Bioinformatics, 38(13), 2022, 3407–3414
https://doi.org/10.1093/bioinformatics/btac329
Advance Access Publication Date: 17 May 2022
Original Paper

Systems biology

PersonaDrive: a method for the identification and prioritization of personalized cancer drivers

Cesim Erten[1,*], Aissa Houdjedj[1,2], Hilal Kazan [1,*] and Ahmed Amine Taleb Bahmed[3]

[1]Department of Computer Engineering, Antalya Bilim University, Antalya 07190, Turkey, [2]Department of Computer Engineering, Akdeniz University, Antalya 07070, Turkey and [3]Institute of Postgraduate Education, Electrical and Computer Engineering Graduate Program, Antalya Bilim University, Antalya 07190, Turkey

*To whom correspondence should be addressed.
Associate Editor: Anthony Mathelier

Received on September 30, 2021; revised on May 6, 2022; editorial decision on May 9, 2022; accepted on May 11, 2022

Abstract

Motivation: A major challenge in cancer genomics is to distinguish the driver mutations that are causally linked to cancer from passenger mutations that do not contribute to cancer development. The majority of existing methods provide a single driver gene list for the entire cohort of patients. However, since mutation profiles of patients from the same cancer type show a high degree of heterogeneity, a more ideal approach is to identify patient-specific drivers.

Results: We propose a novel method that integrates genomic data, biological pathways and protein connectivity information for personalized identification of driver genes. The method is formulated on a personalized bipartite graph for each patient. Our approach provides a personalized ranking of the mutated genes of a patient based on the sum of weighted 'pairwise pathway coverage' scores across all the samples, where appropriate pairwise patient similarity scores are used as weights to normalize these coverage scores. We compare our method against five state-of-the-art patient-specific cancer gene prioritization methods. The comparisons are with respect to a novel evaluation method that takes into account the personalized nature of the problem. We show that our approach outperforms the existing alternatives for both the TCGA and the cell line data. In addition, we show that the KEGG/Reactome pathways enriched in our ranked genes and those that are enriched in cell lines' reference sets overlap significantly when compared to the overlaps achieved by the rankings of the alternative methods. Our findings can provide valuable information toward the development of personalized treatments and therapies.

Availability and implementation: All the codes and data are available at https://github.com/abu-compbio/PersonaDrive, and the data underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.6520187.

Contact: cesim.erten@antalya.edu.tr or hilal.kazan@antalya.edu.tr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cancer is an evolutionary process arising in many cases from the dysregulation of gene sequence, expression, genomic and transcriptomic alterations, in which abnormal cells divide uncontrollably and can invade nearby tissues. The mutations that give a cancer cell a fundamental growth advantage and promote cancer development are called *driver mutations* and the corresponding genes subject to alteration are called *driver genes*. On the other hand, there may exist several other mutations that occur randomly in a tumor sample but that are not directly associated with cancer development. Such mutations are the so called *passenger mutations*. A major challenge in cancer genomics is to distinguish the drivers that are causally linked to cancer from the passenger mutations.

Due to the advances in high-throughput DNA sequencing technology, projects such as The Cancer Genome Atlas (TCGA) (Weinstein *et al.*, 2013) and The Cancer Cell Line Encyclopedia (CCLE) (Barretina *et al.*, 2012) have been able to systematically generate genomic data for thousands of tumors and cell lines across many cancer types providing useful data sources for many cancer driver identification studies. Several related versions of the cancer driver identification problem have been proposed. In one version the goal is to come up with a set of driver modules where each module consists of genes expected to be important in the development of the cancer under study (Ahmed *et al.*, 2020; Baali *et al.*, 2020; Ciriello *et al.*, 2012; Kim *et al.*, 2015; Leiserson *et al.*, 2015; Vandin *et al.*, 2012). Another problem version sets the goal as that of ranking the altered genes with respect to their potentials for being cancer

drivers. The methods suggested for this purpose can further be classified into two depending on whether the goal is to provide a cohort-level driver gene prioritization or a patient-specific driver gene prioritization.

The cohort-level driver gene identification and prioritization methods usually integrate multi-dimensional genomic data (Cheng *et al.*, 2016; Pham *et al.*, 2021). They can be categorized into different classes such as frequency-based, hotspot-based and network-based methods (Han *et al.*, 2019). MutsigCV is one of the notable approaches among the first class. It investigates mutational heterogeneity patterns by correcting for variation using patient-specific mutation frequency and spectrum, and gene-specific background mutation rates incorporating expression level and replication time (Lawrence *et al.*, 2013). A representative of the methods of the second class is OncodriveCLUST which detects driver genes with a significant bias toward mutations clustered within specific protein sequence regions (Tamborero *et al.*, 2013). Network-based methods constitute a class of promising methods to prioritize low-frequency and high-frequency cancer genes due to their power to elucidate molecular mechanisms and their ability to model gene interactions (Wei *et al.*, 2020). DriverNet is among the notable approaches that prioritize mutated genes based on their degrees of network connectivity to the differentially expressed genes (DEGs) in tumor samples (Bashashati *et al.*, 2012). The model used by DriverNet to relate the set of mutated genes to the set of DEGs has inspired many subsequent driver identification methods (Bertrand *et al.*, 2015; Erten *et al.*, 2021; Wei-Feng *et al.*, 2019).

1.1 Patient-specific driver ranking methods

It is well-known that the mutation profiles of patients from the same cancer type show a high degree of heterogeneity. Since each patient may have a distinct set of driver genes, a more ideal approach is to identify patient-specific drivers. There are two main challenges associated with personalized driver gene rankings; some patients may carry too many known driver genes, necessitating narrowing down the driver genes to the true drivers for that patient while on the other hand, some patients do not possess any known drivers, so it is necessary to discover novel drivers for those patients (Dinstag and Shamir, 2020). There have been several attempts to address these problems and similar to some cohort-level network-based driver identification methods, many of the methods within the personalized setting of the problem are also inspired by the concept underlying the DriverNet approach which relates mutations to their consequent effects on transcription and gene expression by utilizing an underlying interaction network or pathway information. DawnRank prioritizes personalized driver genes by quantifying the impact of the mutated gene on the DEGs using a random walk process (Hou and Ma, 2014). OncoIMPACT identifies driver genes as those that explain DEGs with small graph-theoretical distances to mutated genes in the gene interaction network for at least a certain proportion of patients (Bertrand *et al.*, 2015). The methods SSN, SCS and PNC share a common overall framework. In all three, first a personalized network is constructed based on gene expression data and an underlying biological network. Such a network is then analyzed by means of different heuristics for appropriate graph-theoretical problems. In SSN a sample-specific network is constructed based on a statistical perturbation analysis of the sample against a set of control samples (Liu *et al.*, 2016). Graph-theoretical shortest path distances between genes in the SSN and the mutated genes are then computed and checked for significance. The single-sample controller strategy (SCS) implements network control theory for personalized driver gene identification by searching for the minimal set of mutated genes to control the maximal coverage of individual DEGs in a directed biological interaction network (Guo *et al.*, 2018). On the other hand PNC creates the personalized network by constructing a differential co-expression network based on the data from tumor and normal sample, and applies a greedy heuristic for minimum vertex cover on the resulting network (Wei-Feng *et al.*, 2019). We note that unlike other personalized driver identification methods considered in this study, PNC does not output 'driver gene rankings' but rather 'sets' of driver genes. Prodigy employs the expression and mutation profiles of the patient along with data on known pathways and the connectivity information to prioritize personalized cancer driver genes using the Steiner tree model based on the impact of mutated genes on dysregulated pathways which are significantly enriched in DEGs (Dinstag and Shamir, 2020).

We observe two shortcomings of current personalized driver prioritization methods. One stems from the way they employ the available data. Although each one employs its algorithm on a specific sample from some cohort, it does so by neglecting the availability of data pertaining to other samples in the cohort. This is a waste of valuable data, as data from other samples may guide the personalized driver rankings of a specific sample. The emphasis of the personalized setting of the problem should not concern the used input data but rather solely the provided outputs so that the provided gene rankings should be patient-specific. Yet another important drawback is the lack of evaluation methodology consistent with the personalized setting of the problem. Mainly two different approaches are used in evaluations comparing alternative personalized driver ranking methods. In the first one, the relevant method is used on each sample from an available cohort and then the quality of the output is determined by cohort-wise aggregation of the patient-specific outputs. Usually a Condorcet voting is used for such an aggregation and the resulting gene set is compared against a set of known reference genes such as the COSMIC Cancer Gene Census (CGC) database (Tate *et al.*, 2019). DawnRank, SCS and IMCDriver (Zhang *et al.*, 2021b) are among the methods following this approach. Such an evaluation based on *ranking-aggregation-evaluation (RAE)* is flawed for the obvious reason; the patient-specific findings are lost in cohort-wise aggregation and a cohort-based prioritization method may provide better results than a patient-specific driver ranking method. Another evaluation strategy used by PRODIGY is to compare the rankings of each sample against the reference set of known drivers separately and then aggregate the results by averaging values for the entire cohort as a function of the top N ranked genes. If a sample has less than N ranked genes, the last value for that sample is duplicated so that all quality measure vectors for all patients are of length N. Although not as severe as the RAE strategy, this evaluation strategy based on *ranking-evaluation-aggregation (REA)* suffers from certain shortcomings as well. The first issue is that any set of reference known cancer drivers, although appropriate for cohort-based driver prioritization settings, is not an appropriate ultimate golden standard for the personalized setting of the problem. If used in such a setting, the evaluations must be supported by other strategies that can emphasize the personalized nature of the problem. Secondly, even if the issue with the golden standard is neglected, the duplication procedure of the REA strategy is problematic; a hypothetical prioritization providing a single gene such as TP53 as its output for every sample would superficially outperform the personalized driver identification and ranking methods.

1.2 Novelties of PersonaDrive

In this work, we propose a novel method called PersonaDrive that integrates genomic data, transcriptomic data, protein–protein interaction and biological pathway data for the identification and prioritization of personalized driver genes. Unlike other network-based personalized driver identification and ranking methods, PersonaDrive takes into account the data available from the whole cohort while producing the personalized ranking for every specific patient. A *pairwise patient similarity* score based on the amount of overlap between the set of DEGs of the pair is used in determining the degree of the influence each patient's data has on the other's personalized driver ranking. Although the idea of exploiting data from the whole cohort exists in methods based on machine learning such as the IMCDriver, network-based methods not employing learning have thus far not incorporated the concept. Furthermore, we note that a learning-based method like IMCDriver that employs the RAE strategy for the evaluations suffers even more than its non-learning-based counterparts, since the reference set of known drivers is already part of the input features used in learning. In fact, the only other set of features used in the learning stage of the method is

simply the mutations data of the samples. This makes the prioritizations of such methods highly biased toward the reference set and evaluating such prioritizations based on the RAE strategy applied with the same set of reference genes may lead to incorrect conclusions. A second novelty of PersonaDrive is the concept of *pairwise pathway coverage* based on which the rankings of mutated genes are determined. Inspired by the original DriverNet model building on the influence of mutated genes on the DEGs and the subsequent methods DawnRank, PRODIGY and SCS all of which employ a similar idea, we also create a network model of such influences. However different from these methods, PersonaDrive's ranking score is based on the amount of coexistence of the mutated gene and the DEG pair in biological pathways, that is pairwise pathway coverage. Informally, a mutated gene of a sample coexisting with many DEGs in many pathways gets a higher score and this score is magnified even more if it does so in many *similar* samples. Finally, another contribution of this work is with regard to the proposed evaluation framework. Notably, we propose a modification to the REA strategy of PRODIGY to resolve the issues arising from the duplication idea. More importantly, we propose a novel framework for the evaluation of personalized driver identification and ranking methods. This framework is based on the idea of creating a separate reference golden standard set for each sample. Since the cell line data is usually richer than the tumor data available through TCGA in terms of personalized features, the framework is built on employing such data. A reference gene set is constructed for each cell line sample by incorporating the drug sensitivity together with the drug targets data.

2 Materials and methods

We first provide a description of the PersonaDrive method, a novel personalized cancer driver identification and ranking algorithm. Next, we provide a description of the introduced framework for the evaluation of cancer driver identification and prioritization outputs. The proposed evaluation framework includes certain novelties directed specifically toward the personalized setting of the problem.

2.1 PersonaDrive algorithm

Figure 1 provides an overview of the PersonaDrive algorithm which consists of three main steps: (i) constructing personalized bipartite networks; (ii) computing edge weights in the bipartite networks; (iii) personalized ranking of the genes. Below we provide a description of each of these steps in detail.

2.1.1 Personalized bipartite networks

Similar to DriverNet, PersonaDrive constructs a bipartite graph to model the relationship between the set of mutated genes and DEGs. However different from DriverNet, PersonaDrive constructs a personalized network for each sample.

Let $G = (V, E)$ denote the PPI network, where V denotes the set of nodes corresponding to the genes and E denotes the set of edges corresponding to pairwise protein–protein interactions. Assume the genes are denoted with $g_1, g_2, \ldots g_{|V|}$. Note that we use the same notation to denote a node of G and its corresponding gene. Let $P = \{P_1, P_2, \ldots, P_r\}$ denote the set of biological pathways and $S = \{S_1, \ldots S_n\}$ denote the set of samples. If a gene g_x is mutated in sample S_i we create an instance of g_x denoted with m_x^i and denote the set of all such instances of genes mutated in sample S_i with M_i. Similarly, if a gene g_y is a DEG for sample S_i, we create an instance of g_y denoted with d_y^i and denote the set of all such instances of DEGs of sample $S_i$with D_i. For each sample, a gene is regarded as a DEG if its z-score > 0.5 or z-score < -0.5 based on its expression value in the sample as compared to its expression values in the whole population. Note that this is similar to the DriverNet's *outlying gene* concept except that DriverNet employs a threshold of 2 rather than 0.5. Unlike the definitions used by most personalized driver identification and ranking methods such as DawnRank, PRODIGY and SCS, this definition of a DEG does not rely on the existence of paired normal/tumor data or background gene expression distributions from healthy samples of the same tissue of origin.

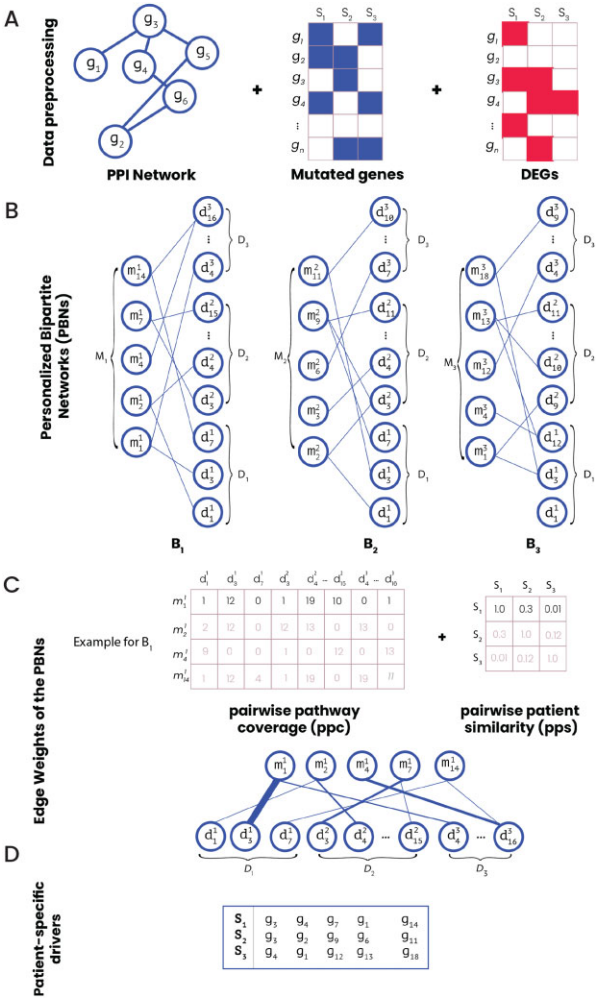

Fig. 1. A depiction of the main steps of the PersonaDrive algorithm. (**A**) Integrating the relevant genomic data (mutations and DEGs) with the interaction network. (**B**) Constructing the personalized bipartite network B_i for each sample S_i in the cohort S. One partition of B_i is the set M_i consisting of m_x^i which is an instance of the gene g_x mutated in S_i. Each d_y^i in the second partition corresponds to an instance of a DEG g_y of sample S_i. (**C**) Computing the pairwise pathway coverage and the pairwise patient similarity values. The edge weights in each personalized bipartite network (PBN) B_i is then assigned based on these values. (**D**) The final output is a ranking of the mutated genes in M_i based on their influence scores which in turn are computed with respect to the weights of incident edges in the PBNs

For each sample S_i we construct a bipartite graph B_i with the edge set E_i. B_i has two node partitions, the first partition consists of nodes corresponding to the set of mutated genes M_i. The second partition has nodes corresponding to the union of instances of DEGs of all the samples, that is $D_1 \cup D_2 \cup \ldots D_n$. Note that a gene may appear multiple times in the second partition as instances of different DEG sets from different samples. For $m_x^i \in M_i, d_y^i \in D_i$, there exists an edge $(m_x^i, d_y^i) \in E_i$, if $(g_x, g_y) \in E$ and $m_x^i \in M_i$. In other words, an instance of a gene g_x mutated in sample S_i has a bipartite edge with an instance of a gene g_y determined to be a DEG in sample S_i, if they interact in the PPI network and g_x is a mutated gene in sample S_i as well. The subsequent steps assume all nodes in T_i^2 have degree greater than zero in B_i. Therefore, we remove all zero-degree nodes from the second partition.

2.1.2 Edge weights of the personalized bipartite networks

The weight of an edge $(m_x^i, d_y^i) \in E_i$ is computed with respect to a normalized value of the *pairwise pathway coverage (ppc)* score of

incident nodes m_x^i and d_y^j. The score $\text{ppc}(m_x^i, d_y^j)$ is defined as the number of pathways $P_k \in P$ such that $g_x, g_y \in P_k$. The idea behind the *ppc* scoring is that the more the pathways shared by a mutated gene and DEG pair of a sample, the larger the potential for the mutated gene to be a cancer driver in that sample. Note that this is quite different from simply counting the number of pathways a mutated gene is involved in. For a mutated gene in a sample and every interacting partner that is a DEG in the sample, the pathways they coincide are counted toward the driver potential of the mutated gene in that sample. Furthermore, since it is not necessarily the case that $i=j$, an edge weight is not computed in isolation, the data from a single sample being the sole determinant. It is also exposed to the mutation/dsyregulation patterns of the rest of the samples in the whole cohort, up to certain degrees formalized by the proposed normalization.

The normalization of the weight of an edge $(m_x^i, d_y^j) \in E_i$ is with respect to the *pairwise patient similarity (pps)* value of the samples S_i and S_j. Such a similarity is defined based on the overlap of the DEGs of the two patients. More specifically, for samples S_i, S_j we define $\text{pps}(i,j) = \frac{|D_i \cap D_j|^2}{|D_i| \times |D_j|}$. We note that similar overlap score definitions have been used previously in measuring amount of overlaps among pairs of groups in various studies on clustering biological networks (Baali *et al.*, 2020; Nepusz *et al.*, 2012). Incorporating the proposed normalization we define the edge weight of $(m_x^i, d_y^j) \in E_i$ as, $w(m_x^i, d_y^j) = \text{ppc}(m_x^i, d_y^j) \times \text{pps}(i,j)$.

2.1.3 Prioritizing mutated genes

For each mutated gene g_x in sample S_i, we define an influence score $\text{Infl}(m_x^i)$ for the instance m_x^i based on the weights of incident edges in the corresponding bipartite graph B_i:

$$\text{Infl}(m_x^i) = \sum_{\forall (m_x^i, d_y^j) \in E_i} w(m_x^i, d_y^j). \quad (1)$$

All mutated genes of S_i are prioritized with respect to this influence score and output in the relevant order by the PersonaDrive algorithm.

2.2 Data collection and preprocessing

We employ the TCGAbiolinks R package to compile mutation and gene expression data for tumor samples and gene expression data for normal samples from TCGA (Colaprico *et al.*, 2016). The colon adenocarcinoma (COAD) cohort contains 396 tumor and 41 normal samples, the lung adenocarcinoma (LUAD) cohort contains 508 tumor and 58 normal samples, and the breast adenocarcinoma (BRCA) cohort contains 974 tumor and 113 normal samples. We use the DepMap database version 20Q1 (Ghandi *et al.*, 2019) to gather the mutations and gene expression data of the cancer cell line samples. The ids of the cell lines are available in Supplementary Table S1. Regarding the expression data of normal cell line samples we use the normal human bronchial epithelial (NHBE) cell line data from Petryszak *et al.* (2016) for LUAD and the CCD-18Co human normal colon myofibroblasts data from Ferrer-Mayorga *et al.* (2019) for COAD. For both the TCGA and the CCLE datasets, we filter out the silent mutations as classified by Mutect2 (Cibulskis *et al.*, 2013) from the somatic mutation data.

Our main results are obtained with the STRING network (v11.5) (Szklarczyk *et al.*, 2019) as the input PPI network and KEGG pathways (Kanehisa *et al.*, 2021) for the input set of biological pathways. For the PPI network, we include experimentally validated physical interactions with confidence score greater than 0.7 from the STRING network (v11.5) resulting in 10 196 nodes and 163 627 edges. Hereafter this network is simply referred to as the STRING network. We provide two more control studies with additional results using alternative PPI networks and pathway databases used by the methods compared against PersonaDrive. Both control studies utilize the KEGG pathways used in Dinstag and Shamir (2020). In the first control study (CS1) we employ the version of the STRING network used in Dinstag and Shamir (2020) whereas in the second control study (CS2), we employ the DawnRank network (Hou and Ma, 2014). Further details regarding data preprocessing and the used networks are available in Supplementary Section S1 of the Supplementary Document.

2.3 Comparative evaluation framework

We propose two main types of comparative evaluations emphasizing the personalized aspect of the problem. The first type of evaluations employs the reference sets used in previous work and a modified version of the REA strategy proposed in Dinstag and Shamir (2020). The second type of evaluations in line with the personalized nature of the problem proposes the use of novel personalized reference sets that are based on cell line data. We first construct reference sets by incorporating drug sensitivity data specific to each cell line sample and repeat the evaluation methodology with respect to the modified REA strategy. In order to complement the findings, we propose an additional evaluation method on cell line data that abandons the modified REA strategy as a way to quantify the matches between reference sets and output driver prioritizations but rather introduces a more flexible matching based on pathway enrichments.

2.3.1 Evaluations with reference sets relevant for cohort studies

The first type of evaluations is based on appropriate modifications of the REA strategy. Note that the REA strategy is used by PRODIGY after producing a *personalized reference set* for each sample. The personalized reference set of a sample is considered to be the intersection of the set of mutated genes of the sample and an appropriate reference set of known cancer genes. The original strategy first fixes a certain desired output size N. In the case of PRODIGY evaluations N is set to 20. For every increment of k from 1 to N, for a benchmark method $\mathcal{M}$ it compares the first k genes output by M in the ranking of each sample S_i against the personalized reference set of S_i. It then aggregates the evaluation scores, usually measured in terms of precision and recall, by averaging the values over the entire cohort. If a method $\mathcal{M}$ provides k' ranked genes for a certain sample S_i for $k' < N$, the evaluation score for S_i at instance k' is repeatedly duplicated until N in the evaluation scores of $\mathcal{M}$.

We propose two main modifications to the original strategy. First of all, since not all samples are expected to have the same number of personalized cancer drivers, the first modification to this strategy consists of assigning N dynamically based on the sizes of the used personalized reference sets. We define N to be twice the median of the sizes of the personalized reference sets after excluding the samples with reference set sizes less than three from the evaluations. Secondly, due to the issues stemming from the duplication procedure of the original strategy, rather than duplicating the scores of $\mathcal{M}$ for S_i at instance k' if $k' < N$, we simply remove the sample S_i from all the successive average evaluation score calculations for the instances $k'+1, \ldots, N$.

2.3.2 Evaluations based on cell line data

We first apply the first type of evaluations on the experiments employing cell line data. More specifically, the personalized reference set for each cell line is defined as the intersection of the set of mutated genes of the sample and the appropriate reference set of known drivers, and the modified REA strategy is applied for quantification of matches between reference sets and the output prioritizations. We call these evaluations $CLEV_1$.

The second type of evaluations are based on the observation that even the modified REA strategy may fail to provide conclusive findings if the personalized reference sets are constructed from known driver databases appropriate for population-based studies, such as the CGC, NCG (Repana *et al.*, 2019) or CancerMine (Lever *et al.*, 2019). In order to resolve this issue and construct potential personalized reference sets based on data specific to each sample we propose the use of cell line data coupled with drug sensitivity data. For this type of evaluations, for each available cell line we define a novel reference gene set by compiling the target genes of drugs that are found to be sensitive based on data from GDSC and DepMap

databases. The reference set is further filtered with the CGC set of known drivers. Once the personalized reference sets are determined, the rest of the modified REA strategy described above is applied. We call these evaluations $CLEV_2$. The statistical information of the targets in GDSC and the DepMap reference sets for each cancer type under study are available in Supplementary Tables S2–S4. The GDSC cell line drug sensitivity is retrieved from the GDSC2 dataset (Yang *et al.*, 2013). We label a cell line as *sensitive* to a drug if the z-score value is less than 0. For DepMap cell line drug sensitivity data, we employ the values in 'primary-screen-replicate-collapsed-log-fold-change.csv' file (Corsello *et al.*, 2020). This dataset consists of the results of pooled-cell line chemical-perturbation viability screens. We label a cell line as sensitive to a drug if the collapsed log fold-change value is less than -0.8.

Finally, to extend the personalized emphasis proposed by the second type of evaluations we introduce one last type of evaluations again based on the same types of data. The extension is based on the observation that directly comparing overlaps of sensitive drug target genes with the prioritized genes of a cell line can be too strict. Therefore, we also evaluate the methods based on KEGG and Reactome (Fabregat *et al.*, 2018) enrichment analysis by checking the amounts of overlaps between the pathways enriched significantly in the genes output by some personalized prioritization method and those that are enriched in cell line reference sets constructed from drug sensitivity data. In pathway analysis to avoid bias related to size it is suggested to employ gene sets of size at least 10 (Cirillo *et al.*, 2017). Therefore, we filter out the samples with less than 20 genes in their reference sets. Furthermore, samples with less than 20 genes in the output prioritization of any method under comparison are also excluded from the analysis. We find the set of enriched KEGG or Reactome pathways using the g: GOSt tool (the core of g: Profiler tool) of Raudvere *et al.* (2019) which maps genes to known functional information sources and detects statistically significantly enriched terms. Note that all the enrichment analyses are performed with respect to KEGG Release 101 and Reactome (BioMart 2022-01-03 version). For each cell line sample S_i, we identify $\mathcal{E}_i^{\mathcal{M}}$, the set of pathways enriched significantly in the output prioritized gene set of a method $\mathcal{M}$ and $\mathcal{E}_i^{\mathcal{R}}$, the set of pathways enriched significantly in the reference set of S_i. The set of mutated genes M_i are used as the background set in both cases. We then compute the *enriched pathway overlap (EPO)* score of a method $\mathcal{M}$ as the average of $|\mathcal{E}_i^{\mathcal{M}} \cap \mathcal{E}_i^{\mathcal{R}}|/|E_i^{\mathcal{R}}|$ over all samples S_i.

3 Results

We first compare PersonaDrive results against those of five existing personalized driver prioritization methods: Prodigy, SCS, DawnRank, OncoIMPACT and SSN. Note that SSN is not a driver ranking algorithm per se, since it outputs candidate driver sets. However, since the output driver sets are constructed solely based on the degrees of the genes in the proposed SSN network, an explicit ranking can be produced by ranking the genes with respect to their degrees. We include two versions of SSN where SSN_{orig} is the original output of the SSN method and SSN_{mut} is its output filtered by us to contain only mutated genes. On the other hand, we do not include PNC, another driver set identification algorithm similar to SSN, in our evaluations since its output set construction method does not imply a ranking of genes. Our evaluations are based on patient data from TCGA project and cell line data from the CCLE project. Followed by the comparative evaluations we provide a more detailed analysis of the output prioritizations provided by the PersonaDrive algorithm.

3.1 Comparative evaluations

Prodigy, SCS, DawnRank and OncoIMPACT assume the same type of input data as PersonaDrive. Therefore none of these algorithms has any additional advantage of extra information. More specifically, all these algorithms assume data from a cohort which include somatic mutation data and the gene expression data of the tumor samples, PPI network data and pathway data. The only exception is SSN which employs all this data except the somatic mutation data. PRODIGY and PersonaDrive make explicit use of the pathway information, whereas DawnRank, SCS, SSN and OncoIMPACT use it implicitly in their used *gene networks* which aggregate protein–protein interactions and pathway interactions. Actually, PersonaDrive is even less restricted than the other benchmark methods in terms of its input requirements as it requires the expression data only from the tumor samples. On the other hand all the other methods additionally require gene expression data from the normal samples. The experimental results presented in the main document are those obtained with the STRING network, whereas the analogous results with the networks of the control studies CS1 and CS2 are provided in the Supplementary Document.

3.1.1 Comparisons with reference sets relevant for cohort studies

The personalized reference sets are constructed with respect to several relevant reference sets of known cancer genes. The first reference set employs CGC and is denoted with CGC_{all}. We further extract cancer type-specific known drivers from it by checking the 'Tumor Types (Somatic)' information. The resulting set is denoted with $CGC_{specific}$. We also compile cancer type-specific genes from the Network of Cancer Genes (NCG) by filtering the *primary site* column. This reference set is denoted with NCG_{all}. We further identify a subset of it as the intersection of NCG_{all} and CGC_{all}. Hereafter, this reference gene set is referred to as NCG_{CGC}. The third repository, CancerMine, uses text-mining to catalogue cancer associated genes where it also extracts information about the type of the cancer. We compile a list of genes that have at least two citations in literature and denote the resulting reference set with $CancerMine_{all}$. Note that for BRCA we select a more strict threshold of at least three citations due to the large size of the reference set obtained otherwise. The last reference set is constructed by intersecting $CancerMine_{all}$ cancer-specific genes with the set of all CGC genes CGC_{all} and is denoted with $CancerMine_{CGC}$ The number of genes in each reference is provided in Supplementary Table S5.

We evaluate each method based on how well it recovers the personalized reference set of each sample which is constructed by intersecting the set of mutated genes of the sample and one of the reference sets described above. The evaluation strategy is described in Section 2.3.1. Figure 2 shows the F1 score values for all the used methods for the COAD, LUAD and BRCA datasets from TCGA (see Supplementary Figs S1–S3 for the mean precision, and recall values). Here, the results are obtained with the STRING network and $CGC_{specific}$ reference set. We note that N is determined as 8, 6 and 6 based on the sizes of the used personalized reference sets for the COAD, LUAD and BRCA datasets, respectively. For all the cohorts, we observe that PersonaDrive achieves a higher performance than the alternatives in terms of F1 score. It is followed by

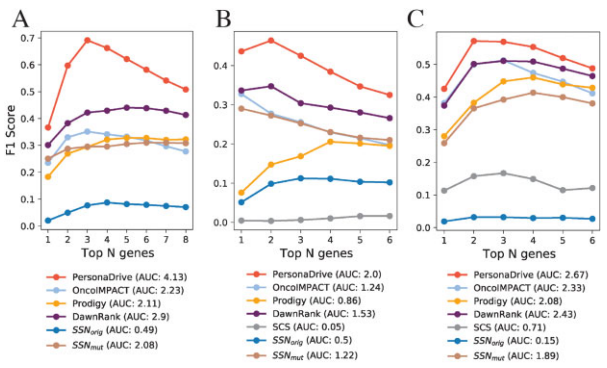

Fig. 2. Comparison of the PersonaDrive outputs with those of the five alternative methods, DawnRank, SCS, PRODIGY, SSN (two versions of SSN where SSN_{orig} is the original output of the SSN method and SSN_{mut} is its output filtered by us to contain only mutated genes), and OncoIMPACT in terms of average F1 values for (**A**) TCGA COAD dataset, (**B**) TCGA LUAD dataset and (**C**) TCGA BRCA dataset. STRING network is used as the input interaction network and $CGC_{specific}$ is used as the reference set

DawnRank and OncoIMPACT. The ranking of the other methods varies depending on the cohort where SCS and SSN_{orig} show the worst two performances. Furthermore, we confirm that the difference between the F1 scores of PersonaDrive and each alternative method obtained with k for $k = 1 \ldots N$ is significant using Wilcoxon signed-rank test (Supplementary Table S6). The only cases where PersonaDrive is worse than the alternative method for the majority of k values is its comparisons with DawnRank for TCGA LUAD dataset with CGC_{all} and $CancerMine_{all}$ reference sets. Note that SCS is not included for some evaluations as it does not execute until completion within a reasonable amount of time. We also observe that the results remain consistent even if we vary N up/down by 25% (Supplementary Tables S7–S9). When we construct each personalized reference set with respect to the NCG_{CGC} genes or $CancerMine_{CGC}$ genes, we still observe that PersonaDrive performs better than its alternatives (Supplementary Figs S1–S3). We evaluate the methods with three additional reference sets: CGC_{all}, NCG_{all} and $CancerMine_{all}$ (Supplementary Figs S4–S6). PersonaDrive achieves the top performance for all datasets and metrics.

We repeat the analogous experiments with the network data of CS1 and CS2; see Supplementary Figures S7–S9 for the former and Supplementary Figures S10–S12 for the latter. The results show that PersonaDrive's top performance do not depend on the used interaction network. The only case where PersonaDrive gives a slightly worse performance than DawnRank is for the LUAD dataset under CGC_{all} and $CancerMine_{all}$ reference sets.

3.1.2 Comparisons based on cell line data

Next, we compare the performances of the methods based on the personalized reference sets constructed with respect to the cell line data and evaluated with respect to $CLEV_1$ and $CLEV_2$. We note that our cell line evaluations do not include BRCA since there are only five cell lines with at least three genes in their references for this cancer type. Figure 3A and B shows the mean F1 values for all the methods for the COAD and LUAD cell lines's data respectively, where evaluations are with respect to $CLEV_1$, the STRING network is used as input and the references are constructed based on $CGC_{specific}$; see Supplementary Figures S13 and S14 for the analogous results under the rest of the reference sets and Supplementary Figures S15–S18 for the analogous results on the control studies CS1 and CS2. Similar to the results obtained with TCGA data, PersonaDrive outperforms the alternative methods by a large margin where evaluations are with respect to $CLEV_1$. Figure 3C and D shows the mean F1 values for all the used methods for the COAD and LUAD cell lines' data respectively, where evaluations are with respect to $CLEV_2$, the STRING network is used as input; see Supplementary Figure S19 for the mean precision and recall values. We note that N is determined as 30 and 10 for the COAD and LUAD datasets, respectively. PersonaDrive achieves the top performance in terms of all three metrics on the COAD dataset. It is followed by DawnRank, Prodigy, SSN_{mut}, OncoIMPACT, SSN_{orig} and SCS in the order of decreasing performance. Similarly, for the LUAD cell lines, PersonaDrive outperforms the alternatives in terms of area under the curve. It is followed by DawnRank, Prodigy, SSN_{mut}, OncoIMPACT, SCS, SSN_{orig} in the order of decreasing performance. Wilcoxon signed-rank test results show that PersonaDrive's F1 scores are significantly higher than those of other methods for all the cases except for its comparison with DawnRank in Figure 3D (Supplementary Table S10). We also show that the top performance of PersonaDrive is robust to varying N by 25% up or down (Supplementary Table S11). We repeat the analogous experiments with the network data of CS1 and CS2; see Supplementary Figures S20 and S21. The results show that PersonaDrive achieves the best performance. SSN_{orig} consistently ranks the worst for all the comparisons whereas the ranking of the other methods varies.

We perform an additional evaluation on the cell line datasets where we utilize pathway enrichment analysis as described in the second part of Section 2.3.2. This evaluation is performed only on the COAD dataset since the set of pathways enriched significantly in cell line references constructed from drug sensitivity data is empty for 19 out of 27 LUAD cell lines. For COAD dataset, PersonaDrive performs the best in terms of EPO scores. Its EPO scores are 0.57 and 0.36 for KEGG and Reactome pathways, respectively. The second best performer is Prodigy with EPO scores of 0.18 and 0.21, respectively for the KEGG and Reactome pathways. The analogous scores of DawnRank and OncoIMPACT are 0.1, 0.3 and 0.02, 0, respectively. The analogous scores of SSN_{orig} and SSN_{mut} are 0.02, 0.02 and 0.03, 0.11, respectively. Lastly, the EPO scores of SCS are 0 and 0, respectively for the KEGG and Reactome pathways. Results for control studies CS1 and CS2 are similar and are available in Supplementary Figure S22. In addition, we investigate the most commonly enriched KEGG pathways across the cell lines with respect to their reference sets and determine whether these pathways are also found enriched within the sets of genes output by the methods under consideration. We observe that PersonaDrive output genes achieve the largest overlap for the majority of the commonly enriched KEGG and Reactome pathways (Supplementary Tables S12–S17).

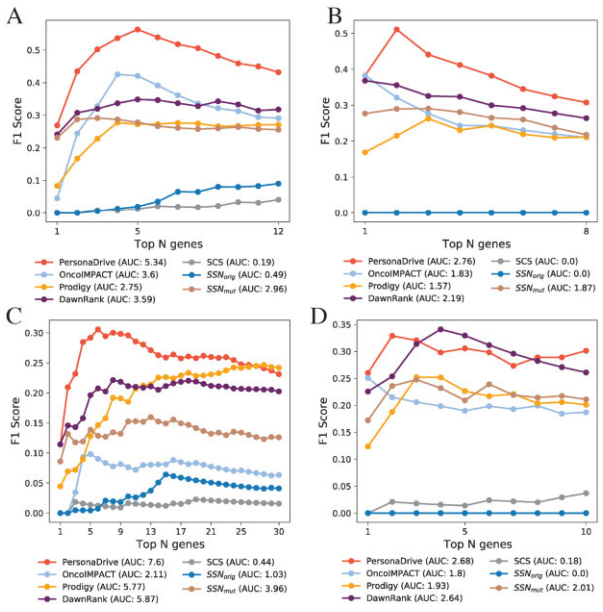

Fig. 3. Comparison of the PersonaDrive outputs with those of the five alternative methods, DawnRank, SCS, PRODIGY, SSN (two versions of SSN where SSN_{orig} is the original output of the SSN method and SSN_{mut} is its output filtered by us to contain only mutated genes), and OncoIMPACT in terms of average F1 values for (**A**) CCLE COAD cell lines ($CLEV_1$), (**B**) CCLE LUAD cell lines ($CLEV_1$), (**C**) CCLE COAD cell lines ($CLEV_2$), (**D**) CCLE LUAD cell lines ($CLEV_2$). The STRING network is used as the input interaction network

3.2 Detailed analysis of PersonaDrive gene rankings

We provide an in-depth analysis of the prioritizations provided by the PersonaDrive algorithm. For this purpose, we first assess whether PersonaDrive can identify rare drivers. We employ the TCGA dataset for this assessment since the number of samples in the cell line dataset is not appropriate for an analysis on rare drivers. We investigate the top k genes ranked by PersonaDrive for $1 \leq k \leq 20$ and count the number of genes that belong to the following bins defined by the mutation frequencies and the CGC reference set of known drivers: $< 2\%(CGC_{all}), 2-5\% \quad (CGC_{all}), > 5\%(CGC_{all}), < 2\%, 2-5\%, > 5\%$. The first three bins correspond to the genes that belong to the set CGC_{all} and each of the rest of the bins contains genes not in CGC_{all}. For all datasets, the top gene appears in the $> 5\%(CGC_{all})$ category most frequently. As N increases from 1 to 20 we see a decrease in the size of this category, whereas the sizes of the categories for rare known drivers and the categories for unknown drivers increase. Interestingly, for the LUAD and BRCA datasets, we observe more genes in rare categories (both in CGC_{all} and not in CGC_{all}) compared to the results obtained from the COAD dataset. Overall, these results show that

PersonaDrive is able to prioritize rarely mutated genes and detect both rare and frequent known drivers (Supplementary Fig. S23).

Secondly, employing cell line data we implement two complementary types of analysis investigating the novel genes provided by PersonaDrive. In the first one the focus is on the distinctive genes provided for each of the cell lines used in the evaluations, whereas in the second one the focus is on the common genes provided for most of the cell lines. With regard to the first type of analysis, we relate the personalized drivers identified and ranked by PersonaDrive with the gene *essentiality* scores of DepMap computed based on systematic loss of function screens on cell lines (Meyers *et al.*, 2017). More specifically, for each sample we analyze the top ranked genes provided by PersonaDrive and check whether the novel ones among them, that is those not in the CGC_{all} are verified for their driver potential by an external source such as DepMap essentiality. In order to emphasize the personalized nature of the problem setting, rather than simply using the essentiality scores we make use of the *preferentially essential genes* of a cell line as provided by DepMap. A similar evaluation has been performed under the cohort setting of the problem in Schulte-Sasse *et al.* (2021). For a cell line and a gene, the *mean-subtracted score* is calculated by subtracting the mean score for that gene across all cell lines from the score of the gene in that cell line. Those with the lowest mean-subtracted scores are defined as the preferentially essential genes of the cell line. For each cell line, we compute the intersection between our top 20 novel genes and the top 500 preferentially essential genes determined for that cell line. For LUAD, this intersection includes the genes *SOS1* (ACH-000787), *XAB2* (ACH-000587), *CPSF1* (ACH-000343), *NDUFS2* (ACH-000888) and *PTK2* (ACH-000861) where the corresponding cell lines are indicated in parentheses. Among these genes, there is experimental evidence for SOS1's oncogenic activity in lung cancer (Cai, 2019). In addition, upregulation of NDUFS2, a core subunit of mitochondrial complex I, has been shown to promote invasion of lung cancer cells (Liu *et al.*, 2019). XAB2 mutations have been significantly associated with the risk of non-small cell lung cancer in Chinese population (Pei *et al.*, 2015). Similarly, PTK2 is a member of the non-receptor tyrosine kinase family and regulates cell survival, proliferation, migration and invasion (Nana *et al.*, 2019). Its inhibition has been explored as a potential therapy in both small cell and non-small cell lung cancer (Tong *et al.*, 2019). For COAD, the corresponding intersection includes the genes *HK3* (ACH-001458), *GSK3B* (ACH-000957), *PTK2* (ACH-000943). The complete list can be found in Supplementary Table S18. Among these genes, GSK3 inhibitors (GSK3i) have shown a strong synergistic effect with PARP inhibitors in a panel of colorectal cancer (CRC) cell lines with diverse genetic backgrounds (Zhang *et al.*, 2021a). HK3's overexpression has been found to be associated with epithelial-mesenchymal transition in colorectal cancer (Pudova *et al.*, 2018) and overexpression of PTK2 is correlated with metastatic colon cancer (Lark *et al.*, 2003; Tai *et al.*, 2016).

With regard to the second type of analysis, we perform a batch study on the novel genes proposed by PersonaDrive. For this, we take PersonaDrive's most frequently prioritized genes among the top 20 ranking genes across all cell lines and explore the literature for associations with cancer. For lung cancer, Fibronectin 1 (FN1) which is ranked among the top 20 for five cell lines has been found to play critical roles in driving lung cancer (Spada *et al.*, 2021). For colon cancer, IRS1 which appears among the top 20 genes for nine cell lines is a key transducer of carcinogenesis (Esposito *et al.*, 2012). In addition, polymorphisms in IRS1 are found to change the risk for colorectal cancer (Pechlivanis *et al.*, 2007). Integrin 4 (ITGB4), which appears among the top 20 genes for seven cell lines, has been shown to play an important role in the regulation of cancer stem cells (CSC). Also, immune targeting of ITGB4 inhibits tumor growth and decrease metastasis in colon cancer cell lines (Ruan *et al.*, 2020).

4 Conclusion

Several crucial design choices are made in the PersonaDrive algorithm such as keeping the bipartite network and the set of pathways static throughout the ranking procedure, the z-score threshold used to determine the set of DEGs, the calculation of pairwise patient similarity (pps) scores. We discuss each such choice and compare against the plausible alternatives in the Supplementary Document Section S3. Finally, to verify that the achieved results presented in the main document are not artifacts of the choice of the datasets and the specific evaluation strategies used, we repeated our evaluations on the same TCGA COAD dataset as that used in the most recent benchmark study, PRODIGY. Furthermore, for these evaluations we used the unmodified REA strategy which is the original evaluation strategy proposed in the same study. We show that our conclusions remain almost the same in this controlled setting as well; see Supplementary Document Section S4. We acknowledge some limitations of the proposed method. PersonaDrive's DEG definition ignores the directionality and the magnitude of gene expression changes. A more detailed DEG definition incorporating these types of information has potential to improve the accuracy of the method. In addition, similar to the other existing methods, PersonaDrive is affected by the incompleteness of PPI and biological pathway databases. Increased completeness of these sources will decrease the number of false negative predictions of PersonaDrive. One future direction of research is to extend our analyses to additional cancer types as experimental data on a larger set of cell lines become available. Another future direction is to utilize recent clinical datasets that include drug response data where several drugs are administered to different subsets of patients. In this way, we can define personalized reference sets for patients as well, by using a similar procedure to produce personalized reference sets for the cell lines.

Acknowledgements

The authors are listed in the alphabetical order of their lastnames. They thank Gal Dinstag for providing details regarding the PRODIGY method.

Funding

This work was supported by the Scientific and Technological Research Council of Turkey [117E879 to H.K. and C.E.] and Health Institutes of Turkey [2019-TA-01-4069 to H.K. and C.E.].

Conflict of Interest: none declared.

References

Ahmed,R. *et al.* (2020) MEXCOwalk: mutual exclusion and coverage based random walk to identify cancer modules. *Bioinformatics*, **36**, 872–879.

Baali,I. *et al.* (2020) DriveWays: a method for identifying possibly overlapping driver pathways in cancer. *Sci. Rep.*, **10**, 21971.

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Bashashati,A. *et al.* (2012) DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.*, **13**, R124.

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

Cai,D. *et al.* (2019) Identification and characterization of oncogenic SOS1 mutations in lung adenocarcinoma. *Mol. Cancer Res.*, **17**, 1002–1012.

Cheng,F. *et al.* (2016) Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Brief. Bioinform.*, **17**, 642–656.

Cibulskis,K. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, **31**, 213–219.

Ciriello,G. *et al.* (2012) Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, **22**, 398–406.

Cirillo,E. *et al.* (2017) A review of pathway-based analysis tools that visualize genetic variants. *Front. Genet.*, **8**, 174.

Colaprico,A. *et al.* (2016) TCGAbiolinks: an R/bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71.

Corsello,S.M. *et al.* (2020) Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat. Cancer*, **1**, 235–248.

Dinstag,G. and Shamir,R. (2020) PRODIGY: personalized prioritization of driver genes. *Bioinformatics*, **36**, 1831–1839.
Erten,C. *et al.* (2021) Ranking cancer drivers via betweenness-based outlier detection and random walks. *BMC Bioinformatics*, **22**, 62.
Esposito,D.L. *et al.* (2012) The insulin receptor substrate 1 (IRS1) in intestinal epithelial differentiation and in colorectal cancer. *PLoS One*, **7**, e36190.
Fabregat,A. *et al.* (2018) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
Ferrer-Mayorga,G. *et al.* (2019) Vitamin D and Wnt3A have additive and partially overlapping modulatory effects on gene expression and phenotype in human Colon fibroblasts. *Sci. Rep.*, **9**, 8085.
Ghandi,M. *et al.* (2019) Next-generation characterization of the cancer cell line encyclopedia. *Nature*, **569**, 503–508.
Guo,W.-F. *et al.* (2018) Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics*, **34**, 1893–1903.
Han,Y. *et al.* (2019) DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic Acids Res.*, **47**, e45.
Hou,J.P. and Ma,J. (2014) DawnRank: discovering personalized driver genes in cancer. *Genome Med.*, **6**, 56.
Kanehisa,M. *et al.* (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
Kim,Y.-A. *et al.* (2015) MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. *Bioinformatics*, **31**, i284–i292.
Lark,A.L. *et al.* (2003) Overexpression of focal adhesion kinase in primary colorectal carcinomas and colorectal liver metastases: immunohistochemistry and real-time PCR analyses. *Clin. Cancer Res.*, **9**, 215–222.
Lawrence,M.S. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.
Leiserson,M.D.M. *et al.* (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, **47**, 106–114.
Lever,J. *et al.* (2019) CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat. Methods*, **16**, 505–507.
Liu,L. *et al.* (2019) S100a4 alters metabolism and promotes invasion of lung cancer cells by up-regulating mitochondrial complex i protein ndufs2. *J. Biol. Chem.*, **294**, 7516–7527.
Liu,X. *et al.* (2016) Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res.*, **44**, e164.
Meyers,R.M. *et al.* (2017) Computational correction of copy number effect improves specificity of CRISPR-cas9 essentiality screens in cancer cells. *Nat. Genet.*, **49**, 1779–1784.
Nana,A. *et al.* (2019) Therapeutic potential of focal adhesion kinase inhibition in small cell lung cancer. *Mol. Cancer Ther.*, **18**, 17–27.
Nepusz,T. *et al.* (2012) Detecting overlapping protein complexes in protein–-protein interaction networks. *Nat. Methods*, **9**, 471–472.
Pechlivanis,S. *et al.* (2007) Insulin pathway related genes and risk of colorectal cancer: INSR promoter polymorphism shows a protective effect. *Endocr. Relat. Cancer*, **14**, 733–740.
Pei,N. *et al.* (2015) XAB2 tagSNPs contribute to non-small cell lung cancer susceptibility in Chinese population. *BMC Cancer*, **15**, 560.
Petryszak,R. *et al.* (2016) Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.*, **44**, D746–D752.
Pham,V.V.H. *et al.* (2021) Computational methods for cancer driver discovery: a survey. *Theranostics*, **11**, 5553–5568.
Pudova,E.A. *et al.* (2018) HK3 overexpression associated with epithelial-mesenchymal transition in colorectal cancer. *BMC Genomics*, **19**, 5–13.
Raudvere,U. *et al.* (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, **47**, W191–W198.
Repana,D. *et al.* (2019) The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. *Genome Biol.*, **20**, 1.
Ruan,S. *et al.* (2020) Integrin 4–targeted cancer immunotherapies inhibit tumor growth and decrease metastasis. *Cancer Res.*, **80**, 771–783.
Schulte-Sasse,R. *et al.* (2021) Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nat. Mach. Intell.*, **3**, 513–526.
Spada,S. *et al.* (2021) Fibronectin as a multiregulatory molecule crucial in tumor matrisome: from structural and functional features to clinical practice in oncology. *J. Exp. Clin. Cancer Res.*, **40**, 102.
Szklarczyk,D. *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
Tai,Y.-L. *et al.* (2016) Activation of focal adhesion kinase through an interaction with 4 integrin contributes to tumorigenicity of Colon cancer. *FEBS Lett.*, **590**, 1826–1837.
Tamborero,D. *et al.* (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, **29**, 2238–2244.
Tate,J.G. *et al.* (2019) COSMIC: the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **47**, D941–D947.
Tong,X. *et al.* (2019) Protein tyrosine kinase 2: a novel therapeutic target to overcome acquired EGFR-TKI resistance in non-small cell lung cancer. *Respir. Res.*, **20**, 270.
Vandin,F. *et al.* (2012) De novo discovery of mutated driver pathways in cancer. *Genome Res.*, **22**, 375–385.
Wei,T. *et al.* (2020) An efficient and easy-to-use network-based integrative method of multi-omics data for cancer genes discovery. *Front. Genet.*, **11**, 613033.
Wei-Feng,G. *et al.* (2019) A novel network control model for identifying personalized driver genes in cancer. *PLoS Comput. Biol.*, **15**, e1007520.
Weinstein,J.N. *et al.*; Cancer Genome Atlas Research Network. (2013) The Cancer Genome Atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
Yang,W. *et al.* (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, **41**, D955–961.
Zhang,N. *et al.* (2021a) Glycogen synthase kinase 3 inhibition synergizes with PARP inhibitors through the induction of homologous recombination deficiency in colorectal cancer. *Cell Death Dis.*, **12**, 1–18.
Zhang,T. *et al.* (2021b) Identifying driver genes for individual patients through inductive matrix completion. *Bioinformatics*, **37**, 4477–4484.