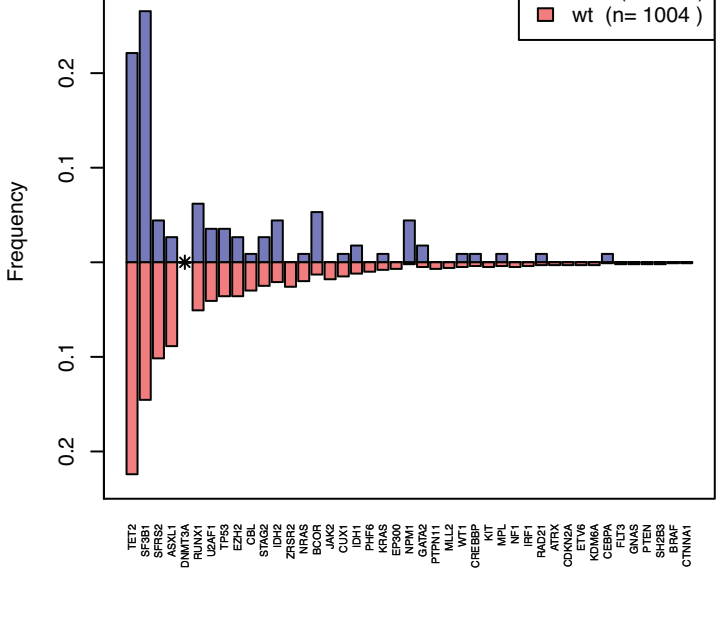
Supplementary figure 5

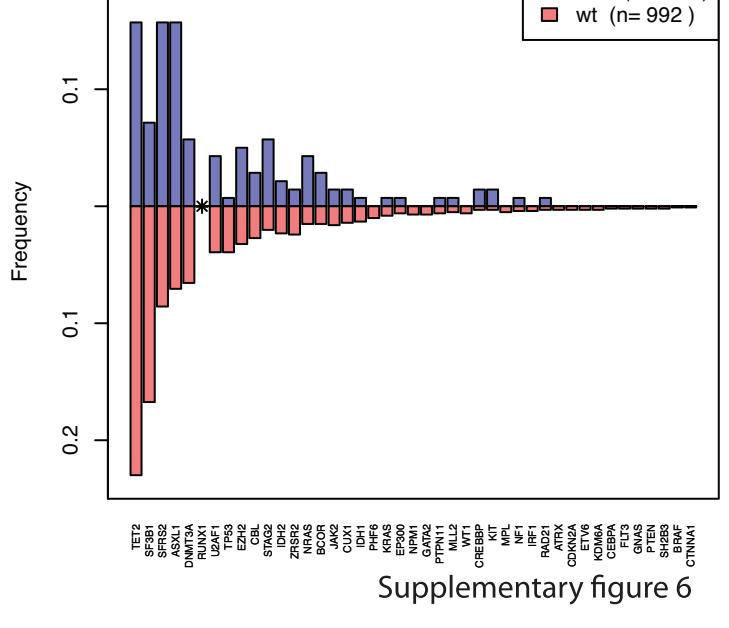




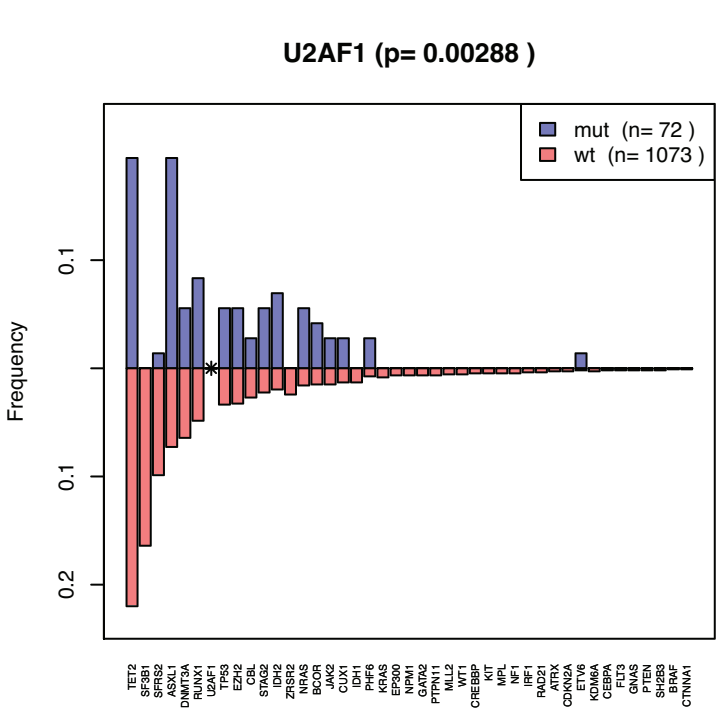


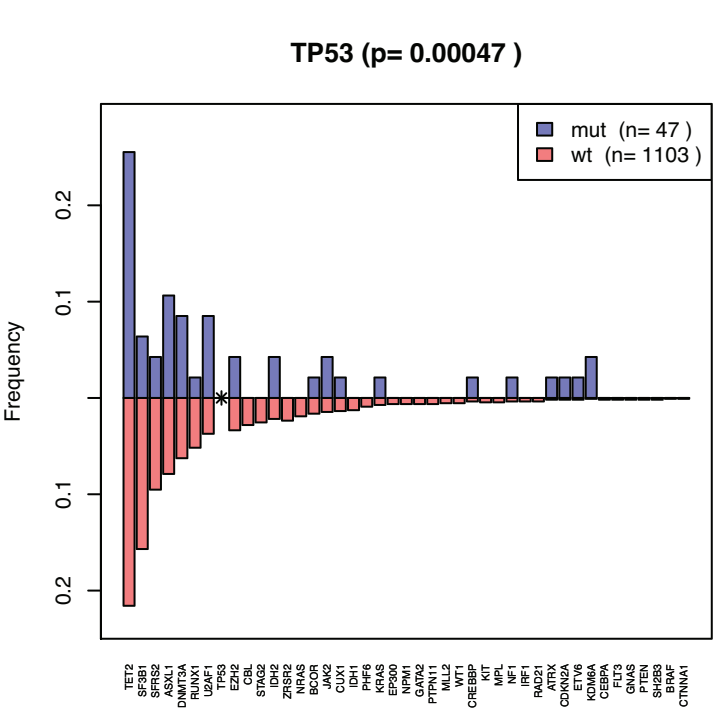


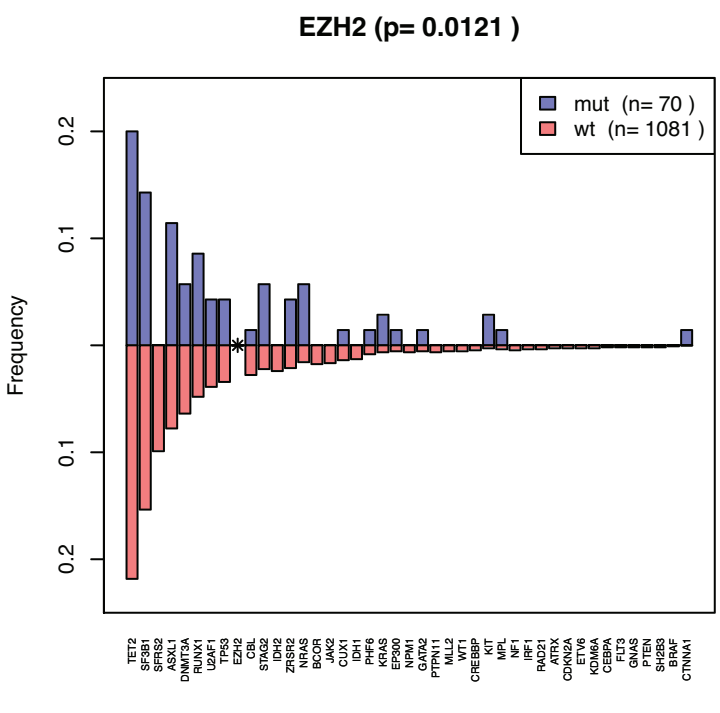


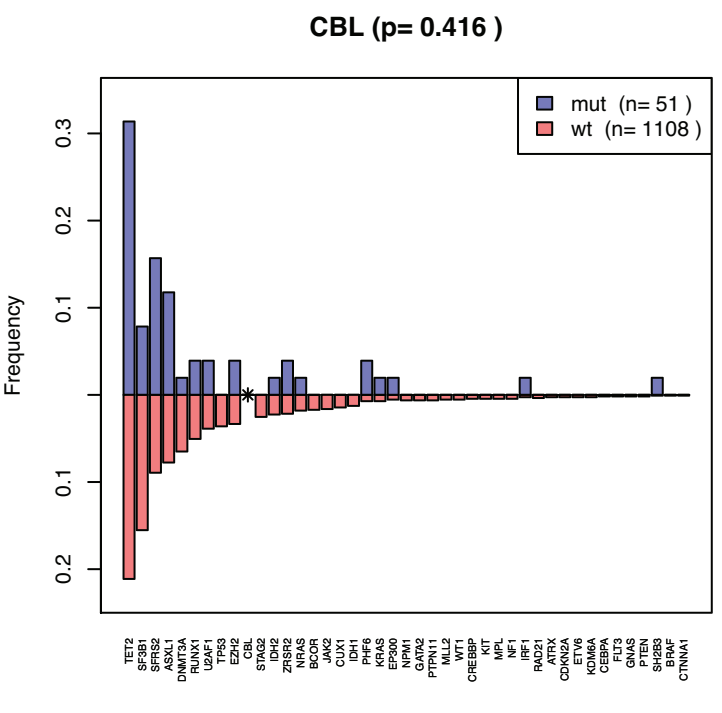


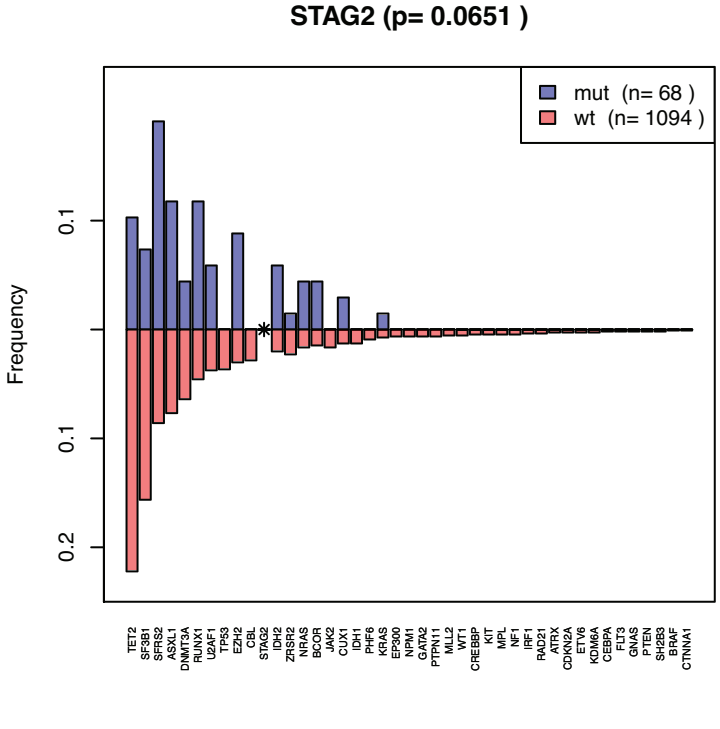
Supplementary figure 6

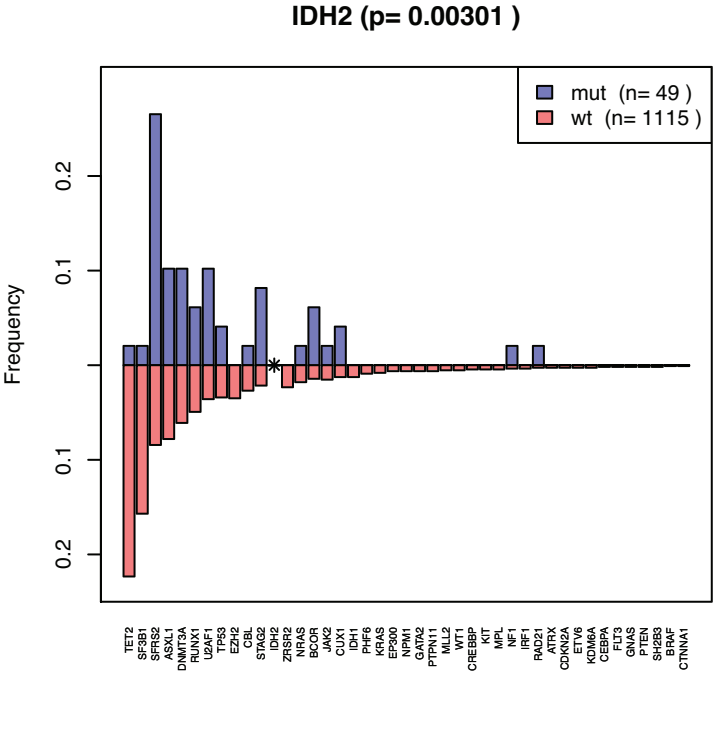

U2AF1 (p= 0.00288)
mut (n= 72)
wt (n= 1073)
Frequency
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

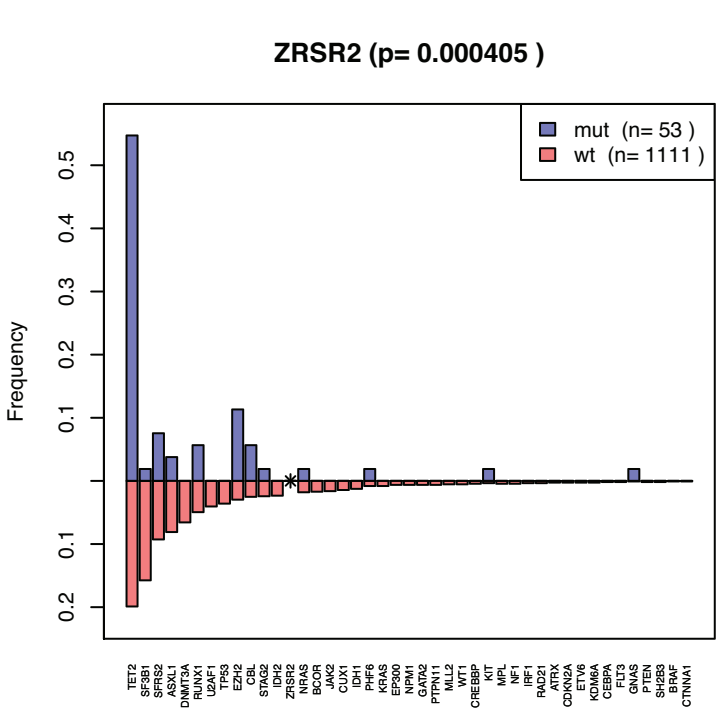

TP53 (p= 0.00047)
mut (n= 47)
wt (n= 1103)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

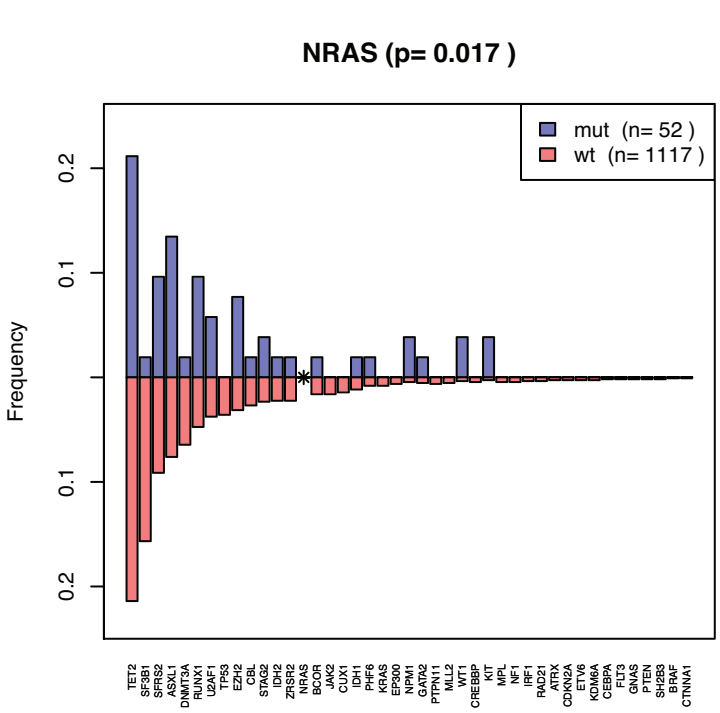

EZH2 (p= 0.0121)
mut (n= 70)
wt (n= 1081)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

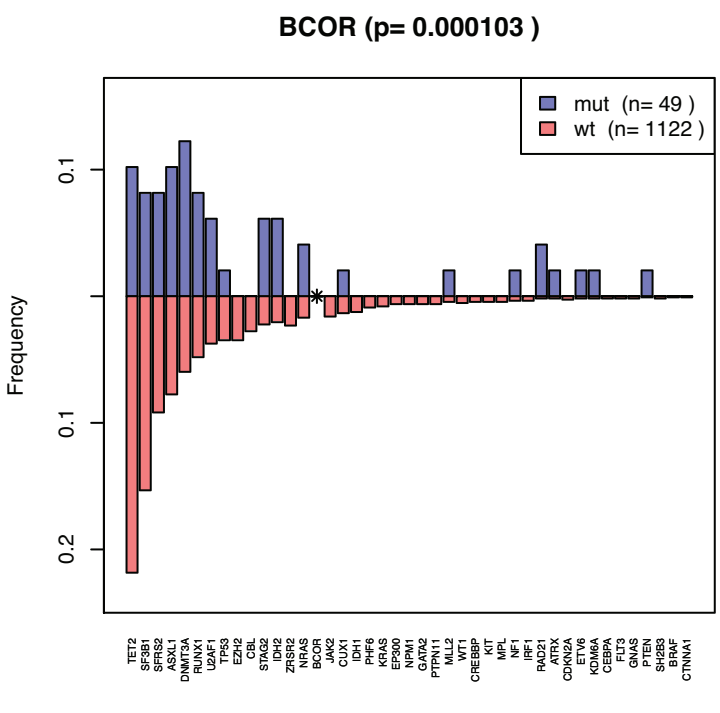

CBL (p= 0.416)
mut (n= 51)
wt (n= 1108)
Frequency
0.3
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

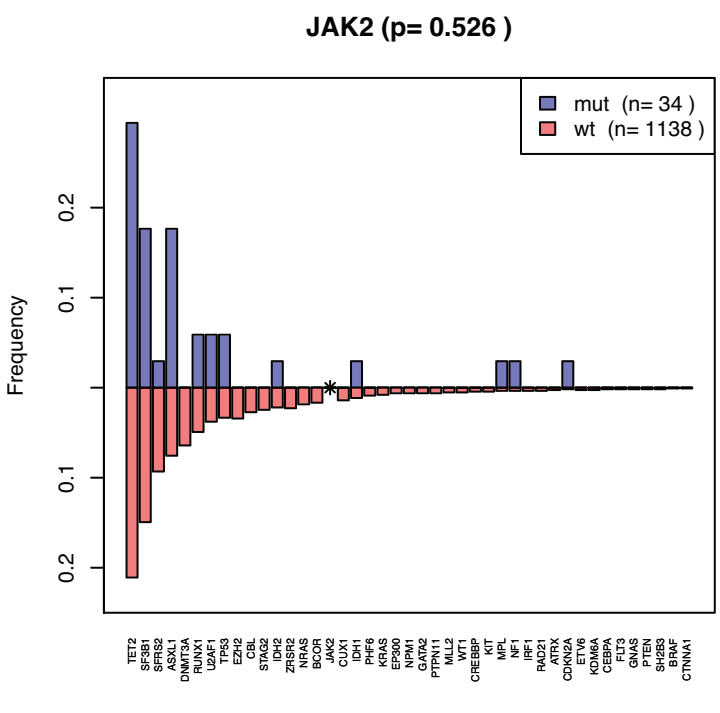

STAG2 (p= 0.0651)
mut (n= 68)
wt (n= 1094)
Frequency
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

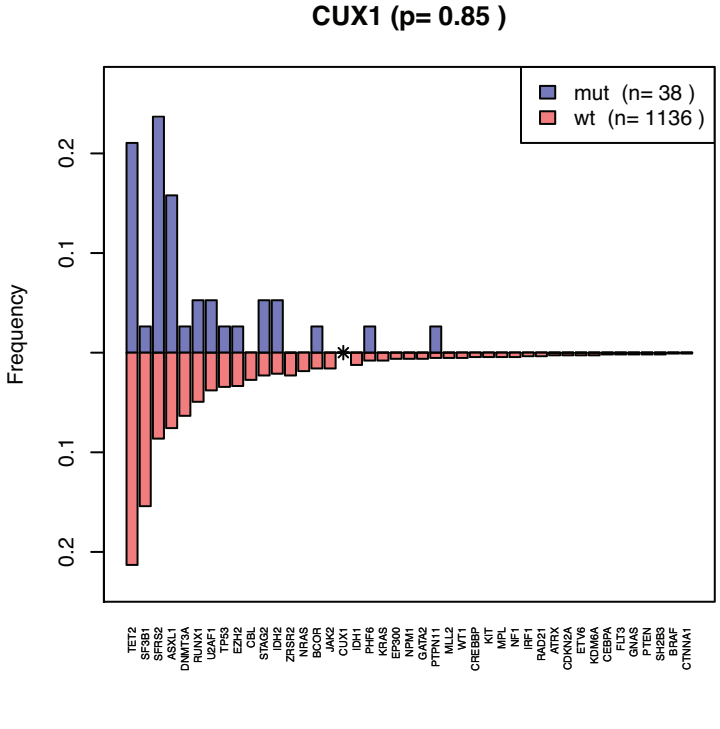

IDH2 (p= 0.00301)
mut (n= 49)
wt (n= 1115)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

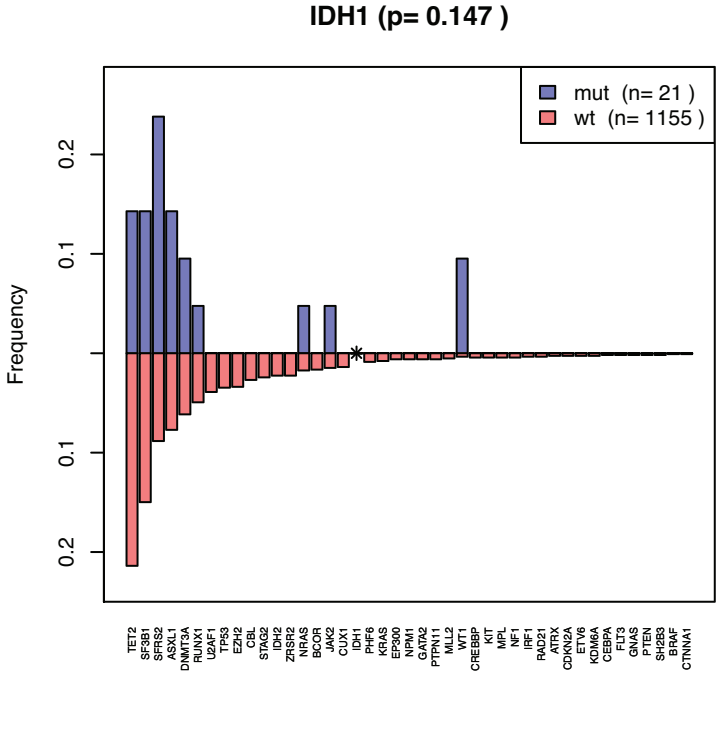

ZRSR2 (p= 0.000405)
mut (n= 53)
wt (n= 1111)
Frequency
0.5
0.4
0.3
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

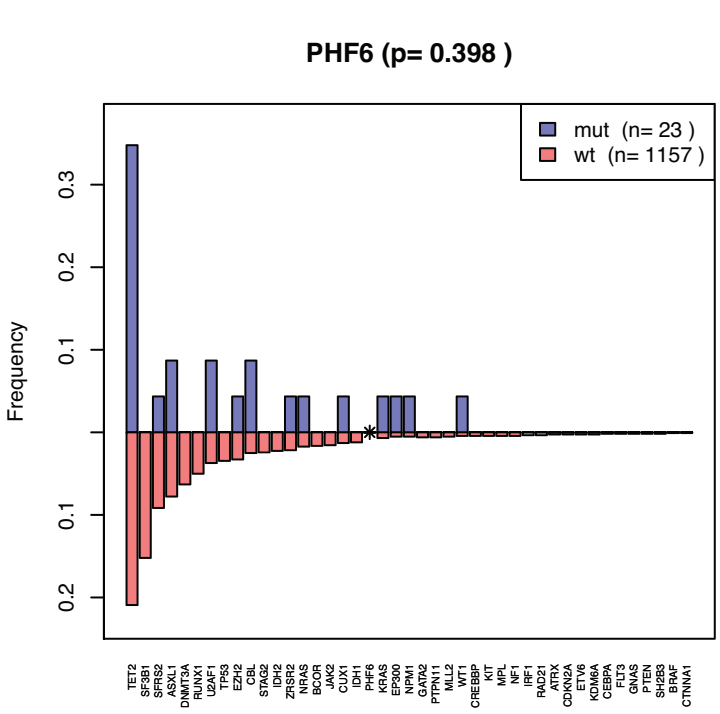

NRAS (p= 0.017)
mut (n= 52)
wt (n= 1117)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

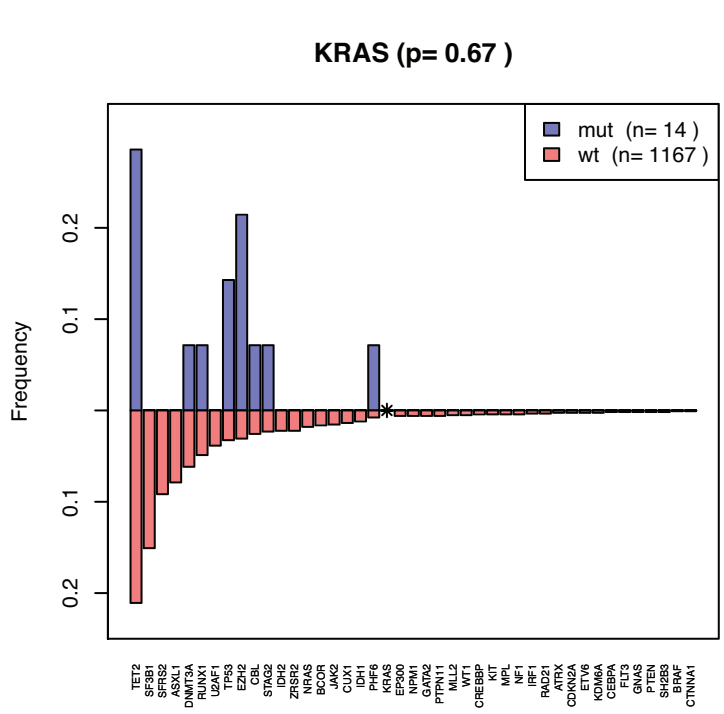

BCOR (p= 0.000103)
mut (n= 49)
wt (n= 1122)
Frequency
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

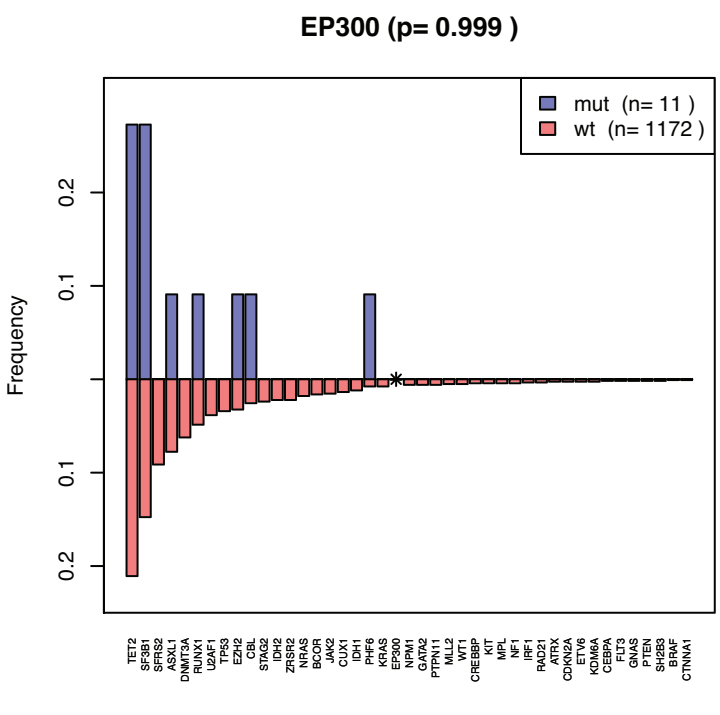

JAK2 (p= 0.526)
mut (n= 34)
wt (n= 1138)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

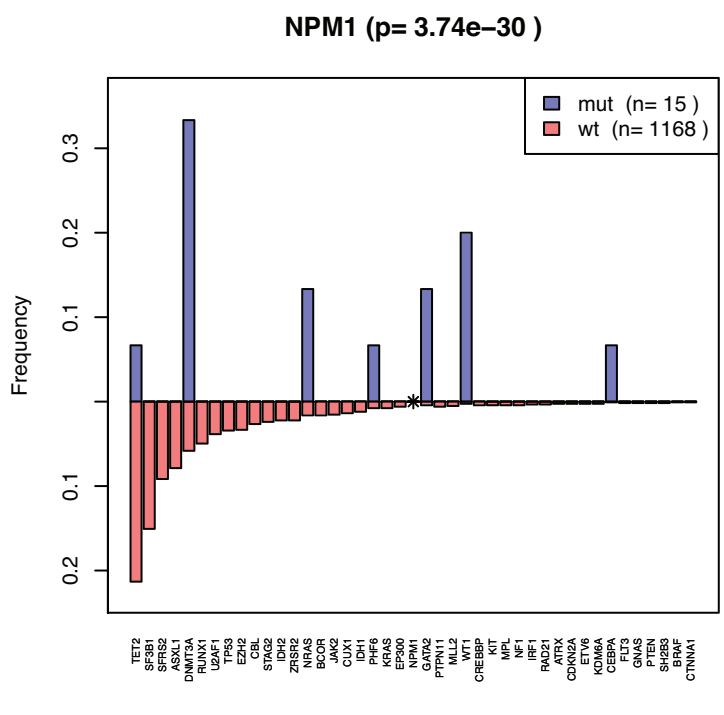

CUX1 (p= 0.85)
mut (n= 38)
wt (n= 1136)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

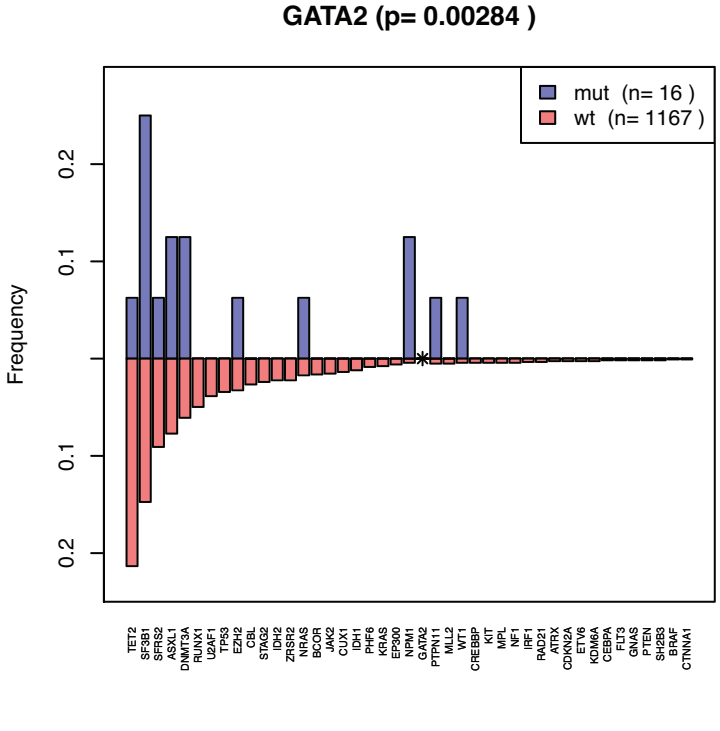

IDH1 (p= 0.147)
mut (n= 21)
wt (n= 1155)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

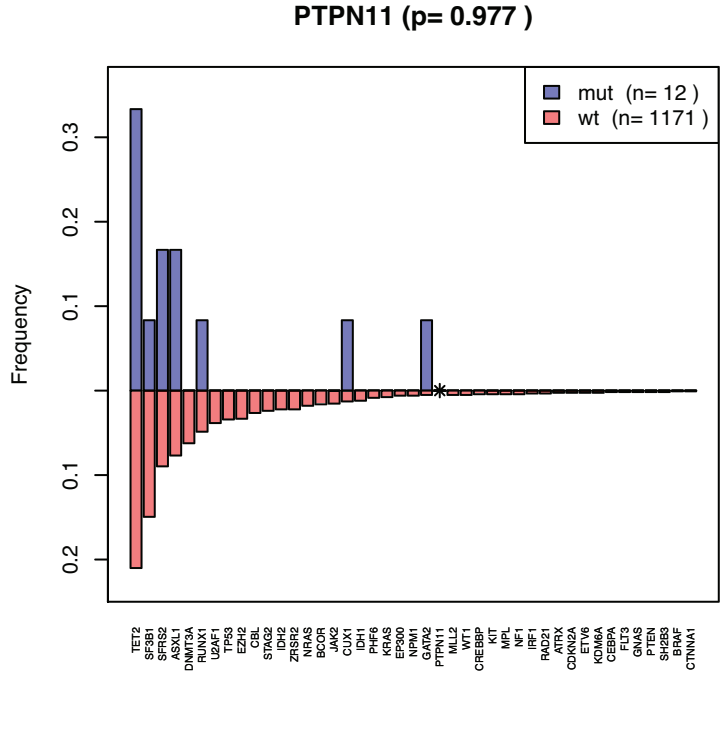

PHF6 (p= 0.398)
mut (n= 23)
wt (n= 1157)
Frequency
0.3
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

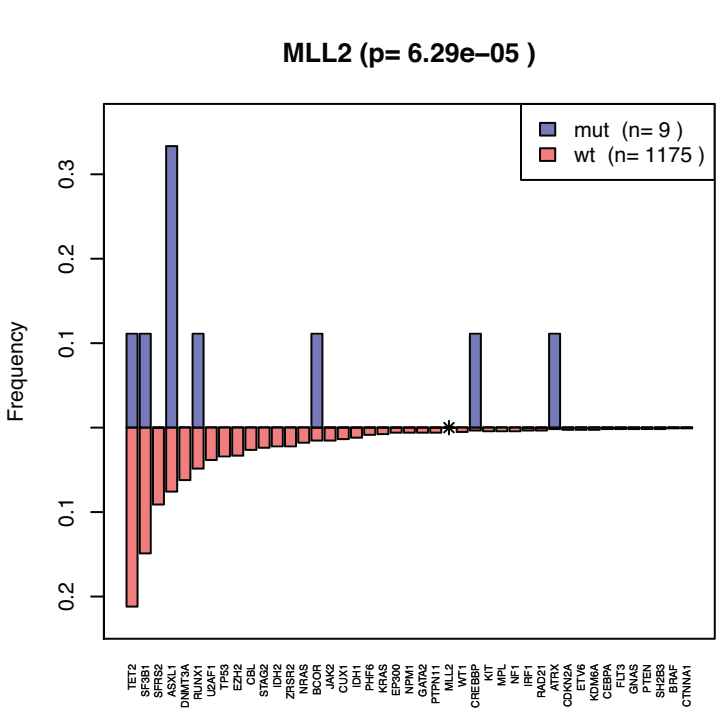

KRAS (p= 0.67)
mut (n= 14)
wt (n= 1167)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

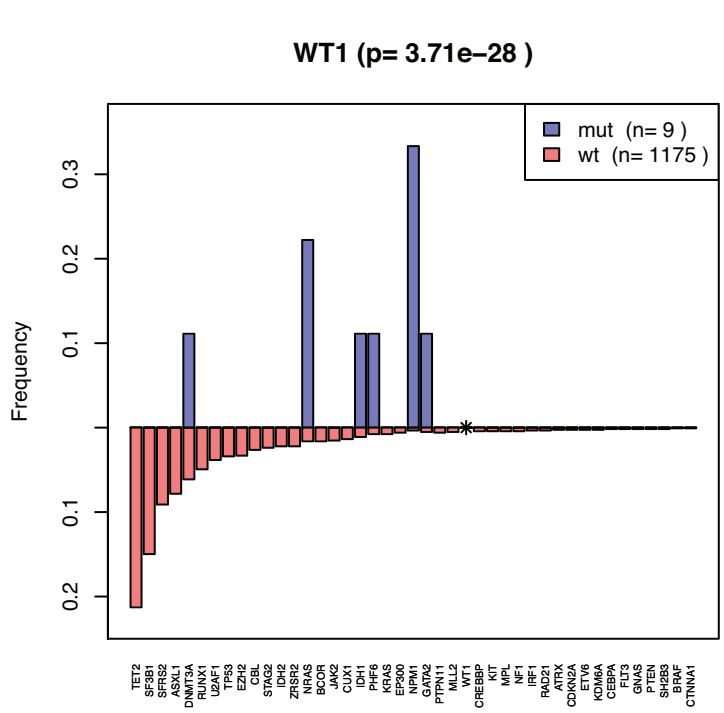

EP300 (p= 0.999)
mut (n= 11)
wt (n= 1172)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

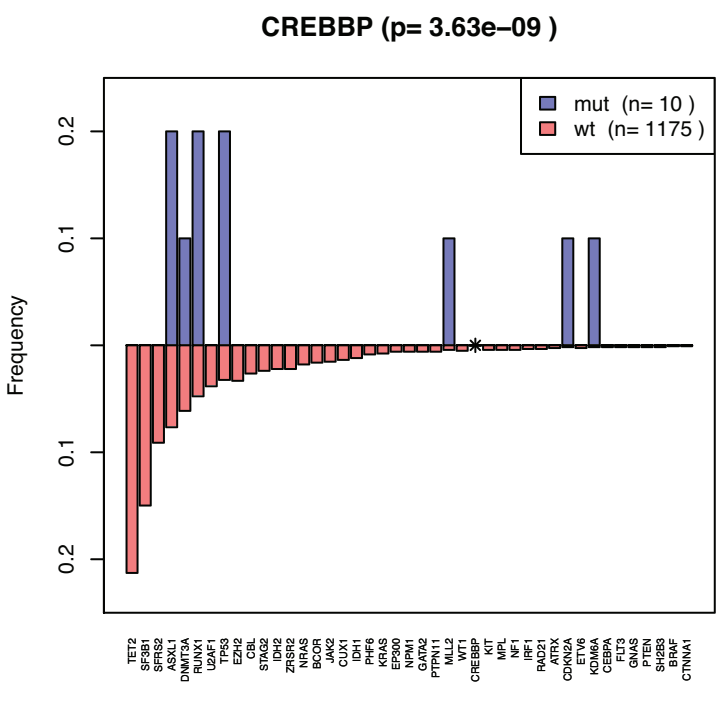

NPM1 (p= 3.74e−30)
mut (n= 15)
wt (n= 1168)
Frequency
0.3
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

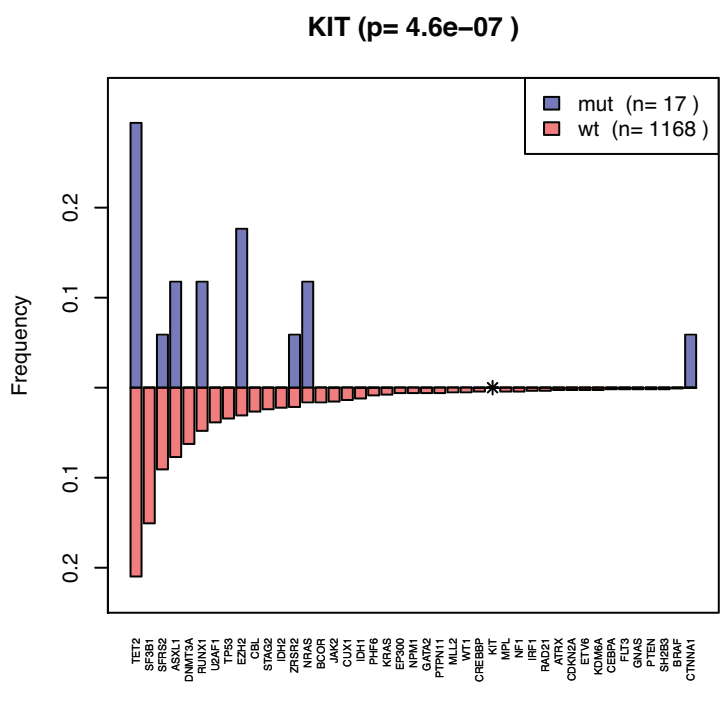

GATA2 (p= 0.00284)
mut (n= 16)
wt (n= 1167)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

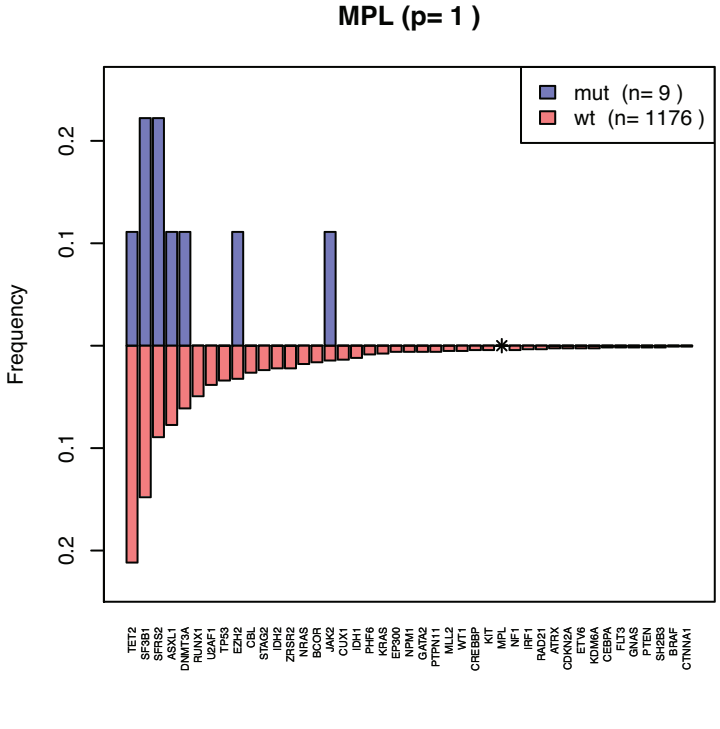

PTPN11 (p= 0.977)
mut (n= 12)
wt (n= 1171)
Frequency
0.3
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

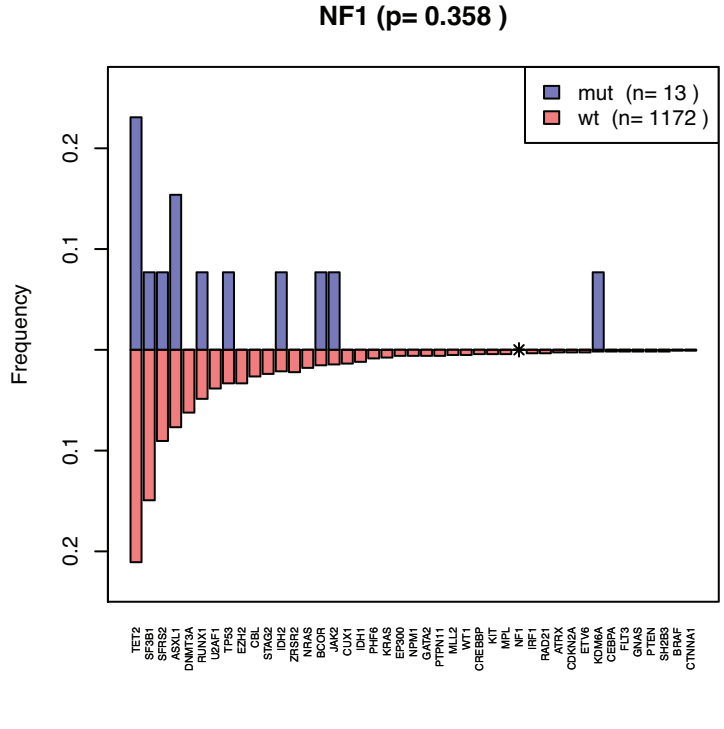

MLL2 (p= 6.29e−05)
mut (n= 9)
wt (n= 1175)
Frequency
0.3
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

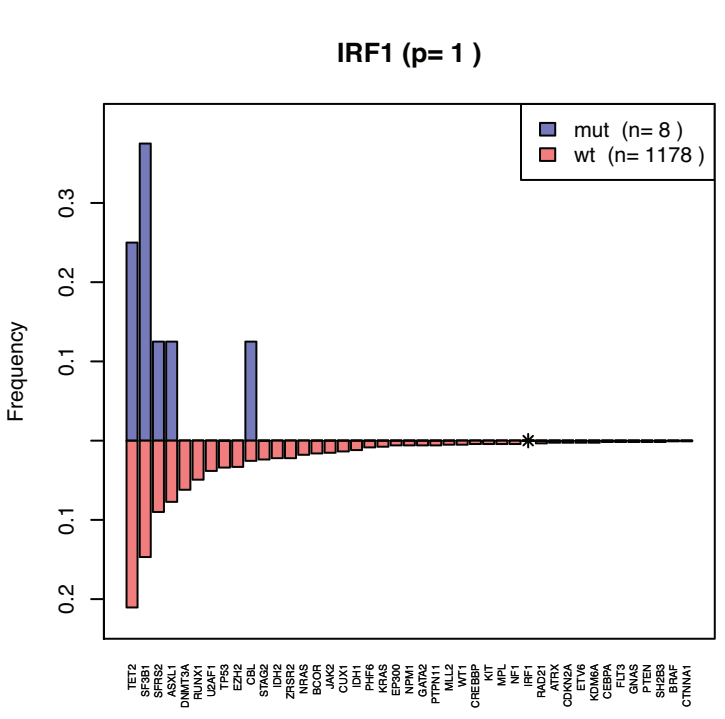

WT1 (p= 3.71e−28)
mut (n= 9)
wt (n= 1175)
Frequency
0.3
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

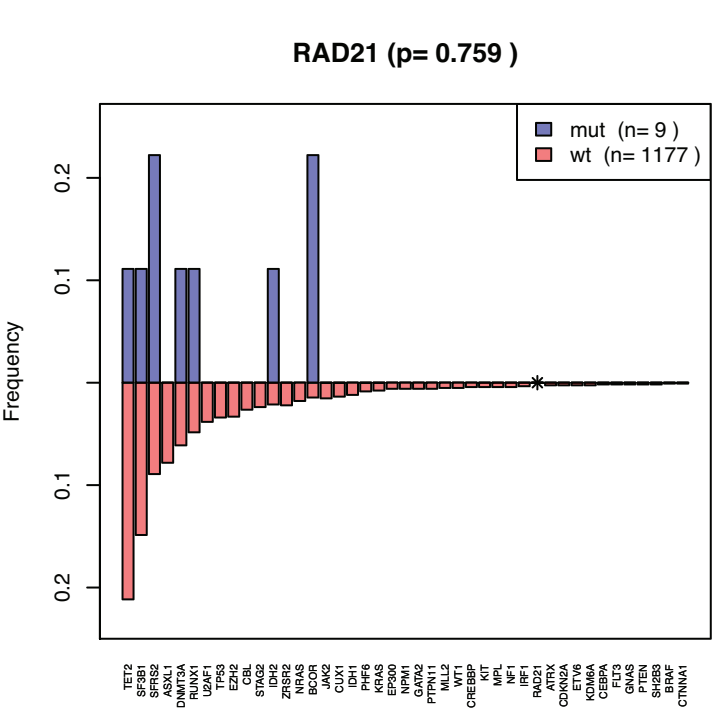

CREBBP (p= 3.63e−09)
mut (n= 10)
wt (n= 1175)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

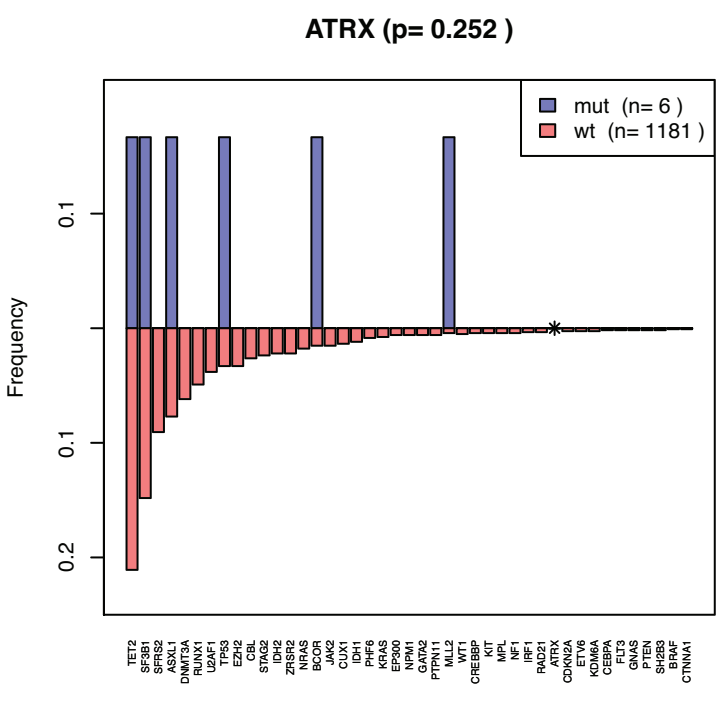

KIT (p= 4.6e−07)
mut (n= 17)
wt (n= 1168)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

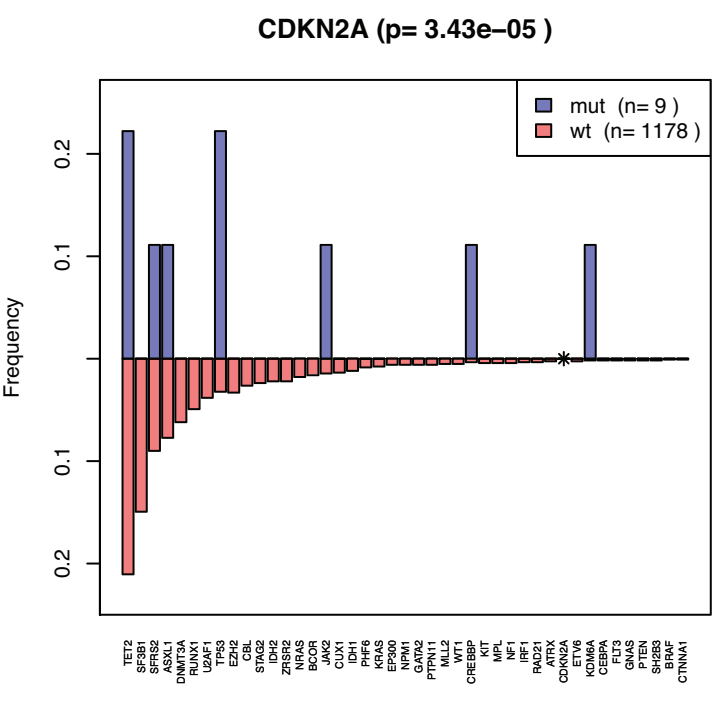

MPL (p= 1)
mut (n= 9)
wt (n= 1176)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

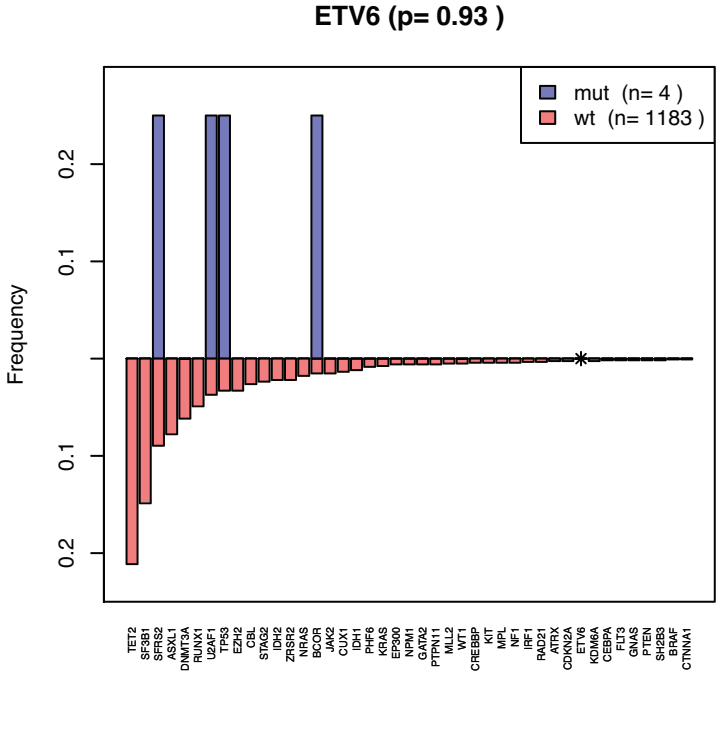

NF1 (p= 0.358)
mut (n= 13)
wt (n= 1172)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

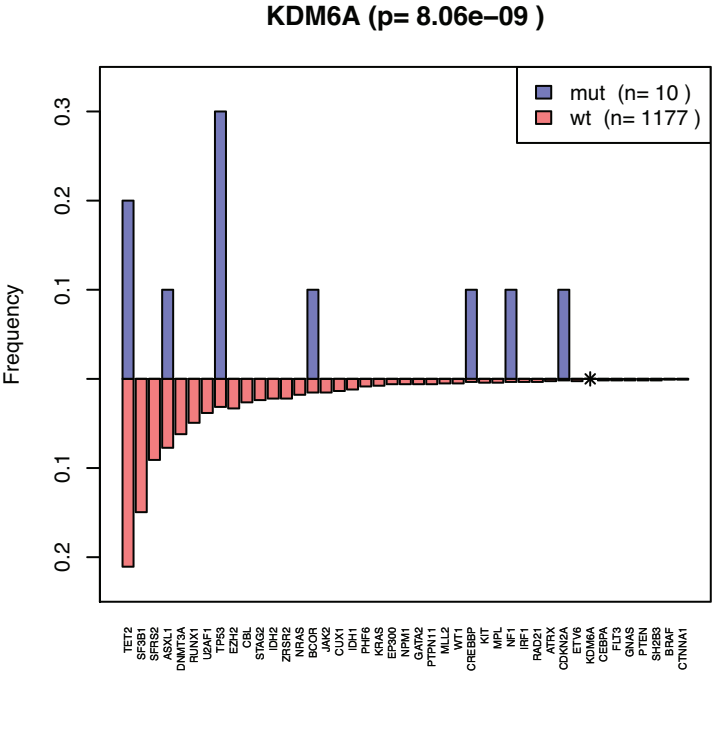

IRF1 (p= 1)
mut (n= 8)
wt (n= 1178)
Frequency
0.3
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

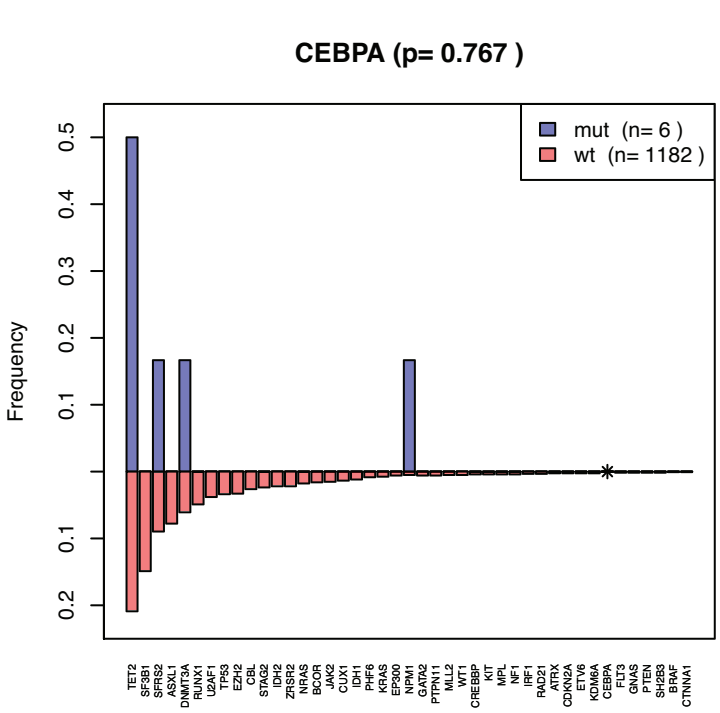

RAD21 (p= 0.759)
mut (n= 9)
wt (n= 1177)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

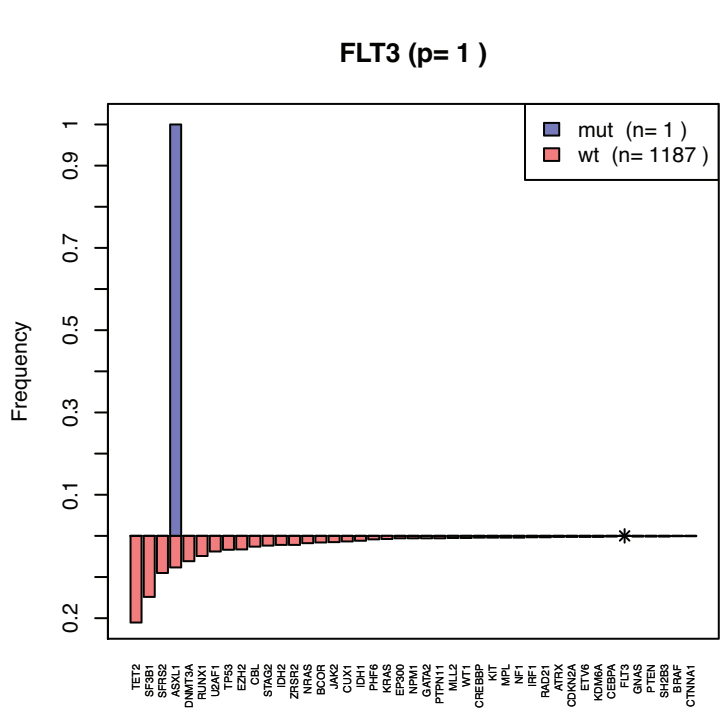

ATRX (p= 0.252)
mut (n= 6)
wt (n= 1181)
Frequency
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

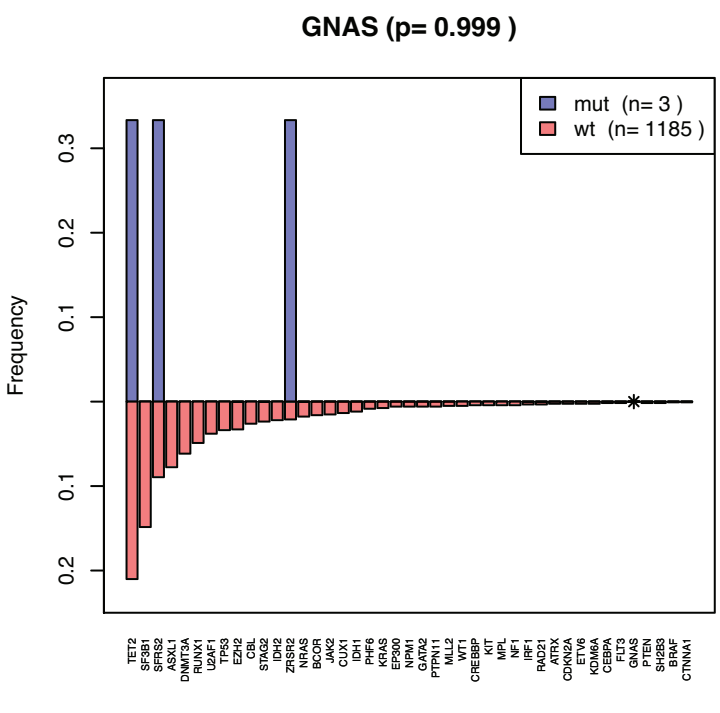

CDKN2A (p= 3.43e−05)
mut (n= 9)
wt (n= 1178)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

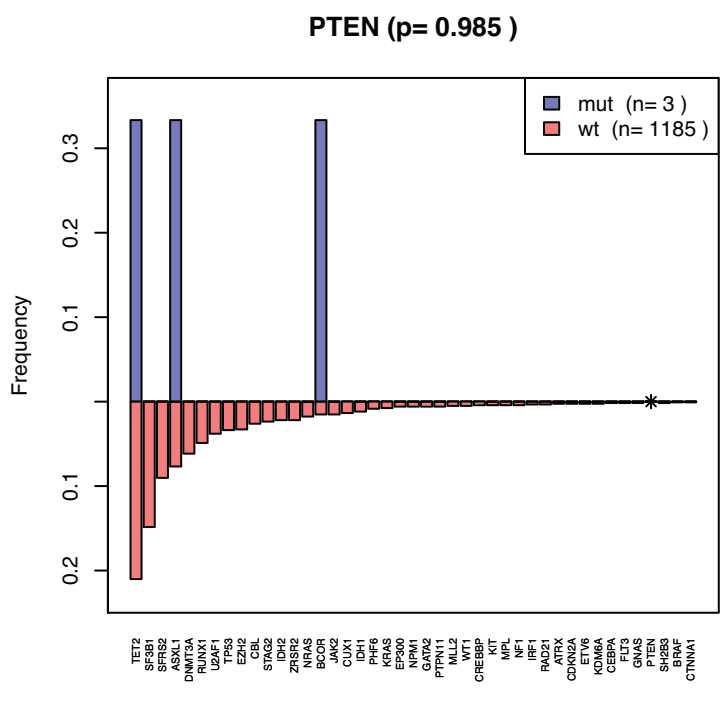

ETV6 (p= 0.93)
mut (n= 4)
wt (n= 1183)
Frequency
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

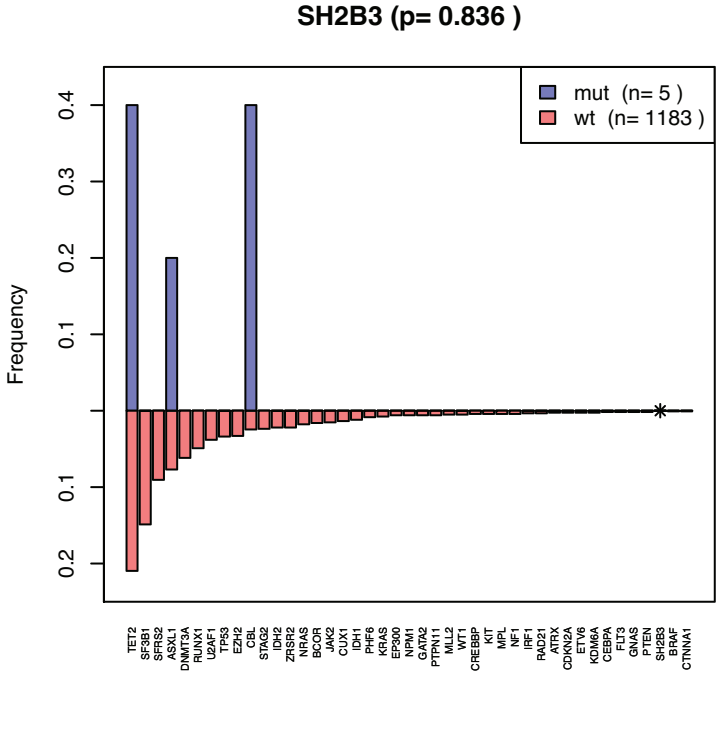

KDM6A (p= 8.06e−09)
mut (n= 10)
wt (n= 1177)
Frequency
0.3
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

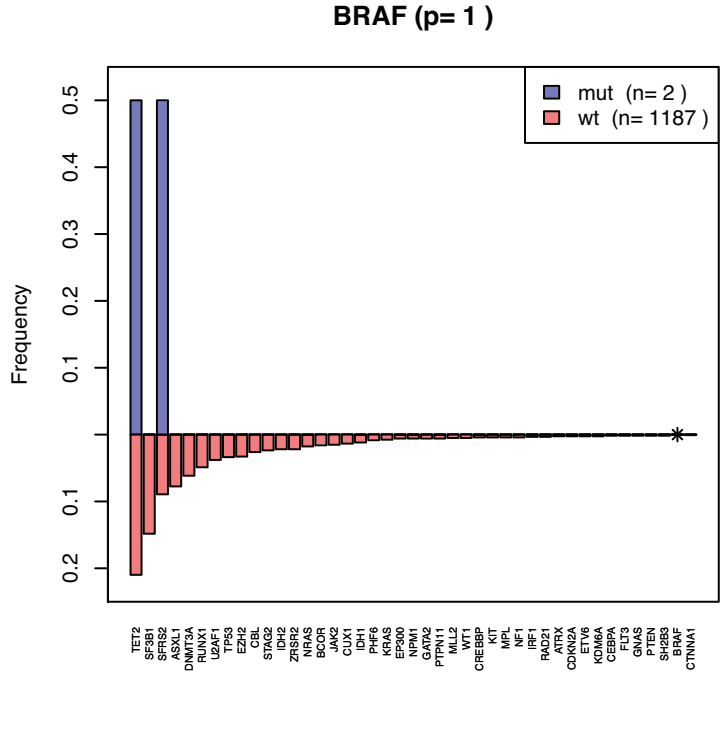

CEBPA (p= 0.767)
mut (n= 6)
wt (n= 1182)
Frequency
0.5
0.4
0.3
0.2
0.1
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

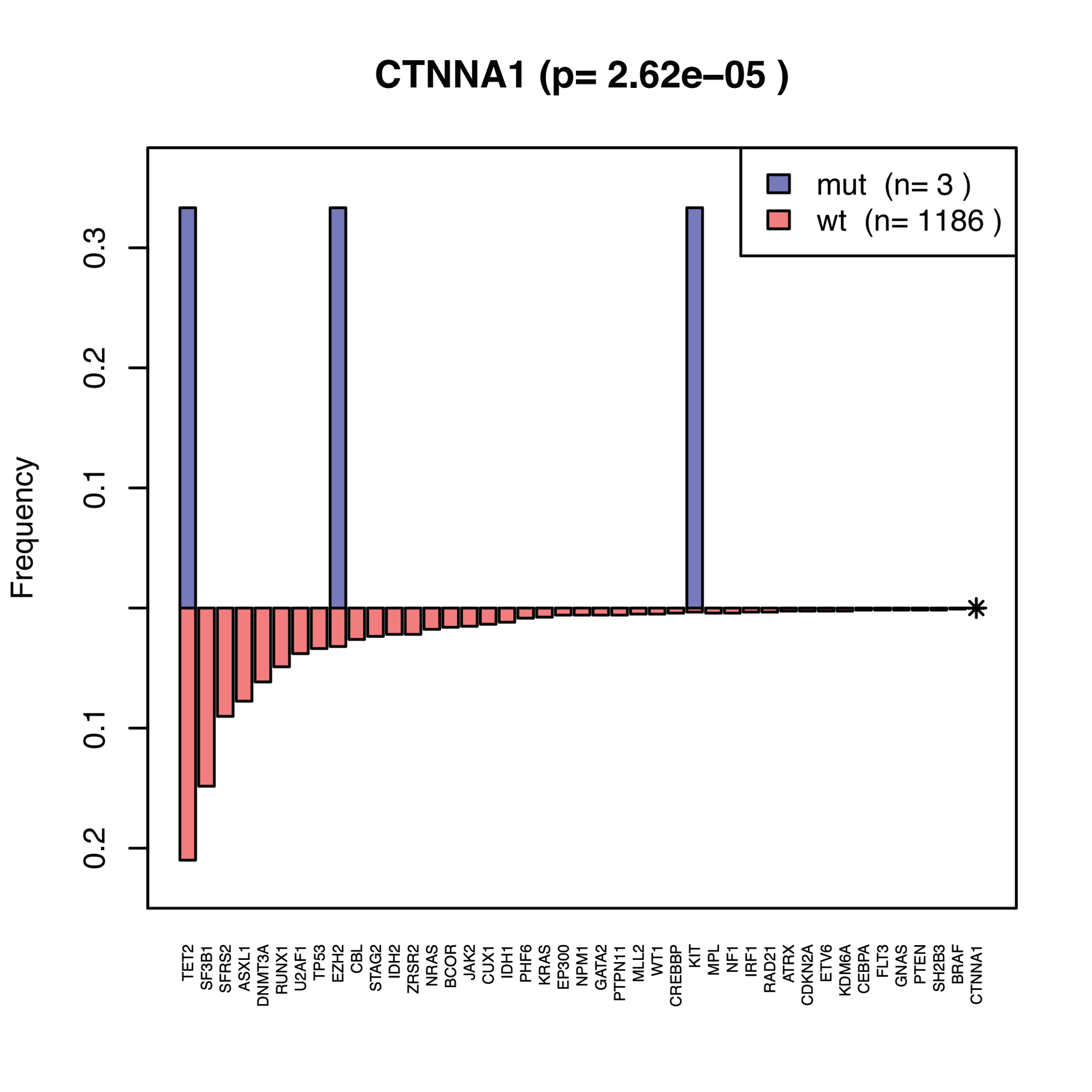

FLT3 (p= 1)
mut (n= 1)
wt (n= 1187)
Frequency
1
0.9
0.7
0.5
0.3
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

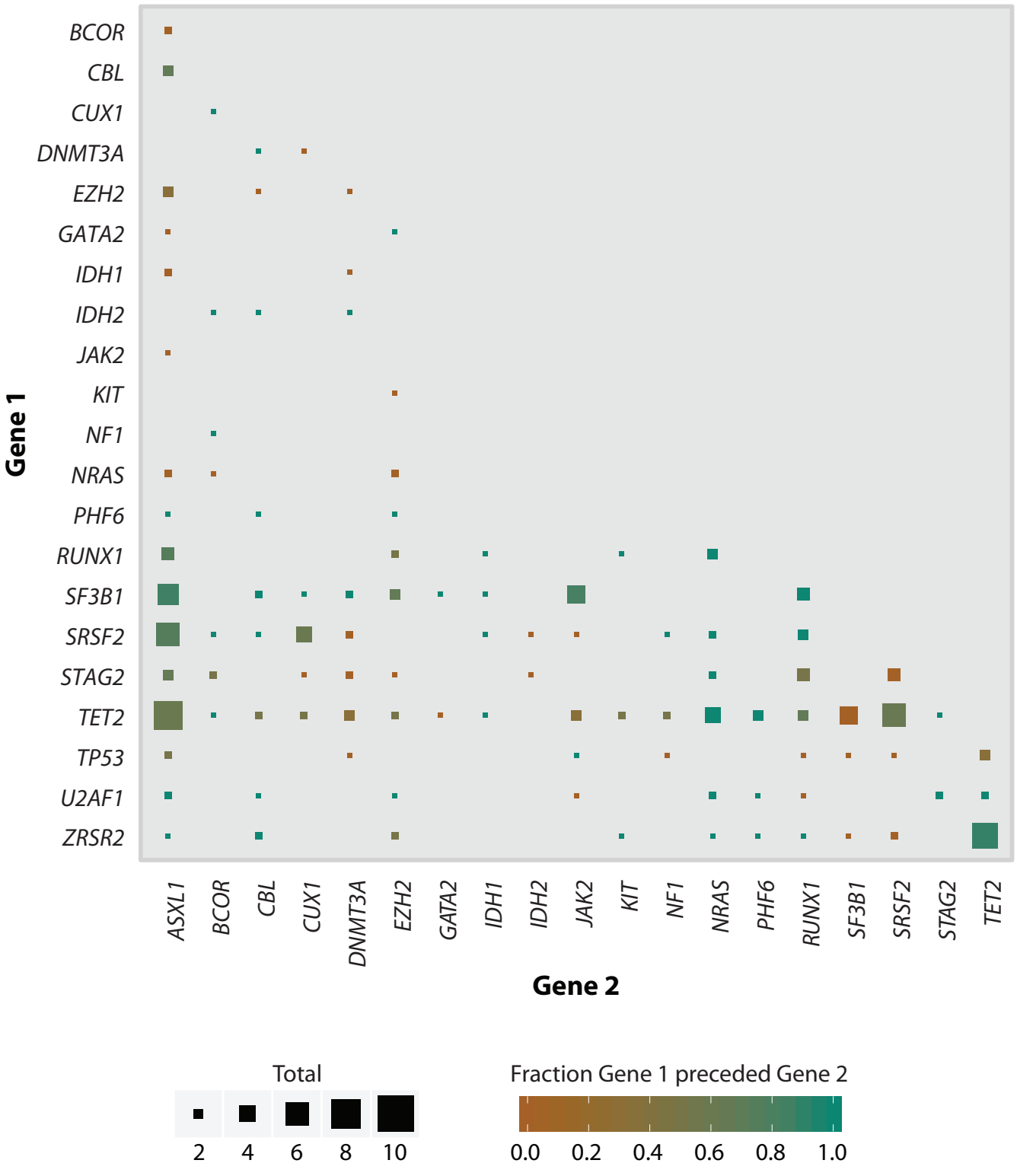

GNAS (p= 0.999)
mut (n= 3)
wt (n= 1185)
Frequency
0.3
0.2
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1


PTEN (p= 0.985)
mut (n= 3)
wt (n= 1185)
Frequency
0.3
0.2
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

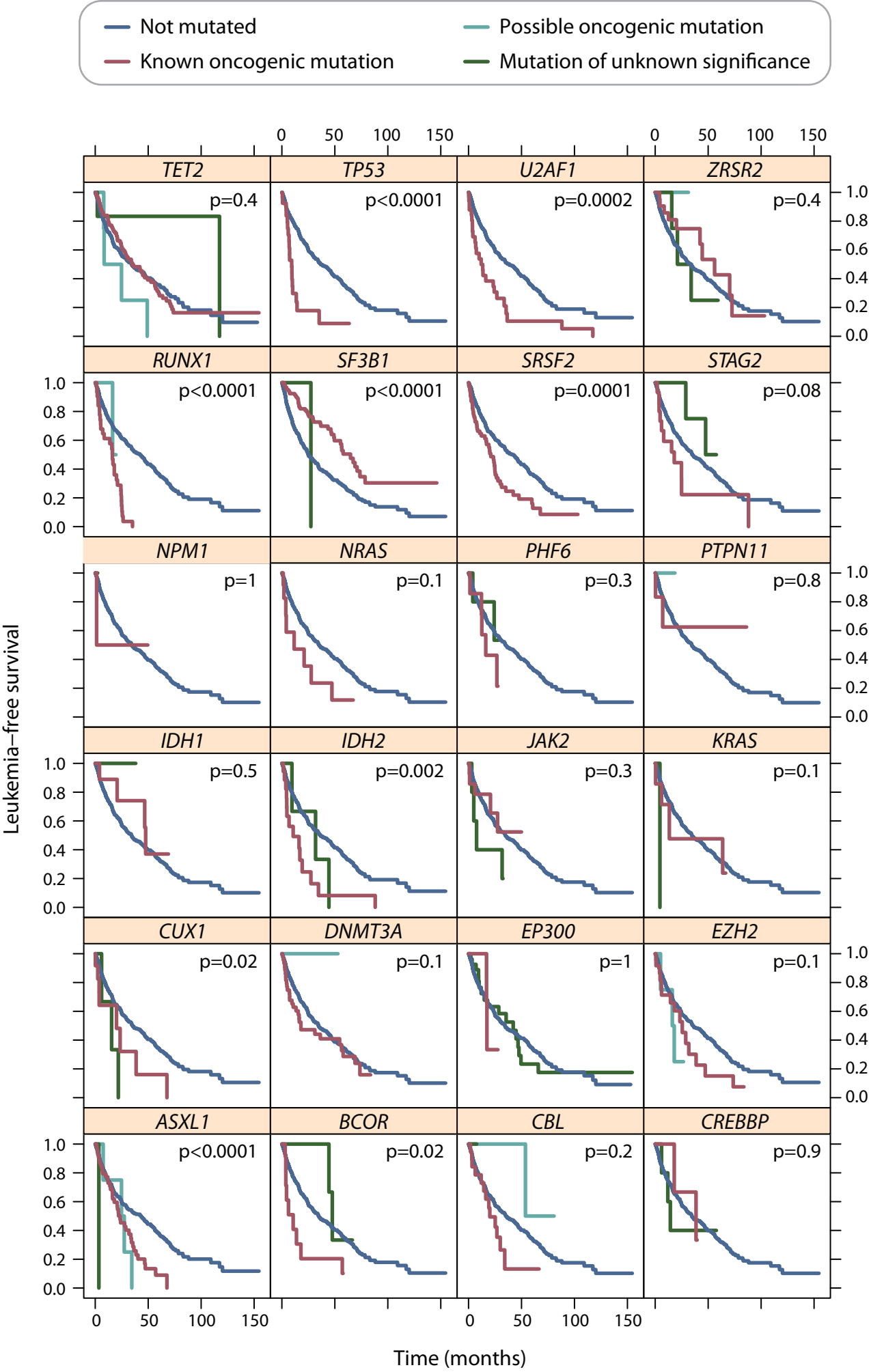

SH2B3 (p= 0.836)
mut (n= 5)
wt (n= 1183)
Frequency
0.4
0.3
0.2
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1


BRAF (p= 1)
mut (n= 2)
wt (n= 1187)
Frequency
0.5
0.4
0.3
0.2
0.1
0.2
TET2 SF3B1 SFRS2 ASXL1 DNMT3A RUNX1 U2AF1 TP53 EZH2 CBL STAG2 IDH2 ZRSR2 NRAS BCOR JAK2 CUX1 IDH1 PHF6 KRAS EP300 NPM1 GATA2 PTPN11 MLL2 WT1 CREBBP KIT MPL NF1 IRF1 RAD21 ATRX CDKN2A ETV6 KDM6A CEBPA FLT3 GNAS PTEN SH2B3 BRAF CTNNA1

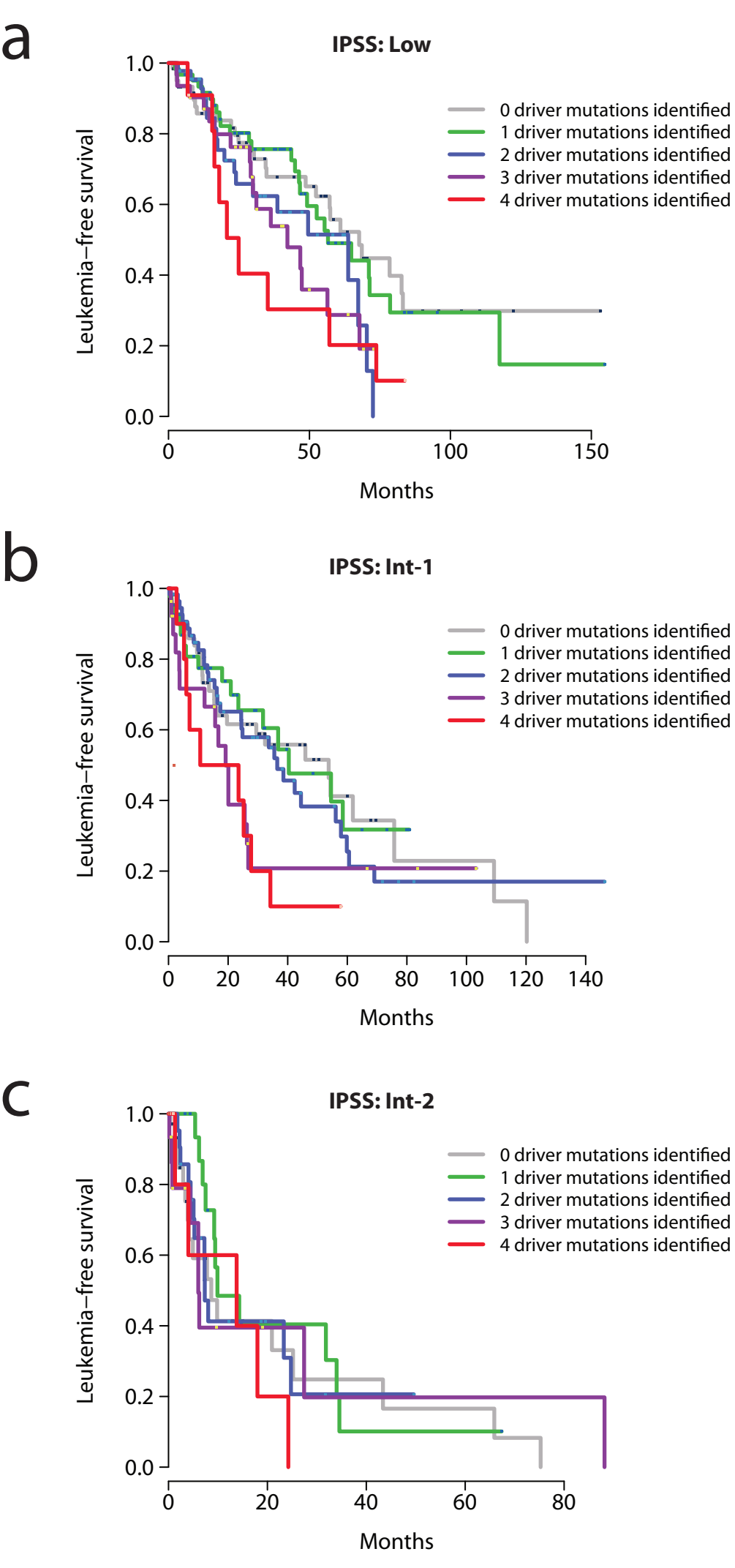

CTNNA1 (p= 2.62e−05)
mut (n= 3)
wt (n= 1186)
Frequency
0.3
0.2
0.1
0.1
0.2
TET2
SF3B1
SFRS2
ASXL1
DNMT3A
RUNX1
U2AF1
TP53
EZH2
CBL
STAG2
IDH2
ZRSR2
NRAS
BCOR
JAK2
CUX1
IDH1
PHF6
KRAS
EP300
NPM1
GATA2
PTPN11
MLL2
WT1
CREBBP
KIT
MPL
NF1
IRF1
RAD21
ATRX
CDKN2A
ETV6
KDM6A
CEBPA
FLT3
GNAS
PTEN
SH2B3
BRAF
CTNNA1



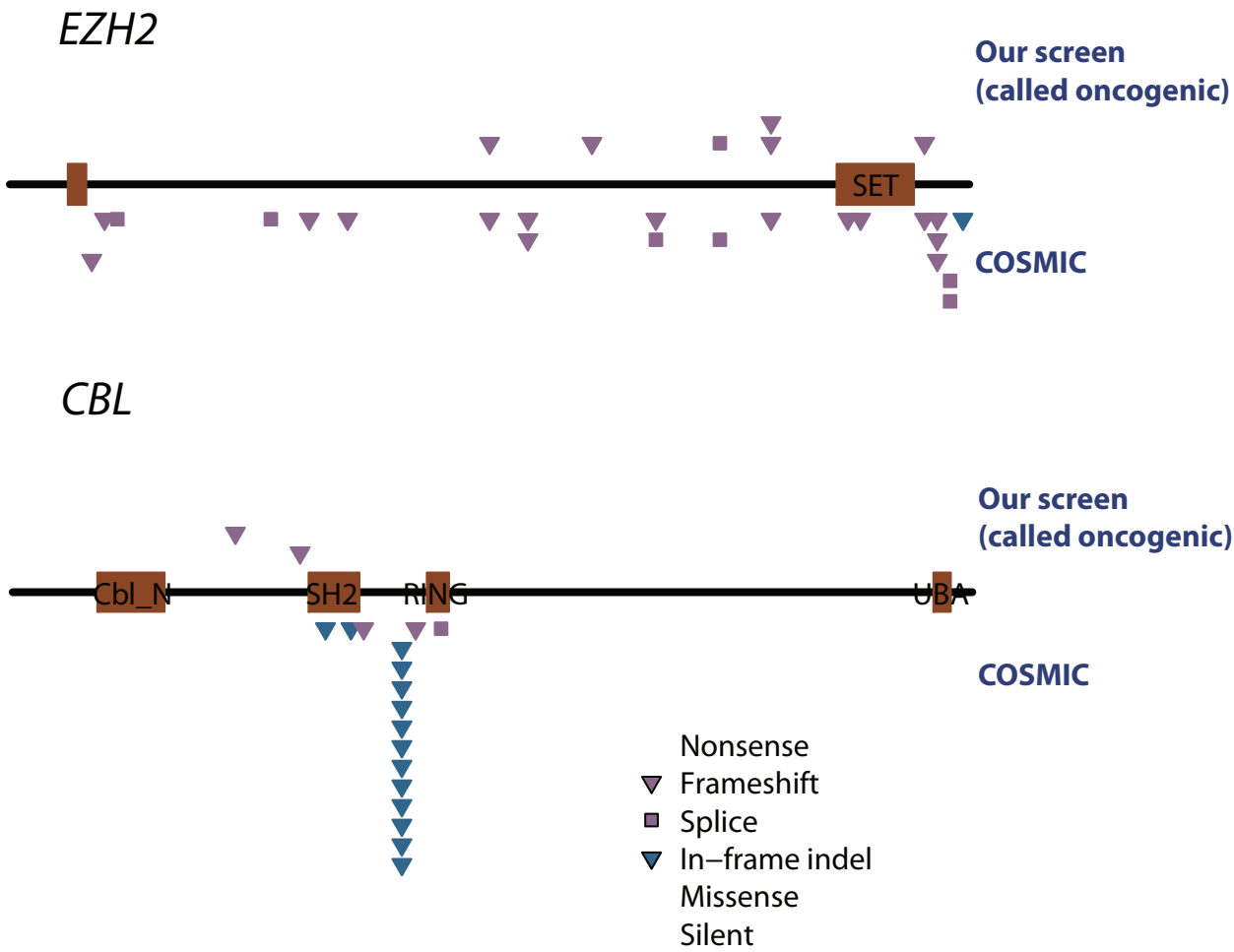
Supplementary figure 7



Supplementary figure 8



Supplementary figure 9



Supplementary figure 10