APSiC: Analysis of Perturbation Screens for the Identification of Novel Cancer Genes

Hesam Montazeri[1,2*], Mairene Coto-Llerena[2*], Gaia Bianco[2*], Ehsan Zangeneh[1], Stephanie Taha-Mehlitz[3], Viola Paradiso[2], Sumana Srivatsa[4,5], Antoine de Weck[6], Guglielmo Roma[6], Manuela Lanzafame[2], Martin Bolli[3,7], Niko Beerenwinkel[4,5], Markus von Flüe[3,7], Luigi M. Terracciano[2], Salvatore Piscuoglio[2,3,7#] and Charlotte K. Y. Ng[2,8#]

[1]Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.

[2]Institute of Pathology, University Hospital Basel, Basel, Switzerland;

[3]Visceral surgery research laboratory, Clarunis, Department of Biomedicine, University of Basel, Basel, Switzerland;

[4]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland;

[5]SIB Swiss Institute of Bioinformatics, Basel, Switzerland;

[6]Novartis Institutes for BioMedical Research, Novartis Pharma AG, Basel, Switzerland;

[7]Clarunis Universitäres Bauchzentrum Basel, Basel, Switzerland;

[8]Department for BioMedical Research, University of Bern, Bern, Switzerland.

*equal author contribution

[#]Correspondence to charlotte.ng@dbmr.unibe.ch and s.piscuoglio@unibas.ch

Keywords: APSiC, perturbation screen, cancer genes, *LRRC4B*, oncogene, tumor suppressor.

ABSTRACT

Systematic perturbation screens provided comprehensive resources for the elucidation of cancer driver genes. However, few algorithms have been developed to robustly interrogate such datasets, particularly with limited number of samples. Here we developed a computational tool called APSiC (*A*nalysis of *P*erturbation *S*creens for *i*dentifying novel *C*ancer genes) and applied it to the large-scale deep shRNA screen DRIVE[1] to unveil novel genetic and non-genetic driver genes. APSiC identified both well-known and novel drivers across all cancer types and within individual cancer types. The analysis of individual cancer types revealed that cancer drivers segregate by cell of origin and that genes involved in mRNA splicing may be oncogenic or tumor suppressive depending on the cancer type. We discovered and functionally demonstrated that *LRRC4B* is a novel putative tumor suppressor gene in breast cancer. The analysis of DRIVE using APSiC is provided as a web portal and represents a valuable resource for the discovery of novel cancer genes.

Advances in large-scale functional screening technologies have enabled the discovery of gene requirements across diverse cancer entities[2,3]. Systematic perturbation screens assess how genetic alterations or expression modulation of individual genes lead to phenotypic changes, revealing novel factors in carcinogenesis. McDonald *et al*. carried out the project DRIVE (deep RNAi interrogation of viability effects in cancer), a large perturbation screen targeting 7,837 genes in 398 cancer cell lines across a variety of malignancies to generate a comprehensive atlas of cancer dependencies[1]. In DRIVE[1], gene dependencies were evaluated using the raw cell viability readout of knockdown/out experiments (**Fig. 1a**, left) and tested using normality likelihood and correlation tests in a pan-cancer setting.

We introduce APSiC (***A***nalysis of ***P***erturbation ***S***creens for ***i***dentifying novel ***C***ancer genes), a novel tool for the systematic and robust interrogation of large-scale perturbation screens to discover gene dependencies for individual cancers even with limited number of samples. Instead of the raw cell viability readout, we compute a rank profile for each gene by first ranking all genes by their viabilities upon knockdown in a given sample to the range of [0, 1] then aggregating the normalized ranks for a given gene across all samples (**Fig. 1a**, **Online Methods**). Thus ranks close to zero represent reduced viability while the ranks close to one indicate cell growth upon knockdown.

Incorporating mutation and copy number status of the samples, APSiC identifies potential genetic and non-genetic cancer genes by assessing deviation of the distribution of normalized ranks from what is expected by chance using the Bates and Irwin-Hall tests. The use of the rank-based statistics with the Bates and Irwin-Hall distributions provides enhanced statistical power when the number of cell lines is limited. We consider three classes of genetic drivers (mutation oncogenes, amplification oncogenes, and mutation tumor suppressor genes) and two classes of non-genetic drivers (non-genetic oncogenes and tumor suppressor genes; **Online Methods**). We define mutation and amplification oncogenes as genes for which reduced cell viabilities are preferentially observed in samples with missense mutations and

copy number amplifications, respectively, while mutation tumor suppressor genes are those for which increased viabilities are preferentially observed in samples with deleterious mutations. To identify such genetic drivers, we test, for a given gene, whether ranks of the samples with and without the specific class of genetic alteration are significantly different using a one-sided Bates test (**Fig. 1b**). For mutation and amplification oncogenes, we compute the lower-tailed *P* values (i.e. the ranks preferentially suggest reduced viability upon gene knockdown), while for mutation tumor suppressor genes, we compute the upper-tailed *P* values (i.e. the ranks preferentially suggest increased cell viability upon knockdown). For the non-genetic drivers, we test whether gene knockdown in samples without genetic alteration in the gene has any impact on cell viability by computing lower and upper-tailed Irwin-Hall test *P* values for oncogenes and tumor suppressor genes, respectively (**Fig. 1c**). Optionally, we further test whether the expression of candidate non-genetic oncogenes or tumor suppressor genes is respectively enhanced or repressed in human tumors compared to the corresponding normal tissue type.

We applied APSiC to the DRIVE perturbation screens and the genetic data from the Cancer Cell Line Encyclopedia[4] to identify genetic driver genes. The dataset consists of 383 cell lines across 26 cancer types (**Fig. 2a**). In a pan-cancer analysis, APSiC reassuringly identified the well-known mutation oncogenes *BRAF, CTNNB1, KRAS, NRAS, PIK3CA* and *TP53* as the top candidates (**Figs. 2b-c**, **Supplementary Table S1**). Additionally, *DDX27, DCAF8L2* and *RBM39* were detected as mutation oncogenes (**Supplementary Fig. S1 and Supplementary Table S1**). The top amplification oncogene were *KRAS, BRAF, CDK4, YAP1, IL6* and *HAS2* (**Figs. 2b, d and Supplementary Table S1**), while the only mutation tumor suppressor was *ARID1A* (**Figs. 2b, e and Supplementary Table S1**). However, the identification of mutation tumor suppressor genes in a knockdown screen is likely to have limited utility given that mutation tumor suppressor genes are frequently associated with loss of the wild-type allele.

One of the main strengths of APSiC is the identification of dependencies in small sample sets. We therefore applied APSiC to the DRIVE data to identify genetic driver genes for individual cancer types. Across the 26 cancer types, we found at least one mutation oncogene, mutation tumor suppressor gene and amplification oncogene in 15, 14 and 11 cancer types, respectively (**Fig. 2f**). *KRAS*, *BRAF* and *TP53* were identified as a mutation oncogene in 4, 2 and 2 cancer types, respectively. We identified *MCL1* as the top amplification oncogene in the squamous subtype of non-small cell lung cancer and *BRCA1* as the top mutation tumor suppressor in breast cancer, but no amplification oncogene or mutation tumor suppressor was identified in more than one cancer type.

While whole-exome sequencing of 10,000+ cancers has revealed the global landscape of genetic driver genes[5], a systematic analysis of non-genetic driver genes (*i.e.* driver genes for which the basis for oncogenicity is non-genetic) is lacking. By assessing the rank profiles of cell lines wild-type for a given gene using APSiC, we evaluated the non-genetic dependencies across cancer types in the DRIVE data. Consensus clustering of the most variable genes in terms of APSiC *P* values revealed that such non-genetic dependencies segregate by organ systems or cell-of-origin into four clusters (**Fig. 3a**). In particular, non-epithelial cancers including leukemias/lymphomas, sarcomas, gliomas and neuroblastomas form a cluster distinct from epithelial cancers including those of the lungs, the breasts and gastrointestinal tract. This is consistent with the observation that multi-omics cancer classification is primarily driven by cell-of-origin and anatomic regions[6]. Furthermore, the top-level segregation of the cancer types was largely driven by the context-dependency of mRNA-splicing genes. We observed that mRNA-splicing genes such as *PRPF6* (Pre-MRNA Processing Factor 6) and *SART3* (Spliceosome Associated Factor 3, U4/U6 Recycling Protein) were tumor suppressive in the cluster enriched for non-epithelial cancers while they were oncogenic in the epithelial cancer cluster (**Fig. 3b**). The context-dependency highlights the divergent role of mRNA-splicing in carcinogenesis between cancer types. Our results also underscore the necessity

for an algorithm powerful enough to analyze perturbation screens for small numbers of samples in a cancer type-specific manner.

Based on the DRIVE screen alone, we identified a median of 28 non-genetic oncogenes (range 6-557) and 35 non-genetic tumor suppressor genes (range 1-471) per cancer type. However, we reasoned that the many non-genetic onco- and tumor suppressor genes would also be over- and under-expressed, respectively, in the corresponding cancer types. For the 12 cancer types for which gene expression data for the cancer and corresponding non-cancer counterparts were available from the TCGA (**Supplementary Fig. S2**), we further restricted the putative non-genetic onco- and tumor suppressor genes to those that were over- and under-expressed, respectively, relative to their non-cancer counterparts. After this filtering step, there were a median of 13 non-genetic oncogenes (range 2-117) and 3 non-genetic tumor suppressor genes (range 0-42, **Fig. 4**) per cancer type. We identified several well-known oncogenes, including *CDK1* (a master regulator of cell cycle) and *SMC1A* (a component of the cohesin complex involved in cell cycle checkpoint and genome stability)[7], and some that have been shown to have oncogenic properties in some cancer types, such as *MKI67IP* (or *NFIK*)[8]. We also identified *TEAD3*, a lesser described member of the TEAD family involved in hippo signalling, as oncogenic in liver cancer[9]. Among the top candidate tumor suppressors were *FOXP2* in endometrial cancer and *XRCC5* in kidney carcinoma. *FOXP2* knockdown has been shown to promote tumor initiation and metastasis in breast cancer[10] while *XRCC5*, encoding the protein Ku80, is a key DNA damage repair protein. However, we also identified many genes that have not been associated with carcinogenesis.

As a proof-of-concept to validate APSiC, we selected *LRRC4B,* one of the top putative non-genetic tumor suppressor genes in breast cancer that has not been associated with carcinogenesis. Nearly all breast cancer cell lines displayed significantly increased cell viability upon *LRRC4B* knockdown and breast cancers in TCGA showed lower expression compared to normal breast tissue (**Supplementary Fig. S3a**). We selected the breast cancer cell lines

MDA-MB231, BT-549 and MCF-7 with high, moderate and low endogenous *LRRC4B* expression to investigate its role in breast carcinogenesis (**Supplementary Fig. S3b**). We silenced *LRRC4B* in MDA-MB231 and BT-549 using siRNA, reducing LRRC4B protein expression by 40% and 60%, respectively, 72 hours post-transfection (**Figs. 5a, e**). In both models, *LRRC4B* downregulation significantly increased the proliferation and migration rates (**Figs. 5b-c, f-g**). By contrast, *LRRC4B* overexpression significantly reduced proliferation and migration in MCF-7 (**Figs. 5i-k**).

LRRC4, an important paralog of *LRRC4B*, has been shown to have an oncosuppressor role in glioma[11,12,13,14], suppressing cell proliferation by delaying cell cycle in late G_1 phase[11,15]. To test whether *LRRC4B* may play the same role in breast, we analyzed cells with *LRRC4B* overexpression or downregulation stained with DAPI by flow cytometry (FACS). *LRRC4B* knockdown in MDA-MB231 and BT-549 promoted cell transition into S phase (**Figs. 5d, h**), while *LRRC4B* overexpression in MCF-7 significantly retained cells in G_1 phase (**Fig. 5l**), suggesting a similar mechanism.

A common mechanism of oncogenicity is resistance to apoptosis[16]. To test whether modulation of apoptosis is a mechanism of action of *LRRC4B* as an oncosuppressor, we induced apoptosis with doxorubicin and measured it using Annexin V and propidium iodide co-staining followed by FACS analysis (**Fig. 6a**). Forty eight hours after treatment, *LRRC4B*-overexpressing MCF-7 cells showed 10% more apoptotic and 10% fewer live cells, suggesting that *LRRC4B* overexpression could sensitize cells to doxorubicin-induced apoptosis (**Fig. 6b**). By contrast, *LRRC4B*-downregulating MDA-MB231 and BT-549 cells showed increased resistance to doxorubicin and had 25% and 10% fewer apoptotic and 25% and 10% more live cells, respectively (**Fig. 6b**). Our results provide compelling evidence that APSiC identified *LRRC4B* as a novel oncosuppressor gene in breast cancer.

Here we provide a powerful statistical analysis for the scientific community to explore and functionally characterize genes that may be involved in carcinogenic processes and may pave the way for the discovery of novel cancer-related biomarkers and drug targets.

Methods

The APSiC algorithm

We introduce a novel computational tool called APSiC for ***A***nalysis of ***P***erturbation ***S***creens for ***i***dentifying novel ***C***ancer genes. To begin, we briefly describe some necessary definitions and background material from ordered statistics.

We consider the knockdown experiments of *p* genes across *N* cell lines. Let v_{ij}be viability of cell line $i \in \{1,\ldots,N\}$ upon knocking down gene $j \in \{1,\ldots, p\}$ and m_{ij} be a binary variable indicating whether a specific genetic alteration (i.e. mutation or copy number alteration) is present in gene *j* of cell line *i*. In this study, we only consider deleterious (e.g. nonsense, frameshift, splice site and mutations affecting start or stop codons) and missense mutations. Waterfall plots are often used to show viabilities of knockdown experiments for a single gene across different cell lines and are aimed to illustrate different gene dependencies. As an example, waterfall plot for gene *TP53* is shown in **Fig. 1a** (left). Each vertical bar corresponds to a cell line and is colored by the pre-existing mutation types present in *TP53*. **Fig. 1a** indicates cell lines with the presence of deleterious or missense mutations in *TP53* tend to have lower viabilities upon knockdown of this gene. While waterfall plot is a useful visualization tool for demonstrating gene dependencies, it lacks sufficient interpretability in certain cases, particularly when the number of cell lines is limited. In this paper, we introduce a new waterfall plot, named rank viability profile or simply rank profile, to address this issue.

To make viability scores comparable across cell lines, we compute normalized rank values per cell lines denoted as r_{ij}, representing the rank of viability for gene *j* among all knockdown experiments in cell line *i*. For mathematical convenience and without loss of generality, we normalized ranks to the range of [0, 1]. When the number of knockdown genes is high, normalized ranks have many distinct levels in the interval [0, 1] and we assume normalized ranks are continuous. Let $R_{1,A}, R_{2,A}, \ldots, R_{N,A}$denote random variables associated to ranks of a

gene *A* in *N* cell lines. We drop subscript A and denote ranks as $R_1, R_2, \ldots, R_N$ for the simplicity of notation. By placing ranks, R_i, in ascending orders and renaming them, we obtain $Y_1 < Y_2 < \ldots < Y_N$ where Y_i is called *i*th ordered statistic. It is easy to see that $Y_1 = min(R_1, \ldots, R_N)$ and $Y_N = max(R_1, \ldots, R_N)$. The probability density function of ordered statistic Y_i in general is given as

$$f(y_i) = N f(r) \binom{N-1}{i-1} F(r)^{i-1} (1 - F(r))^{N-i}$$

where *f(r)* and *F(r)* denote probability density and cumulative distribution functions, respectively. If there is no dependency between knocking down of a gene and the viability of the cell, we can assume $R_i \sim U(0,1)$ for $i = 1, \ldots, N$, hence we have $Y_i \sim Beta(i, N - i + 1)$. Using this result, we can construct a no-change viability band at statistical significance *α* using the quantiles of Y_i at the $\alpha/2$ and $1 - \alpha/2$ for $i = 1, \ldots, N$. Now we define a new waterfall plot, called rank viability profile or simply rank profile, as a waterfall plot using normalized ranks, realizations of R_i for a gene, overlaid with no-change viability band (**Fig. 1a**).

The APSiC algorithm identifies potential cancer genes by assessing deviation of respective rank profiles from what is expected by chance. The algorithm can identify both genetic and non-genetic drivers (**Fig. 1b-c**). We consider three categories for genetic drivers.

- Mutation oncogene: defined as genes for which reduced viabilities are observed preferentially in samples with missense mutation.
- Amplification oncogene: defined as genes for which reduced viabilities are observed preferentially in samples with copy number amplification.
- Mutation tumor suppressor: defined as genes for which increased viabilities are observed preferentially in samples with deleterious mutation.

We consider two categories for non-genetic drivers, namely

- Non-genetic oncogene: defined as genes for which reduced viabilities are observed in samples without a genetic alteration in the respective gene.

- Non-genetic tumor suppressor: defined as genes for which increased viabilities are observed preferentially in samples without a genetic alteration in the respective gene.

For genetic drivers, the APSiC algorithm considers rank profiles of mutated and wild-type samples with respect to an input gene g (**Fig. 1b**). Then, it performs a one-sided statistical test to determine whether rank scores of the two groups of samples are significantly different in the direction of interest, according to the genetic feature of interest. Suppose $R_1^{wt}, R_2^{wt}, \ldots, R_m^{wt}$ and $R_1^{mu}, R_2^{mu}, \ldots, R_n^{mu}$ are random variables denoting rank scores upon knockdown of gene g for m wild-type and n mutated samples, respectively. Let $\overline{R^{wt}}$ and $\overline{R^{mu}}$ denote the average of ranks for the wild-type and mutated samples, respectively. We define $S = \overline{R^{mu}} - \overline{R^{wt}}$as the test statistic and s_{obs} as the observed test statistic. We assume the null hypothesis is that the knockdown of gene g does not have any impact on the viability of samples and therefore there is no difference in average of ranks for two groups, i.e. S =0. The general formula for the distribution of any weighted sum of uniform random variables is given in Kamgar-Parsi, 1994[17]. We simplify the general formula thereby and obtain the null distribution of the test statistic Sas

$$P(S \leq s) = \frac{(-1)^m}{n^n m^m (n+m)!} \sum_{k=0}^{n} \sum_{p=0}^{m} (-1)^{p+k} \frac{n}{k} \frac{m}{p} (s + \frac{k}{n} - \frac{p}{m})^{m+n} \Theta(s + \frac{k}{n} - \frac{p}{m})$$

where

$$\theta(t) = \begin{cases} 0 & t < 0 \\ \frac{1}{2} & t = 0 \\ 1 & t > 0 \end{cases}$$

For mutation and amplification oncogenes, we compute lower-tailed *P* values, $P(S \leq s_{obs})$ while for mutation tumor suppressors, we compute upper-tailed *P* values, $P(S \geq s_{obs})$. Due to numerical issues, it is impractical to use the exact null distribution formula for large values of *m* and *n* (*m*+*n* >20). In this case, we compute an approximation for the null distribution of *S* as

follows. Under the null hypothesis, $\overline{R^{mu}}$ and $\overline{R^{wt}}$ follow *Bates* distributions *Bates(m)* and *Bates(n),* respectively. The Bates distribution is a distribution that represents the mean of a number of independent uniform random variables on the unit interval. For large values of *m* and *n,* $\overline{R^{mu}}$ and $\overline{R^{wt}}$ are approximately distributed by $N(\frac{1}{2}, \frac{1}{12m})$ and $N(\frac{1}{2}, \frac{1}{12n})$. Hence, under the null hypothesis the test statistic *S* is approximately distributed as $N(0, \frac{1}{12}(\frac{1}{m} + \frac{1}{n}))$.

To identify non-genetic drivers, we only consider the wild-type (i.e. without non-synonymous mutations and without copy number amplification (GISTIC copy number state 2) or deep deletions (GISTIC copy number state -2)) samples with respect to an input gene g (**Fig. 1c**). The null hypothesis is that the knockdown of a gene *g* does not have any impact on the viability of the samples. We define the test statistic as $T = R_1 + R_2 + \ldots + R_m$ and t_{obs} as the observed test statistic. Under the null hypothesis, *T* follows an Irwin-Hall distribution $T \sim IH(m)$, which represents the summation of m independent uniform random variables on the unit interval. For large values of *m*, *S* is approximately distributed as $N(m/2, m/12)$. To identify non-genetic oncogenes, we require significant lower-tailed *P* values, $P(T \leq t_{obs})$ for wild-type cell lines with respect to the input gene. Additionally, for the respective tissue type, the overall expression at the RNA level of a putative oncogene in tumor samples is required to be significantly higher than the one in normal tissue samples using the *t* test. On the contrary, for identifying non-genetic tumor suppressors, we require significant upper-tailed *P* values, $P(T \geq t_{obs})$, for wild-type cell lines with respect to the input gene as well as lower RNA expression of tumor samples in comparison to normal tissue samples using the *t* test.

Downloading and preprocessing of DRIVE and TCGA data

We considered the viability profiles of 383 cell lines in the project DRIVE[1] for which their genetic profiles were available at the Cancer Cell Line Encyclopedia (**Fig. 2a**)[4]. We computed aggregated gene-level viability scores for each experiment by the RSA and ATARiS algorithms[18,19]. The RSA and ATARiS scores are available for 7726 and 6557 genes,

respectively. We used the same method as defined in the project DRIVE to remove essential genes, defined as genes with an RSA value of ≤ 3 in more than half of cell lines. After the removal of 184 essential genes by this method, for pan-cancer analysis, the ATARiS scores are available for 6373 genes in 383 cell lines in 39 cancer types over 17 primary tissues. For the genetic drivers, we considered only genes for which there are at least 2 samples harbored a genetic alteration of the corresponding class in individual cancer types. For the non-genetic drivers, we only considered genes for which there are at least 2 samples wild-type for the gene in individual cancer types. For pan-cancer analysis, the above threshold was at least 4 samples for both genetic and non-genetic drivers. For the identification of mutation tumor suppressor genes, mutations annotated as *In_Frame_Ins*, *In_Frame_Del*, *Frame_Shift_Ins*, *Frame_Shift_Del*, *Nonsense_Mutation*, *Splice_Site*, *Start_Codon_Del*, *Stop_Codon_Del*, *Stop_Codon_Ins*, *Start_Codon_Ins* were considered deleterious. For the analysis of individual cancer types, we considered the 26 cancer types for which more than four cell lines are available in the DRIVE data.

TCGA gene expression data was obtained for 12 cancers for which data are available for tumor and normal tissues (**Supplementary Fig. S2**). The data were downloaded using the TCGAbiolinks package in R[20]. The normalized expression level of genes in CPM (counts per million) were used for the identification of non-genetic drivers.

Multiple testing

To address the multiple comparisons problem, we chose a significance level such that the expected number of false positives due to multiple testing for each cancer and feature is equal to one. To this end, we chose a significance level of $1/n$, or 0.05 if $1/n$>0.05, where n is the number of genes tested for identification of drivers. Using this approach, we were able to keep many interesting hits while keeping the number of false positive cases low.

Clustering and pathway analysis

Clustering was performed using ConsensusClusterPlus[21] using 1-Spearman correlation as the distance metric and the Ward hierarchical clustering algorithm. The number of clusters was determined based on the relative change in area under the consensus cumulative distribution function over the number of evaluated clusters. Pathway analysis was performed using g:Profiler[22].

Software and data availability

A web portal using the Shiny framework in R has been developed to visualize rank profiles of the DRIVE shRNA screen and corresponding gene expression data from TCGA at https://apsic.shinyapps.io/APSiC/. The code for the APSiC algorithm is available at https://github.com/hesmon/APSiC/.The raw shRNA data has already published as a part of the project DRIVE (https://data.mendeley.com/datasets/y3ds55n88r/4) and copy number and mutation profiles of the cell lines are available at the Cancer Cell Line Encyclopedia portal (https://portals.broadinstitute.org/ccle/home). The gene expression data from the TCGA are available at the TCGA Genomics Data Commons data portal (https://portal.gdc.cancer.gov/).

Cell lines

Breast cancer derived cell lines (MCF-7, BT-549 and MDA-MB231) were maintained in a 5% CO_2-humidified atmosphere at 37°C and cultured in DMEM supplemented with 10% FBS, 1% Pen/Strep (Bio-Concept) and 1% MEM-NEAA (MEM non-essential amino acids, ThermoFisher Scientific). All cell lines were confirmed negative for mycoplasma infection using the PCR-based Universal Mycoplasma Detection kit (American Type Culture Collection, Manassas, VA) as previously described[23].

Transient gene knockdown by siRNAs

Transient gene knockdown was conducted using ON-TARGET plus siRNA transfection. ON-TARGET plus SMARTpool siRNAs against human *LRRC4B* (Dharmacon, CO; #L-023786-

01-0005), ON-TARGET plus SMARTpool non-targeting control and DharmaFECT transfection reagent (Dharmacon, CO; #T-2001-03) were all purchased from GE Dharmacon. Transfection was performed according to the manufacturer's protocol. Briefly, log-phase breast cancer cells were seeded at approximately 60% confluence. Because antibiotics affects the knockdown efficiency of ON-TARGET plus siRNAs, growth medium was removed as much as possible and replaced by antibiotic-free complete medium. siRNAs were added to a final concentration of 25 nM. Cells were incubated at 37°C in 5% CO_2 for 24-48-72 hours for 48-72 hours for protein analysis. To avoid cytotoxicity, transfection medium was replaced with complete medium after 8 hours.

Protein extraction and western blot

Proteins were extracted using Co-IP buffer (100 mmol/L NaCl, 50 mmol/L Tris pH 7.5, 1 mmol/L EDTA, 0.1% Triton X-100) supplemented with 1x protease inhibitors (cOmplete Mini, EDTA-free Protease Inhibitor Cocktail, Roche, CO, #4693159001) and 1x phosphatase inhibitors (PhosSTOP #4906837001, Merck). Cell lysates were then treated with 10x reducing agent (NuPAGE Sample Reducing Agent, Invitrogen, #NP0009), 4x loading buffer (NuPAGE LDS Sample Buffer, Invitrogen, #NP0007), boiled and loaded into neutral pH, pre-cast, discontinuous SDS-PAGE mini-gel system (NuPAGE 10% Bis-Tris Protein Gels, ThermoFisher). The proteins were then transferred to nitrocellulose membranes using Trans-Blot Turbo Transfer System (Bio-Rad). The membranes were blocked for 1 hr with Sure Block (Lubio Science) and then probed with primary antibodies overnight at 4°C. Next day, the membranes were incubated for 1 hr at RT with fluorescent secondary goat anti-mouse (IRDye 680) or anti-rabbit (IRDye 800) antibodies (both from LI-COR Biosciences). Blots were scanned using the Odyssey Infrared Imaging System (LI-COR Biosciences) and band intensity was quantified using ImageJ software. The ratio of proteins of interest/loading control in treated samples were normalized to their counterparts in control cells. Antibodies against

LRRC4B (PA5-23529, Thermofisher) and B-actin (A5441, Sigma) were used at dilution 1:1000 and 1:5000, respectively.

Proliferation assay

Cell proliferation was assayed using the xCELLigence system (RTCA, ACEA Biosciences, San Diego, CA, USA) as previously described.[24] Cells were first seeded and transfected in 6 well plates and 24 h after transfection $5x10^3$ cells were resuspended in 100 µl of medium and plated in each well of an E-plate 16. Background impedance of the xCELLigence system was measured for 12 s using 50 µl of room temperature cell culture media in each well of E-plate 16. The final volume in each well was then 150 µl. The impedance signals were recorded every 15 minutes until 96/120 h and expressed as cell index values, calculated automatically and normalized by the RTCA Software Package v1.2. The values were defined as mean ± standard deviation. Mann-Whitney test was used for statistical analysis with GraphPad software.

Migration assay

Migration assays were performed using the CIM-plate of the xCELLigence Real-Time Cell Analysis (RTCA, ACEA Biosciences, San Diego, CA, USA) system. Cells were first transfected in 6-well plates and 24 h after transfection, they were harvested and seeded in the CIM-plate. Every well of the bottom chamber was filled with 160 µl of the corresponding medium at 10% FBS concentration. After placing the upper chamber on top of the lower chamber, 50 µl of serum free medium was added on each CIM well for the background measurement. After 3x PBS washing, $3x10^4$ cells re-suspended in 100 µl of the corresponding medium at 1% FBS concentration were seeded in each well of the upper chamber. The measurements were taken every 15 minutes until 24 h after seeding and expressed as cell index values. Mann-Whitney U test was used for statistical analysis with GraphPad software.

Cell cycle analysis

Seventy-two hour after transfection, cells were collected, stained with DAPI and analyzed by flow cytometry using the BD FACS Canto II cytometer (BD Biosciences, USA). Briefly cells were harvested and washed 2X in PBS to get rid of serum proteins at 1200 rpm for 5 minutes. Pellets (up to $3x10^6$ cells) were resuspended in 1.2 ml PBS (Ca and Mg free). For crosslinking proteins 3.0 ml of 95% ice cold EtOH was added dropwise while vortexing. Cells were fixed in this final 70% Et-OH solution for at least 30 minutes or over night. The Et-OH/cell suspension was then diluted with 12 ml of PBS (for a total volume of 15 ml) and centrifuge at 2000-2200 rpm for 10 min. Cells were then washed once more with 15 ml PBS and then resuspended in 0.5-2.0 ml of DAPI stain solution (0.1% TritonX 100 and 10 ug/ml). After 30 min of incubation on ice cells were analyzed by flow cytometry, measuring the fluorescence emission at 461 nm. Data were analysed using the FlowJo software version 10.5.3 (https://www.flowjo.com).

Apoptosis analysis by flow cytometry

BT-549 and MDA-MB231 cells were transfected with siRNA (control or against *LRRC4B*) and MCF-7 cells were transfected with *LRRC4B* overexpressing plasmid or control plasmid. Eight hours after transfection medium was changed and doxorubicin added according to the respective IC50 for each cell line[25,26]. Cells were collected 60 hours post siRNA transfection or *LRRC4B* overexpression and 48 hours post treatment with doxorubicin respectively, stained with annexin V (Annexin V-FITC conjugate; Invitrogen, CO; #V13242) and propidium iodide (PI; Invitrogen, CO; #V13242), and analyzed by flow cytometry using the BD FACS Canto II cytometer (BD Biosciences, USA). Briefly, cells were harvested after incubation period and washed twice by centrifugation (1,200 g, 5 min) in cold phosphate-buffered saline (DPBS; Gibco, CO; #14040133). After washing, cells were resuspended in 0.15 ml AnnV binding buffer 1X (ABB 5X, Invitrogen, CO; #V13242; 50 mM HEPES, 700 mM NaCl, and 12.5 mM CaCl2 at pH 7.4) containing fluorochrome-conjugated AnnV and PI (PI to a final concentration of 1 ug/ml) and incubated in darkness at room temperature for 15 min. As soon as possible cells were analyzed by flow cytometry, measuring the fluorescence emission at 530 nm and >575 nm. Data were analysed using the FlowJo software version 10.5.3 (https://www.flowjo.com).

Acknowledgments

Development of APSiC was performed at sciCORE scientific computing center at the University of Basel and at the scientific computing center of the Department of Bioinformatics, University of Tehran. V.P. was supported by the Swiss Centre for Applied Human Toxicology. L.M.T. and S.P. were supported by the Swiss Cancer League (KLS-3639-02-2015 and KFS-3995-08-2016, respectively). S.P. was supported by Swiss National Science Foundation (PZ00P3_168165). N.B. was supported by European Research Council Synergy Grant 609883. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Contributions

H.M., S.P. and C.K.Y.N. conceived the study. S.P. and C.K.Y.N. supervised the study. H.M. and C.K.Y.N. developed the methodology. H.M. performed the bioinformatic analyses. E.Z. contributed to data presentation and visualization. H.M. and E.Z. implemented the Shiny web application. M.C.-L., G.B., S.T.-M., V.P. and M.L. performed *in vitro* experiments. S.S., A.dW., G.R., M.B., N.B., M.vF. and L.M.T. critically discussed the results. H.M., M.C.-L., G.B., S.P. and C.K.Y.N. interpreted the results and wrote the manuscript. All authors agreed to the final version of the manuscript.

Competing interests

A.dW. and G.R. are employees of Novartis Pharma AG. The other authors declare no competing interests.

References

1. McDonald, E. R., 3rd *et al.* Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. *Cell* **170**, 577–592.e10 (2017).
2. Cheung, H. W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 12372–12377 (2011).
3. Marcotte, R. *et al.* Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov.* **2**, 172–189 (2012).
4. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
5. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385.e18 (2018).
6. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291–304.e6 (2018).
7. Yi, F. *et al.* Structural Maintenance of Chromosomes protein 1: Role in Genome Stability and Tumorigenesis. *Int. J. Biol. Sci.* **13**, 1092–1099 (2017).
8. Lin, T.-C. *et al.* The nucleolar protein NIFK promotes cancer progression via CK1α/β-catenin in metastasis and Ki-67-dependent cell proliferation. *Elife* **5**, (2016).
9. Perra, A. *et al.* YAP activation is an early event and a potential therapeutic target in liver cancer development. *Journal of Hepatology* **61**, 1088–1096 (2014).
10. Cuiffo, B. G. *et al.* MSC-regulated microRNAs converge on the transcription factor FOXP2 and promote breast cancer metastasis. *Cell Stem Cell* **15**, 762–774 (2014).
11. Zhang, Q.-H. *et al.* Study of a novel brain relatively specific gene LRRC4 involved in glioma tumorigenesis suppression using the Tet-on system. *Acta Biochim. Biophys. Sin.* **37**, 532–540 (2005).
12. Wu, M. *et al.* LRRC4 inhibits glioblastoma cell proliferation, migration, and angiogenesis

by downregulating pleiotropic cytokine expression and responses. *J. Cell. Physiol.* **214**, 65–74 (2008).

13. Wang, Z. *et al.* The D Domain of LRRC4 anchors ERK1/2 in the cytoplasm and competitively inhibits MEK/ERK activation in glioma cells. *Journal of Hematology & Oncology* **9**, (2016).
14. Tang, H. *et al.* Interaction of hsa-miR-381 and glioma suppressor LRRC4 is involved in glioma growth. *Brain Res.* **1390**, 21–32 (2011).
15. Wang, R. *et al.* LRRC4 inhibits the proliferation of human glioma cells by modulating the expression of STMN1 and microtubule polymerization. *J. Cell. Biochem.* **112**, 3621–3629 (2011).
16. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
17. Kamgar-Parsi, B., Kamgar-Parsi, B. & Brosh, M. Distribution and moments of the weighted sum of uniforms random variables, with applications in reducing monte carlo simulations. *Journal of Statistical Computation and Simulation* **52**, 399–414 (1995).
18. König, R. *et al.* A probability-based approach for the analysis of large-scale RNAi screens. *Nat. Methods* **4**, 847–849 (2007).
19. Shao, D. D. *et al.* ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome Res.* **23**, 665–678 (2013).
20. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **44**, e71–e71 (2016).
21. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
22. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
23. Ng, C. K. Y. *et al.* Intra-tumor genetic heterogeneity and alternative driver genetic alterations in breast cancers with heterogeneous HER2 gene amplification. *Genome Biol.* **16**, 107 (2015).

24. Andreozzi, M. *et al.* HMGA1 Expression in Human Hepatocellular Carcinoma Correlates with Poor Prognosis and Promotes Tumor Growth and Migration in in vitro Models. *Neoplasia* **18**, 724–731 (2016).
25. Pilco-Ferreto, N. & Calaf, G. M. Influence of doxorubicin on apoptosis and oxidative stress in breast cancer cell lines. *Int. J. Oncol.* **49**, 753–762 (2016).
26. Inao, T. *et al.* Bcl-2 inhibition sensitizes triple-negative human breast cancer cells to doxorubicin. *Oncotarget* **9**, 25545–25556 (2018).
27. Tang, J. *et al.* DEAD-box helicase 27 promotes colorectal cancer growth and metastasis and predicts poor survival in CRC patients. *Oncogene* **37**, 3006–3021 (2018).
28. Tsukamoto, Y. *et al.* Expression of DDX27 contributes to colony-forming ability of gastric cancer cells and correlates with poor prognosis in gastric cancer. *Am. J. Cancer Res.* **5**, 2998–3014 (2015).
29. Thomas, R. & Majeti, R. No Matter How You Splice It, RBM39 Inhibition Targets Spliceosome Mutant AML. *Cancer Cell* **35**, 337–339 (2019).
30. Wang, E. *et al.* Targeting an RNA-Binding Protein Network in Acute Myeloid Leukemia. *Cancer Cell* **35**, 369–384.e7 (2019).
31. Han, T. *et al.* Anticancer sulfonamides target splicing by inducing RBM39 degradation via recruitment to DCAF15. *Science* **356**, (2017).
32. Mai, S. *et al.* Global regulation of alternative RNA splicing by the SR-rich protein RBM39. *Biochim. Biophys. Acta* **1859**, 1014–1024 (2016).
33. Cheung, T. H. *et al.* Alteration of cyclin D1 and CDK4 gene in carcinoma of uterine cervix. *Cancer Lett.* **166**, 199–206 (2001).
34. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
35. Curtin, J. A. *et al.* Distinct sets of genetic alterations in melanoma. *N. Engl. J. Med.* **353**, 2135–2147 (2005).
36. Wong, D. D. *et al.* MDM2/CDK4 gene amplification in large/deep-seated 'lipomas': incidence, predictors and clinical significance. *Pathology* **48**, 203–209 (2016).

37. Lopez-Mejia, I. C. *et al.* CDK4 Phosphorylates AMPKα2 to Inhibit Its Activity and Repress Fatty Acid Oxidation. *Mol. Cell* **68**, 336–349.e6 (2017).

38. Kim, J.-A. *et al.* Amplification of TLK2 Induces Genomic Instability via Impairing the G2-M Checkpoint. *Mol. Cancer Res.* **14**, 920–927 (2016).

39. Kim, J.-A. *et al.* Comprehensive functional analysis of the tousled-like kinase 2 frequently amplified in aggressive luminal breast cancers. *Nat. Commun.* **7**, 12991 (2016).

40. Lin, M., Yao, Z., Zhao, N. & Zhang, C. TLK2 enhances aggressive phenotypes of glioblastoma cells through the activation of SRC signaling pathway. *Cancer Biol. Ther.* **20**, 101–108 (2019).

41. He, G. *et al.* Identification of liver cancer progenitors whose malignant progression depends on autocrine IL-6 signaling. *Cell* **155**, 384–396 (2013).

42. Hartman, Z. C. *et al.* Growth of triple-negative breast cancer cells relies upon coordinate autocrine expression of the proinflammatory cytokines IL-6 and IL-8. *Cancer Res.* **73**, 3470–3480 (2013).

43. Yun, M. R. *et al.* ERK-dependent IL-6 autocrine signaling mediates adaptive resistance to pan-PI3K inhibitor BKM120 in head and neck squamous cell carcinoma. *Oncogene* **37**, 377–388 (2018).

44. Liu, G. *et al.* Prostate-specific IL-6 transgene autonomously induce prostate neoplasm through amplifying inflammation in the prostate and peri-prostatic adipose tissue. *J. Hematol. Oncol.* **10**, 14 (2017).

45. Li, Y., Li, L., Brown, T. J. & Heldin, P. Silencing of hyaluronan synthase 2 suppresses the malignant phenotype of invasive breast cancer cells. *Int. J. Cancer* **120**, 2557–2567 (2007).

46. Shen, Y. N. *et al.* Inhibition of HAS2 induction enhances the radiosensitivity of cancer cells via persistent DNA damage. *Biochem. Biophys. Res. Commun.* **443**, 796–801 (2014).

47. Preca, B.-T. *et al.* A novel ZEB1/HAS2 positive feedback loop promotes EMT in breast cancer. *Oncotarget* **8**, 11530–11543 (2017).

48. Zhang, H.-Y. *et al.* In Vitro Effects of HAS-2 Gene Silencing on the Proliferation and Apoptosis of the MCF-7 Human Breast Cancer Cell Line. *Cell. Physiol. Biochem.* **40**, 807–817 (2016).

FIGURES AND FIGURE LEGENDS

Fig.1

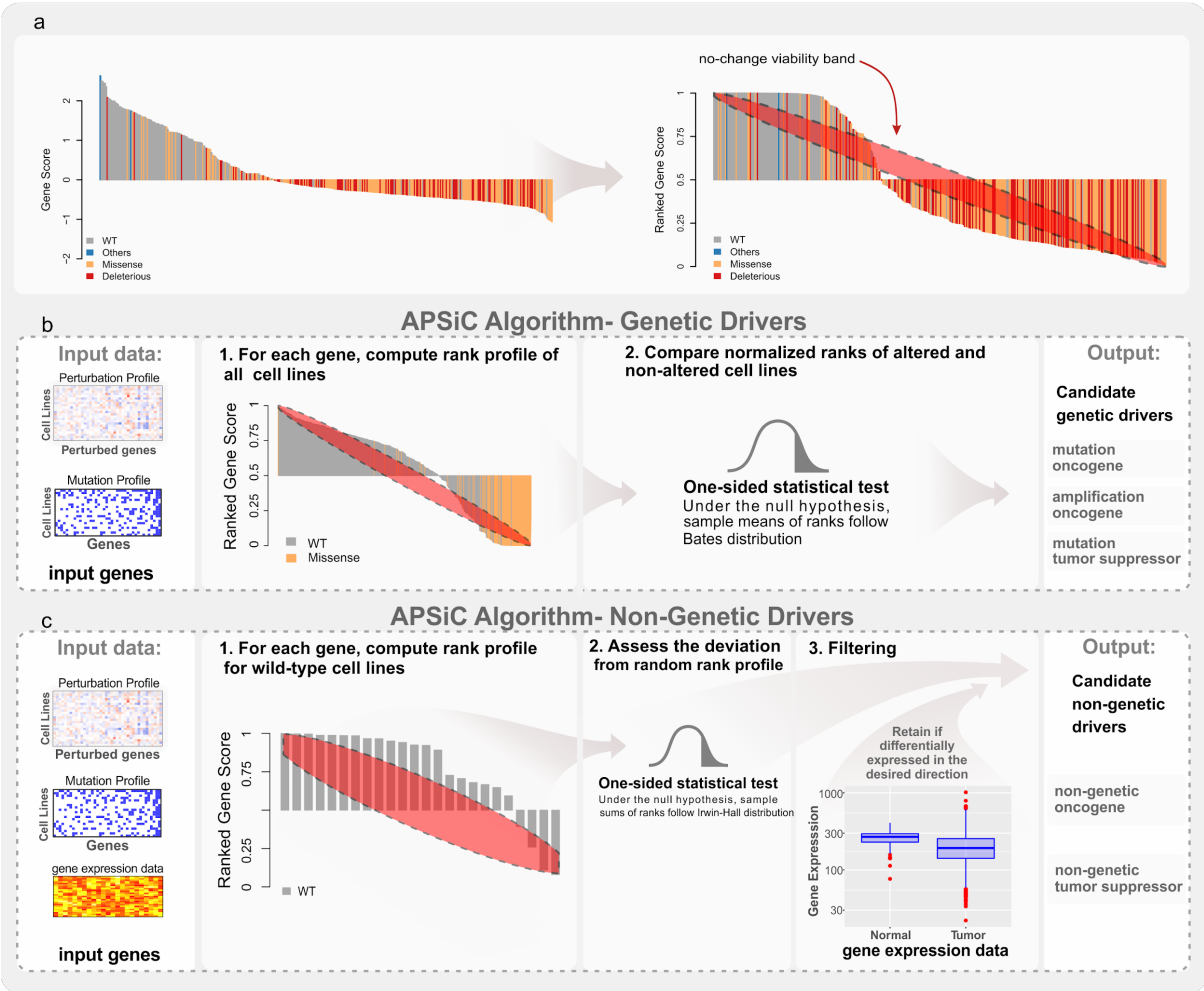

Fig. 1. Overview of the APSiC algorithm. a, Illustration of the transformation from the raw cell viability scores for a given gene (left, using *TP53* as an example) to the rank profile (right). Each bar of the waterfall plots represents one sample and is colored by the mutational status of the given gene in the sample. The red ellipse in the rank profile (right) represents a no-change (random) viability band. **b-c,** Schematic representation of the APSiC algorithm for identifying **(b)** genetic and **(c)** non-genetic drivers (See Online Methods for details).

Fig. 2

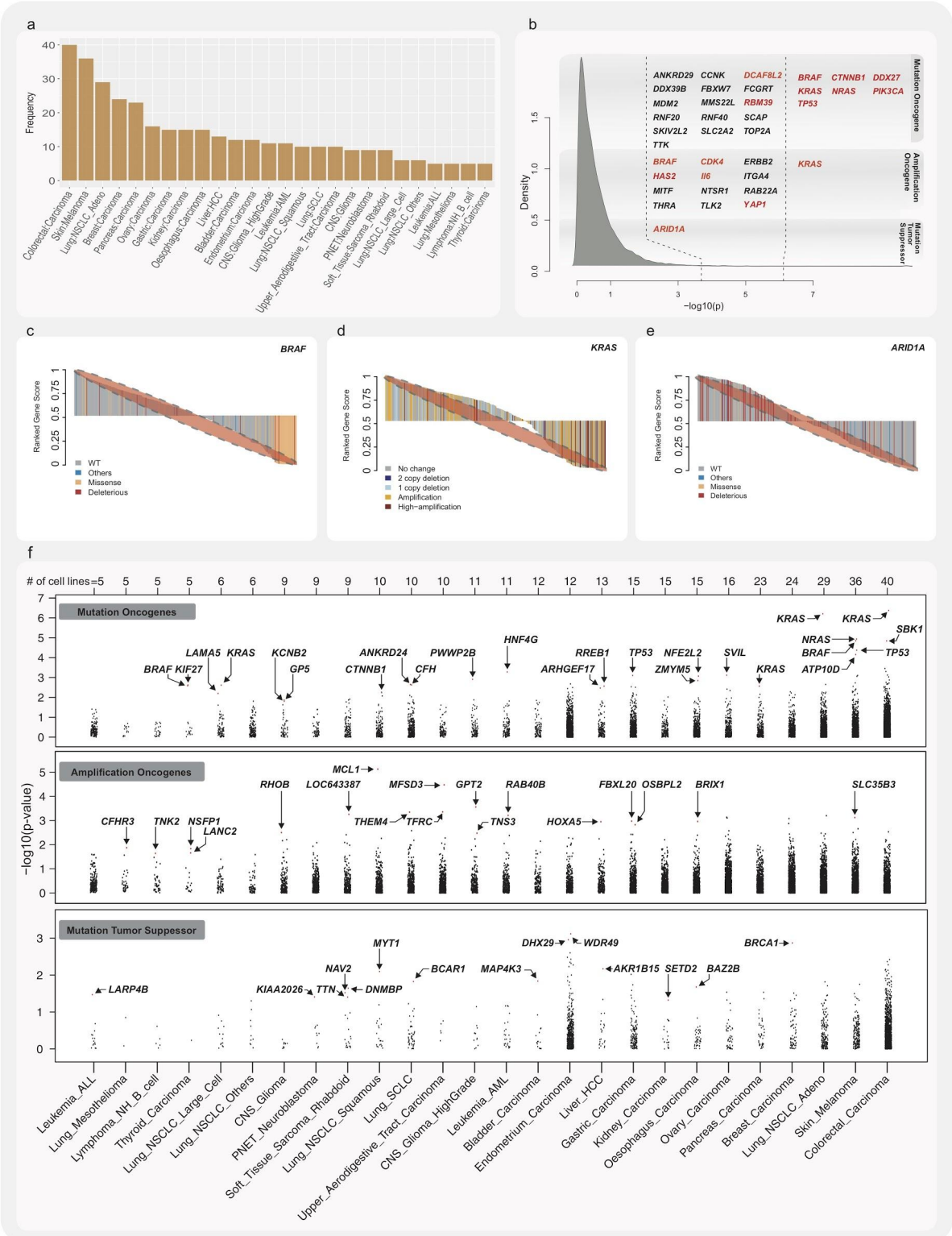

Fig. 2. Analysis of the genetic drivers in the DRIVE perturbation screen. a, The number of cell lines available for 26 cancer subtypes in the DRIVE perturbation screen. **b,** Kernel density estimation of the *P* values (on a -$\log_{10}$ scale) for genetic drivers using the APSiC

algorithm in a pan-cancer analysis. Candidate mutation oncogenes, amplification oncogenes, and mutation tumor suppressors identified by the APSiC are shown, with genes reaching significance level after multiple testing corrections highlighted in red. **c,** Rank profile for a mutation oncogene (*BRAF*) colored by mutation status. **d,** Rank profile of an amplification oncogene (*KRAS*) colored by copy number status. **e,** Rank profile of a mutation tumor suppressor (*ARID1A*) colored by mutation status. **f,** Dot plots of the *P* values (on a $-\log_{10}$ scale) for genetic drivers using the APSiC algorithm in a cancer type-specific analysis, for (top) mutation oncogenes, (middle) amplification oncogenes and (bottom) mutation tumor suppressors. Genes reaching significance level after multiple testing corrections are highlighted in red. Cancer types are sorted by the number of cell lines.

Fig. 3

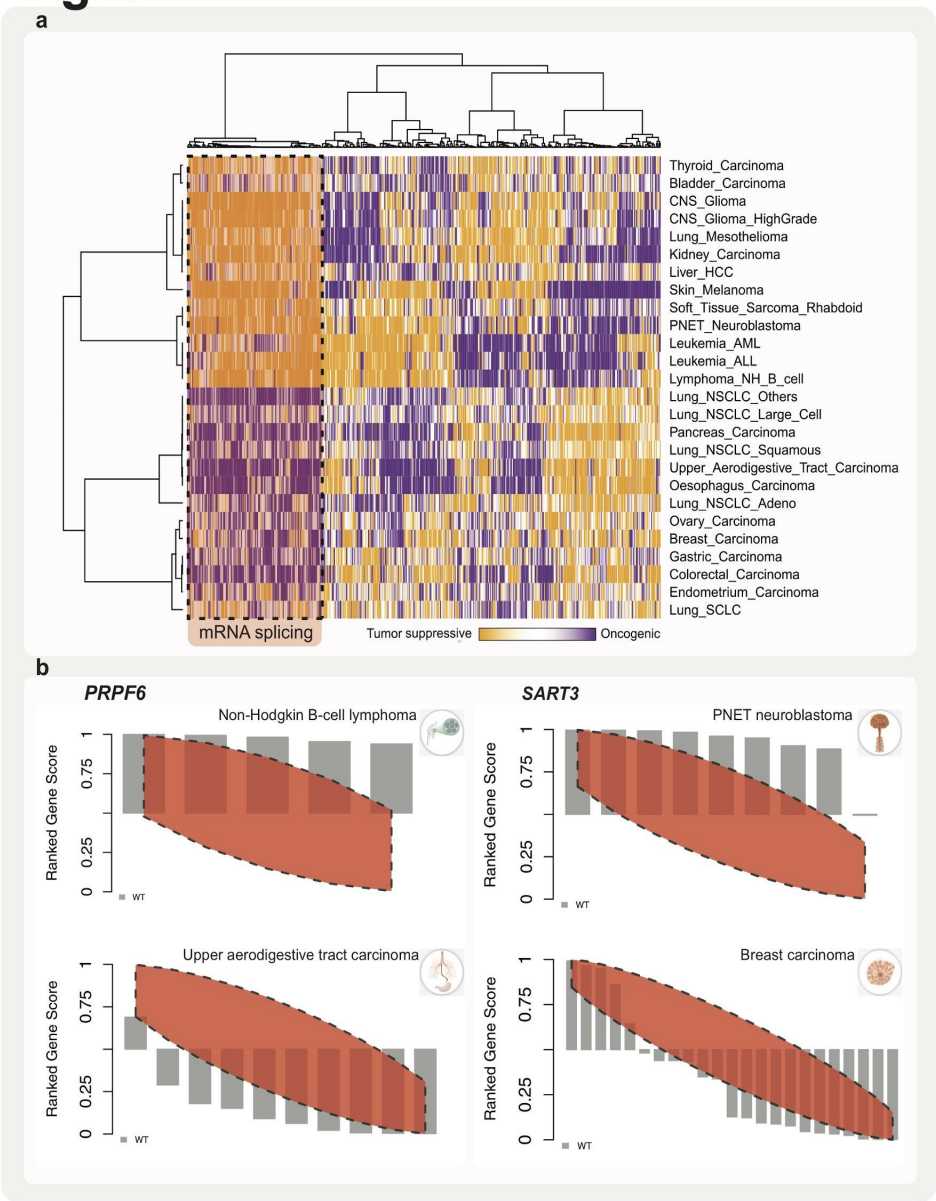

Fig. 3: Pan-cancer analysis of the non-genetic drivers in the DRIVE perturbation screen. **a,** Heatmap illustrates consensus clustering of the 500 most variable *P* values for non-genetic drivers using the APSiC algorithm for the 26 cancer types in the DRIVE perturbation screen. Consensus clustering identified 4 clusters across cancer types and 5 clusters across the 500 genes. One of the clusters was enriched for genes involved in mRNA splicing and processing. **b,** Rank profiles for *PRPF6* in (top) non-Hodgkin B-cell lymphoma and (bottom) upper aerodigestive tract carcinoma and for *SART3* in (top) PNET neuroblastoma and (bottom) breast carcinoma.

Fig. 4

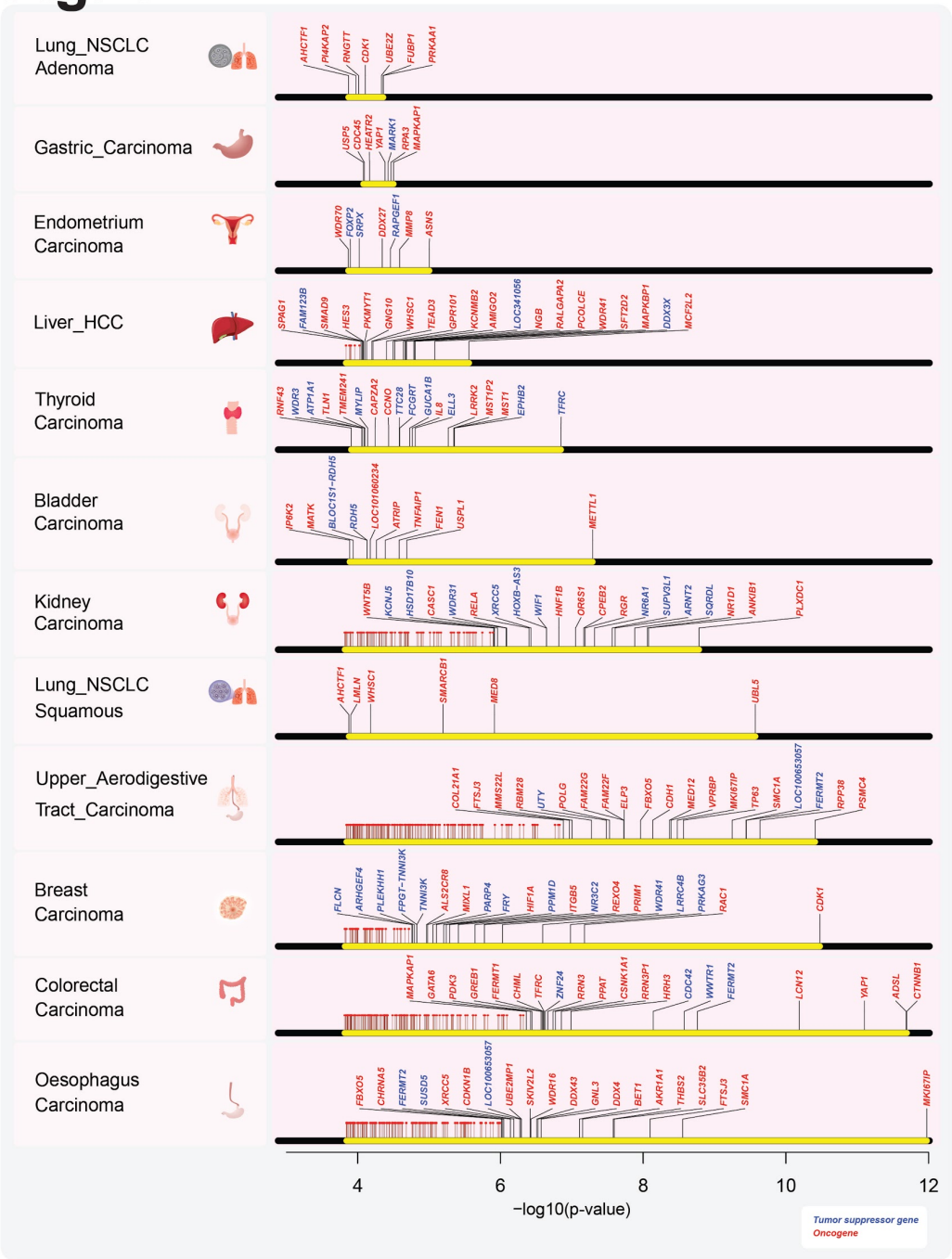

Fig. 4: Cancer type-specific analysis of the non-genetic drivers in the DRIVE perturbation screen. Non-genetic driver genes identified in the 12 cancer types with corresponding gene expression data from the TCGA. APSiC *P* values are shown (on a -$\log_{10}$ scale) for driver genes significant after multiple testing corrections and over- or underexpressed in human cancers for oncogenes (in red) and tumor suppressor genes (in blue), respectively. The top 20 genes for each cancer type are labelled.

Fig. 5

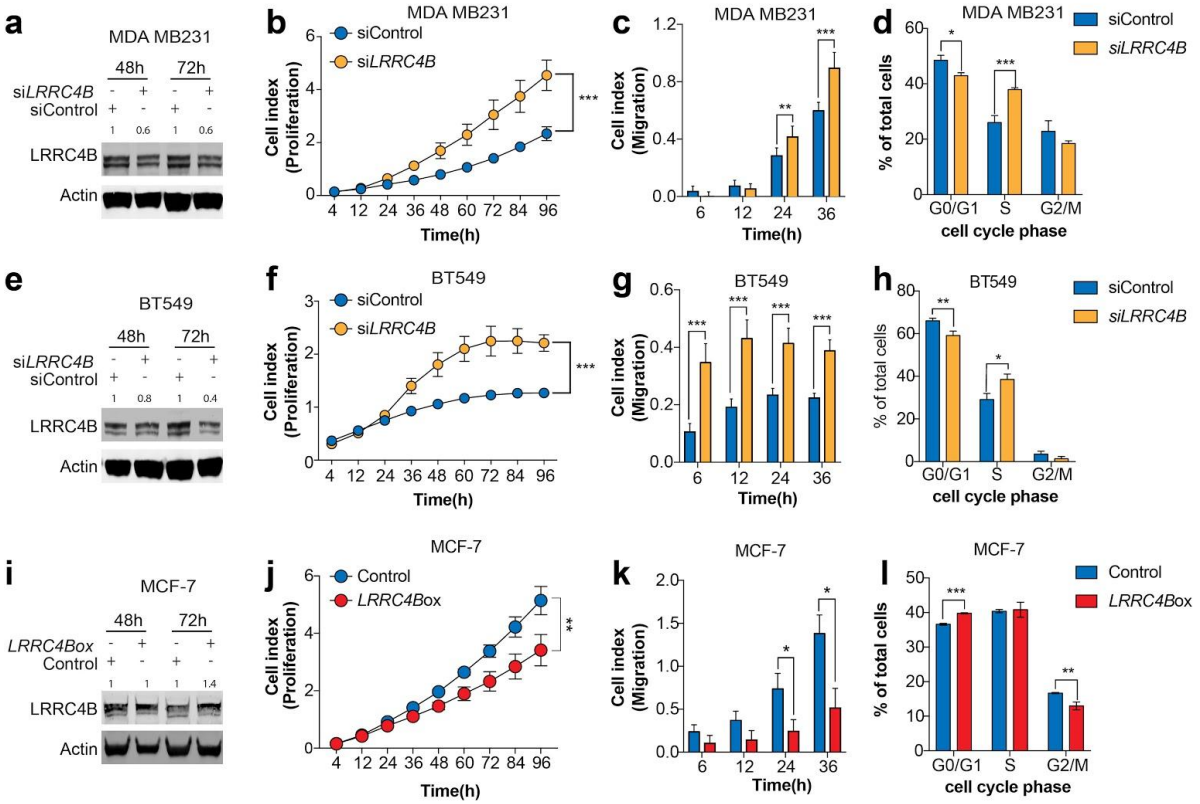

Fig. 5: *LRRC4B* has tumor suppressor-like properties in *in-vitro* models of breast cancer. a, e, i, Western blotting showing LRRC4B protein level in (**a**) MDA-MB231, (**e**) BT-549 and (**i**) MCF-7 cell lines 48 and 72 hours post transfection. Actin was used as a loading control and for normalization. **b, f, j,** Proliferation kinetics of (**b**) MDA-MB231, (**f**) BT-549 and (**j**) MCF-7 cells upon (**b, f**) downregulation or (**j**) upregulation of *LRRC4B* compared with the control. **c, g, k,** Migration potential of **(c)** MDA-MB231, (**g**) BT-549 and (**k**) MCF-7 cells upon (**c, g**) downregulation or (**k**) upregulation of *LRRC4B* compared with the control. **d, h, i,** Cell cycle analysis of (**d**) MDA-MB231, (**h**) BT-549 and (**i**) MCF-7 cells upon (**d, h**) downregulation or (**i**) upregulation of *LRRC4B* compared with the control. Error bars represent standard deviation from three independent experiments. For all experiments, statistical significance was assessed by multiple t-tests (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$).

Fig. 6

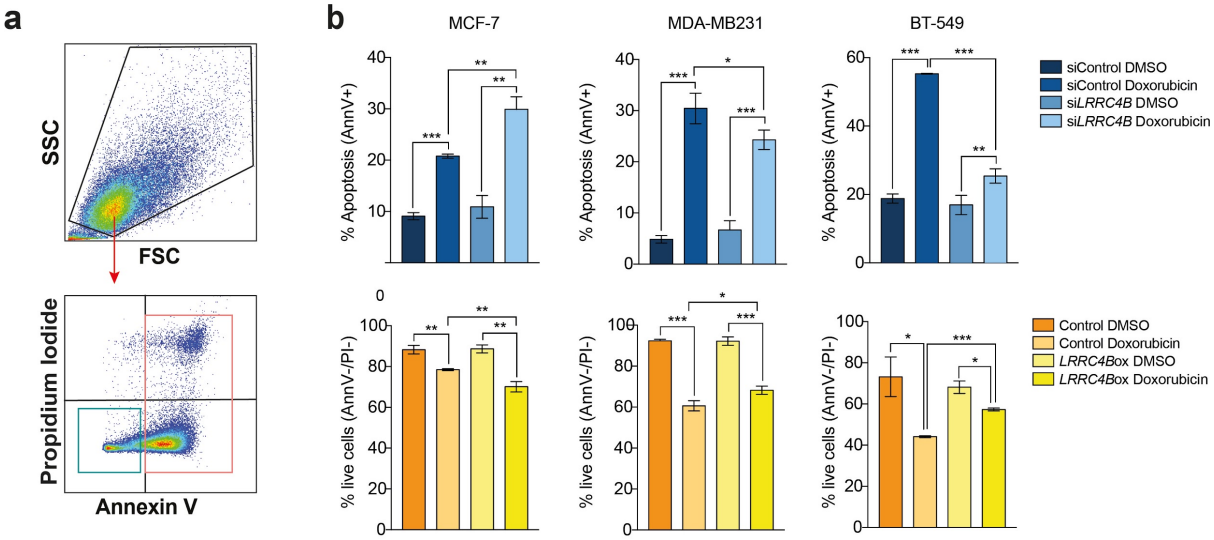

Fig. 6: *LRRC4B* expression modulates induction of apoptosis by Doxorubicin. a, Dot plot illustrating the flow cytometry gating strategy used to assess cell viability and apoptosis using Annexin V and propidium iodide staining of MDA-MB231, BT-549 and MCF-7 cells upon downregulation/ upregulation of *LRRC4B* compared with the control cells, with and without Doxorubicin. **b,** Quantification of the mean (+/- SD) percentage of apoptotic cells (AnnV+) and live cells (AnnV-/PI-) across the different groups. Error bars represent standard deviation from three independent experiments. For all experiments, statistical significance was assessed by multiple t-tests (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$).