COHORT INTEGRATION (CI) THEORY

A. Differentiation of the genotype-phenotype relationship (Evolutionary Action)

We approach the problem of quantifying meaningful mutations in cancer populations from first principles. The genotype-phenotype relationship is modeled as a potential function (f) that gives to each genotype location of the fitness landscape (γ) a phenotype fitness value (φ) (Fig. 1A) (Katsonis and Lichtarge 2014):

$$f(\gamma) = \varphi \quad (1).$$

To first order approximation, a genetic variant ($\Delta\gamma$) displaces the genome in the landscape, causing a change in the fitness potential ($\Delta\varphi$). We call this fitness effect of the variant its Evolutionary Action (EA), and model by differentiating Eq. 1 to obtain (Fig. 1B):

$$f'(\gamma) \cdot \Delta\gamma \approx \Delta\varphi = EA \quad (2),$$

where f' is the gradient (i.e., derivative) of the fitness potential f. To evaluate Eq. 2 for missense mutations, we approximate the two terms on the right-hand side. Namely, we quantify the fitness sensitivity of a sequence position to variations, f', with ranks of the Evolutionary Trace algorithm (Lichtarge et al. 1996; Mihalek et al. 2004), and we quantify the magnitude of a mutation $\Delta\gamma$ with ranks of the amino acid substitution odds (Fig. 1B-C) (Katsonis and Lichtarge 2014). Multiplication and normalization yield $\Delta\varphi$, a continuous Evolutionary Action (EA) score that ranges from 0 to 100, where 0 means no functional effect and 100 means maximal disruption with total loss of function. Intermediate values, such as 50±20, suggest significant perturbations that modify functionality rather than totally disabling it.

B. Evolutionary Action yields the fitness effects of genetic variants

In order to assess whether EA predicts the effect of coding variants in the context of cancer, as suggested before in *TP53* (Neskey et al. 2015; Osman et al. 2015a; Osman et al. 2015b), we tested its agreement with experimental and clinical data on *MLH1* (Raevaara et al. 2005), *BRCA1*, and *BRCA2* (Spurdle et al. 2012) mutations (Supplementary Fig. 1 A-C, p-value<0.01, Mann-Whitney *U* test). EA

could separate benign from deleterious variants better than alternative methods (Supplementary Fig. 1 D-F). Moreover, thousands of *TP53* mutations measured for their deleterious effect on p21WAF1 promoter response (Kato et al. 2003) showed that EA made 40-70% fewer false-positive predictions (Miosge et al. 2015) than leading methods, depending on the threshold (Supplementary Fig. 1 G-H). These data show that EA evaluates mutational impact with greater specificity than the other methods. This performance is consistent with objective assessments by independent judges who alone possess the experimental gold standard data (Katsonis and Lichtarge 2017; Xu et al. 2017; Zhang et al. 2017; Katsonis and Lichtarge 2019).

C. Cohort Integration (CI) of the fitness effects

Next, we complemented this differential model of mutations as individual micro-steps in the fitness landscape with an integration model, by summing all coding mutations in a cohort of individuals who share a trait of interest, namely cancer in this study. In theory, integration should reverse differentiation and yield back the genotype and phenotype relationship between genotype and phenotype, i.e. integrating Eq. 2 should solve f. We proceed as follows. Individuals from a cancer cohort experienced variants that, in aggregate, resulted in a displacement away from the normal equilibrium location in the fitness landscape. The variants driving this displacement must necessarily be impactful in some affected individuals (Martincorena et al. 2017), so that genes with unexpectedly impactful variants in the cohort may be drivers of that phenotype. For cancer, tumor suppressor genes would be expected to harbor inactivating variants with large EA scores, oncogenes should harbor less impactful function altering variants, i.e., gain-of-function variants with intermediate EA scores, and passenger mutations in innocent genes should appear random and unbiased. To find these differences we may integrate EA:

$$\int f'(\gamma)\, d\gamma = \varphi(C_j) \qquad (3),$$

where the integral is performed over all the somatic mutations in all genes from all cancer patients in a cohort of individuals, C_j, with cancer type j. In practice, (3) is evaluated numerically through summation:

$$\sum^{somatic\ mutations\ of\ C_j} f'(\gamma) \cdot d\gamma - \sum^{random\ mutations} f'(\gamma) \cdot d\gamma = \varphi(C_j) \qquad (4),$$

where the negative term is an integration constant, chosen here to zero out random passenger mutations in unrelated to cancer genes. Next, since genes are the functional units of genotype, we rearrange the summations in Eq. 4 gene by gene, denoted by the index k:

$$\sum_{C_j} f_k{}'(\gamma) \cdot d\gamma - \sum_{C_j}^{random\ mutations} f_k{}'(\gamma) \cdot d\gamma = \varphi_k(C_j) \quad (5).$$

Eq. (5) states that a given gene k will make no contribution to the cancer phenotype $\varphi_k(C_j)$ when its somatic mutations in the patient cohort C_j are no different than random. However, if gene k is a cancer driver gene, it should have a non-zero value in (5) and will contribute to $\varphi(C_j)$ (Fig. 1D). Therefore, the cohort integration approach detects the gene-by-gene selective pressure associated to the trait common in C_j (Fig. 1D, E).

SUPPLEMENTARY METHODS

Experimental and clinical assessment of variant impact in cancer-associated genes. The 28 *MLH1* missense coding variants and their pathogenic significance were obtained from Raevaara et al. (Raevaara et al. 2005). The 107 *BRCA1* missense coding variants and the 90 *BRCA2* missense coding variants and their pathogenicity were collected from the BRCA Exchange database (http://brcaexchange.org/). The transactivity of $p21^{WAF1}$ response-element for the 2,314 human *TP53* point mutants assayed in yeast were obtained from Kato et al. (Kato et al. 2003).

Coding variations from the 1000 Genomes Project and The Cancer Genome Atlas (TCGA). The 23,810,164 human coding missense germline variations were obtained from the 1000 Genomes Project, phase 3 (http://www.internationalgenome.org/). Cancer somatic variants used in the primary analyses presented in this manuscript were obtained from The Cancer Genome Atlas (TCGA) data portal (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp) in January 2015, when variants from 5,996 tumors across 20 cancers were available. The variant calls may vary between genome sequencing centers (GSCs) due to different exome capture technologies, filtering strategies, and variant calling methods each center employed. To acquire the highest confident variant calls, for the genomes that were analyzed by more than one GSC, only the calls that agreed in a majority of GSCs were considered for analysis. All variants were then re-annotated using ANNOVAR package (Wang et al. 2010). 645,359 missense, 51,759 nonsense, 272,468 silent mutations, and 43,584 frameshift indels or other somatic point mutations were obtained. Updated MC3 TCGA variant information was downloaded on April 24, 2020 from National Cancer Institute Genomic Data Commons (https://gdc.cancer.gov/about-data/publications/mc3-2017). MC3 variants were similarly re-annotated with ANNOVAR.

Evaluation of functional impact of coding variants. The functional impact scores of coding variants in *MLH1*, *BRCA1*, and *BRCA2* as evaluated by PolyPhen-2 (PPH2), MutationAssessor (MA), and CADD were obtained from the corresponding servers (PPH2: http://genetics.bwh.harvard.edu/pph2/; MA: http://mutationassessor.org; CADD: http://cadd.gs.washington.edu/). For CADD scores, we used the TransVar annotator (Zhou et al. 2015) (http://bioinformatics.mdanderson.org/transvarweb/) to get the

genome position and nucleotide change of each variant and then obtained the scaled CADD score (v1.3). For missense variants matching multiple genome positions and nucleotide changes, we calculated the average CADD score to represent the functional impact of the variant.

Comparison of Cohort integrals of the coding variants. To compare the effects of coding variants and simulated random nucleotide changes on fitness, we performed the two-sample, one-sided Kolmogorov–Smirnov (K-S) test and calculated the p-values, i.e., the lowest attainable significance levels at which the null hypothesis that the two samples are drawn from the same distribution can be rejected. K-S is a nonparametric test of the equality of continuous, one-dimensional probability distributions, measuring the maximum distance between the empirical distribution functions of the two samples. Therefore, the two-sample K-S test is sensitive to differences in both the location and shape of the empirical cumulative distribution functions. Because the K-S test is sensitive to sample size, for large datasets including The 1000 Genomes Project and TCGA, instead of using all coding variants, we randomly sampled 1000 variants from the set of variants and performed the one-sided K-S test against 1000 simulated random variants. These analyses were repeated 1000 times to obtain average p-values.

Random nucleotide changes and cancer-specific mutational signatures. To assess the impact of cancer-specific mutational signatures on the EA distributions, we calculated the frequency of each of the 12 nucleotide substitutions (e.g. A to C) for the somatic mutations found in each of the 20 cancer types. Then, we obtained N random missense nucleotide changes (N was chosen to be 50, 100, 200, 400, 800, 1600, 3200, and 6400, in 8 independent tests) and N missense nucleotide changes randomly weighted according to the mutational signature of each cancer type, throughout the human genome. The corresponding EA distributions of random versus the randomly-weighted missense variants were compared according to the two-sided, two-sample K-S test. Each test was repeated 50 times to obtain better representation of the expected p-value. None of the experiments resulted in significant p-values.

Evaluation of the performance of CI. We evaluated the performance of CI against 10 other cancer gene identification methods on the same set of input cancer genomes by adapting, modifying, and enhancing previously suggested guidelines (Tokheim et al. 2016). The criteria were: (1) overlap with COSMIC

Cancer Gene Census (*CGC Overlap*); (2) overlap with predictions from the other methods (*Method Consensus*); (3) area under the Receiver Operating Characteristic curve (*AUC-ROC*) ; (4) area under the Precision-Recall curve (*AUPRC*); (5) discovery consistency over random two-way splits of the cohort from the top one to 100 genes (*Consistency*); (6) deviation from the expected uniform p-value distribution after removal of common predicted drivers (*p-value deviation*). Discovery gene lists for 2020+, TUSON, OncodriveFML, MutsigCV, MuSiC, OncodriveClust, OncodriveFM, and ActiveDriver were obtained from Dataset_S04 of Tokheim et al 2016. To generate the discovery gene list and additional performance metrics for dNdScv, a publicly available script was utilized (downloaded from GitHub on September 10, 2021, https://github.com/im3sanger/dndscv). The MutPanning discovery gene list and performance metrics were generated on and performed according to the guidelines provided on the GenePattern plaform (https://www.genepattern.org/)(Dietlein et al. 2020). The Python library scikit-learn (version 0.24.1) was used to compute *AUC-ROC* and *AUPRC* values. For the *AUC-ROC* and *AUPRC* analyses, the true positive cancer genes were defined as the current Cancer Gene Census COSMIC Tier 1 somatic cancer genes (downloaded June 30, 2020; Table S2) and the predictions were restricted to genes ranked by all 10 methods plus the CI and CI (with INDEL).

Update of the Cancer Gene Census in COSMIC.

To avoid using recently added cancer genes as true negatives in the Receiver Operating Characteristic analysis, we downloaded a more recent version of the COSMIC Cancer Gene Consensus on June 30, 2020. Inclusion criteria for cancer associated genes was defined by the presence of somatic mutations (missense, frame shifts, nonsense, splice site). Genes annotated as translocations, large deletions, amplifications, or other mutations were excluded unless they also met the somatic mutation inclusion criteria. We also excluded the gene *TENT5C,* because it was not ranked by all methods. This resulted into selecting 238 true positive cancer genes (see Table S2).

GSEA Hallmark Gene Sets enrichment analysis. The GSEA Hallmark Gene Sets (Liberzon et al. 2015) were obtained from the Molecular Signatures Database (MSigDB) (Subramanian et al. 2005). The enrichment of the 460 candidate driver genes for the GSEA Hallmark Gene Set was calculated with the

hypergeometric test against all genes. False discovery rates (q-values) were calculated with the Benjamini–Hochberg procedure to correct for multiple testing (Storey and Tibshirani 2003).

Ingenuity Pathway Analysis (IPA). The 460 candidate driver genes were analyzed by the canonical pathway analysis and the molecular and cellular functions analysis of the IPA software (Kramer et al., 2014). The significance of the association between the candidate driver genes and the canonical pathways or the biological functions and diseases was measured by a Fisher's exact test. The multiple-hypothesis testing was corrected with the Benjamini–Hochberg procedure (Storey and Tibshirani 2003).

Similarity tree of cancer types and the candidate driver genes. The dendrogram of cancer types was based on the Jaccard distance of candidate driver genes, which is 1 - intersection over union of the candidate driver gene sets. The dendrogram of candidate driver genes was based on the CI q-values and it was calculated for 56 candidate cancer genes that were identified in multiple cancer types. Both dendrograms were constructed using the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) hierarchical clustering method.

Gold-standard gene set. A gold-standard cancer gene list (Table S4) is collected from: (a) the Cancer Gene Census in COSMIC database (v79), considering only the genes with mutation types as somatic, missense and nonsense, (b) the cancer genes predicted by TUSON explorer (Davoli et al. 2013) (manual confidence=4, Table S7A, S7B from the paper) (c) the 127 significantly mutated genes in all cancer types and 12 separate cancer types from Kandoth *et al*. (Kandoth et al. 2013) (Table S4 from the paper), (d) the 125 driver genes from Vogelstein *et al*. (Vogelstein et al. 2013) (Table S2A from the paper), and (e) the 260 predicted cancer genes from Lawrence *et al*. (Lawrence et al. 2014) (Table S2 from the paper). For gold-standard tumor suppressors and oncogenes (Table S11), we collected 54 "gold-standard" tumor suppressor genes and 18 "gold-standard" oncogenes that were identified by CI and were consistently agreed by two authoritative studies (Davoli et al. 2013; Vogelstein et al. 2013), or annotated as tumor suppressors or oncogenes in the Cancer Gene Census in COSMIC database (v79). Genes that were annotated as "oncogene/TSG" or marked as "germline", "A (amplification)", "D (large deletion)", "T (translocation)", or "O (other)" in CGC were excluded.

Support of cancer association (1): PubMed associations to cancer.

PubMed association to cancer was determined by querying PubMed for each gene name along with the terms "cancer," "protein," and "gene." The enrichment of a gene group for PubMed association was calculated with a hypergeometric test against the PubMed association of all genes with the same terms.

Support of cancer association (2): Graph-based Information Diffusion (GID).

We applied graph-based information diffusion (GID) (Shin et al. 2007; Venner et al. 2010; Lisewski et al. 2014) to determine the association of a given candidate to known cancer genes in the protein-protein interaction network. The protein-protein network was obtained from the STRING database (Szklarczyk et al. 2015)(https://string-db.org/, human experimental network, v10). Gold-standard cancer genes (Table S4) that were present in the network were labeled (437 genes), and those labels were diffused to the genes in the network to compute the cancer association score,

$$f = (I + \alpha L)^{-1} y,$$

where $f = \{f_1, \ldots, f_n\}$ is the score vector of each gene, I is the identity matrix, α is defined as $1/\|L\|_1$(Lisewski and Lichtarge 2010), L is the symmetric normalized Laplacian of the protein-protein network, and y is the label vector where the gold-standard cancer genes were set as 1 and the rest of genes were set as 0. Greater f score represents stronger associations of the given gene to known cancer genes. Since genes with higher connectivity tend to acquire greater f, we further normalized f based on the network connectivity. For each gene we sequentially selected "control genes" of similar connectivity, starting from the genes with the same connectivity, and then with connectivity increment or decrement by 1 each time, until at least 50 "control genes" were collected. The f of these "control genes" (f_c) for the given gene were then used to compute the normalized score of the gene, a

$$s_{gene} = \frac{(f_{gene} - \mu_c)}{\sigma_c},$$

where μ_c is the mean of f_c, and σ_c is the standard deviation of f_c. We considered the candidate genes with $s_{gene} \geq 1$ as strongly associated with known cancer genes.

Support of cancer association (3): *Measuring positive selection with d_N/d_S ratio.*

The nonsynonymous to synonymous rate (dN/dS) ratio was used to indicate the presence of selection on non-synonymous mutations, and values of dN/dS ratio above 1 indicate positive selection. The dN/dS ratio is calculated as the ratio of the number of nonsynonymous cancer somatic mutations per non-synonymous site to the number of synonymous cancer somatic mutations per synonymous site. The numbers of nonsynonymous and synonymous sites were quantified based on the simplistic assumption of unbiased codon usage and an equal rate for transition and transversions.

Confidence score of cancer association. The confidence score of cancer association was based on agreement across the five cancer gene lists selected for constructing the "gold-standard" cancer gene list and support from PubMed literature, network linkage, and dN/dS ratio. Genes that were listed in the "gold-standard" cancer gene list were considered as high confidence, and confidence scores were based on agreement upon the cancer gene lists: genes appearing in all five lists were assigned a confidence score of "8"; in four, were scored "7"; in three, were scored "6"; in two, were scored "5"; and those appearing in only one list received a score of "4". For genes not present in any of the five cancer gene lists, the confidence score was based on support from the literature (with more than 10 literature associations with cancer), association with known cancer genes in a network (GID normalized score, $s \geq 1$), and dN/dS ratio above 1. Genes supported by all three types of evidence were assigned a confidence score of "3"; supported by two were "2"; and supported by one were "1". The remaining genes were assigned a confidence score of "0."

CI oncogenes validation with the Avana CRISPR screen from the DepMap database. The data were downloaded from the DepMap database (https://depmap.org/) and consisted of the CRISPR (Broad Avana) screen data (Doench et al. 2016; Meyers et al. 2017), the cell lines in Avana table, and the merged mutations calls (coding region, germline filtered) from the CCLE (Barretina et al. 2012; Cancer Cell Line Encyclopedia and Genomics of Drug Sensitivity in Cancer 2015). Variants were scored using the EA equation and classified as "moderate EA" (30 <= EA score < 70) or "others" (Low EA variants 0 <= EA < 30, high EA variants 70 <=EA<100, nonsense variants, and other uncategorized variants) categories. Gene sets include: (1) random: a random gene set consists of 100 randomly chosen genes from the 17,632 genes

screened in the Avana set from all cell lines, (2) COSMIC: Oncogenes listed in the Cancer Gene Census (v79) in COSMIC database (Futreal et al. 2004), genes that were annotated as “oncogene/TSG” or marked as “germline”, “A (amplification)”, “D (large deletion)”, “T (translocation)”, or “O (other)” were excluded, (3) CI PAN: oncogenes identified by CI in Pan-cancer analysis, and (4) CI Indiv.: oncogenes identified by CI in individual cancers. For CI Indiv oncogene set, the mapping between TCGA cancer identifier and DepMap cell line primary disease annotation are as follows: BLCA ('Bladder Cancer'), BRCA ('Breast Cancer'), CESC ('Cervical Cancer'), COAD ('Colon Cancer'), GBM ('GBM/Brain Cancer'), HNSC ('Head and Neck Cancer'), LIHC ('Liver Cancer'), OV ('Ovarian Cancer'), SKCM ('Skin Cancer'), STAD ('Gastric Cancer'), THCA ('Soft Tissue/ Thyroid Cancer', 'Thyroid Cancer'), UCEC ('Endometrial Cancer'). We compared the shift in essentiality within each gene set between variants in the “moderate EA” category (n=358, 474, 432, 115 for random, COSMIC, CI PAN, CI Indiv., respectively) and variants in the “others” category (n= 1309, 832, 795, 122 for random, COSMIC, CI PAN, CI Indiv., respectively), and the significance of the shift in Ceres Scores was measured using the Mann-Whitney *U* (Wilcoxon rank-sum) test. To test the tissue specificity of CI prediction, we compared the shift in essentiality of CI Indiv. gene set between the target cancer cell lines (n=115) and other cell lines data (n=122). In both cases only variants in the moderate EA category were considered and significance was measured using the Mann-Whitney *U* test. See Table S12 for the Avana CRISPR screen data.

Functional testing of CUL3 overexpression. A cDNA encoding human CUL3 (NM_003590) in the pENTR™221 vector (Thermo Fisher Scientific) was subcloned into the lentivrial destination vector pLenti7.3/V5 (Invitrogen), co-expressing the Emerald Green fluorescent protein marker, through Gateway recombination and then verified by DNA sequencing. The resulting CUL3 lentiviral vector or control lentiviral vector expressing bacterial lacZ were transfected into 293FT packaging cells with ViraPower packaging mix plasmids (Invitrogen) to produce replication defective virus. Equivalent titers (pre-determined by tittering GFP expression) of CUL3 or lacZ expressing virus were inoculated into target cells by spinoculation after seeding cells the night before. Positively infected cells were purified by flow cytometry using the GFP marker (excluding the top and bottom 10% brightest or dimmest cells) and

allowed to recover for 3 days before plating into clonogenic assays or performing western blotting. For clonogenic assays, 1000 cells/well were seeded into 6-well plates and allowed to grow for 9 to 12 days before staining with crystal violet and counting colonies ≥ 50 cells with the assistance of ImageJ software. Antibodies used in western blots were anti-CUL3 (Cell Signaling Technology; #10450), anti-NRF2 (Cell Signaling Technology; #8882), and anti-total/cleaved PARP (Cell Signaling Technology; #9542). For cell cycle analysis, cells were trypsinized, washed in PBS, fixed at room temperature by addition of cold 70% ethanol, and stored at 4ºC before adding propidium idodide/RNAase and analyzing by flow cytometry.

Functional testing of DUSP16 overexpression. BT474, and HEK293 were obtained from ATCC and maintained in DMEM medium supplemented with high glucose (Life Technologies), 10% fetal bovine serum (FBS) (Thermo Fisher Scientific) with 50 units/ml penicillin and 50 µg/ml streptomycin (Life Technologies) in a humidified atmosphere of 5% CO2 at 37ºC. The DUSP16 plasmid (pLX302-MKP7-V5 puro) was purchased from Addgene (#87771). The DUSP16 ORF was amplified with forward primer GGGGACAAGTTTGTACAAAAAAGCAGGCTTCGAAGGAGATAGAACCATGGCCCATGAGATG ATTGGAACTCA and reverse primer GGGGACCACTTTGTACAAGAAAGCTGGGTTCTACGTAGAATCGAGACCGAGGA and cloned into an inducible lentiviral expression system (from Dr. Zhu Songyang(Kim et al. 2017)) using Gateway BP and LR Clonase II (Life Technologies) according to the manufacturer's instructions. The inducible lentiviral expression plasmid was amplified in DH5α E. coli (Life Technologies). Lentiviral particles were produced by transient transfection of pLend_empty and pLent_DUSP16 (1 µg) vectors along with packaging vectors pMD2.G (750 ng) and psPAX2 (250 ng) in 60% confluent HEK293T cells seeded in 6 well plates. The lentiviral supernatant stocks were collected at 24 h and 48 h, pooled and passed through 0.45 µm size filters. Lentiviral particles were incubated with BT474 and HEK293 cell lines for 24 h and then removed. Cells were subsequently selected by G418 (500 µg/ml) for 2 weeks to obtain stable cell lines. Selected cells were grown in media containing 100 µg/ml of G418 for further study. *Cell proliferation assay.* Stable BT474 and HEK293 cells were trypsinized and plated in 100 µl of growth medium at a density of 1,000 cells per well of a 96 well plate. Cells were incubated for 24h at 37ºC, added

with 0.5 µg/ml of doxycycline (Fisher BioReagents) and further incubated at 37◦C up to 3 days. Cell density was monitored with CCK-8 (Dojindo Molecular technologies, Inc) according to the manufacturer's protocols. Because doxycycline has a slight effect on growth rates (Moullan et al. 2015), cell counts were normalized to growth rates of doxycycline-treated control lines without an inducible vector. *Colony-formation assays.* Stable cells containing empty vector or a wildtype DUSP16 gene were seeded at a density of 1000 cells/well in 6 well plates in 3 ml of medium in the presence or absence of 1 µg/ml of doxycycline. Following incubation for 1-2 weeks, the colonies were stained with 0.5% crystal violet and scanned with an EPSON 4180 photo stylus scanner. *In vitro scratch assay and Cell migration assay.* Stable BT474_vehicle control and BT474_DUSP16 cells were grown in 12-well plates until about 70–80% confluency was reached at which point a 200 µL pipette tip was used to create a scratch/wound with clear edges across the width of a well. Wells containing cells were treated with 0.5 µg/ml doxycycline and photomicrographs were taken over a 48 h time period. An Olympus CK40 inverted microscope was used to measure and photograph the cell migration from the wound/scratch edge every 24 hr. After stable BT474 cells were trypsinized, cells (50,000 cells) were resuspended in DMEM containing 0.1% BSA with or without 0.5 ug/ml of doxycycline, added to the top of a Transwell (Corning, NY) migration chamber (24 well, 8 µm pore) and allowed to migrate for 18 h in the presence of DMEM containing 10% FBS. Residual cells were removed from top of the membrane with cotton ball and the cells on the underside of membrane were stained with crystal violet for 5 min and then visualized using a bright field microscope. *Expression of DUSP16, p53, PARP, pcJUN (S63), pJNK and JNK.* Cells were lysed with lysis buffer (40 mM Tris-Cl, pH 7.5, 150 mM NaCl, 0.6% CHAPS, 0.5 mM EDTA, 0.2% NP-40 and 1% glycerol) with protease inhibitor cocktail (Calbiochem) and phosphatase inhibitor cocktail (Sigma Aldrich). The protein concentration of cell lysate was measured with the Bio Rad reagent (Bio-Rad). The same amount of protein was boiled for 5 min in Laemmli sample buffer and separated by SDS–PAGE. Proteins were separated by 4–12% bis-Tris NuPAGE (Invitrogen) as indicated, transferred to nitrocellulose membranes, and immunoblotted with specified antibodies. Antibodies against DUSP16, p53, PARP, pcJUN (S63), pJNK and total JNK were purchased from Cell Signaling Technologies (#5523; #9982; #9532; #9261;

#4668; #9252, respectively). Anti-GAPDH was purchased from Thermo Scientific (MA5-15738).

Quantitative real time PCR (qRT-PCR). Stable cells containing empty vector or a DUSP16 gene were grown in 12 well plate until 70-80% of confluency and treated with 100 nM of PMA (phobol-12-myristate-13-acetate, from EMD MilliporeSigma) for 1 h. Cells were lysed with using TRIzol® Reagent (Life Technologies) and total RNA was extracted. RNA was reverse transcribed using the qScriptTM cDNA SuperMix (Quanta BioSciences) as per manufacturer instructions. qRT-PCR analysis was performed in a StepOne Plus system with SYBR Green (Applied Biosystems). Gene expression levels were normalized against β-actin and analyzed using the ΔΔCt method based on the manufacturer's manual. *Generation of DUSP16 depleted stable cell lines using CRISPR-Cas9.* The DUSP16 CRISPR-Cas9 constructs (pLenti-U6-sgRNA-SFFV-CAS9-2A-Puro) were purchased from ABM (K0443005). The lentiviral plasmids were amplified in DH5α E. coli (Life Technologies). Lentiviral particles were produced by transient transfection of each plasmid (1 µg) along with packaging vectors pMD2.G (750 ng) and psPAX2 (250 ng) in 60% confluent HEK293T cells seeded in 6 well plates. The lentiviral supernatant stocks were collected at 24 h and 48 h, pooled and passed through 0.45 µm size filters. Lentiviral particles were incubated with stable BT474-DUSP16 and HEK293-DUSP16 cell lines for 24 h and then removed. Cells were subsequently selected by puromycin (2 µg/ml) for 2 weeks to obtain stable cell lines. Selected cells were grown in media containing 0.5 µg/ml of puromycin for further study.

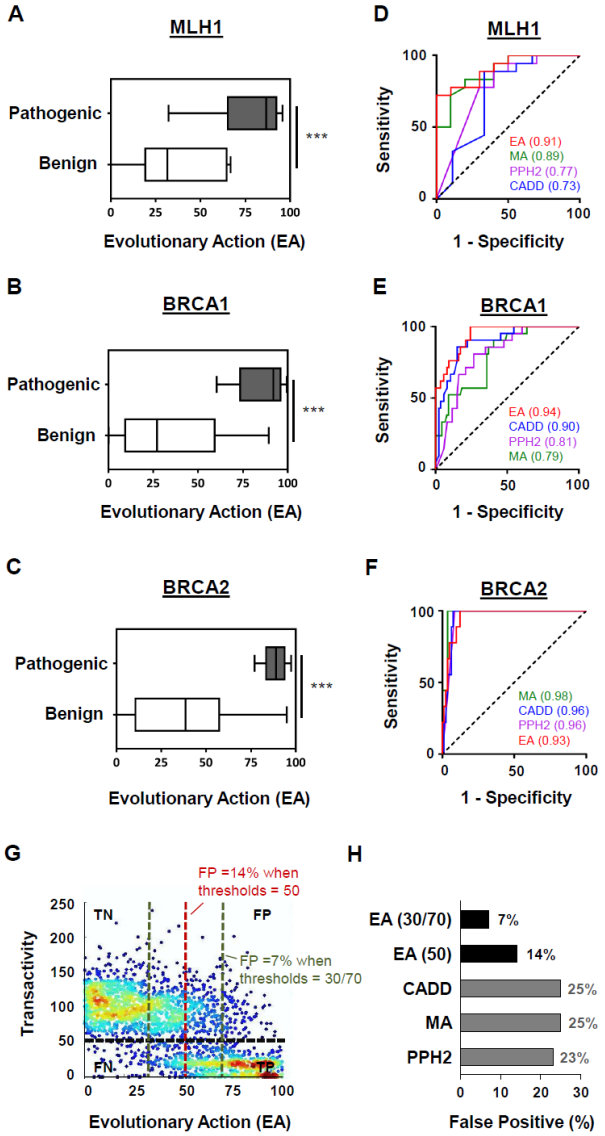
SOFTWARE AVAILABILITY

The source code of the Cohort Integration (CI) algorithm is available as a python script supplemental file: CohortInteg_SupplementalMaterial.py

Instructions for installation and running CI are available as a text supplemental file: README.txt.

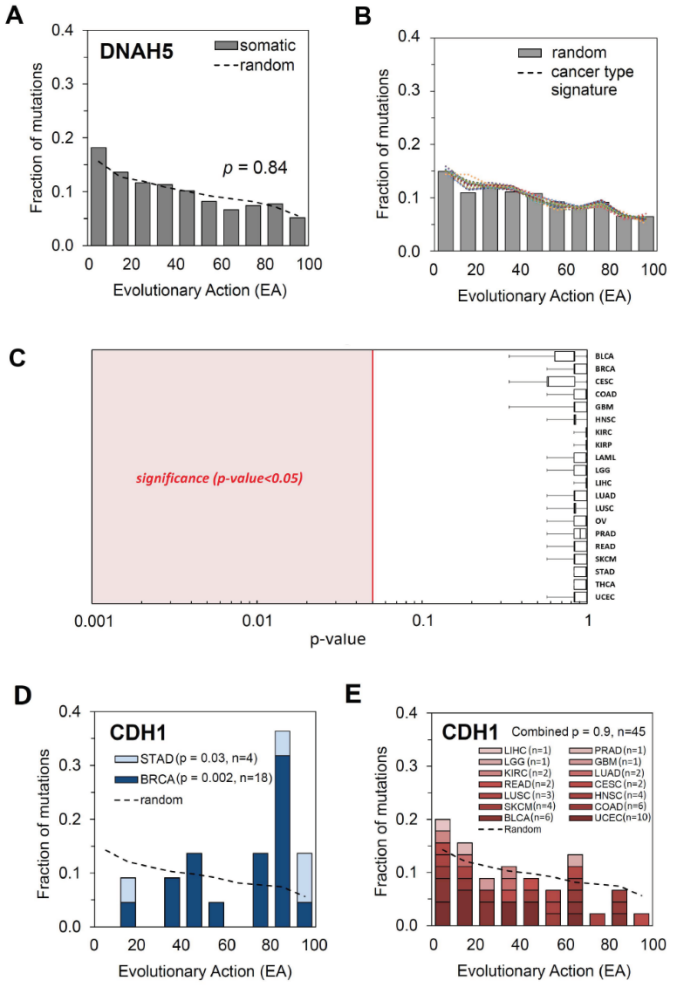
The method is also available at: http://cohort.lichtargelab.org/.

SUPPLEMENTARY FIGURES



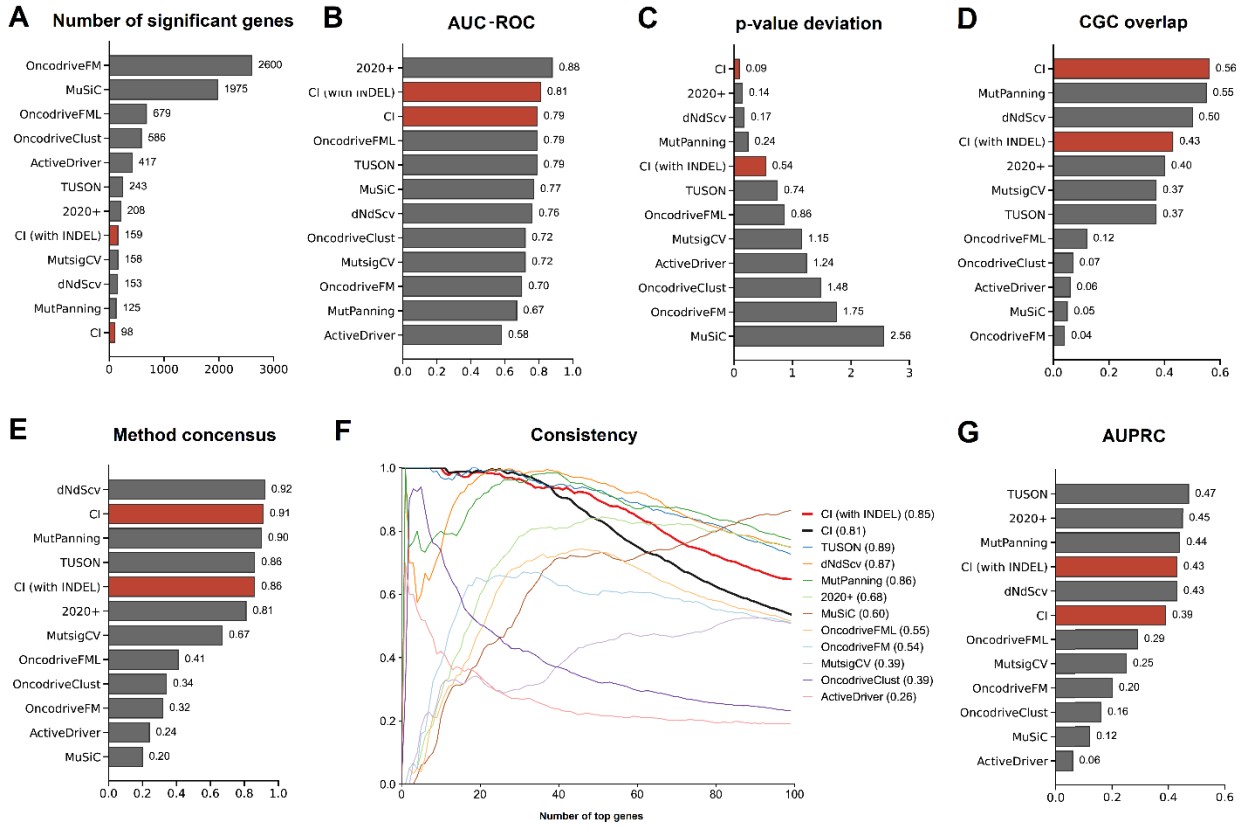
Supplementary Figure 1. Evolutionary Action (EA) correlates with the functional and clinical impact of mutations in cancer-associated genes. a-c, Boxplots comparing the EA scores of the pathogenic (grey) and benign (white) coding variants in **a,** MLH1, **b,** BRCA1, and **c,** BRCA2. **d-e,** The receiver operating characteristic curves for the separation of pathogenic and benign variants in **d,** MLH1, **e,** BRCA1, and **f,** BRCA2 by EA (red), PolyPhen-2 (Adzhubei et al. 2010) (PPH2, purple), MutationAssessor (Reva et al. 2011) (MA, green), and CADD (Kircher et al. 2014) (blue) scores. **g,** The transactivation activity of 2,314 human *TP53* point mutants assayed in yeast (*y*-axis) for the p21WAF1 promoter as function of the EA score, using a single cutoff at EA=50 (red), and dual cutoff of EA≤30 for benign mutations and EA≥70 for deleterious mutations (green) to calculate false positives (FP). **h,** The percentage of false positives of EA with single cutoff ("EA (50)") or dual cutoff ("EA (30/70)"), CADD,

MA, and PPH2.



Supplementary Figure 2. The distribution of EA scores of cancer somatic mutations. a, The distribution of EA scores for cancer somatic mutations in the non-cancer gene *DNAH5*. The dashed lines correspond to simulated random amino acid changes in DNAH5. **b-c,** The distributions of EA scores for 2,578 simulated random nucleotide changes that resulted in missense *TP53* substitutions (gray bars) and for 6,400 simulated missense substitutions generated with nucleotide changes following the mutational signatures (ratios of single nucleotide transitions) observed in each of 20 cancer types in **b**. Nucleotide changes following trinucleotide mutational signatures also resulted in non-significant p-values (data not shown). The two-sided Kolmogorov–Smirnov *p*-values for comparing each distribution of the cancer type signature with the distribution of simulated random nucleotide changes in **c**. The whisker plots represent the variability of 50 independent runs for each cancer type. **d-e,** The distribution of the EA scores of the CDH1 cancer somatic missense mutations compared to random CDH1 mutations (dashed line) in **d,** breast invasive carcinoma (BRCA, *p*-value=0.002) and stomach adenocarcinoma (STAD, *p*-value=0.03), and **e,**

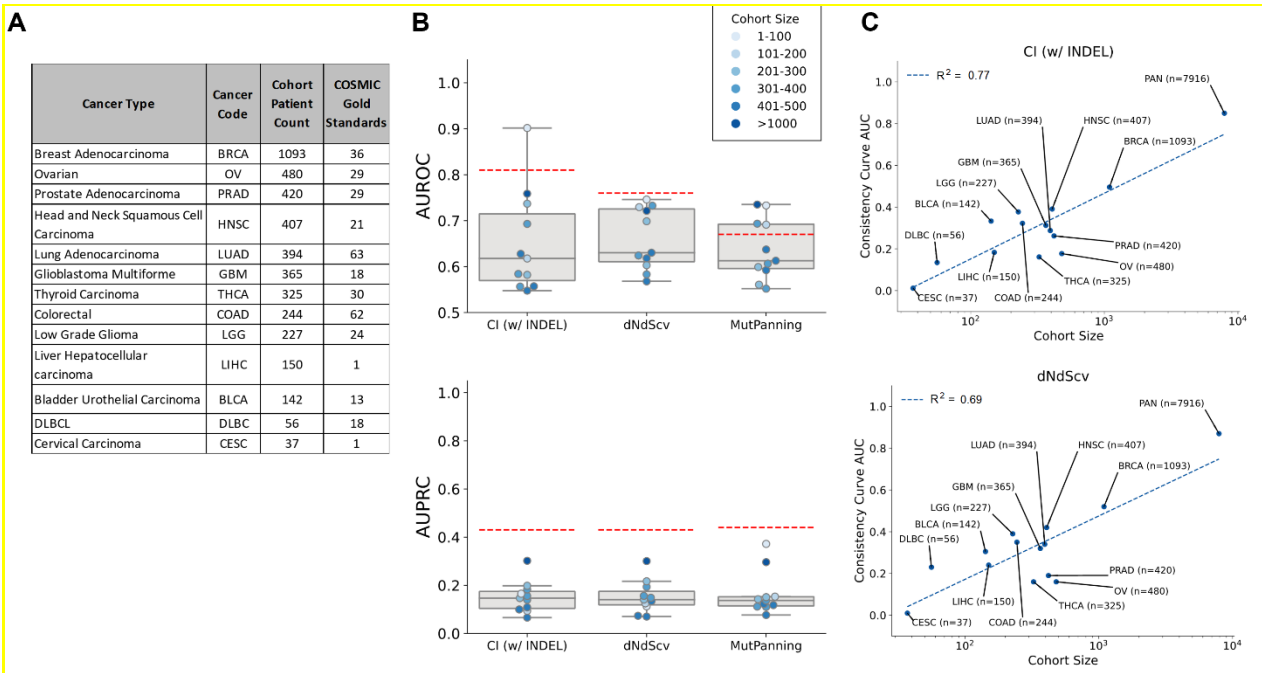
the other 18 cancer types (combined *p*-value=0.9). The *p*-values of cohort integral differences between the observed and simulated random mutations were calculated by the two-sample Kolmogorov–Smirnov test.



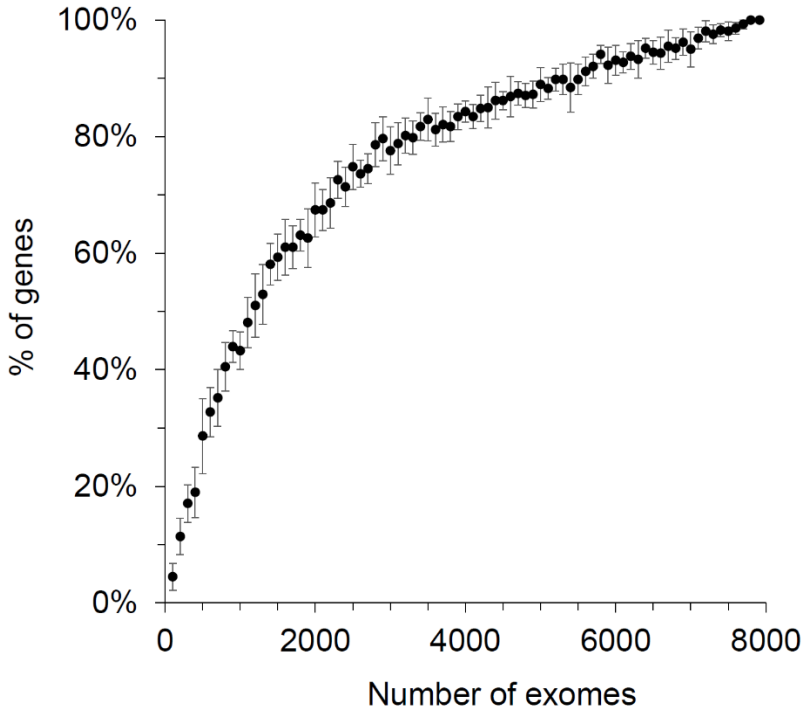
Supplementary Figure 3. Performance of CI (with and without INDEL) and ten state-of-the-art methods in identifying cancer-driving genes. a, The number of cancer genes predicted by each method. **b,** The area under the Receiver Operating Characteristic curve for each method. **c,** The deviation between observed and theoretical *p*-values for each method. **d,** The fraction of predicted cancer genes overlapped with the Cancer Gene Census (Forbes et al. 2016) for each method. **e,** The fraction of predicted cancer genes that were also predicted by one or more other methods. **f,** Top genes consistency as the number of the top genes varies from 1 to 100 for each method. The area under the curve of each evaluated method is given in the parentheses. **g,** The area under the Precision-Recall curve for each method.

A

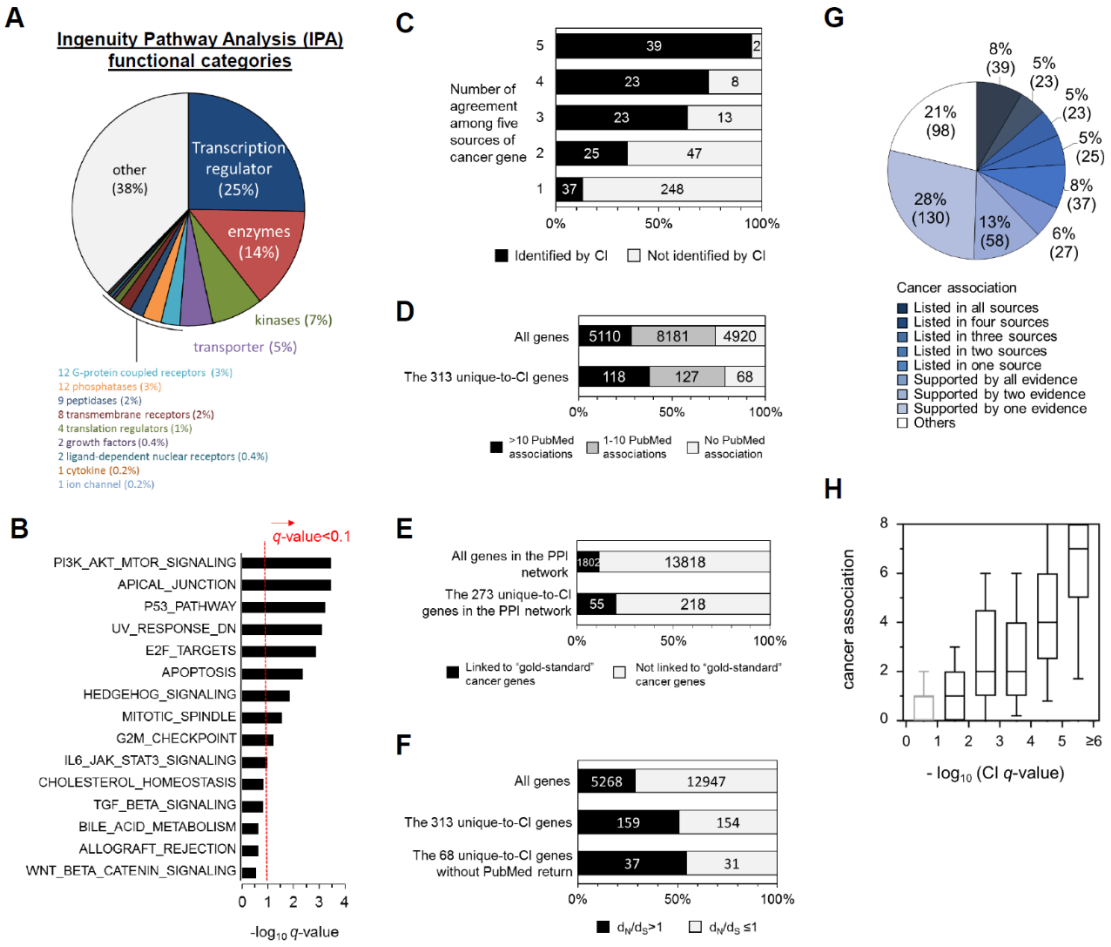
Cancer Type	Cancer Code	Cohort Patient Count	COSMIC Gold Standards
Breast Adenocarcinoma	BRCA	1093	36
Ovarian	OV	480	29
Prostate Adenocarcinoma	PRAD	420	29
Head and Neck Squamous Cell Carcinoma	HNSC	407	21
Lung Adenocarcinoma	LUAD	394	63
Glioblastoma Multiforme	GBM	365	18
Thyroid Carcinoma	THCA	325	30
Colorectal	COAD	244	62
Low Grade Glioma	LGG	227	24
Liver Hepatocellular carcinoma	LIHC	150	1
Bladder Urothelial Carcinoma	BLCA	142	13
DLBCL	DLBC	56	18
Cervical Carcinoma	CESC	37	1



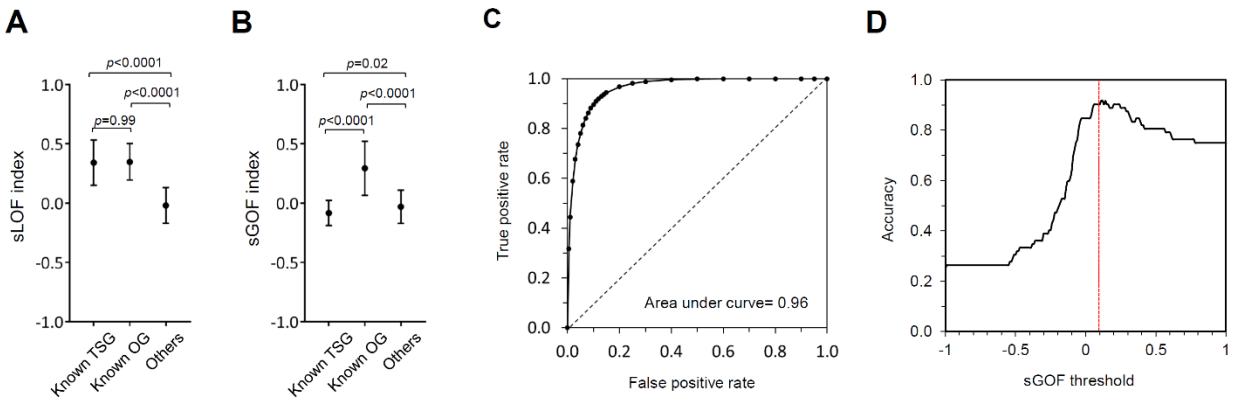
Supplementary Figure 4. Cancer type specific benchmark analysis. a, Patient counts and number of Tier 1 and Tier 2 COSMIC cancer specific genes for 13 cancer types found within the test population from Tokheim et al 2016. **b,** Cancer specific AUROC (top plot) and AUPRC (bottom plot) values for CI (w/ INDEL), dNdScv, and MutPanning presented in boxplot format showing the first, median, and third quartile values across 11 cancer types (LIHC and CESC have a single COSMIC cancer specific gene and were omitted from these analyses). Whiskers extend to include points within 1.5 times the interquartile range. Red dashed lines indicate method performance across the full test population (pan-cancer analysis shown in Figure 3A). Markers are shaded according to cancer cohort size. True positive genes were defined using COSMIC genes from panel a. **c,** Consistency Curve AUC values for CI (w/ INDEL) (top plot) and dNdScv (bottom plot) across 13 cancer types plus the full test population (pan-cancer = PAN). A logarithmic curve was fitted to the data and the correlation coefficient R^2 is shown in the plot legend.



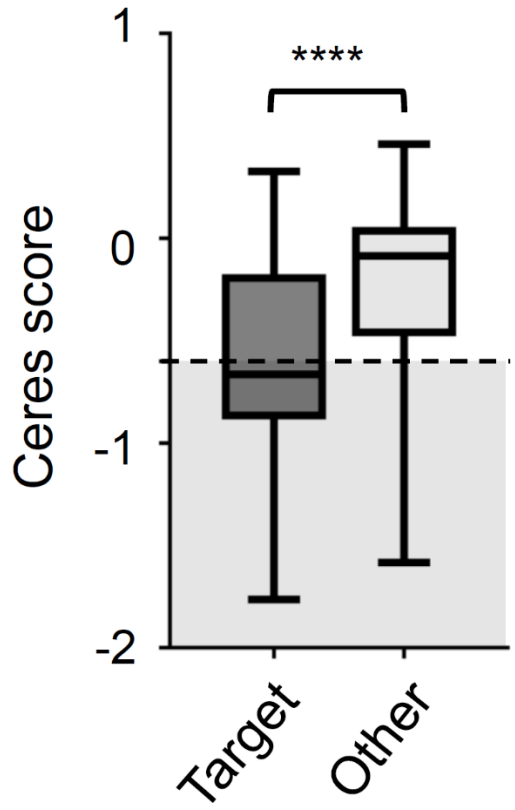
Supplementary Figure 5. Down-sampling analysis of CI. We performed 10 random samplings for increments of 100 tumor exomes (x-axis) and compared the driver gene predictions using these subsets to the predictions using the whole input exomes. Only genes that were reported as cancer associated by at least four of five reference sources were considered (n=58, see Table S4). The y-axis shows the percentage of true predictions from a given subset over true predictions from all available exomes.



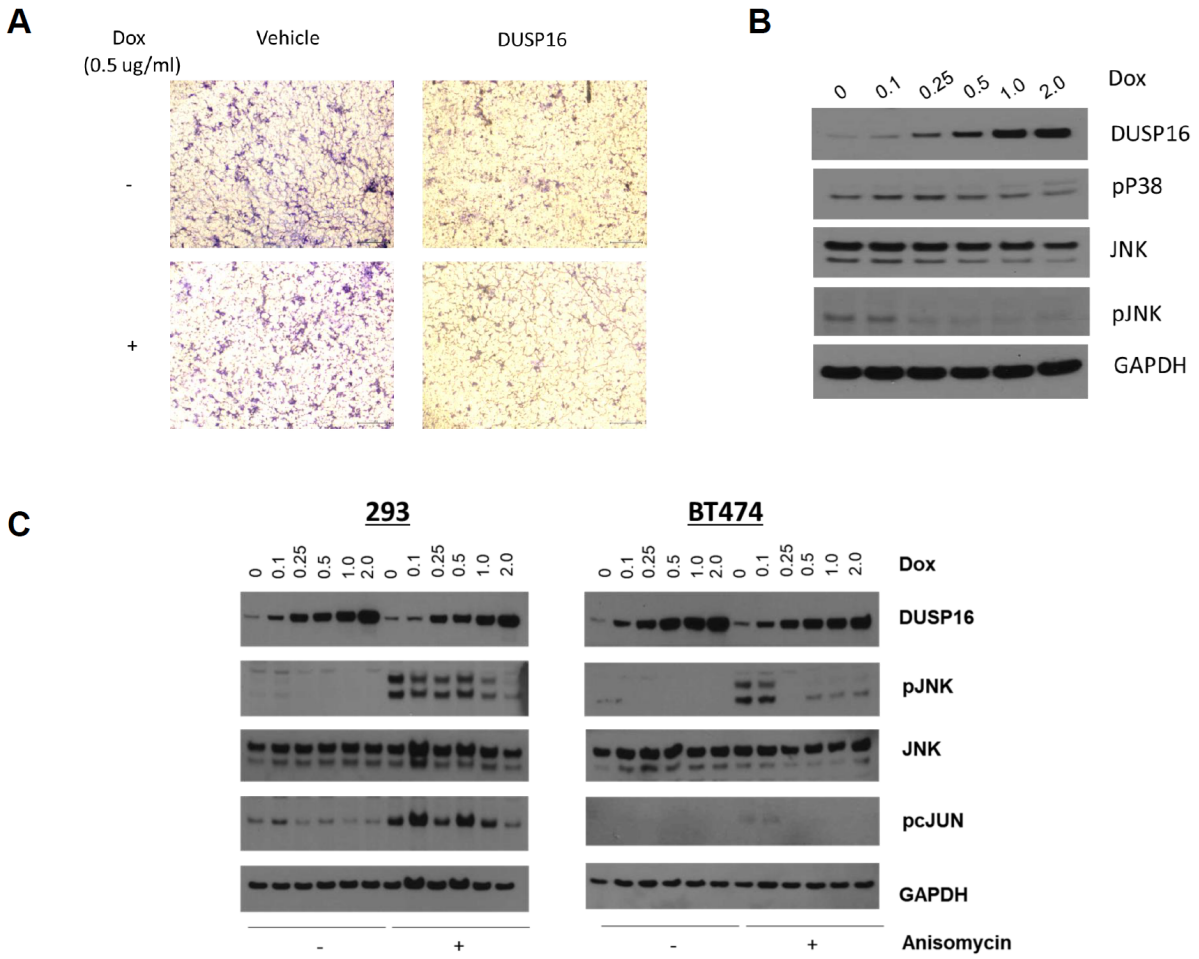
Supplementary Figure 6. 460 candidate cancer driver genes identified by CI. **a,** GSEA Hallmark Gene Sets enrichment of all 460 candidate genes identified in pan-cancer or in specific cancers. **b,** Ingenuity Pathway Analysis (IPA) functional categories of the 460 candidate cancer driver genes identified by CI. **c,** The fraction of gold-standard genes identified by CI, when the gold-standard genes were defined by the overlap of five or less references (Table S4). **d,** The fraction of genes with more than ten PubMed literature associations to cancer (black), one to ten PubMed associations (grey), and no PubMed association (white) for all genes and for the 313 that are unique to CI. **e,** The number of genes linked (black bar) or not (white bar) to the "gold-standard cancer genes" (Table S4) according to the STRING protein-protein interaction network, when we considered all STRING genes and the 273 that are unique to CI and have a STRING entry (40 genes did not match any STRING database entry). **f,** The number of genes with d_N/d_S value either greater than 1 (black bar) or not (white bar) for all genes, the 313 unique-to-CI genes, and the 68 unique-to-CI genes with no PubMed literatures association to cancer. **g,** Confidence of cancer association for the 460 candidate genes. **h,** Confidence of cancer association for different levels of CI q-value. Non-significant genes (q-value ≥ 0.1) are shown in grey. The box extends from the 25th to 75th percentiles, and the whiskers from the 10th to 90th percentiles.



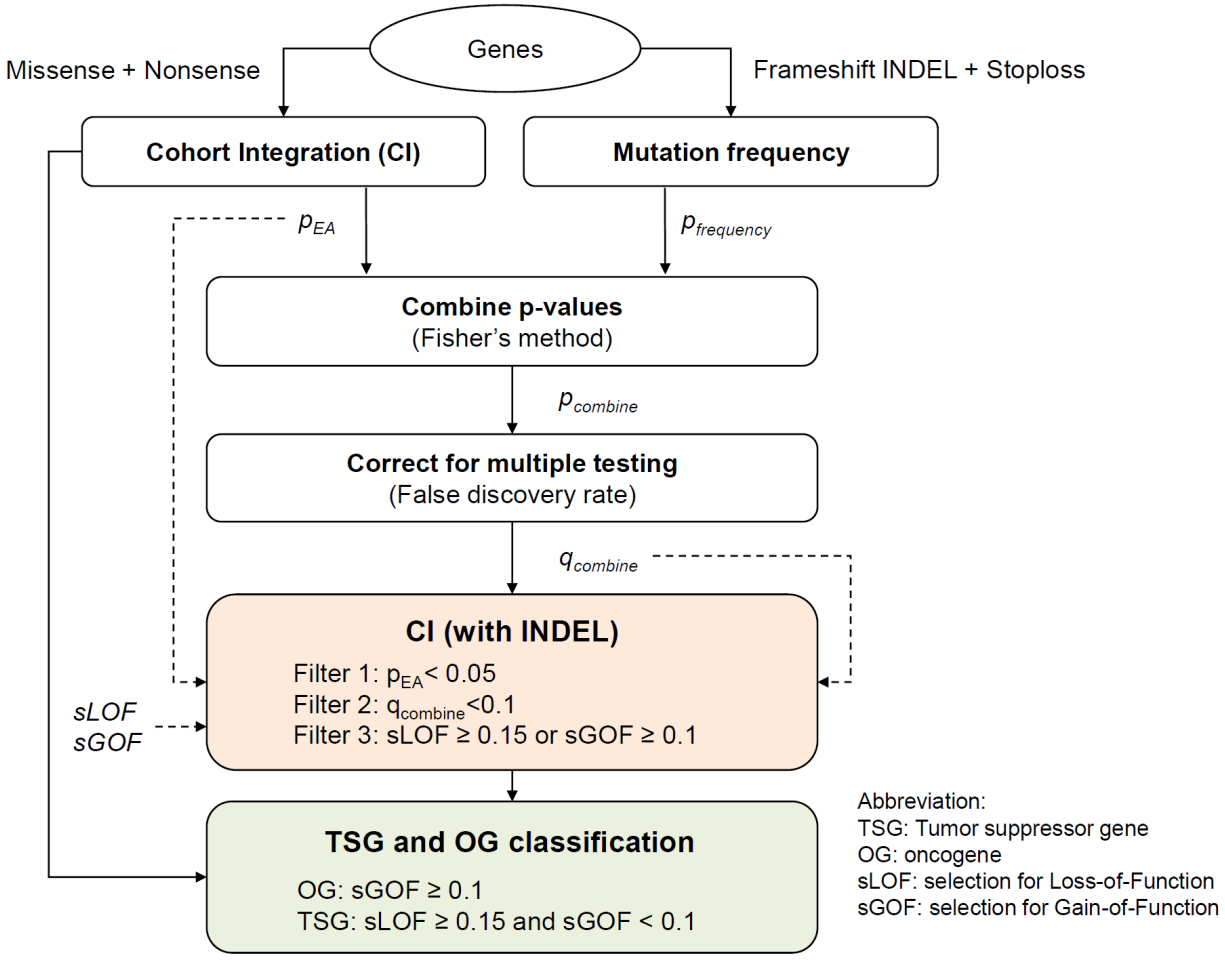
Supplementary Figure 7. Classifying tumor suppressors and oncogenes with sLOF and sGOF. a, The average sLOF index and **b,** the average sGOF index of 54 known tumor suppressor genes (TSG), 18 known oncogenes (OG), and 18,143 other genes. The p-values were calculated by Tukey's multiple comparisons test. The error bars represent the standard deviation. **c,** The receiver operating characteristic curve of classifying tumor suppressors and oncogenes with sGOF index. **d,** Accuracy of classifying tumor suppressors and oncogenes by a binary separation with sGOF threshold that varies from -1 to 1.



Supplementary Figure 8. Genome-scale CRISPR gene dependency screen (Tsherniak et al. 2017) validated tissue specificity for CI oncogenes. Cell lines harboring variants of moderate EA score in CI individual cancer genes showed a statistically significant shift toward essentiality in target individual cancer types than other cancer lines. Statistical significance was calculated with Mann–Whitney U test. ****: $p<0.0001$.



Supplementary Figure 9. Overexpression of DUSP16 inhibited cellular migration by dephosphorylation phosphor-JNK and its downstream target c-JUN. a, Overexpression of DUSP16 inhibited cellular migration in BT474 cells by Transwell migration assay. **b,** Effect of dose-dependent expression of DUSP16 on p38 and JNK phosphorylation. Increased expression inhibited JNK and p38 phosphorylation proportionally. Inhibition of JNK phosphorylation was more significant than p38 phosphorylation. **c,** Overexpression of DUSP16 inhibit JNK phosphorylation and c-JUN phosphorylation at basal and anisomycin stimulation.



Supplementary Figure 10. CI flowchart. A flowchart delineating the steps to identify cancer driver genes with CI and to classify tumor suppressors and oncogenes with sLOF and sGOF indices.

SUPPLEMENTARY TABLES

Table S1. Nucleotide substitution ratios observed in each of 20 cancer types (mutational signatures)

Cancer Type Signatures (nucleotide rubstitution ratios)													
	A->C	A->G	A->T	C->A	C->G	C->T	G->A	G->C	G->T	T->A	T->C	T->G	Ti/Tv
BLCA	0.011	0.043	0.015	0.04	0.12	0.23	0.3	0.16	0.045	0.01	0.024	0.0095	1.45
BRCA	0.054	0.058	0.022	0.065	0.084	0.2	0.24	0.1	0.07	0.018	0.041	0.046	1.17
CESC	0.012	0.024	0.0082	0.047	0.11	0.25	0.31	0.15	0.048	0.0056	0.02	0.011	1.54
COAD	0.03	0.076	0.015	0.069	0.018	0.3	0.31	0.017	0.069	0.013	0.065	0.022	2.97
GBM	0.018	0.074	0.023	0.049	0.039	0.3	0.33	0.037	0.046	0.021	0.051	0.016	3.03
HNSC	0.014	0.067	0.032	0.069	0.078	0.23	0.25	0.095	0.084	0.02	0.04	0.012	1.45
KIRC	0.045	0.092	0.053	0.092	0.056	0.16	0.18	0.06	0.089	0.055	0.081	0.036	1.06
KIRP	0.046	0.1	0.057	0.08	0.068	0.15	0.17	0.07	0.066	0.057	0.095	0.042	1.06
LAML	0.015	0.078	0.019	0.06	0.041	0.29	0.33	0.033	0.048	0.026	0.047	0.019	2.85
LGG	0.017	0.11	0.021	0.039	0.048	0.27	0.32	0.04	0.04	0.017	0.058	0.02	3.13
LIHC	0.034	0.14	0.087	0.089	0.043	0.14	0.15	0.043	0.11	0.055	0.072	0.029	1.02
LUAD	0.017	0.056	0.057	0.14	0.066	0.14	0.14	0.086	0.22	0.031	0.033	0.013	0.59
LUSC	0.017	0.064	0.047	0.13	0.072	0.16	0.16	0.091	0.18	0.027	0.035	0.012	0.73
OV	0.031	0.081	0.048	0.081	0.085	0.18	0.19	0.092	0.093	0.037	0.052	0.032	1.01
PRAD	0.028	0.087	0.027	0.051	0.043	0.28	0.28	0.04	0.061	0.023	0.051	0.027	2.33
READ	0.046	0.046	0.012	0.11	0.013	0.26	0.3	0.015	0.12	0.012	0.037	0.032	1.79
SKCM	0.0094	0.022	0.011	0.0092	0.0085	0.51	0.37	0.0077	0.008	0.011	0.021	0.011	12.18
STAD	0.038	0.071	0.016	0.064	0.021	0.29	0.3	0.023	0.065	0.016	0.061	0.032	2.63
THCA	0.017	0.11	0.037	0.047	0.07	0.23	0.21	0.064	0.056	0.091	0.044	0.022	1.47
UCEC	0.043	0.062	0.0081	0.12	0.0084	0.26	0.28	0.0078	0.13	0.0067	0.05	0.033	1.83
Pancancer	0.025	0.057	0.023	0.071	0.043	0.29	0.28	0.051	0.084	0.018	0.041	0.021	1.99

Table S2. COSMIC Tier 1 Cancer Gene Census genes (downloaded on June 30, 2020)

Cancer Gene Census gold-standard genes											
ABL1	*BAP1*	*CDC73*	*DDX3X*	*FGFR2*	*IDH1*	*LATS2*	*MYD88*	*PIK3R1*	*PTPRC*	*SMAD2*	*TGFBR2*
ACVR1	*BARD1*	*CDH1*	*DICER1*	*FGFR3*	*IDH2*	*LEF1*	*MYOD1*	*PIM1*	*PTPRT*	*SMAD3*	*TNFAIP3*
ACVR2A	*BAX*	*CDK12*	*DNM2*	*FGFR4*	*IKBKB*	*LRP1B*	*NCOR1*	*PLCG1*	*QKI*	*SMAD4*	*TNFRSF14*
AKT1	*BCL6*	*CDKN1B*	*DNMT3A*	*FLT3*	*IL7R*	*LZTR1*	*NCOR2*	*POLD1*	*RAC1*	*SMARCA4*	*TP53*
ALK	*BCL9L*	*CDKN2A*	*DROSHA*	*FOXA1*	*IRS4*	*MAP2K1*	*NF1*	*POLE*	*RAD21*	*SMARCB1*	*TP63*
AMER1	*BCOR*	*CEBPA*	*EGFR*	*FOXL2*	*JAK1*	*MAP2K2*	*NF2*	*POLQ*	*RB1*	*SMARCD1*	*TRAF7*
APC	*BCORL1*	*CHD4*	*EP300*	*FUBP1*	*JAK2*	*MAP2K4*	*NFE2L2*	*POT1*	*RBM10*	*SMO*	*TRRAP*
AR	*BIRC3*	*CIC*	*EPAS1*	*GATA1*	*JAK3*	*MAP3K1*	*NFKBIE*	*PPM1D*	*RET*	*SOCS1*	*TSC1*
ARHGAP26	*BRAF*	*CNOT3*	*ERBB2*	*GATA2*	*KCNJ5*	*MAP3K13*	*NOTCH1*	*PPP2R1A*	*RHOA*	*SPEN*	*TSC2*
ARID1A	*BRCA1*	*COL2A1*	*ERBB3*	*GATA3*	*KDM5C*	*MAPK1*	*NOTCH2*	*PPP6C*	*RNF43*	*SPOP*	*TSHR*
ARID1B	*BRCA2*	*CREBBP*	*ERBB4*	*GNA11*	*KDM6A*	*MAX*	*NPM1*	*PRDM1*	*RPL10*	*SRC*	*U2AF1*
ARID2	*BTK*	*CRLF2*	*ESR1*	*GNAQ*	*KDR*	*MED12*	*NRAS*	*PREX2*	*RPL5*	*SRSF2*	*UBR5*
ASXL1	*CACNA1D*	*CSF3R*	*ETNK1*	*GNAS*	*KEAP1*	*MEN1*	*NT5C2*	*PRKACA*	*SALL4*	*STAG2*	*USP8*
ATM	*CALR*	*CTCF*	*EZH2*	*GRIN2A*	*KIT*	*MET*	*PAX5*	*PRKAR1A*	*SDHA*	*STAT3*	*VHL*
ATP1A1	*CARD11*	*CTNNB1*	*FAS*	*H3F3A*	*KLF4*	*MLH1*	*PBRM1*	*PTCH1*	*SETBP1*	*STAT5B*	*WT1*
ATR	*CASP8*	*CUX1*	*FAT1*	*H3F3B*	*KLF6*	*MPL*	*PDGFRA*	*PTEN*	*SETD2*	*STK11*	*XPO1*
ATRX	*CBL*	*CXCR4*	*FAT4*	*HIF1A*	*KMT2C*	*MSH2*	*PHF6*	*PTK6*	*SF3B1*	*SUFU*	*ZFHX3*
AXIN1	*CBLB*	*CYLD*	*FBXO11*	*HIST1H3B*	*KMT2D*	*MSH6*	*PHOX2B*	*PTPN11*	*SFRP4*	*TBL1XR1*	*ZRSR2*
AXIN2	*CD79A*	*DAXX*	*FBXW7*	*HNF1A*	*KRAS*	*MTOR*	*PIK3CA*	*PTPN13*	*SH2B3*	*TBX3*	*TENT5C **
B2M	*CD79B*	*DDR2*	*FES*	*HRAS*	*LATS1*	*MYCN*	*PIK3CB*	*PTPRB*	*SIX1*	*TET2*	
* excluded from the ROC analysis											

Table S3. Method evaluation

Method	Number of Significant Genes	CGC Overlap	Method Consensus	p-value Deviation	Consistency	AUROC	AUPRC
CI	98	0.56	0.91	0.09	0.81	0.79	0.39
CI (with INDEL)	159	0.43	0.86	0.54	0.85	0.81	0.43
MutPanning	125	0.55	0.90	0.24	0.86	0.67	0.44
dNdScv	153	0.50	0.92	0.17	0.87	0.76	0.43
MutsigCV	158	0.37	0.67	1.15	0.40	0.72	0.25
2020+	208	0.40	0.81	0.14	0.69	0.88	0.45
TUSON	243	0.37	0.86	0.74	0.89	0.79	0.47
ActiveDriver	417	0.06	0.24	1.24	0.26	0.58	0.06
OncodriveClust	586	0.07	0.34	1.48	0.39	0.72	0.16
OncodriveFML	679	0.12	0.41	0.86	0.56	0.79	0.29
MuSiC	1975	0.05	0.20	2.56	0.61	0.77	0.12
OncodriveFM	2600	0.04	0.32	1.75	0.55	0.70	0.20

Table S4. Gold-standard cancer genes

See Supplemental_Table_S4.xlsx file

Table S5. 460 candidate drivers

See Supplemental_Table_S5.xlsx file

Table S6. Candidate driver genes for the MC3 version of the TCGA data

See Supplemental_Table_S6.xlsx file

Table S7: Overlap of candidate genes between the original and the MC3 versions for the CI with INDEL analysis

Cancer Type	Number Significant Genes in MC3 Analysis	Number Significant Genes in Original Analysis	Number Overlapping MC3 Genes with Original Analysis	Fraction of Original Genes Recovered in MC3 Analysis	Hypergeometric p-value	Hypergeometric -log10 (p-value)
BLCA	56	19	16	0.84	1.84E-38	37.73
BRCA	29	40	25	0.63	5.87E-66	65.23
CESC	21	20	11	0.55	1.43E-28	27.85
COAD	95	49	31	0.63	2.17E-60	59.66
GBM	13	11	9	0.82	4.52E-28	27.34
HNSC	38	45	32	0.71	1.22E-82	81.91
KIRC	10	13	9	0.69	1.42E-27	26.85
KIRP	6	3	3	1.00	9.28E-11	10.03
LAML	2	13	2	0.15	8.35E-06	5.08
LGG	17	13	13	1.00	7.20E-41	40.14
LIHC	19	7	5	0.71	6.20E-14	13.21
LUAD	79	23	17	0.74	3.32E-36	35.48
LUSC	40	26	10	0.38	8.46E-21	20.07
OV	10	6	4	0.67	2.46E-12	11.61
PRAD	13	9	9	1.00	9.05E-29	28.04
READ	19	17	10	0.59	2.14E-26	25.67
SKCM	554	75	53	0.71	4.53E-63	62.34
STAD	83	3	3	1.00	1.10E-07	6.96
THCA	8	5	5	1.00	7.69E-16	15.11
UCEC	372	72	42	0.58	1.33E-52	51.88
PAN	539	249	143	0.57	1.82E-156	155.74

Table S8. IPA canonical pathway

See Supplemental_Table_S8.xlsx file

Table S9. Molecule Cell Function

See Supplemental_Table_S9.xlsx file

Table S10. IPA 313 novel genes

See Supplemental_Table_S10.xlsx file

Table S11. Gold-standard tumor suppressor oncogene

See Supplemental_Table_S11.xlsx file

Table S12. Avana CRISPR screen data

See Supplemental_Table_S12.xlsx file

Table S13. HNSC CI predictions

See Supplemental_Table_S13.xlsx file

Table S14. CI predicted oncogenes

See Supplemental_Table_S14.xlsx file

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nature methods* **7**(4): 248-249.

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D et al. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**(7391): 603-607.

Cancer Cell Line Encyclopedia C, Genomics of Drug Sensitivity in Cancer C. 2015. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528**(7580): 84-87.

Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, Elledge SJ. 2013. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**(4): 948-962.

Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R et al. 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34**(2): 184-191.

Forbes SA, Beare D, Bindal N, Bamford S, Ward S, Cole CG, Jia M, Kok C, Boutselakis H, De T et al. 2016. COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. *Curr Protoc Hum Genet* **91**: 10 11 11-10 11 37.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. *Nat Rev Cancer* **4**(3): 177-183.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA et al. 2013. Mutational landscape and significance across 12 major cancer types. *Nature* **502**(7471): 333-339.

Kato S, Han SY, Liu W, Otsuka K, Shibata H, Kanamaru R, Ishioka C. 2003. Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proceedings of the National Academy of Sciences of the United States of America* **100**(14): 8424-8429.

Katsonis P, Lichtarge O. 2014. A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome research* **24**(12): 2050-2058.

-. 2017. Objective assessment of the evolutionary action equation for the fitness effect of missense mutations across CAGI blinded contests. *Hum Mutat* **38**(9): 1072–1084.

-. 2019. CAGI5: Objective performance assessments of predictions based on the Evolutionary Action equation. *Hum Mutat*.

Kim H, Li F, He Q, Deng T, Xu J, Jin F, Coarfa C, Putluri N, Liu D, Songyang Z. 2017. Systematic analysis of human telomeric dysfunction using inducible telosome/shelterin CRISPR/Cas9 knockout cells. *Cell Discov* **3**: 17034.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework

for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**(3): 310-315.

Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**(7484): 495-501.

Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**(6): 417-425.

Lichtarge O, Bourne HR, Cohen FE. 1996. An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology* **257**(2): 342-358.

Lisewski AM, Lichtarge O. 2010. Untangling complex networks: risk minimization in financial markets through accessible spin glass ground states. *Physica A* **389**(16): 3250-3253.

Lisewski AM, Quiros JP, Ng CL, Adikesavan AK, Miura K, Putluri N, Eastman RT, Scanfeld D, Regenbogen SJ, Altenhofen L et al. 2014. Supergenomic network compression and the discovery of EXP1 as a glutathione transferase inhibited by artesunate. *Cell* **158**(4): 916-928.

Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, Davies H, Stratton MR, Campbell PJ. 2017. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**(5): 1029-1041 e1021.

Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV, Montgomery PG, Cowley GS, Pantel S et al. 2017. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat Genet* **49**(12): 1779-1784.

Mihalek I, Res I, Lichtarge O. 2004. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *Journal of molecular biology* **336**(5): 1265-1282.

Miosge LA, Field MA, Sontani Y, Cho V, Johnson S, Palkova A, Balakishnan B, Liang R, Zhang Y, Lyon S et al. 2015. Comparison of predicted and actual consequences of missense mutations. *Proceedings of the National Academy of Sciences of the United States of America* **112**(37): E5189-5198.

Moullan N, Mouchiroud L, Wang X, Ryu D, Williams EG, Mottis A, Jovaisaite V, Frochaux MV, Quiros PM, Deplancke B et al. 2015. Tetracyclines Disturb Mitochondrial Function across Eukaryotic Models: A Call for Caution in Biomedical Research. *Cell reports* **10**(10): 1681-1691.

Neskey DM, Osman AA, Ow TJ, Katsonis P, McDonald T, Hicks SC, Hsu TK, Pickering CR, Ward A, Patel A et al. 2015. Evolutionary Action Score of *TP53* Identifies High-Risk Mutations Associated with Decreased Survival and Increased Distant Metastases in Head and Neck Cancer. *Cancer research* **75**(7): 1527-1536.

Osman AA, Monroe MM, Ortega Alves MV, Patel AA, Katsonis P, Fitzgerald AL, Neskey DM, Frederick MJ, Woo SH, Caulin C et al. 2015a. Wee-1 kinase inhibition overcomes cisplatin resistance associated with high-risk *TP53* mutations in head and neck cancer through mitotic arrest followed by senescence. *Molecular cancer therapeutics* **14**(2): 608-619.

Osman AA, Neskey DM, Katsonis P, Patel AA, Ward AM, Hsu TK, Hicks SC, McDonald TO, Ow

TJ, Alves MO et al. 2015b. Evolutionary Action Score of *TP53* Coding Variants Is Predictive of Platinum Response in Head and Neck Cancer Patients. *Cancer research* **75**(7): 1205-1215.

Raevaara TE, Korhonen MK, Lohi H, Hampel H, Lynch E, Lonnqvist KE, Holinski-Feder E, Sutter C, McKinnon W, Duraisamy S et al. 2005. Functional significance and clinical phenotype of nontruncating mismatch repair variants of MLH1. *Gastroenterology* **129**(2): 537-549.

Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research* **39**(17): e118.

Shin H, Lisewski AM, Lichtarge O. 2007. Graph sharpening plus graph integration: a synergy that improves protein functional classification. *Bioinformatics* **23**(23): 3217-3224.

Spurdle AB, Healey S, Devereau A, Hogervorst FB, Monteiro AN, Nathanson KL, Radice P, Stoppa-Lyonnet D, Tavtigian S, Wappenschmidt B et al. 2012. ENIGMA--evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Hum Mutat* **33**(1): 2-7.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**(16): 9440-9445.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**(43): 15545-15550.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP et al. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**(Database issue): D447-452.

Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. 2016. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A* **113**(50): 14330-14335.

Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM et al. 2017. Defining a Cancer Dependency Map. *Cell* **170**(3): 564-576 e516.

Venner E, Lisewski AM, Erdin S, Ward RM, Amin SR, Lichtarge O. 2010. Accurate protein structure annotation through competitive diffusion of enzymatic functions over a network of local evolutionary similarities. *PLoS One* **5**(12): e14286.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. 2013. Cancer genome landscapes. *Science* **339**(6127): 1546-1558.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**(16): e164.

Xu Q, Tang Q, Katsonis P, Lichtarge O, Jones D, Bovo S, Babbi G, Martelli PL, Casadio R, Lee GR et al. 2017. Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4. *Human mutation* **38**(9): 1123-1131.

Zhang J, Kinch LN, Cong Q, Weile J, Sun S, Cote AG, Roth FP, Grishin NV. 2017. Assessing

predictions of fitness effects of missense mutations in SUMO-conjugating enzyme UBE2I. *Human mutation* **38**(9): 1051-1063.

Zhou W, Chen T, Chong Z, Rohrdanz MA, Melott JM, Wakefield C, Zeng J, Weinstein JN, Meric-Bernstam F, Mills GB et al. 2015. TransVar: a multilevel variant annotator for precision genomics. *Nat Methods* **12**(11): 1002-1003.