Computational tools for prioritizing cancer driver candidates

Nathalie A.M. Nijs and Joanna von Berg[1]

[1] Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, 3584 CS Utrecht, the Netherlands

Keywords: Cancer driver candidates, prioritization, somatic mutations, computational tools

Abstract

Cancer is a very complex disease driven by DNA alternations, called somatic mutations. The identification of cancer-related genes and their contribution to the initiation and development of cancer is needed to make an accurate diagnosis and treatment. For this reason, high throughput DNA sequencing techniques are frequently applied, resulting in dozens or even hundreds of cancer driver candidate genes. However, identifying a small number of cancer driver mutation genes from a much greater number of passenger mutation genes is still highly challenging and the experimental validation of all these candidates would be too expensive and time-consuming. Therefore, multiple different cancer driver candidates prioritization tools are developed to tackle this problem. In this review, different computational tools to prioritize cancer driver candidates will be evaluated. The tools are divided into three different groups; methods based on literature, methods based on machine learning and network-based approaches. This review aims to inform the reader about the key features of different cancer gene prioritization tools. The reader can, based on this, select a tool that is suitable for their specific purpose and avoid possible pitfalls of the methods.

Layman's Summary

We all know that cancer is a leading cause of death worldwide. Cancer is a very complex disease driven by mutations in your genome. Differences in your DNA can lead to the initiation and development of cancer. Therefore it is needed to identify the cancer-related genes to make an accurate diagnosis and treatment. There are cancer driver mutation genes and passenger mutation genes. The cancer driver mutation genes initiate and develop cancer, while the passenger mutation genes do not affect functionality. There are only a small number of cancer driver genes in a big pool of passenger genes. Therefore, the identification of these driver genes candidates is challenging and the experimental validation of all these candidates is expensive and time-consuming. For this reason, multiple different cancer driver candidates prioritization tools are developed. In this review, different computational tools to prioritize cancer driver candidates will be evaluated.

Introduction

Cancer is a very complex disease driven by alternations to a cell's DNA. These alterations are also known as somatic mutations. A lot of research has been done to characterize somatic mutations in a patient tumor using next-generation sequencing technologies. Well-known projects of this kind are The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC). All this research results in more than 30.000 cancer whole exomes or genomes that have been sequenced and thousands of unique somatic mutations from a broad range of cancer types that are identified. Most

of the found somatic mutations have no biological effect nor phenotypical consequences. These mutations don't have a positive or negative selection during cell division and are so-called passenger mutations. On the other hand, among the somatic mutations, there is a small fraction called driver mutations. Such somatic driver mutations have a functional effect and also positive selection with respect to neighbouring cells. This leads to the exponential growth or survival of a clone. Although the discovery of new unique somatic mutations is going very rapidly, it is still challenging to distinguish passenger mutations from driver mutations.

There are multiple different somatic mutations; single-nucleotide variants (SNVs), indels and structural variants (SVs) including copy number alterations (CNAs) (Pleasance et al., 2010). Single-nucleotide variants (SNVs) are variants where a single nucleotide is substituted (Spencer et al., 2015). When an SNV is present in a protein-coding region, the nucleotide substitution can result in a translation of another amino acid, which can affect the protein structure and function. This is called a non-synonymous change or a missense variant. There are multiple codons for the same amino acid, which makes it possible that the amino acid does not change, even though there was a nucleotide substitution. This is called a synonymous change. The third change that can happen by a nucleotide substitution is the initiation of a stop codon, this mostly results in a non-functional protein and is also called a nonsense variant mutation.

Indels are small insertions or deletions from the genome, usually less than 50 base pairs (Mullaney et al., 2010). In the process of translation, three bases (= 1 codon) of the mRNA sequence are read at the time. A frameshift will occur when the length of the indel is not a multiple of three. This will rearrange the reading frame and all the remaining bases will be translated differently than intended. Resulting in a totally different product of translation, which most of the time leads to a targeted decay of the mRNA.

Structural variants (SVs) are large rearrangements of more than 50 base pairs in the genome (Feuk et al., 2006). This can be an insertion, deletion, duplication, inversion, translocation or even a combination of these kinds. A large subtype of SVs are copy number alterations (CNAs), a duplication or deletion that alternates the number of copies of sections of DNA in the genome (Taylor et al., 2008). This can change the expression level of genes in the affected genomic regions, causing tumor progression (Bhattacharya et al., 2020).

We have seen that different alterations in the genome can lead to changes in protein function. These mutations can cause a protein to lose its native function (loss of function, LoF) or it can obtain a new function (gain of function, GoF) (Schroeder et al., 2014) (Jung et al., 2015). Loss of function mutations disrupts the normal activity of a protein by reducing its expression level or by impairing its function. These mutations are often associated with tumor suppressor genes, which are genes that normally prevent cancer development by regulating cell growth and division. On the other hand, gain of function (GoF) mutations enhances the activity of a protein by increasing its expression level or by changing its function. These mutations are often associated with oncogenes, which are genes that promote cancer development by stimulating cell growth and division.

The identification of cancer driver genes is needed for targeted drug therapy. Already some cancer mutations can be targeted therapeutically. For instance, the oral drug Crizotinib, which is a tyrosine kinase inhibitor targeting the anaplastic lymphoma kinase (ALK) in ALK-positive lung cancer patients (Sahu et al., 2013). Another targeted drug is Sorafenib. This is a multi-kinase inhibitor that targets the extracellular signal-regulated kinases (ERK) and other pathways that are involved in cell proliferation in tumors (Shaw et al., 2013) (Pinter et al., 2009) (Dimitrakopoulos & Beerenwinkel, 2017). The

identification of more cancer driver genes will help us develop mutation-specific drugs and will improve the treatment of cancer.

There are multiple (good) tools available to discover cancer driver candidates, but how do we find out if these candidates are in practice actual cancer driver genes and not passenger genes? (Martínez-Jiménez et al., 2020) (Gonzalez-Perez & Lopez-Bigas, 2012) (Han et al., 2019) There are too many different cancer driver candidates to validate individually in the lab. An additional step, between the characterization of cancer driver candidates and the experimental validation, is needed. A step that can prioritize the cancer driver candidates in such a way that it is easier to select the ones that we want to validate experimentally.

In this review, different computational tools to prioritize cancer driver candidates will be evaluated. The tools are divided into three different groups; methods based on literature, methods based on machine learning and network-based approaches. (Fig. 1) This review aims to inform the reader about the key features of different cancer gene prioritization tools. So that the reader can, based on this, select a tool that is suitable for their specific purpose and avoid possible pitfalls of the methods. For this review, we focused on 'newer' available tools, in other words, tools that have been developed since 2017. This was done to eliminate outdated tools and to (hopefully) found more in-depth approaches.

Main

In this review, we evaluate tools that use different approaches but have the same goal: the prioritization of a list of cancer driver candidates. (Table 1) We categorized the tools into three groups: Literature-based tools, Network-based tools and Machine learning based tools. (Fig. 1) Sometimes a method does not fit perfectly into one group; it might have some overlap between the groups. Overall, it is helpful to have this categorization, because it is easier to follow this way and the tools are easier to compare.

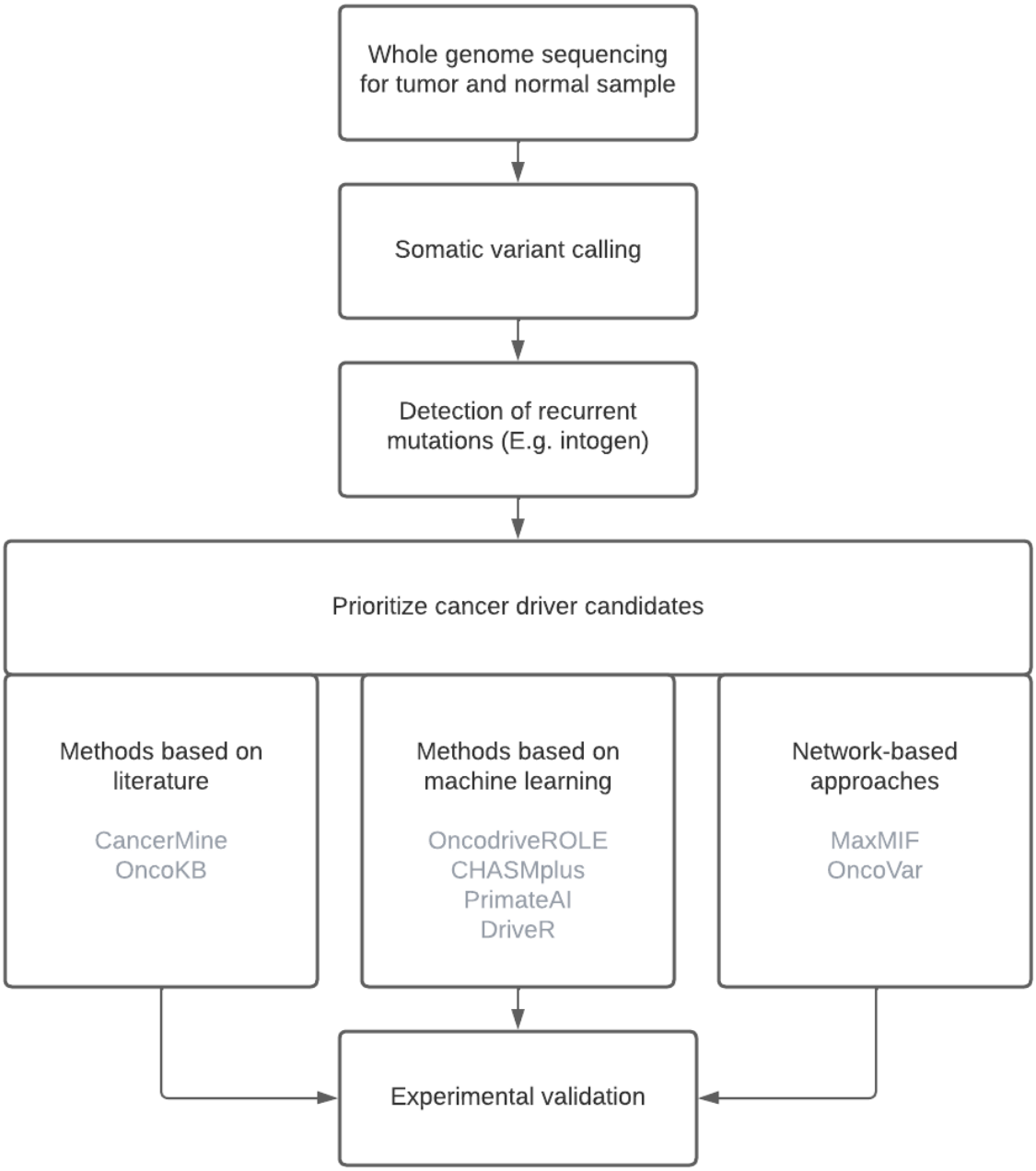

Figure 1. Process of detecting cancer driver genes. First, the tumor and normal samples need to be sequenced. Then somatic mutations, such as single-nucleotide variants, indels and CNVs are detected from the sequencing data using variant calling. From this list of variants, the passenger mutations need to be filtered out and the candidate cancer driver genes need to be detected. An often used tool for this is Intogen (Martínez-Jiménez et al., 2020). The next step is to prioritize these cancer driver candidates. They can broadly be categorized into three groups: Methods based on literature, Methods based on machine learning and Network-based approaches. The last step is to validate the findings experimentally.

Tool	Article	Accessibility	Year
Literature based tools			
CancerMine	Link (Lever et al., 2019)	Online tool	2019
Extracts various types of cancer-related information from scientific publications, creating a database of drivers, oncogenes and tumor suppressors in different types of cancer.			
OncoKB	Link (Chakravarty et al., 2017)	Online tool	2017
Summarizes all the already known cancer drivers from 7 different sources.			
Machine learning based tools			
OncodriveROLE	Link (Schroeder et al., 2014)	Available in intogen pipeline	2014
Predicts if cancer driver genes have a loss of function (LoF) or a gain of function (GoF).			
CHASMplus	Link (Tokheim & Karchin, 2019)	Available via OpenCRAVAT	2019
Calculates a score that represents the fraction of decision trees which vote for the mutation being a driver.			
PrimateAI	Link (Sundaram et al., 2018)	Available for academic and non-profit use	2018
Uses a deep neural network trained on non-human primate data to predict the functional impact of genetic variants.			
DriveR	Link (Ülgen & Sezerman, 2021)	Available via GitHub	2021
Uses a multi-task learning model with the combination of genomics information and prior biological knowledge.			
Network-based tools			
MaxMIF	Link (Hou et al., 2018)	Download via website	2018
Calculate a maximal mutational impact function score by integrating somatic mutation data and protein–protein interaction (PPI) data.			
OncoVar	Link (Wang et al., 2021)	Online tool	2020
Calculates a 'driverness' score based on a Gaussian mixture model (GMM).			
Tools specific for personalized use			
PDGPCS	Link (Zhang et al., 2022)	Available via GitHub	2022
Identifies and prioritizes personalized cancer driver genes based on the Prize-Collecting Steiner tree model.			
PersonaDrive	Link (Erten et al., 2022)	Available via GitHub	2022
Calculates a weighted 'pairwise pathway coverage' score to prioritize the mutated genes of a patient.			

Table 1. Overview of different computational approaches to prioritize cancer driver candidates. The tool name, reference, accessibility, year of publication and a brief description is given. The accessibility shows the easiest way of using the tools. Some tools are publicly available as an online platform, while others provide the source code to run the tool, for instance via GitHub.

Literature based tools

The first category is literature-based tools, which are tools that use information from the scientific literature to prioritize cancer driver gene candidates. A tool that summarizes all the already known cancer drivers is available, called OncoKB (Chakravarty et al., 2017). It uses information from various sources, such as different sequencing panels, the Sanger Cancer Gene Census and Vogelstein et al. (2013). In particular, the inclusion of the genes in 4 different sequencing panels was evaluated, named MSK-IMPACT™ (Cheng et al., 2015), MSK-IMPACT Heme™ (Ptashkin et al., 2019), FoundationOne®CDx (*FoundationOne CDx | Foundation Medicine*, z.d.), FoundationOne®Heme (Foundation Medicine, 2020). Besides that, they checked if the gene was part of the published Vogelstein et al. (2013) cancer gene list, which is a list of 125 'known' cancer driver genes identified by Vogelstein et al. and part of the Sanger Cancer Gene Census Tier 1, which included 576 genes. Based on this, the gene can be covered in 7 different sources and thus can have a score between 0 to 7 sources. The higher the number of sources for a particular gene, the more evidence there is that this gene is actually a cancer driver gene. Ranking the genes based on the number of sources seems to be a good way to start the prioritization of the cancer driver candidate genes.

Another literature-based tool is called CancerMine developed by the University of California, San Diego (UCSD) (Lever et al., 2019). CancerMine extracts various types of cancer-related information from scientific publications, creating a database of drivers, oncogenes and tumor suppressors in different types of cancer. This is done by a combination of rule-based and machine learning algorithms. First CancerMine preprocesses the input text by removing noise and irrelevant information, like formatting and citation information. The next step is to identify and extract mentions of relevant entities, such as genes, from the preprocessed text. Then CancerMine uses machine learning algorithms to identify relationships between those relevant entities, like gene mutations associated with specific cancer types. A logistic regression classifier was trained on word frequencies and description features, to learn the characteristics of sentences. After, a prediction of the pairs of genes and the associated roles in different cancer types was done using the classifier. To minimize the effect of incorrect associations, a high threshold on the classifier scores is applied. This will decrease the number of false positives, but also increase the number of false negatives, this is something to take for granted. Finally, the extracted information will be normalized by mapping entities and relationships to standard terminologies and ontologies, such as the Unified Medical Language System (UMLS), a repository of biomedical vocabularies developed by the US National Library of Medicine (Bodenreider, 2004). This results in a frequently updated database of drivers, oncogenes and tumor suppressors, available via a web viewer or download link.

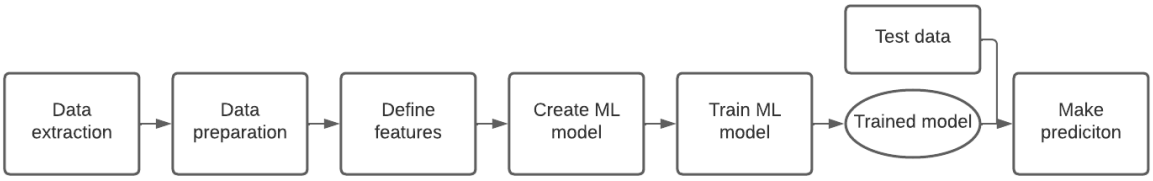

Figure 2. General overview of machine learning approaches. First, the data will be extracted and prepared. This involves removing irrelevant or duplicate data, handling missing values and transforming the data into the desired format. Then, the most relevant features are defined that help the model learn patterns and make accurate predictions. Now, the machine learning model can be created and trained. Once the model has been trained and tested, it can be deployed to make predictions on new, unseen data (test data).

Machine-learning based tools

In the previous section, we saw examples of machine learning being applied to literature data, but there are also tools that apply machine learning to mutation data as input by the user (Fig. 2).

OncodriveRole can predict if cancer driver genes have a loss of function (LoF) or a gain of function (GoF) (Schroeder et al., 2014). Loss of function mutations disrupts the normal activity of a protein by reducing its expression level or by impairing its function. These mutations are often associated with tumor suppressor genes, which are genes that normally prevent cancer development by regulating cell growth and division. On the other hand, gain of function mutations enhances the activity of a protein by increasing its expression level or by changing its function. These mutations are often associated with oncogenes, which are genes that promote cancer development by stimulating cell growth and division. OncodriveROLE predicts these LoF and GoF mutations based on several features, like the location of the mutation within the protein sequence, the potential impact of the mutation on the protein structure and function, and the potential effect of the mutation on the protein-protein interactions or signaling pathways. The output given is the probabilities per gene if it is a LoF or a GoF and a label assigned with Activating (GoF), Loss of Function (LoF) or No class. For this tool, the default threshold for labelling a gene Activating or Loss of Function is at a probability of >0.7. Nevertheless, it is possible to change this threshold, it depends on how strict of a classification the user wants for their list of cancer driver genes.

The tool CHASMplus is based on an earlier algorithm called CHASM (Cancer-specific High-throughput Annotation of Somatic Mutations), which was developed by the same research group (Carter et al., 2009) (Tokheim & Karchin, 2019). CHASM is a machine learning algorithm to analyze various features of somatic mutations, such as their location within the genome and the type of amino acid substitution. Eventually, it will assign a score to each mutation, reflecting its likelihood of being a driver mutation. In comparison to CHASM, CHASMplus includes some improvements, like the usage of more features and a machine learning-based approach that can incorporate multiple types of data. First, the user needs to provide the protein sequence and the somatic missense mutation of interest. Then CHASMplus will extract various features from the protein sequence and the mutation to use in the prediction model. These features include information on evolutionary conservation, physicochemical properties of the substituted amino acids, and the location of the mutation in the protein structure. Then a Random Forest classier is used to learn the relationship between the extracted features and the functional impact of the mutation. This classifier is trained on a large set of known driver mutations and neutral mutations. Now, CHASMplus can predict the functional impact of the somatic missense mutation by calculating a score that reflects the likelihood that the mutation is a driver mutation. The output is not only this likelihood score, but it gives also the location of the mutation in the protein and the conservation of the affected amino acid. Based on this likelihood, the cancer driver candidates can be ranked, from most likely to be a cancer driver gene to least likely.

Sundaram et al. (2018) demonstrated that missense variants that are common in other primate species are largely clinically benign in humans. For this reason, it is possible to systematically identify pathogenic mutations by the process of elimination. Their tool PrimateAI uses a deep neural network that was trained on a dataset consisting of ~380.000 common missense mutations from human and six non-human primate species (Sundaram et al., 2018). This training was done using a semi-supervised benign versus unlabeled training approach. Once the algorithm is trained, it can be used to predict the functional impact of genetic variants in your own genomic data. The algorithm takes in the genomic sequence of a variant and outputs the prediction of its functional impact, such as whether it is likely to be benign or pathogenic. This is in the form of a prediction score, where the founders of the tool recommend a threshold of > 0.8 for likely pathogenic classification, <0.6 for likely benign and between

0.6 and 0.8 for genes that could not be specified as either. For approximately 70 million human missense variants the prediction scores are already publicly available, which makes this tool very accessible.

DriveR does not only utilize somatic mutation information but also incorporates prior biological knowledge (Ülgen & Sezerman, 2021). The DriveR algorithm uses a two-step approach to prioritize cancer driver candidate genes. First, they employ a feature selection method to identify the most informative genomic features that distinguish diseased from healthy individuals. The features can include gene expression, protein-protein interactions and pathway information. The second step is to assign a score to each gene based on its association with the disease of interest, this is done using a Support Vector Machine (SVM). The higher the score, the more likely the gene is to be involved in the disease. The algorithm also incorporates a network-based approach to identify genes that are highly connected to other genes already known to be involved in cancer. It is possible to customize the analysis by selecting specific data sources or features to be included in the analysis and the weights that are assigned to each feature can be adjusted.

Network-based tools

The last tools that we saw are machine learning based. DriveR is a machine learning based tool that also incorporates a network-based approach. Besides this combination, there are also tools that mainly use network-based approaches. These are tools that use biological networks to evaluate the relationships between genes based on their interactions within biological pathways or networks.

One of these tools is called MaxMIF (Hou et al., 2018). MaxMIF consists of three steps (Fig. 3). First, based on somatic mutation data, a mutation score for each candidate driver gene for its role in driving cancer is computed. An earlier study showed that there are only a few driver mutation genes in comparison to the total number of mutated genes in the sample (Vogelstein et al., 2013). To avoid potential biases caused by samples with a large number of mutated genes, this approach ensures that each cancer sample contributes equally to the mutation score. This allows genes to be stratified based on their resulting mutation scores. The second step is calculating a mutational impact function (MIF) value for every candidate genes pair. Here, their mutational impact is measured according to their relationship in protein-protein interaction (PPI) networks. If they both have a high mutation score and are close to each other in the PPI networks, the two genes should have a strong mutational impact (Cheng et al., 2015). For that reason, the MIF value of two genes is defined to be proportional to the product of their mutation scores but inversely proportional to the square of the distance between them in PPI networks ($MIF(a,c) = \frac{M(a)M(c)}{r_{a,c}^2}$). Lastly, a novel maximal mutational impact function value for each candidate gene is computed by taking all its neighbours in the PPI networks into account ($S_{MaxMIF}(a) = \max_{j\in\{b,c,e,f,h\}} MIF(a,j)$. The candidate genes are then ranked based on their maximal mutational impact function value (MaxMIF value). The MaxMIF score represents the predicted functional impact of a mutation, where higher scores indicate a greater likelihood of the mutation causing a significant change in protein function. The maximal MaxMIF score (MaxMIF value) is the highest predicted score for a given mutation, among all possible amino acid substitutions at that position. Prioritizing the cancer driver candidates based on this MaxMIF value is helpful as mutations with high MaxMIF values are more likely to have a functional impact and may be more relevant to the disease.

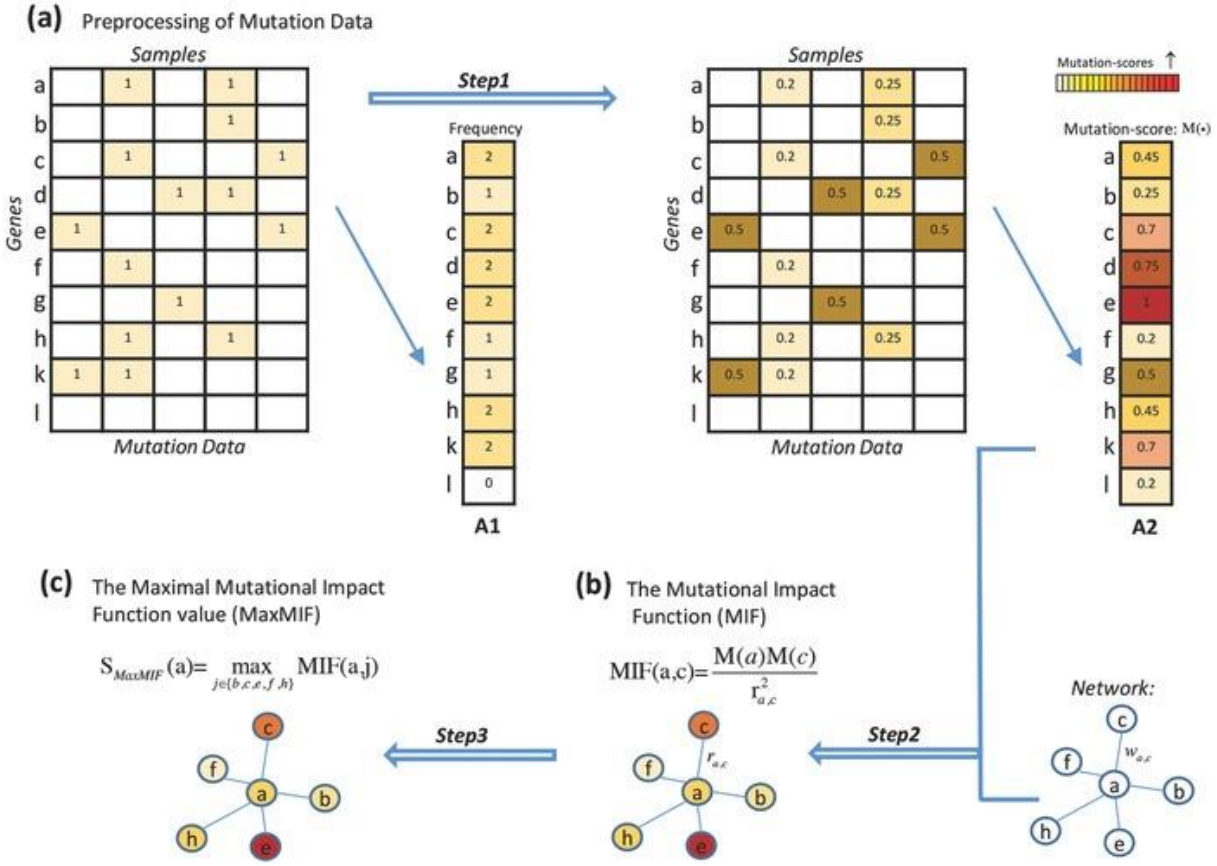

Figure 3. Visual representation of the MaxMIF approach. **a:** First the somatic mutation data matrix is standardized using the frequency of the samples per gene and the number of mutated genes per sample. Then, the mutation score of each gene is calculated by taking the sum of the rows. **b:** The mutational impact function (MIF) value of two candidate genes is computed as the product of their mutation scores divided by the square of the interaction distance between them in the PPI networks. **c:** The maximal MIF score for each candidate gene is calculated by taking all of its neighbors in the networks into account.
Adapted from Hou et al., 2018.

The tool OncoVar works quite differently, the approach calculates a scoring based on a Gaussian mixture model (GMM) (Wang et al., 2021). This score is called the 'driverness' of each mutation, gene and pathway. A probabilistic model is used to estimate the distribution of genomic features for driver and non-driver genes separately. OncoVar uses somatic mutations, copy number alterations, and gene expression data to calculate the driverness score. First, the likelihood of the gene's genomic features is calculated given the driver and non-driver GMMs separately. Then, the posterior probability of the gene being a driver gene is calculated using the Bayes' theorem. This takes the likelihood of a gene's genomic features given the prior probability of that gene being a driver into account. Eventually, the driverness score is the difference between the posterior probability and the prior probability of the gene being a driver gene. Genes with higher driverness scores are more likely to be cancer driver genes. OncoVar is incorporated into a user-friendly platform, where you can find the driverness score of more than twenty thousand mutations.

Tools for personalized use

The tools that we evaluated before, are tools that work with large cohorts. Other available tools use patient-specific data to discover personalized cancer driver genes. This is very helpful for providing diagnosis and target drugs for individual patients. PDGPCS (Personalized cancer Driver Genes based on

the Prize-Collecting Steiner tree model) is a tool that uses a network-based approach (Zhang et al., 2022). It integrates gene expression, mutation, copy number alteration, and protein-protein interaction data to predict driver genes specific to individual cancer patients. A personalized driver gene network is constructed for each patient using a Prize-Collecting Steiner tree model. This model aims to find the minimum subnetwork covering the maximum number of significant cancer-related genes. Then, the candidate driver genes are ranked based on their impact on the personalized driver genes network and their association with cancer-related pathways.

PersonaDrive is a patient-specific tool with three main steps (Erten et al., 2022). First, a personalized bipartite graph for each patient is constructed to model the relationship between the set of mutated genes and the differentially expressed genes (DEGs). Secondly, the edge weights in the bipartite networks are calculated. Eventually, the cancer driver candidates are ranked, based on the sum of the weighted 'pairwise pathway coverage' scores of all of the samples. Here, the coverage scores are normalized using the relevant pairwise patient similarity scores as weights. Interestingly, PersonaDrive takes whole cohort data into account, when it produces the personalized ranking for every individual patient. The degree of how much influence each individual patient data has on the other personalized driver ranking is determined using a pairwise patient similarity score, which is based on the amount of overlap between the set of DEGs of the pair.

Discussion and Conclusion

Studies from the Pediatric Cancer Genome Project (PCGP) have revealed that pediatric cancer differs from adult cancer in several ways (Downing et al., 2012) (Bandopadhayay & Meyerson, 2018). Unlike adult cancer, pediatric cancer has a lower rate of mutation and typically involves a single cancer-driving mutation (Kattner et al., 2019). In contrast, adult cancer usually involves multiple cancer drivers. While adult cancer shares certain mutational characteristics, pediatric cancer has its own unique set of mutational features. Furthermore, even within the same type of pediatric cancer, the frequency of a particular mutation can vary depending on the patient's age. For this reason, the applicability of the tools with regard to pediatric cancer needs to be discussed. The literature-based tools OncoKB and CancerMine contain mixed adult and pediatric cancer data. With the tool CancerMine it is possible to filter on specific childhood cancers. Most of the machine learning based tools are trained on adult data. Because of the differences between grown-up cancers and pediatric cancers, it logically seems that the outcome will differ if you train only on adult cancers, only on pediatric cancers or on both. The effects of using different training sets are not in the scoop of this literature review, but it is definitely something to keep in mind while considering the different tools for pediatric cancers. Network-based tools mostly use protein-protein interaction (PPI) networks, which are also applicable to pediatric cancers. For example, MaxMIF uses two independently developed PPI networks, HumanNet24 and STRINGv10.25 as a default (Hou et al., 2018). It is also possible to run MaxMIF with other networks, making it a tool that can be used for pediatric cancer data.

In the past twenty years, the biomedical field has seen significant progress in high-throughput technologies. As a result, there has been a constant increase in the amount of biomedical knowledge, the diversity of evidence sources, and the completeness of these sources. Therefore, there are now large and comprehensive data portals that collect gene and gene-disease associations for humans and other organisms. To continue improving the field of gene prioritization in the future, it is crucial to incorporate new or rarely used evidence sources and to enhance the quality of the available data.

We discussed two tools specific for personalized use. The use of specific patient data is very useful for the improvement of establishing a diagnosis and the development of targeted drugs specific to individual patients. The use of these tools seems to be a revolution in this field, but there is a downside. At the moment, there is a limited number of personalized omics data available. In the future, when it becomes easier to sequence the genome per patient, these tools could be very helpful.

There is a difference in how the various tools are presented by the developers. Some were described as a black box, while others were very open about the process behind their tool. Besides that, some tools use a web tool to visualize their results and scores per gene, while others only state the process and not the specific scores resulting from this process. This results in some tools being more user-friendly than others. You can go to the web tool, find the genes that you want to rank and save the scores per gene. A downside to this is that your gene of interest might not be included in the already-generated results, then you still need to use the tool itself to generate the scores. It was not in the scope of this review, but it would be interesting to compare these tools based on how many genes are incorporated in the already formed scoring schema and how likely it is to find your gene of interest in this list.

By the evaluation of the named prioritization tools, it was very helpful that the results and scores per gene were given for some tools. This made it easier to get a feeling of the tools. I would advise future developers of computational tools to include some kind of web interface, where the user interactively can play with the tool. Furthermore, a tool is way more appealing to use when there is a clear description of the input data and when the full source code is on GitHub. Ideally, a tutorial on how to use the tool would be available.

All in all, it is hard to point out which tool is better, given the scope of this literature review. To really say something about the quality and performance of the different tools, a more practical approach is needed. Testing all of the different tools on the same data is a possibility and some kind of scoring mechanism to compare the tools should be determined. This is something that could be done for further research.

References

Bandopadhayay, P., & Meyerson, M. (2018). Landscapes of childhood tumours. *Nature*, *555*(7696), 316–317. https://doi.org/10.1038/d41586-018-01648-4

Bhattacharya, A., Bense, R. D., Urzúa-Traslaviña, C. G., de Vries, E. G. E., van Vugt, M. A. T. M., & Fehrmann, R. S. N. (2020). Transcriptional effects of copy number alterations in a large set of human cancers. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-14605-5

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(Database issue), D267–D270. https://doi.org/10.1093/nar/gkh061

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., & Karchin, R. (2009). Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational Prediction of Driver Missense Mutations. *Cancer Research*, *69*(16), 6660–6667. https://doi.org/10.1158/0008-5472.CAN-09-1133

Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J., Rudolph, J. E., Yaeger, R., Soumerai, T., Nissan, M. H., Chang, M. T., Chandarlapaty, S., Traina, T. A., Paik, P. K., Ho, A. L., Hantash, F. M., Grupe, A., Baxi, S. S., Callahan, M. K., … Schultz, N. (2017). OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, *1*, 1–16. https://doi.org/10.1200/PO.17.00011

Cheng, D. T., Mitchell, T. N., Zehir, A., Shah, R. H., Benayed, R., Syed, A., Chandramohan, R., Liu, Z. Y., Won, H. H., Scott, S. N., Brannon, A. R., O'Reilly, C., Sadowska, J., Casanova, J., Yannes, A., Hechtman, J. F., Yao, J., Song, W., Ross, D. S., … Berger, M. F. (2015). Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *The Journal of Molecular Diagnostics*, *17*(3), 251–264. https://doi.org/10.1016/j.jmoldx.2014.12.006

Cheng, F., Liu, C., Lin, C.-C., Zhao, J., Jia, P., Li, W.-H., & Zhao, Z. (2015). A Gene Gravity Model for the Evolution of Cancer Genomes: A Study of 3,000 Cancer Genomes across 9 Cancer Types. *PLOS Computational Biology*, *11*(9), e1004497. https://doi.org/10.1371/journal.pcbi.1004497

Cosmic. (2022). *Cancer Gene Census*. http://cancer.sanger.ac.uk/census

Dimitrakopoulos, C. M., & Beerenwinkel, N. (2017). Computational approaches for the identification of cancer genes and pathways. *WIREs Systems Biology and Medicine*, *9*(1), e1364. https://doi.org/10.1002/wsbm.1364

Downing, J. R., Wilson, R. K., Zhang, J., Mardis, E. R., Pui, C.-H., Ding, L., Ley, T. J., & Evans, W. E. (2012). The Pediatric Cancer Genome Project. *Nature Genetics*, *44*(6), Article 6. https://doi.org/10.1038/ng.2287

Erten, C., Houdjedj, A., Kazan, H., & Taleb Bahmed, A. A. (2022). PersonaDrive: A method for the identification and prioritization of personalized cancer drivers. *Bioinformatics*, *38*(13), 3407–3414. https://doi.org/10.1093/bioinformatics/btac329

Feuk, L., Marshall, C. R., Wintle, R. F., & Scherer, S. W. (2006). Structural variants: Changing the landscape of chromosomes and design of disease studies. *Human Molecular Genetics*, *15*(suppl_1), R57–R66. https://doi.org/10.1093/hmg/ddl057

Foundation Medicine. (2020). *FoundationOne®Heme*. https://www.foundationmedicine.com/test/foundationone-heme

FoundationOne CDx | Foundation Medicine. (z.d.). Geraadpleegd 6 maart 2023, van https://www.foundationmedicine.com/test/foundationone-cdx

Gonzalez-Perez, A., & Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Research*, *40*(21), e169. https://doi.org/10.1093/nar/gks743

Han, Y., Yang, J., Qian, X., Cheng, W.-C., Liu, S.-H., Hua, X., Zhou, L., Yang, Y., Wu, Q., Liu, P., & Lu, Y. (2019). DriverML: A machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic Acids Research*, *47*(8), e45. https://doi.org/10.1093/nar/gkz096

Hou, Y., Gao, B., Li, G., & Su, Z. (2018). MaxMIF: A New Method for Identifying Cancer Driver Genes through Effective Data Integration. *Advanced Science (Weinheim, Baden-Wurttemberg, Germany)*, *5*(9), 1800640. https://doi.org/10.1002/advs.201800640

Jung, S., Lee, S., Kim, S., & Nam, H. (2015). Identification of genomic features in the classification of loss- and gain-of-function mutation. *BMC Medical Informatics and Decision Making*, *15 Suppl 1*(Suppl 1), S6. https://doi.org/10.1186/1472-6947-15-S1-S6

Kattner, P., Strobel, H., Khoshnevis, N., Grunert, M., Bartholomae, S., Pruss, M., Fitzel, R., Halatsch, M.-E., Schilberg, K., Siegelin, M. D., Peraud, A., Karpel-Massler, G., Westhoff, M.-A., & Debatin, K.-M. (2019). Compare and contrast: Pediatric cancer versus adult malignancies. *Cancer and Metastasis Reviews*, *38*(4), 673–682. https://doi.org/10.1007/s10555-019-09836-y

Lever, J., Zhao, E. Y., Grewal, J., Jones, M. R., & Jones, S. J. M. (2019). CancerMine: A literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nature Methods*, *16*(6), Article 6. https://doi.org/10.1038/s41592-019-0422-y

Li, Q., Ren, Z., Cao, K., Li, M. M., Wang, K., & Zhou, Y. (2022). CancerVar: An artificial intelligence–empowered platform for clinical interpretation of somatic mutations in cancer. *Science Advances*, *8*(18), eabj1624. https://doi.org/10.1126/sciadv.abj1624

Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., Gonzalez-Perez, A., & Lopez-Bigas, N. (2020). A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, *20*(10), Article 10. https://doi.org/10.1038/s41568-020-0290-x

Milbury, C. A., Creeden, J., Yip, W.-K., Smith, D. L., Pattani, V., Maxwell, K., Sawchyn, B., Gjoerup, O., Meng, W., Skoletsky, J., Concepcion, A. D., Tang, Y., Bai, X., Dewal, N., Ma, P., Bailey, S. T., Thornton, J., Pavlick, D. C., Frampton, G. M., … Vietz, C. (2022). Clinical and analytical validation of FoundationOne®CDx, a comprehensive genomic profiling assay for solid tumors. *PLOS ONE*, *17*(3), e0264138. https://doi.org/10.1371/journal.pone.0264138

Mullaney, J. M., Mills, R. E., Pittard, W. S., & Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Human Molecular Genetics*, *19*(R2), R131–R136. https://doi.org/10.1093/hmg/ddq400

Pinter, M., Sieghart, W., Graziadei, I., Vogel, W., Maieron, A., Königsberg, R., Weissmann, A., Kornek, G., Plank, C., & Peck-Radosavljevic, M. (2009). Sorafenib in unresectable hepatocellular carcinoma from mild to advanced stage liver cirrhosis. *The Oncologist*, *14*(1), 70–76. https://doi.org/10.1634/theoncologist.2008-0191

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordóñez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., … Stratton, M. R. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, *463*(7278), Article 7278. https://doi.org/10.1038/nature08658

Ptashkin, R. N., Benayed, R., Ziegler, J., Rema, A. B., Sadowska, J., Kiecka, I., Ho, C., Yao, J., Moung, C., Petrova-Drus, K., Nafa, K., Batlevi, C., Tallman, M., Levine, R., Giralt, S., Younes, A., Ladanyi, M., Berger, M., Zehir, A., & Arcila, M. E. (2019). MSK-IMPACT Heme: Validation and clinical experience of a comprehensive molecular profiling platform for hematologic malignancies. *Cancer Research*, *79*(13_Supplement), 3409. https://doi.org/10.1158/1538-7445.AM2019-3409

Sahu, A., Prabhash, K., Noronha, V., Joshi, A., & Desai, S. (2013). Crizotinib: A comprehensive review. *South Asian Journal of Cancer*, *2*(2), 91–97. https://doi.org/10.4103/2278-330X.110506

Schroeder, M. P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A., & Lopez-Bigas, N. (2014). OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. *Bioinformatics*, *30*(17), i549–i555. https://doi.org/10.1093/bioinformatics/btu467

Shaw, A. T., Kim, D.-W., Nakagawa, K., Seto, T., Crinó, L., Ahn, M.-J., De Pas, T., Besse, B., Solomon, B. J., Blackhall, F., Wu, Y.-L., Thomas, M., O'Byrne, K. J., Moro-Sibilot, D., Camidge, D. R., Mok, T., Hirsh, V., Riely, G. J., Iyer, S., … Jänne, P. A. (2013). Crizotinib versus Chemotherapy in Advanced ALK-Positive Lung Cancer. *New England Journal of Medicine*, *368*(25), 2385–2394. https://doi.org/10.1056/NEJMoa1214886

Spencer, D. H., Zhang, B., & Pfeifer, J. (2015). Chapter 8—Single Nucleotide Variant Detection Using Next Generation Sequencing. In S. Kulkarni & J. Pfeifer (Red.), *Clinical Genomics* (pp. 109–127). Academic Press. https://doi.org/10.1016/B978-0-12-404748-8.00008-3

Sundaram, L., Gao, H., Padigepati, S. R., McRae, J. F., Li, Y., Kosmicki, J. A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J., Xu, J., Batzoglou, S., Li, X., & Farh, K. K.-H. (2018). Predicting the clinical impact of human mutation with deep neural networks. *Nature genetics*, *50*(8), 1161–1170. https://doi.org/10.1038/s41588-018-0167-z

Taylor, B. S., Barretina, J., Socci, N. D., DeCarolis, P., Ladanyi, M., Meyerson, M., Singer, S., & Sander, C. (2008). Functional Copy-Number Alterations in Cancer. *PLOS ONE*, *3*(9), e3179. https://doi.org/10.1371/journal.pone.0003179

Tokheim, C., & Karchin, R. (2019). CHASMplus reveals the scope of somatic missense mutations driving human cancers. *Cell systems*, *9*(1), 9-23.e8. https://doi.org/10.1016/j.cels.2019.05.005

Ülgen, E., & Sezerman, O. U. (2021). driveR: A novel method for prioritizing cancer driver genes using somatic genomics data. *BMC Bioinformatics*, *22*(1), 263. https://doi.org/10.1186/s12859-021-04203-7

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, *339*(6127), 1546–1558. https://doi.org/10.1126/science.1235122

Wang, T., Ruan, S., Zhao, X., Shi, X., Teng, H., Zhong, J., You, M., Xia, K., Sun, Z., & Mao, F. (2021). OncoVar: An integrated database and analysis platform for oncogenic driver variants in cancers. *Nucleic Acids Research*, *49*(D1), D1289–D1301. https://doi.org/10.1093/nar/gkaa1033

Zhang, S.-W., Wang, Z.-N., Li, Y., & Guo, W.-F. (2022). Prioritization of cancer driver gene with prize-collecting steiner tree by introducing an edge weighted strategy in the personalized gene interaction network. *BMC Bioinformatics*, *23*(1), 341. https://doi.org/10.1186/s12859-022-04802-y