CSH PRESS GENOME RESEARCH

A general calculus of fitness landscapes finds genes under selection in cancers

Teng-Kuei Hsu, Jennifer Kay Asmussen, Amanda Michele Koire, et al.

Genome Res. published online March 17, 2022
Access the most recent version at doi:10.1101/gr.275811.121

A general calculus of fitness landscapes finds genes under selection in cancers

Teng-Kuei Hsu[1*], Jennifer Asmussen[2*], Amanda Koire[3], Byung-Kwon Choi[2], Mayur A. Gadhikar[4], Eunna Huh[5], Chih-Hsu Lin[3], Daniel M. Konecki[3], Young Won Kim[6], Curtis R. Pickering[4], Marek Kimmel[7,8], Lawrence A. Donehower[9], Mitchell J. Frederick[10], Jeffrey N. Myers[4], Panagiotis Katsonis[2**], and Olivier Lichtarge[1-3,5,6,11,**]

[1] Department of Biochemistry & Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX, 77030

[2] Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX, 77030

[3] Program in Quantitative and Computational Biosciences, Baylor College of Medicine, One Baylor Plaza, Houston, TX, 77030

[4] Department of Head and Neck Surgery, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd, Houston, TX 77030

[5] Department of Pharmacology and Chemical Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX, 77030

[6] Program in Integrative Molecular and Biomedical Sciences, Baylor College of Medicine, One Baylor Plaza, Houston, TX, 77030

[7] Departments of Statistics and Bioengineering, Rice University, 6100 Main Street, Houston, TX 77005, USA

[8] Systems Engineering Group, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

[9] Department of Molecular Virology and Microbiology, Baylor College of Medicine

[10] Department of Otolaryngology - Head and Neck Surgery, Baylor College of Medicine, One Baylor Plaza, Houston, TX, 77030

[11] Computational and Integrative Biomedical Research Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX, 77030

* These authors contributed equally to this work.

** Corresponding authors: katsonis@bcm.edu and lichtarge@bcm.edu

ABSTRACT

Genetic variants drive the evolution of traits and diseases. We previously modeled these variants as small displacements in fitness landscapes and estimated their functional impact by differentiating the evolutionary relationship between genotype and phenotype. Conversely, here we integrate these derivatives to identify genes steering specific traits. Over cancer cohorts, integration identified 460 likely tumor-driving genes. Many have literature and experimental support but had eluded prior genomic searches for positive selection in tumors. Beyond providing cancer insights, these results introduce a general calculus of evolution to quantify the genotype-phenotype relationship and discover genes associated with complex traits and diseases.

INTRODUCTION

From short term disease risk to long term species evolution, the genotype-phenotype relationship describes how genetic variations induce biological change (Manolio et al. 2009). Experimental screens tracking the effect of these variations include RNA interference (RNAi) (Boutros and Ahringer 2008), CRISPR-Cas9 knockouts (Mali et al. 2013; Koike-Yusa et al. 2014; Wang et al. 2014b; Hart et al. 2015), and deep mutational scans (Fowler et al. 2014) within the limitations of achievable perturbations and assays (Mak and Justman 2017). Alternately, statistical analyses of genome-wide association identify over-represented variants in case-control studies. These presumably influence the phenotype common to cases subjects (Hirschhorn and Daly 2005; Hardy and Singleton 2009) although sample size, signal quality (McCarthy et al. 2008), and rate biases (Korte and Farlow 2013) may limit accuracy.

Here, we propose a different approach to recover the genotype-phenotype relationship, which is based on representing genetic variations as moves in the fitness landscape (Wright 1932). Prior theory suggests that these moves should generally be small and nearly neutral (Nei 2005). Against this background, we hypothesize that gene mutations driving new phenotypes are the result of abnormally large moves in the fitness landscape. Testing this hypothesis requires a metric for motions in the fitness landscape. We propose to use the Evolutionary Action (EA) of mutations on fitness described in prior work as the derivative of the genotype-phenotype relationship (Katsonis and Lichtarge 2014). In practice, the EA score correlates with the experimental effects of mutations (Katsonis and Lichtarge 2014) and consistently performs well in blinded assessments of predictions of deleterious mutations against state-of-the-art statistical and machine learning methods (Katsonis and Lichtarge 2017; Katsonis and Lichtarge 2019). A limitation of EA, however, is that it describes only the impact of single mutations, or individual moves in the fitness landscape. This is not sufficient to interpret complex polygenic phenotypes due to multiple causal variants. To identify groups of gene variations that in aggregate drive patients to a disease region of the fitness landscape, we therefore propose a new operation, called Cohort Integration (CI), that sums the individual effects of variants measured with EA over all genes and over all patients. Calculus suggests that this summation will reverse the differential operation that led to EA in the first place and thus recover the genotype-phenotype relationship, meaning, it will uncover genes that drive cohort-specific traits.

Here, we test this model in cancer. Tumor genomes evolve (Greaves and Maley 2012) by acquiring advantageous somatic mutations and which, when considered collectively across a cohort of cancer patients, should be associated with a large displacement in the fitness landscape. The average number of coding mutations per tumor can be as small as about eight in leukemia but is more often confoundingly large, such as about 1,600 in colorectal cancers (Vogelstein et al. 2013). Among these, however, the number of cancer-driving somatic mutations are relatively few, three to five by some estimates (Tomasetti et al. 2015), and finding these cancer drivers remains difficult, and critical for personalized therapy (Chin et al. 2011). To search for cancer genes that harbor these driver mutations, state-of-the-art methods (Dees et al. 2012; Davoli et al. 2013; Vogelstein et al. 2013; Lawrence et al. 2014; Tokheim et al. 2016; Martincorena et al. 2017; Bailey et al. 2018; Dietlein et al. 2020; Martinez-

Jimenez et al. 2020) pool statistics and machine learning to search for signs of positive of selection in cancer genes, including mutation frequency (Dees et al. 2012; Lawrence et al. 2014), surrounding nucleotide context (Dietlein et al. 2020), inactivation bias (Greenman et al. 2007; Van den Eynden et al. 2015; Martincorena et al. 2017), functional impact (Gonzalez-Perez and Lopez-Bigas 2012; Davoli et al. 2013), and structural or functional clustering (Tamborero et al. 2013; Porta-Pardo and Godzik 2014). Some of the challenges to identify driver genes include inaccurate background mutation rates (Lawrence et al. 2013), too few mutations per gene (Van den Eynden et al. 2015), and unbalanced distributions of passenger mutations (Bignell et al. 2010) that lead to a repertoire of mutational signatures (Alexandrov et al. 2013; Alexandrov et al. 2020). Notably, as much as 60 to 80% of genes identified by one method are not found by others (Tokheim et al. 2016), and a large number of rare cancer drivers in individual patients remain hidden for lack of a population-wide role (Garraway and Lander 2013; Chang et al. 2016). The sequencing of diverse types of somatic tumor tissue from a large number of patients by The Cancer Genome Atlas (The Cancer Genome Atlas Research Network 2008; The Cancer Genome Atlas Research Network et al. 2013; Tomczak et al. 2015; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020) yields a rich dataset for discovering new cancer genes. Here, we used CI to prioritize the collective fitness impact of variants across all genes and all patients in a cohort in order to discover candidate cancer driver genes, and compare this approach to other cancer gene identification techniques (Greenman et al. 2007; Dees et al. 2012; Gonzalez-Perez and Lopez-Bigas 2012; Davoli et al. 2013; Tamborero et al. 2013; Lawrence et al. 2014; Porta-Pardo and Godzik 2014; Van den Eynden et al. 2015; Tokheim et al. 2016; Martincorena et al. 2017; Bailey et al. 2018).

RESULTS

Cohort Integration (CI) of the fitness effects detects selection on a variant set

As detailed in Supplementary Information, our evolutionary calculus model hypothesizes a formal genotype-phenotype relationship. Differentiating this relationship yields the EA equation for the fitness effects of genetic variants. Integrating the numerical values of the gene fitness effects given by EA over a patient cohort should measure the relationship of each gene with the phenotype, as illustrated in the visual summary of the evolutionary calculus theory (Fig. 1A-E). As a basic test of EA in cancer, the experimental and clinical effects of *TP53* (Kato et al. 2003), *MLH1* (Raevaara et al. 2005), *BRCA1*, and *BRCA2* (Spurdle et al. 2012) mutations agree with EA fitness scores better than alternative methods (Supplementary Fig. 1).

For proof of concept of the cohort integration approach, we used mutation data from distinct populations and from random simulations. The cohort integral of nearly 24 million germline missense variants from more than 2,500 diverse, healthy individuals in the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015) showed an exponentially decreasing distribution as a function of EA, biased toward lower EA values than simulated random nucleotide substitutions (Fig. 2A, $p = 3.8\times10^{-29}$, see Methods). This EA distribution bias is consistent with negative selection against germline variants with large fitness effects among healthy-born individuals. By contrast, the cohort integral of 645,359 cancer somatic missense mutations from 5,996 patient tumor genomes across 20 cancers from The Cancer Genome Atlas (TCGA) (Tomczak et al. 2015) was exponentially decreasing at a slower rate that was indistinguishable from computer-simulated random substitutions of nucleotides in protein coding regions (Fig. 2B, $p = 0.41$). This is consistent with most cancer somatic missense mutations being random and under little selection (Greenman et al. 2007). A single gene illustration is the TCGA-wide cohort integral of *DNAH5*, a large and frequently mutated gene that has not, to the best of our knowledge, been associated with cancer and which appears free of any selection pressure (Supplementary Fig. 2A, $p = 0.84$). The distribution of EA values was also indistinguishable between random nucleotide substitutions and computer-simulated variants generated according to mutational signatures of mono- and tri- nucleotide substitution ratios observed in somatic mutations from 20 different cancer types (Table S1; Supplementary Fig. 2B-C). These observations suggest that the distribution of EA values is insensitive to biases related to the mechanism that generates the mutations and is specific to fitness selection forces.

In contrast, cohort integrals of representative well-established tumor suppressors, such as *TP53*, *CDKN2A*, *PTEN*, and *NOTCH1*, computed over all TCGA somatic missense mutations found in any cancer type, were all strongly biased toward high EA scores, usually above 70 and consistent with complete loss of function (Fig. 2C, $p = 2.8\times10^{-303}$, 2×10^{-7}, 9.7×10^{-50}, and 1.4×10^{-11}, respectively). This difference indicates positive selection and fits the expectation that cancer-causing mutations often weaken tumor suppression. Likewise, the cohort integrals of representative oncogenes, such as *PIK3CA*, *BRAF*, *KRAS*, and *NRAS*, were all also highly significant (Fig. 2D, $p = 2.2\times10^{-57}$, 9.4×10^{-179}, 1.1×10^{-63}, and 9.6×10^{-45}, respectively). Their peak EA scores, however, fell between 30 and 70, consistent with gain-of-function variants that repurpose rather than destroy a protein. We may also apply cohort integration on individual cancer types. For example, *CDH1* is known to drive breast and stomach cancers (Corso et al. 2020). This gene has a significant, non-neutral cohort integral in these cancer types (Supplementary Fig. 2D, $p = 0.002$ and 0.03, respectively), but not in any other cancer type individually or collectively (Supplementary Fig. 2E, $p = 0.9$). These data show that cohort integration identifies populations and genes under different selection pressures. Typically, inherited coding variants in healthy individuals are under negative selection, passenger coding variants in cancer populations are under random (or no) selection pressure, and well-known cancer driver genes are under positive selection, with intermediate EA values in oncogenes, and larger EA values in tumor suppressors. These differences suggest that cohort integration provides a new method to identify genes under positive selection.

CI recovery of cancer-driving genes performs on par with or better than existing state-of-the-art methods

To assess performance on recovering cancer-driver genes, we compared CI to ten state-of-the-art algorithms using multiple criteria of success. All algorithms were evaluated over the same test population of 7,916 exomes from 34 cancer types, collected by another study (Tokheim et al. 2016). Nevertheless, the result should be cautiously interpreted since the included methods may rely on different variant types and engage external complementary information. We adapted and enriched previously used benchmark measures (Tokheim et al. 2016) to evaluate each method for: precision measured by the overlap of the discovered genes with a gold-standard gene set (*CGC Overlap*, Table S2) and with the consensus of genes discovered by the other algorithms (*Method Consensus*); the combined true and false positive rates measured by area under the Receiver Operating Characteristic (*AUC-ROC*) and Precision-Recall (*AUPRC*) curves; the overlap of the discovered genes when using each half of the test population (*Consistency*); the discrepancy between observed and theoretical *p*-value distributions (*p-value deviation*). Using only the missense mutations as input to the CI approach resulted in 61 significantly non-random genes after multiple testing correction (*q-value* < 0.1), with a *CGC Overlap* of 0.64, *AUPRC* of 0.32, and *AUC-ROC* of 0.73. Although this *CGC Overlap* was the best compared to the other methods, the small number of discovered genes resulted in intermediate *AUC-ROC* and *AUPRC* performance (Supplementary Fig. 3). This likely stems from the fact that other methods used additional mutations, such as nonsense variants and frame-shift insertions and deletions (fs-indels). However, accounting for nonsense variants in the CI framework is straight forward since random nucleotide changes provide the expected rate of nonsense mutations and we can score them with the highest impact for a loss of function variant (EA=100). CI of missense plus nonsense mutations thus discovered 98 genes after multiple testing correction (*q*-value < 0.1), which increased *AUC-ROC* (0.79) and *AUPRC* (0.39) values compared to using missense variants only, and precision (*CGC Overlap* of 0.56) remained better than the state-of-the-art methods (Fig. 3A, Supplementary Fig. 3, and Table S3, “CI”). To also account for fs-indels in our approach, we further combined the CI *p*-value of each gene with the probability that fs-indels appear in that gene by chance (see Methods). This modified CI approach prevents the discovery of genes due to fs-indel variants only and results in the inclusion of borderline CI analysis genes (*p*-value < 0.05 but *q*-value > 0.1) when these genes have sufficient support from the fs-indel variants analysis. This raised the number of discovered genes to 159 and further improved the *AUC-ROC* (0.81) and *AUPRC* (0.43) values (Fig. 3A, Supplementary Fig. 3, Table S3, “CI (with INDEL)”). Using fs-indels variants alone recovers 349 candidate cancer genes with a 0.28 *AUPRC* and 0.19 *CGC overlap*. Additionally, we assessed the performance of CI with fs-indels across 13 cancer types within test population. As expected, the robustness of the CI with fs-indels method, as measured by *AUC-ROC*, *AUPRC*, and *Consistency*, decreases with cancer cohort size (Supplementary Fig. 4). A comparable performance reduction is observed for other state of the art methods. This benchmark analysis establishes that the calculus-based CI methods have robust performance compared to state-of-the-art methods for identifying candidate cancer genes. Additionally, CI *p*-values for missense and nonsense mutations can be combined with *p*-values based on an independent test for fs-indels with no appreciable loss of performance, therefore, CI with fs-indels is used hereafter.

To examine the power of CI we tested the robustness of its results as we reduced the number of test population exomes by down-sampling. First, we defined a target set of 58 genes, as those identified by CI using all 7,916 test exomes and also reported as cancer associated by at least four of five reference sources (Davoli et al. 2013; Kandoth et al. 2013; Vogelstein et al. 2013; Lawrence et al. 2014; Forbes et al. 2017) (see Table S4). Then, for increments of 100 tumor exomes, we calculated the percentage overlap of the genes identified by the subset and the target genes. For each increment, we generated 10 subsets by resampling exomes and calculated the average and standard deviation of the overlap (Supplementary Fig. 5). Using half of the 7,916 exomes was sufficient to detect 86% of the target genes, while 1,150 exomes (14%) were sufficient to detect half of the target genes. These results show that CI is robust and powerful, producing reliable results even with a fraction of the input data, suggesting that it may also be capable of identifying driver genes in specific cancer types.

CI detects genes under positive selection across cancer mutational landscapes

Next, to discover new candidate cancer-driving genes, we applied CI prospectively and separately over the genomic sequencing data from 5,996 TCGA tumor genomes across 20 cancers from TCGA (see Methods). CI detected 326 candidates that were implicated in at least one TCGA cancer type (Fig. 3B). Moreover, 134 additional candidate pan-cancer genes were found by analyzing all cancer genomes together (Table S5, q-value < 0.1). CI was also applied to an updated release of TCGA tumor exomes (Table S6) with significantly overlapping findings (Table S7). The 460 genes identified in any and all of the 20 cancer types included 116 transcription regulators, 65 enzymes, and 33 kinases (Supplementary Fig. 6A). Gene set enrichment analysis (GSEA) using Hallmark Gene Sets (Liberzon et al. 2015) (see

Methods) found nine significant associations (Supplementary Fig. 6B), including to PI3K-AKT-MTOR signaling (q-value = 3×10^{-4}), apical junction (q-value = 3×10^{-4}), and p53 pathway (q-value = 6×10^{-4}). Ingenuity Pathway Analysis (Kramer et al. 2014) (see Methods) found significant overlap with 261 overlapping canonical pathways (Table S8, q-value < 0.1) most of which are cancer signaling pathways. The top molecular and cellular functions (Table S9) included gene expression (q-value = 4×10^{-15} - 4×10^{-38}), cell death and survival (q-value = 9×10^{-11} - 6×10^{-25}), cell cycle (q-value = 3×10^{-11} - 7×10^{-23}), cell growth and proliferation (q-value = 6×10^{-11} - 3×10^{-22}), and cellular development (q-value = 8×10^{-11} - 4×10^{-18}).

We hypothesized that if CI robustly identified cancer-driver genes then it could point out which patient cohorts shared tumors of similar genetic etiology. This was tested with a similarity tree of cancer types based on the candidate drivers that were predicted independently in each type (see Methods) (Fig. 3C). As expected, patients with renal papillary carcinoma (KIRP) clustered together with those suffering from renal clear cell carcinoma (KIRC); rectal adenocarcinoma (READ) with colon adenocarcinoma (COAD); and lower grade glioma (LGG) with glioblastoma multiforme (GBM). Head-neck (HNSC), cervical (CESC), bladder (BLCA), and lung squamous cell carcinoma (LUSC) clustered in one branch, in agreement with a study proposing LUSC, HNSC, and BLCA as a squamous-like subtype (Hoadley et al. 2014). Together these data suggest that across distinct patient cohorts CI robustly identifies biologically reasonable candidate cancer-drivers in specific cancers and can indicate which genes associate more broadly with the tissue-of-origin.

To assess novelty, we compared the 460 genes to a gold-standard control list of genes already associated with cancer in established sources (Table S4): the COSMIC Cancer Gene Census (Forbes et al. 2017) and candidate genes suggested by 20/20 (Vogelstein et al. 2013), MutSigCV (Lawrence et al. 2014), TUSON (Davoli et al. 2013), and MuSiC (Kandoth et al. 2013). Of the 465 genes from these combined sources, CI recovered 147 (p-value = 1×10^{-124}, hypergeometric test), with mutual agreement coverage rising from 13% (37 of 285, p-value = 3×10^{-16}) for genes imputed by a single source only to 95% (39 of 41, p-value = 8×10^{-61}) for genes agreed upon by all five (Supplementary Fig. 6C). For the 313 candidates identified by CI but not part of this gold-standard list, cancer was the top Diseases or Functions Annotation according to Ingenuity Pathway Analysis (Table S10). Moreover, 215 genes had independent cancer associations from three lines of evidence: appearance in at least ten cancer papers (Supplementary Fig. 6D, n = 118, p-value = 0.0001), significant diffusion to known cancer drivers in the STRING protein-protein interaction network (Shin et al. 2007; Lisewski et al. 2014; Szklarczyk et al. 2015) (Supplementary Fig. 6E, n = 55, p-value = 2×10^{-5}), and a non-synonymous to synonymous mutation ratio d_N/d_S greater than 1 and consistent with positive selection (Supplementary Fig. 6F, n = 159, p-value = 2×10^{-14}). Overall, 27 genes had support from all three types of evidence, 58 from two, and 130 from one, leaving 98 genes (21%) with no prior support from any of these sources (Supplementary Fig. 6G). Moreover, when we quantified a confidence score for cancer association by weighing the support sources (see Methods), the CI significance q-value correlated strongly with the cancer association confidence scores (Supplementary Fig. 6H, p-value < 0.0001).

CI profiles distinguish tumor suppressor genes from oncogenes

Since tumor suppressors and oncogenes have distinct peaks in their EA distributions (Fig. 2C-D), we could measure the selection pressure for complete Loss of Function (sLOF) with the skew of the distribution towards the maximum EA of 100, expected for tumor suppressors, and the selection pressure for Gain of Function (sGOF) by the skew towards the intermediate EA of 50, expected for oncogenes. For example, known tumor suppressor genes and oncogenes (Table S11) had significantly greater sLOF or sGOF scores, respectively, than other genes (Supplementary Fig. 7A), enabling a separation of tumor suppressor genes from oncogenes (Supplementary Fig. 7B) that was measurable with an area under the receiver-operator characteristic curve of 0.96 (Supplementary Fig. 7C). At a binary separation threshold of 0.1 the accuracy was 90% (Supplementary Fig. 7D), classifying correctly 93% (50 out of 54) of tumor suppressors and 83% (15 out of 18) of oncogenes (Fig. 4A). These data show that CI may sort oncogenes from tumor suppressors based on distinct selection profiles in the cancer mutational landscape. We could then classify the 460 candidate cancer genes identified by CI into 357 likely tumor suppressor genes and

103 likely oncogenes (Table S5). Among the 98 (21%) of new genes identified by CI, 18 were likely oncogenes and 80 were tumor suppressors.

Experimental support for oncogenes and tumor suppressors

Next, we verified these genes against the CRISPR essentiality screen from the Cancer Dependency Map (DepMap) database (Tsherniak et al. 2017), which used Cas9-mediated DNA cleavage to observe the effects of gene inactivation on cell proliferation. Since certain cancer cell lines are dependent on activated oncogenes to maintain their malignant properties, known as oncogene addiction (Weinstein and Joe 2008), this screen should be a good test system to independently validate the role of the 103 putative oncogenes. As expected, cancer cell lines are dependent on CI oncogenes when they harbor mutations with EA scores between 30 and 70, which is the range for gain-of-function mutations, and a similar dependency was observed in the known oncogenes from COSMIC but not in random genes (Fig. 4B and Table S12). Moreover, CI oncogenes identified in individual cancers had significantly stronger essentiality responses in their corresponding cancer types, suggesting tissue specificity (Supplementary Fig. 8). These results together show that CI is able to classify oncogenes and tumor suppressor genes and can also identify oncogenic gain-of-function variants in the context of specific cancer types.

CI provides insight into the biology of specific cancers

Head and neck squamous carcinoma (HNSC)

For illustration, we initially focus our attention on head and neck squamous carcinoma (HNSC). Based on the somatic mutations from 508 HNSC tumors, CI identified 45 candidate genes (Fig. 5A and Table S13), and the majority (39 out of 45) were predicted to be tumor suppressors, consistent with the genomic landscape of this cancer reported by others (Agrawal et al. 2011; Stransky et al. 2011; The Cancer Genome Atlas Research Network 2015). CI identified the major candidate tumor suppressors previously published for HNSC, including *TP53*, *FAT1*, *NOTCH1*, *CDKN2A*, *CASP8*, *PTEN*, *RB1*, *FBXW7*, *AJUBA*, and *NSD1*. CI also identified two well-known oncogenes for HNSC (*PIK3CA* and *HRAS*), as well as known oncogenes in other cancers but new in HNSC (*NEF2L2*, *RAC1*, and *MAPK1*) (Table S5). Of the 45 candidates, 28 have already been associated with cancer (though not necessarily HNSC) in COSMIC or by other state-of-the-art methods, and 17 genes were predictions unique to CI. Nine of which were associated with cancer in literature (Table S13).

HNSC candidate genes were involved in pathways associated with oncogenesis (Fig. 5B). Cell differentiation is a key pathogenic pathway in HNSC that includes the well-known HNSC drivers *NOTCH1*, *TP63, FAT1*, and *ZNF750* (Agrawal et al. 2011; Pickering et al. 2013; The Cancer Genome Atlas Research Network 2015). CI recovered three of these known drivers and predicted three more candidates from the same pathway, *NOTCH2*, *PDZRN4*, and *KDF1*. We have previously detected *NOTCH2* as a driver in aggressive cutaneous squamous cell carcinoma (Pickering et al. 2014) and its homology to the known tumor suppressor gene *NOTCH1* lends further support to the notion that *NOTCH2* is also a HNSC driver gene. *PDZRN4* belongs to the LNX gene family of homologous RING type E3 ubiquitin ligases (Katoh and Katoh 2004), and two related LNX family members have been shown to bind *NUMB*, a negative NOTCH regulator (Rice et al. 2001). Exogenous expression of *PDZRN4* has been shown to inhibit growth of hepatocellular carcinoma cell lines (Hu et al. 2015a), suggesting a tumor suppressor function consistent with the CI classification. *KDF1* regulates *TP63* and has been identified as a regulator of squamous differentiation through both human and mouse genetic studies (Lee et al. 2013). Another key pathway in HNSC is epigenetic regulation. *CREBBP*, *EP300*, *KMT2D* (*MLL2*), and *NSD1* have previously been identified as frequently mutated epigenetic regulators in HNSC (The Cancer Genome Atlas Research Network 2015), and we have identified *ARID2* and *KMT2B* (*MLL4*) as additional HNSC cancer-related epigenetic regulators. Other important pathways are the DNA damage and cell stress processes, and CI identified the DNA damage response gene *HNRNPUL1* and the nonsense-mediated RNA decay (NMD) factor *UPF2*, the latter can be targeted by Pateamine A (Dang et al. 2009). Lastly, we identified *NF2*, which is a proven tumor suppressor gene and is responsible for hereditary neurofibromatosis type II but has not been previously linked to HNSC. These results suggest that CI can identify new candidate driver genes in individual cancer types, although further experimental studies are required to validate them as true drivers.

We validated experimentally the *CUL3* gene, which was neither in our gold-standard list nor an established HNSC gene but was identified by CI as a tumor suppressor in this cancer type (CI *q*-value = 0.07; sGOF = -0.13; sLOF = 0.66). *CUL3* is the core component of E3 ubiquitin-protein ligase complexes that down-regulates *NFE2L2*, a known driver of HNSC. Previous work shows that the disruption of such adaptor-substrate recognition promotes malignancy (Ohta et al. 2008) and has been associated with poor prognosis in HNSC (Martinez et al. 2015). *CUL3* somatic mutations in TCGA HNSC samples are located at functionally and structurally important sites (Fig. 6A), as predicted by the Evolutionary Trace (ET) algorithm (Lichtarge et al. 1996; Mihalek et al. 2004). To assess how *CUL3* activity impacts cell growth and survival, four HNSC cell lines were engineered to ectopically express wild-type protein. Of the four cell lines, clonogenic potential was diminished by wild-type *CUL3* ectopic expression in UMSCC25 cells ($p < 0.0001$; Fig. 6B) compared to expression of the control gene LacZ. UMSCC25 was the only cell line to exhibit a prominent higher molecular weight *CUL3* band following ectopic expression (Fig. 6C), which is suggestive of the post-translational neddylation (Hori et al. 1999) that is required for *CUL3* ligase activity (Pintard et al. 2004). Further investigation revealed that ectopic expression of *CUL3* in UMSCC25 cells substantially reduced the percentage of G1 cells and produced a concomitant increase in the subG1 fraction (Fig. 6D), indicating cell death. There was also an increase in cleaved PARP (Fig. 6E), consistent with apoptosis, and, as expected, decreased levels of NFE2L2 (also known as NRF2), the known HNSC driver (Fig. 6F). Collectively, these data demonstrate the ability of CI to discern cancer-driving genes in the HNSC TCGA population, even if the mutation is uncommon. The rarity of actual *CUL3* mutations in HNSC could be explained by the cancer cells simply preventing *CUL3* neddylation instead, which would still disable the protein.

Breast Cancer (BRCA)

Similar observations are made when CI is applied to 977 BRCA tumor samples. Of the 40 candidate genes identified in BRCA, 32 are predicted to be tumor suppressors, and six of these have not been identified by other studies (Table S5). One of these six genes is a dual specificity protein phosphatase, a family of genes known to both positively and negatively regulate cellular pathways associated with tumorigenesis (Patterson et al. 2009; Meeusen and Janssens 2018). Based on these observations, we elected to focus additional *in vitro* experimental analyses on dual specificity protein phosphatase 16 (*DUSP16*).

DUSP16 (CI q-value = 0.09; sGOF = -0.53; sLOF = 0.62) has been implicated as a candidate negative growth regulator and potential tumor suppressor, at least in part through negative regulation of JNK signaling (Domotor 1989; Hoornaert et al. 2003; Keyse 2008). To confirm the role of *DUSP16* as a negative growth regulator we generated clonal lines of HEK293 and BT474 human cells containing transduced doxycycline-inducible *DUSP16* (Fig. 6G). Addition of doxycycline to both cell lines resulted in significant reduction (p-value < 0.001) in cell division relative to the same clones in the absence of doxycycline (Fig. 6H). The ability of doxycycline-induced HEK293 and BT474 cells to form colonies after plating individual cells at low density was also significantly reduced (HEK293, about 40%. n = 5, p-value = 0.014 and BT474, about 50 %, n = 3, p-value = 0.036) (Fig. 6I). Additionally, cell migratory activity was shown to be reduced in cells with elevated *DUSP16* (Fig. 6J and Supplementary Fig. 9A). Elevation of *DUSP16* was associated with increased cellular apoptotic markers in BT474 cells (Fig. 6K) and enhanced dephosphorylation of JNK (Supplementary Fig. 9B) as well as JUN at Ser63, a major proliferative transcription factor in AP-1 signaling (Fig. 6L and Supplementary Fig. 9C). This effect on JUN and potentially AP-1 signaling was shown in the reduction of AP-1 transcription targets JUN and FOS in the presence and absence of phorbol ester (an AP-1 signaling activator) stimulation (Fig. 6M-N). Finally, we also demonstrated that the *DUSP16*-induced inhibition of cell division in HEK293 and BT474 cells can be reversed when DUSP16 protein expression is downregulated by CRISPR (Fig. 6O). These results provide evidence that *DUSP16* is a potent tumor suppressor and provides additional support for the ability of CI to identify novel cancer driver genes.

Potential therapeutic applications

With a view to personalized therapy, we noted that 70 CI oncogenes were not found in the gold-standard list (Table S14), of which 13 may be druggable, including nine that can be directly targeted by drugs and four that belong to a targetable gene family. For example, the candidate *PDE10A*, a phosphodiesterase involved in PKG and PKA signaling (Soderling et al. 1999), is unique to CI in skin cutaneous melanoma (SKCM) and is predicted to be an oncogene. Consistent with a possible oncogenic role, its increased expression stimulates cell growth by activating the Wnt/β-catenin pathway (Li et al. 2015). Selective *PDE10A* inhibitors, such as papaverine, PQ-10, and Pf-2545920, have already been shown to suppress colon tumor cell growth (Lee et al. 2016). Another unique candidate with an oncogenic CI profile is *DYRK1A* in endometrial cancer (UCEC). This dual-specificity protein kinase regulates cell cycle progression (Fernandez-Martinez et al. 2015) and its drug-induced inhibition decreases proliferation and colony formation *in vitro* and suppresses tumor growth *in vivo* (Radhakrishnan et al. 2016). These examples show that CI oncogenes could suggest potential therapeutic targets.

DISCUSSION

The genotype to phenotype relationship is central to biology (Fisher 1930; Wright 1932). It hinges on the fitness effect of mutations, namely, their molecular, cellular, and physiological impact on the population structure (Orr 2009) and location in the fitness landscape (Wright 1932). Mutational scan experiments explore these fitness landscapes to a resolution limited by mutations, assays, and epistasis (Cunningham and Wells 1989; Pal et al. 2003; Pal et al. 2006; Datta et al. 2008). Alternately, we show here that evolutionary comparative sequence analysis across species can probe a vastly greater set of mutational trials coupled to evolutionary fitness via phylogenetic divergences (Lichtarge et al. 1996; Mihalek et al. 2004; Wilkins et al. 2010). EA first tapped these evolutionary data through differential analysis of fitness landscapes (Katsonis and Lichtarge 2014) to compute the effect of single mutations relatively accurately (Katsonis and Lichtarge 2017; Katsonis and Lichtarge 2019) and usefully (Neskey et al. 2015; Osman et al. 2015a; Osman et al. 2015b; Chun et al. 2019; Clarke et al. 2019). The hypothesis is that genotype (γ) and phenotype (φ) are linked through a continuous and differentiable evolutionary fitness function f, such that $f(\gamma) = \varphi$. EA exploited the computation of the gradient (∇f), although f itself remained unknown. Now, we show how to integrate ∇f in the fitness landscape to compute the path-determining genes driving a population to phenotype regions of interest. In effect, Cohort Integration solves f numerically for each gene, in line with the anti-derivative property stemming from the fundamental theorem of calculus. Integration thus completes the calculus model introduced with EA. This calculus has two consequences: it reveals evolutionary constraints and the genetic determinants of complex phenotypes (Kim et al. 2020; Koire et al. 2021).

The completion of the EA calculus points to a physics-based description of evolution. Although interpretations may differ (Doebeli et al. 2017), one possibility is to view *f* as a *potential* function describing the ability of a genotype to perform the *evolutionary work* required for reproduction in the fitness landscape. The gradient of such a potential, ∇f, would then define a field describing the *evolutionary force* at each point of the genome space acting against a substitution in the direction of each of the 19 alternate amino acids. Together, this space and its field are akin to the "fitness landscape" and its slope proposed by Wright (Wright 1932). The EA equation can now be interpreted as the product of a force times a displacement, namely the evolutionary *work* of a mutation pulling a genotype across the landscape against the *evolutionary force field*. Cohort Integration, by summing the work of all mutations minus random background fluctuations, is the energy of driver gene mutations propelling the patient cohort along paths to phenotypic traits. This physical interpretation makes testable predictions: the random background mutational energy of populations at steady state should follow the Boltzmann exponential distribution of energy (Landau and Lifshitz 1980). This model fits observations (Fig. 2 and Supplementary Fig. 2) and matches the distribution of fitness effect Fisher anticipated (Fisher 1930) and expanded more recently by others (Sella and Hirsh 2005). In cancer, this equilibrium distribution of mutational energy is illustrated by the background EA distribution of unconstrained random passenger mutations in a gene such as *DNAH5*. By contrast, *TP53*, *PIK3CA* and other cancer drivers are under selection, rather than fluctuating randomly around an equilibrium, and their mutational distributions differ drastically from a decaying exponential form.

In practice, our data show that in cancer this approach complements others to identify driver genes (Tokheim et al. 2016). It finds 460 genes under positive selection across 20 different types of cancer cohorts sequenced by TCGA. Almost one third of these drivers are well-known driver genes (32%). Nearly half (47%) are known in other types of cancers or supported by publications, interactions with cancer-associated genes, or other evidence of positive selection. The remainder are less studied genes (21%). Among the most implicated pathways were chromatin remodeling, nuclear receptor signaling, apoptosis, protein elongation, transcription, and G-protein signaling. This is consistent with cancer biology and progression, but it also includes many genes that were not previously reported by other large genetic studies (Dees et al. 2012; Davoli et al. 2013; Kandoth et al. 2013; Vogelstein et al. 2013; Bailey et al. 2018). For example, we identified SNF/SWI members *SMARCA2* (pan-cancer), *SMARCA1* (colon and pan-cancer), *SMARCC2* (pan-cancer), and *BRD7* (skin and pan-cancer) which are unreported in other large cancer sequencing studies. Among the latter, *SMARCA1* and *SMARCC2* have been shown to function as tumor suppressors in gastric cancers (Takeshima et al. 2015). Transcription factors belonging to the Forkhead (i.e. *FOXP1*), Homeobox (*MSX2*, *HOXB7*, *HOXB9*, *TGIF1*, and *POU4F1*), and Kruppel-like (i.e., *KLF5*, and *KLF3*) families were found in the pan-cancer analysis (*TGIF1* and *KLF5* were also found in colon and bladder cancer, respectively), and members from each of these classes of transcription factors have been extensively linked to cancer previously (Tetreault et al. 2013; Joo et al. 2016; Miller et al. 2016). Two protein elongation factors, *EEF1A1* and *EIF1AX* were predicted to be oncogenes in the pan-cancer analysis, consistent with many reports that elongation factors as a class are overexpressed in multiple tumor types and control proliferation and cell death (Abbas et al. 2015). Three genes involved in G-protein signaling: *RGS22* (pan-cancer) encodes a GTPase activating protein previously found to function as a tumor suppressor in pancreatic and esophageal cancers (Hu et al. 2011; Hu et al. 2015b); *ARHGAP5* (colon and pan-cancer) encodes a Rho GTPase activating protein shown to increase invasiveness of lung cancer cell lines (Wang et al. 2014a); and *RGL2* (pan-cancer) encodes a guanine nucleotide exchange factor that positively regulates growth of lung cancer cell lines (Santos et al. 2016).

Cohort Integration also captures the distinct selection profiles of tumor suppressors and oncogenes. The former have complete loss of function mutations with EA scores above 70 and the latter have gain of function mutations with EA scores between 30 and 70. During validation of two predicted tumor suppressors, intact versions of *CUL3* in head and neck cancer and *DUSP16* in breast cancer, each inhibited cancer cell growth in their corresponding cell lines. For *CUL3*, this data is consistent with a large-scale cancer gene identification study that pooled multiple predictive algorithms (Bailey et al. 2018). For oncogenes, we found strong agreement between our predicted oncogenes and a large-scale tumor dependency screen (DepMap data, Fig. 4B and Table S12). Of note, of the 70 candidate oncogenes we find, a third are currently categorized as druggable targets.

EA calculus has several limitations. It does not currently account for common and important non-coding cancer mutations, such as genomic rearrangements (Mitelman et al. 2007; Mertens et al. 2015), epigenetic (Esteller 2008) or gene copy number variations (Shlien and Malkin 2009; Yu and Shao 2009). Also, our approach does not account for the fact that oncogenic drivers are often seen at hot spot residues (Chang et al. 2018) or near each other in protein sequence and structure (Tamborero et al. 2013). Together, with the fact that oncogenic drivers have milder EA scores than tumor suppressor drivers, this may lead to lower sensitivity in identifying oncogenes. However, *p*-values that account for mutation clustering and recurrence, suggested by existing methods, could be integrated through Fisher's combined probability test (Fisher 1948) as illustrated here for frameshift indels (see Methods and Supplementary Fig. 10). Other integrative approaches employing machine learning are of interest and could reflect even broader types of data (Guan et al. 2012; Hu et al. 2015b; Yu et al. 2016; Yu et al. 2017). Additionally, the number of available homologous sequences for some genes may be insufficient to produce EA scores of high accuracy. For these genes, somatic mutations may mimic a random distribution and the genes will not be identified as cancer-driver candidates. Although this limitation may reduce sensitivity, it should not affect the precision of our approach. The sensitivity of the approach depends logarithmically on the sample size. Therefore, small sample sizes are sufficient for finding prominent cancer genes, while the discovery of very rare drivers remain a challenge. Differential mutation probabilities across sites might be an important confounder of the substitution odds that approximate $\Delta\gamma$, since this calculation involves aligned sequences of various pair identities that correspond to different branching lengths. Finally, epistasis is implicit in estimates of ∇f with ET, but its explicit modeling would require the fitness cross-effect of dual genome positions x and y, namely $\partial^2 f/\partial x \partial y$, a second order mixed derivative term beyond the scope of this first order analytic theory. Formal questions of differentiability and integrability in fitness landscapes are also not trivial (Carneiro and Hartl 2010). Here, differentiability of f is consistent with evolvability (Wagner and Altenberg 1996) through smooth mutational steps and, more strictly, with Schwartz distributions (Schwartz 1963). The discrete domain of integration which consists of many somatic variants, in many genes and over many patients cannot support Riemann integration, but is consistent with Lebesgue-Stieltjes integration (Carter and Brunt 2012) and, more strictly, a stochastic integral of Ito or Stratonovich type (LeGall 2016).

In summary, we propose a general analytic and evolutionary framework for the genotype-phenotype relationship. The approach uses differential and integral calculus to interpret the mutational burden in patient cohorts in light of the couplings between mutations, selection, and divergence throughout evolution. This method sidesteps the difficulty of measuring mutational effects from incomplete and complex descriptions of protein structures and their dynamics, interactions, and pathways in and across cells. Instead, it only requires three quantities f, $\Delta\gamma$, and $\int f'$. We showed previously how to compute f' and $\Delta\gamma$ (Katsonis and Lichtarge 2014) and now by computing $\int f' = CI$ we complete this calculus of fitness landscape. The result may appear surprising, but is consistent with the ability of calculus to routinely solve otherwise seemingly intractable problems (Penrose 2007) and is also consistent with the view that biology is likely to follow statistical thermodynamic rules for the large scale behavior of complex systems that eschews details of internal structures and forces (Sella and Hirsh 2005; Koonin 2011). Indeed, we find that mutations in genes that are under no specific selective force follow a random distribution. By contrast, genes under selection such as those that drive a group of individuals to a specific location of the fitness landscape will be typified by a non-random distribution. In principle, this is a general approach to identify the genotype determinants of the phenotype specific to a population (Kim et al. 2020; Koire et al. 2021), as we show here for cancer.

METHODS

Calculation of the Evolutionary Action score of coding missense variants. The Evolutionary Action (EA) is an untrained and formal *model* to measure the fitness effect of missense variants analytically, using protein evolution data from homologous sequences. Thus, EA predictions of fitness effects are not based on or adjusted according to experimental values. As described in detail elsewhere (see Supplementary Material and (Katsonis and Lichtarge 2014), in order to estimate the fitness effect of variants, EA considers the fitness landscape to be a mapping from genotype (γ) to phenotype fitness values (φ) via the evolutionary function $f: f(\gamma) = \varphi$ (Eq. 1). In essence, Eq. 1 describes the genotype-phenotype relationship. Assuming genotypes are evolvable, f should be differentiable, such that to first order a sufficiently small genotype variant ($\Delta\gamma$) due to a mutation would lead to a fitness perturbation ($\Delta\varphi$) given by: $f'(\gamma) \cdot \Delta\gamma \approx \Delta\varphi = EA$ (Eq. 2), where *EA* is the evolutionary action of the mutation $\Delta\gamma$ on fitness and $f'(\gamma)$ is the functional sensitivity of the mutated site to genotype variants. Although the function f in Eq. 1 is unknown, we may still compute Eq. 2 in the special context of amino acid substitution variants by approximating its two terms. First, $f'(\gamma)$ is approximated with Evolutionary Trace (ET) ranks of importance (Lichtarge et al. 1996). ET measures the functional importance of protein residues by accounting for the phylogenetic distances ($\Delta\varphi$) between homologous sequences given a variation ($\Delta\gamma$) at a given residue position. Thus, positions that tend to vary mostly between phylogenetically distant homologous sequences are ranked as more important. Second, $\Delta\gamma$ is approximated for coding variants from amino acid substitution odds. These substitution odds reflect the differences in various physicochemical properties of amino acids and are calculated from numerous homologous sequence pairs, for the specific functional importance of the mutated positions. Thus, Alanine to Serine has greater substitution odds, and lesser contribution to EA than Alanine to Tryptophan, in keeping the greater similarities between the former pair than the latter. These are first order approximations and local functional and structural features of proteins will impact their accuracy. But these local effects are treated as second order effects, and for now ignored. In practice, this is justified since EA performs well in objective tests against state-of-the-art methods to assess the deleterious impact of mutations (Katsonis and Lichtarge 2017; Katsonis and Lichtarge 2019). EA scores for human gene variants are available for non-profit use at http://eaction.lichtargelab.org/.

Cohort Integration (CI). The methodological innovation in this work is to integrate Eq. 2 in order to recover the genotype-phenotype relationship from Eq. 1. In all generality, integrals are sums of terms, called integrands, each one evaluated from a function of a variable over a range, called domain. Here, the variable is the genotype variation, the function is Eq. 2, the integrands are the EA scores, and the domain is all mutations in a patient cohort compared to control mutations (see Supplementary Materials for details). Therefore, the integral will be a sum of EA scores for patient variants compared to the sum of random variants. Considering gene-specific distributions of EA scores, we may calculate integrals gene by gene, and with a one-tailed, two-sample Kolmogorov–Smirnov test we may find whether these integrals differ significantly from zero. Insignificant differences indicate genes with coding variants in the cancer cohort that are consistent with random mutational events, not under selection. However, significant Kolmogorov-Smirnov tests indicate genes with variants that experienced selection in the cancer cohort, i.e. genes with genotype perturbations linked to the phenotype of the cohort. To eliminate genes with EA distributions that are not biased to high or intermediate EA scores (different than those seen in Fig. 2C-D), we asked that either sLOF index is greater than 0.15 or sGOF index is greater than 0.1. This integral analysis only used missense and nonsense mutations compared to random nucleotide changes. Frameshift insertions and deletions (fs-indels) were treated separately because they require different control variants. For fs-indels, we used the binomial probability test to compare their frequency in each gene to their frequency in the genome, assuming that most tumor mutations do not contribute to cancer (passengers) and that the rate of passenger mutations is constant throughout the genome. As such, the fs-indel mutation rate was calculated in the context of each cancer type, by dividing the total number of fs-indel mutations by the combined size of the genes with at least one mutation. We combined for each gene the CI *p*-value (missense and nonsense mutations) with the *p*-value for fs-indels according to Fisher's combined probability (Fisher 1948), and then corrected for multiple-hypothesis testing (*q*-value < 0.1) (Storey and Tibshirani 2003). Because the fs-indels analysis is very sensitive to sequencing quality and in order to avoid potentially false positive findings, we further filtered out genes with missense and nonsense CI *p*-value less than 0.05. In practice, this filtering prevented discoveries based on fs-indel variants only and acted to lower the stringent threshold of the missense and nonsense variant CI analysis. A flowchart of the CI approach is provided in Supplementary Fig. 10.

Tumor suppressor and oncogene classification with sLOF and sGOF index. We defined the sLOF index and sGOF index to quantify the bias of coding variants towards high and intermediate EA scores, respectively. To do so, we considered the distribution of EA scores that corresponds to variants obtained from all possible nucleotide changes in each gene. Then, we calculated a reference curve of how the average of the distribution changes when we exclude any fractions of the variants with the lowest (positive selection) or the highest (negative selection) EA scores. A positive sLOF index equal to x indicates that the average EA score of the given coding variants is equal to the averaged EA score of all possible nucleotide substitutions in a gene minus fraction x of the substitutions with the lowest EA scores, while a negative sLOF index $-x$ indicates that the given coding variants have an average EA equal to that of all possible nucleotide substitutions minus fraction x of the substitutions with the highest EA. To account for bias to intermediate EA scores, we considered the minimum difference of each EA score from 0 or from 100 (DEA values). The DEA values for all possible nucleotide changes in a gene created a new distribution scaled from 0 to 50. We calculated a reference curve of how the average of the distribution changes when we excluded any fractions of the variants with the lowest (positive selection) or the highest (negative selection) DEA values. The positive sGOF index equal to y indicates that the average DEA value of given coding variants is equal to the average DEA value of all possible nucleotide substitutions minus the fraction *y* of the substitutions with the lowest DEA values (most distant from an EA of 50), and a negative sGOF index -*y* indicates that the given coding variants have average EA equal to that of all possible nucleotide substitutions minus the fraction *y* of the substitutions with the highest DEA values (closest to 50). In sum, a positive sLOF indicates the selection characteristic of a mutated tumor suppressor gene, while a positive sGOF indicates the selection characteristic of a mutated oncogene. We then annotate genes as tumor suppressors when their sLOF index > 0.15 and as oncogenes when the sGOF index > 0.1. Genes that satisfy both the tumor suppressor and oncogene criteria are annotated as oncogenes, since oncogenes typically also have positive sLOF indices (due to biases of the genetic code to conservative substitutions), whereas tumor suppressors typically have negative sGOF indices.

SOFTWARE AVAILABILITY

The Cohort Integration (CI) software is available method is available in Supplemental Material together with a README.txt file that contains the instructions for installation and execution. The method is also available through the web server: http://cohort.lichtargelab.org/.

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (GM079656 and GM066099 to O.L., R01 DE14613 to M.K., and DE025181 to J.N.M.); the National Science Foundation (DBI1356569 to O.L.). J.A. was supported by a training fellowship from the Gulf Coast Consortia, on the NLM Training Program in Biomedical Informatics & Data Science (T15LM007093). A.K. was supported by RP160283 - Baylor College of Medicine Comprehensive Cancer Training Program, and the Baylor Research Advocates for Student Scientists (BRASS).

AUTHOR CONTRIBUTIONS

T.-K.H., P.K., and O.L. designed the study and interpreted the results. T.-K.H., P.K., M.K., and O.L developed the methodology. T.-K.H., J.A., P.K., C.-H.L., E.H., and D.M.K. analyzed the data. B.-K. C., M.A.G. performed the experiments. T.-K.H., P.K., J.A., C.-H.L., D.M.K., B.-K. C., C.R.P., M.J.F., and O.L. wrote the manuscript. T.-K.H., P.K., J.A., A.K., C.R.P., M.J.F., L.A.D., M.K., J.N.M., and O.L. reviewed and edited the manuscript. P.K. and O.L. supervised the research.

FIGURE LEGENDS

Figure 1. Evolutionary calculus of the genotype-phenotype relationship. **(A)** Our hypothesis is that the genotype (γ) is linked to phenotype (φ) through a continuous evolutionary function (f). **(B)** The fitness effect ($\Delta\varphi$), or Evolutionary Action (EA), of each mutation is the product of evolutionary gradient (f') and genotypic change ($\Delta\gamma$). Evolutionary gradient is the sensitivity of the mutated position to substitution approximated by Evolutionary Trace (ET), which accounts for the phylogenetic distances between homologous sequences ($\Delta\varphi$) that vary at a residue position ($\Delta\gamma$); the size of genotypic change can be approximated by substitution odds. **(C)** EA score calculation for the R175H variant of *TP53*. The evolutionary gradient (f') of position 175 was measured by ET ranks of importance and the genotypic change ($\Delta\gamma$) of R to H substitution was measured by ranks of context dependent substitution log-odds. These two terms were then multiplied and normalized to yield EA score. **(D)** To identify genes under positive selection in a trait-associated cohort, for every gene k, we compared the cohort integral, represented by the distribution of EA scores, of its cancer somatic mutations in the patient cohort (C_j) with the cohort integral of random mutations. A non-random cohort integral indicates the gene k harbors cancer driver mutations and therefore is a cancer driver gene. **(E)** Cancer driver genes are identified by non-random cohort integrals. Cohort integral of somatic mutations was significantly

different from random (one-tailed two-sample Kolmogorov–Smirnov test) for the cancer gene *TP53*, in contrast to the non-cancer gene *DNAH5*.

Figure 2. Coding variants under selection have non-random cohort integrals. The cohort integrals for **(A)** all germline variants in the 1000 Genomes Project and **(B)** all somatic mutations from The Cancer Genome Atlas (TCGA). The somatic TCGA mutations were also shown for **(C)** the tumor suppressor genes *TP53*, *CDKN2A*, *PTEN*, and *NOTCH1*, and for (*D*) the oncogenes *PIK3CA*, *BRAF*, *KRAS*, and *NRAS*. The dashed lines correspond to simulated random amino acid changes in all human genes (A-B) or the respective single gene (C-D). The *p*-values of cohort integral differences between the observed and simulated random mutations were calculated by the one-tailed two-sample Kolmogorov–Smirnov test.

Figure 3. CI recovers cancer-driving genes in benchmarking and prospective analyses. **(A)** The performance of CI against ten state-of-the-art methods over the same input samples. The heatmap represents the relative performance of the methods (red means the best and white means the worst performance) for 6 evaluation metrics: area under the Receiver Operating Characteristic curve, deviation from the expected *p*-value distribution, overlap with the COSMIC Cancer Gene Census, overlap with the consensus of all the other methods, the consistency among cohort subsamples, and the area under the Precision-Recall curve. **(B)** CI identified 460 genes under positive selection in TCGA tumors. The number of tumor suppressors (blue) and oncogenes (red) identified by CI in each cancer type. **(C)** Heatmap representation of CI q-value of the 56 candidate genes that were identified in two or more cancer types. The significance level is represented by a color scale from red (more significant) to yellow (less significant). The cancer types were ordered according to a dendrogram of pairwise distances based on the overlap of predicted driver genes (see Methods).

Figure 4. CI distinguished tumor suppressor genes from oncogenes. **(A)** The selection for gain-of-function (sGOF) and loss-of-function (sLOF) indices for 54 known tumor suppressor genes (left) and 18 known oncogenes (right). Genes were plotted according to the sLOF index (*x*-axis) and sGOF index (*y*-axis) for the cancer type with the most significant *q*-value, and the circle color indicates the CI *q*-value for the most significant cancer type, which is represented by a color scale from red (more significant) to yellow (less significant). Genes located in the red rectangular area are classified as oncogenes, and genes located in the blue rectangular area are tumor suppressor genes. **(B)** Genome-scale CRISPR gene dependency screen validated CI oncogenes. Oncogenes from the COSMIC database (COSMIC), identified by the CI method across all cancers (CI PAN) or from individual cancer types (CI Indiv.) show a statistically significant shift toward essentiality (Ceres Score <= -0.6) when harboring variants of moderate EA range (30 <= EA score < 70) as opposed to other mutations, including low EA variants (0 <= EA < 30), high EA variants (70 <=EA<100), nonsense variants, and other uncategorized variants. No difference was observed for random genes. Statistical significance was calculated with Mann–Whitney *U* test.

Figure 5. CI identified 45 genes under positive selection in head and neck squamous carcinoma (HNSC). **(A)** The 45 HNSC candidate genes were plotted according to their sLOF index (*x*-axis) and sGOF index (*y*-axis) as circles. The color of each circle indicates the CI *q*-value in HNSC in a color scale of red (more significant) to yellow (less significant). Genes located in the red rectangular area were classified as oncogenes (names shown in red), and genes located in the blue rectangle area were classified as tumor suppressor genes (names shown in blue). **(B)** Pathways associated with CI-identified candidates in HNSC.

Figure 6. Overexpressing *CUL3* in HNSC cell lines suppresses tumor growth. **(A)** Somatic missense mutations in *CUL3* are shown as spheres in the structure of a homologous protein

(chain C of PDB: 2HYE, sequence identity = 39.82%), and the EA score of the given mutation is shown in parentheses. The importance of position was evaluated by Evolutionary Trace (ET), and is represented by a color scale from red (more important) to green (less important). **(B)** Clonogenic assay following infection of the cell lines UMSCC25, HN31, UMSCC22A, and MDA1386LN with lentivirus to express wild-type *CUL3* or the control gene *LacZ*. **(C)** Western blot shows the ectopic expression of wild-type *CUL3* after lentivirus infection. **(D)** Cell cycle distribution for the UMSCC25 cell line following expression of *CUL3*. **(E)** *PARP* levels, a marker of apoptotic death, increase in UMSCC25 cells expressing *CUL3*, **(F)** while levels of NRF2, the HNSC-driving protein, are reduced. Overexpression of *DUSP16* inhibited cell proliferation, colony formation, migration, and induced apoptosis by inhibition of JNK pathway. **(G)** *DUSP16* overexpression was assessed by induction with doxycycline (1µg/ml) to stable HEK293 and BT474 cells containing inducible dusp16 construct. **(H)** Overexpression of *DUSP16* inhibited cell proliferation of established cells. **(I)** Clonogenic assay: overexpression of *DUSP16* inhibited colony formation of HEK293 and BT474 cells. The experiment was performed in triplicates and the density of the stained cells was measured at 630 nm after extraction with 10% acetic acid. **(J)** In vitro scratch assay. Cells were plated and incubated until confluent, then scratched with pipette tip and further incubated to compare recovery of scratched area. **(K)** Overexpression of *DUSP16* induced apoptosis in stable BT474 cells. **(L)** Overexpression of *DUSP16* inhibited c-Jun phosphorylation in stable HEK293 and BT474 cells. Overexpression of *DUSP16* induced down-regulation of Jun and Fos transcription in **(M)** HEK293 cells and **(N)** BT474 cells. **(O)** Depletion of *DUSP16* by CRISPR antagonized cell growth inhibition by overexpression of *DUSP16*. * p-value < 0.05, ** p-value < 0.005 and *** p-value < 0.001.

REFERENCES

Abbas W, Kumar A, Herbein G. 2015. The eEF1A Proteins: At the Crossroads of Oncogenesis, Apoptosis, and Viral Infections. *Front Oncol* **5**: 75.

Agrawal N, Frederick MJ, Pickering CR, Bettegowda C, Chang K, Li RJ, Fakhry C, Xie TX, Zhang J, Wang J et al. 2011. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science* **333**(6046): 1154-1157.

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN et al. 2020. The repertoire of mutational signatures in human cancer. *Nature* **578**(7793): 94-101.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL et al. 2013. Signatures of mutational processes in human cancer. *Nature* **500**(7463): 415-421.

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B et al. 2018. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**(2): 371-385.

Bignell GR, Greenman CD, Davies H, Butler AP, Edkins S, Andrews JM, Buck G, Chen L, Beare D, Latimer C et al. 2010. Signatures of mutation and selection in the cancer genome. *Nature* **463**(7283): 893-898.

Boutros M, Ahringer J. 2008. The art and design of genetic screens: RNA interference. *Nature reviews Genetics* **9**(7): 554-566.

Carneiro M, Hartl DL. 2010. Colloquium papers: Adaptive landscapes and protein evolution. *Proc Natl Acad Sci U S A* **107 Suppl 1**: 1747-1751.

Carter M, Brunt BV. 2012. The Lebesgue-Stieltjes Integral: A Practical Introduction. *Springer Science & Business Media, , Chapter 4, page 49-67*.

Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, Gao J, Socci ND, Solit DB, Olshen AB et al. 2016. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol* **34**(2): 155-163.

Chang MT, Bhattarai TS, Schram AM, Bielski CM, Donoghue MTA, Jonsson P, Chakravarty D, Phillips S, Kandoth C, Penson A et al. 2018. Accelerating Discovery of Functional Mutant Alleles in Cancer. *Cancer Discov* **8**(2): 174-183.

Chin L, Andersen JN, Futreal PA. 2011. Cancer genomics: from discovery science to personalized medicine. *Nat Med* **17**(3): 297-303.

Chun YS, Passot G, Yamashita S, Nusrat M, Katsonis P, Loree JM, Conrad C, Tzeng CD, Xiao L, Aloia TA et al. 2019. Deleterious Effect of RAS and Evolutionary High-risk *TP53* Double Mutation in Colorectal Liver Metastases. *Annals of surgery* **269**(5): 917-923.

Clarke CN, Katsonis P, Hsu TK, Koire AM, Silva-Figueroa A, Christakis I, Williams MD, Kutahyalioglu M, Kwatampora L, Xi Y et al. 2019. Comprehensive Genomic Characterization of Parathyroid Cancer Identifies Novel Candidate Driver Mutations and Core Pathways. *Journal of the Endocrine Society* **3**(3): 544-559.

Corso G, Montagna G, Figueiredo J, La Vecchia C, Fumagalli Romario U, Fernandes MS, Seixas S, Roviello F, Trovato C, Guerini-Rocco E et al. 2020. Hereditary Gastric and Breast Cancer Syndromes Related to CDH1 Germline Mutation: A Multidisciplinary Clinical Review. *Cancers* **12**(6).

Cunningham BC, Wells JA. 1989. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* **244**(4908): 1081-1085.

Dang Y, Low WK, Xu J, Gehring NH, Dietz HC, Romo D, Liu JO. 2009. Inhibition of nonsense-mediated mRNA decay by the natural product pateamine A through eukaryotic initiation factor 4AIII. *J Biol Chem* **284**(35): 23613-23621.

Datta D, Scheer JM, Romanowski MJ, Wells JA. 2008. An allosteric circuit in caspase-1. *Journal of molecular biology* **381**(5): 1157-1167.

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.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER et al. 2012. MuSiC: identifying mutational significance in cancer genomes. *Genome research* **22**(8): 1589-1598.

Dietlein F, Weghorn D, Taylor-Weiner A, Richters A, Reardon B, Liu D, Lander ES, Van Allen EM, Sunyaev SR. 2020. Identification of cancer driver genes based on nucleotide context. *Nature genetics* **52**(2): 208-218.

Doebeli M, Ispolatov Y, Simon B. 2017. Towards a mechanistic foundation of evolutionary theory. *eLife* **6**.

Domotor E. 1989. Intensive postoperative mydeton therapy of traumatological patients. *Ther Hung* **37**(4): 230-233.

Esteller M. 2008. Epigenetics in cancer. *N Engl J Med* **358**(11): 1148-1159.

Fernandez-Martinez P, Zahonero C, Sanchez-Gomez P. 2015. DYRK1A: the double-edged kinase as a protagonist in cell growth and tumorigenesis. *Mol Cell Oncol* **2**(1): e970048.

Fisher R. 1930. *The genetical theory of natural selection*.

Fisher R. 1948. Questions and answers #14. *The American Statistician* **2**(5): 30-31.

Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole CG, Ward S, Dawson E, Ponting L et al. 2017. COSMIC: somatic cancer genetics at high-resolution. *Nucleic acids research* **45**(D1): D777-D783.

Fowler DM, Stephany JJ, Fields S. 2014. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat Protoc* **9**(9): 2267-2284.

Garraway LA, Lander ES. 2013. Lessons from the cancer genome. *Cell* **153**(1): 17-37.

Gonzalez-Perez A, Lopez-Bigas N. 2012. Functional impact bias reveals cancer drivers. *Nucleic acids research* **40**(21):e169.

Greaves M, Maley CC. 2012. Clonal evolution in cancer. *Nature* **481**(7381): 306-313.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C et al. 2007. Patterns of somatic mutation in human cancer genomes. *Nature* **446**(7132): 153-158.

Guan Y, Gorenshteyn D, Burmeister M, Wong AK, Schimenti JC, Handel MA, Bult CJ, Hibbs MA, Troyanskaya OG. 2012. Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS Comput Biol* **8**(9): e1002694.

Hardy J, Singleton A. 2009. Genomewide association studies and human disease. *N Engl J Med* **360**(17): 1759-1768.

Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S et al. 2015. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**(6): 1515-1526.

Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nature reviews Genetics* **6**(2): 95-108.

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V et al. 2014. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**(4): 929-944.

Hoornaert I, Marynen P, Goris J, Sciot R, Baens M. 2003. MAPK phosphatase DUSP16/MKP-7, a candidate tumor suppressor for chromosome region 12p12-13, reduces BCR-ABL-induced transformation. *Oncogene* **22**(49): 7728-7736.

Hori T, Osaka F, Chiba T, Miyamoto C, Okabayashi K, Shimbara N, Kato S, Tanaka K. 1999. Covalent modification of all members of human cullin family proteins by NEDD8. *Oncogene* **18**(48): 6829-6834.

Hu T, Yang H, Han ZG. 2015a. PDZRN4 acts as a suppressor of cell proliferation in human liver cancer cell lines. *Cell Biochem Funct* **33**(7): 443-449.

Hu Y, Xing J, Chen L, Zheng Y, Zhou Z. 2015b. RGS22 inhibits pancreatic adenocarcinoma cell migration through the G12/13 alpha subunit/F-actin pathway. *Oncol Rep* **34**(5): 2507-2514.

Hu Y, Xing J, Wang L, Huang M, Guo X, Chen L, Lin M, Zhou Y, Liu Z, Zhou Z et al. 2011. RGS22, a novel cancer/testis antigen, inhibits epithelial cell invasion and metastasis. *Clin Exp Metastasis* **28**(6): 541-549.

Joo MK, Park JJ, Chun HJ. 2016. Impact of homeobox genes in gastrointestinal cancer. *World J Gastroenterol* **22**(37): 8247-8256.

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.

Katoh M, Katoh M. 2004. Identification and characterization of PDZRN3 and PDZRN4 genes in silico. *Int J Mol Med* **13**(4): 607-613.

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.

Katsonis P, Lichtarge O. 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.

Katsonis P, Lichtarge O. 2019. CAGI5: Objective performance assessments of predictions based on the Evolutionary Action equation. *Hum Mutat*.

Keyse SM. 2008. Dual-specificity MAP kinase phosphatases (MKPs) and cancer. *Cancer Metastasis Rev* **27**(2): 253-261.

Kim YW, Al-Ramahi I, Koire A, Wilson SJ, Konecki DM, Mota S, Soleimani S, Botas J, Lichtarge O. 2020. Harnessing the paradoxical phenotypes of APOE varepsilon2 and APOE varepsilon4 to identify genetic modifiers in Alzheimer's disease. *Alzheimers Dement* **17***:* 831-846.

Koike-Yusa H, Li Y, Tan EP, Velasco-Herrera Mdel C, Yusa K. 2014. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol* **32**(3): 267-273.

Koire A, Katsonis P, Kim YW, Buchovecky C, Wilson SJ, Lichtarge O. 2021. A method to delineate de novo missense variants across pathways prioritizes genes linked to autism. *Sci Transl Med* **13**(594):eabc1739.

Koonin EV. 2011. Are there laws of genome evolution? *PLoS Comput Biol* **7**(8): e1002173.

Korte A, Farlow A. 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9**: 29.

Kramer A, Green J, Pollard J, Jr., Tugendreich S. 2014. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**(4): 523-530.

Landau LD, Lifshitz EM. 1980. *Statistical Physics. Course of Theoretical Physics*. Oxford: Pergamon Press.

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.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**(7457): 214-218.

Lee K, Lindsey AS, Li N, Gary B, Andrews J, Keeton AB, Piazza GA. 2016. beta-catenin nuclear translocation in colorectal cancer cells is suppressed by PDE10A inhibition, cGMP elevation, and activation of PKG. *Oncotarget* **7**(5): 5353-5365.

Lee S, Kong Y, Weatherbee SD. 2013. Forward genetics identifies Kdf1/1810019J16Rik as an essential regulator of the proliferation-differentiation decision in epidermal progenitor cells. *Dev Biol* **383**(2): 201-213.

LeGall J-F. 2016. Brownian Motion, Martingales, and Stochastic Calculus. *Springer*

International Publishing, Graduate Texts in Mathematics, Springer, New York, Chapter 5.

Li N, Lee K, Xi Y, Zhu B, Gary BD, Ramirez-Alcantara V, Gurpinar E, Canzoneri JC, Fajardo A, Sigler S et al. 2015. Phosphodiesterase 10A: a novel target for selective inhibition of colon tumor cell growth and beta-catenin-dependent TCF transcriptional activity. *Oncogene* **34**(12): 1499-1509.

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, 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.

Mak HC, Justman Q. 2017. Genotype-Phenotype Mapping Meets Single Cell Biology. *Cell Syst* **4**(1): 1-2.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-guided human genome engineering via Cas9. *Science* **339**(6121): 823-826.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**(7265): 747-753.

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.

Martinez-Jimenez F, Muinos F, Sentis I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, Mularoni L, Pich O, Bonet J, Kranas H et al. 2020. A compendium of mutational cancer driver genes. *Nature reviews Cancer* **20**(10): 555-572.

Martinez VD, Vucic EA, Thu KL, Pikor LA, Lam S, Lam WL. 2015. Disruption of KEAP1/CUL3/RBX1 E3-ubiquitin ligase complex components by multiple genetic mechanisms: Association with poor prognosis in head and neck cancer. *Head Neck* **37**(5): 727-734.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews Genetics* **9**(5): 356-369.

Meeusen B, Janssens V. 2018. Tumor suppressive protein phosphatases in human cancer: Emerging targets for therapeutic intervention and tumor stratification. *The international journal of biochemistry & cell biology* **96**: 98-134.

Mertens F, Johansson B, Fioretos T, Mitelman F. 2015. The emerging complexity of gene fusions in cancer. *Nature reviews Cancer* **15**(6): 371-381.

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.

Miller KR, Patel JN, Ganapathi MK, Tait DL, Ganapathi RN. 2016. Biological role and clinical implications of homeobox genes in serous epithelial ovarian cancer. *Gynecol Oncol* **141**(3): 608-615.

Mitelman F, Johansson B, Mertens F. 2007. The impact of translocations and gene fusions on cancer causation. *Nature reviews Cancer* **7**(4): 233-245.

Nei M. 2005. Selectionism and neutralism in molecular evolution. *Mol Biol Evol* **22**(12): 2318-2342.

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.

Ohta T, Iijima K, Miyamoto M, Nakahara I, Tanaka H, Ohtsuji M, Suzuki T, Kobayashi A, Yokota J, Sakiyama T et al. 2008. Loss of Keap1 function activates Nrf2 and provides advantages for lung cancer cell growth. *Cancer Res* **68**(5): 1303-1309.

Orr HA. 2009. Fitness and its role in evolutionary genetics. *Nature reviews Genetics* **10**(8): 531-539.

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.

Pal G, Kossiakoff AA, Sidhu SS. 2003. The functional binding epitope of a high affinity variant of human growth hormone mapped by shotgun alanine-scanning mutagenesis: insights into the mechanisms responsible for improved affinity. *Journal of molecular biology* **332**(1): 195-204.

Pal G, Kouadio JL, Artis DR, Kossiakoff AA, Sidhu SS. 2006. Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. *J Biol Chem* **281**(31): 22378-22385.

Patterson KI, Brummer T, O'Brien PM, Daly RJ. 2009. Dual-specificity phosphatases: critical regulators with diverse cellular targets. *The Biochemical journal* **418**(3): 475-489.

Penrose MD. 2007. Laws of Large Numbers in Stochastic Geometry with Statistical Applications. *Bernoulli* **13**(4): 1124-1150.

Pickering CR, Zhang J, Yoo SY, Bengtsson L, Moorthy S, Neskey DM, Zhao M, Ortega Alves MV, Chang K, Drummond J et al. 2013. Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. *Cancer Discov* **3**(7): 770-

781.

Pickering CR, Zhou JH, Lee JJ, Drummond JA, Peng SA, Saade RE, Tsai KY, Curry JL, Tetzlaff MT, Lai SY et al. 2014. Mutational landscape of aggressive cutaneous squamous cell carcinoma. *Clin Cancer Res* **20**(24): 6582-6592.

Pintard L, Willems A, Peter M. 2004. Cullin-based ubiquitin ligases: Cul3-BTB complexes join the family. *EMBO J* **23**(8): 1681-1687.

Porta-Pardo E, Godzik A. 2014. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics* **30**(21): 3109-3114.

Radhakrishnan A, Nanjappa V, Raja R, Sathe G, Puttamallesh VN, Jain AP, Pinto SM, Balaji SA, Chavan S, Sahasrabuddhe NA et al. 2016. A dual specificity kinase, DYRK1A, as a potential therapeutic target for head and neck squamous cell carcinoma. *Sci Rep* **6**: 36132.

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.

Rice DS, Northcutt GM, Kurschner C. 2001. The Lnx family proteins function as molecular scaffolds for Numb family proteins. *Mol Cell Neurosci* **18**(5): 525-540.

Santos AO, Parrini MC, Camonis J. 2016. RalGPS2 Is Essential for Survival and Cell Cycle Progression of Lung Cancer Cells Independently of Its Established Substrates Ral GTPases. *PLoS One* **11**(5): e0154840.

Schwartz L. 1963. Some applications of the theory of distributions, In: Lectures on Modern Mathematics. **1**: 23-58.

Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci U S A* **102**(27): 9541-9546.

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

Shlien A, Malkin D. 2009. Copy number variations and cancer. *Genome Med* **1**(6): 62.

Soderling SH, Bayuga SJ, Beavo JA. 1999. Isolation and characterization of a dual-substrate phosphodiesterase gene family: PDE10A. *Proc Natl Acad Sci U S A* **96**(12): 7071-7076.

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.

Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, Kryukov GV, Lawrence MS, Sougnez C, McKenna A et al. 2011. The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**(6046): 1157-1160.

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 research* **43**(Database issue): D447-452.

Takeshima H, Niwa T, Takahashi T, Wakabayashi M, Yamashita S, Ando T, Inagawa Y, Taniguchi H, Katai H, Sugiyama T et al. 2015. Frequent involvement of chromatin remodeler alterations in gastric field cancerization. *Cancer Lett* **357**(1): 328-338.

Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. 2013. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**(18): 2238-2244.

Tetreault MP, Yang Y, Katz JP. 2013. Kruppel-like factors in cancer. *Nature reviews Cancer* **13**(10): 701-713.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.

The Cancer Genome Atlas Research Network. 2015. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**(7536): 576-582.

The Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**(7216): 1061-1068.

The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics* **45**(10): 1113-1120.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. Pan-cancer analysis of whole genomes. *Nature* **578**(7793): 82-93.

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.

Tomasetti C, Marchionni L, Nowak MA, Parmigiani G, Vogelstein B. 2015. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc Natl Acad Sci U S A* **112**(1): 118-123.

Tomczak K, Czerwinska P, Wiznerowicz M. 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* **19**(1A): A68-77.

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.

Van den Eynden J, Fierro AC, Verbeke LP, Marchal K. 2015. SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering. *BMC Bioinformatics* **16**: 125.

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

Wagner GP, Altenberg L. 1996. Perspective: Complex Adaptations and the Evolution of Evolvability. *Evolution* **50**(3): 967-976.

Wang J, Tian X, Han R, Zhang X, Wang X, Shen H, Xue L, Liu Y, Yan X, Shen J et al. 2014a. Downregulation of miR-486-5p contributes to tumor progression and metastasis by targeting protumorigenic ARHGAP5 in lung cancer. *Oncogene* **33**(9): 1181-1189.

Wang T, Wei JJ, Sabatini DM, Lander ES. 2014b. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**(6166): 80-84.

Weinstein IB, Joe A. 2008. Oncogene addiction. *Cancer Res* **68**(9): 3077-3080; discussion 3080.

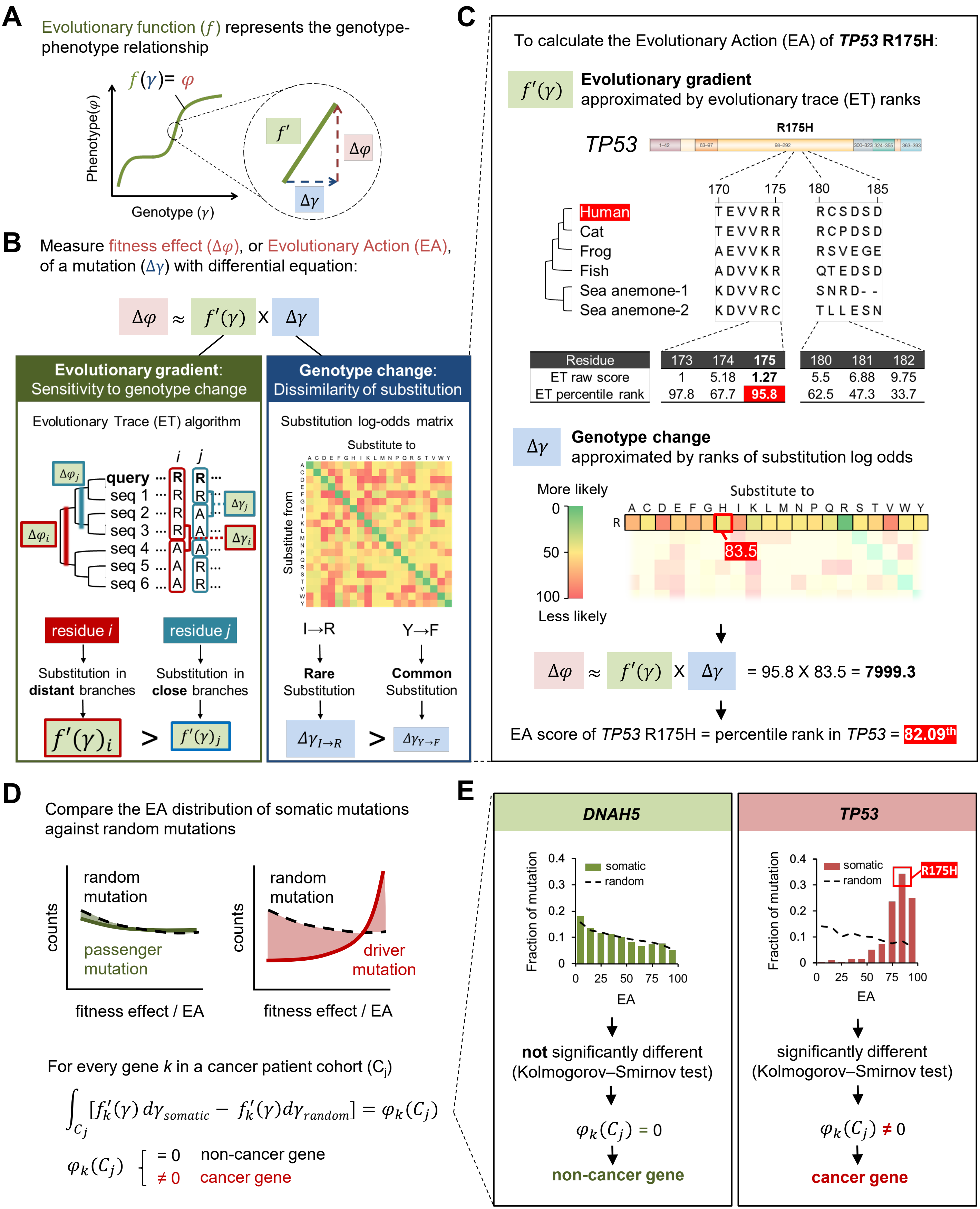
Wilkins AD, Lua R, Erdin S, Ward RM, Lichtarge O. 2010. Sequence and structure continuity of evolutionary importance improves protein functional site discovery and annotation. *Protein Sci* **19**(7): 1296-1311.

Wright S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics* **1**: 356-366.

Yu KD, Shao ZM. 2009. Genetic matters of CYP2D6 in breast cancer: copy number variations and nucleotide polymorphisms. *Nature reviews Cancer* **9**(11): 842.

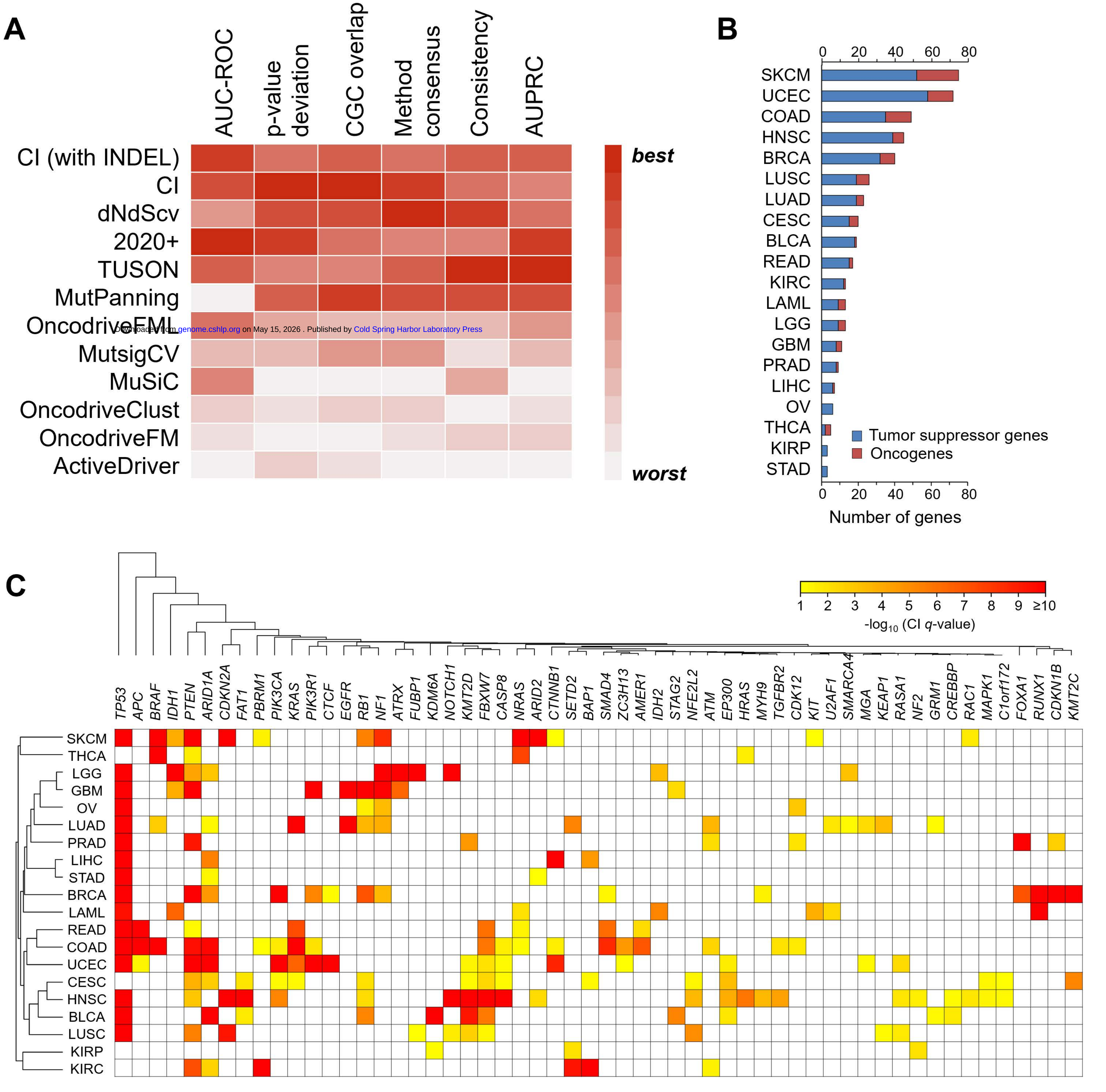
Yu KH, Berry GJ, Rubin DL, Re C, Altman RB, Snyder M. 2017. Association of Omics Features with Histopathology Patterns in Lung Adenocarcinoma. *Cell Syst* **5**(6): 620-627.

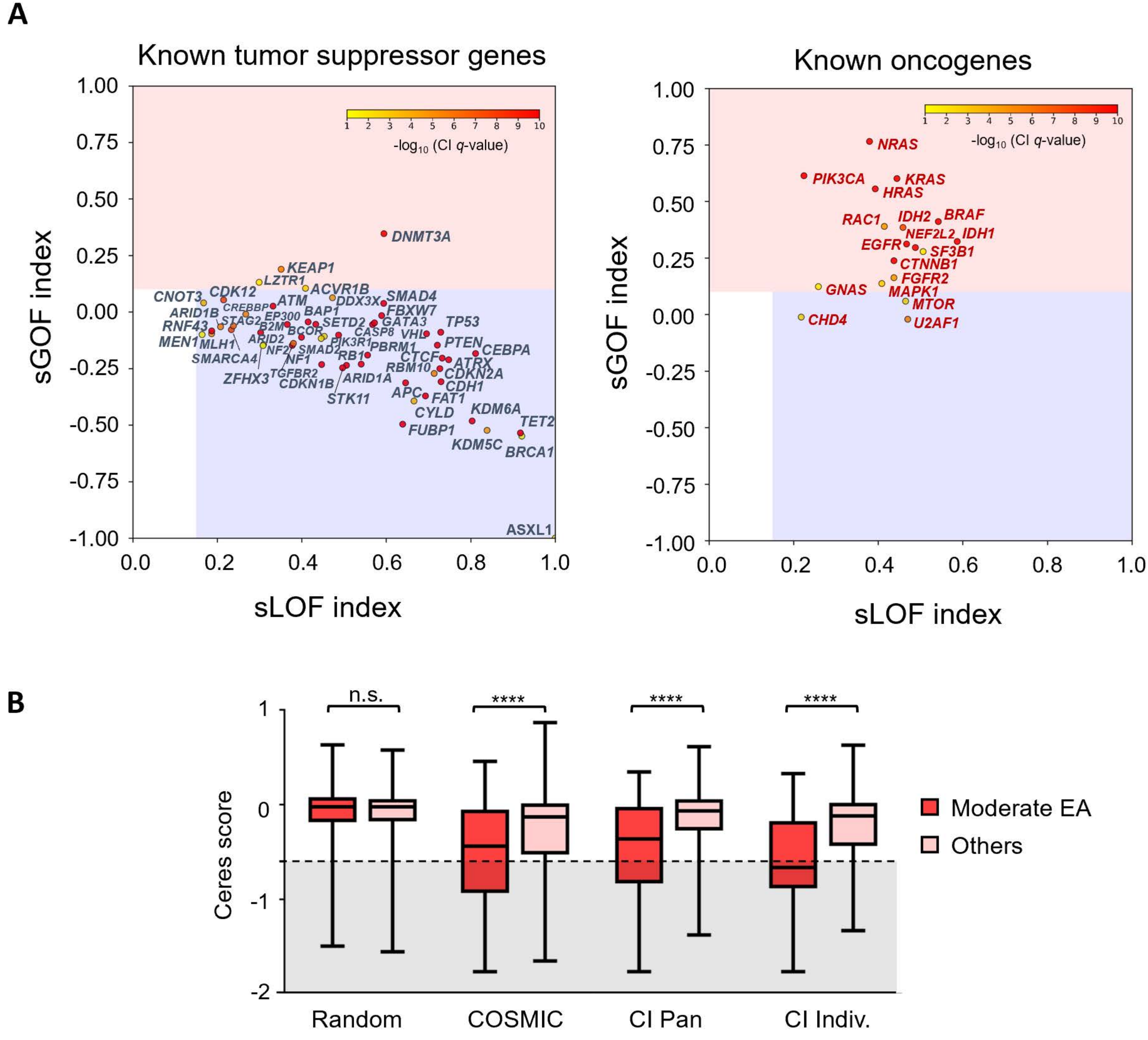
Yu KH, Zhang C, Berry GJ, Altman RB, Re C, Rubin DL, Snyder M. 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat Commun* **7**: 12474.



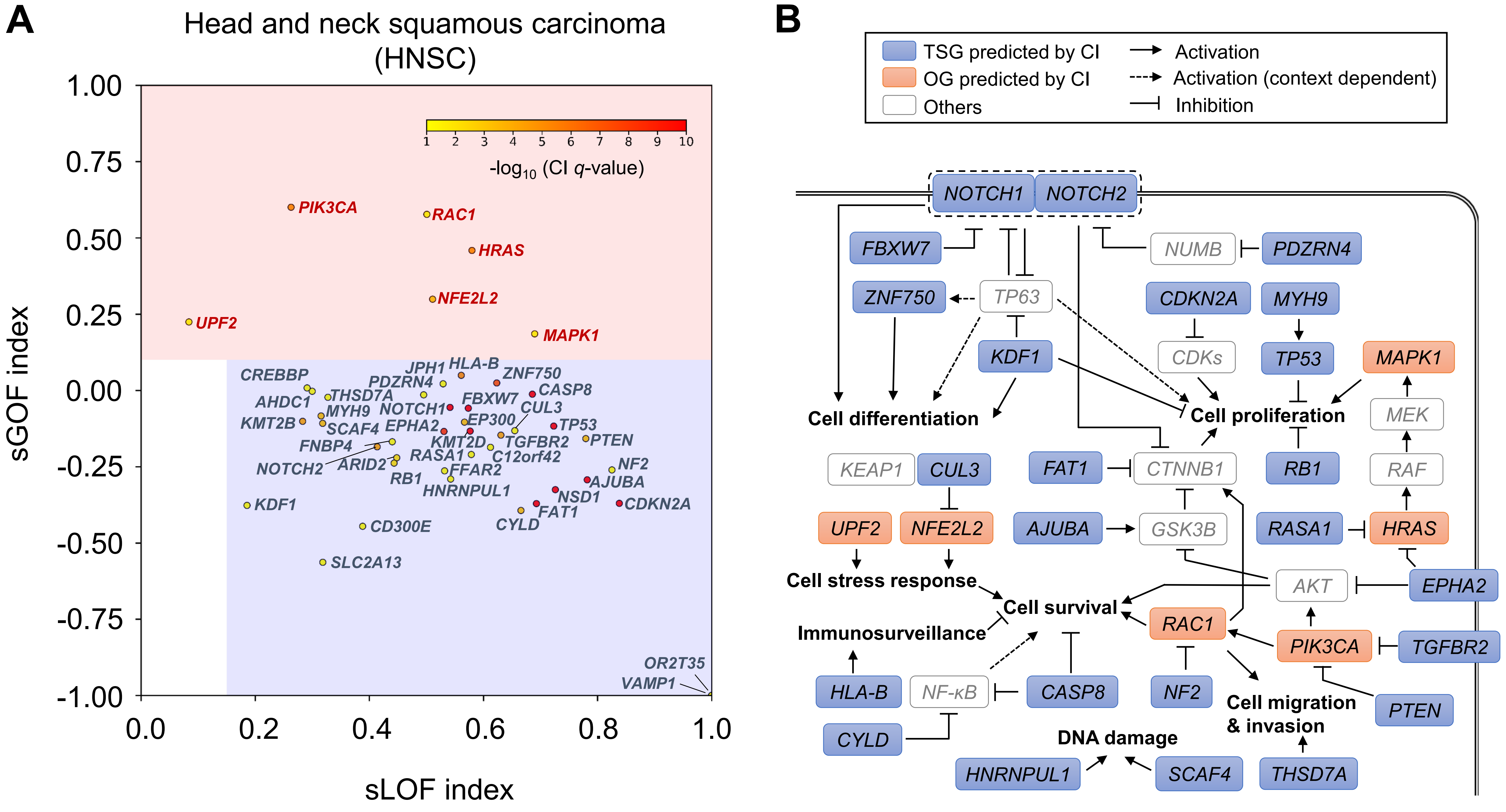
A
Evolutionary function (f) represents the genotype-phenotype relationship
f(γ)= φ
Phenotype(φ)
Genotype (γ)
f′
Δφ
Δγ
B
Measure fitness effect (Δφ), or Evolutionary Action (EA), of a mutation (Δγ) with differential equation:
Δφ ≈ f′(γ) X Δγ
Evolutionary gradient: Sensitivity to genotype change
Evolutionary Trace (ET) algorithm
query
seq 1
seq 2
seq 3
seq 4
seq 5
seq 6
residue i
residue j
Substitution in distant branches
Substitution in close branches
f′(γ)i > f′(γ)j
Genotype change: Dissimilarity of substitution
Substitution log-odds matrix
Substitute to
Substitute from
I→R
Y→F
Rare Substitution
Common Substitution
Δγ I→R > Δγ Y→F
C
To calculate the Evolutionary Action (EA) of TP53 R175H:
Evolutionary gradient
approximated by evolutionary trace (ET) ranks
TP53
R175H
Human
Cat
Frog
Fish
Sea anemone-1
Sea anemone-2
TEVVRR RCSDSD
TEVVRR RCPDSD
AEVVKR RSVEGE
ADVVKR QTEDSD
KDVVRC SNRD- -
KDVVRC TLLESN
Residue 173 174 175 180 181 182
ET raw score 1 5.18 1.27 5.5 6.88 9.75
ET percentile rank 97.8 67.7 95.8 62.5 47.3 33.7
Genotype change
approximated by ranks of substitution log odds
More likely
Less likely
Substitute to
83.5
Δφ ≈ f′(γ) X Δγ = 95.8 X 83.5 = 7999.3
EA score of TP53 R175H = percentile rank in TP53 = 82.09th
D
Compare the EA distribution of somatic mutations against random mutations
random mutation
passenger mutation
driver mutation
counts
fitness effect / EA
For every gene k in a cancer patient cohort (Cj)
$\int_{C_j} [f_k'(\gamma)\, d\gamma_{somatic} - f_k'(\gamma) d\gamma_{random}] = \varphi_k(C_j)$
$\varphi_k(C_j)$ = 0 non-cancer gene
≠ 0 cancer gene
E
DNAH5
TP53
somatic
random
R175H
Fraction of mutation
EA
not significantly different (Kolmogorov–Smirnov test)
significantly different (Kolmogorov–Smirnov test)
$\varphi_k(C_j)$ = 0
$\varphi_k(C_j)$ ≠ 0
non-cancer gene
cancer gene

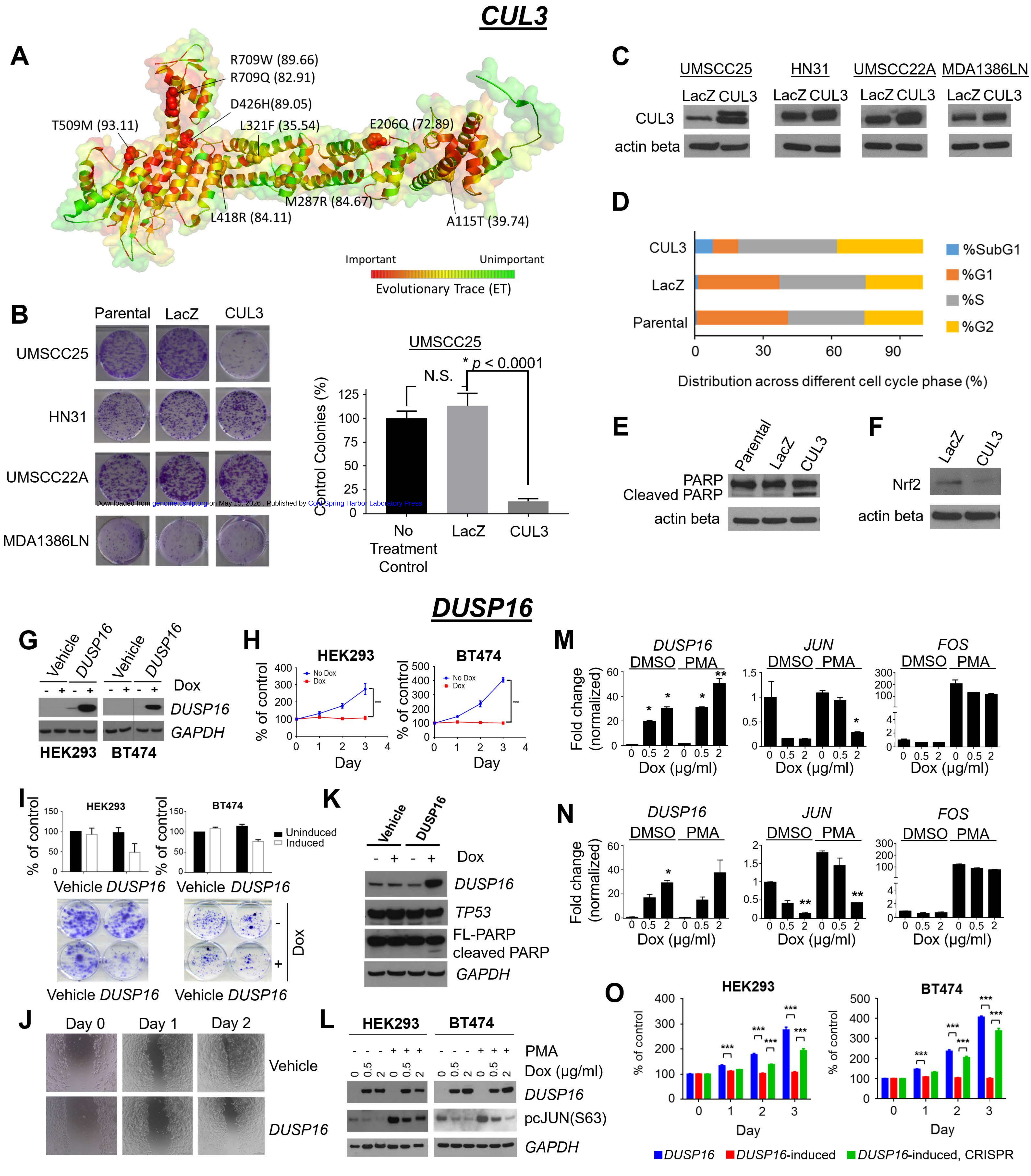

A
1000 Genomes
negative selection
Fraction of missense variants
1000 Genomes
Random
p=3.8×10-29
Evolutionary Action (EA)
B
Cancer somatic
random
Fraction of missense mutations
Cancer somatic
Random
p=0.41
Evolutionary Action (EA)
C
Tumor suppressor genes - positive selection for high impact mutations
somatic
random
Fraction of missense mutations
TP53
p=2.8×10-303
CDKN2A
p=2×10-7
PTEN
p=9.7×10-50
NOTCH1
p=1.4×10-11
Evolutionary Action (EA)
D
Oncogenes - positive selection for intermediate impact mutations
somatic
random
Fraction of missense mutations
PIK3CA
p=2.2×10-57
BRAF
p=9.4×10-179
KRAS
p=1.1×10-63
NRAS
p=9.6×10-45
Evolutionary Action (EA)


A
AUC-ROC
p-value deviation
CGC overlap
Method consensus
Consistency
AUPRC
CI (with INDEL)
CI
dNdScv
2020+
TUSON
MutPanning
OncodriveFML
MutsigCV
MuSiC
OncodriveClust
OncodriveFM
ActiveDriver
best
worst
B
0
20
40
60
80
SKCM
UCEC
COAD
HNSC
BRCA
LUSC
LUAD
CESC
BLCA
READ
KIRC
LAML
LGG
GBM
PRAD
LIHC
OV
THCA
KIRP
STAD
Tumor suppressor genes
Oncogenes
Number of genes
C
1
2
3
4
5
6
7
8
9
≥10
-log10 (CI q-value)
TP53 APC BRAF IDH1 PTEN ARID1A CDKN2A FAT1 PBRM1 PIK3CA KRAS PIK3R1 CTCF EGFR RB1 NF1 ATRX FUBP1 KDM6A NOTCH1 KMT2D FBXW7 CASP8 NRAS ARID2 CTNNB1 SETD2 BAP1 SMAD4 ZC3H13 AMER1 IDH2 STAG2 NFE2L2 ATM EP300 HRAS MYH9 TGFBR2 CDK12 KIT U2AF1 SMARCA4 MGA KEAP1 RASA1 NF2 GRM1 CREBBP RAC1 MAPK1 C1orf172 FOXA1 RUNX1 CDKN1B KMT2C
SKCM
THCA
LGG
GBM
OV
LUAD
PRAD
LIHC
STAD
BRCA
LAML
READ
COAD
UCEC
CESC
HNSC
BLCA
LUSC
KIRP
KIRC



A
Known tumor suppressor genes
sGOF index
sLOF index
-log10 (CI q-value)
DNMT3A
KEAP1
LZTR1
ACVR1B
CNOT3
CDK12
ATM
DDX3X
SMAD4
ARID1B
CREBBP
EP300
BAP1
FBXW7
RNF43
STAG2
B2M
SETD2
GATA3
TP53
MEN1
MLH1
ARID2
BCOR
CASP8
VHL
NF2
SMAD2
PIK3R1
PBRM1
PTEN
CEBPA
SMARCA4
NF1
RB1
CTCF
ATRX
TGFBR2
RBM10
CDKN2A
ZFHX3
CDKN1B
ARID1A
CDH1
APC
FAT1
STK11
CYLD
KDM6A
TET2
FUBP1
KDM5C
BRCA1
ASXL1
Known oncogenes
NRAS
PIK3CA
KRAS
HRAS
RAC1
IDH2
BRAF
NEF2L2
IDH1
EGFR
SF3B1
CTNNB1
FGFR2
GNAS
MAPK1
MTOR
CHD4
U2AF1
B
n.s.
****
Ceres score
Moderate EA
Others
Random
COSMIC
CI Pan
CI Indiv.


A
Head and neck squamous carcinoma
(HNSC)
-log10 (CI q-value)
1 2 3 4 5 6 7 8 9 10
PIK3CA
RAC1
HRAS
NFE2L2
UPF2
MAPK1
CREBBP
JPH1
HLA-B
ZNF750
PDZRN4
CASP8
AHDC1
THSD7A
FBXW7
MYH9
NOTCH1
CUL3
KMT2B
EP300
TP53
SCAF4
EPHA2
KMT2D
TGFBR2
PTEN
FNBP4
RASA1
C12orf42
NOTCH2
ARID2
FFAR2
NF2
RB1
AJUBA
HNRNPUL1
NSD1
KDF1
CDKN2A
FAT1
CD300E
CYLD
SLC2A13
OR2T35
VAMP1
sGOF index
sLOF index
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
0.0
0.2
0.4
0.6
0.8
1.0
B
TSG predicted by CI
OG predicted by CI
Others
Activation
Activation (context dependent)
Inhibition
NOTCH1
NOTCH2
FBXW7
NUMB
PDZRN4
ZNF750
TP63
CDKN2A
MYH9
KDF1
CDKs
TP53
MAPK1
Cell differentiation
Cell proliferation
MEK
KEAP1
CUL3
FAT1
CTNNB1
RB1
RAF
UPF2
NFE2L2
AJUBA
GSK3B
RASA1
HRAS
Cell stress response
AKT
EPHA2
Cell survival
RAC1
Immunosurveillance
PIK3CA
TGFBR2
HLA-B
NF-κB
CASP8
NF2
Cell migration
& invasion
PTEN
CYLD
DNA damage
HNRNPUL1
SCAF4
THSD7A


CUL3
A
R709W (89.66)
R709Q (82.91)
D426H(89.05)
T509M (93.11)
L321F (35.54)
E206Q (72.89)
M287R (84.67)
L418R (84.11)
A115T (39.74)
Important
Unimportant
Evolutionary Trace (ET)
B
Parental
LacZ
CUL3
UMSCC25
HN31
UMSCC22A
MDA1386LN
UMSCC25
N.S.
* p < 0.0001
Control Colonies (%)
No Treatment Control
LacZ
CUL3
C
UMSCC25
HN31
UMSCC22A
MDA1386LN
LacZ CUL3
CUL3
actin beta
D
%SubG1
%G1
%S
%G2
Parental
Distribution across different cell cycle phase (%)
E
PARP
Cleaved PARP
actin beta
F
Nrf2
DUSP16
G
Vehicle
DUSP16
Dox
GAPDH
HEK293
BT474
H
% of control
No Dox
Dox
Day
I
Uninduced
Induced
Vehicle DUSP16
J
Day 0
Day 1
Day 2
Vehicle
K
TP53
FL-PARP
cleaved PARP
L
PMA
Dox (µg/ml)
pcJUN(S63)
M
JUN
FOS
DMSO PMA
Fold change (normalized)
N
O
DUSP16-induced
DUSP16-induced, CRISPR