Pan-cancer detection of driver genes at the single-patient resolution

Joel Nulsen[1,2], Hrvoje Misetic[1,2], Christopher Yau[3,4], Francesca D. Ciccarelli[1,2,*]

[1]Cancer Systems Biology Laboratory, The Francis Crick Institute, London NW1 1AT, UK

[2]School of Cancer and Pharmaceutical Sciences, King's College London, London SE11UL, UK

[3]School of Health Sciences, University of Manchester, Manchester M13 9PL, UK

[4]The Alan Turing Institute, London NW1 2DB, UK

* = corresponding author (francesca.ciccarelli@crick.ac.uk)

ABSTRACT

Identifying the complete repertoire of genes that drive cancer in individual patients is crucial for precision oncology. Established methods for driver detection focus mostly on genes that are recurrently altered across cohorts of cancer patients. However, mapping these genes back to patients leaves a sizeable fraction with few or no driver events, hindering our understanding of cancer mechanisms and limiting the choice of therapeutic interventions. Here we present sysSVM2, a tool based on machine learning that integrates somatic alteration data with systems-level gene properties to predict drivers in individual patients. We develop sysSVM2 for pan-cancer applicability, demonstrating robust performance on real and simulated cancer data. We benchmark its performance against other driver detection methods and show that sysSVM2 has a lower false positive rate and superior patient driver coverage. Applying sysSVM2 to 7,646 samples from 34 cancer types, we find that predicted drivers are often rare or patient-specific. However, they converge to disrupt well-known cancer-related processes including DNA repair, chromatin organisation and the cell cycle. sysSVM2 is a resource to enhance personalised predictions of cancer driver events with possible use in research and clinical settings. Code to implement sysSVM2 and the trained models in simulated cancer-agnostic data as well as in 34 cancer types are available at https://github.com/ciccalab/sysSVM2.

INTRODUCTION

Cancer is characterised by the acquisition of somatic alterations of the genome, the majority of which are thought to have little or no phenotypic consequence for the development of the disease. Identifying the genes whose alterations instead have a role in driving cancer (cancer drivers) is one of the major goals of cancer genomics and numerous methods have been developed so far to achieve this.

Most of these methods work at the cohort-level, which means that they identify driver genes within a cohort of patients. For example, recurrence-based methods such as MutSigCV[1] and MuSiC[2] search for genes whose mutation rate (single nucleotide variants (SNVs) and small insertions or deletions (indels) per nucleotide) is above the background level. This is because mutations in cancer drivers are more likely to become fixed and recur across samples than those in non-driver genes. GISTIC2[3] adopts a similar approach for recurrent copy number variants (CNVs). OncodriveCLUST[4] and ActiveDriver[5] look specifically for mutations clustering in hotspots or encoding post-translational modification sites. TUSON[6] and 20/20+[7] predict new drivers based on features of canonical oncogenes and tumour suppressors, including the proportion of missense or loss-of-function to silent mutations occurring across patients. dNdScv[8] computes the nonsilent to silent mutation ratio to identify gene mutations under positive selection, while OncodriveFM[9] focuses on biases towards variants of high functional impact. Finally, network-based methods like HotNet2[10] incorporate gene interaction networks to identify significantly altered modules of genes within the cohort. Albeit with different approaches, all these methods rely on the comparison of alterations and/or altered genes across patients.

Cohort-level methods have been of great value leading to the identification of more than 2,000 well-established (canonical) or candidate cancer driver genes[11,12]. However, these approaches fail to identify rare driver events that occur in small cohorts or even in single patients because of low statistical power. Moreover, they are not ideal for application in the clinical setting because they return lists of drivers in entire cohorts, rather than predictions in individual patients.

Patient-level methods ideally predict cancer drivers in each patient but are more challenging to implement. A few attempts such as OncoIMPACT[13], DriverNet[14] and DawnRank[15] combine transcriptomic and genomic data to identify gene network deregulations in individual samples. Such methods require user-specified gene networks and deregulation thresholds, which can affect their results[13]. In addition, matched exome and transcriptome data from the same sample are not always available. Alternative approaches such as PHIAL[16] match the patient mutations with databases of known clinically actionable or driver alterations, with a limited capacity to identify as-yet unknown driver alterations. To overcome this limitation, iCAGES[17] combines deleteriousness predictions and curated database annotations to learn features of true positive and true negative driver alterations.

We recently developed sysSVM, a patient-level driver detection method based on one-class support vector machines (SVMs)[18]. sysSVM learns the distinct molecular features (damaging somatic alterations) and systems-level features (gene properties) of canonical drivers. It then predicts as drivers the altered genes in individual patients that best resemble these features. When applied to 261 patients with oesophageal adenocarcinomas, sysSVM successfully identified the driver events in every patient[18].

Here, we further develop sysSVM to be applied to any cancer type and benchmark it against other available approaches, showing that it has a lower false positive rate and better patient driver coverage. We also develop optimal models for identifying driver genes in all 34 cancer types available in The Cancer Genome Atlas (TCGA)[19]. These optimised models and their associated driver predictions are provided as a resource that can be used to identify and study driver events in cancers at the single patient resolution.

RESULTS

Simulation of pan-cancer datasets

The sysSVM approach to driver detection prioritises genes with features similar to those of canonical cancer drivers (Figure 1A, **Supplementary Note**). First, sysSVM first identifies a set of true positive canonical drivers damaged within a cohort of samples. It then uses their features to train four one-class SVMs (linear, radial, sigmoid, polynomial). Finally, it ranks the remaining damaged genes in individual cancer patients based on a combined score from the four SVMs. Highly ranked genes are most similar to canonical drivers and will be the cancer drivers for that patient.

In order to optimise the use of sysSVM for any cancer type, we simulated 1,000 cancer-agnostic samples starting from all TCGA tumours with matched mutation, CNV and gene expression data (**Methods**). We ensured that the tumour mutation and copy number burdens were similar between real and simulated samples (Figure 1B) and that gene mutation and copy number status in the simulated dataset was the same of TCGA (Supplementary Figure 1). As a result, the frequency of damaging alterations in known oncogenes and tumour suppressors was comparable between the two datasets, with *TP53*, *PIK3CA* and *CDKN2A* among the most frequently altered genes in both (Figure 1C).

The simulated cohort for sysSVM optimisation (hereafter referred to as the reference cohort) was composed of 1,000 samples with 18,455 genes damaged 309,427 times. Of these, 686 were canonical drivers with an experimentally proven role in cancer[12,20], 1,605 were candidate cancer genes from 273 cancer screens[11], 43 were known false positive predictions of driver detection methods[1,21], and 16,121 were the remaining damaged genes (Figure 1D, Supplementary Table 1). We annotated seven molecular and 25 systems-level features of all damaged genes (Supplementary Table 2) and used these features for training and prediction. As a training set, we selected 457 of 686 canonical drivers with proven roles as oncogenes (236) or tumour suppressors (221). We used the remaining 229 damaged canonical drivers, 43 false positives and 16,121 remaining damaged genes (hereafter referred to as the rest of genes) for prediction and performance assessment (Figure 1D).

sysSVM2 optimisation on the pan-cancer reference cohort

Using the reference cohort, we optimised sysSVM in terms of data normalisation, parameter tuning and feature selection (Figure 1A). Not to bias the optimisation with a particular set of kernel parameters, we implemented 512 models with parameter combinations representing a sparse coverage of a standard grid search (**Supplementary Note**). We then measured the ability of each of these 512 models to prioritise 229 canonical drivers not used for training over the rest of damaged genes or the false positives. This was done by computing the Area Under the Curve (AUC) in each sample and considering the median AUC over the whole cohort as representative of the whole cohort.

First, we derived the optimal settings for data normalisation in terms of centered and un-centered data (**Supplementary Note**). All models robustly prioritised canonical drivers above the rest using either centered or un-centered data but showed lower performance in distinguishing canonical drivers from false positives (Figure 2A). We reasoned that false positives from recurrence-based driver detection methods[1] share some features with canonical drivers. For example, they encode long and multi-domain proteins. When removing protein length and number of domains from the feature list (Supplementary Table 2), the performance substantially improved particularly for un-centred data (Figure 2B).

Second, we selected the optimal sets of parameters in each kernel. Hyper-parameter choice is known to have substantial impacts on classification but it is an open problem for one-class SVMs[22]. Since the parameters for each kernel needed to be selected separately (**Supplementary Note**), we could not use AUC for assessment. Instead, we used the sensitivity of the 512 models to predict canonical drivers calculated from three-fold cross-validation on the training set. Sensitivity was indeed a good predictor of the AUC of canonical drivers over the rest of genes (Figure 2C) and false positives (Figures 2D). We therefore developed an approach to select the parameters that conferred the highest sensitivity in multiple iterations of cross-validation (**Methods**). In the reference cohort, parameters chosen in this way converged within 2,000 cross-validation iterations for all kernels (Supplementary Figure 2A).

Finally, we systematically assessed the correlation between all 25 systems-level features since the presence of highly correlated features can hinder SVM performance[23]. Four features (selective mRNA expression, ubiquitous mRNA expression, limited protein expression, and central position in the protein interaction network) exhibited a significant degree of inter-correlation (Supplementary Figure 2B). Removing them led to earlier convergence of kernel parameters (Supplementary Figure 2A) and improved performance (Supplementary Figure 2C).

We further assessed the performance of the optimised sysSVM in prioritising cancer drivers over other genes. We confirmed that, overall, the prediction scores of 229 canonical drivers outside the training set were significantly higher than those of any other gene category (Figure 2E). Candidate cancer genes also scored significantly higher than the rest of genes, indicating that they were also in top ranking positions. We also measured the relative ranks of genes in individual samples using Receiver Operating Characteristic (ROC) curves. Comparing canonical drivers to the rest of genes and to false positives gave AUCs of 0.73 and 0.93, respectively (Figure 2F), demonstrating that canonical drivers were prioritised above the rest of genes and especially above false positives.

Based on these results, we chose the default settings for the cancer-agnostic SVM classifier, which we named sysSVM2. By default, data are un-centered but scaled to have unit standard deviation. Six of the original systems-level features are excluded resulting in a total of seven molecular and 19 systems-level features (Supplementary Table 2). Finally, kernel parameters optimised on the reference cohort are provided as a default (Supplementary Figure 2A), although users may perform cross-validation iterations on their own cohorts.

Effect of training cohort size on sysSVM2 performance

The sample size of patient cohorts can highly vary across cancer types. For example, in TCGA it ranges from 32 samples for diffuse large B-cell lymphoma (DLBC) to 726 for breast cancer (BRCA, Supplementary Table 3), with a median of 201 samples. We therefore sought to address how the sample size of the training cohort affected sysSVM2 performance.

Starting from all TCGA samples and using the previously described approach, we simulated 40 training cohorts, ten of which were composed of ten samples, ten of 100 samples, ten of 200 samples, and ten of 1,000 samples. We then trained sysSVM2 on each of these 40 cohorts independently and used the resulting models to rank damaged genes in the reference cohort and to compare their performance.

The distributions of AUCs of canonical drivers over the rest of genes or false positives were high for all four cohort sizes (Figure 3A). This suggested that sysSVM2 is overall very effective in prioritising cancer genes independently of the training cohort size. We then compared the composition of prioritised gene list in each sample across models of a given size. We measured a composition score of the top five genes accounting for the number and position of canonical drivers, candidate cancer genes and false positive genes (**Methods**). Similar to the AUC, the composition score of the top five genes was also very similar across training cohorts (Figure 3B). However, a few models trained on ten or 100 samples returned false positives in the top five positions while no false positives were predicted by models trained on larger cohorts of 200 or 1,000 samples. Finally, irrespective of the training cohort size >80% and ~50% of samples had at least one and two canonical drivers in the top five genes, respectively (Figure 3C).

Since we used the same reference cohort for prediction, we could directly compare the gene ranks in each patient across models, thus assessing their prediction stability. To do so, we measured the Rank-Biased Overlap (RBO) score that compares two ranked lists giving greater weight to the higher-ranked positions[24]. The distributions of RBO scores of the top five genes were significantly higher for large training cohorts compared to those composed of ten samples (Figure 3D). Moreover, models trained on large cohorts showed overall higher gene overlap in the top five genes (Figure 3E).

These results show that, although sysSVM2 successfully separates canonical drivers from other genes independently of the training cohort size, small cohorts lead to occasional false positive predictions and to unstable gene ranking. Since the median cohort size of TCGA cancers is 201 samples, sysSVM2 is likely to separate canonical drivers from the rest of genes with a very low false positive rate and stable gene rankings for most cancer cohorts.

Benchmark of sysSVM2 against existing methods

Next, we sought to compare the predictions of sysSVM2 on real cancer data to those of other driver detection methods. To do this, we used 657 Gastro-Intestinal (GI) adenocarcinomas from TCGA (73 oesophageal, 279 stomach, 219 colon and 86 rectal cancers, Supplementary Table 3). Overall, this cohort had 17,122 unique damaged genes, including 438 tumour suppressors and oncogenes used for sysSVM2 training (Supplementary Table 1). After ranking the remaining 16,684 damaged genes, we confirmed the overall ability of sysSVM2 to prioritise the 228 canonical drivers not used for training over the rest of damaged genes and false positives also in real data (Figure 4A).

To identify the list of cancer drivers of each patient, we adopted a top-up approach. Starting from the GI canonical drivers[11] damaged in each sample, sysSVM2 predictions were progressively added based on their rank to reach five drivers per patient (**Methods**). This was based on the assumption that each cancer requires at least five driver events to fully develop, in concordance with recent quantifications of the amount of excess mutations arising from positive selection in cancer[8,25]. One-hundred and fifty-four patients had damaging alterations in five or more GI canonical drivers, while 503 patients (77%) needed at least one prediction (Figure 4B). This resulted in 564 unique sysSVM2 drivers.

We then predicted the drivers in the same GI samples using two cohort-level (PanSoftware[26] and dNdScv[8]) and two patient-level (OncoIMPACT[13] and DriverNet[14]) detection methods. PanSoftware integrates 26 computational driver prediction tools and we took the list of 40 damaged drivers directly from the original publication[26], given that we used a large subset (87%) of the same TCGA GI samples. We ran the other three methods with default parameters (**Methods**) and obtained 25 predicted drivers with dNdScv, 607 with DriverNet and 1,345 with OncoIMPACT.

We compared sysSVM2 to the four other methods in terms of recall rates of canonical drivers or false positives, proportion of novel predictions and patient driver coverage. Overall, cohort-level methods had higher recall rates of GI canonical drivers, fewer novel predictions and a comparably low false positive recall than

sysSVM2 (Figure 4C). However, unlike sysSVM2, neither cohort-level method predicted drivers in all patients, leaving the vast majority of them with less than five predictions and some with no predictions (Figure 4D).

Compared to sysSVM2, the other two patient-level methods had higher recall rates of the 228 canonical drivers, a comparably high proportion of novel predictions but higher false positive rate (Figure 4C). Namely, sysSVM2 made only one false positive prediction in one patient while DriverNet and OncoIMPACT predicted four and seven false positives in 124 and 306 patients, respectively (Supplementary Figure 3A). Overall, all three methods had high patient driver coverage, but sysSVM2 outperformed the other two with only one sample where it predicted less than five drivers (Figure 4D). Interestingly, the overlap of predictions between sysSVM2 and the other patient-level methods was statistically significant (Supplementary Figure 3A) even when only top-up predictions were considered (Supplementary Figure 3B). This suggested that the majority of predictions converged to the same genes.

These results showed that cohort-level methods have high specificity and sensitivity to identify cancer-specific canonical drivers but often fail to find drivers in a substantial subset of patients. Compared to other patient-level detection methods, sysSVM2 outperforms them in terms of specificity and patient coverage.

Compendium of sysSVM2 models and patient-level drivers in 34 cancer types

In order to provide a comprehensive resource of trained models and patient-level drivers, we sought to apply sysSVM2 to 7,646 TCGA samples of 34 cancer types with at least one somatically damaged gene (**Methods**).

To find the best training setting for the algorithm on real cancer samples, we compared the performance of sysSVM2 trained on the whole pan-cancer cohort as well as on the 34 cancer types separately. In the pan-cancer setting, we used all 477 tumour suppressors and oncogenes damaged across the whole cohort. In the cancer-specific setting, we used instead only the subsets of these genes damaged in each cancer type (Supplementary Table 3). We then predicted on the remaining damaged genes and applied the top-up approach as described above, starting from the cancer-specific canonical drivers damaged in each patient (Supplementary Table

3). We found that 6,067 samples (79%) required at least one sysSVM2 prediction in order to reach five drivers (Figure 5A). These corresponded to 4,369 and 4,548 unique genes in the pan-cancer and cancer-specific settings, respectively, with a significant overlap of predictions (3,896, $p < 2.2 \times 10^{-16}$, two-sided Fisher's exact test).

We then compared the performance of pan-cancer and cancer-specific settings of sysSVM2 in prioritising canonical drivers over rest of genes or false positives. The AUCs differed significantly (FDR <0.05) and substantially (|difference in medians| >0.05) in only five cancer types (Figure 5B, Supplementary Figure 4A). All of them were composed of small cohorts with <200 samples and in all cases the pan-cancer setting showed better performance than the cancer-specific setting. The composition score of the top five predictions also differed significantly and substantially (|difference in medians| >1) in only three cancer types (Figure 5C, Supplementary Figure 4A). All these cancer types were again characterised by small training cohorts and showed higher performance in the pan-cancer setting. Overall, these results confirmed the trend observed in the simulated data and indicated that the pan-cancer and cancer-specific settings performed similarly well in most cases, except for small cohorts where the pan-cancer model performed better.

Based on these results, we used the pan-cancer setting for cancer types with small cohorts (N <200) and the cancer-specific setting for the others, as this could reflect cancer-type specific biology without jeopardising performance or stability. The trained models for each cancer type are provided as a resource for driver prediction.

The final list of patient-specific predictions in 34 cancer types was composed of 4,470 unique genes, the vast majority of which (93%) were rare (<10 patients) or patient-specific (Figure 5D, Supplementary Table 4). A gene set enrichment analysis on these genes revealed 984 enriched pathways overall (Reactome level 2 or above, FDR <0.01, Supplementary Table 5). Interestingly, when mapping these pathways to broader biological processes (Reactome level 1), a few processes were widely enriched in almost all cancer types (Figure 5E). These included well-known cancer-related processes such as chromatin organisation[27], DNA repair[28], cell cycle[29] and signal transduction[30]. Therefore, although not recurring across patients, sysSVM2 predictions converged to perturb similar biological processes that are known to contribute to cancer.

DISCUSSION

Identifying the complete repertoire of driver events in each cancer patient holds great potential for furthering the molecular understanding of cancer and ultimately for precision oncology. While many recurrent driver genes have now been identified, the highly heterogeneous long tail of rare drivers still poses great challenges for detection, validation and therapeutic intervention.

In this work, we developed a cancer-agnostic algorithm, sysSVM2, for identifying driver genes in individual patients. By refining the machine learning approach upon which the original algorithm was built[18], we broadened its applicability to the pan-cancer range of malignancies represented in TCGA. sysSVM2 successfully and stably prioritises canonical driver genes for most publicly available cancer cohorts. For those composed of fewer samples, the models optimised on the whole pan-cancer dataset offer a valid alternative. Moreover, compared to other patient-level driver detection methods, sysSVM2 has better patient coverage, and a particularly low rate of predicting established false positives.

It is ultimately the predictions that do not have a firmly established role in cancer that are of greatest interest. We found that these genes converged to well-known cancer-related biological processes and further studies could potentially use these predictions to investigate particular aspects of cancer biology, such as driver clonality and their progressive acquisition during cancer evolution. Extending the algorithm with additional sources of data is another avenue for future work. For example, transcriptomic and epigenomic data could enhance the ability of sysSVM2 to identify driver events. Additionally, recent efforts have identified a large number of driver events in non-coding genomic elements[25]. Given such a training set of true positives, sysSVM2 could be further developed to identify non-coding drivers in individual patients, as long as appropriate features could be identified. The general approach of identifying drivers using a combination of molecular and systems-level properties affords great flexibility for such developments.

It is increasingly common for sequencing studies to integrate multiple tools for driver detection[11,26], since building a consensus can make results robust to the weaknesses of individual methods. sysSVM2 also has its weaknesses. For example, while systems-level properties distinguish cancer genes as a set, there are some

cancer genes that do not follow this trend[11], and are thus likely to be missed by the algorithm. Our approach in the current work of topping up known driver genes with predictions from sysSVM2 is a simple example of how sysSVM2 can be used in conjunction with other approaches. More broadly, it is likely the case that patient-level driver detection will eventually rely on an entire ecosystem of different methods. In this work, we have demonstrated that there is a place for sysSVM2 in such an ecosystem.

METHODS

Pan-cancer data simulation

Sequence mutations (SNVs and indels) for 9,079 samples were obtained from the MC3 release of TCGA[26] and annotated with ANNOVAR[31] (downloaded April 2018) and dbNSFP v3.0[32]. Only mutations identified as exonic or splicing were retained. Damaging mutations included (1) truncating (stopgain, stoploss, frameshift) mutations; (2) missense mutations predicted by at least five out of seven functional prediction methods (SIFT[33], PolyPhen-2 HDIV[34], PolyPhen-2 HVAR, MutationTaster[35], MutationAssessor[36], LRT[37] and FATHMM[38]) , at least two out of three conservation-based methods (PhyloP[39], GERP++RS[40] and SiPhy[41]); (3) splicing mutations predicted by one of two splicing-specific methods (ADA[32] and RF) and (4) hotspot mutations identified by OncodriveCLUST[4] v1.0.0.

Copy number data (array intensities) for 11,379 samples were obtained from the Genomic Data Commons portal (https://portal.gdc.cancer.gov/). Copy Number Variant (CNV) segments, sample ploidy and sample purity values were obtained using ASCAT[42] v2.5.2, with 9,873 samples passing quality control. Segments were intersected with the exonic coordinates of 19,549 human genes that were derived as previously described[11] in the reference genome hg38, and converted to hg19 coordinates using the UCSC liftOver tool (http://genome.ucsc.edu/). Genes were considered to have undergone a CNV if at least 25% of their transcribed length was covered by a segment. RNA-Seq data were used to filter out false positive CNVs. Fragments per kilobase million (FPKM) values were obtained for 10,974 samples and only samples with matched copy number and expression data were retained. Damaging CNVs included homozygous gene losses (copy number 0 and FPKM <1 over mean cancer type purity) and gene amplifications (copy number >2 x sample ploidy).

Considering only one sample per patient, a total of 7,630 samples had matched mutation, CNV and RNA-Seq data that passed ASCAT quality controls and had at least one damaged gene. These comprised 535,615 damaging mutations and 2,041,598 damaging CNVs in 18,784 genes (Supplementary Table 1).

To simulate samples that reproduced the molecular features of real TCGA samples, the damaging mutation burden and ploidy were measured for the whole

TCGA cohort. Then, 1,000 random combinations of damaging mutation burden and ploidy were extracted and assigned to the simulated samples. To preserve the same frequency of damaging alterations, damaged genes were extracted from within TCGA samples that had similar values of damaging mutation burden (+/-10%, for mutations) and ploidy (+/-0.1, for CNVs) and assigned to the simulated samples. Overall, the final simulated dataset of 1,000 samples contained 69,269 damaging mutations and 252,409 damaging CNVs in 18,455 genes.

As a training set, 220 tumour suppressors with 2,433 loss-of-function alterations (truncating mutations, missense or splicing damaging mutations, homozygous deletions, or double hits) and 236 oncogenes with 5,539 gain-of-function alterations (hotspot mutations, missense or splicing damaging mutations or gene amplifications) were retained. For *TP53* both loss- and gain-of-function alterations (n=352) were considered to be driver events and included in the training set. 301,103 damaging alterations in the remaining 17,998 human genes not present in the training set were used for prediction (Supplementary Table 1).

Annotation of systems-level properties

Systems-level properties of human genes were obtained from various sources. Duplicated gene loci (genes with 60% protein sequence shared with another locus) were identified as previously described[11], and ohnolog status was obtained from Nakatani *et al.*[43]. Data on gene essentiality in cell lines were obtained from the PICKLES (September 2017)[44] and OGEE v2[45] databases. Genes in PICKLES were considered to be essential in a cell line if they had Bayes factor >3, while original annotations of essentiality from OGEE were retained. mRNA expression data for healthy human tissues were downloaded from GTEx v7[46] and the Protein Atlas v18[47], and a gene was considered to be expressed in a tissue if median TPM >1 in both databases (or in one database if that tissue was not included in both). Categorical protein expression data (indicating whether or not a protein was expressed) for healthy human tissues was taken from the Protein Atlas v18[47]. The protein-protein interaction network (PPIN) was constructed as previously described[11], from the union of BioGRID v3.4.157[48], MIntAct v4.2.10[49], DIP (February 2018)[50] and HPRD v9[51], and topological properties (degree, betweenness and centrality) were calculated using custom scripts. Genes were identified as

participating in complexes using data taken from CORUM (July 2017)[52], HPRD v9[51] and Reactome v63[53]. The number of miRNAs targeting a gene was calculated using data from miRTarBase v7.0[54] and miRecords v4.0[55]. The evolutionary origin (age) of genes was identified as previously described[56], using data from EggNOG v4.5.1[57]. Protein length was obtained from RefSeq[58] release 94, and the number of protein domains were obtained from InterPro[59] annotations in UniProt[60] release 2019_07. From these properties, 25 systems-level features were derived to be used for classification, of which 19 were retained after feature selection (Supplementary Table 2).

For each feature, missing values were imputed using the median or mode (for numeric and categorical features, respectively) of available data for canonical drivers and the rest of genes separately. All features used in the model significantly differentiated cancer drivers from other human genes (Supplementary Table 2).

Kernel parameter tuning

sysSVM2 uses four one-class Support Vector Machines (SVMs), with linear, radial, sigmoid and polynomial kernels following three steps: feature mapping, model selection, and training and prediction (Figure 1A, **Supplementary Note**). sysSVM2 is implemented in R[61], and SVM training and prediction are carried out using the e1071[62] package.

Model selection was updated from the original sysSVM formulation to improve convergence. A parameter grid search was carried out for each kernel separately, for a total of 196 kernel-parameter combinations (**Supplementary Note**). The aim was to select parameters that resulted in a model with a high sensitivity and stability (low standard deviation of sensitivity). The model sensitivity for each parameter combination was assessed on the simulated training set using three-fold cross-validation for 5,000 iterations. For each kernel k and parameter combination i, the mean μ_{ki} and standard deviation σ_{ki} of the sensitivity were calculated across the cross-validation iterations. These were then converted into z-scores $z_{ki}^{(\mu)}$ and $z_{ki}^{(\sigma)}$, which measured the relative values of mean and standard deviation between the different parameter combinations such that:

$$\sum_i z_{ki}^{(\mu)} = \sum_i z_{ki}^{(\sigma)} = 0$$

and

$$\text{Variance}_i\left(z_{ki}^{(\mu)}\right) = \text{Variance}_i\left(z_{ki}^{(\sigma)}\right) = 1.$$

Finally, we defined the Δ_z score as:

$$\Delta_z = z_{ki}^{(\mu)} - z_{ki}^{(\sigma)}$$

High Δ_z scores corresponded to parameter combinations that had high mean sensitivity and low standard deviation relative to the other combinations for that kernel. The four parameter combinations (one per kernel) with the highest Δ_z scores were selected and used to train the four kernels on the entire training set.

Performance assessment

The performances of all driver prediction models tested in sysSVM2 were measured using four metrics: Area Under the Curve (AUC); composition score; Rank-Biased Overlap (RBO) score and overlap of the top five predictions between models.

For the AUC, Receiver Operating Characteristic (ROC) curves were first derived for each sample individually based on the ranks of all damaged genes. This was done comparing the ranks of canonical drivers both to known false positives, and to the rest of genes. To derive the median ROC curve, for each value of a false positive rate (x-axis) the median true positive rate (y-axis) across samples was calculated.

The composition score assessed the top five predictions in each sample, and measured the prevalence and ranks of different types of genes. The score *S* was calculated as a weighted sum according to the following formula:

$$S = \sum_{g=1}^{5} w_g \times t_g.$$

The weight w_g of each gene *g* in the top five was such that higher-ranked genes were assigned greater weight. Specifically, $w_g = 6 - r_g$ where r_g is the rank of gene g (with 1 being the highest). The type contribution t_g of gene g was defined for

different gene categories as follows: cancer-specific canonical drivers ($t_g = 3$); other canonical drivers (2); cancer-specific candidate genes (1.5); other candidate cancer genes (1); false positives (-1); other genes (0).

The RBO score[24] was used to assess the similarity of the top five predictions from pairs of models. It measures the overlap of ranked lists at incrementally increasing depths using a convergent series. Including a correction for finite lists of length five, it was calculated according to the formula:

$$\mathrm{RBO} = \frac{1-p}{1-p^5}\sum_{d=1}^{5} p^{d-1} \times A_d.$$

where p was set to 0.9[24], d indicates depth in the rankings (starting from the top-ranked elements), and A_d is the overlap of the two lists, restricted to depth d.

TCGA sample analysis

Starting from the TCGA samples with matched mutation, copy number and RNA-Seq data, more stringent filters were applied for homozygous deletions and gene amplifications. Genes with CN=0 that either (1) had one or more mutations in their sequence, or (2) were expressed at greater than 1 FPKM over sample purity (as opposed to mean cancer type purity) were re-annotated as having CN=1. Amplifications were filtered by applying Wilcoxon tests to compare expression levels between amplified and non-amplified cases, for each gene separately. Only genes for which amplification was associated with overexpression at FDR <0.05 were retained as having damaging amplifications. The resulting dataset comprised 7,651 samples at least one damaging alteration. To prevent sysSVM2 training sets from being biased by individual samples, five samples each accounting for more than 50% of the damaged genes in their respective cancer types were removed, leaving 7,646 samples for further analysis. For the purposes of measuring performance and overlap of the cancer-specific and pan-cancer sysSVM2 settings, 752 samples with fewer than ten damaged genes were removed.

To tune the kernel parameters of sysSVM2, 10,000 three-fold cross validations were run in each cancer type. In cases where parameters did not converge, additional 15,000 cross-validations were run only for the non-converging parameters, with all other parameters fixed at their converged values. This was done

for gamma of the sigmoid kernel in adrenocortical carcinoma (ACC), oesophageal adenocarcinoma (OAC), colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (KIRC) and stomach adenocarcinoma (STAD), as well as for the degree parameter of the polynomial kernel in lung squamous cell carcinoma (LUSC). Prediction were assessed using the same metrics as for simulated data (AUC, composition score, RBO score and overlap of top five predictions).

Lists of cancer driver genes in individual TCGA samples were identified as follows. First, a list of canonical driver genes for each of the 34 cancer types in TCGA was obtained from NCG[11]. Cancer-specific canonical drivers damaged in each sample were considered as cancer drivers for that sample. In samples with five or more such drivers, no further prediction was done. Otherwise, the highest-ranked genes from sysSVM2 were added so that there were five drivers in total. Samples with five or fewer damaged genes overall were not considered for the pan-cancer analysis. For the purpose of comparing sysSVM2 to other driver detection methods on gastro-intestinal cancer data one sample with less than five damaged genes (TCGA-FP-8210, stomach cancer) was included for completeness.

sysSVM2 predictions in 657 gastro-intestinal (GI) cancer samples were compared to those of PanSoftware[26]; dNdScv[8]; OncoIMPACT[13]; and DriverNet[14]. Forty genes identified as GI cancer drivers by PanSoftware was taken from the original publication[26]. These genes were considered as drivers in every sample in which they were damaged. dNdScv was implemented with default parameters, taking all mutations as input. Genes were considered to exhibit significant signs of positive selection if they had a FDR <0.05. OncoIMPACT was implemented with default parameters, using the gene network provided on GitHub (https://github.com/CSB5/OncoIMPACT). Nonsilent mutations were used as input for mutations. Copy number amplifications were filtered as described above, and copy number losses included heterozygous and homozygous deletions. In each sample, all genes that shared the highest score were taken as the predicted drivers. DriverNet was implemented with default parameters, using the same gene network and data inputs as for OncoIMPACT. The events identified in each sample were taken as the predicted drivers.

Human pathways for gene set enrichment analysis were obtained from Reactome v72[63]. Before testing, pathways were restricted to those of level 2 or

higher, and with between 10 and 500 genes (total of 1,429 pathways containing 10,178 genes). In each cancer type separately, the unique set of top-up predictions from sysSVM2 was tested for enrichment in pathways containing at least one prediction, using one-sided hypergeometric tests The resulting p-values across all cancer types were corrected for False Discovery Rate (FDR) using the Benjamini-Hochberg method[64], as a single set.

REFERENCES

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

2 Dees, N. D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res* **22**, 1589-1598, doi:10.1101/gr.134635.111 (2012).

3 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41, doi:10.1186/gb-2011-12-4-r41 (2011).

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

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

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

7 Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A* **113**, 14330-14335, doi:10.1073/pnas.1616440113 (2016).

8 Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e1021, doi:10.1016/j.cell.2017.09.042 (2017).

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

10 Leiserson, M. D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* **47**, 106-114, doi:10.1038/ng.3168 (2015).

11 Repana, D. *et al.* The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol* **20**, 1, doi:10.1186/s13059-018-1612-0 (2019).

12 Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696-705, doi:10.1038/s41568-018-0060-1 (2018).

13 Bertrand, D. *et al.* Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res* **43**, e44, doi:10.1093/nar/gku1393 (2015).

14 Bashashati, A. *et al.* DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol* **13**, R124, doi:10.1186/gb-2012-13-12-r124 (2012).

15 Hou, J. & Ma, J. DawnRank: discovering personalized driver genes in cancer. *Genome Medicine* **6**, doi:10.1186/s13073-014-0056-8 (2014).

16 Van Allen, E. M. *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* **20**, 682-688, doi:10.1038/nm.3559 (2014).

17 Dong, C. *et al.* iCAGES: integrated CAncer GEnome Score for comprehensively prioritizing driver genes in personal cancer genomes. *Genome Med* **8**, 135, doi:10.1186/s13073-016-0390-0 (2016).

18 Mourikis, T. P. *et al.* Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. *Nat Commun* **10**, 3101, doi:10.1038/s41467-019-10898-3 (2019).

19 Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271-281 e277, doi:10.1016/j.cels.2018.03.002 (2018).

20 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

21 An, O. *et al.* NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database (Oxford)* **2014**, bau015, doi:10.1093/database/bau015 (2014).

22 Wang, S., Liu, Q., Zhu, E., Porikli, F. & Yin, J. Hyperparameter selection of one-class support vector machine by self-adaptive data shifting. *Pattern Recognition* **74**, 198-211, doi:10.1016/j.patcog.2017.09.012 (2018).

23 Weston, J. *et al.* in *Conference on Neural Information Processing Systems* (2001).

24 Webber, W., Moffat, A. & Zobel, J. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems* **28**, doi:10.1145/1852102.1852106. (2010).

25 Consortium, I. T. P.-C. A. o. W. G. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93, doi:10.1038/s41586-020-1969-6 (2020).

26 Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e318, doi:10.1016/j.cell.2018.02.060 (2018).

27 Morgan, M. A. & Shilatifard, A. Chromatin signatures of cancer. *Genes Dev* **29**, 238-249, doi:10.1101/gad.255182.114 (2015).

28 Helleday, T., Petermann, E., Lundin, C., Hodgson, B. & Sharma, R. A. DNA repair pathways as targets for cancer therapy. *Nat Rev Cancer* **8**, 193-204, doi:10.1038/nrc2342 (2008).

29 Otto, T. & Sicinski, P. Cell cycle proteins as promising targets in cancer therapy. *Nat Rev Cancer* **17**, 93-115, doi:10.1038/nrc.2016.138 (2017).

30 Sever, R. & Brugge, J. S. Signal transduction in cancer. *Cold Spring Harb Perspect Med* **5**, doi:10.1101/cshperspect.a006098 (2015).

31 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164, doi:10.1093/nar/gkq603 (2010).

32 Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**, 235-241, doi:10.1002/humu.22932 (2016).

33 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814, doi:10.1093/nar/gkg509 (2003).

34 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20, doi:10.1002/0471142905.hg0720s76 (2013).

35 Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575-576, doi:10.1038/nmeth0810-575 (2010).

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

37 Chun, S. & Fay, J. Identification of deleterious mutations within three human genomes. *Genome Res* **19**, doi:10.1101/gr.092619.109 (2009).

38 Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* **34**, 57-65, doi:10.1002/humu.22225 (2013).

39 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121, doi:10.1101/gr.097857.109 (2010).

40 Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025, doi:10.1371/journal.pcbi.1001025 (2010).

41 Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62, doi:10.1093/bioinformatics/btp190 (2009).

42 Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-16915, doi:10.1073/pnas.1009843107 (2010).

43 Nakatani, Y., Takeda, H., Kohara, Y. & Morishita, S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**, 1254-1265, doi:10.1101/gr.6316407 (2007).

44 Lenoir, W. F., Lim, T. L. & Hart, T. PICKLES: the database of pooled in-vitro CRISPR knockout library essentiality screens. *Nucleic Acids Res* **46**, D776-D780, doi:10.1093/nar/gkx993 (2018).

45 Chen, W. H., Lu, G., Chen, X., Zhao, X. M. & Bork, P. OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Res* **45**, D940-D944, doi:10.1093/nar/gkw1013 (2017).

46 Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat Genet* **45** (2013).

47 Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419, doi:10.1126/science.1260419 (2015).

48 Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Res* **45**, D369-D379, doi:10.1093/nar/gkw1102 (2017).

49 Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358-363, doi:10.1093/nar/gkt1115 (2014).

50 Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**, D449-451, doi:10.1093/nar/gkh086 (2004).

51 Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-772, doi:10.1093/nar/gkn892 (2009).

52 Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Res* **38**, D497-501, doi:10.1093/nar/gkp914 (2010).

53 Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46**, D649-D655, doi:10.1093/nar/gkx1132 (2018).

54 Chou, C. H. *et al.* miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res* **46**, D296-D302, doi:10.1093/nar/gkx1067 (2018).

55 Xiao, F. *et al.* miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* **37**, D105-110, doi:10.1093/nar/gkn851 (2009).

56 D'Antonio, M. & Ciccarelli, F. D. Modification of gene duplicability during the evolution of protein interaction network. *PLoS Comput Biol* **7**, e1002029, doi:10.1371/journal.pcbi.1002029 (2011).

57 Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* **44**, D286-293, doi:10.1093/nar/gkv1248 (2016).

58 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745, doi:10.1093/nar/gkv1189 (2016).

59 Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res* **47**, D351-D360, doi:10.1093/nar/gky1100 (2019).

60 The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

61 R: a language and environment for statistical computing (R foundation for statistical computing, Vienna, Austria, 2020).

62 e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071) (Vienna University of Technology, Vienna, Austria, 2019).

63 Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* **48**, D498-D503, doi:10.1093/nar/gkz1031 (2020).

64 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B* **57** (1995).

LIST OF SUPPLEMENTARY MATERIAL

Supplementary Note. Algorithm description

Supplementary Figure 1. Comparison of simulated and TCGA samples

Supplementary Figure 2. Parameter convergence and feature selection

Supplementary Figure 3. Patient-level comparison of driver detection methods

Supplementary Figure 4. Setting comparison for sysSVM2 training on TCGA data

Supplementary Table 1. Cohorts and genes used in the study

Supplementary Table 2. Features of genes in the simulated reference cohort

Supplementary Table 3. Application of sysSVM2 to TCGA samples

Supplementary Table 4. Driver predictions in 7,646 TCGA samples

Supplementary Table 5. Gene set enrichment analysis of TCGA predictions

ACKNOWLEDGMENTS

This work was supported by Cancer Research UK [C43634/A25487], the Cancer Research UK King's Health Partners Centre at King's College London [C604/A25135], the Cancer Research UK City of London Centre [C7893/A26233] and innovation programme under the Marie Skłodowska-Curie grant agreement No CONTRA-766030. JN is supported by the Doctoral Training Centre for Cross-Disciplinary Approaches to Non-Equilibrium Systems, funded by the EPSRC [EP/L015854/1]. The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. We thank Damjan Temelkovski for testing sysSVM2.

AUTHOR CONTRIBUTIONS

F.D.C. conceived and directed the study. J.N. performed data simulation and algorithm assessment and optimisation under the supervision of F.D.C. and C.Y. H.M. carried out the pan-cancer TCGA analysis with the help of J.N. F.D.C. and J.N. wrote the manuscript. All authors approved the manuscript.

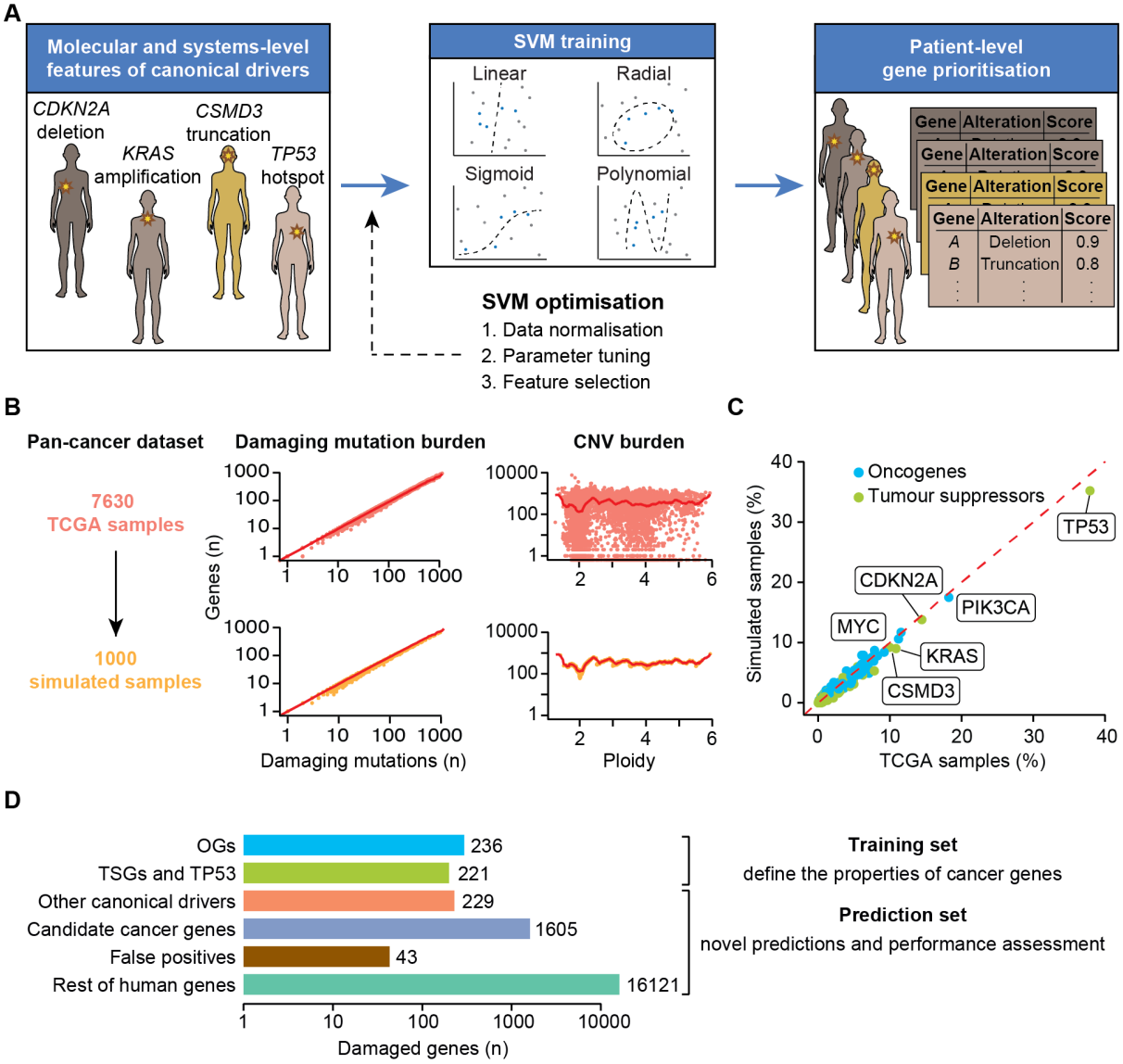
COMPETING INTERESTS

The authors declare no competing interests.

SOFTWARE AND DATA AVAILABILITY

Code to implement sysSVM2, along with a README file and an example dataset, is freely available at https://github.com/ciccalab/sysSVM2. The recommended settings as described in this manuscript are set as default values. However, users can modify many aspects of the implementation, including selection of features, data normalisation and kernel parameters. Models trained in pan-cancer and cancer-specific settings in 34 TCGA cancer types are also provided. This software code is protected by copyright. No permission is required from the rights-holder for non-commercial research uses. Commercial use will require a license from the rights-holder. For further information contact translation@crick.ac.uk.

Figure 1. sysSVM approach for driver prioritisation



A. Overview of sysSVM. Molecular (somatic SNVs, indels and mutation burden) and systems-level features (Supplementary Table 2) of damaged canonical drivers in the analysed samples are used for training. The best models of support vector machines (SVMs) with four kernels are selected using cross-validation and trained on the whole set of damaged canonical drivers. Finally, a combined score is used to prioritise driver genes in individual patients. The SVM implementation was generalised for optimal performance on a simulated cancer-agnostic dataset through data normalisation, parameter tuning and feature selection.

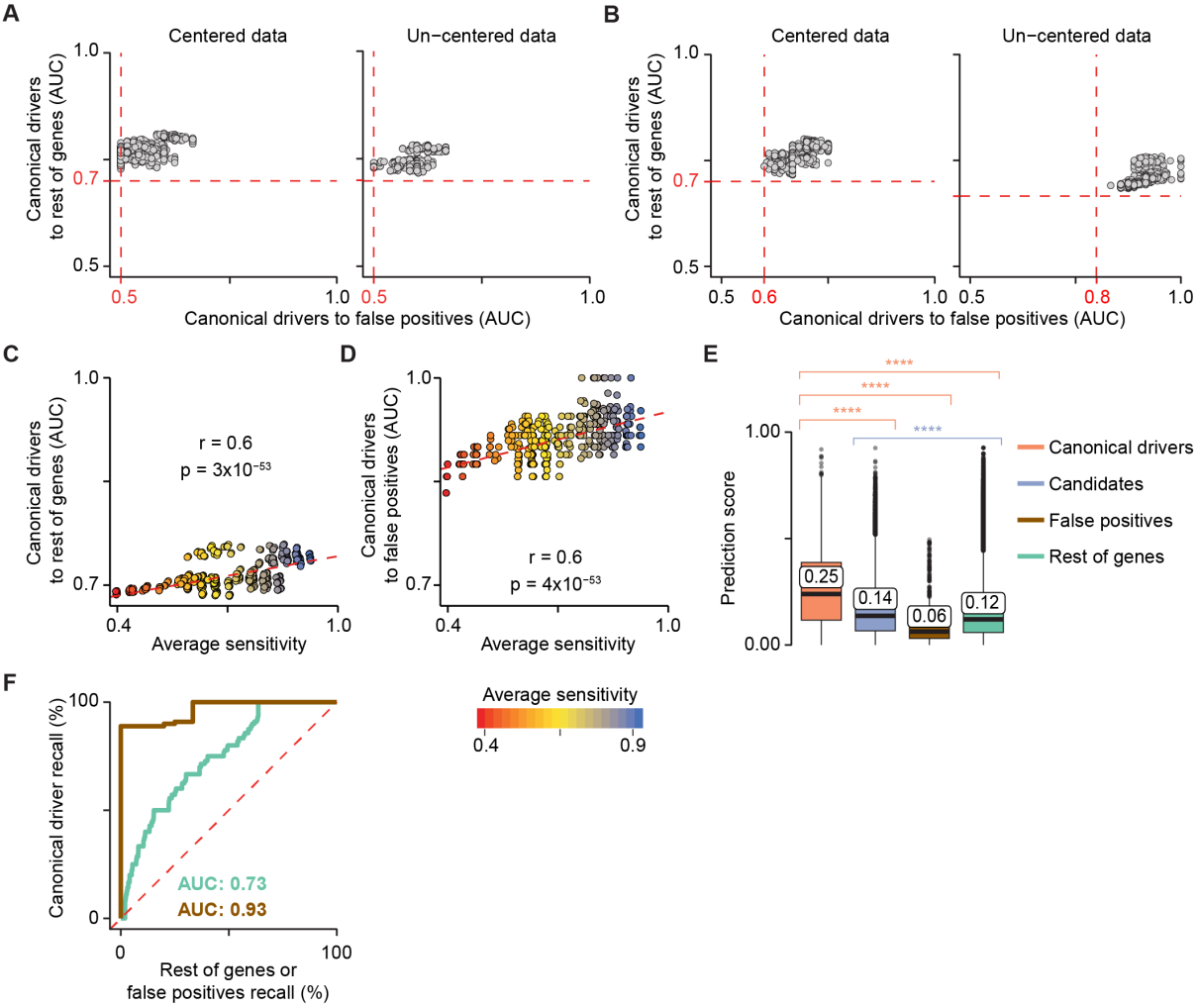
B. Generation of a simulated reference cohort from TCGA data. Values of damaging mutation burden and ploidy were randomly assigned to samples. Damaged genes were then extracted from real samples with similar values of damaging mutation

burden (+/-10% for mutations), and ploidy (+/-0.1 for CNVs). Dots represent individual TCGA (orange) or simulated (yellow) samples. Red lines indicate average numbers of genes with damaging mutations or CNVs in TCGA samples, for each given values of damaging mutation burden or ploidy.

C. Frequencies of canonical drivers in real and simulated samples. Oncogene gain-of-function, tumour suppressor loss-of-function and both types of *TP53* alterations were considered.

D. Gene sets used for sysSVM optimisation. The training set included oncogenes (OGs) and tumour suppressor genes (TSGs), as well as *TP53*. All other damaged genes were used for prediction and assessment. These included other canonical drivers (without a proven OG or TSG role), candidate cancer genes from published cancer sequencing screens, known false positives of established driver detection methods, and the remaining damaged genes. Bars indicate the number of unique damaged genes across the reference cohort of 1,000 simulated samples.

Figure 2. sysSVM optimisation on the simulated reference cohort



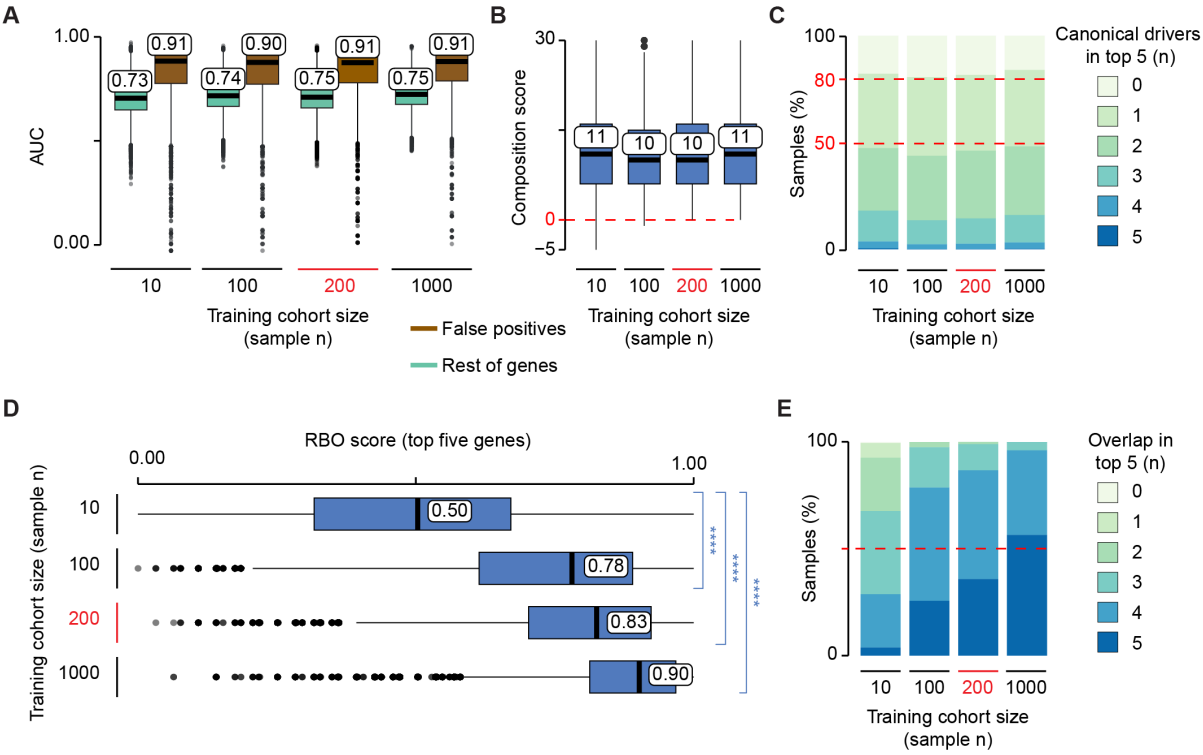
Model performances on the reference cohort using centered (left) and un-centered (right) data with all 25 systems-level features (**A**) or excluding protein length and number of protein domains (**B**). A sparse grid of 512 parameter combinations in the four kernels was tested. The performance of each model was measured using the Area Under the Curve (AUC), comparing the ranks of canonical drivers to the rest of genes and false positives. Median AUC values across all samples were plotted. Red dotted lines represent the minimum AUC values.

Correlation between model average sensitivity and AUCs of canonical drivers over the rest of genes (**C**) or false positives (**D**). The sensitivity of each kernel was measured on the training set over 100 three-fold cross-validation iterations. The median values over the four kernels are plotted. R and p-values from Pearson's correlation test are reported. Dotted red lines indicate the linear regression curves of best fit.

E. Distributions of sysSVM2 prediction scores for different types of damaged genes in the reference cohort. Whiskers extend to 1.5 times the Inter-Quartile Range (IQR). Statistical significance was measured using two-sided Wilcoxon tests. The median values of the distributions are labelled. **** = $p < 2.2 \times 10^{-16}$.

F. Receiver Operating Characteristic (ROC) curves, comparing canonical drivers to the rest of genes (green) and to false positives (brown). Recall rates were calculated for each sample separately, and the median ROC curve across samples was plotted. Median Areas Under the Curve (AUCs) for both comparisons are also indicated.

Figure 3. Effect of cohort size on sysSVM2 performance



A. Distributions of AUCs comparing the ranks of canonical drivers to the rest of genes (green) and False Positives (brown). Models were trained on ten simulated cohorts composed of ten, 100, 200 and 1,000, for a total of 40 simulated cohorts. These were then used to predict on the same reference cohort of 1,000 samples. Tthe AUC was measured for each set of predictions in each sample.

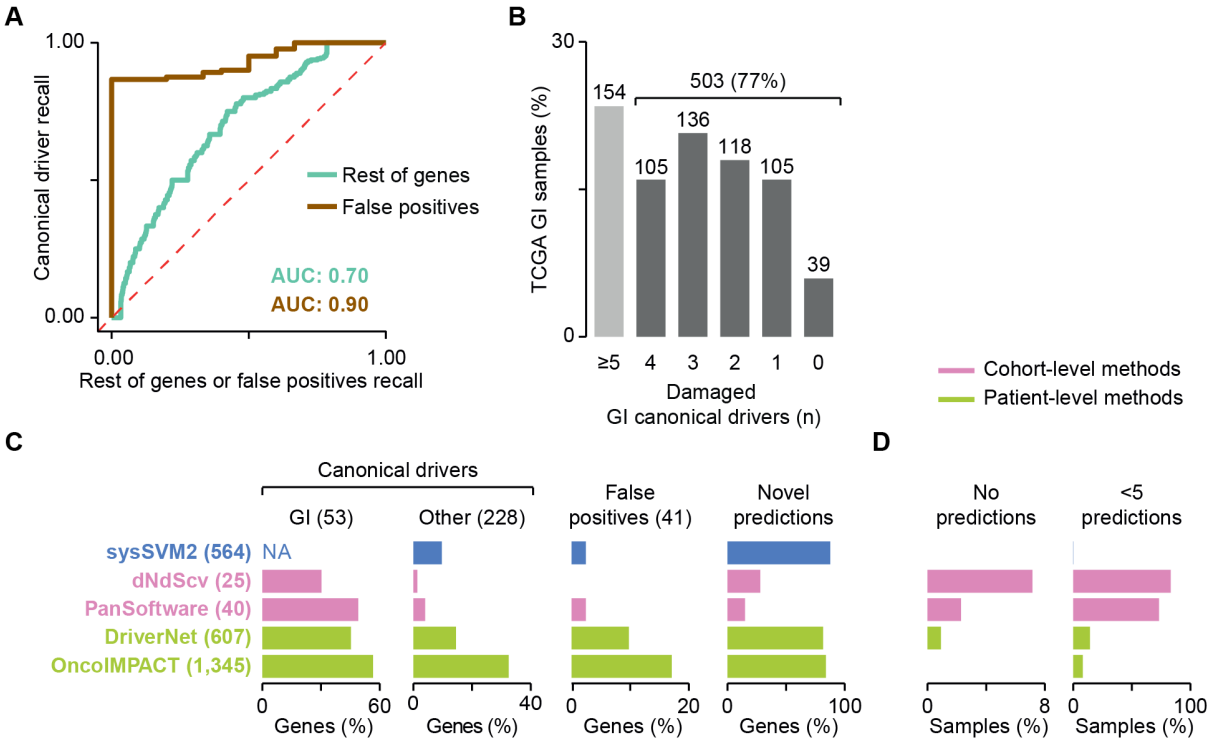
B. Distributions of composition scores of the top five predictions in terms of canonical drivers, candidate cancer genes, false positives and rest of genes (**Methods**). The composition score was measured for each set of predictions in each sample. Six training cohorts of size ten and two cohorts of size 100 gave negative composition scores in at least one sample, indicating highly ranked false positive genes.

C. Distribution of the number of canonical drivers in the top five predictions. The percentages were calculated out of the total of ten sets of predictions for each sample.

D. Rank-Biased Overlap (RBO) score of the top five predictions in each sample (**Methods**). RBO scores measured the similarity between the predictions from every possible pair of models trained on cohorts of a particular sample size. Statistical significance was measured using two-sided Wilcoxon tests. **** = $p < 2.2 \times 10^{-16}$.

E. Distribution of the number of top five predictions shared between models trained with the same cohort size. The overlap was calculated between each pair of predictions in each sample.

Figure 4. sysSVM2 benchmark on TCGA gastro-intestinal cancers



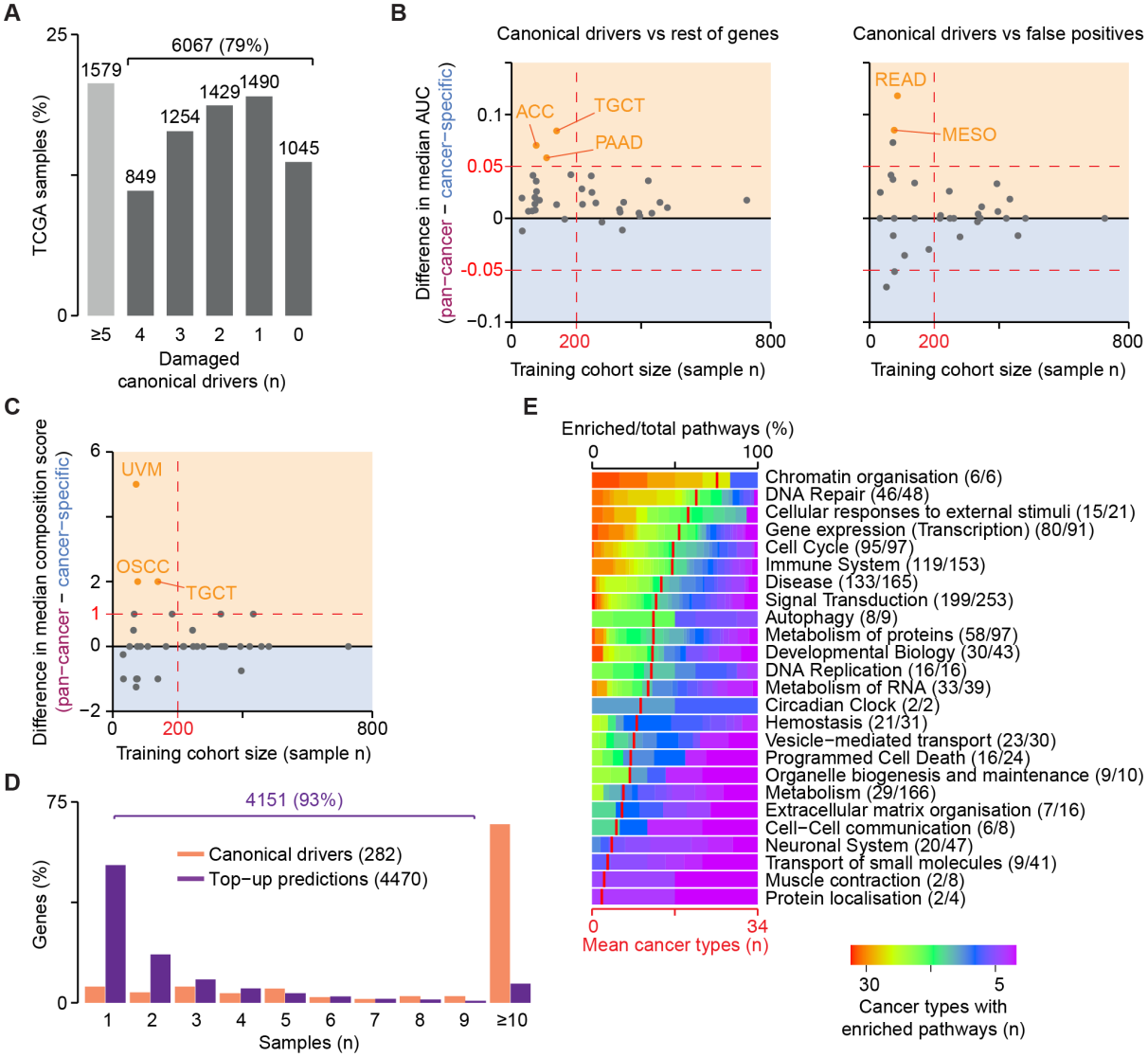
A. Median Receiver Operating Characteristic (ROC) curves across 657 Gastro-Intestinal (GI) samples from TCGA. Curves compare the ranks of canonical drivers to the rest of genes or to false positives. The median Areas Under the Curve (AUCs) are also indicated.

B. Distribution of GI canonical drivers across the GI cohort. Lists of canonical drivers for each GI cancer type were obtained from NCG6[11] and mapped to samples of the corresponding cancer type where they were damaged. Numbers of samples are indicated above each bar. Samples with five or more GI drivers did not require additional driver predictions.

C. Comparison of performance between sysSVM2 and four other driver detection methods. The set of unique drivers predicted by each approach were compared in terms of recall of GI canonical drivers, other canonical drivers (non-GI and outside the sysSVM2 training set) and false positives and proportion of novel predictions not previously associated with a cancer driver role. The number of genes in each category is reported in brackets. The recall of GI canonical drivers could not be assessed for sysSVM2 because these were part of the training set. They were however considered as drivers by default, rather than predicted by the algorithm. NA = Not Applicable.

D. Proportions of 657 GI samples left with no predicted drivers (left) or fewer than 5 predictions. The one sample left with fewer than 5 predictions by sysSVM2 (TCGA-FP-8210, stomach cancer) had four damaged genes overall.

Figure 5. sysSVM2 predictions in 34 cancer types



A. Number of damaged canonical drivers per sample. Lists of canonical drivers for each cancer type were obtained from NCG[11] and mapped to samples of the corresponding cancer type. 6,067 samples with less than five canonical drivers damaged underwent the top-up procedure to reach five drivers.

Difference in Areas Under the Curve (AUCs) between the pan-cancer and cancer-specific settings in ranking canonical drivers over the rest of human genes and false positives (**B**), and in the composition score of the top five predictions (**C**). The median values of the distributions in each cancer type were used for comparison, with the yellow and blue regions indicating better performance in the pan-cancer and cancer-specific settings, respectively. The number of samples used for training is indicated on the x-axis. Colour dots represent cancer types where the two settings differ both significantly (FDR <0.05, Wilcoxon rank-sum test) and substantially

(|difference in medians| >0.05 for AUCs, >1 for composition score). ACC, adrenocortical carcinoma; TGCT, testicular germ cell tumours; PAAD, pancreatic adenocarcinoma; READ, rectum adenocarcinoma; MESO, mesothelioma; UVM, uveal melanoma; and OSCC, oesophageal squamous cell carcinoma.

D. Recurrence of damaging alterations in 282 canonical driver genes and 4,470 sysSVM2 top-up predictions across 7,646 samples.

E. Gene set enrichment analysis of sysSVM2 top-up genes, grouped in broad biological processes (Reactome level 1). Numbers of pathways enriched in at least one cancer type out of the total pathways tested are reported in brackets. Red vertical strokes indicate the mean number of cancer types that pathways from each broad process are enriched in (bottom x-axis).