Supplementary Materials:
Evaluating the Evaluation of Cancer Driver Genes

Authors

Collin J. Tokheim, Nickolas Papadopoulis, Kenneth W. Kinzler, Bert Vogelstein, and Rachel Karchin

Evaluated driver gene prediction methods

In addition to 20/20+, we selected several methods that cover alternate methodological approaches: mutational clustering, mutation functional impact, and significantly mutated genes. All methods were run in-house, using the latest version of the software provided by the authors. Detailed output from each method is in Dataset S2.

Oncodrive Suite We evaluated three methods in the Oncodrive suite, namely OncodriveFM [1], OncodriveFML, and OncodriveClust [2], that identify driver genes based on mutation functional impact bias (OncodriveFM, OncodriveFML) and mutational clustering (OncodriveClust). OncodriveFM functional impact scores for missense mutations were obtained using PolyPhen2 (HumVar) [3], SIFT [4], and MutationAssessor [5] from the pre-computed scores in the hg19 UCSC genome browser database [6]. Since PolyPhen2, SIFT, and MutationAssessor score missense mutations, silent mutations were encoded as the least damaging for each method (0, 1, and -2, respectively) and inactivating mutations (nonsense, frameshift indel, lost stop, lost start, and splice site) were assigned the most damaging score (1, 0, and 3.5, respectively) in accordance with recommendation for OncodriveFM [1]. Because low scores for SIFT correspond with more damage, SIFT scores were adjusted to be more damaging with higher scores by taking one minus the original SIFT score. OncodriveFM version 0.6.0 (https://bitbucket.org/bbglab/oncodrivefm) was executed with default parameters using all three scores as a measure of functional impact bias. OncodriveFML used pre-computed CADD scores [7] for functional impact bias (fetched via OncodriveFML), and a CDS regions file from the OncodriveFML website (https://bitbucket.org/bbglab/oncodrivefml). OncodriveFML v1.2 was run using default parameters. OncodriveClust version 0.4.1, installed as per directions for python2 environment (https://bitbucket.org/bbglab/oncodriveclust/overview), was also executed with default parameters on missense mutations, except for the minimum mutations were lowered to 1 (5 default) for cancer type specific analysis due to few genes passing the threshold.

MutsigCV We evaluated MutsigCV, a significantly mutated gene method [8], which adjusts for known covariates of mutation rate. The latest publically available MutsigCV version (v1.4) was executed according to recommended practices (https://www.broadinstitute.org/cancer/cga/mutsig_run). Briefly, expression, replication time, and HiC features obtained from the Broad website (https://www.broadinstitute.org/cancer/cga/mutsig_run) were used as mutation rate covariates. Additionally, the recommended exome coverage file (http://www.broadinstitute.org/cancer/cga/sites/default/files/data/tools/mutsig/reference_files/exome_full192.coverage.zip) was utilized, as precise coverage information was not available from the two studies originating the obtained data [9,10]. Further, non-coding mutations were removed from the data set and the exome coverage file was adjusted to reflect an absence of sequencing non-coding bases.

TUSON (TUmor Suppressor and ONcogenes) Explorer. TUSON uses mutational clustering and relative amount of damaging mutations, as indicators of oncogenes and tumor suppressor genes [10]. TUSON combines p-values (Stouffer-liptak method [11]) for each feature with a weight optimized by grid-search, using known oncogenes and tumor suppressor genes, to yield a single combined p-value for each of oncogene and tumor suppressor gene. An updated version of TUSON was obtained directly from the authors

(June 6, 2014) and executed according to recommendations from the authors. Required PolyPhen2 HumVar scores were obtained from the version 2.2.2 whole exome database (ftp://genetics.bwh.harvard.edu/pph2/whess/polyphen-2.2.2-whess-2011_12.sqlite.bz2).

ActiveDriver ActiveDriver identifies driver genes by significantly high mutation rates in particular annotated protein sites, *e.g.*, phosphorylation sites or domains [12]. We used annotations of protein sequences, phosphorylation sites, and predicted disordered protein sequences employed in Reimand *et al.* [12] to identify driver genes with significant phosphorylation site mutations. ActiveDriver version 0.0.10 with corresponding annotations (downloaded from http://individual.utoronto.ca/reimand/ActiveDriver/) was executed using default settings. We used the negative-binomial model rather than Poisson model (default) because the negative-binomial model ranked better overall in our evaluations.

MuSiC MuSiC (v0.4) is a method to identify significantly mutated genes. We used the recommended regions of interest file (ROI) file for hg19 (obtained https://github.com/ding-lab/calc-roi-covg) with a full coverage wig file. To improve the calculation of background mutation rate, we performed a search for the background mutation rate groups parameter (1(default) through 5), and chose 5 based on best performance. Other MuSiC parameters were left as default. Since MuSiC reports three p-values all of which assess significantly mutated genes, we chose the Fischer's combined p-value test (FCPT) as it performed better than other single p-values or a combination approach of the three.

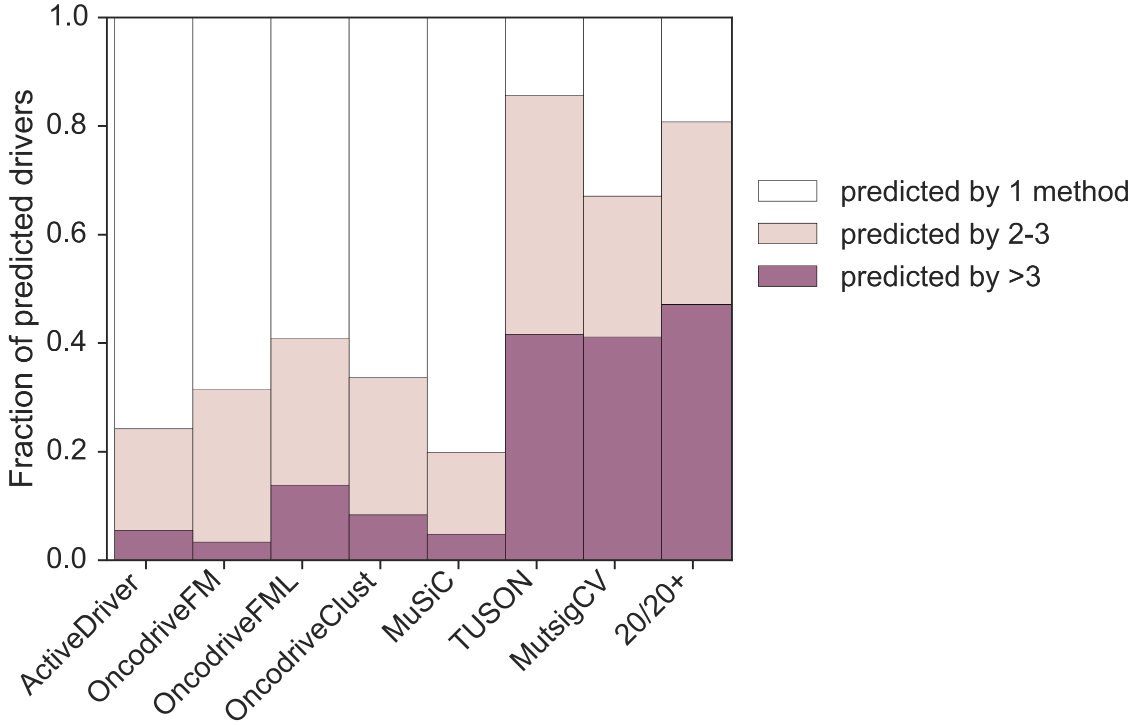

Figure S1. **Fraction of predicted driver genes for each method by consensus among methods.** Fraction of predicted drivers unique to each method, predicted by 2-3 methods or predicted by >3 methods are shown. A predicted driver gene is defined by Benjamini-Hochberg adjusted p-value (q≤0.1).

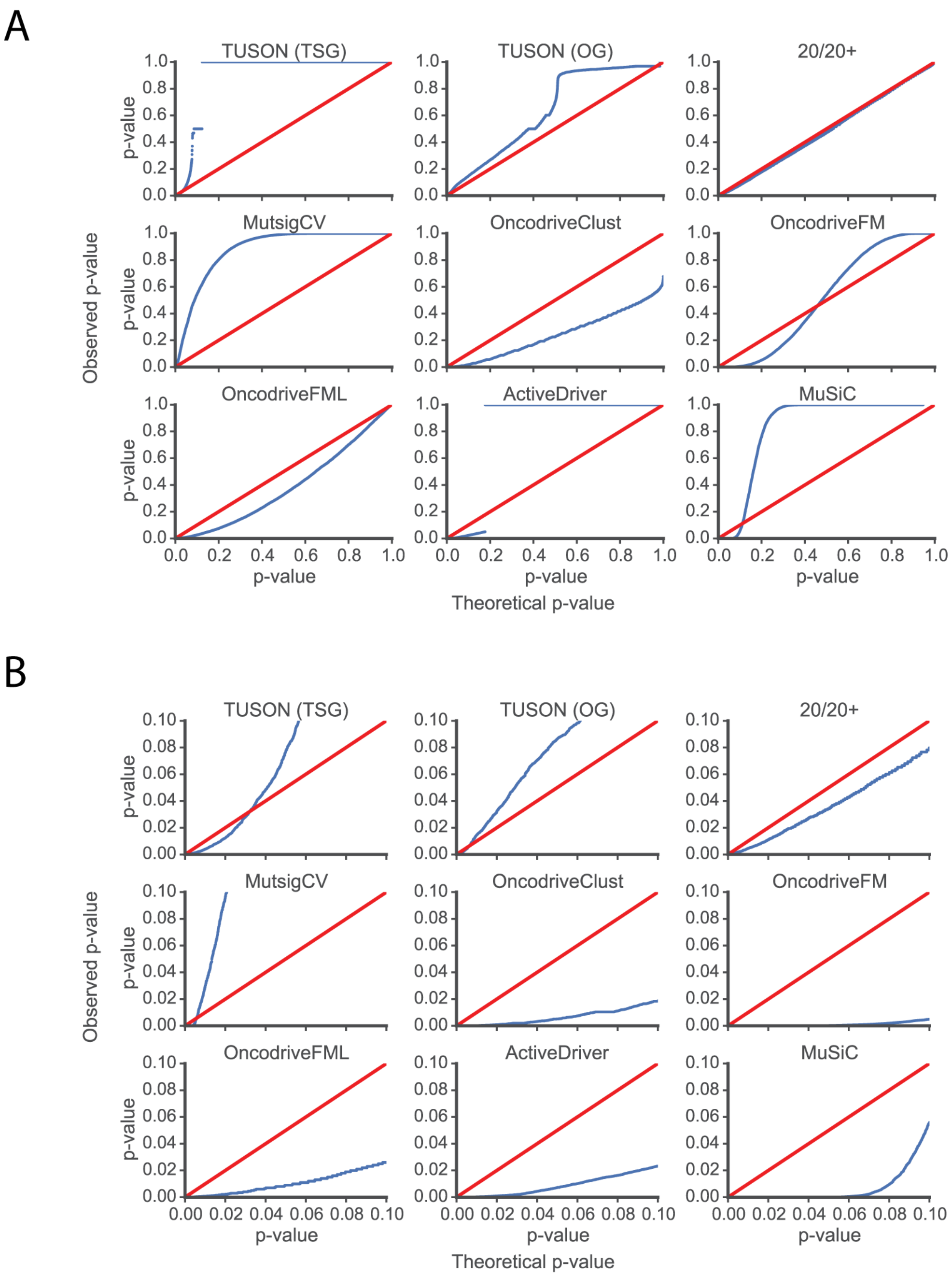

Figure S2. **Quantile-Quantile plots comparing observed and theoretical p-values for the tested methods. A.** Full p-value range from 0 to 1. **B**. Blow-up of p-values from 0. to 0.1. Observed p-values for the methods (blue) are compared to those expected from a uniform distribution (red). Genes predicted as drivers by at least 3 methods were removed along with genes in the Cancer Gene Census. TUSON oncogene and tumor suppressor gene p-values are shown separately.

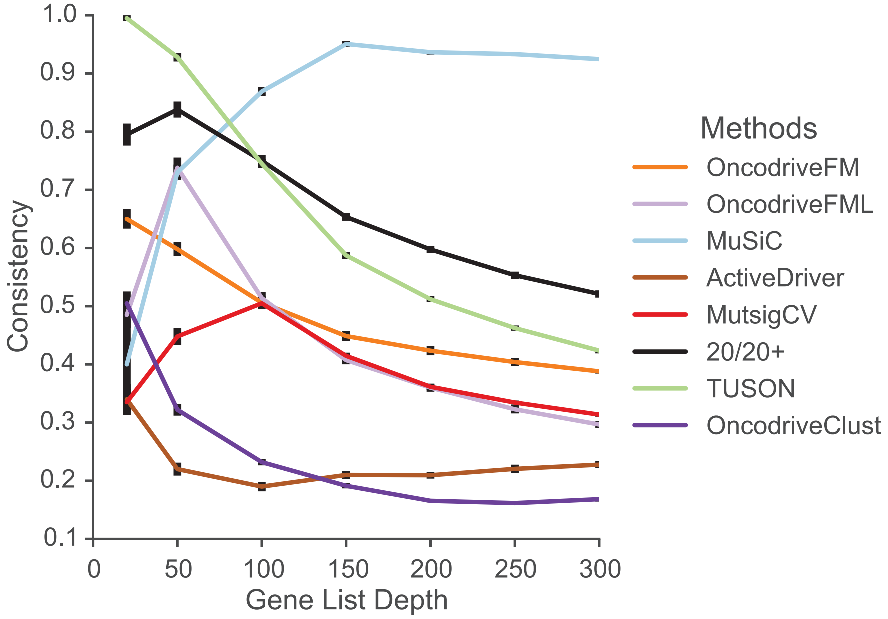

Figure S3. TopDrop consistency of pan-cancer driver gene predictions as depth threshold is varied. The consistency of each evaluated method is shown as depth threshold varies from 20 to 300. Error bars indicate ±1 SEM (standard error of the mean) across 10 repeated splits of the data.

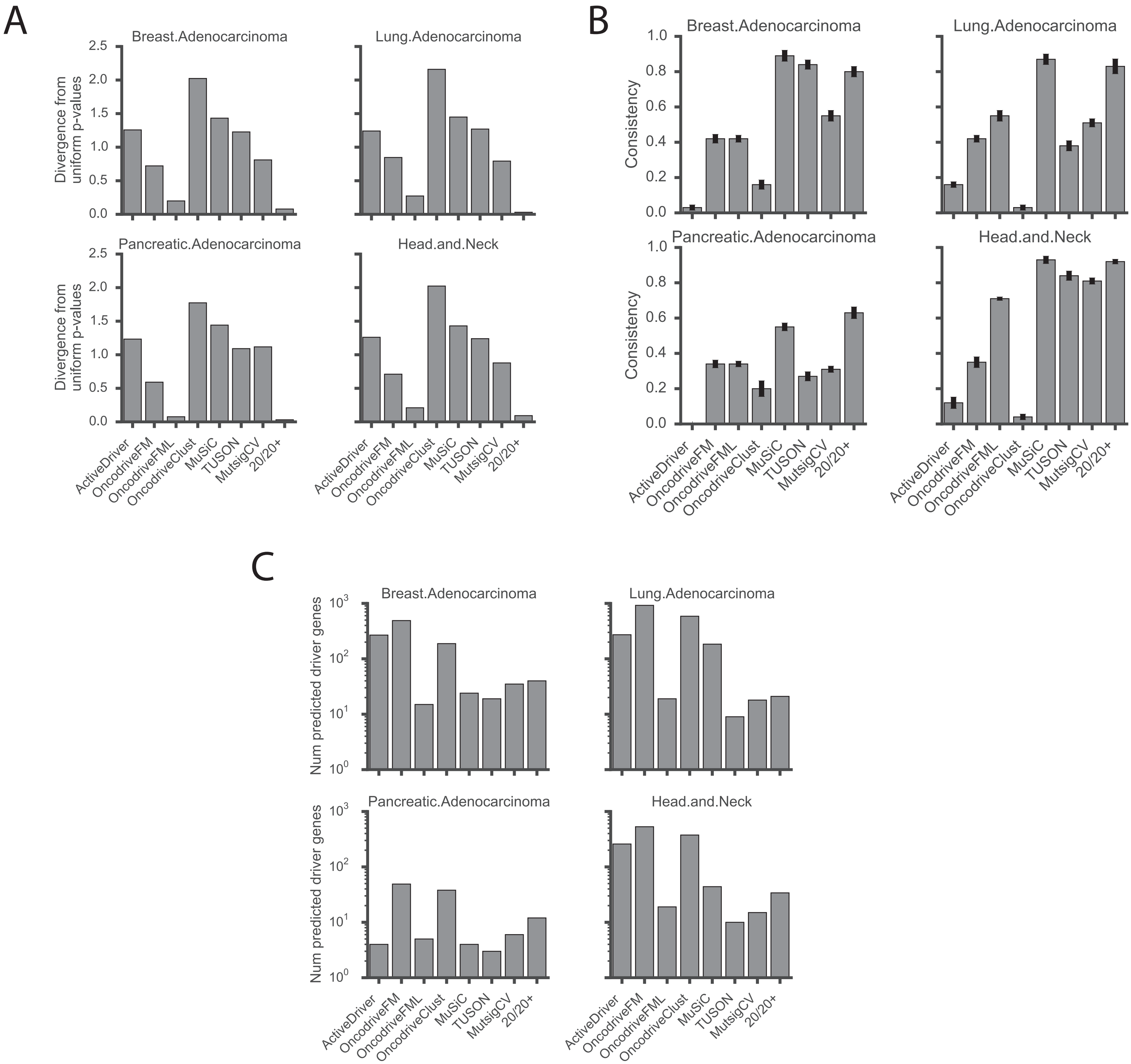

Figure S4. Evaluation of the eight methods on four different cancer types

Methods were evaluated for Mean Log Fold Change (MLFC), TDC 10 (TopDrop Consistency at a gene rank depth of 10) and number of drivers predicted (q≤0.1). 20/20+ and OncodriveFML have the lowest MLFC (least divergence between observed and theoretical p-values). MuSiC, 20/20+ and TUSON have the highest TDC 10 (consistency in gene rankings across matched random partitions of each tumor type). The four cancer types: pancreatic carcinoma (PDAC), breast adenocarcinoma (BRAC), head and neck squamous carcinoma (HNSCC) and lung adenocarcinoma (LUAD), have background somatic mutation rates ranging from moderate to high.

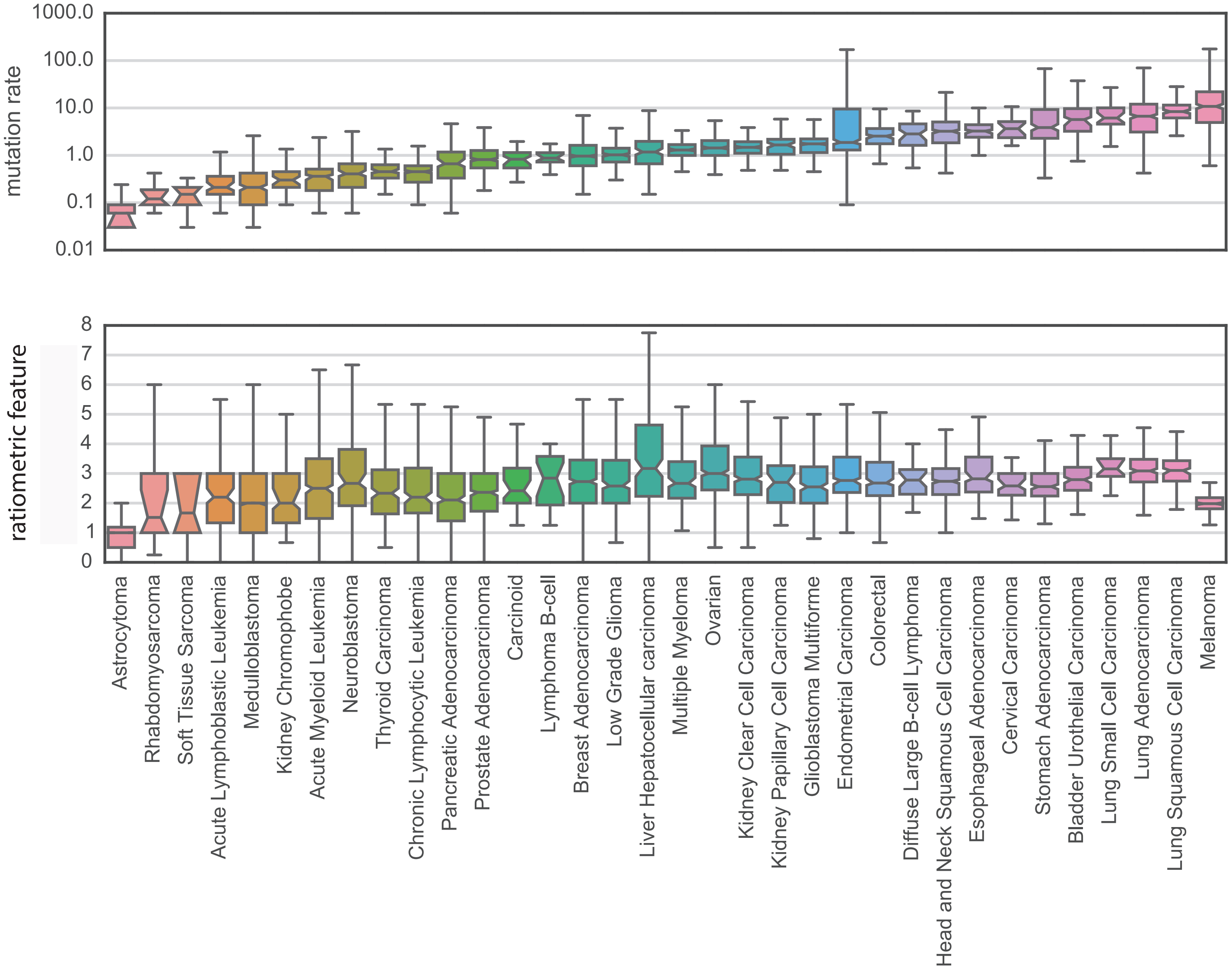

Figure S5. Background mutation rate is more variable than the ratiometric non-silent to silent mutation ratio across the 34 cancer types. The top boxplot for mutation rate is on a log10 scale and shows the mutation rate in coding sequence for the samples in our pan-cancer dataset. The bottom boxplot shows the non-silent to silent mutation ratio in coding sequence for the same samples. A pseudo-count for a silent mutation was added for each sample to avoid dividing by zero. Notches indicate bootstrap 95% confidence interval (1,000 iterations) for the median. Outliers, defined as 1.5*IQR away from the first and third quartile, are not shown.

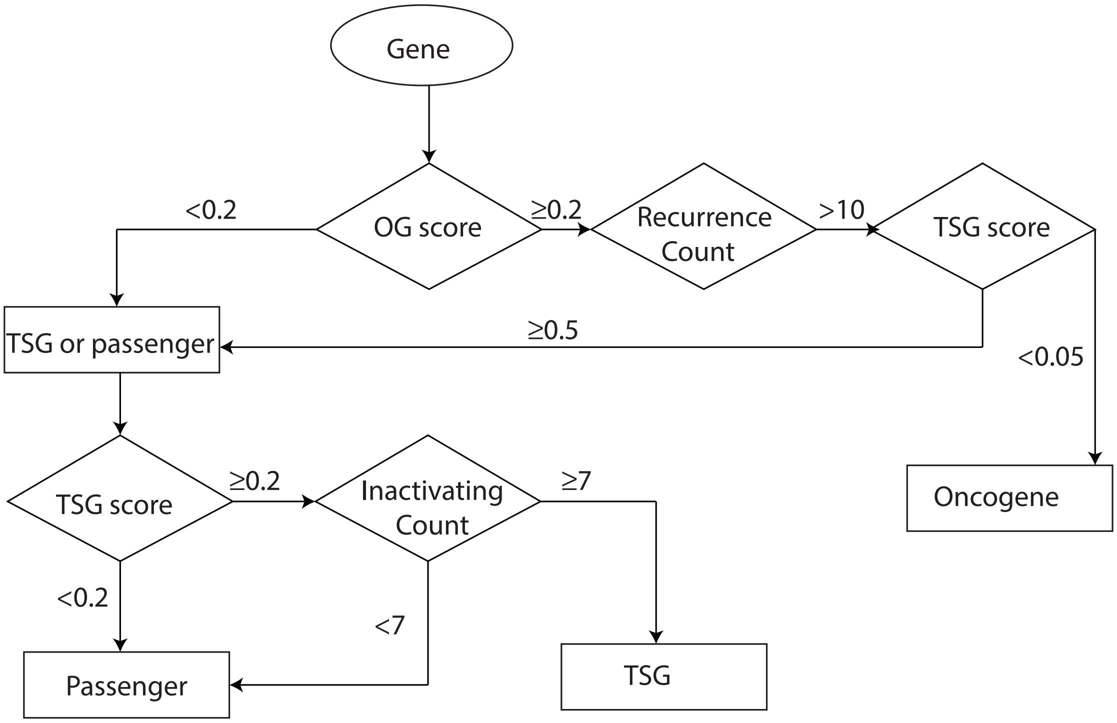

Figure S6. Decision tree underlying 20/20 rule. Each gene is input into the tree and Oncogene (OG) and Tumor suppressor gene (TSG) score computed (Online methods). Thresholds of each score and the numerator of the OG score (Recurrence count) and TSG score (Inactivating count) are used to determine whether a gene is an OG, TSG or passenger.

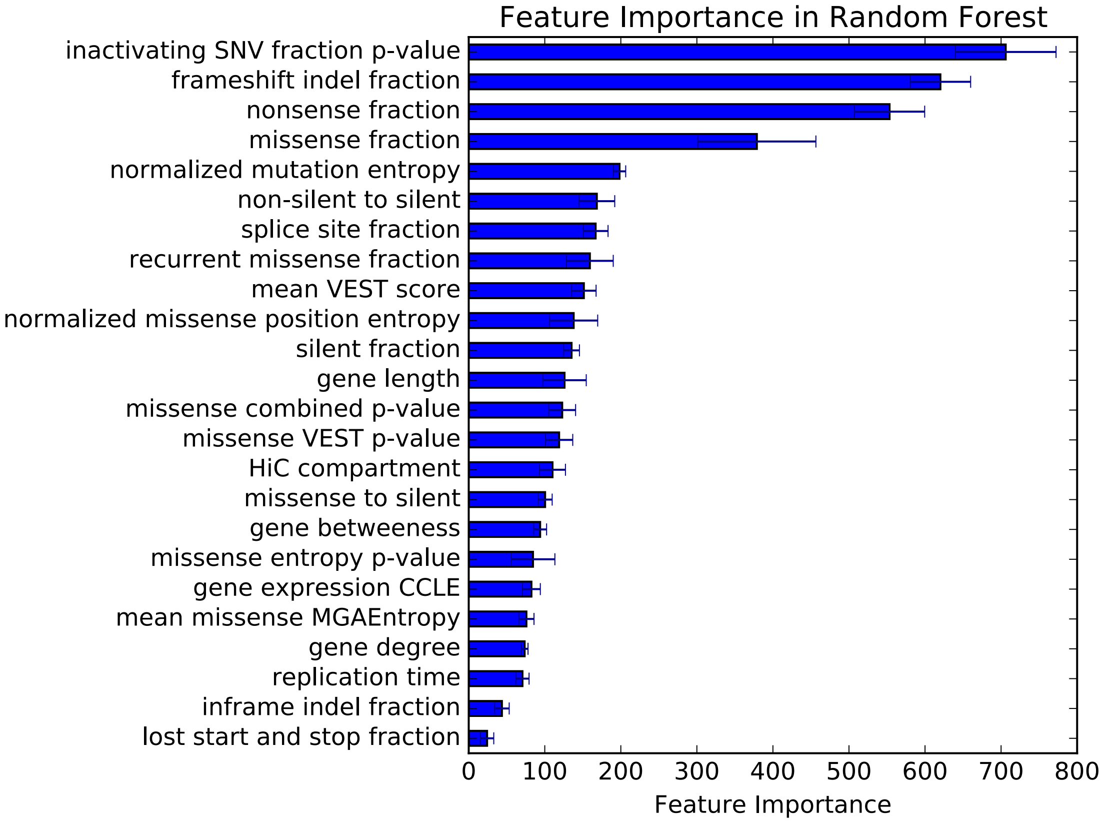

Figure S7. Random forest feature importance ranking for the 24 predictive features. The mean decrease in Gini index is plotted for each feature. Error bars indicate standard deviation when feature importance calculation was repeated on 10 different cross-validation partitions. SNV = single nucleotide variant. VEST = Variant Effect Scoring Tool [13]. HiC = 3D chromatin interaction capture [14]. CCLE = Cancer Cell Line Encyclopedia [15]. MGAEntropy = Shannon entropy in column of a vertebrate genome 46-way alignment corresponding to location of the mutation [16].

Table S1. Features used in 20/20+. Features use mutations that are small somatic variants, including single base substitutions and small insertions/deletions. CCLE = cancer cell line encyclopedia. SNV = single nucleotide variant. SNVBox = database of features of single nucleotide variants. Biogrid = database of gene networks.

Name	Source	Description
silent fraction	Calculated from mutations	Fraction of mutations that are silent mutations
nonsense fraction	Calculated from mutations	Fraction of mutations that are nonsense mutations
splice site fraction	Calculated from mutations	Fraction of mutations that are 2bp splice site mutations
missense fraction	Calculated from mutations	Fraction of mutations that are missense mutations
recurrent missense fraction	Calculated from mutations	Fraction of mutations that are recurrent missense
frameshift indel fraction	Calculated from mutations	Fraction of mutations that are frameshift indel mutations
inframe indel fraction	Calculated from mutations	Fraction of mutations that are inframe indel mutations
lost start and stop fraction	Calculated from mutations	Fraction of mutations that are either lost start or lost stop mutations
normalized missense position entropy	Calculated from mutations	See Materials and Methods
missense to silent	Calculated from mutations	Ratio of missense to silent mutations. A pseudo count is added to silent mutations to avoid divide by zero.
non-silent to silent	Calculated from mutations	Ratio of non-silent to silent mutations. A pseudo count is added to silent mutations to avoid divide by zero.
normalized mutation entropy	Calculated from mutations	Normalized entropy score (see Materials and Methods). Missense mutations are binned together based on codon position (see Materials and Methods). Each silent mutation is regarded in its own bin. Potentially inactivating mutations (nonsense, splice site, lost stop, and lost start) mutations are grouped into a single bin.
mean missense MGAEntropy	Calculated from mutations. MGAEntropy scores obtained from SNVBox [16].	Mean MGAEntropy score for missense mutations [16]. MGAEntropy for a missense mutation is the entropy of the column for a protein-translated version of UCSC's 46-way vertebrate alignment
mean VEST score [13]	Calculated from mutations	Mean score. Score for missense mutations are taken as the VEST score, silent mutations receive a score of 0, and other mutations receive a score of 1.
inactivating SNV p-value	Calculated from mutations	See Materials and Methods. SNV=single nucleotide variant.
missense entropy p-value	Calculated from mutations	See Materials and Methods
missense VEST p-value [13]	Calculated from mutations	See Materials and Methods
missense combined p-value	Calculated from mutations	Combined p-value composed of missense entropy and missense VEST p-value using Fisher's method
gene degree	BioGrid [17]	Number of other genes that are connected in the BioGrid interaction network
gene betweenness centrality	BioGrid [17]	Fraction of shortest paths that pass through a gene's node in the BioGrid interaction network
gene length	Longest SNVBox transcript [16]	CDS length of reference transcript
expression CCLE	MutsigCV [8]	Average expression of a gene in the Cancer Cell Line Encyclopedia [15]
replication time	MutsigCV [8]	DNA replication time during cell cycle
HiC compartment	MutsigCV [8]	HiC measures open vs closed chromatin [14]

Table S2. Eight evaluated methods and p-value of overlap of predicted drivers and Cancer Gene Census genes. The overlap is highly significant for all methods (one-tailed Fisher's Exact Test).

Method	P value
TUSON	<2.2e-16
20/20+	<2.2e-16
MutsigCV	<2.2e-16
MuSiC	<2.2e-16
OncodriveClust	<2.2e-16
OncodriveFM	<2.2e-16
OncodriveFML	<2.2e-16
ActiveDriver	5.50E-07

References

1 Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic acids research* **40**, e169, doi:10.1093/nar/gks743 (2012).

2 Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238-2244, doi:10.1093/bioinformatics/btt395 (2013).

3 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).

4 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* **31**, 3812-3814 (2003).

5 Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research* **39**, e118, doi:10.1093/nar/gkr407 (2011).

6 Speir, M. L. *et al.* The UCSC Genome Browser database: 2016 update. *Nucleic acids research* **44**, D717-725, doi:10.1093/nar/gkv1275 (2016).

7 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310-315, doi:10.1038/ng.2892 (2014).

8 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

9 Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501, doi:10.1038/nature12912 (2014).

10 Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948-962, doi:10.1016/j.cell.2013.10.011 (2013).

11 Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. & Williams Jr, R. M. *The American soldier: adjustment during army life. (Studies in social psychology in World War II, Vol. 1.).* (Princeton Univ. Press, 1949).

12 Reimand, J. & Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular systems biology* **9**, 637, doi:10.1038/msb.2012.68 (2013).

13 Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC genomics* **14 Suppl 3**, S3, doi:10.1186/1471-2164-14-S3-S3 (2013).

14 Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268-276, doi:10.1016/j.ymeth.2012.05.001 (2012).

15 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

16 Wong, W. C. *et al.* CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics* **27**, 2147-2148, doi:10.1093/bioinformatics/btr357 (2011).

17 Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic acids research* **43**, D470-478, doi:10.1093/nar/gku1204 (2015).