Computational and Structural Biotechnology Journal 21 (2023) 5028–5038

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Review article

Identifying cancer driver genes in individual tumours

Rhys Gillman [a,b], Matt A. Field [a,b,c,d], Ulf Schmitz [a,b], Rozemary Karamatic [e,f], Lionel Hebbard [a,b,g,h,*]

[a] Department of Biomedical Sciences and Molecular and Cell Biology, College of Public Health, Medical, and Veterinary Sciences, James Cook University, Townsville, Queensland, Australia
[b] Centre for Tropical Bioinformatics and Molecular Biology, James Cook University, Cairns, Queensland, Australia
[c] Immunogenomics Lab, Garvan Institute of Medical Research, Darlinghurst, New South Wales, Australia
[d] Menzies School of Health Research, Charles Darwin University, Darwin, Northern Territory, Australia
[e] Gastroenterology and Hepatology, Townsville University Hospital, PO Box 670, Townsville, Queensland 4810, Australia
[f] College of Medicine and Dentistry, Division of Tropical Health and Medicine, James Cook University, Townsville, Queensland, Australia
[g] Storr Liver Centre, Westmead Institute for Medical Research, Westmead Hospital and University of Sydney, Sydney, New South Wales, Australia
[h] Australian Institute for Tropical Health and Medicine, Townsville, Queensland, Australia

ARTICLE INFO

Keywords:
Cancer
Driver gene
Gene interaction network
Machine learning
Precision medicine

ABSTRACT

Cancer is a heterogeneous disease with a strong genetic component making it suitable for precision medicine approaches aimed at identifying the underlying molecular drivers within a tumour. Large scale population-level cancer sequencing consortia have identified many actionable mutations common across both cancer types and sub-types, resulting in an increasing number of successful precision medicine programs. Nonetheless, such approaches fail to consider the effects of mutations unique to an individual patient and may miss rare driver mutations, necessitating personalised approaches to driver-gene prioritisation. One approach is to quantify the functional importance of individual mutations in a single tumour based on how they affect the expression of genes in a gene interaction network (GIN). These GIN-based approaches can be broadly divided into those that utilise an existing reference GIN and those that construct *de novo* patient-specific GINs. These single-tumour approaches have several limitations that likely influence their results, such as use of reference cohort data, network choice, and approaches to mathematical approximation, and more research is required to evaluate the *in vitro* and *in vivo* applicability of their predictions. This review examines the current state of the art methods that identify driver genes in single tumours with a focus on GIN-based driver prioritisation.

1. Background

While the development of novel approaches to treating cancer has improved survival prospects for patients with some cancer types, other cancer types have witnessed dramatically increased mortality rates. For example, in Australia and the US, cancers of the liver, pancreas, thyroid, and uterus have been associated with steadily increasing mortality over the past four decades, and their mortality rates are projected to rise sharply into the future [1,2]. Unsurprisingly, these cancers are frequently inoperable and lack effective chemotherapeutic options. In the case of liver and pancreatic cancers, front-line chemotherapeutics increase life expectancy by only a matter of months, with the only potentially curative treatment option being surgical resection, for which less than 25% of patients are eligible due to co-morbidities, late

Abbreviations: CCG, Cancer Census Genes; CIViC, Clinical Interpretation of Variants in Cancer; CSN, Cell-Specific Network Construction; DEG, Differentially Expressed Gene; DFVS, Directed Feedback Vertex Set; DNA, Deoxyribonucleic Acid; FVS, Feedback Vertex Sets; GIN, Gene Interaction Network; IMC, Inductive Matrix Completion; KEGG, Kyoto Encyclopedia of Genes and Genomes; LIONESS, Linear Interpolation to Obtain Network Estimates for Single Samples; MDS, Minimum Dominating Sets; MMS, Maximum Matching Sets; MOSCATO, Molecular Screening for Cancer Treatment and Optimization; NCI PID, National Cancer Institute Pathway Interaction Database; NCI-MATCH, National Cancer Institute's Molecular Analysis for Therapy Choice; NCUA, Nonlinear Control Of Undirected Networks Algorithm; PCC, Pearson Correlation Coefficient; PCST, Prize-Collecting Steiner Tree; PPI, Protein-Protein Interaction; PRODIGY, Personalised Ranking Of Driver Genes Analysis; RWR, Random Walker with Restart; SCS, Single-Sample Controller Strategy; SSN, Single-Sample Network; STRING, Search Tool for the Retrieval of Interacting Genes/Proteins.

* Correspondence to: The Science Place, Bd. 142, 1 James Cook Drive, 4811 Townsville, Queensland, Australia.
E-mail address: lionel.hebbard@jcu.edu.au (L. Hebbard).

https://doi.org/10.1016/j.csbj.2023.10.019
Received 28 July 2023; Received in revised form 10 October 2023; Accepted 11 October 2023
Available online 13 October 2023

detection and progression [3–5]. These cancers are typically treated with either non-targeted cytotoxic chemotherapies, which include DNA synthesis and repair inhibitors, and topoisomerase inhibitors, or broad acting multi-tyrosine kinase inhibitors [6–8]. These drug classes function with little specificity by killing the overactive and rapidly-replicating tumour cells faster than normal body cells.

To direct chemotherapeutic treatment more specifically towards cancer cells, much effort has been directed to the development of molecularly targeted chemotherapeutics which target cells with a specific gene alteration. However, the widespread applicability of these therapeutics is impeded by the extensive heterogeneity observed in some cancer types. While targeted therapies are often effective in tumours containing the relevant mutations, the extremely heterogeneous landscape of individual tumours limit their effectiveness. For example, of the 6 most frequently recurring mutations in liver cancer, none are present in more than 31% of patients [9,10]. Reimand and Bader [11] defined this phenomenon as the long-tail hypothesis: cancer mutations are represented by a short list of frequently mutated genes, and a long tail of infrequently mutated genes which collectively constitute most driver mutations.

To address this, personalised approaches that match directed therapies with patients most likely to respond have become crucial for modern cancer therapy. Personalised medicine has many advantages, including improved medication efficacy, reduced side-effects, and cost reduction [12], though, as discussed below, some challenges remain [13]. A basic way of achieving more patient-centric treatment is to identify biomarkers that predict patient response to treatment and then stratify patients accordingly. While this approach has been successful in some cancer types [14,15], for many it has not. By example, for liver cancer, several major multi-omics studies have attempted to molecularly classify tumours [10, 16–18]. However, to date, there are no established biomarkers that reliably predict the response to any treatment [19,20], likely due to the limited actionability of the most recurrent mutations [21]. Furthermore, several large clinical trials have begun to assess the feasibility of true molecular biomarker-based therapy choices for cancer patients, but the results are, thus far, underwhelming. For example, the National Cancer Institute's Molecular Analysis for Therapy Choice (NCI-MATCH) has so far shown that actionable mutations were only present in 37.6% of patients [22], and two published arms of the study have not shown positive results [23,24]. The similarly aimed Molecular Screening for Cancer Treatment and Optimization (MOSCATO) trial also found less than 50% of patients carrying mutations that fit their list of targets, and overall, only 7% of patients benefited from the trial [25].

Evidently, precision care based solely on the presence of a genetic alteration is not sufficient. To circumvent this, further analysis identifying driver genetic alterations may improve these efforts. In short, tumours often harbour thousands of somatic mutations; however, only a small subset of these mutations, termed *driver mutations*, are responsible for driving cancer growth, while the remainder are called *passenger mutations* [26]. Moreover, given that driver mutations are indispensable for neoplastic growth, it is hypothesised that some are present in every intra-tumoural subclone [27]. Thus, driver mutations are the ideal target of precision care.

Problematically, most popular approaches to identifying driver genes work at a cohort-level, by example, MuSiC [28], MutSigCV [29], CHASM [30], HotNet2 [31], and are usually a reflection of mutation frequency. Once such drivers have become sufficiently well-characterised, they are deposited in databases such as the Clinical Interpretation of Variants in Cancer (CIViC) [32], Cancer Gene Census (CGC) [33], Network of Cancer Genes (NCG) [34]. However, it remains unclear whether these *canonical drivers* are drivers in every case, that is, if a mutation is responsible for driving cancer in one patient it may not necessarily drive cancer in a different patient. Indeed, a truly personalised approach to cancer treatment would benefit from the ability to identify drivers based on the effects they are causing in the individual patient.

The focus of this review will be on driver prioritisation methods which attempt to identify personalised drivers by combining genomic and transcriptomic sequencing data to evaluate how a genetic alteration is affecting patterns of expression across a gene interaction network (GIN). The limitations of these methods will be explored, along with some potential solutions to these limitations in the form of emergent methods of differential expression analysis without replicate data. Finally, some alternative machine-learning based approaches that do not rely on transcriptomic data will be briefly discussed. The literature discussed in this review was sourced from the PubMed and Web Of Science databases, primarily using the search terms: ("driver gene" OR "driver mutation") AND ("personalised" OR "precision" OR "individual" OR "patient-specific"). However, this is not a comprehensive systematic review, and instead the included publications represent the selection of papers that best fit the authors criteria of algorithms for the prioritisation of driver genes in individual patients.

2. Network-based driver prioritisation

The authors of DriverNet [35] pioneered the idea that the functional impact of mutations can be quantified by virtue of their effect on the expression of connected genes in a network. The essential concept is outlined in Fig. 1. If a mutation can be functionally linked with large scale changes in gene expression, then it is likely that this mutation is a driver of the phenotypic changes observable in that individual's cancer.

Network-based driver identification methods combine genomic sequencing data with transcriptomic data in the form of GINs. GINs are often visualised as a graph, which is made up of nodes or vertices which are the genes, and edges which indicate an interaction between genes. These edges may be binary (1 = connected, 0 = not connected) or weighted with confidence values or strengths of interaction. Additionally, if the direction of a regulatory relationship between two genes is known, this information can also be shown in the form of a directed graph (Fig. 2). Importantly, we can differentiate a GIN from a protein-protein interaction (PPI) network which only considers physical interactions.

The degree of a node in the network is the number of edges connected to it and importantly, it has been shown that driver genes tend to have a high degree [36,37]. Due to their goal to identify many connections with DEGs, driver prioritisation algorithms are often biased for selection of such high-degree nodes, a phenomenon known as centrality-bias. This presents a problem in detecting lower degree yet important nodes. Paradoxically, to account for this bias, algorithms sometimes penalise high-degree nodes, despite the likelihood that these nodes contain the driver genes of interest. Some, but not all algorithms attempt to account for this bias.

2.1. Network-based driver prioritisation using external reference networks

The methods in Table 1 are driver prioritisation algorithms that have applied this concept to the level of an individual patient, and each of them relies on the provision of an external reference network, by example, STRING [38], Reactome [39], KEGG [40], and NCI PID [41]. It should be noted that all these algorithms are designed for use with bulk RNA-seq data.

DawnRank [36] was one of the first algorithms designed to apply this framework to individual samples. DawnRank maps patient-specific mutations, tumour gene expression and normal gene expression data to a pre-defined GIN with directed and unweighted edges. Mutated genes are subsequently ranked based on their connectivity to differentially expressed genes in the network using a modified version of Google's PageRank algorithm [42]. This random-walk approach works iteratively, whereby at each iteration there is a chance to "walk" to the next node in the network, the probability of which is dependent on the degree of that node and the differential expression level (simply the absolute difference of log tumour *versus* normal expression). Thus, for an

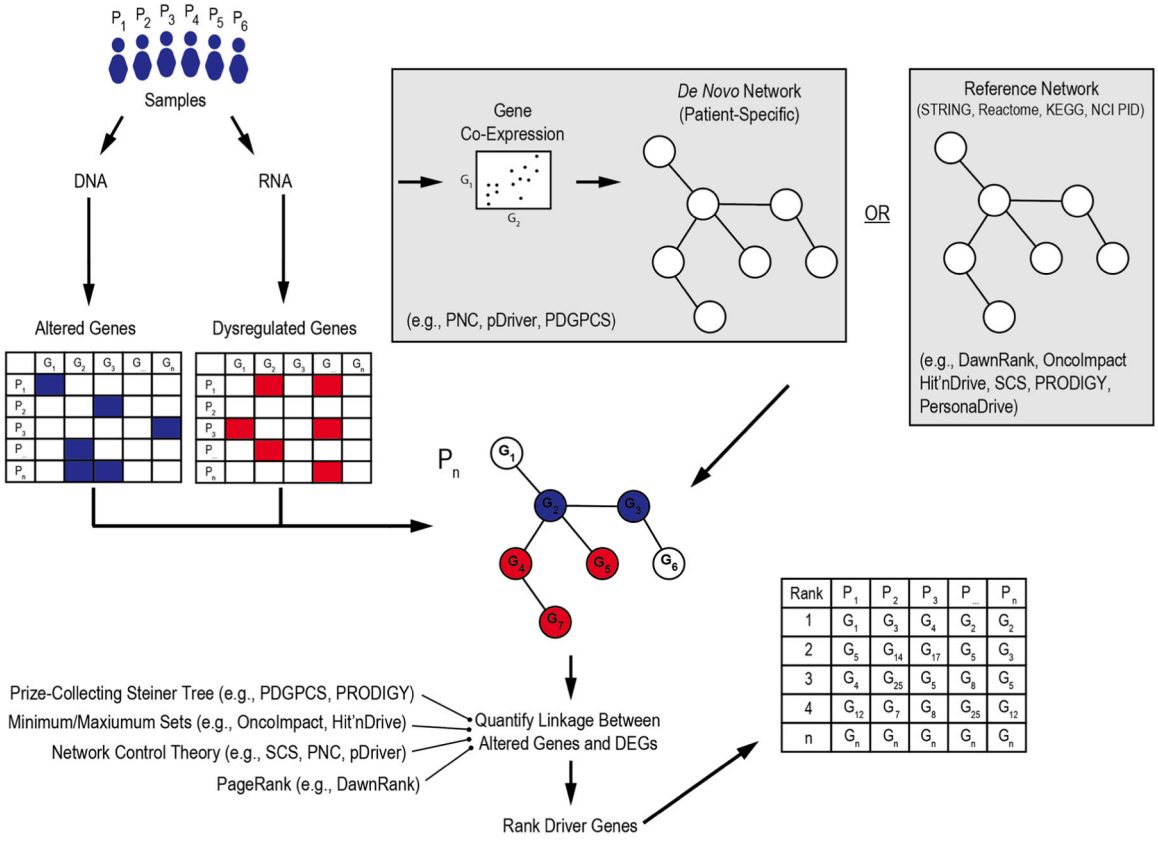

Fig. 1. Overview of network-based driver prioritisation. Given that driver mutations are responsible for creating a cancer phenotype, it is assumed that this phenotype must be the result of an alteration in gene transcription activity. Therefore, by combining mutation data with expression data on a gene interaction network (GIN), these algorithms quantify the likelihood of a mutation being a driver based on how it influences the expression of genes. To begin, genomic (DNA) and transcriptomic (RNA) sequencing data is acquired for each individual patient (P_n), and, using this information, genetic alteration (Blue) and expression dysregulation (Red) is identified for each gene (G_n). These altered and dysregulated genes are then mapped onto a GIN, which may be sourced from an external database (*e.g.*, STRING, Reactome, KEGG, NCI PID) or constructed *de novo* from transcriptomic data. The prioritisation algorithms then utilise various methods to quantify the linkage between altered genes and DEGs, and ultimately rank the altered genes by their overall impact on the network.

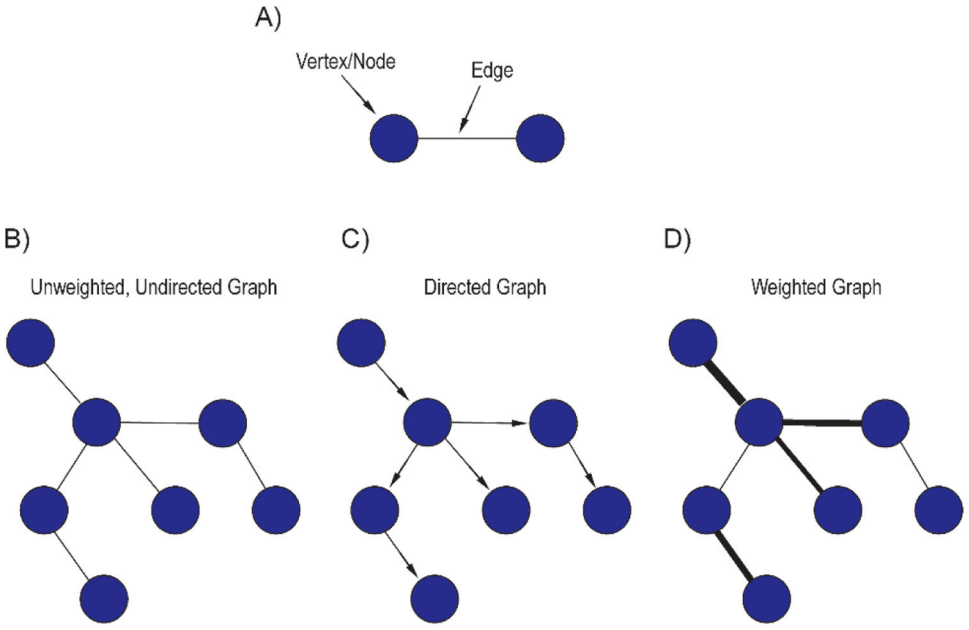

Fig. 2. The basic representations of gene interaction networks. A) The graph is constructed by joining vertices/nodes which represent genes by edges, which represent regulatory relationships. B) Undirected, unweighted graph where interactions are binary and it is unknown which direction, if any, the interaction occurs in. C) Directed graph, which indicates the direction of a relationship between genes. D) Weighted graph, where edges carry weights usually representative of their strength or confidence.

Table 1
Reference network influence algorithms. These programs evaluate the driver potential of a mutation based on its effect on a pre-defined reference network.

Software	Description	Use of Expression Data	Data Type	Primary Language	Year	Ref.
DawnRank	Ranks mutated genes based on connectivity to DEGs using Google's PageRank algorithm [46].	Quantitative	Paired- or unpaired-Tumour/ Normal	R	2014	Hou and Ma [36]
OncoImpact	Ranks mutated genes based on path length to frequently dysregulated genes and finds minimal set.	Binary	Tumour and reference healthy samples (external)	Perl	2015	Bertrand et al. [43]
Hit'nDrive	Finds a minimum set of mutated genes with maximal coverage of a user-defined fraction of DEGs	Binary	Tumour Only Data From Collection of Patients	C++	2017	Shrestha et al. [44]
Single Sample Controller Strategy (SCS)	Creates basic transition network based on log2 fold-change values and identifies minimal transition network controllers.	Binary	Paired-Tumour/ Normal	MATLAB	2018	Guo et al. [45]
Personalised Ranking Of DrIver Genes analYsis (PRODIGY)	Creates confidence-weighted subnetworks including mutated genes and dysregulated pathways and quantifies impact using the prize-collecting Steiner tree (PCST) problem.	Quantitative	Unpaired- Tumour/ Normal	R	2020	Dinstag and Shamir [37]
PersonaDrive	Creates bipartite networks of mutated genes connected to DEGs from the same or similar samples, and ranks mutated genes based on the number of pathways in which it connects with a DEG.	Binary	Tumour Only Data From Collection of Patients	Python	2022	Erten et al. [46]

individual patient, mutated genes which are connected to many dysregulated genes in a network are highly ranked. It should be noted that DawnRank's damping factor increases the priority of genes with a high degree, and there is no correction for non-driver high-connectivity genes, meaning it is likely susceptible to centrality bias.

OncoImpact [43] was another early method in this category which works similarly to DawnRank, though with more stringent filtering of predicted drivers. Rather than utilising the quantitative differential expression level of each gene, OncoImpact converts this information to a binary form (DEG or non-DEG), at the risk of losing valuable information. This information is then mapped to the unweighted directed reference network where dysregulated genes are considered "explained" if they have a short path to a given mutation or copy number alteration. "Phenotype genes" are defined as genes that are frequently dysregulated in the population of samples being tested (>5%), and mutations are ranked by the number of dysregulated phenotype genes they explain. Finally, true drivers are further filtered using the parsimony principle by creating a bipartite graph for patient-specific mutation-phenotype gene associations and finding the minimum number of mutations to cover a maximal number of phenotype genes using a greedy approximation algorithm. It should be noted that OncoImpact, in contrast to DawnRank, discards paths which travel through genes with a degree greater than a certain threshold, which is calculated systematically based on the dataset to account for centrality bias.

Hit'nDrive [44] considers the outliers of gene expression in the individual patient compared with other patients in the dataset as binary DEGs, and utilises the STRING network as an undirected, unweighted PPI network. Similar to DawnRank, Hit'nDrive uses a random-walk approach to calculate the "hitting time" or expected length of walk between two nodes and applies these as edge weights. The data is then split into a bipartite graph of mutated genes and DEGs, and the weighted multiset cover (WMSC) problem is applied to find the smallest subset of mutated genes that sufficiently influence a user-defined fraction of DEGs. As noted by the authors, Hit'nDrive does tend to identify drivers with high degree.

Single-Sample Controller Strategy software (SCS) [45] identifies driver genes by using network-control theory. Briefly, network-control algorithms attempt to find the minimal set of "control" nodes that transition a network from one state to another based on their connectivity with "target" nodes. SCS considers the transition between two phenotypic states, normal and tumour, and mutated genes are considered "controllers" of this transition. DEGs, defined by a log2 fold-change of +/- 1, are the target genes to be controlled, and these DEGs are indicated on a pre-defined GIN. Next, a Random Walk with Restart (RWR) algorithm is used to identify a sub-network around each mutated gene, which is taken to be the network of genes being controlled by the mutated gene. Finally, SCS identifies a minimum set of controllers which cover a maximal set of DEGs, and the mutations are ranked by summing up the overall confidence of edges in these sub-networks.

PRODIGY (Personalised Ranking Of DrIver Genes analYsis) [37] uses a global confidence-weighted undirected network of curated PPIs, meaning that each edge is assigned a value corresponding to the confidence of that interaction. Mutated genes and expression data are taken as input, and DEGs are identified using the R Bioconductor package DESeq2 [47] comparing an individual tumour against the provided normal samples. Dysregulated pathways (using Reactome, KEGG, NCI-PID) are also identified by their significant enrichment of DEGs. To calculate the influence of a mutated gene on a dysregulated pathway, a small subnetwork is constructed for every mutation with every dysregulated pathway. This subnetwork consists of nodes and edges from a mutated gene of interest, the genes in the pathway of interest, and all distance-1 neighbours of these nodes from the global network. Edge weights (costs) are assigned dependant on their confidence scores in this global network, and node weights (prizes) are assigned based on the level of differential expression. Finally, PRODIGY quantifies the influence of a mutated gene on each dysregulated pathway using a variant of the prize-collecting Steiner tree (PCST) problem, which aims to collect as many "prizes" (*i.e.* nodes) as possible, while paying the least cost, in essence to find the most reliable edges that connect a mutated gene with DEGs. Mutated genes are then ranked according to their overall influence score on a pathway. Because nodes with a high degree are assigned negative prize values, PRODIGY accounts for centrality bias.

The most recently developed tool in this category, PersonaDrive [46], builds on the above methods. PersonaDrive, like Hit'nDrive, does not require paired tumour/normal expression data or a reference dataset of normal expression. It does, however, require that the individual sample of interest be part of a relatively large collection of samples, whose information it utilises in a unique way compared with the other methods. DEGs are defined binarily based on their expression *versus* other tumours in the group, and a bipartite graph is constructed consisting of sample-specific mutated genes connected to DEGs that are not necessarily present in the given sample, however are present and connected to the same mutated gene in at least one other sample. In this way, PersonaDrive utilises the information from other samples in the cohort when finding patient-specific drivers. The edge weights of this bipartite graph are determined by the number of known pathways in which a connected mutated gene and DEG exist, and these are normalised by the similarity between the two samples from which the mutated

gene and the DEG are from. This sample similarity is defined by their overlap of DEGs. Finally, mutations in a sample are ranked by their overall edge weights, this being the overall influence of mutations on dysregulated genes in their corresponding pathways.

Currently, few comparisons have been made between these methods other than within the algorithm publications. The authors of PRODIGY noted that both SCS and DawnRank appear to perform similarly to pure measures of network centrality [37], potentially stemming from their lack of control for centrality bias. To the best of our knowledge, none of these tools have been evaluated for their ability to detect drivers *in vitro* or *in vivo*.

2.2. Network-based driver prioritisation using De-Novo networks

The previously described algorithms model gene interactions as a simplified static network, however, in reality such networks are dynamic. It is understood that gene regulatory networks can "re-wire" in different biological conditions [48,49], thus a normal cell and a tumour cell will each have their own unique regulatory networks. Additionally, many interactions are likely cell type-specific. To address this, some algorithms do not rely on a reference GIN but instead build *de novo* patient-specific transition networks to identify driver genes using network control strategies. Transition networks model the changes that occur between normal and tumour networks, where edges between gene pairs are the alterations found to significantly trigger a state transition. Overall, this approach requires two steps, generating patient-specific transition networks and then identifying the controllers of these networks, namely the most likely driver genes.

2.2.1. Creating a personalised transition network

Currently, three main approaches exist for the *de novo* creation of patient-specific transition networks (Table 2). Single-sample network (SSN) [50] works by comparing an individual sample against an independent reference cohort of control samples based on differential co-expression. First, the algorithm creates an aggregate network from the control samples using Pearson correlation coefficients (PCC_n) as edges. Next, the single sample of interest is added to this data and PCC values are recalculated (PCC_{n+1}), and the difference between these correlations is calculated (ΔPCC). Finally, the significance of this difference is determined by.

the distribution and standard deviation of ΔPCC values and by performing a z-test to determine inclusion in the SSN. Thus, the final SSN network shows pairs of genes with significant *differential correlation*, where a negative ΔPCC indicates a loss of correlation, and a positive ΔPCC indicates a gain of correlation. Importantly, the direction of the correlation (positive or negative) is no longer known. While the Pearson correlation is a common measure of association, it should be noted that this is a measure of linear association, while gene-regulatory relationships are often nonlinear [53].

Linear Interpolation to Obtain Network Estimates for Single Samples (LIONESS) [51] is another method for creating patient-specific networks which, similarly to SSN, determines the individual network based on perturbation of an aggregate network. However, LIONESS takes a group of tumour samples and creates a network for each individual sample, meaning that, unlike SSN, LIONESS identifies differences in single tumours *versus* other tumours, rather than comparing against healthy samples. Another difference is that LIONESS acts as a mathematical "wrapper" that is applied after user-specified edge-weights have been calculated, meaning that it is not limited to Pearson correlation coefficients. Firstly, an aggregate network is created based on all samples (n) in the cohort. Next, a single sample is removed from the aggregate network, and the edges are then recalculated (n-1). Because LIONESS considers the aggregate network as the linear average of all networks equally contributed to it, the perturbation of the network (n *versus* n-1) can be used to estimate the individual's network by finding the difference of the two, scaling them by the total sample size, and then adding the perturbed network back. This can be repeated for every individual in the cohort.

The third approach is the cell-specific network (CSN) [52], a tool originally designed for single-cell RNA sequencing data, which can also be applied to bulk RNA sequencing data. Like LIONESS, CSN takes a cohort of samples and produces a CSN for each sample. Briefly, to calculate the significance of an association between any pair of genes, a scatter plot is created for these genes and three boxes are drawn, one around the nearest neighbours of gene x, one around the nearest neighbours of gene y, and the third box is the intersection of the first two. By counting the number of points in each box, a statistical test can be performed which determines: i) whether there is an association between the two genes across the entire cohort, and ii) whether the expression of these genes in a given sample is significantly close to their expected relative values, given the association in the cohort. An advantage of this approach is that this allows CSN to identify nonlinear associations. An edge is drawn between two genes if the statistical value is above a given threshold, and therefore the edges in a CSN network are undirected and binary. This contrasts with LIONESS and SSN, wherein edges are weighted to indicate the strength of differential co-expression.

2.2.2. Identifying the controllers of a transition network

Network control theory is a highly developed mathematical theory used in many engineering applications to identify the nodes that can be altered to produce the transition from one state to another. In a biological context, although biological networks are large, there is presumed to be a relatively small subset of nodes that, if controlled, can drive the entire network to any given state. Driver nodes act in response to input signals, which, in the case of cancer-related networks, are genetic alterations [54]. Identifying such nodes is the goal of network control, and while the intricacies of network control algorithms are outside the scope of this review, their application in identifying the controllers in patient-specific transition networks will be briefly discussed below (Table 3).

The maximum matching sets (MMS) method [55] was one of the first network control methods applied to biological networks. The MMS

Table 2
Sample-specific network construction methods. These tools use patient-specific data to build patient-specific correlation networks between genes.

Software	Data Type	Measure of Association	Year	Ref.
Single-Sample Network (SSN)	Paired-Tumour/Normal	Linear	2016	Liu et al. [50]
Linear Interpolation to Obtain Network Estimates for Single Samples (LIONESS)	Tumour Only	Unspecified	2019	Kuijjer et al. [51]
Cell-Specific Network Construction (CSN)	Tumour Only	Linear and Nonlinear	2019	Dai et al. [52]

Table 3
Transition network-control methods. These methods are designed to identify the key "control" nodes in a network which are responsible for the transition from one state to another.

Software	Network Type	Year	Ref.
Maximum matching sets (MMS)	Directed-Network, Linear	2011	Liu et al. [55]
Minimum dominating sets (MDS)	Undirected-Network, Linear	2012	Nacher and Akutsu [56]
Directed Feedback Vertex Set (DFVS)	Directed-Network, Non-Linear	2017	Zanudo et al. [57]
Nonlinear control of undirected networks algorithm (NCUA)	Undirected-Network, Non-Linear	2019	Guo et al. [54]

method utilises directed-networks and assumes linear dynamics. Firstly, a maximum matching set of edges in the network is identified, in other words, the maximum number of edges in the network such that no two edges share a common vertex. Once this is achieved, any remaining unmatched vertices with direct paths from the input signal to the matched nodes are considered as the driver nodes. It should be noted that MMS is the only approach to network control that is not non-deterministic polynomial-time hard (NP-hard). This means that it is the only method that can be completed in polynomial time without requiring the use of approximations.

Like MMS, the minimum dominating sets (MDS) method [56] considers the network as a linear dynamic network. Unlike MMS however, MDS does not require a directed network, and instead assumes that all edges are bi-directional, and that nodes are able to independently control all their outgoing links. This has been suggested to result in higher costs, and therefore underestimation of controllability [54]. The goal of MDS is to identify drivers as dominating set of nodes where every node in the network is either within the dominating set or is adjacent to it.

Another method for determining the controllers in directed networks is the directed feedback vertex set (DFVS) approach [57]. This approach states that to drive a network to a desired endpoint, one needs to manipulate a set of driver nodes that intersects every feedback loop (cycle) in the network, that is, the feedback vertex set (FVS), as well as any source nodes, which are nodes without any incoming edges in the directed network. Identifying the minimal FVS is NP-hard, but a variety of algorithms exist to find close-to-optimal solutions.

Similar to DFVS, the nonlinear control of undirected networks algorithm (NCUA) [54] identifies transition network controllers based on FVS. In this algorithm, all edges are assumed to be bidirectional, and thus act as feedback loops. From the original network, NCUA constructs a bipartite graph, where the top node set is the original set of nodes, and the bottom node set is the edges of the original graph, and a minimum dominating set of nodes is selected which cover all edges in the graph. In a more recent publication, the authors built on this method using weight-NCUA to overcome the problem associated with their original method that can result in the identification of multiple possible driver node sets. Weight-NCUA takes edge weights of the original network into account, and instead of attempting to capture a maximum number of edges, it captures the maximum edge weights, and therefore is better able to find the optimal driver node set [58].

While we have discussed the creation of transition networks and the identification of network controllers as two separate steps, some methods combine these two analysis steps in one workflow for driver prioritisation (Table 4). For example, personalised network control (PNC) uses paired-SSN, followed by controller identification with NCUA [54]. pDriver on the other hand utilises the LIONESS algorithm to build the network, and finds drivers using MMS [59]. Unlike all the other algorithms discussed, pDriver also considers miRNA driver genes. Finally, PDGPCS [60] is a unique algorithm that utilises paired-SSN to create *de novo* patient networks, but then identifies controllers of this network using a very similar approach to PRODIGY. Namely, dysregulated pathways are identified using pathway enrichment of DEGs (based on a log2 fold-change threshold between paired tumour and normal samples), and then for each mutated gene and each dysregulated pathway, a PCST model is used to rank drivers using differential expression as a node prize and making the paired-SSN edge-weights inversely proportional to edge costs.

While some attempts have been made to assess and compare these methods, none of these progressed beyond *in silico* validation approaches, with no patient data benchmarked. Guo et al. [54] compared all the individual network construction and control methods in combination with each other. They found CSN and SSN to be superior to LIONESS, and the undirected network control methods (MDS and NCUA) performed better than the directed control methods for identifying driver genes in the CGC database. Importantly, the authors also identified that results were highly dependent on the network type (directed or undirected). Bhuva et al. [48] recently produced a framework for the evaluation of differential co-expression measures based on simulated network data, which is directly applicable to the measures used to generate patient-specific networks. They concluded that while Pearson correlation can be robust for analysis of differential co-expression, entropy-based methods may be a better alternative, particularly when dealing with lower sample sizes.

2.3. Limitations of network-based driver prioritisation

2.3.1. General limitations

The tools discussed in this section offer exciting prospects in this area. However, each method carries its own limitations. Already discussed throughout are the issues of centrality bias and the quantitative utilisation of expression data. OncoImpact and PRODIGY are the only two algorithms to directly account for centrality bias by penalising the selection of drivers with very high degrees. Additionally, many of the methods take expression data and convert this into binary classification of DEGs *vs* non-DEGs, which discards useful quantitative information that could be utilised in the analysis as is done by PRODIGY and DawnRank.

Another fundamental limitation of many of the reference-network based approaches is their use of PPI networks as a substitute for regulatory GINs. In extension, by considering the linkage of mutated genes with DEGs in a network, the implication is that the edges in this network are expected to indicate *regulatory* interactions. However, PPI networks do not necessarily indicate regulatory interactions. The most appropriate type of network would be a functional interaction network, which includes PPIs that activate or inhibit proteins, along with molecular interactions representing expression regulation. Unfortunately, such relationships are less well-known, and this means that these networks are usually far smaller.

2.3.2. Limitations of application

In addition to these limitations, there are other factors that affect the usability of these algorithms and the ability to compare them, contributing to poor adoption. For example, while all the algorithms aim to prioritise drivers in an individual patient, it is still a requirement that these individuals either be a member of a collection of patients, or that some external set of reference samples is available. This adds further complications, due to the need to account for batch effects in the data when integrating multiple cohorts, including differing sample preparation and composition, for example, stromal and immune cell content. In addition, PNC, Hit'nDrive, and the *de novo* methods do not rank their

Table 4
Driver prioritisation methods that utilise *de novo* networks.

Software	*De novo* Network Strategy	Driver-Prioritisation / Network Control Strategy	Data Type	Primary Language	Year	Ref.
Personalised Network Control (PNC)	Paired-SSN	NCUA	Paired-Tumour/Normal	MATLAB	2019	Guo et al. [54]
pDriver	LIONESS	MMS	miRNA and mRNA, Tumour Only	R	2021	Pham et al. [59]
PDGPCS	Paired-SSN	PCST	Paired-Tumour/Normal	MATLAB	2022	Zhang et al. [60]

results like the others but only provide lists of predicted drivers. If these lists are very long, confidence in the validity of individual drivers becomes low and it is difficult to compare the results against ranked methods. In this case, user-defined ranks could be designed based on factors such as network-degree of the drivers with DEGs. Additionally, lists of driver genes produced by all these algorithms do not give indication as to whether those drivers are over-active or underactive in the cancer. This is a crucial missing piece of information when using these predicted drivers to forecast the efficacy of target therapies.

Almost all approaches for driver-gene prioritisation involve the solution of NP-hard problems, meaning that the time taken to find the true underlying solution increases exponentially with input size. Because of this, most of the algorithms use heuristic approaches to approximate the solution, which usually give very good results with much lower computational complexity, but do not guarantee an optimal solution. Instead of using these approximations, Hit'nDrive and PNC formulate the driver prioritisation problems as Integer Linear Programming (ILP) problems which require mathematical optimisers, namely CPLEX [61] and Gurobi [62] respectively, in order to find optimal solutions using variants of the branch and bound algorithm. These optimisers are proprietary licenced software which limits user uptake.

2.3.3. Validation

So far, rigorous benchmarking studies using either *in vitro* or *in vivo* data to determine the efficacy of these tools are lacking. In general, the algorithms are created and then validated based on their ability to detect "known" driver genes from a "gold-standard" database. Indeed, a previous study had performed a comparison of many driver prioritisation tools, however performance was measured based on the successful prediction of 'canonical' drivers, and additionally did not focus on personalised methods [63]. The problem with this approach is that these driver genes do not act in isolation, and thus in a different genetic context in a different patient they may not carry the same potential to drive tumour growth. Additionally, this approach diminishes the potential to identify rare or novel drivers in a patient.

One way to potentially improve current validation approaches is by using simulated data such as that created by Bhuva et al. [48], which simulates the effect of genetic mutations under the user's control. However, such approaches often fail to generalise and result in over-fitted algorithms. Another approach, proposed by the creators of PersonaDrive [46], is to incorporate cell line drug-sensitivity data into the validation protocol. However, for these algorithms to progress to clinical use, their ability to identify driver genes *in vitro* or *in vivo* is specifically required.

2.3.4. Differential expression in single samples

As mentioned before, although these methods are intended to identify drivers in individual patients, some of them require a cohort of patients as input, and this is due to the requirements of identifying DEGs. The typical approach to identifying DEGs is to compare gene expression in paired tumour and normal samples. However, the commonly used differential expression analysis tools have been designed for cohort-level statistical analyses with a minimum requirement of three biological replicates. While useful for identifying DEGs at a cohort level, attempting to calculate the statistical significance of a change in expression of any gene in a single tumour requires a fundamentally different approach. Methods like DawnRank, OncoImpact, SCS, PRODIGY, and PersonaDrive attempt this through using a 1-*versus*-all approach, or by assigning a log2 fold-change cut-off. However, there are a variety of other approaches (Table 5).

Some tools designed for cohort-level DEG analysis have implemented features into their existing algorithms to deal with non-replicate data, including DESeq2 [47], NOISeq [64] and GFOLD [65]. These tools only work with paired tumour and normal samples from a single patient to make this comparison. To address the lack of biological replicates, DESeq2 assumes that most genes will not be DEGs, and, therefore, uses tumour and normal samples as their own replicates to calculate mean variance, which is then used to calculate statistical significance of individual genes. GFOLD and NOISeq-Sim attempt to simulate variance based on Poisson and multinomial distributions, respectively. These solutions, however, are only an estimate of technical replicates at best, and cannot give any indication of biological variance of gene expression.

RankComp [66] and, more recently PenDA [67], are two tools with similar approaches that identify DEGs based on their expression in rank order. By doing so, these tools improve the statistical power of their analyses because they consider a great number of genes rather than comparing single genes in isolation. Both methods require a collection of tissue-specific, normal reference samples to compare against a single tumour sample. RankComp converts the expression of genes in these reference samples into a rank-ordered list and performs a pairwise comparison of all genes to identify gene-pairs with stable ordering in normal samples, and subsequently identifies genes in tumour samples for which this stable ordering is reversed, using a Fisher's Exact test to find whether the gene is consistently up or down-regulated relative to its stable pairs. PenDA uses the reference samples to construct a local list of genes whose expression is higher or lower than a given gene. The rank-ordered list of genes in the tumour sample is reviewed and PenDA considers whether the local list of genes with higher or lower expression than a given gene has changed, and if this meets a statistical threshold. These approaches are advantageous in that they do not require matched normal samples for a patient. Of course, any time that external references samples are utilised in this way, systemic issues will likely arise due to batch effects. As such, neither of these tools have been validated *in vitro*, but instead only by their ability to detect DEGs *in silico* in controlled, simulated data. Regardless, it could be of interest to combine these approaches with the driver-prioritisation algorithms.

Table 5
Methods for identifying differentially expressed genes in individual samples.

Software	Description	Data Type	Year	Ref.
GFOLD	Combines fold change and statistical significance, assumes Poisson distribution in the absence of biological replicates and estimates uncertainty.	Paired-Tumour/Normal	2012	Feng et al. [65]
DESeq2	In the absence of biological replicates, assumes that most genes will not be differentially expressed, and uses the two conditions (tumour/normal) as their own replicates to calculate mean variance.	Paired-Tumour/Normal	2014	Love et al. [47]
RankComp	Creates ranked-ordered list of genes, looks for gene pairs with stable ordering across reference samples and then finds genes with reversed order.	Tumour Only	2015	Wang et al. [66]
NOISeq-Sim	Simulates technical replicates assuming a multinomial distribution. Only a simulation of technical replicates.	Paired-Tumour/Normal	2015	Tarazona et al. [64]
PenDA	Creates ranked-ordered list of genes, compares local ordering of a gene in a sample of interest *versus* a set of reference samples.	Tumour Only	2020	Richard et al. [67]

3. Machine learning-based driver prioritisation

Given the rise in prevalence of machine learning in the life sciences, one may expect to find this implemented in the field of driver-prioritisation. Indeed, this approach has shown promise for large

cohort data with several tools being recently developed (DriverML [68], DeepDriver [69], CHASM [30], and AI-Driver [70]). However, only a few machine learning approaches have been developed for personalised driver-identification (Table 6). These methods involve training models to recognise driver mutations based on features or descriptors of genes, and are advantageous compared with network-based approaches, because they often only rely on genomic data. Additionally, once the model has been trained, these algorithms can be applied to single samples.

For example, support vector machines (SVMs) are a popular supervised machine-learning method for binary classification. In essence, SVM systematically transforms data into higher dimensional spaces until it finds a space in which the two classes of interest (in this case, driver and non-driver) can be separated with a maximal margin by a hyperplane. iCAGES [71] is a SVM method trained on 11 ANNOVAR [75] annotation features, and requires both true positive and true negative annotation of driver genes for training, which is based on external annotation of canonical drivers. iCAGES is limited by the ability of its ANNOVAR features, which are almost all measures of sequence conservation, to differentiate drivers from non-drivers, and by the fact that true-negative drivers are paradoxically impossible to define when considering the issue of personalised, rare driver detection. sysSVM2 [72] is the updated version of sysSVM [76], which improves on the iCAGES method by utilising a semi-supervised one-class SVM classifier. Supplied with true positives only, one-class SVM can build a model that creates a boundary around true positives to separate them from the other data. Therefore, this removes the requirement to be able to label true negative driver genes. sysSVM2 is trained using 26 gene-based features including information about mutations, copy number, tissue expression, essentiality, network interactions, and genetic evolution.

Similarly, driveR [73] is another machine learning model trained on 26 genetic features including a metaprediction of mutation neutrality, copy number alteration, cancer-sepcific phenolyzer scores, COSMIC hotspot mutations and membership in cancer-related KEGG pathways. Instead of SVMs, driveR uses a supervised multi-task learning model for logistic regression to classify mutated genes as “driver genes” or “non-driver genes” in a cancer-type specific manner, unlike the previous two examples whose predictions are cancer-type agnostic. However, driveR suffers the same limitation as iCAGES in that both “true positives” and “true negatives” need to be defined for the training data.

An alternative machine learning approach is utilised by IMCDriver [74]. This tool applies inductive matrix completion (IMC), to identify candidate driver genes in an individual. Matrix completion techniques attempt to impute missing values in a matrix, however, these techniques are unable to generalise to new samples that are not present in the training dataset, meaning they are transductive. The purpose of IMC is to introduce external information about similarity between items in the matrix, which will allow the model to infer missing values when a new sample is added to the matrix, making it inductive. In this case, IMC is employed to address the incompleteness of known driver genes. Nonetheless, IMCDriver is limited by its approach in obtaining external similarity information. Since it is often prohibitively difficult to acquire high quality external similarity information, methods have been proposed to allow inductive matrix completion using only the data available in the matrix [77]. This is the approach utilised by IMCDriver, which puts emphasis upon and defines functional similarity of genes by their co-mutation. In reality, genes of interest are often the ones that are mutated mutually exclusively. Genes that are functionally similar are rarely mutated together in a patient [78]. It would be of great interest to build on this method in future by incorporating a more suitable measure of gene-similarity, such as co-expression. Finally, it should be noted that IMCDriver only considers genes that are mutated in at least one sample in the training group, thus limiting its ability to detect rare or novel drivers.

Table 6
Driver prioritisation methods that utilise machine learning.

Software	Learning Model	Training Features	Primary Language	Year	Ref.
iCAGES	Support Vector Machine (SVM)	11 ANNOVAR mutation annotations	Perl	2016	Dong et al. [71]
sysSVM2	One-Class Support Vector Machine (SVM)	26 Features including: ANNOVAR mutation annotations, copy number, essentiality, tissue expression, genetic evolution and network interaction	R	2021	Nulsen et al . [72]
driveR	Lasso Regression Multi-Task Learning (MTL)	26 Features including: mutation annotation metaprediction, copy number, hotspot mutations, tissue-specific Phenolyzer score, KEGG cancer pathway membership	R	2021	Ulgen and Sezerman [73]
IMCDriver	Inductive Matrix Completion (IMC)	No external training features. Trained using similarity between samples (shared mutated genes) and similarity between genes (co-mutation across samples)	Python	2021	Zhang et al. [74]

3.1. Limitations of machine learning-based driver prioritisation

Taken together, the identification of driver genes using machine-learning approaches has previously been considered an ineffective approach [12]. First, given the reliance on gold-standards, all these approaches are only as reliable as the available sets of canonical drivers. Additionally, gene-expression data is highly variable, which can greatly affect the accuracy of the model, which is likely why current machine-learning implementations do not utilise this data. Furthermore, machine learning often requires very large datasets. Also, deep-learning and machine-learning approaches are often criticized for being “black-box” approaches, a challenge in large complex systems. In essence, as the training data becomes larger and more complex, as is the case for gene network data, the biological interpretation of the results becomes convoluted. It therefore becomes nearly impossible to understand the underlying logic used by the machine learning algorithm to identify driver genes. As such, Frohlich et al. [12] describe algorithms on a spectrum of interpretability, with machine-learning models on one end and fully mechanistic models on the other.

4. Considerations for performing driver gene prioritisation

The limitations and advantages of both the general categories of algorithms and the individual algorithms themselves have been broadly summarised in Table 7. Due to the lack of benchmarking and comparison of these algorithms that has been performed to date, and the limited

Table 7
Overall advantages and limitations of driver prioritisation approaches.

Driver Prioritisation Approach	Advantages	Limitations
Network-Based Driver Prioritisation Using External Reference Networks	• Utilises patient-specific changes in gene expression to evaluate the effect of a mutation in a patient-centric manner	• Requires genomic and transcriptomic information • Dependent on external reference network • Requires a cohort of patients for making comparisons which may need to be batch-corrected • Network-based approaches in general are susceptible to centrality bias
DawnRank	• Uses differential expression information quantitatively	• Requires paired tumour/normal expression data
OncoImpact	• Has measures to combat centrality bias	• Uses differential expression information qualitatively • Requires reference healthy expression data
Hit'nDrive	• Utilises integer linear programming • Requires only tumour expression data	• Requires additional licenced software (CPLEX) • Coded in proprietary, licenced language (MATLAB) • Uses differential expression information qualitatively
SCS		• Uses differential expression information qualitatively • Requires reference healthy expression data
PRODIGY	• Uses differential expression information quantitatively • Incorporates additional pathway information • Has measures to combat centrality bias	• Requires reference healthy expression data
PersonaDrive	• Requires only tumour expression data • Incorporates information from other similar samples • Incorporates additional pathway information	• Uses differential expression information qualitatively
Network-Based Driver Prioritisation Using De-Novo Networks	• Utilises patient-specific changes in gene expression to evaluate the effect of a mutation in a patient-centric manner • Does not rely on external reference networks	• Requires genomic and transcriptomic information • Requires a cohort of patients for making comparisons which may need to be batch-corrected • Network-based approaches in general are susceptible to centrality bias
PNC	• Utilises integer linear programming	• Requires additional licenced software (Gurobi) • Coded in proprietary, licenced language (MATLAB) • Requires paired tumour/normal expression data
pDriver	• Also considers miRNA drivers	• Requires miRNA expression data
PDGPCS	• Incorporates additional pathway information	• Coded in proprietary, licenced language (MATLAB) • Requires paired tumour/normal expression data
Machine Learning-Based Driver Prioritisation	• Only requires genomic information • Can theoretically be expanded to include more features • Once the model is trained, truly requires only a single patient	• Models are reliant on accuracy of "known driver gene" databases • "Black-box" approaches without mechanistic interpretability
iCAGES		• Training model requires true positive and true negative drivers • Only uses ANNOVAR annotations as training features
sysSVM2	• Only true positive drivers are required for training • Expanded list of training features	
driverR	• Expanded list of training features • Cancer-type specific	• Training model requires true positive and true negative drivers
IMCDriver		• Does not utilise any external similarity features • Similarity based on co-mutation

utilisation of these algorithms in *in vitro* and *in vivo* experiments, recommendations for researchers can only be based on theoretical capabilities. To this end, of the network-based approaches, PRODIGY and PersonaDrive incorporate the most information. Thus, this in theory gives them the best chances of identifying true drivers. Similarly, sysSVM2 and driveR are the machine learning models trained on the most varied features. Therefore, these should provide the best results from this category. However, clearly future work is required to validate these opinions.

4.1. Future directions

Given the growing recognition of a need for personalised approaches to cancer treatment, it is surprising to find that although a diverse range of elegant approaches to driver prioritisation have been implemented, they appear to remain in their relative infancy with several limitations to overcome. Inherent to all these approaches, is the limitation of statistical power when working with single samples. Moreover, as discussed, all the algorithms mentioned in this review that utilise transcriptomic information are designed for use with bulk-RNAseq data. With the recent rapid growth of single-cell sequencing technologies with ever-improving accuracy and depth capability, it seems inevitable that these technologies will revolutionise our ability to investigate changes in single patients with improved confidence, and thus new tools are required to harness these innovations. Additionally, it is worth noting the emerging appreciation for consensus methodologies in bioinformatics [79,80]. It is likely that the most accurate tool for identifying driver genes will be one that considers a consensus output from multiple diverse approaches, such as that used by ConsensusDriver [63] which combines driver ranks from several cohort-level methods.

5. Conclusion

In recent years, significant advances have been made in the field of

cancer driver-prioritisation, but less attention has been assigned to methods which focus on single tumours. This review discussed some of the computational approaches designed for this purpose, most of which are network-based approaches that combine genomic and transcriptomic sequencing data, but also include machine-learning approaches which require only genomic data. While each new method builds upon the last and attempts to address their inherent limitations, minimal independent evaluation of these methods exists regarding whether the drivers predicted are biologically meaningful *in vitro* and *in vivo,* and whether they can be therapeutically targeted. Given the potential of these methods to progress the realisation of personalised medicine for individual cancer patients, it is essential that such evaluations are performed.

Funding

James Cook University Postgraduate Research Scholarship (R.G.); Tour De Cure Postgraduate PhD Scholarship (RSP-379-FY2023, R.G.); Tropical Australian Academic Health Centre Limited - Research Seed Grant (SF000121, L.H.); Townsville Hospital and Health Service - Study Education Research Trust Account (THHSSERTA_RPG1 2023, R.K., M.F., L.H.); National Health and Medical Research Council Investigator Grant (#1196405, U.S.); Cancer Council NSW project grant (RG20-12, U.S.); Tropical Australian Academic Health Centre project grant (SF0000321, U.S.).

Author statement

In the following pages, we have given responses and/or outlined specific changes to the manuscript in response to each of the reviewers' comments. We feel the key concern held by all reviewers was that the review lacked recommendations for the reader as to which tool to use in their research. While an important consideration, this is largely due to the lack of comprehensive benchmarking studies of these tools to date in the literature on which to base any such recommendations. Regardless, we have now incorporated an additional section titled "Considerations for Performing Driver Gene Prioritisation" wherein we briefly discuss and present a summary table that highlights potential advantages and limitations of each tool. Other significant changes are the inclusion of additional algorithms for discussion. Changes made to the manuscript are coloured in red.

CRediT authorship contribution statement

Rhys Gillman: Conceptualization, Visualisation, Investigation, Writing – original draft, Funding acquisition. **Matt Field:** Supervision, Writing – review & editing. **Ulf Schmitz:** Supervision, Writing – review & editing. **Rozemary Karamatic:** Writing – review & editing, Funding acquisition. **Lionel Hebbard:** Supervision, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rhys Gillman reports financial support was provided by Tour de Cure Ltd. Lionel Hebbard reports financial support was provided by Townsville Hospital and Health Service. Lionel Hebbard reports financial support was provided by Tropical Australian Academic Health Centre.

Acknowledgements

We would like to acknowledge Casey Toft for his feedback and suggestions for the manuscript.

References

[1] Luo Q, Steinberg J, O'Connell DL, Grogan PB, Canfell K, Feletto E. Changes in cancer incidence and mortality in Australia over the period 1996-2015. BMC Res Notes 2020;13(1):561.
[2] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019;69(1):7–34.
[3] Giuliano K, Ejaz A, He J. Technical aspects of pancreaticoduodenectomy and their outcomes. Chin Clin Oncol 2017;6(6):64.
[4] Martin SP, Wang XW. The evolving landscape of precision medicine in primary liver cancer. Hepat Oncol 2019;6(2):HEP12.
[5] Taieb J, Prager GW, Melisi D, Westphalen CB, D'Esquermes N, Ferreras A, et al. First-line and second-line treatment of patients with metastatic pancreatic adenocarcinoma in routine clinical practice across Europe: a retrospective, observational chart review study. ESMO Open 2020;5:1.
[6] Jayarangaiah A, Sidhu G, Brown J, Barrett-Campbell O, Bahtiyar G, Youssef I, et al. Therapeutic options for advanced thyroid cancer. Int J Clin Endocrinol Metab 2019;5(1):26–34.
[7] Llovet JM, Ricci S, Mazzaferro V, Hilgard P, Gane E, Blanc JF, et al. Sorafenib in advanced hepatocellular carcinoma. New Engl J Med 2008;359(4):378–90.
[8] Yau CC, Leeds J. Managing inoperable pancreatic cancer: the role of the pancreaticobiliary physician. Frontline Gastroenterol 2022;13(e1):e88–93.
[9] Llovet JM, Montal R, Sia D, Finn RS. Molecular therapies and precision medicine for hepatocellular carcinoma. Nat Rev Clin Oncol 2018;15(10):599–616.
[10] Wheeler DA. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell 2017;169(7):1327–41. e23.
[11] Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol 2013;9:637.
[12] Frohlich H, Balling R, Beerenwinkel N, Kohlbacher O, Kumar S, Lengauer T, et al. From hype to reality: data science enabling personalized medicine. BMC Med 2018;16(1):150.
[13] Field MA. Bioinformatic challenges detecting genetic variation in precision medicine programs. Front Med (Lausanne) 2022;(9):806696.
[14] Hayward NK, Wilmott JS, Waddell N, Johansson PA, Field MA, Nones K, et al. Whole-genome landscapes of major melanoma subtypes. Nature 2017;545(7653):175–80.
[15] Wilmott JS, Field MA, Johansson PA, Kakavand H, Shang P, De Paoli-Iseppi R, et al. Tumour procurement, DNA extraction, coverage analysis and optimisation of mutation-detection algorithms for human melanoma genomes. Pathology 2015;47(7):683–93.
[16] Boyault S, Rickman DS, de Reynies A, Balabaud C, Rebouissou S, Jeannot E, et al. Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. Hepatology 2007;45(1):42–52.
[17] Llovet JM, Villanueva A, Lachenmayer A, Finn RS. Advances in targeted therapies for hepatocellular carcinoma in the genomic era. Nat Rev Clin Oncol 2015;12(7):408–24.
[18] Wu Y, Liu Z, Xu X. Molecular subtyping of hepatocellular carcinoma: a step toward precision medicine. Cancer Commun (Lond) 2020;40(12):681–93.
[19] EASL. EASL clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol 2018;69(1):182–236.
[20] Marisi G, Cucchetti A, Ulivi P, Canale M, Cabibbo G, Solaini L, et al. Ten years of sorafenib in hepatocellular carcinoma: are there any predictive and/or prognostic markers. World J Gastroenterol 2018;24(36):4152–63.
[21] Llovet JM, Hernandez-Gea V. Hepatocellular carcinoma: reasons for phase III failure and novel perspectives on trial design. Clin Cancer Res: J Am Assoc Cancer Res 2014;20(8):2072–9.
[22] Flaherty KT, Gray RJ, Chen AP, Li S, McShane LM, Patton D, et al. Molecular landscape and actionable alterations in a genomically guided cancer clinical trial: national cancer institute molecular analysis for therapy choice (NCI-MATCH). J Clin Oncol 2020;38(33):3883–94.
[23] Krop IE, Jegede OA, Grilley-Olson JE, Lauring JD, Mitchell EP, Zwiebel JA, et al. Phase II study of taselisib in PIK3CA-mutated solid tumors other than breast and squamous lung cancer: results from the NCI-MATCH ECOG-ACRIN Trial (EAY131) Subprotocol I. JCO Precis Oncol 2022;6:e2100424.
[24] Chae YK, Hong F, Vaklavas C, Cheng HH, Hammerman P, Mitchell EP, et al. Phase II Study of AZD4547 in patients with tumors harboring aberrations in the FGFR pathway: results from the NCI-MATCH Trial (EAY131) Subprotocol W. J Clin Oncol 2020;38(21):2407–17.
[25] Massard C, Michiels S, Ferte C, Le Deley MC, Lacroix L, Hollebecque A, et al. High-throughput genomics and clinical outcome in hard-to-treat advanced cancers: results of the MOSCATO 01 Trial. Cancer Discov 2017;7(6):586–95.
[26] Greaves M, Maley CC. Clonal evolution in cancer. Nature 2012;481(7381):306–13.
[27] Yap TA, Gerlinger M, Futreal PA, Pusztai L, Swanton C. Intratumor heterogeneity: seeing the wood for the trees. Sci Transl Med 2012;4(127):127ps10.
[28] Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: identifying mutational significance in cancer genomes. Genome Res 2012;22(8):1589–98.
[29] 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.
[30] Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res 2009;69(16):6660–7.
[31] 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 2015;47(2):106–14.

[32] Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet 2017;49(2):170–4.
[33] Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. Nat Rev Cancer 2018;18(11):696–705.
[34] Repana D, Nulsen J, Dressler L, Bortolomeazzi M, Venkata SK, Tourna A, et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol 2019;20(1):1.
[35] 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):R124.
[36] Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. Genome Med 2014;6(7):56.
[37] Dinstag G, Shamir R. PRODIGY: personalized prioritization of driver genes. Bioinformatics 2020;36(6):1831–9.
[38] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47(D1):D607–13.
[39] Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, et al. Reactome pathway analysis: a high-performance in-memory approach. BMC Bioinforma 2017;18(1):142.
[40] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28(1):27–30.
[41] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the pathway interaction database. Nucleic Acids Res 2009;37(Database issue):D674–9.
[42] Page L., Brin S., Motwani R., Winograd T. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab; 1999.
[43] Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BK, Sia YY, 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.
[44] Shrestha R, Hodzic E, Sauerwald T, Dao P, Wang K, Yeung J, et al. HIT'nDRIVE: patient-specific multidriver gene prioritization for precision oncology. Genome Res 2017;27(9):1573–88.
[45] Guo WF, Zhang SW, Liu LL, Liu F, Shi QQ, Zhang L, et al. Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. Bioinformatics 2018;34(11):1893–903.
[46] Erten C., Houdjedj A., Kazan H., Taleb Bahmed A.A. PersonaDrive: A Method for the Identification and Prioritization of Personalized Cancer Drivers. bioRxiv. 2021: 2021.10.11.463919.
[47] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.
[48] Bhuva DD, Cursons J, Smyth GK, Davis MJ. Differential co-expression-based detection of conditional relationships in transcriptional data: comparative analysis and application to breast cancer. Genome Biol 2019;20(1):236.
[49] Pe'er D, Hacohen N. Principles and strategies for developing network models in cancer. Cell 2011;144(6):864–73.
[50] Liu X, Wang Y, Ji H, Aihara K, Chen L. Personalized characterization of diseases using sample-specific networks. Nucleic Acids Res 2016;44(22):e164.
[51] Kuijjer ML, Tung MG, Yuan G, Quackenbush J, Glass K. Estimating sample-specific regulatory networks. iScience 2019;14:226–40.
[52] Dai H, Li L, Zeng T, Chen L. Cell-specific network constructed by single-cell RNA sequencing data. Nucleic Acids Res 2019;47(11):e62.
[53] Jiang J, Lai YC. Irrelevance of linear controllability to nonlinear dynamical networks. Nat Commun 2019;10(1):3961.
[54] Guo WF, Zhang SW, Zeng T, Li Y, Gao J, Chen L. A novel network control model for identifying personalized driver genes in cancer. PLoS Comput Biol 2019;15(11):e1007520.
[55] Liu YY, Slotine JJ, Barabasi AL. Controllability of complex networks. Nature 2011;473(7346):167–73.
[56] Nacher JC, Akutsu T. Minimum dominating set-based methods for analyzing biological networks. Methods 2016;102:57–63.
[57] Zanudo JGT, Yang G, Albert R. Structure-based control of complex networks with nonlinear dynamics. Proc Natl Acad Sci 2017;114(28):7234–9.
[58] Guo WF, Zhang SW, Feng YH, Liang J, Zeng T, Chen L. Network controllability-based algorithm to target personalized driver genes for discovering combinatorial drugs of individual patients. Nucleic Acids Res 2021;49(7):e37.
[59] Pham VVH, Liu L, Bracken CP, Nguyen T, Goodall GJ, Li J, et al. pDriver: a novel method for unravelling personalised coding and miRNA cancer drivers. Bioinformatics 2021;37(19):3285–92.
[60] Zhang SW, Wang ZN, Li Y, Guo WF. Prioritization of cancer driver gene with prize-collecting steiner tree by introducing an edge weighted strategy in the personalized gene interaction network. BMC Bioinforma 2022;23(1):341.
[61] Laborie P, Rogerie J, Shaw P, Vilím P. IBM ILOG CP optimizer for scheduling. Constraints 2018;23(2):210–50.
[62] Gurobi Optimization L. Gurobi Optimizer Reference Manual. 2023.
[63] Bertrand D, Drissler S, Chia BK, Koh JY, Li C, Suphavilai C, et al. ConsensusDriver improves upon individual algorithms for predicting driver alterations in different cancer types and individual patients. Cancer Res 2018;78(1):290–301.
[64] Tarazona S, Furio-Tari P, Turra D, Pietro AD, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res 2015;43(21):e140.
[65] Feng J, Meyer CA, Wang Q, Liu JS, Shirley Liu X, Zhang Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. Bioinformatics 2012;28(21):2782–8.
[66] Wang H, Sun Q, Zhao W, Qi L, Gu Y, Li P, et al. Individual-level analysis of differential expression of genes and pathways for personalized medicine. Bioinformatics 2015;31(1):62–8.
[67] Richard M, Decamps C, Chuffart F, Brambilla E, Rousseaux S, Khochbin S, et al. PenDA, a rank-based method for personalized differential analysis: application to lung cancer. PLoS Comput Biol 2020;16(5):e1007869.
[68] Han Y, Yang J, Qian X, Cheng WC, Liu SH, Hua X, et al. DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies. Nucleic Acids Res 2019;47(8):e45.
[69] Luo P, Ding Y, Lei X, Wu FX. deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. Front Genet 2019;10:13.
[70] Wang H, Wang T, Zhao X, Wu H, You M, Sun Z, et al. AI-Driver: an ensemble method for identifying driver mutations in personal cancer genomes. NAR Genom Bioinform 2020;2(4). lqaa084.
[71] 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.
[72] Nulsen J, Misetic H, Yau C, Ciccarelli FD. Pan-cancer detection of driver genes at the single-patient resolution. Genome Med 2021;13(1):12.
[73] Ulgen E, Sezerman OU. driveR: a novel method for prioritizing cancer driver genes using somatic genomics data. BMC Bioinforma 2021;22(1):263.
[74] Zhang T, Zhang SW, Li Y. Identifying driver genes for individual patients through inductive matrix completion. Bioinformatics 2021.
[75] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38(16):e164.
[76] Mourikis TP, Benedetti L, Foxall 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.
[77] Zhang M., Chen Y. Inductive Matrix Completion Based on Graph Neural Networks2019 April 01, 2019:[arXiv:1904.12058 p.]. Available from: ⟨https://ui.adsabs.harvard.edu/abs/2019arXiv190412058Z⟩.
[78] Cisowski J, Bergo MO. What makes oncogenes mutually exclusive. Small GTPases 2017;8(3):187–92.
[79] Waardenberg AJ, Field MA. consensusDE: an R package for assessing consensus of multiple RNA-seq algorithms with RUV correction. PeerJ 2019;7:e8206.
[80] Field MA, Cho V, Andrews TD, Goodnow CC. Reliably detecting clinically important variants requires both combined variant calls and optimized filtering strategies. PLoS One 2015;10(11):e0143199.