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

Joel Nulsen¹², Hrvoje Misetic¹², Christopher Yau³⁴ and Francesca D. Ciccarelli¹²*¹

Abstract

Background: Identifying the complete repertoire of genes that drive cancer in individual patients is crucial for precision oncology. Most established methods identify driver genes that are recurrently altered across patient cohorts. However, mapping these genes back to patients leaves a sizeable fraction with few or no drivers, hindering our understanding of cancer mechanisms and limiting the choice of therapeutic interventions.

Results: We present sysSVM2, a machine learning software that integrates cancer genetic alterations with gene systems-level properties to predict drivers in individual patients. Using simulated pan-cancer data, we optimise sysSVM2 for application to any cancer type. We benchmark its performance on real cancer data and validate its applicability to a rare cancer type with few known driver genes. We show that drivers predicted by sysSVM2 have a low false-positive rate, are stable and disrupt well-known cancer-related pathways.

Conclusions: sysSVM2 can be used to identify driver alterations in patients lacking sufficient canonical drivers or belonging to rare cancer types for which assembling a large enough cohort is challenging, furthering the goals of precision oncology. As resources for the community, we provide the code to implement sysSVM2 and the pre-trained models in all TCGA cancer types (https://github.com/ciccalab/sysSVM2).

Keywords: Cancer genomics, Cancer driver genes, Systems-level properties, Patient-level driver detection

Background

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 hotspot positions 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 2000 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 OncoMPACT [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, especially in clinical settings where shotgun transcriptomic sequencing is still rare. Alternative approaches such as PHIAL [16] match the patient mutations with databases of known clinically actionable or driver alterations but have 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 coverage. We also develop optimal models for identifying driver genes in all 34 cancer types available in The Cancer Genome Atlas (TCGA) [19] and validate them in osteosarcoma, a rare cancer type that was not part of TCGA. The software, 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.

Implementation
The sysSVM approach to driver detection prioritises genes with features similar to those of canonical cancer drivers, i.e. genes whose modifications have experimentally proven roles in cancer initiation and progression (Additional file 1: Supplementary Note). Canonical drivers differ from other human genes by an array of systems-level properties that define them as a group and do not strictly depend on the function of the single gene. These properties include gene duplicability in the human genome [20] and through vertebrate whole-genome duplications [21], gene essentiality across cell lines [11], breadth of expression in healthy tissues at the gene and protein levels [11, 22, 23], protein connectivity and global topology in the protein-protein interaction network [20], participation in protein complexes [22], number of targeting miRNAs [21], gene evolutionary origin [21] and protein length and domain organisation [22, 23] (Additional file 2: Table S1). Canonical drivers can also be described using molecular properties that reflect the somatic alterations that they acquire in cancer. These include alterations with predicted damaging effects on protein function (copy number gains and losses as well as truncating, non-truncating damaging and hotspot mutations) and overall mutational burden and copy number of the gene (Additional file 2: Table S1).

To leverage the systems-level and molecular properties of canonical drivers, sysSVM first identifies a set of true positive canonical drivers damaged within a cohort of patients (Fig. 1a). It then uses the features of this positive set to train one-class SVMs based on four kernels (linear, radial, sigmoid, polynomial). Finally, it ranks the remaining damaged genes in individual cancer patients with a combined score that weights the kernels based on their sensitivity (Additional file 1: Supplementary Note). Highly ranked genes have the most similar properties to those of canonical drivers and will be then considered the cancer drivers for that patient. We use one-class SVMs for sysSVM because, while canonical drivers represent a reliable set of true positives, identifying a true negative set of non-cancer genes is not possible. For example, possible negative genes could be known false positives of driver gene detection methods [1, 22]. However, these genes are representative of false positives rather than true negatives, so training a classifier on them is likely to introduce unwanted bias. A one-class support vector machine for novelty detection is therefore an optimal way to solve this issue.

Results
Simulation of pan-cancer datasets
In order to optimise the use of sysSVM for any cancer type, we simulated 1000 cancer-agnostic samples starting from all TCGA tumours with matched mutation, CNV
and gene expression data (Additional file 1: Supplementary Methods). We ensured that the tumour mutation and copy number burdens were similar between real and simulated samples (Fig. 1b) and that gene mutation and copy number status in the simulated dataset was the same of TCGA (Additional file 1: Figure S1A). 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 (Fig. 1c). We further verified that gene alteration frequencies in the simulated data were not significantly biased by cancer types with large cohort sizes in TCGA (Additional file 1: Figure S1B), confirming the suitability of the simulated data as a representative pan-cancer cohort.

The simulated cohort for sysSVM optimisation (hereafter referred to as the reference cohort) was composed of 1000 samples with 18,455 genes damaged 309,427 times. Of these, 686 were canonical drivers with an experimentally proven role in cancer [12, 24], 1605 were candidate cancer genes from 273 cancer screens [11], 43 were known false positive predictions of driver detection methods [1, 25] and 16,121 were the remaining damaged genes (hereafter referred to as the rest of genes; Fig. 1d, Additional file 2: Table S2). We annotated seven molecular and 25 systems-level features of all damaged genes (Additional file 2: Table S1) and used these features for training and prediction. As a training set, we selected 457 of the 686 canonical drivers with proven roles as oncogenes (236) or tumour suppressors (221). We restricted somatic alterations of oncogenes and tumour suppressors to gain-of-function or loss-of-function alterations, respectively (Additional file 1: Supplementary Methods). Since we could not reliably define the remaining 229 damaged canonical drivers as either oncogenes or tumour suppressors, we could not restrict their somatic alterations to the appropriate type. Therefore, we did not use them for training but could still use them for prediction and performance assessment (Fig. 1d), together with 43 false positives and 16,121 the rest of genes.

**sysSVM optimisation on the pan-cancer reference cohort**

Using the reference cohort, we optimised sysSVM in terms of data normalisation, parameter tuning and feature selection (Fig. 1a). So as 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 (Additional file 1: Supplementary Note). We then measured the ability of each of these 512 models to prioritise the 229 canonical drivers not used for training over the rest of damaged genes or the false positives. We did this by computing the area under the curve (AUC) in each sample and taking the median AUC as representative of the whole cohort (Additional file 1: Supplementary Methods).

First, we derived the optimal settings for data normalisation in terms of centred and un-centred data (Additional file 1: Supplementary Note). All models robustly prioritised canonical drivers above the rest using either centred or un-centred data but showed lower performance in distinguishing canonical drivers from false positives (Fig. 2a). We reasoned that false positives from recurrence-based driver detection methods [1] shared some features with canonical drivers. For example, they encoded long and multi-domain proteins. When removing protein length and number of domains from the feature list (Additional file 2: Table S2), the performance substantially improved particularly for un-centred data (Fig. 2b). We therefore removed protein length and number of domains from the model.

Second, we selected the optimal sets of parameters in each kernel. Hyper-parameter choice is known to have substantial impacts on classification and it is an open problem for one-class SVMs [26]. Since the parameters for each kernel needed to be selected separately (Additional file 1: Supplementary Note), we could not use AUC of the combined multi-kernel model for assessment. Instead, we used the sensitivity of each kernel to predict canonical drivers calculated from three-fold cross-validation on the training set. Sensitivity was indeed a good predictor of the overall AUC of canonical drivers over the rest of genes (Fig. 2c) and false positives (Fig. 2d). We therefore developed an approach to select the parameters that conferred the highest sensitivity in multiple iterations of cross-validation (Additional file 1: Supplementary Methods). In the reference cohort, parameters chosen in this way converged within 2000 cross-validation iterations for all kernels (Additional file 1: Figure S3A).

Finally, since the presence of highly correlated features can hinder SVM performance [27], we performed systematic feature selection by assessing the pairwise correlations between all 25 systems-level features. Four features (gene expression in 1 ≤ tissues ≤ 6 and in ≥ 37 tissues; protein expression in 0 ≤ tissues ≤ 8 and central position in the protein-protein interaction network) exhibited a significant degree of inter-correlation (Pearson |r| > 0.5, FDR < 0.05, Additional file 1: Figure S3B). Removing them led to faster convergence of kernel parameters (Additional file 1: Figure S3A) and improved performance overall (Additional file 1: Figure S3C).

Based on these results, we chose the default settings for the cancer agnostic SVM classifier, which we named sysSVM2 [28]. By default, data are un-centred 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 (Table 1). Finally, kernel parameters optimised on the reference cohort are provided as a default (Additional file 1: Figure S3A), although users
may perform specific cross-validation iterations on their own cohorts.

We then assessed the performance of sysSVM2 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 (Fig. 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 (Fig. 2f), demonstrating that canonical drivers were prioritised above the rest of genes and especially above false positives. This was not surprising as the properties of canonical drivers differ substantially from those of false positives.

**Fig. 2** sysSVM optimisation on the simulated reference cohort. Model performances on the reference cohort using centred (left) and un-centred (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 interquartile range (IQR). Statistical significance was measured using two-sided Wilcoxon tests. The median values of the distributions are labelled. ****p < 2.2 × 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.
(Additional file 1: Figure S3D), further supporting that known false positives are not representative of non-cancer genes.

**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, Additional file 2: Table S3), 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 1000 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 (Fig. 3a). This suggested that sysSVM2 was overall very effective in prioritising cancer genes independently of the training cohort size. We then compared the composition of the 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 (Additional file 1: Supplementary Methods). Similar to the AUC, the composition score of the top five genes was also very similar across training cohorts (Fig. 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

---

**Table 1** Twenty-six features derived from molecular and systems-level properties of genes and used to predict cancer drivers in sysSVM2. Molecular properties describe gene alterations in individual cancer samples. Systems-level properties are global gene properties (see also Additional file 2: Table S1)

| Category                        | Property                   | Feature                          | Type   |
|---------------------------------|----------------------------|----------------------------------|--------|
| **Molecular**                   | Gene mutation              | Mutational load (n)              | Continuous |
|                                 |                            | Non-truncating damaging mutations (n) | Continuous |
|                                 |                            | Truncating mutations (n)         | Continuous |
|                                 |                            | Hotspot mutations (n)            | Continuous |
|                                 | Gene copy number           | Gene copy number (n)             | Continuous |
|                                 |                            | Gene is amplified                | Binary   |
|                                 |                            | Gene is deleted                  | Binary   |
| **Systems-level**               | Gene duplication           | Gene is duplicated               | Binary   |
|                                 |                            | Gene is an ohnolog               | Binary   |
|                                 | Gene essentiality          | Cell lines in which gene is essential (%) | Continuous |
|                                 |                            | Gene is essential                | Binary   |
|                                 | Gene expression            | Tissues expressing gene (n)      | Continuous |
|                                 |                            | Gene is expressed in 0 tissues    | Binary   |
|                                 |                            | Gene is expressed in 7 ≤ tissues ≤ 36 | Binary   |
|                                 | Protein expression         | Tissues expressing protein (n)   | Continuous |
|                                 |                            | Protein is expressed in ≥ 41 tissues | Binary   |
| **Protein-protein interaction network (PPIN)** | | PPIN degree                     | Continuous |
|                                 |                            | Protein is a PPIN hub            | Binary   |
|                                 |                            | PPIN betweenness                 | Continuous |
|                                 |                            | PPIN clustering coefficient       | Continuous |
| **Protein complexes**           | Complexes the protein is part of (n) | Continuous |
| **miRNA interactions**          | miRNAs targeting the gene (n) | Continuous |
| **Gene evolutionary origin**    | Pre-metazoan origin         | Binary                           |
|                                 | Metazoan origin            | Binary                           |
|                                 | Vertebrate origin           | Binary                           |
|                                 | Post-vertebrate origin      | Binary                           |

PPIN protein-protein interaction network, miRNA micro RNA
Finally, we measured the ratio between observed and expected canonical drivers and false positives in the top five genes (Fig. 3c, Additional file 1: Supplementary Methods). Independently of the training cohort size, false positives in the top five genes were always lower than expected, confirming that sysSVM2 efficiently distinguished false positives from drivers. The number canonical drivers in the top five genes was more than twice the expected number in > 85% of samples and more than five times the expected value in around 65% of samples. As with the other metrics, the performance of sysSVM2 did not change substantially with the size of the training cohort.

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 [29] (Additional file 1: Supplementary Methods). The distributions of RBO scores of the top five genes were significantly higher for large training cohorts compared to those composed of ten samples (Fig. 3d). Moreover, models trained on large cohorts showed overall higher gene overlap in the top five genes (Fig. 3e).

These results showed that, although sysSVM2 successfully separates canonical drivers from other genes independent 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 gastrointestinal (GI) adenocarcinomas from TCGA (73 oesophageal, 279 stomach, 219 colon and 86 rectal cancers, Additional file 2: Table S3). Overall, this cohort had 17,122 unique damaged genes, including 438 tumour suppressors and oncogenes used for sysSVM2 training (Additional file 2: Table S2). 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 (Fig. 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, we added sysSVM2 predictions progressively based on their rank to reach five drivers per patient (Additional file 1: Supplementary 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, 30]. While 154 patients had damaging alterations in five or more GI canonical drivers, 503 patients (77%) needed at least one prediction (Fig. 4b), highlighting the need for additional cancer driver predictions. This resulted in 564 unique sysSVM2 drivers.

We then predicted the drivers in the same GI samples using two cohort-level (PanSoftware [31] and dNdScv...
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 onco-genes damaged across the whole cohort. In the cancer-specific setting, we used instead only the subsets of these genes damaged in each cancer type (Additional file 2: Table S3). 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 (Additional file 2: Table S3). We found that 6067 samples (79%) required at least one sysSVM2 prediction in order to reach five drivers (Fig. 5a). These corresponded to 4369 and 4548 unique genes in the pan-cancer and cancer-specific settings, respectively, with a significant overlap of predictions (3896, p < 2.2 × 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 (Fig. 5b, Additional file 1: Figure S5A and S5B). 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 (Fig. 5c, Additional file 1: Figure S5C). All these cancer types were again characterised by small training cohorts and showed higher performance in the pan-cancer setting. Predictions of cancer-specific models and the pan-cancer model were mostly similar, with the exception of cancer types with small training cohorts (Additional file 1: Figure S5D and S5E). 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 final list of patient-specific predictions in 34 cancer types was composed of 4470 unique genes, the vast majority of which (93%) were rare (<10 patients) or patient-specific (Fig. 5d, Additional file 2: Table S4). A gene set enrichment analysis on these genes revealed 984 enriched pathways overall (Reactome level 2 or above, FDR < 0.01, Additional file 1: Supplementary Methods, Additional file 2: Table S5). Interestingly, when mapping these pathways to broader biological processes (Reactome level 1), a few processes were widely enriched in almost all cancer types (Fig. 5e).

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

In order to provide a comprehensive resource of trained models [28] and patient-level drivers, we sought to apply sysSVM2 to 7646 TCGA samples of 34 cancer types with at least one somatically damaged gene (Additional file 1: Supplementary 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 onco-genes damaged across the whole cohort. In the cancer-specific setting, we used instead only the subsets of these genes damaged in each cancer type (Additional file 2: Table S3). 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 (Additional file 2: Table S3). We found that 6067 samples (79%) required at least one sysSVM2 prediction in order to reach five drivers (Fig. 5a). These corresponded to 4369 and 4548 unique genes in the pan-cancer and cancer-specific settings, respectively, with a significant overlap of predictions (3896, p < 2.2 × 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 (Fig. 5b, Additional file 1: Figure S5A and S5B). 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 (Fig. 5c, Additional file 1: Figure S5C). All these cancer types were again characterised by small training cohorts and showed higher performance in the pan-cancer setting. Predictions of cancer-specific models and the pan-cancer model were mostly similar, with the exception of cancer types with small training cohorts (Additional file 1: Figure S5D and S5E). 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 final list of patient-specific predictions in 34 cancer types was composed of 4470 unique genes, the vast majority of which (93%) were rare (<10 patients) or patient-specific (Fig. 5d, Additional file 2: Table S4). A gene set enrichment analysis on these genes revealed 984 enriched pathways overall (Reactome level 2 or above, FDR < 0.01, Additional file 1: Supplementary Methods, Additional file 2: Table S5). Interestingly, when mapping these pathways to broader biological processes (Reactome level 1), a few processes were widely enriched in almost all cancer types (Fig. 5e).
such as chromatin organisation [32], DNA repair [33], cell cycle [34] and signal transduction [35]. Therefore, although not recurring across patients, sysSVM2 predictions converged to perturb similar biological processes that are known to contribute to cancer.

**sysSVM2 predictions in an independent cancer cohort**

We finally sought to assess whether the sysSVM2 models trained on TCGA could be applied for driver prediction in a cancer type not included in TCGA. We therefore analysed 36 osteosarcomas from the Pan-
Cancer Analysis of Whole Genomes (PCAWG) consortium [30]. Osteosarcoma is a rare, genetically heterogeneous bone cancer with poor prognosis and only six well-established canonical drivers [36, 37].

We annotated the genomic data of the PCAWG cohort finding 4969 damaged genes overall with a median of 93 damaged genes per sample (Additional file 2: Table S2). Only two of these samples had three damaged osteosarcoma canonical drivers while 19 (53%) of them had no canonical driver (Fig. 6a), highlighting the need for further predictions. Given the small cohort size, we used the TCGA pan-cancer setting to rank the damaged genes in each osteosarcoma. Considering the top five predictions per sample, we got 129 unique genes (Additional file 2: Table S6), which were poorly recurrent across samples (Fig. 6b), reflecting again the genetic heterogeneity of osteosarcoma.

At the cohort level, sysSVM2 predictions included five of the six (83%) osteosarcoma canonical drivers [36, 37]. At the patient level, the six osteosarcoma canonical drivers were damaged 27 times and in 14 of these cases (53%) they were in the top five predictions (Fig. 6c). This proportion rose to 81% when considering the top ten predictions. In addition to osteosarcoma canonical drivers, 26 sysSVM2 predictions were canonical drivers in other cancer types, 16 were candidate cancer driver genes and 81 had no previously known involvement in cancer (Additional file 2: Table S6). Despite this, these 81 genes were enriched in eight pathways (FDR < 0.1), most of which have a known role in cancer (Fig. 6d).

Moreover, they included genes known to promote osteogenesis such as YAP1 and YES1 [38, 39].

These results showed that sysSVM2 is able to identify reliable cancer drivers in individual patients even for cancer types not used for training. This has relevant implications particularly in the case of rare cancers that are poorly studied and have little genomic data available.

**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. Our method allows to identify driver genes in individual patients. 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 [30]. Given such a training set of true positives, sysSVM2 could be further developed to

![Fig. 6 Validation of sysSVM2 in osteosarcoma.](image-url)
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 studies to integrate multiple tools for driver detection [31], 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.

**Conclusions**

In this work, we developed a cancer-agnostic algorithm, sysSVM2, for identifying cancer driver in cancer individual patients [28]. 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. sysSVM2 can be used to identify driver alterations in individual patients and rare cancer types where canonical drivers are insufficient to explain the onset of disease, as we have validated in osteosarcoma. This potentially opens up further research and therapeutic opportunities.

**Availability and requirements**

Project name: sysSVM2

Project home page: https://github.com/ciccalab/sysSVM2

Operating system: Platform independent

Programming language: R

Other requirements: R version greater than 3.5

Licence: Crick Non-commercial Licence Agreement v2.0

Any restrictions to use by non-academics: Commercial use will require a licence from the rights-holder. For further information, contact translation@crick.ac.uk.

---

**Abbreviations**

SNV: Single-nucleotide polymorphism; Indel: Insertion or deletion; CNV: Copy number variant; SVM: Support vector machine; TCGA: The Cancer Genome Atlas; AUC: Area under the curve; ROC: Receiver operating characteristic; DLBC: Diffuse large B cell lymphoma; BRCA: Breast cancer; RBO: Rank-biased overlap; GI: Gastro-intestinal; FDR: False discovery rate; PPIN: Protein-protein interaction network

**Acknowledgements**

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.

**Funding**

This work was supported by Cancer Research UK [C43634/A25487], the Cancer Research UK King’s Health Partners Centre at King’s College London [C606/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].

**Availability of data and materials**

Platform-independent R code to implement sysSVM2, along with a README file and an example dataset, is 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 licence from the rights-holder. For further information, contact translation@crick.ac.uk. All data supporting this study are included in the paper.

Original data for annotating systems-level properties were obtained from the following publicly available sources:

- BioGRID: https://thebiogrid.org/
- CORUM: http://mips.helmholtz-muenchen.de/corum/
- DIP: [42]: http://dip.doe-mbi.ucla.edu/dip/Main.cgi
- EggNOG: [43]: http://eggnogdb.embl.de/#/app/home
- GTEx: [44]: https://www.gtexportal.org/home
- HPRD: [45]: http://www.hprd.org/
- MIntAct: [46]: https://www.ebi.ac.uk/intact/
- miRecords: [47]: http://c1.accurascience.com/miRecords/
- miRBase: [48]: http://mirbase.mbc.nctu.edu.tw/php/index.php

---

**Supplementary Information**

The online version contains supplementary material available at https://doi.org/10.1186/s13073-021-00830-0.

---

**Additional file 1: Supplementary Note, Supplementary Methods, Supplementary Figures**

sysSVM2 rationale and algorithm description. Algorithm implementation and assessment. Figure S1. Comparison of simulated and TCGA samples. Figure S2. Selection of binary features derived for PPIN and tissue expression properties. Figure S3. Parameter convergence and feature selection. Figure S4. Patient-level comparison of driver detection methods. Figure S5. Setting comparison for sysSVM2 training on TCGA data.

**Additional file 2: Table S1.** Features of genes used in sysSVM2. Table S2. Cohorts and genes used in the study. Table S3. Application of sysSVM2 to TCGA samples. Table S4. Driver predictions in 7646 TCGA samples. Table S5. Gene set enrichment analysis of TCGA predictions. Table S6. sysSVM2 driver predictions in PCAWG osteosarcomas.
OGEE [49]: http://ogee.medgenius.info/browse/
PICKLES [50]: https://hartlab.shinyapps.io/pickles/
Protein Atlas [51]: https://www.proteinnatlas.org/
Reactome [52]: https://reactome.org/
RefSeq [53]: https://www.ncbi.nlm.nih.gov/refseq/

Ethics approval and consent to participate
The need for Institutional Review Board Approval at our institutions (King’s College London and The Francis Crick Institute) was waived for this study as all data had previously been generated as part of The Cancer Genome Atlas and the Pan-Cancer Analysis of Whole Genomes projects. None of the results reported in this manuscript can be used to identify individual patients. This study was conducted in accordance with the Helsinki Declaration.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
1Cancer Systems Biology Laboratory, The Francis Crick Institute, London NW1 1AT, UK. School of Cancer and Pharmaceutical Sciences, King’s College London, London SE1 1UL, UK. 2School of Health Sciences, University of Manchester, Manchester M13 9PL, UK. 4The Alan Turing Institute, London NW1 2DB, UK.

Received: 7 September 2020 Accepted: 8 January 2021
Published online: 01 February 2021

References
1. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8.
2. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSC: identifying mutational significance in cancer genomes. Genome Res. 2012;22(8):1589–98.
3. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12(4):R41.
4. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013;29(18):2238–44.
5. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol. 2013;9:637.
6. Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell. 2013;153(4):948–62.
7. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci U S A. 2016;113(50):14330–5.
8. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haese K, Van Loo P, et al. Universal patterns of selection in cancer and somatic tissues. Cell. 2017;171(5):1029–41.e21.
9. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 2012;40(21):e169.
10. Leiserson MD, Vandin F, Wu HT, Dobson JR, Eldridge JV, Thomas JL, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet. 2018;173(2):171–85.
11. Schuster-Böckler B, Sturmer N, Fröhlich H, Lohse M, Haubold B, Kuster B, et al. SCOUT: efficient recognition of driver mutations. Nucleic Acids Res. 2012;40(18):e148.
12. Schuster-Böckler B, Sturmer N, Fröhlich H, Lohse M, Haubold B, Kuster B, et al. SCOUT: efficient recognition of driver mutations. Nucleic Acids Res. 2012;40(18):e148.
13. Schuster-Böckler B, Sturmer N, Fröhlich H, Lohse M, Haubold B, Kuster B, et al. SCOUT: efficient recognition of driver mutations. Nucleic Acids Res. 2012;40(18):e148.
14. Bertrand D, Chng KR, Shergar SF, Kiesel A, Chia BK, Sai Y, et al. Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. Nucleic Acids Res. 2015;43(7):e44.
15. Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol. 2012;13(12):R24.
16. Hou J, Ma J. DawnRank: discovering personalized driver genes in cancer. Genome Med. 2014;6:Article number: 56.
17. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marafon S, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med. 2014;20(6):682–8.
18. Dong C, Guo Y, Yang H, He Z, Liu X, Wang K. ICAGES: integrated Cancer Genome Score for comprehensively prioritizing driver genes in personal cancer genomes. Genome Med. 2016;8(1):135.
19. Mullikin TP, Benedetti L, Foxell E, Temelkovski D, Nulsen J, Perner J, et al. Patient-specific cancer genes contribute to recurrently perturbed pathways and establish therapeutic vulnerabilities in esophageal adenocarcinoma. Nat Commun. 2019;10(1):3101.
20. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. 2018;6(3):271–81.e7.
21. Rambaldi D, Giorgi FM, Capuani F, Ciliberto A, Criccarelli FD. Low duplicability and network fragility of cancer genes. Trends Genet. 2008;24(9):427–30.
22. D’Antonio M, Criccarelli FD. Modification of gene duplicability during the evolution of protein interaction network. PLoS Comput Biol. 2011;7(4):e1002029.
23. An O, Dall’Olio GM, Mourikis TP, Criccarelli FD. NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. Nucleic Acids Res. 2016;44(O1):D992–9.
24. D’Antonio M, Criccarelli FD. Integrated analysis of recurrent properties of cancer genes to identify novel drivers. Genome Biol. 2013;14(5):R52.
25. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546–58.
26. An O, Pendino V, D’Antonio M, Ratti E, Gentilini M, Criccarelli FD. NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. Database (Oxford). 2014;2014b0115.
27. Wang S, Liu Q, Zhu E, Ponkli F, Yin J. Hyperparameter selection of one-class support vector machine by self-adaptive data shifting. Pattern Recogn. 2018;74:198–211.
28. Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. Feature selection for SVMs. Conference on Neural Information Processing Systems; 2001.
29. Nulsen J, Mistic H, You C, Criccarelli FD. sysSMV software. Criccarelli lab 2020. Available from: https://github.com/ciccarelli/sysSMV2. Accessed December 2020.
30. Webber W, Moffat A, zobel J. A similarity measure for indefinite rankings. ACM Trans Inf Syst. 2010;28(4). Article number: 20.
31. Consortium ITP-CAoWG. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):68–93.
32. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173(2):371–85.e18.
33. Morgan MA, Shlafbard A. Chromatin signatures of cancer. Genes Dev. 2015;29(3):238–49.
34. Helladay T, Petermann E, Lundin C, Hodgson B, Sharma RA. DNA repair pathways as targets for cancer therapy. Nat Rev Cancer. 2008;8(3):193–204.
35. Otto T, Sionkowska P. Cell cycle proteins as promising targets in cancer therapy. Nat Rev Cancer. 2017;17(2):93–115.
36. Sever R, Brugge JS. Signal transduction in cancer. Cold Spring Harb Perspect Med. 2015;5(4). Article number: a006098.
37. Chen X, Bahrami A, Pappo A, Easton J, Dalton J, Hedlund E, et al. Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. Cell Rep. 2014;7(1):104–12.
38. Kovac M, Blattmann C, Ribi S, Smida J, Mueller NS, Engert F, et al. Exome sequencing of osteosarcoma reveals mutation signatures reminiscent of BRCA deficiency. Nat Commun. 2015;6:8940.
39. Segelman CD, Mason DE, Dawahare JH, Horan DJ, Vigil GD, Howard SS, et al. Skeletal cell YAP and TAZ combinatorially promote bone development. FASEB J. 2018;32(5):2706–21.
39. Pan JX, Xiong L, Zhao K, Zeng P, Wang B, Tang FL, et al. YAP promotes osteogenesis and suppresses adipogenic differentiation by regulating beta-catenin signaling. Bone Res. 2018;6:18.

40. Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolias NK, et al. Biological general repository for interaction datasets (BioGRID). FAIRsharing. https://doi.org/10.25504/fairsharing.9d5f5r.

41. Ruepp A, Waegeler B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. The comprehensive resource of mammalian protein complexes (CORUM). FAIRsharing. https://doi.org/10.25504/fairsharing.ohbpnw.

42. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins (DIP). FAIRsharing. https://doi.org/10.25504/fairsharing.qje0v8.

43. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. Evolutionary genealogy of genes: non-supervised orthologous groups (EggNOG). FAIRsharing. https://doi.org/10.25504/fairsharing.jtwy7d.

44. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. Genotype-Tissue Expression (GTEx) project. FAIRsharing. bsg-d001206. 2018.

45. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database (HPRD). FAIRsharing. https://doi.org/10.25504/fairsharing.y2qws7.

46. Orchard S, Ammari M, Aranda B, Breuza L, Broackes-Carter F, et al. The molecular interaction (MIntAct) database. FAIRsharing. https://doi.org/10.25504/fairsharing.d05nwx.

47. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords. AccuraScience URL: http://c1.accurascience.com/miRecords/. Accessed February 2018.

48. Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, et al. The microRNA-target interaction database (miRTarBase). FAIRsharing. https://doi.org/10.25504/fairsharing.fbbxfg.

49. Chen WH, Lu G, Chen X, Zhao XM, Bork P. The database of Online Gene Essentiality (OGEE). FAIRsharing. https://doi.org/10.25504/fairsharing.hsy066.

50. Lenort WF, Lin TL, Hart T. The database of pooled in vitro CRISPR knockout library essentiality screens (PLOLES). Hart Lab URL: https://hartlab.shinyapps.io/PLOLES/. Accessed September 2017.

51. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. The human protein atlas (HPA). FAIRsharing. https://doi.org/10.25504/fairsharing.4jy0pe.

52. Fabregat A, Jupe S, Matthews L, Sidiroopoulos K, Gillespie M, Garapati P, et al. Reactome. FAIRsharing. doi: https://doi.org/10.25504/fairsharing.f6dkbj.

53. O’Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. The NCBI reference sequence database (RefSeq). FAIRsharing. https://doi.org/10.25504/fairsharing.4gg0qw.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.