Supplemental Information for:

Systematic functional interrogation of rare cancer variants identifies oncogenic alleles

Eejung Kim[1,2,7], Nina Ilic[1,2,7], Yashaswi Shrestha[1,7], Lihua Zou[1,3,7], Atanas Kamburov[1,3,7], Cong Zhu[1], Xiaoping Yang[1], Rakela Lubonja[1], Nancy Tran[1], Cindy Nguyen[1], Michael S. Lawrence[1], Federica Piccioni[1], Mukta Bagul[1], John G. Doench[1], Candace R. Chouinard[1], Xiaoyun Wu[1], Larson Hogstrom[1], Ted Natoli[1], Pablo Tamayo[1,4], Heiko Horn[1,5], Steven M. Corsello[1,2], Kasper Lage[1,5], David E. Root[1], Aravind Subramanian[1], Todd R. Golub[1,6], Gad Getz[1,3], Jesse S. Boehm[1], William C. Hahn[1,2]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
[2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.
[3]Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA.
[4]Department of Medicine, University of California, San Diego, La Jolla, California, USA.
[5]Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts, USA.
[6]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.
[7]These authors contributed equally to this work.
Correspondence should be addressed to W.C.H. (william_hahn@dfci.harvard.edu).

Running title: Rare variant characterization in cancer.

Corresponding author: William C. Hahn, M.D., Ph.D., 450 Brookline Avenue, Dana 1538, Boston, MA 02215 USA, 617-632-2641 (phone), 617-632-4005 (fax), william_hahn@dfci.harvard.edu

Supplemental Information includes:

SUPPLEMENTARY TABLES (as separate excel files)
SUPPLEMENTARY TABLE LEGEND
SUPPLEMENTARY FIGURES AND FIGURE LEGEND

SUPPLEMENTARY TABLES (as separate excel files)

Supplementary Table S1 Genes and alleles selected for the project (**excel file**)
Supplementary Table S2 Annotation of 1163 ORFs (**excel file**)
Supplementary Table S3 Pool composition of in vivo screen (**excel file**)
Supplementary Table S4 Composition of cells and tumors from the in vivo screen (**excel file**)
- Supplementary Table S4A: Composition of Pre-expansion cell culture
- Supplementary Table S4B: Composition of Pre-injection cell culture
- Supplementary Table S4C: Tumor composition of Pool1
- Supplementary Table S4D: Tumor composition of Pool2
- Supplementary Table S4E: Tumor composition of Pool3
- Supplementary Table S4F: Tumor composition of Pool4
- Supplementary Table S4G: Tumor composition of Pool5
- Supplementary Table S4H: Tumor composition of Pool6
- Supplementary Table S4I: Tumor composition of Pool8
- Supplementary Table S4J: Tumor composition of Pool9
- Supplementary Table S4K: Tumor composition of Pool10
- Supplementary Table S4L: Tumor composition of Pool11
- Supplementary Table S4M: Tumor composition of Pool12
- Supplementary Table S4N: Tumor composition of Pool13

Supplementary Table S5 L1000 gene expression data of 1036 ORFs (**excel file**)
Supplementary Table S6 Comparison to in silico methods (**excel file**)

SUPPLEMENTARY FIGURES

Supplementary Figure S1: Distribution of barcode read representation in pre-expansion and pre-injection samples.
Supplementary Figure S2: Tumor composition of *in vivo* pooled screen, excluding the pools shown in Figure 2.
Supplementary Figure S3: Gene expression differentiates functional alleles.
Supplementary Figure S4: Validation of rare oncogenic alleles, excluding the ones shown in Figure 4.
Supplementary Figure S5: Gene expression signatures of *NFE2L2* wild type and gain-of-function mutants are correlated.
Supplementary Figure S6: Comparison to *in silico* methods.

SUPPLEMENTARY TABLE LEGEND

Supplementary Table S1: Genes and alleles selected for the project.
This table includes description of all the alleles selected for the project, including the ones excluded due to template unavailability and unsuccessful mutagenesis. The meaning of column headings is specified below:

- template_available: TRUE, when the template ORF was available in hORFeme 8.1 collection
- mutagenesis_successful: TRUE, when the mutagenesis was successful
- BarcodedVectorID: identification number given to each vector. (failed_QC: sequencing results were not satisfactory, template_unavailable: template was not available)
- n_AML - n_UCEC: number of times each mutation was found in each cancer type
- n_pancan: sum of columns of n_AML - n_UCEC. This column was used for generating Fig. 1B.

Supplementary Table S2: Annotation of 1163 ORFs.
This table includes description of all the alleles used in the *in vivo* screening and gene expression experiments. Only the mutant alleles (PC_MUT, under category) were included in the in vivo screening (474 total alleles). All of the ORFs were included for the gene expression assay. The meaning of column headings are specified below:

- plate_well_ID: identification number given to each well of the assay plate. This ID is used in Supplementary Table S5.
- clone_ID: identification number identical to BarcodedVectorID in Supplementary Table S1.
- Vector: lentiviral vectors used. PLX_TRC317 is identical to pLEX_307 (https://www.addgene.org/41392/). It has EF1α promoter and puromycin selection marker. PLX_TRC304 is identical to pLX304 (https://www.addgene.org/25890/). It has CMV promoter and blasticidin selection marker.
- open_close: when the C-terminal of the ORFs did not have the stop codon, it resulted in V5 tagging at the C-terminal (annotated as “open”). “close” otherwise.
- gene, protein_change: shows gene and protein change.
- point_mutation: additional point mutation found. “c.262C>T|p.H88Y” shows that nucleotide position 262 was T, not C, which resulted in non-synonymous mutation H88Y.
- indel: additional insertion or deletion found. “1121delG” means nucleotide position 1121 had a single G deletion.
- intended_transcript: shows the intended RefSeq accession number.
- category:
    - PC_MUT: mutant alleles generated for the study. 474 in total.
    - PC_WT: wild type alleles generated for the study. 187 unique alleles, 334 in total due to many alleles having two entries (open and close forms).
    - REF: reference alleles of known biological function. 232 unique alleles, 308 in total due to many alleles having more than one entry.
    - CTL_INRT: negative controls including BFP, eGFP, HcRED, LacZ, and Luciferase. 5 unique alleles, 35 in total due to each allele being included seven times.
    - CTL_L1000: internal expression control for L1000 assay, including DNMT3A, NFE2L2, NFKBIA, RHEB. 4 unique alleles, 12 total due to each allele being included three times.
    - infection_efficiency: infection efficiency shown in percentage. Please refer to Methods.

Supplementary Table S3: Pool composition of *in vivo* screen.
This table shows the composition 14 pools. The first column shows the name of the mutant or control alleles. TRUE mean that the allele belongs to that pool. For example, "A4GALT_p.A272V" belongs to Pool 5 and Pool 14. To search for alleles in each pool, use the filtering function of the Excel (shown as funnel shaped icon).

Supplementary Table S4: Composition of cells and tumors from the *in vivo* screen.
These tables show the composition of pre-expansion and pre-injection cells and tumors in each pool. The numbers are shown in percentage.

- Supplementary Table S4-1: Composition of Pre-expansion cell culture. Supplementary Table S3 describes the pool membership of each Mutation (first column). This table shows the barcode representation immediately after pooling the cells after arrayed infection.
- Supplementary Table S4-2: Composition of Pre-injection cell culture. Supplementary Table S3 describes the pool membership of each Mutation (first column). This table shows the barcode representation after 15 days of *in vitro* culture, right before cells were injected into nude mice. All enrichment analysis was done using this as reference point.
- Supplementary Table S4-3 - Supplementary Table S4-14: Composition of each tumor in the Pool1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13 in that order. Second column to last column headings show tumor ID. Tumor ID "P1M1_L" means Pool1, mouse 1, left flank injection site. "L": left flank, "R": right flank, "T": upper back.

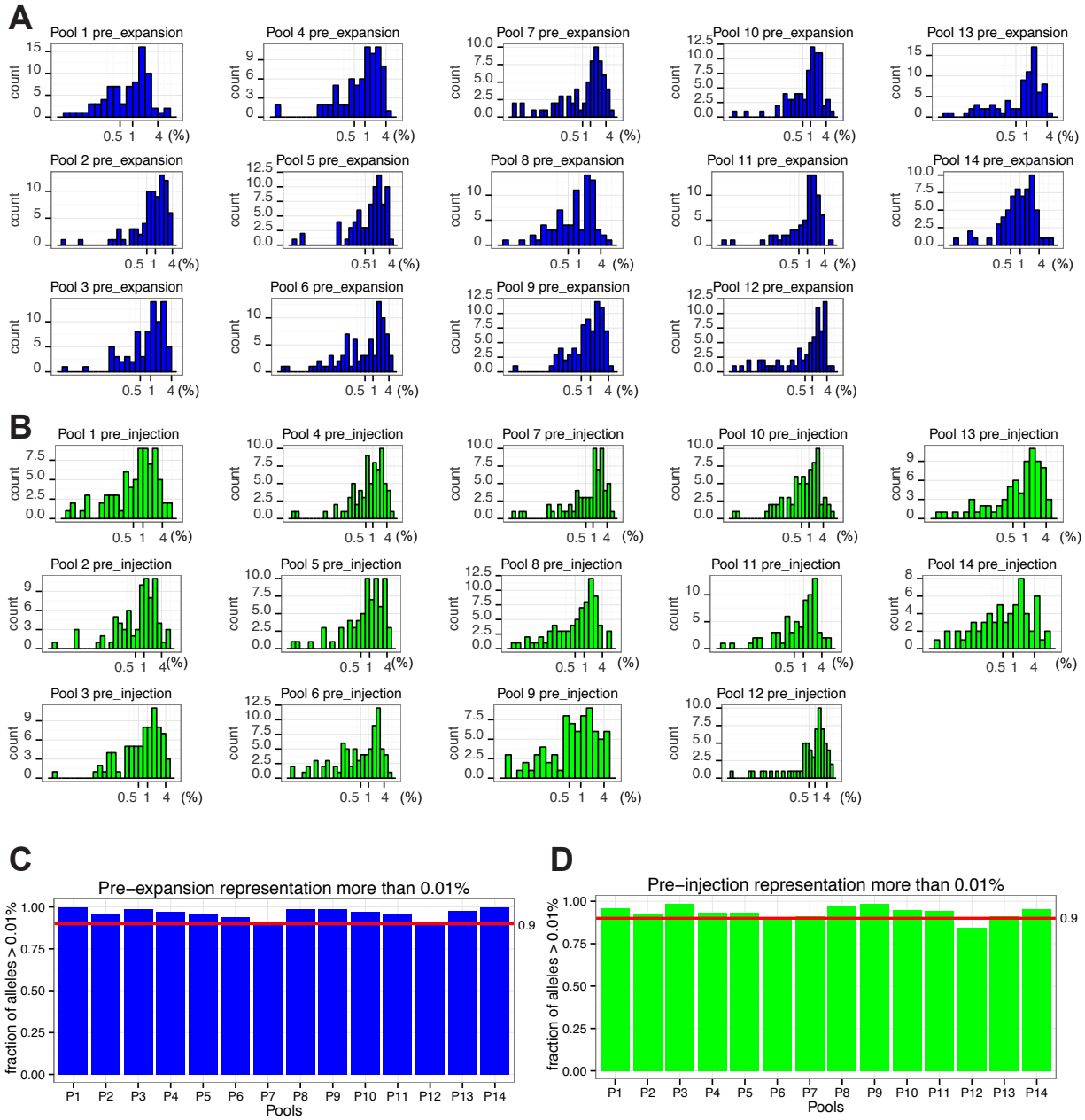
Supplementary Table S5: L1000 gene expression data of 1036 ORFs.
This table shows the L1000 gene expression date of 1036 ORFs that passed 40% infection efficiency cutoff.

- landmark: this column shows the 978 landmark genes, whose expressions are measured in L1000 assay.
- second column to last column: these columns show the plate_well_ID, as specified in Supplementary Table S2.

Supplementary Table S6: Comparison to *in silico* methods.
This table shows the calls of four different in silico methods (Polyphen2, Mutation Assessor, CHASM, and VEST) and comparison to our results. See the methods for description.

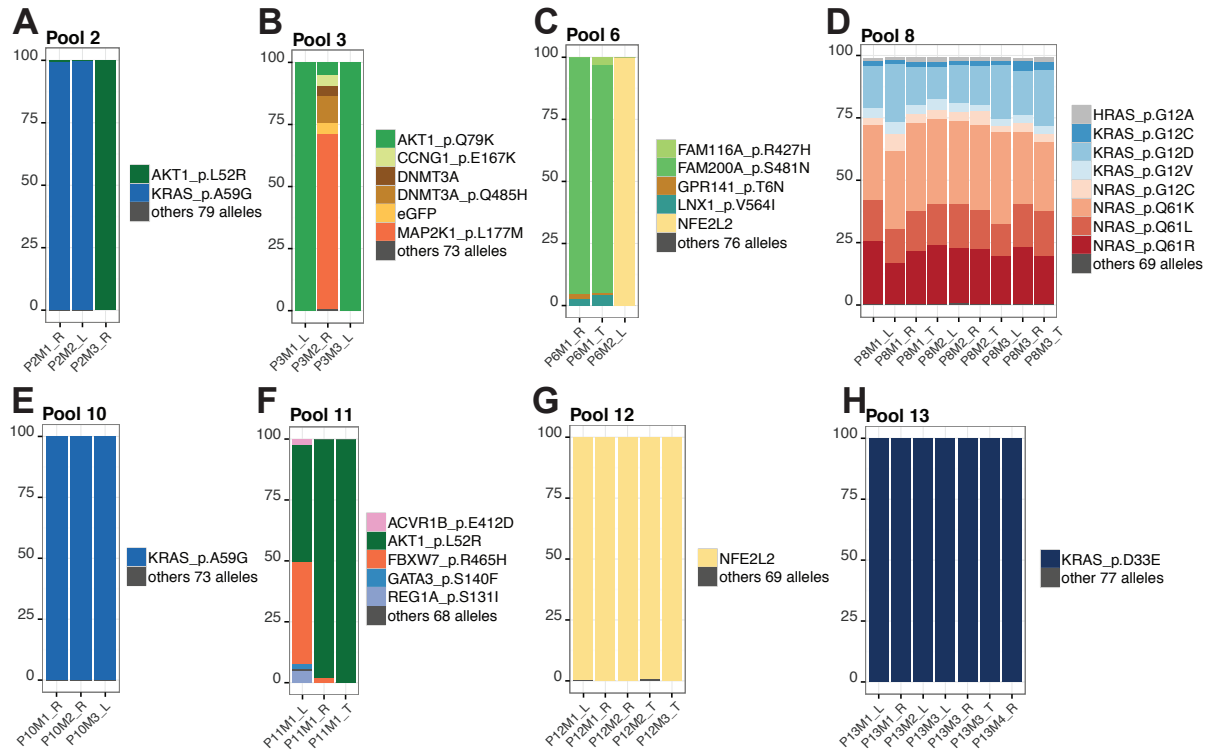
- Mutation: lists alleles
- This Study: functional description from this study. "functional" denotes both gain-of-function and loss-of-function alleles. "neutral" denotes likely passenger mutations.
- Concordance to Polyphen2, Mutation Assessor, CHASM, VEST: "1" if concordant, "0" otherwise.
- Polyphen2 score, Polyphen2 call: output from Polyphen2.
- Mutation Assessor score, Mutation Assessor call: output from Mutation Assessor
- CHASM cancer driver p-value (missense), CHASM FDR (red<0.05): output from CHASM. FDR <0.05 was colored red.
- VEST pathogenicity p-value (non-silent), VEST FDR (red<0.05): output from VEST. FDR <0.05 was colored red.

Supplemetary Figure S1



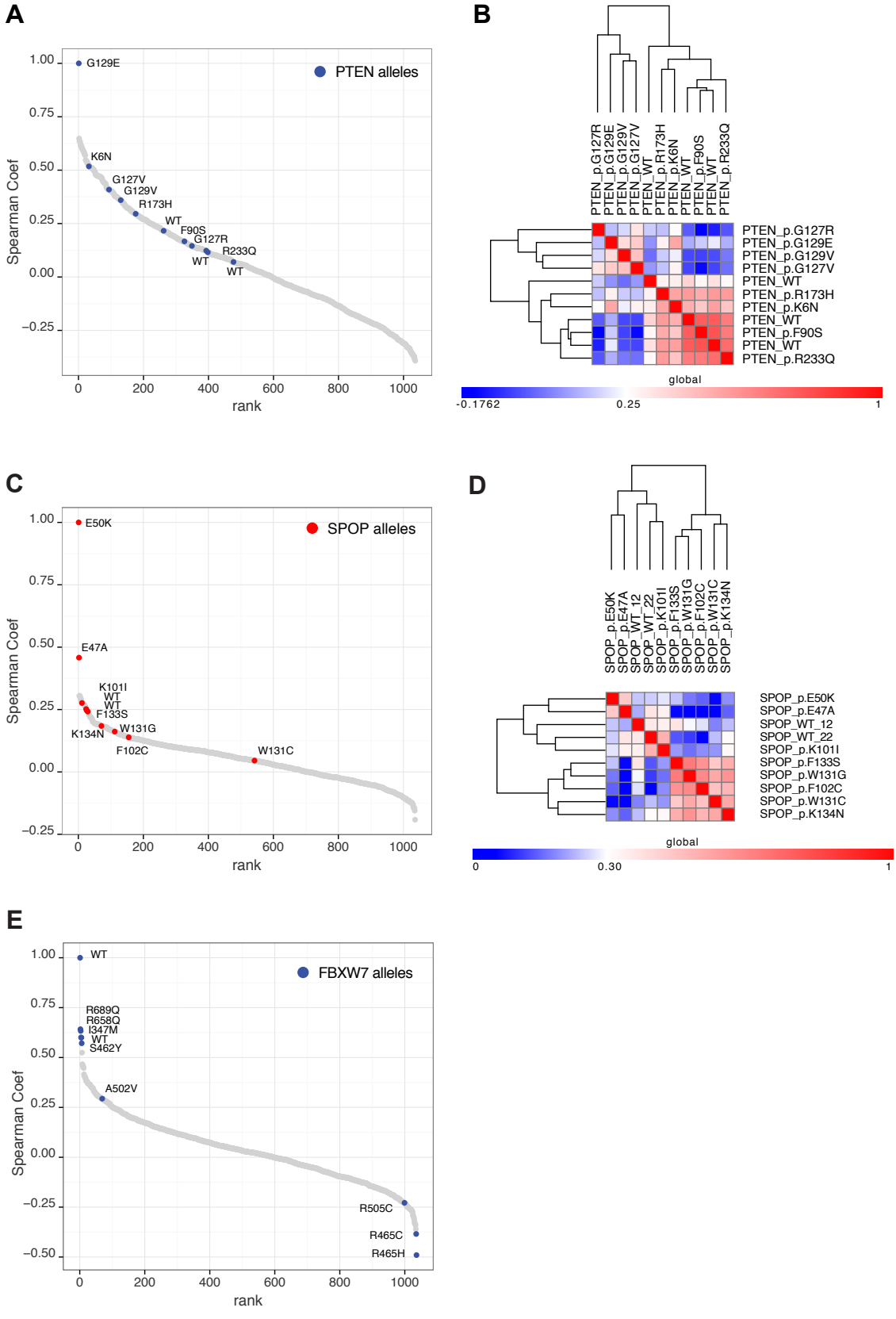
Supplementary Figure S1: Distribution of barcode read representation in pre-expansion and pre-injection samples. (**A**) Allele representation immediately after pooling cells (called “pre-expansion”) according to the pool composition (Supplementary Table S3). Each pool contains ~75 alleles. Majority of alleles were represented at 0.5-4%. The data for this histogram is available in Supplementary Table S4-1. (**B**) Allele representation after 15-day culture, immediately before the injection into nude mice (called “pre-injection”). Majority of alleles were represented at 0.5-4%. The data for this histogram is available in Supplementary Table S4-2. (**C**) Percentage of alleles in each pool that was represented at more than 0.01% in pre-expansion cell pellet. (**D**) Percentage of alleles in each pool that was represented at more than 0.01% in pre-injection cell pellet.

Supplementary Figure S2



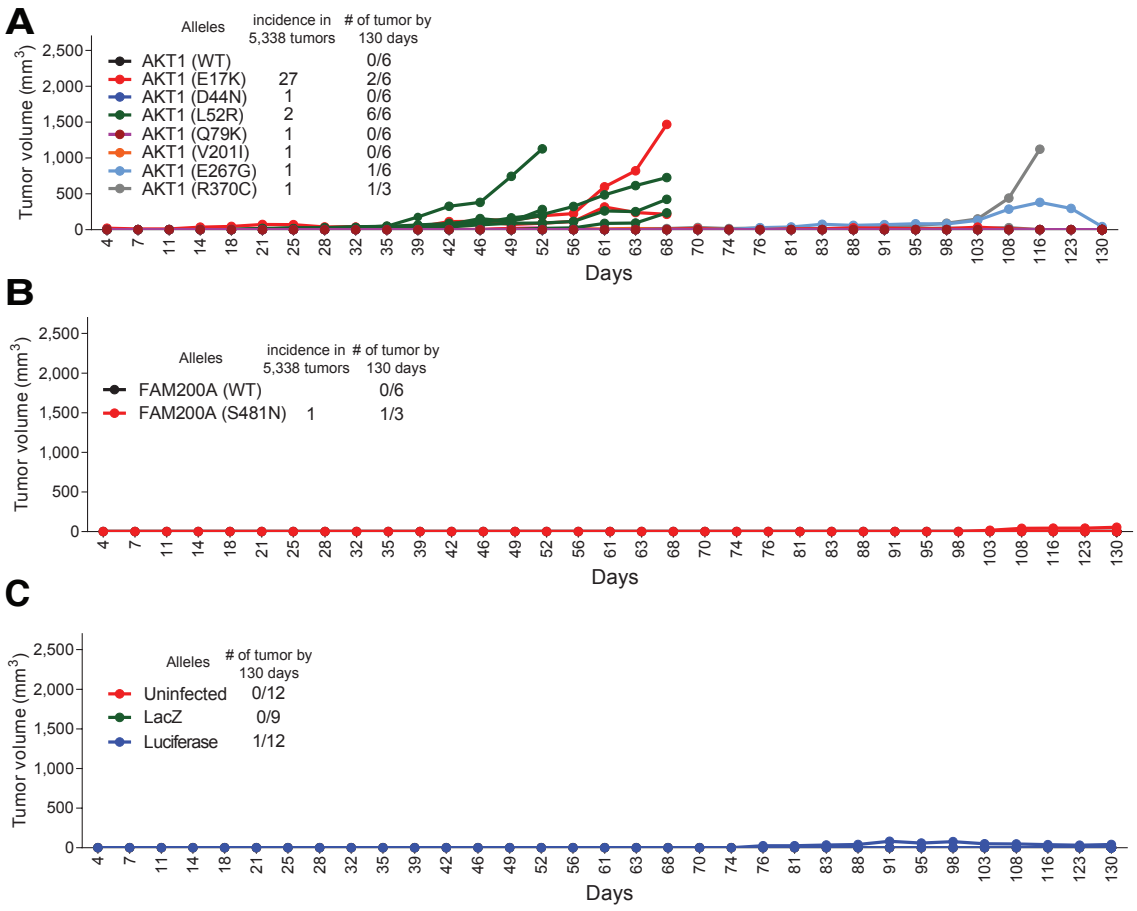
Supplementary Figure S2: Tumor composition of *in vivo* pooled screen, excluding the pools shown in Figure 2 (**A**) Tumor composition of pool 2. *AKT1*^{L52R} and *KRAS*^{A59G} scored. (**B**) Tumor composition of pool 3. *AKT1*^{Q79K} scored. (**C**) Tumor composition of pool 6. *FAM200A*^{S481N} and *NFE2L2*^{WT} scored. (**D**) Tumor composition of pool 8. Tumor composition was analogous to that of pool 1. (**E**) Tumor composition of pool 10. *KRAS*^{A59G} scored. (**F**) Tumor composition of pool 11. *AKT1*^{L52R} scored. (**G**) Tumor composition of pool 12. *NFE2L2*^{WT} scored. (**H**) Tumor composition of pool 13. *KRAS*^{D33E} scored.

Supplementary Figure S3



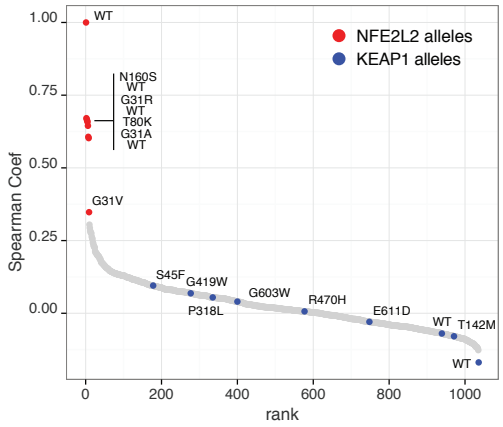
Supplementary Figure S3: Gene expression differentiates functional alleles.
(**A**) When alleles were correlated to *PTEN*^{G129E}, other likely loss-of-function alleles G12V, G129V, and G127R were only moderately correlated. (**B**) When the gene expression changes induced by expression of *PTEN* allelic series were clustered, likely loss-of-function alleles were separated from the likely passenger mutants. (**C**) When alleles were compared to *SPOP*^{E50K}, other likely loss-of-function allele E47A was highly correlated. (**D**) When the gene expression changes induced by expression of *SPOP* allelic series were clustered, likely loss-of-function, dominant negative alleles discovered in prostate cancer were separated from the wild type and likely loss-of-function alleles found in endometrial cancer. (**E**) When alleles were correlated to the *FBXW7* wild type, known dominant interfering alleles (R505C, R465C, R465H) were anti-correlated to the wild type.

Supplementary Figure S4



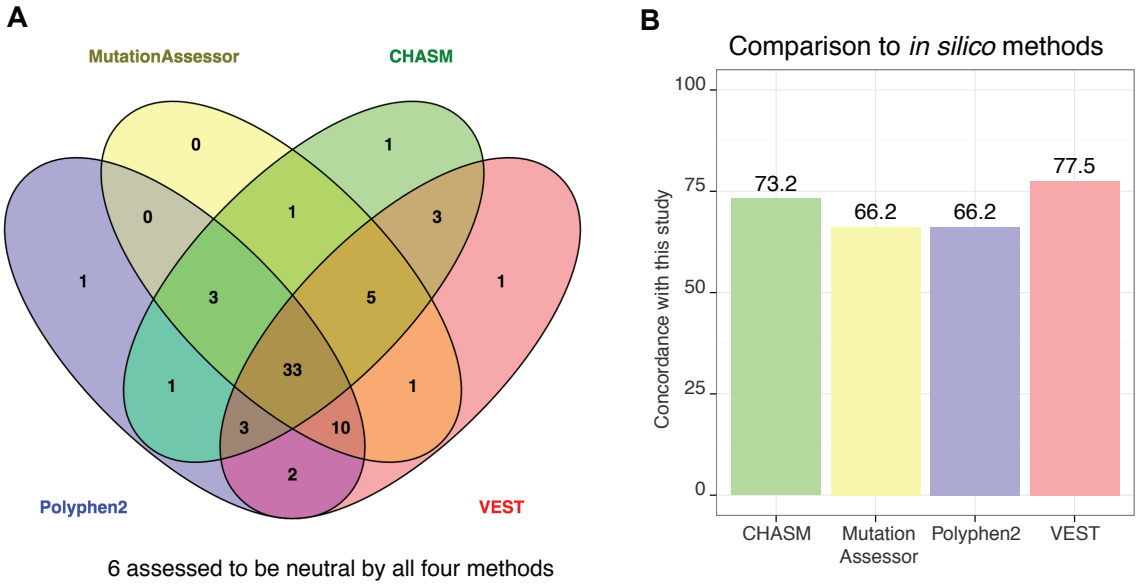
Supplementary Figure S4: Validation of rare oncogenic alleles, excluding the ones shown in Figure 4.
(**A**) Individual tumor validation of *AKT1* alleles. E17K, L52R, E267G, and R370C formed tumors. Q79K did not form tumor. One mouse of *AKT1*^{R370C} died of unknown reason. (**B**) Individual tumor validation of *FAM200A* alleles. *FAM200A*^{S481N} formed one small tumor at later time point. One mouse of *FAM200A*^{S481N} died of unknown reason. (**C**) Negative controls in individual tumor validation. Four mice were used in each of uninfected, LacZ-transduced, and Luciferase-transduced groups. One small tumor formed in Luciferase-transduced groups and regressed spontaneously. One mouse in LacZ-transduced group died of unknown reason.

Supplementary Figure S5



Supplementary Figure S5: Gene expression signatures of *NFE2L2* wild type and gain-of-function mutants are correlated.
Gene expression signatures from *NFE2L2* WT, G31A, G31V, G31R, T80K and N160S were highly correlated. *KEAP1* WT signature was anti-correlated to that of *NFE2L2*.

Supplementary Figure S6



Supplementary Figure S6: Comparison to *in silico* methods.
(**A**) Venn diagram of four different methods showing the overlap of the number of alleles called "functional" in each method. Please refer to Methods for description.
(**B**) Concordance rate of the four different *in silico* methods to the analysis from this study. The concordance rate ranged from 66 – 77%.