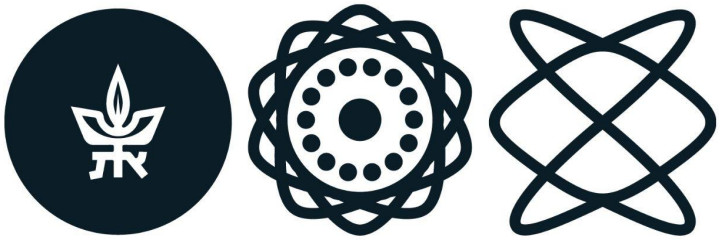


Tel-Aviv University

Raymond and Beverly Sackler Faculty of Exact Sciences

The Blavatnik School of Computer Science

Personalized phylogeny-guided detection of driver genes based on point mutations and copy number changes in the cancer genome

Thesis submitted in partial fulfillment of graduate requirements for

the degree "Master of Sciences" in Tel-Aviv University

School of Computer Science

By

Naama Kadosh

Prepared under the supervision of

Prof. Ron Shamir

June 2022

Acknowledgments

I would like to express my sincere gratitude to the people who accompanied me during this journey and supported me throughout the way.

First and foremost, I wish to thank Prof. Ron Shamir, my outstanding supervisor. It is a true honor to be mentored by a brilliant researcher with exceptional ideas and a kind heart. Thank you for the opportunity to join the computational genomics lab, for the professional guidance and for the persistence that kept me motivated, shaped me as a researcher and allowed me to broaden my horizons.

Secondly, I wish to thank my lab members: Lianrong, David, Tom, Dan C., Nimrod, Hagai, Hadar, Yonatan, Dan F., Omer, Eran, Ron and Maya, for all the help they have provided, for interesting brainstorms and for being true friends. In addition, I deeply thank Gilit Zohar-Oren for her administrative help, always accompanied with a smile.

I would like to thank the agencies that supported my research: Edmond J. Safra Center for Bioinformatics at Tel Aviv University, the Israel Science Foundation (grant 1339/18 and grant 3165/19 within the Israel Precision Medicine Partnership program), German-Israeli Project Cooperation DFG-DIP RE 4193/1–1 and Len Blavatnik and the Blavatnik Family foundation.

Last but not least, I wish to thank my dear family: my husband Mor, my parents Avi and Racheli, my siblings Yoav, Yael and Michal, my parents-in-law Yehudit and David and my brother-in-law Shai. You know that I would not have done this without your endless love and support.

Table of Contents

Abstract

Modern large-scale sequencing technologies have provided us with the ability to collect and analyze a great amount of data from cancer patients. Integration of transcriptomic and genomic data allows better understanding and detection of the factors that are responsible for or contribute to unusual cell proliferation. Driver gene mutations, which promote cancer, are accompanied by many passenger mutations, which have no effect on cell proliferation. Distinguishing between driver and passenger genes is a major challenge, and it is of high importance for understanding cancer mechanisms and for development of potential therapies.

Here we focus on personalized detection of driver genes based on several types of genomic alterations. We incorporate data of single nucleotide variations, copy number variations and genomic translocations in an attempt to pinpoint the most prominent factors that contribute to cancer progression. We also use the inferred evolution of mutations in the tumor of an individual to reweight detected driver genes. This allows us to generate phylogeny-supported results.

Mutation influence scores were computed using the PRODIGY algorithm. PRODIGY ranks the chance of genes with single nucleotide variations to be drivers by their total effect on pathways deregulation. Deregulation scores are computed with the help of the Prize Collecting Steiner Tree method. We extend this work by examining the influence of additional types of mutations.

In order to incorporate a tumor evolution aspect, we use individual phylogenetic trees of mutations deduced by the PhyloWGS algorithm. We use cancer cell fraction (CCF) of nodes in the trees, i.e., the fraction of cells that carry the mutations represented by each node. Mutations with higher CCF likely occurred earlier in the evolution. They are therefore more likely to be drivers and are given higher weights. The prioritization of genes with high CCF improves the precision and recall of detecting drivers.

1. Biological Background

1.1. Genes and proteins

The DNA is composed of two coiled strands with complementary sequences of four units called nucleotides: A for adenine, C for cytosine, G for guanine and T for thymine. It carries all the hereditary information and instructions for cell development, growth and function. A full copy of this information in the form of DNA molecules is present in virtually all human cells. Genes are the most basic hereditary units. These are short segments in the DNA that serve as templates for the creation of RNA molecules in a process called transcription. Some of them are called messenger RNA (mRNA) and are later translated into proteins. Proteins are the most basic functional units of the cell. They are made of building blocks called amino acids. Proteins catalyze metabolic reactions, perform DNA replication, respond to stimuli, transport molecules within the cell and more. During the translation process, mRNA nucleotide triplets (codons) are translated into amino acids, and chains of amino acids form a protein, starting with a start codon (AUG) and ending with a stop codon (UAA, UAG or UGA). U stands for the uracil nucleotide, which is the RNA substitute for thymine.

1.1.1. Mutations

Mutations are alterations in the DNA sequence. These alterations result from errors during DNA replication, mitosis, meiosis or are due to processes that cause damage to the cell such as exposure to ultraviolet radiation. While most mutations are corrected by cellular repair mechanisms, some persist and are carried over to progeny cells in cell divisions, or - for mutations in the germline - to the progeny organisms in future generations. Mutations that occur in non-germline cells are called *somatic mutations*. They may or may not trigger abnormal processes, depending on their landscape and character. For example, harmful mutations that occur within protein binding sites disrupt the transcription process and change

the amount of RNA molecules produced. Mutations that occur within the transcribed regions themselves may result in corrupted RNA products.

DNA mutations could be divided into three main categories [1,2]:

I. **Single nucleotide variations**

This group contains two types of variations:

A. **Single nucleotide substitutions** are changes in single nucleotides. When they occur in gene-coding regions, they can be synonymous or nonsynonymous. A synonymous substitution replaces a codon with another codon that encodes to the same amino acid. It results with no phenotypic effect. A nonsynonymous substitution replaces a codon with another codon that encodes to a different peptide. If it results with a stop codon, the translation terminates earlier than expected and the protein product could be truncated or nonfunctional. This is called a nonsense mutation. Otherwise, the mutation is called missense and the protein product contains an alternative amino acid. The phenotypic effect depends on its characteristics in comparison to the original protein.

B. **Short indels** are insertions and deletions of nucleotides in small amounts. When they occur in coding regions, since each successive triplet in the sequence forms one codon, these mutations might result in a change in the reading frame (partition into triplets) and cause a completely different translation. These are called frameshift mutations. Inframe mutations do not change the reading frame, but result with additional or fewer amino acids in the protein product. The effect depends on the changed amino acids characteristics.

II. **Copy number changes**

Normally, each gene occurs in two copies in the cell. One copy belongs to the

maternal chromosome and one to the paternal chromosome. Large deletions and duplications of genomic segments might contain whole genes. Deleted genes would have a copy number of 1 if the alteration is within one chromosome, otherwise they would be completely erased and have a copy number of 0. Duplications might increase the copy number of a gene to 3, 4 and even >100. These alterations occur often in cancer and they might substantially change transcription processes. **Figure 1** exemplifies deletion and duplication events.

III. **Translocations**

Translocations are particular genomic rearrangements of the chromosomal material. They occur when the DNA breaks in two locations and the broken segments are detached and fused back incorrectly. Unbalanced translocations result with segment loss, while reciprocal translocations occur when two chromosomes exchange segments. In some cases, several chromosomes could be involved. Translocations might affect the transcription process when they involve genes or related DNA regulatory elements. **Figure 1** exemplifies a translocation event.

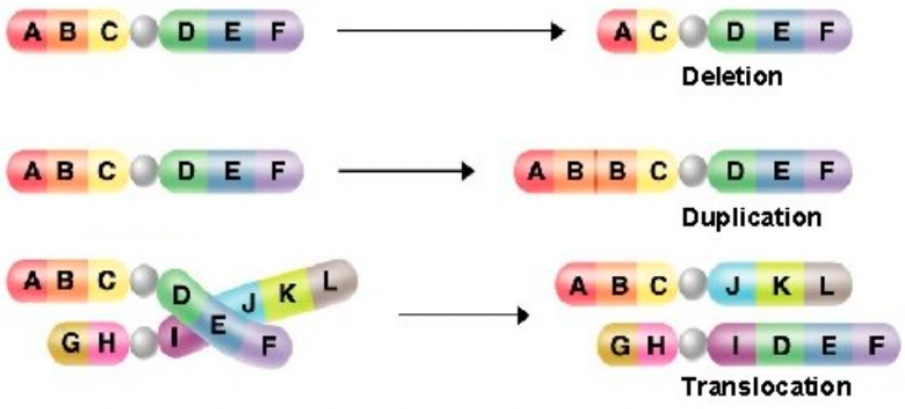


Figure 1: Types of genomic structural mutations: deletions, duplications, and ranslocations in a sample genome are shown (right) in comparison to the reference genome (left). Deletions and duplications result in copy number changes. Source: SlideToDoc, https://slidetodoc.com/chapter-12-dna-rna-i-dna-l-a/

1.1.2. Protein-protein interactions

Proteins tend to interact with each other in a highly specific manner in response to biochemical events [3]. When multiple proteins interact and bind, they often form large complexes that carry out molecular functions and mechanisms [4]. Protein-protein interactions (PPI) could be detected experimentally or predicted using computational methods. Their identification has led to the construction of descriptive networks in which nodes represent proteins and edges represent interactions [5]. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) [6] is a large database for human PPI with more than 4.5 million recorded interactions. It contains both directed edges that describe an effect of one protein on another and undirected edges that indicate interactions of two proteins. Each interaction is annotated with a confidence score that allows filtering uncertain connections by setting some threshold.

1.1.3. Biological pathways

Biological pathways describe the mechanisms by which the cell operates. These are networks of molecules that work together in the cell in response to some stimuli. Pathways could be described by the series of proteins that take part in them and the interactions among them. A pathway has a specific function, such as producing a metabolite or activating a target protein. Cellular processes might be disrupted when a pathway's compartment is incorrect or abnormally expressed.

Pathways could be categorized according to general roles or activation characteristics. The KEGG pathway database [7] identified pathways according to the following categories: metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases and drug development. Another popular source for pathways is Reactome [8]. It groups related reactions of proteins into pathways. **Figure 2** shows the P53 signaling pathway as an example from the KEGG website. It lists all

the proteins known to be involved in the pathway. This pathway is annotated as a cancer pathway under the human diseases category.

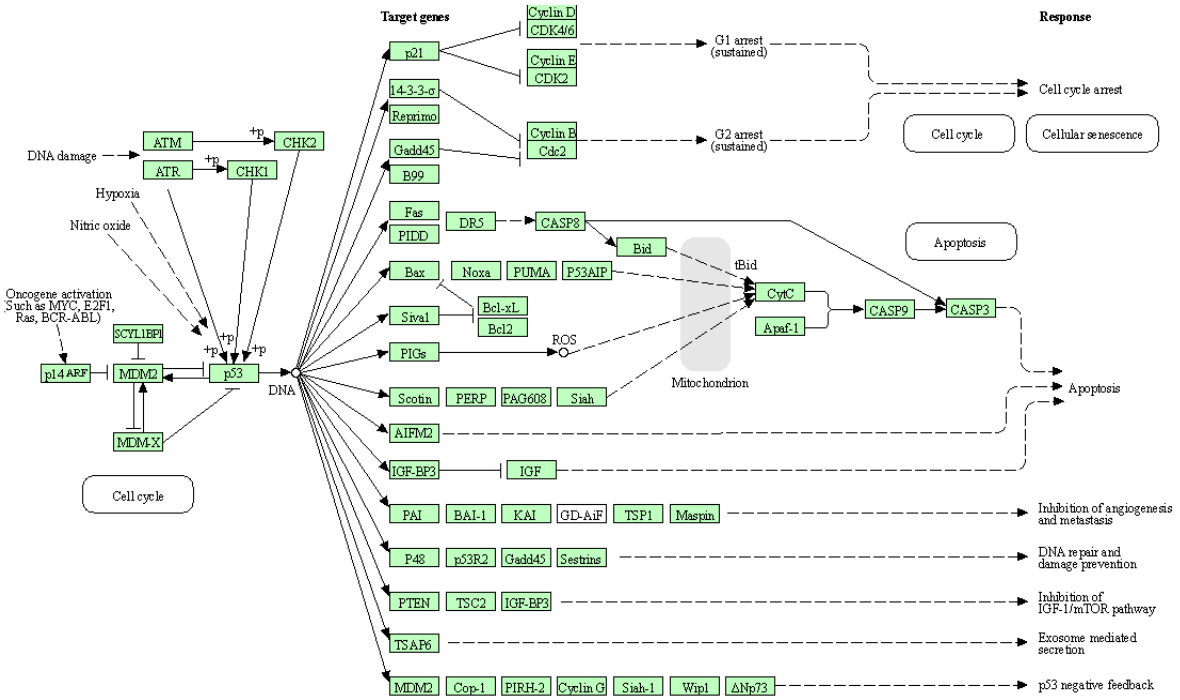


Figure 2: P53 signaling pathway as presented in the KEGG website [7]. p53 activation is induced by a number of stress signals, including DNA damage, oxidative stress and activated oncogenes. The p53 protein is employed as a transcriptional activator of p53-regulated genes. This activation results in three major outputs; cell cycle arrest, cellular senescence or apoptosis. Other p53-regulated gene functions communicate with adjacent cells, repair the damaged DNA or set up positive and negative feedback loops that enhance or attenuate the functions of the p53 protein and integrate these stress responses with other signal transduction pathways.
Source: https://www.genome.jp/entry/hsa04115

1.2. Cancer

Cancer is a complex and heterogeneous disease that originates from one of multiple tissue types in the body. It is a disruption within the cell that results with unsupervised and abnormal proliferation, leading to the formation of tumors that invade beyond normal tissue boundaries and metastasize to distant organs [9]. It is caused by genomic and epigenomic

chromosomal aberrations. According to the World Health Organization (WFO) analysis, cancer is the first or second leading cause of death before the age of 70 in 112 of 183 countries and ranked third or fourth in additional 23 countries as of 2019 [10].

1.2.1. Driver genes

Driver genes are genes that initiate or promote cancer progression upon mutations. Their alteration affects protein products and leads to dysfunction of crucial biological pathways. The disrupted pathways typically regulate three core cellular processes: cell fate, cell survival and genome maintenance [11]. As a result, the cell gains a selective and uncontrolled growth advantage. Typically, driver genes are overexpressed oncogenes or underexpressed tumor suppressors. For example, P53 (**Figure 2**) is a tumor suppressor, and when its function is diminished DNA damage correction is decreased and apoptosis of abnormal cells is less efficient.

Researchers have been using advanced sequencing techniques to characterize and detect abnormalities in cancer patient genomes over the last few decades. Most cancers show a phenomenon of *mountains* and *hills* of driver genes [11]. *Mountains* refer to driver genes that are frequently mutated across patients. This group is relatively small. *Hills* refer to less frequently mutated driver genes. They are observed in much larger numbers. Another group of driver genes is rare and spontaneous mutations that occur in individuals and cause similar cell proliferative effects.

Several projects maintain information on known cancer driver genes. Some of these genes were experimentally validated to be drivers, while others were not validated but are repeatedly observed in tumors. One of the early works is of Futreal et al [12], who listed 291 driver genes in different cancer types. More recent and up to date collections include the COSMIC Cancer Gene Census (CGC) [13] and the Network of Cancer Genes (NCG) [14], both listing about 600 genes (see “Gold standards” under “Methods and Results” for further

information). In all three sources, about 90% of the genes included show somatic mutations, 20% show germline mutations and 10% show both. Some examples of well understood drivers include the fusion of BCR and ABL genes that is caused by a translocation and drives chronic myelogenous leukemia, the duplication of EGFR gene that drives Glioma, and point mutations in TP53 that drive many cancer types.

Tumors usually contain two to eight driver genes [11] (See **Figure 3**). The majority of mutations in tumor cells are *passenger mutations* that do not drive or promote cancer. Distinguishing between driver and passenger mutations is an important mission that allows better understanding of cancer mechanisms and may facilitate personalized medicine treatments. Several driver gene detection methods have been developed recently, some perform analyses at a cohort level and some aim to predict patient specific drivers (see "Computational Background" section).

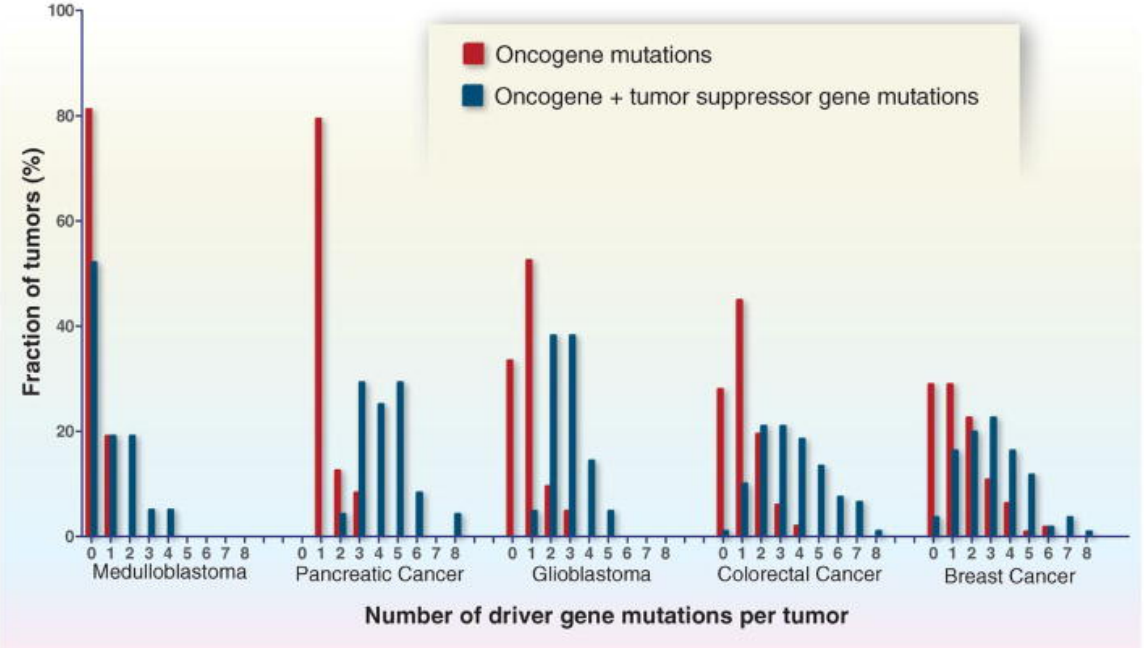


Figure 3: The total number of driver gene mutations in five cancer types, split to oncogene mutations alone and joint oncogene and tumor suppressor mutations. Source: Vogelstein, B.et al [11].

1.2.2. Evolution of cancer

All cancers are thought to follow a Darwinian evolution process [9]. Analogous to the origin of species theory, cancer develops by continuous acquisition of mutations in driver genes of individual cells followed by natural selection (**Figure 4**). The selection fosters cells that gain growth advantage and survive more effectively than their neighbors. Most mutations are passengers and do not affect cell growth. While most positively selected cells become benign growths such as skin moles, cancer cells gain a sufficient advantage that allows them to proliferate beyond normal boundaries and eventually form malignant tumors.

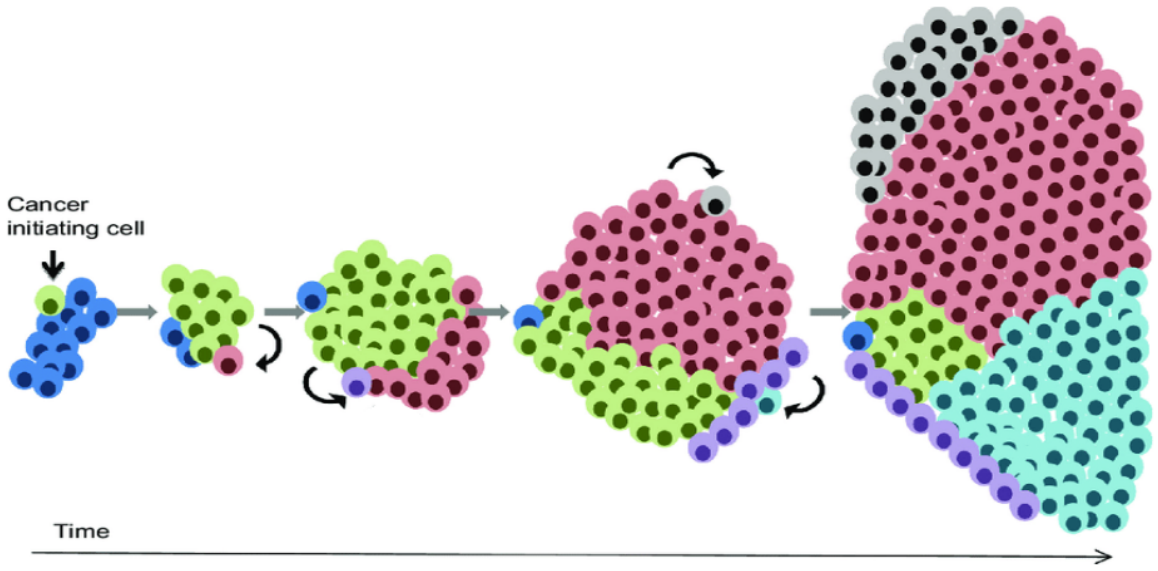


Figure 4: Clonal evolution of cancer cells. A schematic evolution of subclones from a single cell that acquired a growth advantage over its normal counterparts. The initial clone may produce distinct subclones through the course of the disease, originating due to further alterations. Here different subclones are in different colors, and mutated cells are indicated by arrows. Source: Raza et al. doi: 10.2147/AGG.S54184

Tumor evolution could be exemplified by the well studied colorectal cancer [11]. Most colorectal cancers are initiated by a mutation in the APC driver gene of a normal epithelial cell. The mutated cells form a slowly growing cell cluster called adenoma. When another

driver gene such as KRAS is mutated in one of the cells, a new rapidly growing subclone emerges. Cells that carry only the APC mutation may persist, but cells with both mutations outnumber them. Further mutations in genes such as PIK3CA, SMAD4 and TP53 are followed by clonal expansions, forming a malignant tumor that invades normal tissues and could metastasize to lymph nodes and distant organs such as the liver. See **Figure 5** for illustration.

Clonal mutations occur early in evolution. They are found in the vast majority of the tumor cells. Subclonal mutations evolve later and exist in a subset of the tumor cells. Tumors could be highly heterogeneous and contain multiple distinct subclones. Heterogeneity is present within a primary tumor, between two metastases and within metastatic lesions [11]. Analysis performed by Yachida et al. [15] shows that it is not uncommon for one metastasis of a pancreatic cancer to carry 20 mutations that are not shared with other metastases of the same cancer.

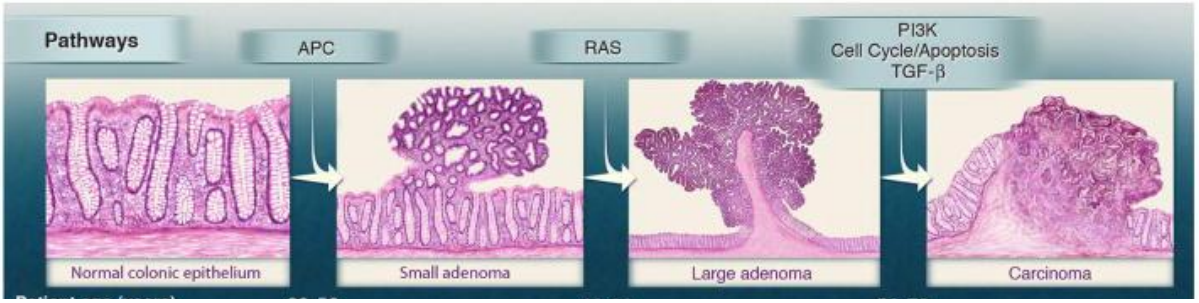


Figure 5: genetic alterations and the progression of colorectal cancer. The major signaling pathways that drive tumorigenesis are shown at the transitions between each tumor stage. One of several driver genes that encode components of these pathways can be altered in any individual tumor. TGF-β, transforming growth factor–β. Source: Vogelstein et al. [11]

2. Computational Background

2.1. Patient-specific driver detection methods

In this section we present four state of the art algorithms for driver gene detection in cancer. They operate in a patient-specific manner by processing data of individuals and outputting lists of predicted driver genes per patient.

Most methods for driver detection operate at a cohort level. They integrate data from all patients in a specific cancer cohort and compute a unified list of the most prominent driver genes. Since our focus here is personalized drivers, we only mention some of the leading cohort-level methods briefly. DriverNet [22] was one of the first cohort-level algorithms to integrate genomic alterations and gene expression data. It outputs a parsimonious set of mutated genes that are linked to genes with differential expression in a PPI network. MEMo [23] identifies small subnetworks of genes that are mutated in a mutually exclusive manner and belong to the same biological pathways. HotNet [24] uses a heat-diffusion process to detect small subnetworks with high frequency of mutated genes in a PPI network. These methods rely on statistical power and provide globally good prediction, but often fail to capture the accurate dysfunctioning processes in individuals. In particular, rare driver genes that promote cancer in a few patients could be missed. Better therapy requires a more delicate and personalized analysis that could be achieved with patient-specific methods.

2.1.1. DawnRank

DawnRank [25] identifies and ranks driver genes in individuals by the extent to which they perturb downstream genes in a PPI network, using an idea from Google's PageRank algorithm. One of its underlying assumptions is that driver genes tend to have higher connectivity in the network. The input data are a directed PPI network, a list of genes with somatic mutations in the individual's tumor, and a differential gene expression for the

individual, showing the difference in expression profile between the normal and the tumor sample.

Let *N* be the number of nodes and let *A* be the adjacency matrix of the network, such that $A_{ji} = 1$ if there is an edge from *j* to *i*, and 0 otherwise. For each gene *j*, Let f_j be its absolute log2-fold change in expression in the tumor and deg_i be its in-degree. Similarly to PageRank, the algorithm iteratively ranks *j* according to the formula:

$$r_j^{t+1} = (1 - d_j) f_j + d_j \sum_{i=1}^{N} \frac{A_{ji} r_i^t}{deg_i}$$

where r_j^t is the rank (score) of gene *j* in iteration *t* and d_j is a damping factor. $r_0 = f$ is initialized according to the differential expression. While PageRank uses two constant damping factors, one for all nodes with positive in-degree and one for all other nodes, DawnRank uses an individual damping factor for each gene, which depends on its connectivity. For gene j:

$$d_j = \frac{deg_j}{deg_j + \mu}$$

The higher the in-degree is, the higher the damping factor is and more connectivity information is incorporated into the ranking. In order to set a default value to *μ*, the genes of 100 random patients were ranked with various values of *μ*, and the ranking of known drivers was tested. The chosen value is $\mu = 3$.

The iterations stop when the overall difference between two consecutive iterations is below a certain threshold:

$$\sum_{i=1}^{N} |r_i^t - r_i^{t-1}| < \epsilon$$

Here, the threshold is set to 0.001. The output is a ranked list of driver genes according to the scores of mutated genes in the final vector r^n.

2.1.2. SCS

SCS (Single-sample Controller Strategy) [26] uses a network control strategy in order to detect and rank driver genes in individuals. The algorithm builds personalized networks based on a directed PPI network, normal gene expression data, tumor gene expression data and gene mutation indications, and attempts to reveal a small set of mutated genes that cause the transition from the normal state to the cancerous state. Mutated genes act as *controllers* and DEGs (differentially expressed genes) act as *targets* in the network. The main steps of SCS are personalized network construction and driver genes detection.

Personalized network construction:

First, the algorithm computes the log2-fold change of each gene out of the paired normal-tumor expression data. Genes with an absolute score greater than 1 are declared as DEGs and a +1 or -1 value is assigned to them, depending on their fold change sign. Then, the Random Walker with Restart algorithm (RWR) generates probabilities of reaching every node in the PPI network by paths starting from the mutated nodes. Revisiting the initial nodes is allowed. Let p^t be a vector in which the ith element holds the probability that the ith gene is reached from the set of initial nodes after *t* steps. Let *r* be the restart probability with a default value of 0.6 and let *W* be the column-normalized adjacency matrix of the PPI network. The RWR formula for transition probabilities at time *t+1* is:

$$p^{t+1} = (1 - r)Wp^t + rp^0$$

Assuming that there are k mutated genes, p^0 holds the probability of $\frac{1}{k}$ for each mutated gene and the probability of 0 for other genes. The RWR algorithm iterates until it reaches a

stationary state, defined by a difference below a predefined threshold between p^t and p^{t+1}. The threshold used in SCS is 10^{-6}.

In order to detect nodes with significant reaching probabilities, the RWR computation is done for additional 100 random networks that are degree preserving. This forms a background distribution of the walk probability for each node. Let p_i be the stationary probability obtained by the original network and let SD_i be the simulated distribution for node i. A z-score is computed as follows:

$$z^i = \frac{p_i - mean(SD_i)}{std(SD_i)}$$

P-values are calculated for all genes based on their z-scores. A personalized network is constructed from the significant genes with p-value<0.05, the mutated genes and the interactions between them.

Driver genes detection:

Next, network control principles are used to detect a minimal set of mutated genes that are linked to DEGs. Instead of using the whole network, the algorithm applies the CTC (Constrained Target Controllability) concept to detect driver genes out of a constrained subset of nodes. The constrained control nodes are mutated genes and the target nodes are DEGs. A greedy algorithm identifies the target controllable subsystem of each mutated gene and the related paths, by the following steps:

I. A bipartite graph is built with the set of nodes $B_0 = L_0 \cup R_0$, where R_0 are the target nodes, L_0 are their neighbors and the edge set is the bipartite edges between them. A maximum matching is computed using the Hopcroft-Karp algorithm. For each target gene, its matched node is recorded as belonging to the path that ends with it. Then, the algorithm iterates over *i* starting from *i*=1, sets R_i to be the set of nodes from L_{i-1} that were matched, sets L_i to be their neighbors, performs

maximum matching and records the matched nodes in the aggregated target paths. The process ends when $L_i = \emptyset$.

II. Another bipartite graph is built with the set of nodes $M \cup D$, where *M* is the set of mutated genes and *D* is the set of DEGs. For each pair $m \in M$, $d \in D$, there is an edge between them if *m* belongs to the previously computed path to *d*. Then, an LP-based classic branch and bound method is applied to solve the minimum set cover problem, where the set is contained in *M* and covers *D*. The genes in this set are called drivers.

III. In order to weight the detected drivers, SCS creates a consensus module for each mutation out of the personalized network. It runs 1000 iterations in which a random process replaces some of the edges found in the maximum matching process (I) by other edges from the personalized network. Then, new control paths between mutated genes and DEGs are extracted. The consensus module of each driver gene is composed of all edges that were found in these control paths in all iterations. Each edge is weighted according to its frequency in the experiments. Finally, drivers are ranked by the sum of edge weights in their modules.

2.1.3. PRODIGY

PRODIGY (Personalized Ranking of DrIver Genes analYsis) [27] ranks driver genes in individual patients according to their overall influence on known biological pathways. Influence scores are derived from gene expression profiles using the Prize Collecting Steiner Tree method (PCST) [28].

For each patient, PRODIGY takes as input its gene expression profile and a binary indication of whether each gene underwent a single nucleotide variation or not. Mutated genes are driver gene candidates. In addition, the algorithm uses a set of biological pathways and a

PPI network. In the first stage, a differential expression analysis is performed using the DeSEQ2 R library. The expression profile of each gene is compared to a background set of normal samples from the cohort and its log2-fold change is measured. Genes that pass an input threshold β and have a significantly different expression at FDR < γ are recorded as DEGs. Then, hypergeometric tests are performed to extract *deregulated pathways*. For each biological pathway, the set of genes that compose it is tested for significant enrichment with the detected DEGs at FDR < . The default parameters for the pre-processing stage are $\beta = 2, \gamma = \delta = 0.05$.

In the main part of the algorithm, influence scores are calculated for each pair of a deregulated pathway *p* and a mutated gene *g*. This includes a new network construction and a PCST computation. The process is presented in **Figure 6**.

Pathway-gene network construction:

Let $G_p = (V_p, E_p)$ be the network of *p* and $G = (V, E, W)$ be the input PPI network. Both networks are undirected. A joint pathway-gene network $G_{p,g} = (V_{p,g}, E_{p,g}, W_{p,g}, P_{p,g})$ is constructed, where:

- The set of nodes is:

$$V_{p,g} = V_p \cup g \cup N(V_p \cup g)$$

  where S is a subset of V and N_s is the set of neighbors of S in G. These are the pathway nodes, the gene node and their first neighbors in G.

- The set of edges is:

$$E_{p,g} = E_p \cup \{(u, v) \mid u, v \in V_{p,g}\ and\ (u, v) \in E\}$$

  These are the pathway edges with the addition of edges in the PPI network that connect other nodes in $V_{p,g}$.

- Edge weights are referred to as *costs* and set as follows:

$$W_{p,g}(u,v)=\begin{cases}0.1, & (u,v)\in E_p\\ 1-W(u,v), & otherwise\end{cases}$$

The cost of each pathway edge is 0.1. For other edges, the cost depends on their weight in the PPI network which represents their confidence level. The higher the confidence of an edge, the lower its cost. The maximum weight of an edge in the PPI network is 0.8, resulting in a minimum cost of 0.2. This gives an advantage to pathway edges as they always cost less.

- Node weights are referred to as *prizes* and set as follows:

$$P_{p,g}=\begin{cases}|\log(|\mathrm{FC}(v)|)|, & v\in \mathrm{DEG}\cap \mathrm{V_p}\\ -degree(v)^{\alpha}, & \mathrm{otherwise}\end{cases}$$

The prize for each DEG that belongs to the pathway is the absolute value of log2-fold change, which reflects the extent to which this node is differentially expressed. Other nodes get negative prizes that depend on their degree in the PPI network. This way hub nodes that have high degrees and have more connections are more severely penalized. The penalties are controlled by α parameter with a default value of 0.05.

Influence score computation:

The influence score of *g* on *p* is calculated with the Prize Collecting Steiner Tree method (PCST). The goal is to find a subtree of $G_{p,g}$ rooted at *g* that maximizes the sum of collected prize nodes while paying as little cost on edges as possible. This way, DEGs are collected along with intermediate nodes with high confidence connections. The optimal tree T_{opt} determines the influence of g in pathway p:

$$infl(g,p) = Score(T_{opt}) = \max\{\sum_{v\in V_T} P(v) - \sum_{(u,v)\in E_T} W(u,v) \mid T \; is \; a \; subtree \; of \; G_{p,g} \; that \; contains \; g\}$$

Pathways for which more than half of the genes have positive scores are excluded, since these are mainly long pathways that could contain many deregulated genes by chance.

Let *DP* be the set of deregulated pathways. The overall influence score of *g* is given by:

$$infl(g) = \sum_{p' \in DP} infl(g, p')$$

In the last stage of the algorithm, drivers are defined as genes with positive influence scores. They are ranked from most to least influential and returned to the user.

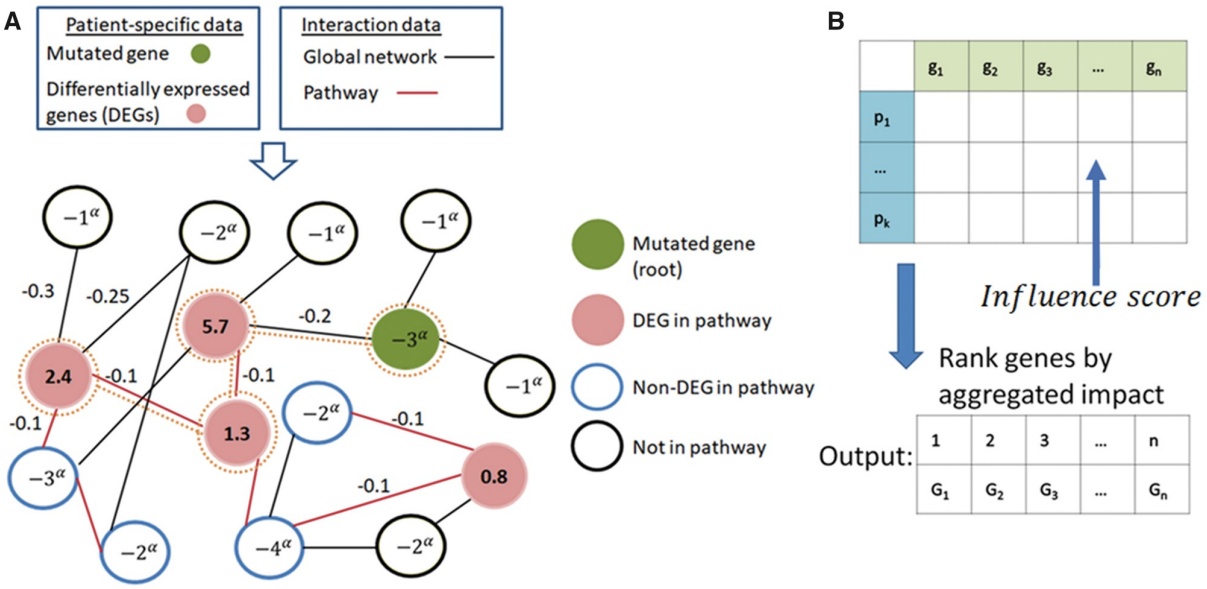


Figure 6: Outline of Prodigy's approach. (A) A pathway-gene network construction with node prizes and edge costs. The goal is to find a maximum weight subtree in the network rooted at the mutated gene. Its weight is the score of the PCST solution. In this example, the subtree marked by orange dotted lines is the PCST solution, of score 9-3α. (B) After calculating the scores for all pairs (p, g), an influence score is computed for each gene by its aggregated effect on all pathways. The final output is a ranked list of genes according to these scores. Source: Dinstag et al. [27].

2.1.4. IMCDriver

IMCDriver [29] detects and ranks driver genes in individuals based on functional similarity scores between patients, using prior information on known driver genes. The inputs per cancer cohort are somatic point mutation data, gene expression data, a PPI network and known driver genes. Then it trains an Inductive Matrix Completion (IMC) model to prioritize mutated genes in a new unseen sample.

Preprocessing and association matrices construction:

In the first stage, the algorithm removes driver candidates that are less likely to influence the expression of downstream genes. For each gene in each sample, the z-score of its expression is calculated relative to the background distribution of its expression across all samples. Genes with |z-score| > 2.0 are called *outlying genes* and mutated genes that are not directly linked to them in the PPI network are filtered out. Then, two association matrices are constructed; $A' \in \mathbb{R}^{N_g X N_s}$ is the *mutated gene-sample* association matrix, where $A'_{ij} = 1$ if the ith gene is mutated in the jth sample and 0 otherwise. $A \in \mathbb{R}^{N_g X N_s}$ is the *driver-sample* association matrix, where $A_{ij} = 1$ if the ith gene belongs to the NCG collection of known driver genes [14] and it is mutated in the jth sample, otherwise $A_{ij} = 0$. This allows paying more attention to known driver genes in the personalized analysis.

Computing similarity between genes/samples:

Next, functional similarities between genes and between samples are computed using the Gaussian interaction profile kernel similarity. The similarity score between sample s_i and sample s_j is given by:

$$G_s(s_i, s_j) = exp(-\gamma_l \mid\mid IP(s_i) - IP(s_j) \mid\mid^2)$$

where $IP(s_i)$ is the ith column of matrix A', which represents the mutational profile of genes in sample *i*. γ_l controls the kernel bandwidth and is set to be:

$$\gamma_l = \frac{m}{\sum_{i=1}^{m} || IP(s_i) ||^2}$$

where m is the number of samples. The similarity score between gene g_i and gene g_j is computed in the same manner, while replacing $IP(s_i), IP(s_j)$ with $IP(g_i), IP(g_j)$ to be the i[th] and the j[th] rows of A'.

Similarity scores are computed in this way for all pairs of genes and all pairs of samples and stored in $X_{orig} \in \mathbb{R}^{N_g X N_g}$ and $Y_{orig} \in \mathbb{R}^{N_s X N_s}$, respectively. In order to reduce the computational time and capture the most prominent similarity features, X_{orig} and Y_{orig} are transformed into $X \in \mathbb{R}^{f_g X N_g}, Y \in \mathbb{R}^{f_s X N_s}$ using the PCA dimensionality reduction method. f_s and f_g were set to 100 in this study, since the top 100 eigenvectors captured more than 95% of the variance of X and Y.

Implementing IMC to identify personalized driver genes:

In the main part of the algorithm, an Inductive Matrix Completion (IMC) model is trained with the functional similarity matrices to reveal driver genes in individual patients. This process transforms the driver-sample association matrix A into a low-rank matrix Z, that is later used to infer unknown driver-sample relationships along with the recovery of known relationships.

As illustrated in **Figure 7**, the goal is to obtain matrices $W \in \mathbb{R}^{f_g X r}, H \in \mathbb{R}^{f_s X r}$ that comprise $Z = WH^T$ and solve the following optimization problem:

$$\min_{W,H} J(W,H) = \frac{1}{2}||A - XWH^T Y^T||_F^2 + \frac{\lambda_1}{2}||W||_F^2 + \frac{\lambda_2}{2}||H||_F^2$$

where λ_1, λ_2 are regularization parameters, $|| \bullet ||_F^2$ is the Frobenius norm of a matrix. W and H are matrices with r columns. In this paper, r was optimized to be 100 and λ_1, λ_2 were set to be 1 following previous works.

The optimization function is non-convex. Therefore, the problem is solved using the alternating minimization strategy that fixes H and W in turns and computes the matrices separately. Specifically, H and W are initialized with random non-negative values and updated iteratively in two consecutive steps:

$$I)\ \ H_{jk} \leftarrow H_{jk} \frac{(Y^T A^T X W)_{jk}}{(Y^T Y H W^T X^T X W + \lambda_2 H)_{jk}}$$

$$II)\ \ W_{ik} \leftarrow W_{ik} \frac{(X^T A Y H)_{ik}}{(X^T X W H^T Y^T Y H + \lambda_1 W)_{ik}}$$

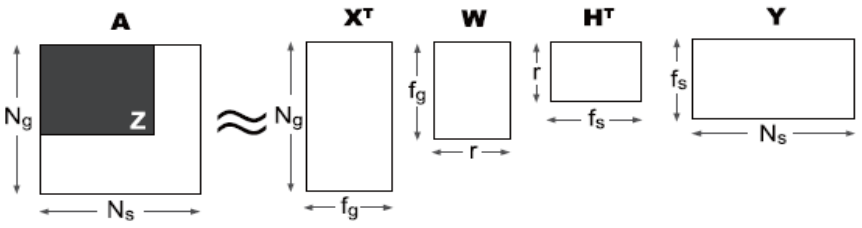
The iterations terminate when $|| A - XWH^TY^T ||_F^2 \leq \epsilon$. Here, $\epsilon = 10^{-6}$.

Finally, driver scores are computed for every mutated gene in a given sample. Let $s_{j'}$ be an unseen sample. Given its gene mutation profile, the algorithm computes its similarity scores vector $Y_{j'}$ with $G_s(s_{j',q})$ for $1 \leq q \leq N_s$ (see previous step). Then, the driver score of every mutated gene g_i is computed as follows:

$$score(i, j') = X_i^T W H^T Y_{j'}$$

The output is the list of genes ranked according to their scores.

This work was tested using the leave-one-out cross-validation (LOOCV) test, where in each iteration the model was trained using all samples but one. Results were evaluated using the precision, recall and F1 metrics.



$$\min_{W,H} \ J = \frac{1}{2}\left\| A - X^T W H^T Y \right\|_F^2 + \frac{\lambda_1}{2}\left\| W \right\|_F^2 + \frac{\lambda_2}{2}\left\| H \right\|_F^2$$

Figure 7: IMC algorithm is used to optimize W and H to predict scores of personalized mutated genes. Source: Zhang et al. [29].

2.2. GISTIC2.0 method for CNV analysis

GISTIC2.0 [18] identifies copy number altered regions in individual genomes of cancer patients and lists the genes that lie within them. The inputs are CEL files that are created by Affymetrix DNA microarray image analysis software. In the first step of the algorithm, these files are processed into segmented copy number profiles using a circular binary segmentation process [30]. Then, a series of computations is performed to detect and classify the underlying somatic events. We describe here main parts of the algorithm that were used in our work.

SCNAs deconstruction:

SCNAs (somatic copy number alterations) cannot be directly deduced from segmented copy number profiles since they may overlap. To this end, the novel ZD algorithm ('Ziggurat Deconstruction') was developed as part of GISTIC2.0. Given an observed chromosomal segmentation profile σ_c and a set of proposed SCNA histories H, the algorithm searches for the history with the maximum likelihood as follows:

$$h_c^* = \arg\max_{h_c \in H} \{Pr(\sigma_c|h_c) + penalty(h_c)\}$$

The penalty is determined by the complexity of the model using the Bayesian Information Criterion. The probability of the history could be expressed as follows, since SCNAs are assumed to occur independently:

$$Pr(\sigma_c|h_c) = \prod_{e_i \in h_c} Pr(e_i) = \prod_{e_i \in h_c} f(l_i, a_i)$$

The probability of the i[th] SCNA e_i is expressed with a function f and depends on its length l_i and its amplitude a_i. All probabilities are initialized using the assumptions that each copy number breakpoint represents a single alteration and that copy number gains are never followed by copy number losses (and vice versa). This implies that events of larger amplitude occurred later and allows a backwards inference of the deconstruction. In each iteration, the probability of each SCNA is computed followed by the most likely history of events. This procedure ends with a list of individual SCNAs along with their amplitude and length.

Scoring SCNAa according to likelihood of occuring by chance:

The algorithm uses a framework to score regions of the genome by the probability that the SCNAs within them did not occur by chance alone. The main interest is in focal SCNAs rather than chromosomal-arm-level events, since the influence of the latter on cancer progression is unclear. Nonetheless, focal and arm-level events might depend on each other. Let $B_i = \{b_1, b_2, ...\}$ be arm-level SCNAs and $F_i = \{f_1, f_2, ...\}$ be focal SCNAs that cover marker i in the genome. Assuming that focal events are independent, the focal GISTIC score FG_i of marker i is defined as:

$$FG_i = -ln(Pr(F_i|B_i)) = -ln(\prod_{f \in F_i} Pr(f|B_i)) = -\sum_{f \in F_i} ln(Pr(f|B_i))$$

Given the length and the amplitude of a focal event, its probability can be estimated by the frequency of other focal events of the same length, amplitude and arm-level event relations. However, focal events containing driver genes tend to be of shorter lengths and higher amplitudes, suggesting that they might compose a biased reference set and each of them could be underestimated. The algorithm formulates alternative computations based on two observations made on a large dataset of more than 3000 patients with various types of cancer. First, it was shown that the frequency of focal events of all lengths is roughly constant, except for the shortest lengths that tend to match driver gene events. Second, the

frequency of focal events decreases exponentially with their amplitude, which is the change in copy number units. The proposed null model for the probability of a focal event f with amplitude A is therefore:

$$Pr(f) = f(A) = \alpha e^{-\alpha A}$$

where $\alpha \in \{\alpha_{amp}, \alpha_{del}\}$ is a positive scaling parameter that is derived from all amplifications or deletions across the samples, in accordance with the focal event type.

Finally, the distribution of total focal event amplitudes as a function of total arm-level event amplitudes was drawn. Focal amplifications appear to be independent of arm-level amplifications, while focal deletions are strongly dependent on arm-level deletions. Let B be the copy number change of the underlying arm-level deletion, which could rarely exceed 1. Then the null model for the conditional probability of focal events is given by:

$$Pr(f|B) = \begin{cases} \alpha_{amp} e^{-\alpha_{amp} A}, & \text{if } A > 0 \\ (1+B)\alpha_{del} e^{\alpha_{del} A}, & \text{if } A < 0 \ \ and \ \ B > -1 \\ \epsilon \alpha_{del} e^{\alpha_{del} A}, & \text{if } A < 0 \ \ and \ \ B \leq -1 \end{cases}$$

These probabilities are computed for each sample and summed up to generate the underlying distribution. Then, the algorithm computes FG_i scores and SCNA p-values with a correction for multiple testing using the Benjamini-Hochberg FDR method. Significant markers are recorded.

In addition to marker SCNA scores, GISTIC2.0 computes gene SCNA scores. These scores account for all the events that affect a single gene. Let *I* be the group of loci that fall within the genomic region of gene *g*. For each sample, the probability of a focal event f_g that includes *g* given arm-level events is calculated as:

$$Pr(f_g|B) = \min_{i \in I} Pr(f_i|B)$$

The logarithms of these gene probabilities are summed up over independent samples to generate a background distribution of gene scores. Then, statistically significant genes are recorded based on their focal event frequency.

Labeling gene SCNAs

GISTIC2.0 further classifies gene SCNAs according to the following categories: '-2' represents homozygous deletion events, '-1' represents single copy deletions, '1' represents low-level amplification and '2' represents high-level amplification.

'-1' and '+1' values are assigned to events that exceed an input low-level threshold that accounts for noise. It is typically equal to the absolute amplitude of 0.1 or 0.3. As for '-2' and '+2' categories, high-level thresholds are computed individually on a sample-by-sample basis. They are the median amplitudes of observed arm-level deletions or arm-level amplifications in each sample.

2.3. Phylogenetic trees

As elaborated in the "Evolution of Cancer" chapter under "Computational Background", cancer development follows a Darwinian model. It can be described by chains of mutational events that cause the formation of new cell subclones out of an initial clone of non-cancerous cells. A convenient and unified representation of these chains is obtained by the construction of phylogenetic trees. For each cancer case, the root node represents the set of mutations that are present in the initial colony of cancer cells, emerging from a normal cell population. Every other node contains mutations that were added on top of those in its parent node, in the cell population corresponding to that subclone. This provides an evolutionary model, in which ancestor node mutations occur earlier than their descendant

mutations. The model assumes that mutations are not reverted in evolution. In every node some mutations are drivers that encourage cell growth, while others are passengers that occurred simultaneously by chance. It is reasonable to speculate that mutations in higher nodes are more likely to be drivers, since they are common to more cancerous cells and can explain their abnormality.

2.3.1. VAF and CCF

VAF (Variant Allele Frequency) is a raw calculation of the percentage of cells that carry a specific variant. The count of a variant is the number of reads the contain it. For allele a, given its reference counts t_ref_count and alternative counts t_alt_count that were obtained by genome sequencing of a cell population, the VAF value is computed as:

$$VAF(a) = \frac{t\_alt\_count}{t\_ref\_count + t\_alt\_count}$$

In normal cells, VAF measures diploid zygosity, where heterozygous loci should have values near 0.5 and homozygous loci should have values near 1. In cancer studies, VAF is used to estimate the extent to which a mutation has spread in a population of cancer cells. Given a point somatic mutation m, we generalize the VAF definition to refer to its mutated and reference alleles.It follows that $VAF(m)$ represents the percentage of mutated alleles within the cells. Since cancer subclones emerge along the tissue development, the larger $VAF(m)$ is, the more likely it is that m occurred early in the evolution.

CCF (Cancer Cell Fraction) is an estimation of the percentage of cancer cells (rather than alleles) that share each mutation while accounting for read count inaccuracies. CCF values are not directly measured out of bulk sequencing data, but inferred out of read counts. They are inferred from cancer phylogenetic trees, where the set of all mutations represented by the same node share the same CCF value. In particular, root node mutations have a CCF

value of 1 as they are common to all cells. We represent in the next section the PhyloWGS algorithm [37], which computes phylogenetic trees and provides CCF values of the mutations they represent.

2.3.2. PhyloWGS

PhyloWGS [37] reconstructs cancer phylogenetic trees based on whole-genome sequencing data of bulk tumor samples. It is one of the first reconstruction methods to account for both SNV and CNV mutations.

Typically, methods for SNV-based tree reconstruction [38,39,40] cluster SNVs by fitting statistical mixture models based on VAF values. The clusters form nodes, and a tree is constructed according to the following assumptions on tumor evolution: 1) The infinite sites assumption, suggesting that each SNV occurs once in the evolution; 2) Strong parsimony, according to which a small number of subpopulations are still present among the cells, suggesting that the number of branch points where the parental subpopulation has a zero frequency should be maximized; 3) Weak parsimony for clusters, suggesting that all SNVs within a VAF cluster are assigned to the same mutation set. **Figure 8** exemplifies an SNV-based tree reconstruction that is based on VAF clusters and could be obtained by one of the reconstruction methods.

Unlike previous methods, PhyloWGS does not assume that a single tree is to be reconstructed. Instead, it may output multiple trees with their probabilities. VAF clustering and tree reconstructions are performed concurrently.

A CNV-based tree reconstruction is a much more difficult problem. The infinite site assumption is often invalid. In addition, there can be more than one solution to the equation for the percentage of mutated cells ϕ and the new copy number C from an observed non-normal copy number x:

$$x = \phi C + (1 - \phi)2$$

Most CNV-based methods only account for clonal CNVs or a small number of subpopulations [41,42].

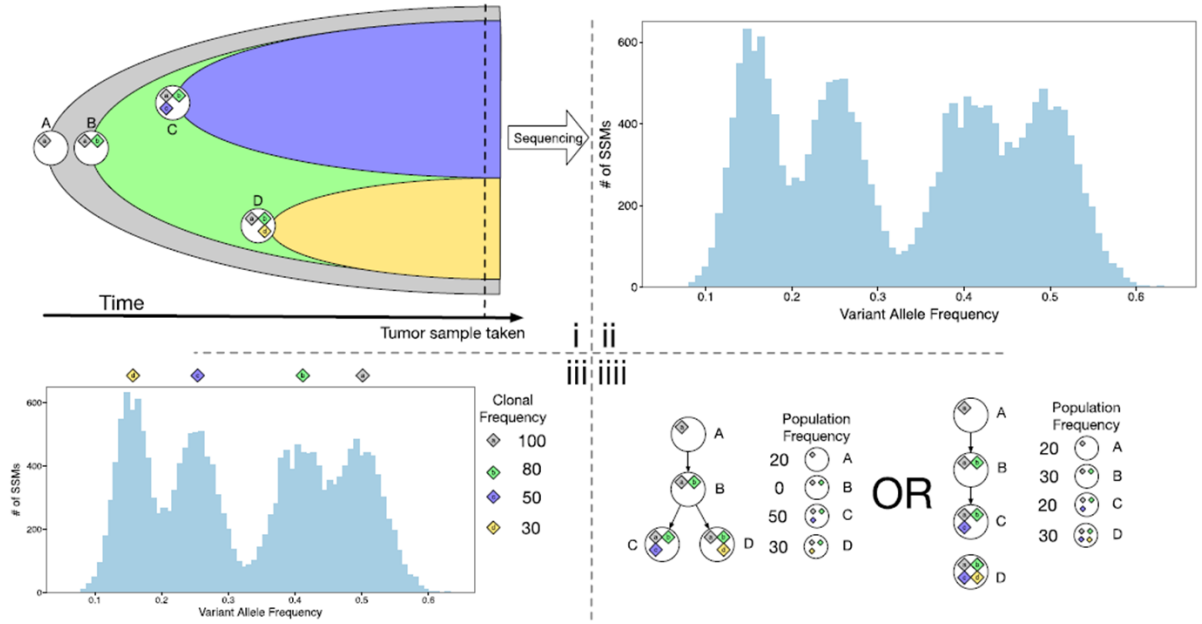


Figure 8: The development of intratumor heterogeneity and SNV-based subclonal reconstruction. (i) Tumor composition over time. Each color corresponds to a subclone and its relative frequency changes over time. The set of all mutations in a subclone is shown inside the node. (ii) The resulting distribution of variant allele frequencies (VAFs). (iii) The result of successful inference of the VAF clusters. (iiii) The desired output of subclonal inference. Source: PhyloWGS [37]

The PhyloSub model for SNV-based reconstruction:

PhyloWGS is based on PhyloSub [39], a previous method for phylogenetic tree reconstruction from SNV data. For a variable genomic position i, let a_i be its reference allele count and b_i be its variant allele count. $d_i = a_i + b_i$ is its total allele count. Let μ_i^r be the probability of sampling a reference allele from the reference population and μ_i^v be the probability of sampling a reference allele from the variant population. μ_i^r and μ_i^v depend on the error rate of the sequencer. μ_i^r ~1 since all alleles in the reference population are not

mutated, while μ_i^v~0.5 since half of the alleles in the reference population are mutated and half are not. Let N be the number of SNVs. PhyloSub generates all SNV population frequencies $\{\tilde{\phi}_i\}_{i=1}^N$ using the Tree-Structured Stick Breaking method (TSSB) [43]. This process produces a discrete set of frequencies out of a prior probability distribution over all SNV frequencies, where the set matches the nodes of an evolutionary tree structure model. The tree structure is obtained using a base distribution H that is updated during the process. For the root node, it is $Uniform(0,1)$ and for every other node v it is $Uniform(0, \phi_{parent(v)} - \sum_{w \in siblings(v)} \phi_w)$. α and γ are parameters that control its height and depth. Then, the generative model for allelic count observations is given by:

$$\varsigma \sim TSSB(\alpha, \gamma, H); \quad \tilde{\phi}_i \sim \varsigma$$

$$a_i \mid d_i, \tilde{\phi}_i, \mu_i^r, \mu_i^v \sim Binomial(d_i, (1-\tilde{\phi}_i)\mu_i^r + \tilde{\phi}_i \mu_i^v)$$

where ς is the prior distribution generated in the TSSB process.

PhyloWGS model for CNV integration:

The relationship between SNVs and their VAF values is complicated when there is an overlapping CNV event. For instance, if an SNV occurred prior to its segment duplication, its observed variant allele count would be two times higher than expected, resulting in a higher VAF value. To this end, PhyloWGS attempts to fix VAFs of SNVs according to overlapping CNVs, and builds a unified phylogenetic tree of both mutation types.

For each CNV j that does not overlap any SNV, a pseudo-SNV is created. Let C_j be its altered copy number, $\tilde{\phi}_j$ be its population frequency, C_j^m be its maternal copy number and C_j^p be its paternal copy number. The pseudo-SNV is represented as a heterozygous binary SNV, where the population frequency is equal to $\tilde{\phi}_j$, the read depth $d_j = a_j + b_j$ depends

on the evidence supporting the CNV and the alternative read depth is $a_j = d_j \frac{\tilde{\phi}_j}{2}$. The VAF value $\frac{b_j}{a_j + b_j}$ is approximated to be $\frac{\tilde{\phi}_j}{2}$. In the first stage, the TSSB process is used to build a phylogenetic tree out of real SNVs and pseudo-SNVs, as described in the PhyloSub section.

The final step is correcting VAF values of real SNVs. For a subpopulation u that is represented by a node in the tree, let η_u be its proportion out of the entire cell population. η_u could be computed based on the frequency of its mutations $\tilde{\phi}_u$ and its children in the tree, in a tree ascending procedure. For instance, let A be a parent node with $\tilde{\phi}_A = 1$ and two children nodes: B with $\tilde{\phi}_B = 0.2$ and C with $\tilde{\phi}_C = 0.5$. Then we set the node subpopulation proportions to be $\eta_B = 0.2, \eta_C = 0.5, \eta_A = 1 - 0.2 - 0.5 = 0.3$. Using these proportions, the number of reference and variant allele copies could be computed while accounting for CNVs as described next, resulting in new VAF values.

Let N_i^r, N_i^v be the number of reference and variant copies of allele i. The algorithm initializes N_i^r, N_i^v to be equal to 0 and updates these values while ascending up the tree. Each node in the tree represents a subpopulation that either contains or lacks this SNV. For each SNV i, the algorithm ascends from the leaves to the root and updates the number of its copies as follows:

- If the current node population u does not contain SNV i (whether it is affected by a CNV or not):

  $N_i^r \leftarrow N_i^r + \eta_u C_i,$

  $N_i^v \leftarrow N_i^v + 0$

- If the current node population u contains SNV i, but is not affected by a CNV or is affected by a CNV that occurred before the SNV:

$$N_i^r \leftarrow N_i^r + \eta_u \cdot max(0, C_i - 1),$$

$$N_i^v \leftarrow N_i^v + \eta_u$$

- If the current node population u contains SNV i and is affected by a CNV that occurred after the SNV (the SNV is on the maternal copy, w.l.o.g):

$$N_i^r \leftarrow N_i^r + \eta_u C_i^p,$$

$$N_i^v \leftarrow N_i^v + \eta_u C_i^m$$

Finally, the new generative model for allelic count observations is given by:

$$\varsigma \sim TSSB(\alpha, \gamma, H); \;\; \tilde{\eta}_i \sim \varsigma$$

$$a_i \mid d_i, \tilde{\eta}_i, \epsilon \sim Binomial(d_i, \frac{N_i^r(1-\epsilon) + N_i^v \epsilon}{N_i^v + N_i^r})$$

where ϵ is the sequencing error probability.

A maximum likelihood solution is used to infer the parameters, and the CCF values correspond to the frequency of the SNVs represented by the node.

3. Methods and Results

In this section we describe several test suites, in which we generated patient-specific driver gene lists using different data types. We show the logics behind the tests, the materials we used and the results of each suite.

3.1. Driver list evaluation

In order to evaluate each test, we compared the generated results to previous PRODIGY run results. Originally, PRODIGY ranked only genes that underwent single nucleotide variations (SNVs) based on DEG-enriched pathways (see “Computational background” section). Here, we made several changes and aimed to improve the performance.

3.1.1. Gold standards

The main resource we rely on in our evaluations is the COSMIC Cancer Gene Census (CGC) [13]. This is a collection of experimentally verified cancer driver genes. Each gene is annotated with the mutation type observed in it, the tumor types it affects and its role in cancer. The genes are divided into two groups; “Tier 1” is composed of validated drivers with both known cancer-related functionality and experimental evidence of causing damage upon mutations. “Tier 2” is composed of genes with high cancer driving potential but no experimental validations. We use only the first tier as a reference set of known driver genes and check overlaps with the genes we detect. It contains about 600 genes.

The evaluations are performed with subsets of the CGC collection. We used two levels of refinement:

- Drivers that are classified according to the type of mutation that triggers them. Here we consider: (1) “SNV known drivers”, i.e., genes that undergo point mutations causing missense mutations or frameshifts, (2) “CNV known drivers”, which undergo

large amplifications or deletions, and (3) “translocation known genes”, namely genes involved in translocations.

- Drivers that are annotated with a specific cancer type.

Table 1 summarizes the different mutation types for five cancers.

	BLCA	BRCA	COAD	HNSC	LUAD	All types
SNV known genes	10	30	38	15	10	248 (37%)
CNV known genes	7	10	11	5	6	65 (10%)
Translocation known genes	-	-	-	-	-	314 (47%)

Table 1: Number of known driver genes from Tier 1 of CGC for each mutation and cancer type. Rows represent mutations and columns represent cancer types. *BLCA* - bladder urothelial carcinoma, *BRCA* - breast invasive carcinoma, *COAD* - colon adenocarcinoma, *HNSC* - head and neck squamous cell carcinoma, *LUAD* - lung adenocarcinoma.

In some of the tests we used the Network of Cancer Genes (NCG) [14] of the Ciccarelli group. Similarly to CGC, it contains about 600 validated driver genes and an additional collection of candidate driver genes with cancerous indications. We used only the validated drivers as a reference set.

3.1.2. Performance assessment

The mission of detecting driver genes could be treated as a binary classification problem, in which we label mutated genes as drivers (“positive”) or non-drivers (“negative”). True positives (TP) stand for known driver genes that were detected. False positives (FP) stand for genes that were positively labeled but are not known drivers. False negatives (FN) stand

for known driver genes that were not detected. Lastly, true negatives (TN) stand for genes that are not known drivers and were negatively labeled. See **Figure 9**.

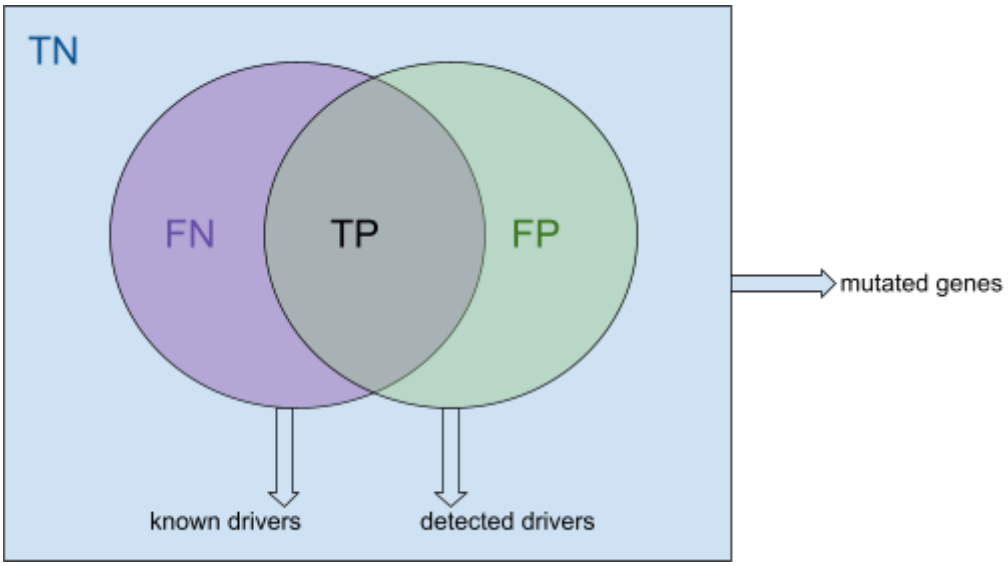


Figure 9: Venn diagram to illustrate the relations between known driver genes and the genes that are labeled as drivers in our experiments. *TN* - true negatives, *FN* - false negatives, *TP* - true positives, *FP* - false positives.

We used three types of measurements to assess the driver lists in each test we performed:

- **Precision** is the percentage of true positive labels out of all positively labeled genes. That is, the percentage of genes that were detected as drivers and are indeed known drivers out of all detected genes.

  $$precision = \frac{\#TP}{\#TP + \#FP}$$

- **Recall** is the percentage of true positive labels out of all positive genes, including those that were mistakenly negatively labeled. That is, the percentage of known drivers that were detected out of all mutated known driver genes. Importantly, the latter set is not constant across patients, since it is an intersection of the gold standard list with the set of a specific patient's mutated genes.

  $$recall = \frac{\#TP}{\#TP+\#FN}$$

- **F1** is a combined score that summarizes both precision and recall at once.

$$f1 = \frac{2 * precision * recall}{precision + recall}$$

The higher these values are, the more accurate our prediction is.

For each patient, for k=1 to 20, we took the top k genes ranked according to their influence scores and computed precision, recall and F1 of the resulting set.

3.2. Test suite 1: Copy number variations

Previous studies have shown that copy number variations (CNVs) in DNA segments are potential contributors to oncogenesis [1,2]. 10% of the known cancer driver genes in CGC tend to trigger cancer due to large amplifications or deletion events. Analyses of TCGA [16] patient genomes reveal a great extent of CNVs.

The aforementioned facts led us to test the cancer driving potential of genes that are involved in CNV events. We used Xena platform [17] to extract the following data for TCGA patients:

I. CNV estimates of genomic segments, including the chromosome number, the start position and the end position. Copy number profiles were measured using the Affymetrix Genome-Wide Human SNP Array 6.0 platform. Segments were deduced using a circular binary segmentation [30] and mapped to hg19 genome assembly at the Broad Institute.

II. Gene-level copy number estimates, computed using the GISTIC2.0 method [18] from genomic segment CNVs.

III. Thresholded gene-level copy number estimates, computed using the GISTIC2.0 method from genomic segment CNVs. The thresholds are -2,-1,0,1, and 2, representing homozygous deletion, single copy deletion, diploid normal copy,

low-level copy number amplification or high-level copy number amplification (see “Computational Background” section).

The original study using PRODIGY ranked only genes that underwent single nucleotide variations (SNVs). Here we applied the algorithm to all genes that underwent any type of copy number change or SNV with the default set of parameters: {network=”STRING”, alpha=0.05, pathwayDB=”reactome”, beta=2, gamma=0.05, delta=0.05}. Note that this test has a different gold standard. When evaluating the list of drivers that resulted from all CNVs and SNVs as candidates, our gold standards included both SNV and CNV known CGC genes (see “Gold standards section” under “Methods and Results”). When evaluating the results of SNV candidates only, we used SNV known genes.

Figure 10 shows Bladder Cancer (BLCA) cohort performance. Other cohorts showed similar results. There is an increase in the average precision. This is probably due to the extended set of gold standards that results with a higher probability to intersect with the detected genes. However, the average recall is very low and as a result, so is the f1 score. The main reason for the drop in recall is the much larger gold standard: the positive set is the set of all CGC genes that contain SNV or are part of a CNV in the sample. The number of genes involved in CNVs is large. For example, for BLCA, the average number of genes with SNVs is 177 while the average number of genes involved in CNVs is 12212. Similar trend is observed in other cancers. See also **Figure 11**, which shows that more than 40% of all genes are involved in CNVs.

Due to the low performance, we used several approaches to identify driver candidates that underwent CNVs and have a higher potential of being true drivers according to specific criteria. These approaches are described in the next sections.

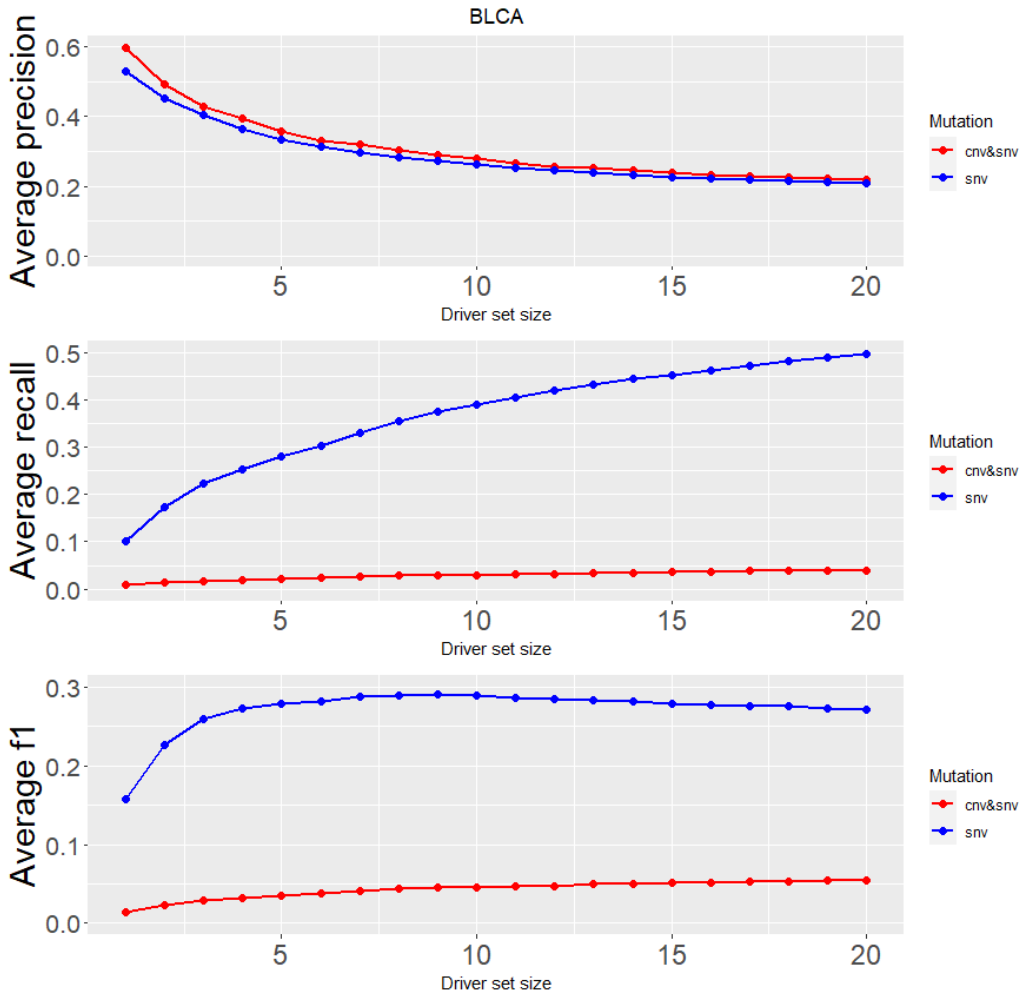


Figure 10: PRODIGY performance when the set of candidate drivers is all CNV and SNV genes in comparison to only SNV genes. Average precision, recall and f1 were measured for x top-scored detected genes in the BLCA cohort of TCGA, for $1 \leq x \leq 20$

3.2.1. Major CNVs

We call homozygous deletions and high-level amplifications, as detected by GISTIC2, *major CNVs*. These events have the potential of causing extensive damage to the cell. **Table 2** and **Figure 11** show statistics of major CNVs vs. all CNV events in five TCGA cohorts. They suggest that testing only major CNVs could filter out a great portion of noise and result with a refined dataset of higher quality.

	# patients	Average	Median	Max
BLCA	408	2.5%	1.8%	19%
BRCA	1080	2.6%	1.7%	22%
COAD	451	0.8%	0.3%	7%
HNSC	522	1.7%	1.2%	14%
LUAD	516	2.2%	1.4%	11%

Table 2: The average, the median and the maximal percentage of genes that underwent major CNVs out of all detected CNV genes for patients in five TCGA cohorts. *BLCA* - bladder urothelial carcinoma, *BRCA* - breast invasive carcinoma, *COAD* - colon adenocarcinoma, *HNSC* - head and neck squamous cell carcinoma, *LUAD* - lung adenocarcinoma.

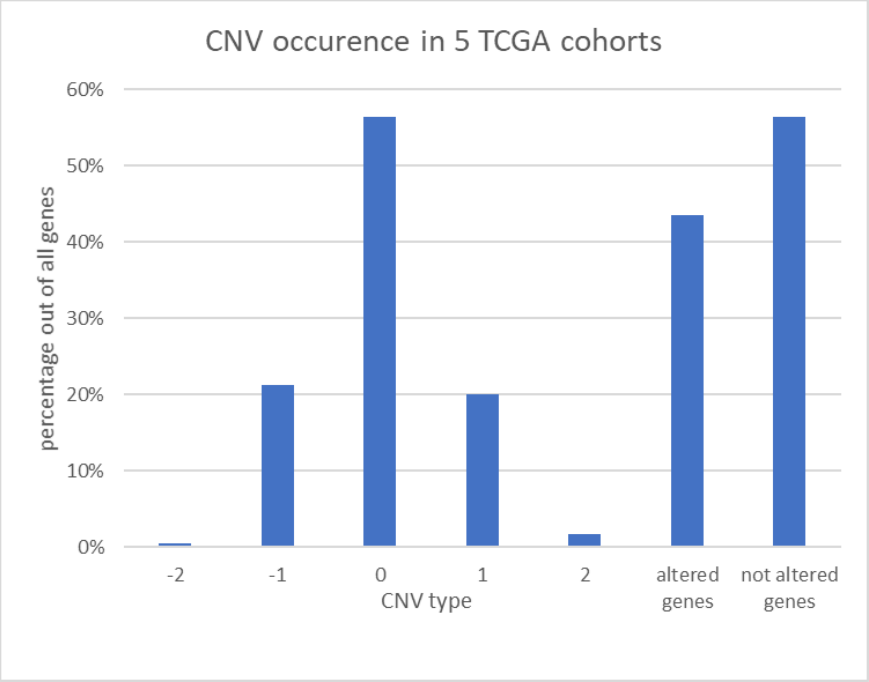


Figure 11: Percentage of genes that underwent different types of copy number variations in five TCGA cohorts: BLCA, BRCA, COAD, HNSC and LUAD (see **Figure s1** in “Supplementary Material” section for cohort-level

statistics). *-2* - homozygous deletion, *-1* - single copy deletion, *0* - normal copy number, *1* - low-level amplification, *2* - high-level amplification. Altered genes are those with any CNV, i.e., type ≠ 0.

We ran PRODIGY with a new collection of altered genes; For each patient, the new driver gene candidates were those with major CNVs or SNVs. The rest of the algorithm remained unchanged. We used default parameters. **Figure 12** shows the performance for the BLCA cohort. Other cohorts showed similar results. The precision for the top ranking genes decreases in comparison to the runs with SNVs only, so fewer detected genes are known drivers, even though the gold standard increases. The recall and the f1 scores are much better than when using all CNVs and SNVs, but are still far lower than the scores when using only SNVs. We conclude that focusing on major CNVs improves the performance and cleans the signal, but is still inferior to using SNVs only.

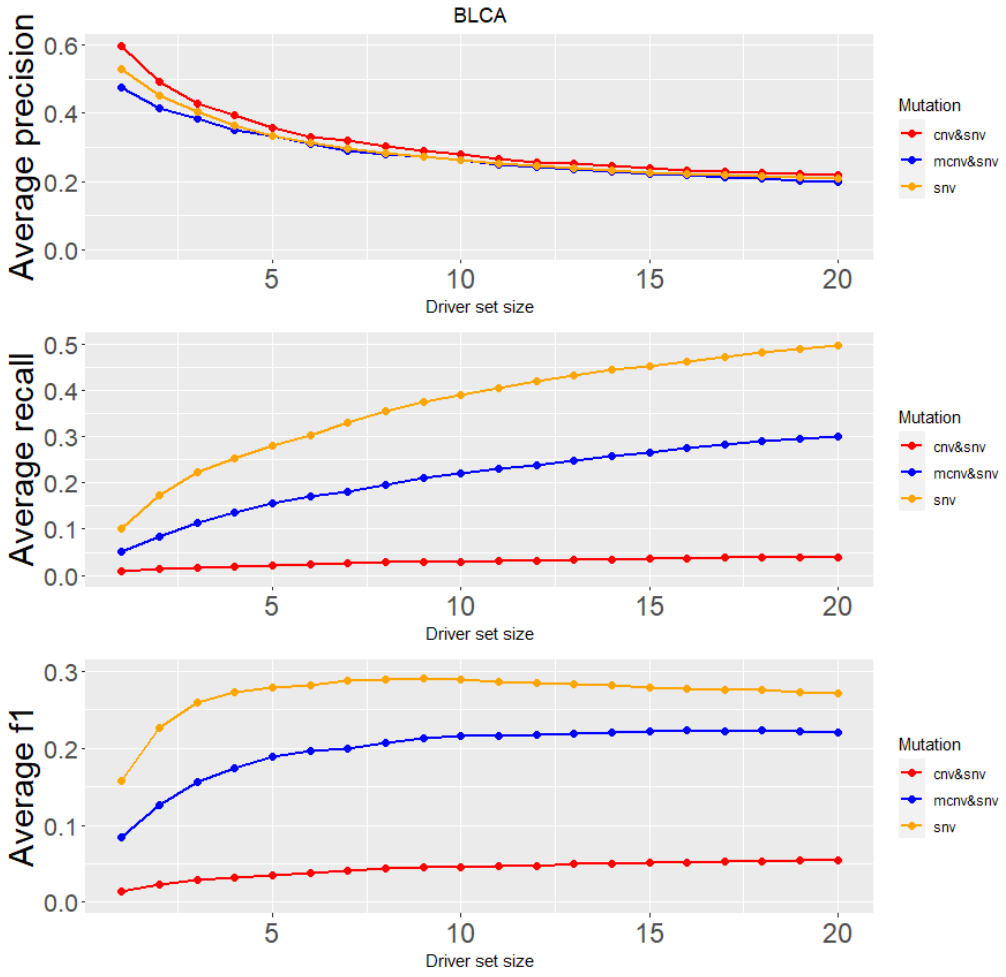


Figure 12: PRODIGY performance when the set of candidate drivers is composed of: (1) major CNVs and SNV genes, (2) all CNVs and SNVs and (3) SNV genes only. Average precision, recall and f1 were measured for x top-scored detected genes in the BLCA cohort of TCGA. *cnv* - all CNVs, *mcnv* - major CNVs.

3.2.2. Jointly altered genes

We call genes that are within the same altered genomic segment *jointly altered genes*. Genes that gain or lose copies in the same event could have a joint effect that may lead to abnormal processes in the cell.

We used CNV estimates of genomic segments from TCGA patients. The segments are of various lengths; some consist of hundreds of millions of base pairs, while the median CNV segment size ranges between 600K and 3M base pairs in the five distinct cancer cohorts that we analyzed. **Figure 13a** shows histograms of these lengths per cohort. Long alterations are clearly rare in comparison to the rest. In order to observe altered segment lengths more closely, we focused on the segments of length shorter than 1Mbp, those shorter than 100Kbp, and those shorter than 1Kbp in each cohort. The histograms in **Figure 13b** show that the amount of segments decreases sharply as lengths are getting longer in the BLCA cohort. This is true for the other cohorts too (see **Figure s2** in “Supplementary Material” section). Note that Mermel et al. [18] found that the frequency of focal copy number alterations of all lengths is roughly constant in the background set (see “GISTIC2.0 method for CNV analysis” under “Computational Background” section). It follows that segments of shorter lengths are more often detected as not occuring by chance alone. The shorter a segment is, the less genes it may contain. This further motivated us to extract jointly altered genes, since when few genes are involved, their influence on cancer progression could be examined more closely.

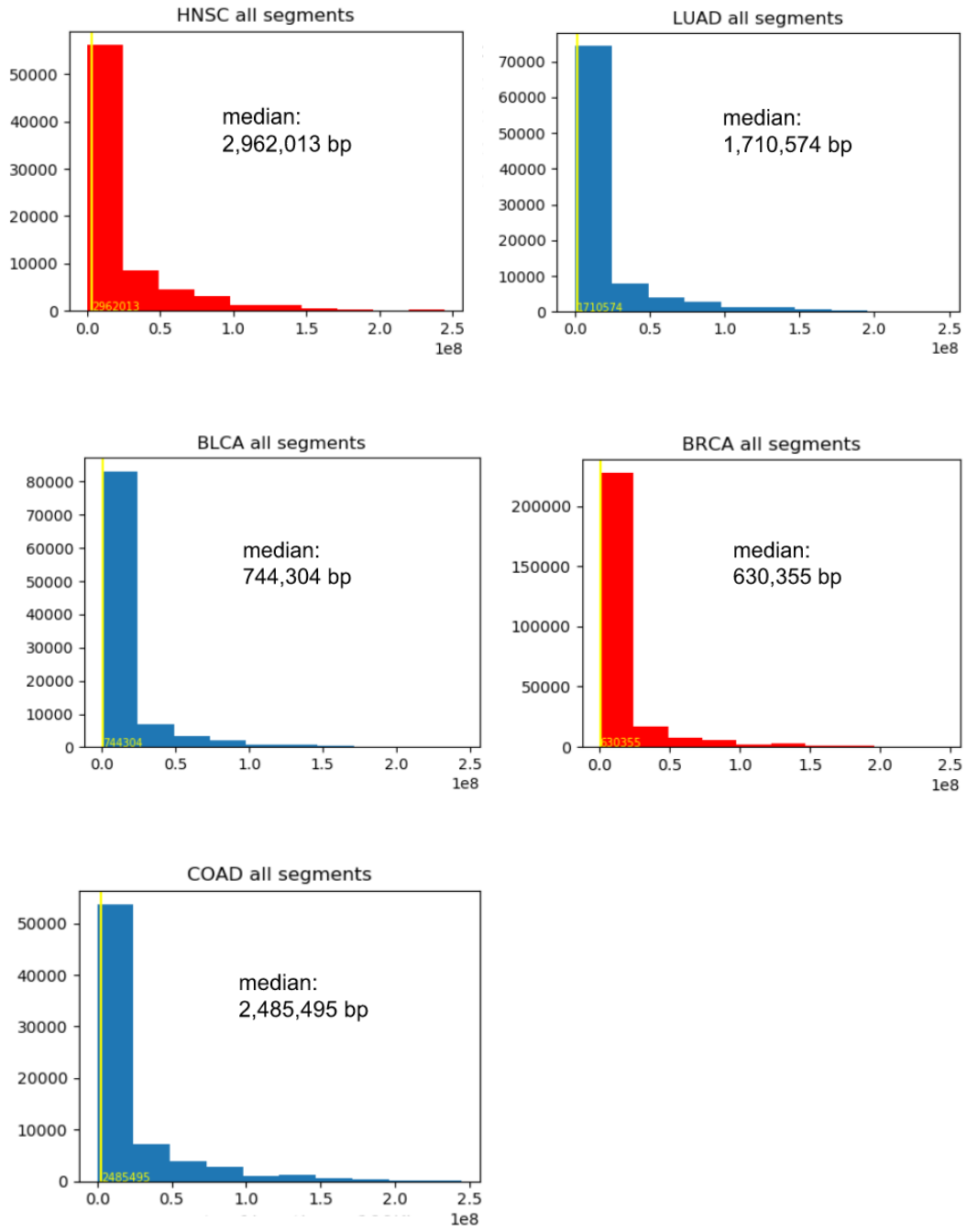


Figure 13a: Distribution of the lengths of all copy number altered segments in basepairs of five cancer cohorts. Median lengths are inlaid .

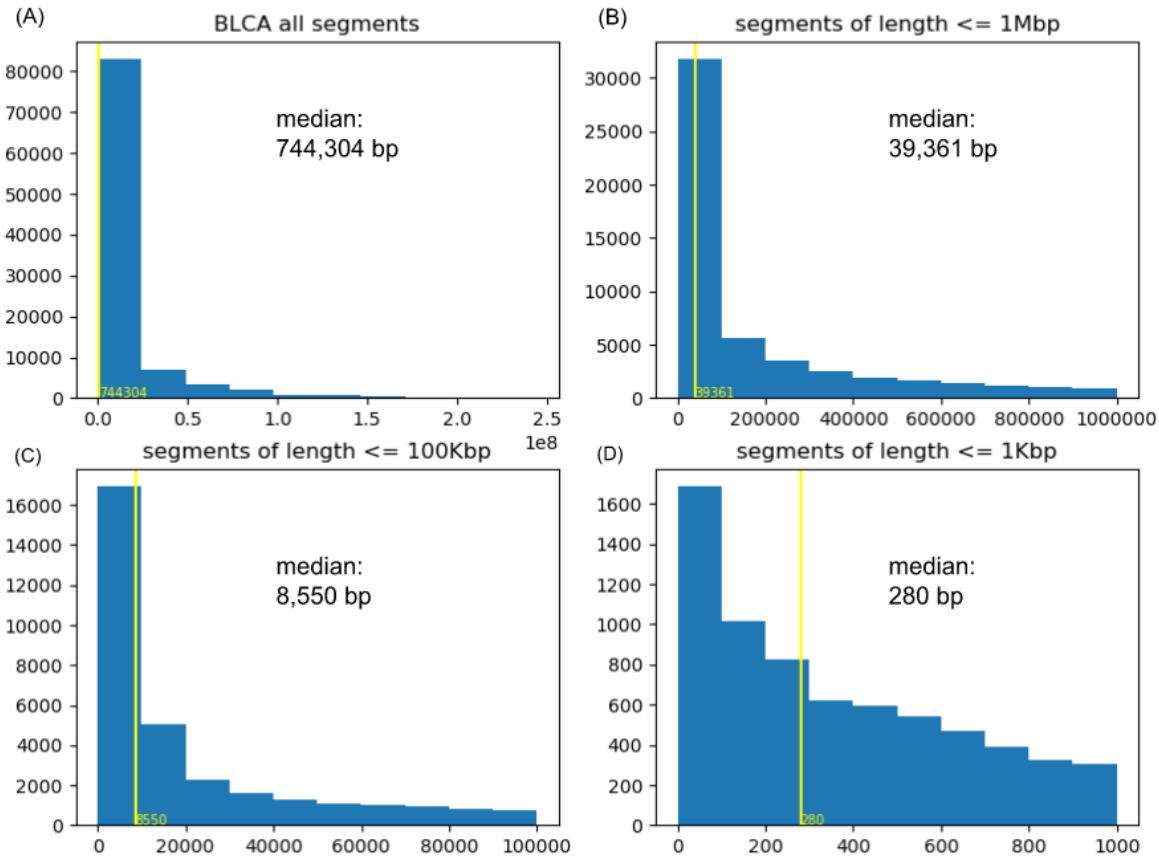


Figure 13b: Distributions of lengths of copy number altered segments in the BLCA cohort. (A) all segment lengths, (B) segments of length ≤ 1Mbp, (C) segments of length ≤ 100Kbp, (D) segments of length ≤ 1Kbp. Median lengths are inlaid.

We identified jointly altered genes using the “GenomicRanges” and “GenomicFeatures” R libraries [19], with hg38 as a reference genome [20]. The chromosome numbers, the start and the end positions of each segment were passed as inputs. **Figure 14** shows the frequency of segments that carry different numbers of genes. Among all altered segments in five TCGA cancer cohorts (**Figure 14a**), 24% contain a single gene. Each set size in the range of 2 to 50 genes has a frequency of 6% or less. When looking only at 10% of the segments in each cohort with the highest alteration values (**Figure 14b**), the high frequency of segments with a single gene is even more prominent. Among these segments, 70% consist of a single gene and other set sizes have a maximum frequency of 8%.

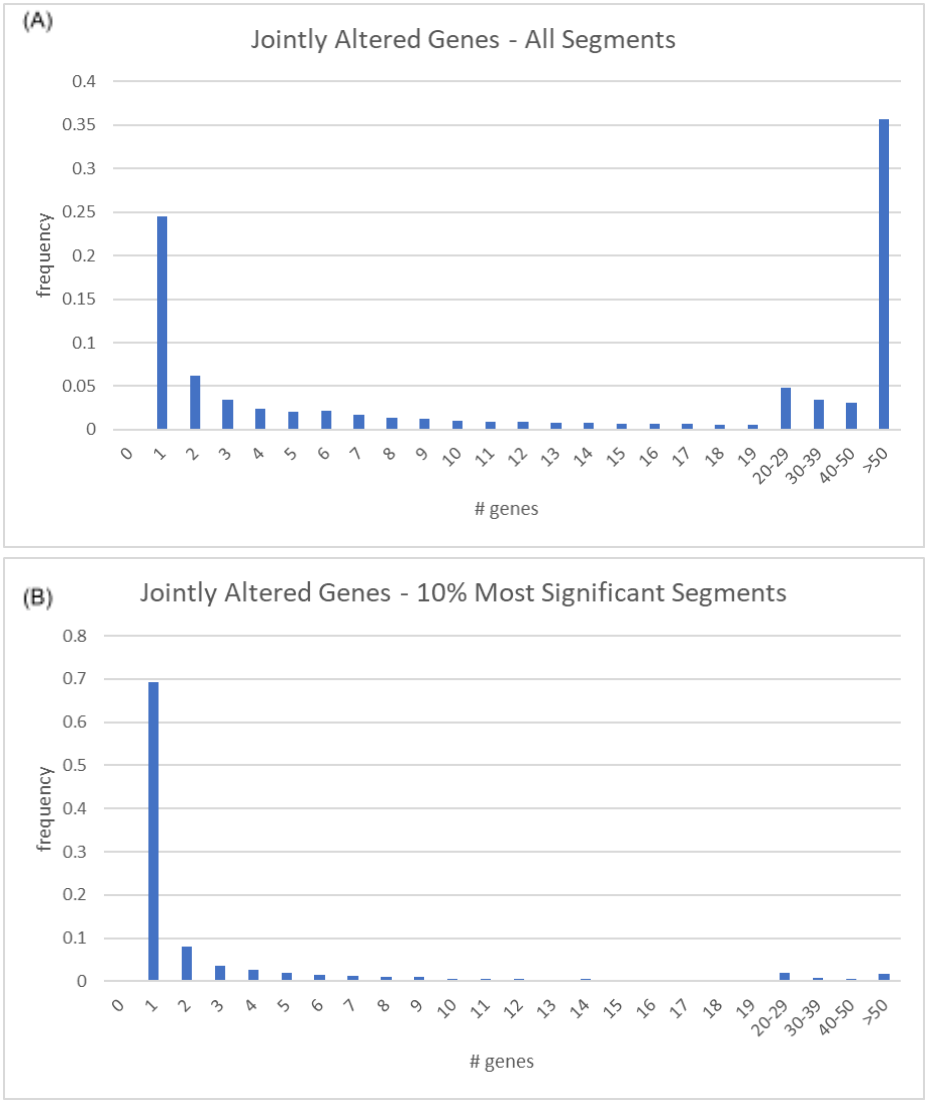


Figure 14: Frequencies of copy number altered segments partitioned by the number of genes they carry in five TCGA cancer cohorts. (A) all altered segments, (B) The 10% of the segments with highest alteration values from each cohort.

3.2.3. Solo altered genes

We call genes that lie within copy number altered segments with no other genes *solo-genes*. As shown in the previous section, these genes are very frequent among the background set of all copy number altered genes, especially when looking at segments with high alteration

values. We suspected that this reveals a positive selection phenomenon and that these genes may tend to be cancer driver genes.

Table 3 describes altered solo-gene statistics in comparison to genes that underwent SNVs in the five TCGA cohorts. Solo-genes are less frequent than SNVs. Their median frequency per patient in one cohort is up to 7 times smaller than the frequency of SNVs. Nevertheless, their influence on biological pathways could be significant. Moreover, some patients carry suspected solo-gene mutations and do not carry suspected SNVs.

		COAD	BRCA	BLCA	LUAD	HNSC
# patients with suspected SNVs		399	980	411	560	506
# patients with solo-genes but no SNVs		83	124	2	12	24
# patients with SNVs but no solo-genes		8	4	2	54	6
Solo-genes	Average per patient	26	45.7	45	27.7	23.4
	Median per patient	20	23	26	19	19
SNVs	Average per patient	342.8	57.6	177.1	203.6	110.5
	Median per patient	94	29	123	136	77

Table 3: Altered solo-gene statistics in comparison to single nucleotide variations (SNVs) in five TCGA cohorts.

We ran PRODIGY with the set of solo-genes and SNVs as driver candidates. The rest of the algorithm remained unchanged. We used default parameters. Influence scores were computed for each candidate and a final ranking was done. **Figure 15** shows the performance on three TCGA cohorts. We compared two runs: one used as input solo-genes

and SNVs, with SNV and CNV known CGC genes as gold standard (see “Gold standards section” in “Methods and Results”). The other one used SNV genes as input and SNV known genes as gold standard. **Figure 15a** shows the results of both runs. Notably, the performance remains the same for the bladder cancer cohort, even though the gold standard set is larger. For the HNSC and LUAD cohorts, results with solo-CNVs were slightly inferior.

Figure 15b shows the evaluation of the same driver lists when using refined gold standards; For each cancer cohort, the gold standards were CGC genes with SNVs annotated to be drivers in that particular cancer type. Among SNV and CNV known drivers, these refined sets were of size 16, 20 and 17 for LUAD, HNSC and BLCA, respectively. Among only SNV known drivers, the sets were of sizes 10, 15 and 10, respectively. This evaluation shows an increased performance for the BLCA cohort when using the solo genes in addition to the SNVs as input. Note that in the first run (**Figure 15a**) 20% of the genes in the gold standard (65 out of 313) were annotated with CNV events. In the second run (**Figure 15b**), 41% of the genes in the BLCA-specific set (7 out of 17) were annotated with CNV events. The fact that both evaluations resulted with stable or improved performance, even though the considered gene sets and the gold standards are notably larger, demonstrates that incorporating solo-gene candidates adds a valuable level of information rather than noise to the driver detection process. If solo-genes were not ranked and detected as drivers, we would expect a drop in recall as shown in **Figure 10.** The results further suggest that BLCA cases tend to be affected by solo-gene events more than other cancer cohorts.

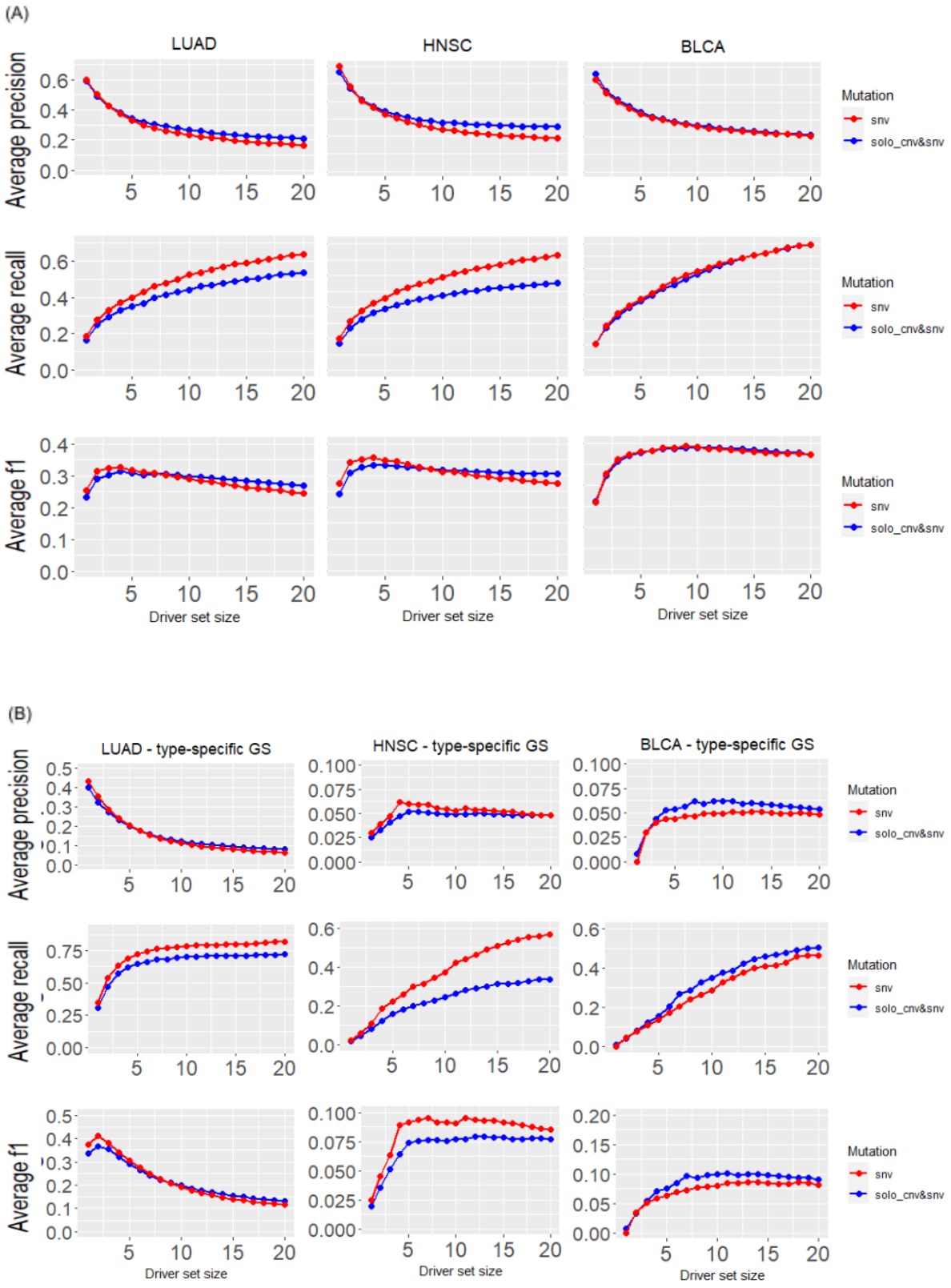


Figure 15: PRODIGY performance when the set of candidate drivers is composed of (1) solo-genes and SNV genes and (2) only SNV genes. Average precision, recall and f1 were measured for x top-scored detected genes in the LUAD, HNSC and BLCA cohort of TCGA. **(A)** Performance when the gold standards are all CNV or a

combination of CNV and SNV genes from CGC. **(B)** Performance when the gold standards are restricted to cancer type-specific genes.

3.3. Test suite 2: Translocations

Genomic segment translocations and gene fusions have the potential of causing expression abnormalities [1,2]. 47% of the known cancer driver genes in the CGC database [13] are associated with translocations. Therefore, we wanted to test the cancer driving potential of these events.

We used The Tumor Fusion Gene Data Portal of The Jackson Laboratory [31], a curated gene fusions dataset of TCGA patients. This study suggested that the translocation and fusion of copy number balanced genes is relatively rare. Instead, most fusions are the result of genomic instabilities. **Figure 16** shows fusion frequencies in different TCGA cohorts and **Table 4** lists the average number of fusions among patients who carry these alterations in five cohorts. In the vast majority of patients, very few fusions were detected and usually only two to four genes are involved. However, almost half of the CGC genes are annotated with translocations as we described earlier (**Table 1**). Therefore, we included both genes involved in translocations and genes with SNV as input in PRODIGY runs with default parameters. We also checked the relative position of genes involved in translocations in the ranked lists to see how dominant they are.

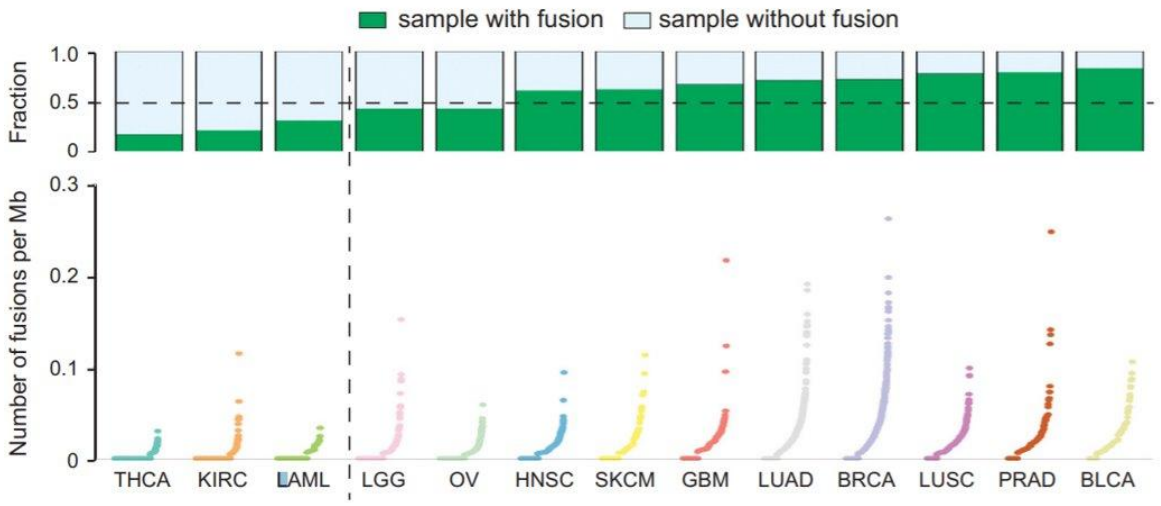


Figure 16: Frequencies of gene fusions across different TCGA cohorts. Source: Yoshihara., Wang, Torres-Garcia et al. [31]

Cancer Type	# patients	Average number of fusions
BLCA	300	3.67
BRCA	838	4.97
COAD	155	1.9
HNSC	342	2.18
LUAD	404	3.63

Table 4: The average number of gene fusions in five TCGA cohorts, as detected by Yoshihara., Wang, Torres-Garcia et al. [31]

Table 5 contains the number of detected driver genes that underwent translocations and their relative ranking when combined in the gold standard set with SNV genes. Most runs resulted with only one translocated driver gene per patient. These drivers were rarely positioned at the top of the ranked lists. Their median position varies between 8 in the BRCA

cohort and 24 in the LUAD cohort. We conclude that in spite of their prominence in CGC, translocations contribute marginally to PRODIGY ranking.

	BLCA	BRCA	COAD	HNSC	LUAD
Median **number** of translocation drivers	1	1	1	0	1
Median **rank** of translocation drivers	23	8	15	8.5	24
Percentage of translocation drivers that are ranked in positions 1 to 10	27%	58%	36%	59%	29%
Percentage of translocation drivers that are ranked in positions 11 to 20	19%	25%	29%	23%	17%

Table 5: Statistics and relative ranking of detected driver genes that underwent translocations. The ranking is when running PRODIGY where the gold standard is SNV genes and genes involved in translocations.

3.4. Test suite 3: Phylogeny-based analysis

As described in the Biological Background section, the earlier a mutation occurred in the evolution of cells, the higher its node would be in the phylogenetic tree. Cancer Cell Fraction (CCF) values describe the fraction of cells that carry each mutation in a certain population, which is correlated with the height of its node in the tree. Mutations that occurred in the same stage of the cancer evolution share the same CCF value.

A simplified version of CCF is the Variant Allele Frequency (VAF), which is a straight-forward calculation based on alternative and reference read counts of point mutations in cancerous cells. For each mutation i:

$$VAF(i) = \frac{\#alternative\ reads\ count}{\#total\ reads\ count}$$

Previous works used mutation VAF or CCF values in the context of cancer driver genes. Wardell et al. [32] filtered out mutated candidates with VAF values lower than 0.15 when detecting driver genes in biliary tract cancers. Ok et al. [33] and Metzeler et al. [34] analyzed VAF values of known driver genes in Acute Myeloid Leukemia patients to support their dominance in comparison to other mutated genes. Hirsch et al. [35] used VAF values to deduce CCF values, reconstruct phylogenetic trees and detect driver genes in pediatric liver cancer. This inspired us to integrate VAF and CCF values in the procedure of driver gene detection as well, with the logic that early occurring mutations are more likely to play a major role in further unusual cell growth. We used patient-specific VAF values and phylogenetic trees and combined them with driver influence scores that are obtained by PRODIGY. We observed a performance improvement thanks to this combination.

Data extraction:

VAF values were obtained from Mutation Annotation Files (MAF) of TCGA cohorts that were downloaded from the Genomic Data Commons (GDC) portal [36]. These files contain the alternative and the reference read counts of each mutation for each patient.

Phylogenetic trees were generated using PhyloWGS Python software [37] (see “Computational Background” section). The inputs we used are alternative and reference read count from MAF files. The outputs are Json format files that describe the number of subclones as nodes in the tree, the sets of mutations that belong to each node, the tree structure and the cellular prevalence of each subclone. Cellular prevalences are given as the percentage of cells that contain the mutations represented by each node.

In order to obtain CCF values, we first calculated the sample purity, i.e. the percentage of cancerous cells in the sample. If the software detects a single tree for the current sample, meaning that all cells emerge from the same clone (root node), the sample purity is the cellular prevalence of the root node. Otherwise, it is the sum of all root node cellular prevalences. Note that the sum of root mutation frequencies is not necessarily equal to 1, as

the sample may be a mixture of normal and tumor cells, and purity is the fraction of tumor cells. Next, CCF values of each node were calculated as follows:

```
CCF(node) = cellular_prevalence(node) / purity
```

All mutations that belong to the same node share its CCF value.

In the rest of this section, we are going to describe our use in the phylogenetic values for personalized driver genes detection.

3.4.1. Removal of low VAF genes

In order to support the conjecture that driver genes tend to have higher VAF values than other mutated genes, we compared the two sets on TCGA cohorts. **Figure 17** shows the distribution of VAF values of known mutated drivers in comparison to all mutated genes in five TCGA cohorts. We analyzed two groups of known CGC drivers; One is all genes with validated driver SNVs, and the other one is the subset of cancer type-specific drivers (see “Gold standards” section under “Methods are Results”). All cohorts show higher VAF values for known driver genes, except for HNSC that shows a similar distribution for all gene groups. Type-specific drivers have the highest values, especially in the BLCA and BRCA cohorts where they form a distinctive and separable distribution in comparison to the background set.

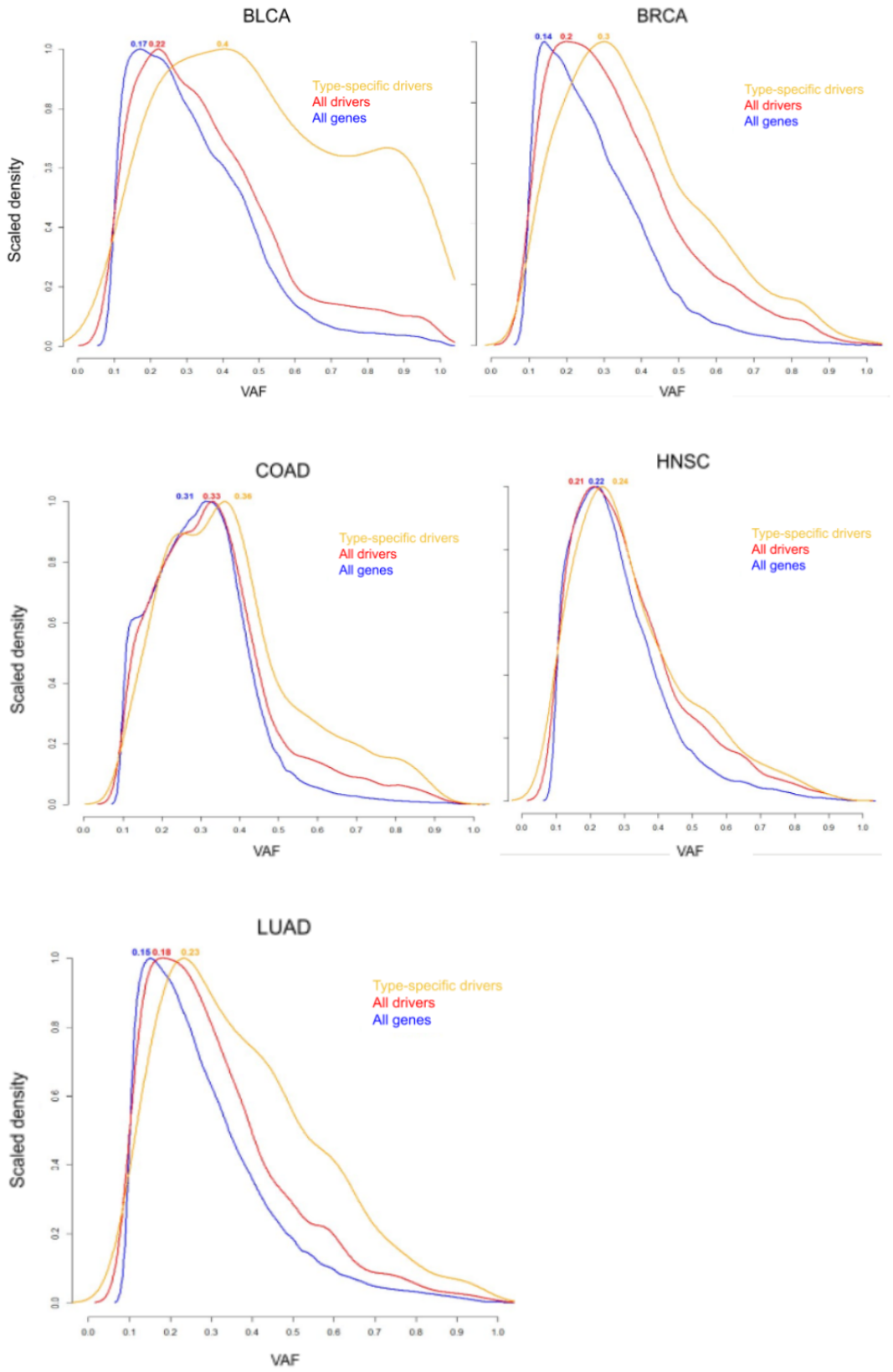


Figure 17: VAF (Variant Allele Frequency) distributions for three types of mutated genes: all CGC known driver genes, cancer type-specific driver genes and all mutated genes. Distributions are shown for five TCGA cancer cohorts. VAF values of highest densities are inlaid.

Inspired by the work of Wardell et al. [32], we removed genes with VAF values lower than 0.15 from the collection of driver candidates for each patient. **Table 6** shows the percentage of disqualified genes out of all candidates. We ran PRODIGY on the reduced gene sets with default parameters. The performance is similar to the case where we use all gene candidates, as shown in **Figure 18** for the BLCA cohort and in **Figure s3** for other TCGA cohorts. We applied lower and higher VAF removal thresholds for comparison, but no improvement was obtained. This suggests that such a removal is not refined enough and causes the loss of eligible candidates. However, the distributions shown in **Figure 17** encouraged us to look for other ways to exploit VAF values for driver genes detection purposes.

	BLCA	BRCA	COAD	HNSC	LUAD
% genes with VAF < 0.15	14%	19%	11%	14%	18%

Table 6: The percentage of mutated genes with VAF values smaller than 0.15, which were removed from the driver candidates collection.

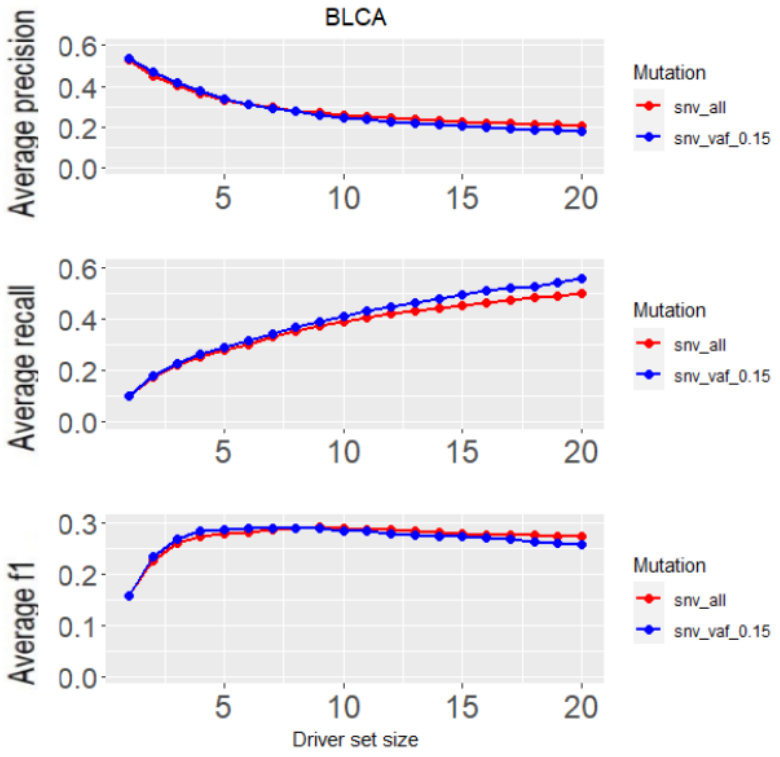


Figure 18: PRODIGY performance when the set of candidate drivers is composed of SNV mutated genes with VAF ≥ 0.15 in comparison to all SNV mutated genes. Average precision, recall and f1 were measured for x top-scored detected genes in the BLCA cohort of TCGA, for 1 ≤ x ≤ 20.

3.4.2. Combined pathway and phylogeny scores

We wished to modify PRODIGY's driver scores by integrating them with VAF values. First, the positive scores were normalized to be on the same scale as the VAF values, which are in range $[0, 1]$. For each driver, we assigned:

$$PRODIGY\_score(driver) := \frac{PRODIGY\_score(driver)}{max_{d \in drivers}(PRODIGY\_score(d))}$$

where $drivers$ is the set of genes with positive PRODIGY scores. Then, we created new scores using two formulas:

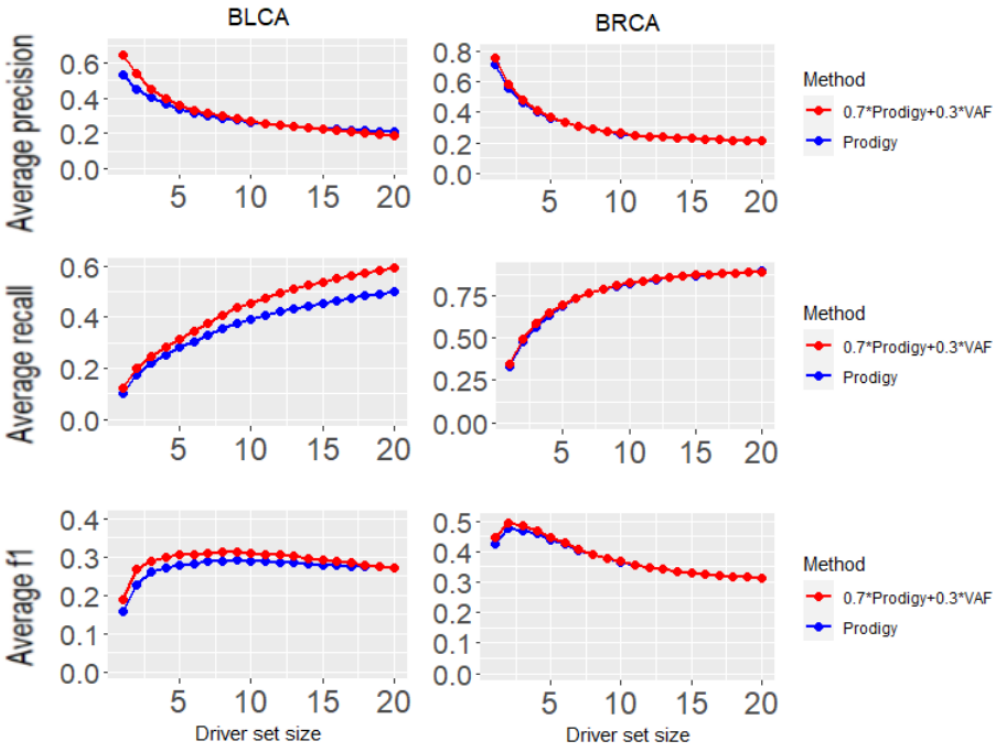
1. For $n \in \{2, 3, 4, 5\}$:

$$new\_score(driver) := PRODIGY\_score(driver) * \sqrt[n]{VAF(driver)}$$

2. For $\alpha \in \{0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9\}$:

$$new\_score(driver) := \alpha * PRODIGY\_SCORE(driver) + (1 - \alpha) * VAF(driver)$$

These combinations rely on PRODIGY scores, while giving priority to mutated genes that are more frequent across the cells. We generated new scores in five cancer cohorts and measured precision, recall and f1 to compare them with plain PRODIGY ranking. Best results were obtained when using the second formula with $\alpha = 0.7$. As shown in **Figure 19**, the performance is improved for all cancer cohorts. The BLCA cohort shows the highest improvement. When using the first formula, best results were obtained with $n = 5$, but they were inferior in comparison to the results using the second formula.



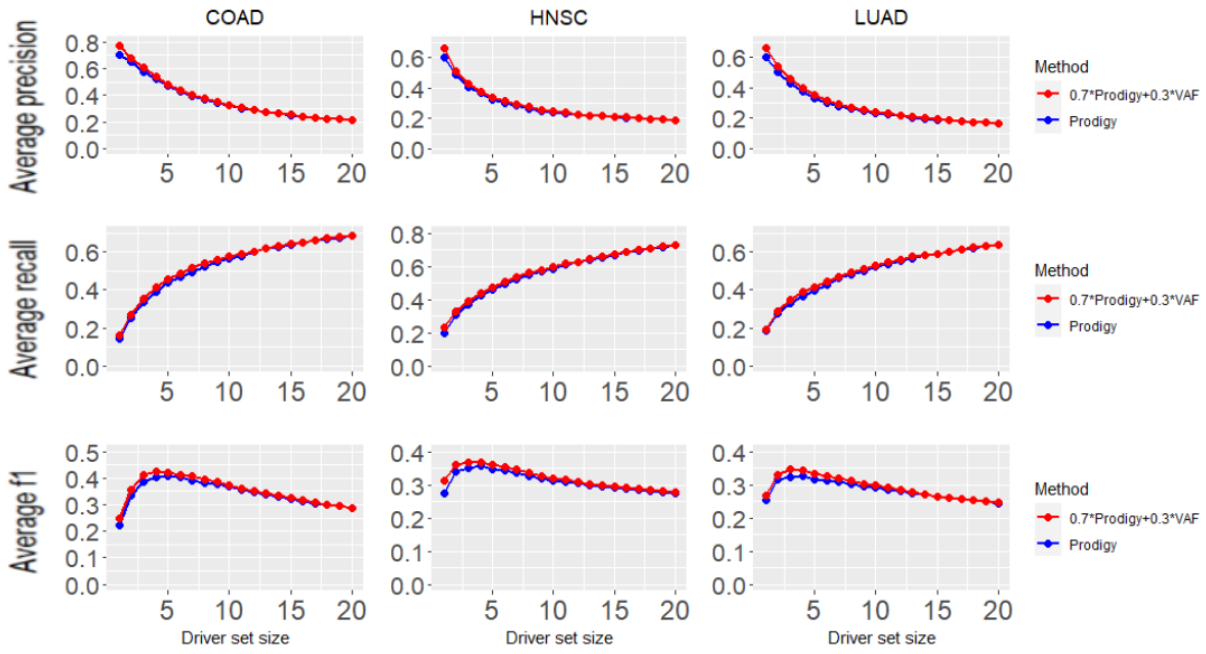


Figure 19: Driver ranking performance when driver scores are a combination of PRODIGY scores and VAF values in comparison to plain PRODIGY scores. Average precision, recall and f1 were measured for x top-scored detected genes in five TCGA cohorts, for $1 \leq x \leq 20$.

As described in the “Computational Background” section, CCF values represent the fraction of cancerous cells that share a specific mutation in a phylogenetic-adjusted manner. We used the PhyloWGS algorithm to compute CCF values based on VAF values in our five TCGA cancer cohorts, in order to derive new driver scores that incorporate the phylogenetic aspect. The PhyloWGS algorithm generated a single tree for all patients except for one LUAD cancer patient, for which two trees were generated. As shown in **Figure 20**, all phylogenetic trees are composed of a small number of nodes ranging between 1 and 6, where most are composed of 3 or 4 nodes. The mutations in each node are represented by a single CCF value. This allows us to replace the noisy, continuous VAF values for individual genes with a small categorical set of values, each applied to all mutated genes in the same tree node..

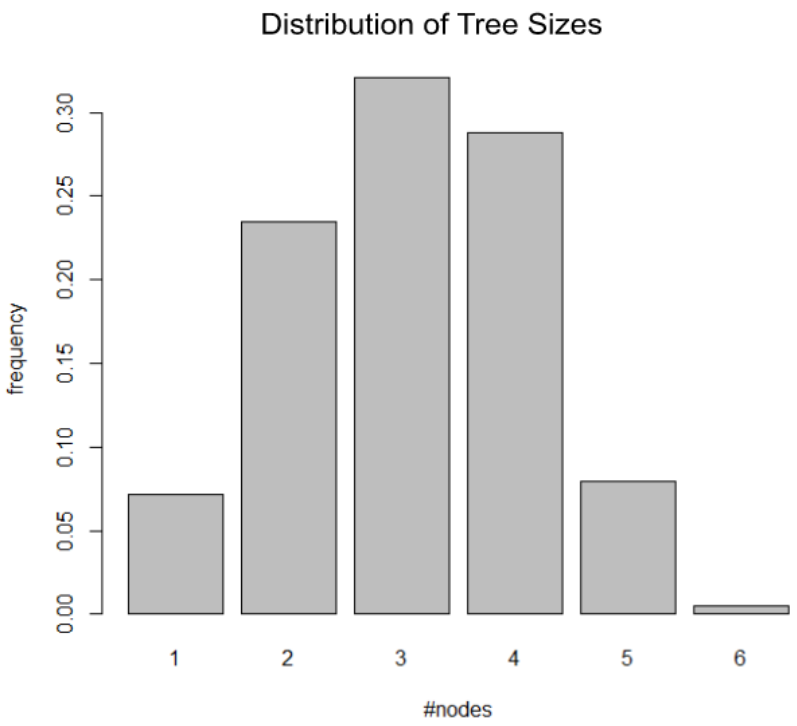


Figure 20: Frequencies of phylogenetic tree sizes in five TCGA cancer cohorts: BLCA, BRCA, COAD, HNSC and LUAD. The trees are patient-specific, as calculated by PhyloWGS with SNV input data. The number of nodes is also the number of distinct CCF values per patient.

We computed new driver scores that incorporate CCF values rather than VAF values, using the aforementioned formula:

$$new\_score(driver) := \alpha * PRODIGY\_SCORE(driver) + (1 - \alpha) * CCF(driver)$$

Figure 21A shows F1 scores of the BLCA cohort when using the full gold standard collection. Other cohort results are in **Figure 4s** under the "Supplementary Material" section. In spite of the diversity reduction across different drivers of the same patient, these scores achieved a similar performance improvement to the VAF incorporated results. Furthermore, when using cancer type-specific gold standards (see "Driver list evaluation" under "Methods and Results" section), the improvement seems to be even more significant. **Figure 21B** shows F1 scores for our TCGA cohorts when using type-specific gold standards.

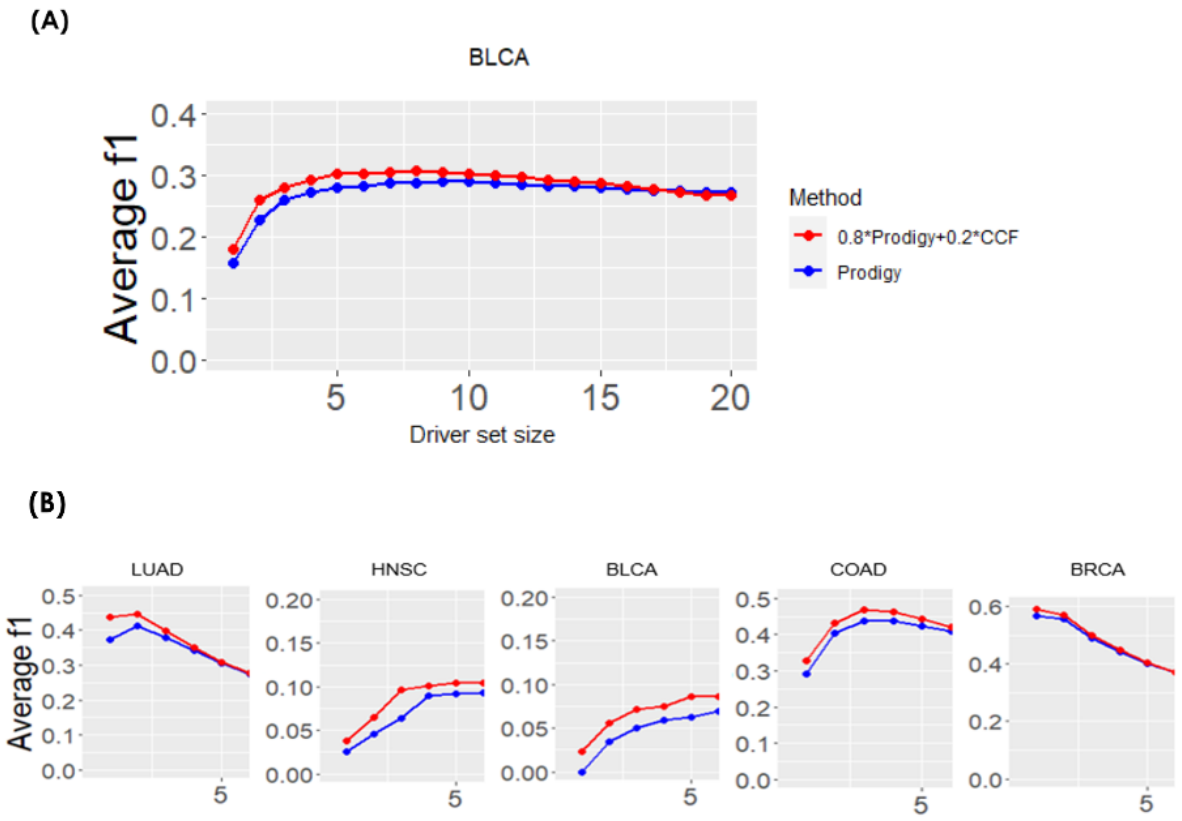


Figure 21: Driver ranking performance when driver scores are a combination of PRODIGY scores and CCF values in comparison to plain PRODIGY scores. F1 was measured for x top-scored detected genes in 5 TCGA cohorts, for 1 ≤ x ≤ 20 or 1 ≤ x ≤ 5. (A) Results when gold standards are the full CGC validated collection. (B) Results when gold standards are the type-specific subsets.

3.4.3. Clonal mutations analysis

Our analyses above show that drivers tend to have higher CCF values than other mutated genes, as expected. Furthermore, drivers that initiate the cancerous process are expected to belong to the root of the phylogenetic tree, since they are among the first to be mutated in the normal cell population. Genes mutated at the root of a tree are called clonal, and satisfy $CCF(root\_gene) = 1$ if only one tree is generated for the patient. Otherwise, the sum of root node CCFs is equal to 1. Since single trees were generated for all patients but one in our five cohorts, we excluded this patient from our analysis.

Table 7 describes the fraction of genes with CCF value of 1 per PRODIGY ranking level across patients. Ranking levels refer to the position of genes in patient driver lists sorted from highest to lowest influence scores. We can see that ranking levels are closely related to gene clonality. The fraction of clonal genes is highest for the top ranking level in all cancer cohorts, and it decreases as the ranking decreases. In other words, clonal drivers tend to have higher influence values.

Ranking level	1	2	3	4	5	6	7	8	9	10
BLCA	0.45	0.41	0.39	0.4	0.38	0.35	0.33	0.34	0.32	0.36
BRCA	0.61	0.47	0.44	0.42	0.38	0.4	0.37	0.35	0.35	0.34
COAD	0.53	0.42	0.41	0.36	0.35	0.34	0.36	0.33	0.37	0.31
HNSC	0.42	0.36	0.35	0.3	0.32	0.33	0.3	0.27	0.28	0.27
LUAD	0.38	0.34	0.34	0.28	0.3	0.25	0.27	0.29	0.29	0.22

Table 7: The fraction of clonal genes per PRODIGY ranking level across cancer patients.

Bearing this in mind, we ran PRODIGY with default parameters using only clonal genes as driver candidates. **Figure 22** shows the performance for the BLCA cohort as a representative example. Precision, recall and F1 scores decreased for all cohorts. This suggests that while clonal drivers tend to have high impact on biological pathways, other drivers that are mutated later in evolution contribute substantially to cancer progression as well.

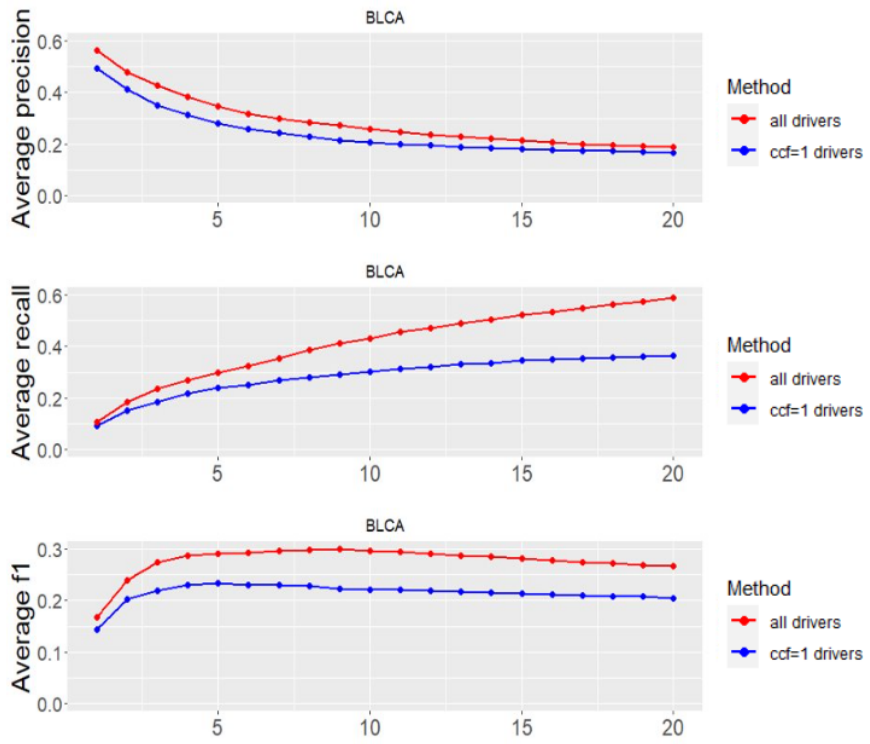


Figure 22: BLCA driver ranking performance when the set of candidates is composed only of clonal genes in comparison to all genes.

Note that the set of clonal genes is markedly reduced, containing on average only 34% of all mutated genes. In view of that, the performance using only them is surprisingly good.

3.4.4. Double layered ranking

In another attempt to combine PRODIGY scores with CCF values, we followed a double layered ranking. We used PRODIGY scores as a primary ranking and internally sorted gene subsets with near equal PRODIGY scores by their CCF values. This allows a better separation of genes that have very similar influence scores but were mutated in different stages during the cancer evolution.

For each patient, we set a separation threshold ε and applied the following steps:

A. *Decreasingly sort genes by their PRODIGY scores in a list* l. *The first element in* l *is* $l[0]$.

B. *Initialize a new ranked list by* $new\_l = []$.

C. *Set* $i = 1$, $G_1 = \{l[0]\}$, $prev = score(l[0])$.

D. *For each (gene) in* l *from* $l[1]$ *to last element:*

  a. *if* $score(gene) + \varepsilon \leq prev$*:*

    i. $G_i = G_i \cup \{gene\}$

  b. *if* $score(gene) + \varepsilon > prev$ *or (gene) is last element in* l*:*

    i. *Internally sort* G_i *by decreasing CCF values.*

    ii. *Append the sorted sublist to* new_l.

    iii. *Assign* $i = i + 1$, $G_i = \{gene\}$

  c. $prev = score(gene)$

E. *Return* new_l.

We used several ε thresholds. Best results were obtained with $\varepsilon = 0.5$. **Figure 23** shows BLCA results. Other cohort results are in **Figure 5s**. Precision, recall and F1 scores were higher than the original PRODIGY ranking. However, they were lower than the best results we achieved with the combined CCF and PRODIGY scores, as described in section 4.4.2.

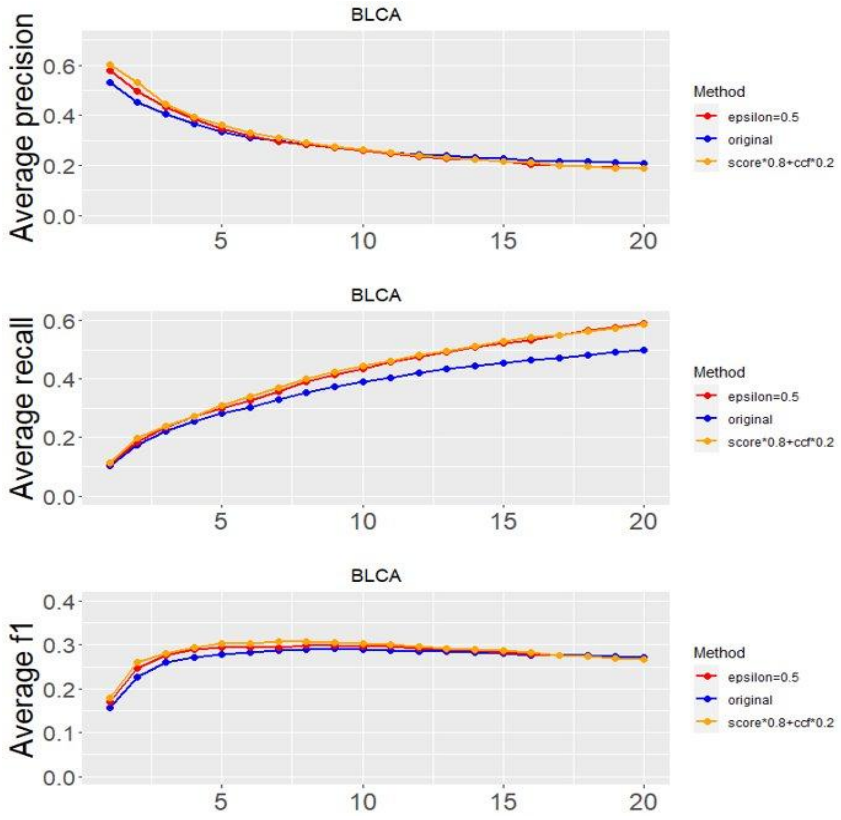


Figure 23: BLCA driver ranking performance with the double layered ranking method and $\varepsilon = 0.5$, in comparison to the original PRODIGY ranking and the combined scores of CCF values and PRODIGY.

4. Discussion

Driver genes trigger or promote cancer through various types of mutations in the genome. While some of the drivers are well known and studied as they are often mutated among patients, other driver mutations are less frequent and some are still unknown today. The latter could also cause abnormal cell growth by affecting key biological processes. In addition, the mutated genes in a particular patient's tumor may contain many known drivers, where only a few of them actually drive the patient's tumorigenesis. Knowledge of the underlying mechanisms and specifically the driver mutations that led to a certain cancer case allows personalized medical treatment. For this reason, developing algorithms that process data of individuals in an unbiased manner is a necessity and a mission of high importance.

This work describes methods to detect patient-specific driver genes and analyses of mutational data. We performed multiple tests in which we characterized different types of mutations and attempted to reveal drivers among them. For the most part, we fed the PRODIGY algorithm [26] with our analyzed data to obtain mutated gene scores and evaluate input candidates. We obtained an improvement over the original results of PRODIGY, which used only SNVs, when incorporating considerations of cancer evolution. Using the inferred cancer phylogenetic trees, we improved the performance of the original PRODIGY with SNVs. We can further suggest the integration of CNVs in a similar way.

The first methods we introduced aim to create a robust set of driver candidates out of CNV data. When trying to use the full collection of genes that underwent copy number alterations, we observed low performance, which suggests that the data is very noisy. This led us to refine the set of driver candidates. As a first attempt, we examined only copy number changes of high amplitudes, called major CNVs. When incorporating them alongside SNVs, the performance increased significantly in comparison to using all CNV candidates, but was still lower than when using only SNVs. Next, we examined genes that are jointly altered

within a CNV segment. We observed that 70% of the segments with significant CNVs contain only a single gene. We termed these as *solo-genes* and tested them as driver candidates together with SNVs. The performance remained similar to runs that contained only SNVs in the BLCA (bladder cancer) cohort of the TCGA, even though the collection of gold standards for comparison increased by 26%. When narrowing down the list of driver genes in the gold standard to include only known driver genes associated with BLCA, we achieved an improvement and outperformed runs containing only SNV candidates. This demonstrates that incorporating solo-genes might reveal additional true influential drivers.

Next, we examined gene pairs that underwent translocations. Even though translocations comprise 47% of the known drivers in The COSMIC Cancer Gene Census (CGC), they are relatively rare among cancer patients in the TCGA cohort. Typically translocations involve only 2 to 4 genes. When we tested these genes as candidates together with SNVs, their median position in the ranked drivers list varied from 8 to 24 in five cancer cohorts. Hence, in spite of their high prevalence in CGC, translocations did not improve our results.

Finally, we used phylogeny to improve the list of obtained SNV driver genes. We introduced two key mutation frequency estimators: VAF (Variant Allele Frequency), which is the fraction of variant alleles out of the total counts of a specific mutation, and CCF (Cancer Cell Fraction), which is the fraction of cells carrying a specific mutation out of all cancer cells. First, since drivers tend to occur early in the evolution of cancer and therefore may have higher VAF values than many other mutations, we filtered out driver candidates with VAFs lower than set thresholds. We did not obtain a performance improvement, suggesting that genes with low VAFs are eligible driver candidates. Instead, we generated gene scores that combined both VAF and PRODIGY influence scores. This combination led to a breakthrough; We noticed performance improvements in comparison to plain PRODIGY runs in all cancer types and most prominently in the BLCA cohort. The performance is further increased when using cancer type-specific gold standards in each of the five cancer cohorts we examined. As an additional step, we used the PhyloWGS algorithm to obtain

phylogenetic trees of SNVs in individual patients. These trees enabled the computation of mutational CCF values. Typically, a single tree with only 1 to 6 subclones was generated, with the majority of them with 3 or 4 nodes. Still, when creating a combined gene score out of CCF values and PRODIGY influence scores, the performance improved similarly to the trend shown with VAF values. This suggests that even though the number of distinct CCF values is very small for a single patient, these values are meaningful and could improve the prediction of driver genes. Mutations of higher tree nodes occur earlier in evolution and indeed tend to be more influential than later occurring mutations.

Other phylogeny related tests included filtering out mutations that are not present in the root node and a new double-layered ranking method of driver genes, which uses CCF values to internally sort drivers with similar PRODIGY influence scores. Root drivers removal decreased the performance. This suggests that highly influential mutations could occur later in evolution even though most of them occur early, similarly to our observation when removing genes with low VAF values. In the double-layered ranking method, we first prioritized genes with high PRODIGY influence scores and then internally sorted them in bins by CCF values. We did not achieve an improvement.

We performed additional tests to improve driver gene predictions that are not elaborated in the text. A lot of effort was invested to obtain phylogenetic trees that account for both CNVs and SNVs rather than only SNVs by PhyloWGS. We wished to incorporate these two types of mutations and run a unified driver detection algorithm. To this end, we ran the Battenberg algorithm [44], that generates copy number inputs for the use of PhyloWGS. Unfortunately, the computational complexity of that algorithm is extremely high and we could not obtain the outputs in a reasonable time. In another attempt, we narrowed down the list of pathways that are checked for perturbations by PRODIGY to contain only cancer-related pathways. The pathways were detected by a hyper-geometric test for enrichment with known cancer driver genes, both from CGC and NCG driver databases. The results were very similar to the experiments where all pathways were used.

Future work could further explore and utilize the phylogenetic relationship of CNV and SNV events and create an inclusive method to detect drivers using evolutionary aspects. In addition, the notation of mutual exclusivity of driver genes could be used, where mutations in different genes within the same pathways could be redundant and cause the same driving effect.

5. References

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* 458, 719– 724 (2009). https://doi.org/10.1038/nature07943

2. Li, Y., Roberts, N. D., Wala, J. A. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* 578(7793), 112–121 (2020). https://doi.org/10.1038/s41586-019-1913-9

3. Jones, S., & Thornton, J. M. Protein-protein interactions: a review of protein dimer structures. *Progress in biophysics and molecular biology*, 63(1), 31-65 (1995). https://doi.org/10.1016/0079-6107(94)00008-W

4. Spirin, V., & Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proceedings of the national Academy of sciences* 100(21), 12123-12128 (2003). https://doi.org/10.1073/pnas.2032324100

5. Rao, V. S., Srinivas, K., Sujini, G. N., & Kumar, G. N. Protein-protein interaction detection: methods and analysis. *International journal of proteomics* (2014). http://dx.doi.org/10.1155/2014/147648

6. Mering, C. V., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., & Snel, B. STRING: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1), 258-261 (2003). https://doi.org/10.1093/nar/gkg034

7. Kanehisa, Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K.; KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 45, D353-D361 (2017). https://doi.org/10.1093/nar/gkw1092

8. Gillespie, M., Vastrik, I., Eustachio, P. D., Schmidt, E. & Bono, B. De. Reactome : a knowledgebase of biological pathways. *Nucleic Acids Res*. 33, 428–432 (2005). https://doi.org/10.1093/nar/gki072

9. Stratton, M. R., Campbell, P. J., & Futreal, P. A. The cancer genome. *Nature*, 458(7239), 719–724 (2009). https://doi.org/10.1038/nature07943

10. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians,* 71(3), 209-249 (2021). https://doi.org/10.3322/caac.21660

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

12. Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., & Stratton, M. R. A census of human cancer genes. *Nature reviews. Cancer*, 4(3), 177–183 (2004). https://doi.org/10.1038/nrc1299

13. Sondka, Z., Bamford, S., Cole, C.G. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* 18, 696–705 (2018). https://doi.org/10.1038/s41568-018-0060-1

14. Dressler, L., Bortolomeazzi, M., Keddar, M.R. *et al.* Comparative assessment of genes driving cancer and somatic evolution in non-cancer tissues: an update of the Network of Cancer Genes (NCG) resource. *Genome Biol* 23, 35 (2022). https://doi.org/10.1186/s13059-022-02607-z

15. Yachida, S. *et al*. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature, 467(7319), 1114–1117 (2012). https://doi.org/10.1038/nature09515

16. Weinstein, J. N. *et al*. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* 45, 1113–1120 (2013). https://doi.org/10.1038/ng.2764

17. Goldman, M.J., Craft, B., Hastie, M. *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* (2020). https://doi.org/10.1038/s41587-020-0546-8

18. Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology*, 12(4), R41 (2011). https://doi.org/10.1186/gb-2011-12-4-r41

19. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, *et al*. Software for Computing and Annotating Genomic Ranges. *PLoS Comput Biol,* 9(8) (2013). https://doi.org/10.1371/journal.pcbi.1003118

20. Schneider, Valerie A., *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research.* 27.5, 849-864 (2017). http://doi.org/10.1101/gr.213611.116

21. Bedard, P. L., Hansen, A. R., Ratain, M. J., & Siu, L. L. Tumour heterogeneity in the clinic. *Nature*, 501(7467), 355-364 (2013).

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

23. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22, 398–406 (2012). http://www.genome.org/cgi/doi/10.1101/gr.125567.111

24. Leiserson, M. D. M. et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114 (2014). https://doi.org/10.1038/ng.3168

25. Hou, J. P. & Ma, J. DawnRank: Discovering personalized driver genes in cancer. *Genome Med.* 6, 1–16 (2014).

26. Guo, W. F. et al. Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. *Bioinformatics* 34, 1893–1903 (2018).

27. Dinstag, G., & Shamir, R. PRODIGY: personalized prioritization of driver genes. *Bioinformatics*, 36(6), 1831-1839 (2020).

28. Ljubić I. et al. An algorithmic framework for the exact solution of the prize-collecting Steiner tree problem. *Math. Program.*, 105, 427–449 (2006).

29. Zhang, T., Zhang, S. W., & Li, Y. Identifying driver genes for individual patients through inductive matrix completion. *Bioinformatics,* 37(23), 4477-4484 (2021).

30. Venkatraman, E. S., & Olshen, A. B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics (Oxford, England)*, 23(6), 657–663 (2007). https://doi.org/10.1093/bioinformatics/btl646

31. Yoshihara, K., Wang, Q., Torres-Garcia, W. *et al*. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene.* 34, 4845–4854 (2015). https://doi.org/10.1038/onc.2014.406

32. Wardell, Christopher P., Fujita, M., Yamada, T. *et al.* Genomic characterization of biliary tract cancers identifies driver genes and predisposing mutations. *Journal of Hepatology*. 68(5), 959-969 (2018). https://doi.org/10.1016/j.jhep.2018.01.009

33. Ok, C.Y., Trowell, K.T., Parker, K.G. *et al.* Chronic myeloid neoplasms harboring concomitant mutations in myeloproliferative neoplasm driver genes (*JAK2/MPL/CALR*) and *SF3B1*. *Mod Pathol* 34, 20–31 (2021). https://doi.org/10.1038/s41379-020-0624-y

34. Metzeler, K. H. *et al.* Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. *Blood, The Journal of the American Society of Hematology*, 128(5), 686-698 (2016). https://doi.org/10.1182/blood-2016-01-693879

35. Hirsch, T. Z., *et al.* Integrated genomic analysis identifies driver genes and cisplatin-resistant progenitor phenotype in pediatric liver cancer. *Cancer discovery* 11(10), 2524-2543 (2021). https://doi.org/10.1158/2159-8290.CD-20-1809

36. Mark, A. J., Vincent F., Robert L. G., Louis M. S. The NCI Genomic Data Commons as an engine for precision medicine. *Blood*, 130(4), 453-459 (2017). https://doi.org/10.1182/blood-2017-03-735654

37. Deshwar, A.G., Vembu, S., Yung, C.K. *et al*. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology* 16, 35 (2015). https://doi.org/10.1186/s13059-015-0602-8

38. Strino, F., Parisi, F., Micsinai, M., & Kluger, Y. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic acids research,* 41(17), e165-e165 (2013). https://doi.org/10.1093/nar/gkt641

39. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L., & Morris, Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics*, 15(1), 1-16 (2014). https://doi.org/10.1186/1471-2105-15-35

40. Hajirasouliha, I., Mahmoody, A., & Raphael, B. J. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30(12), i78-i86 (2014). https://doi.org/10.1093/bioinformatics/btu284

41. Oesper, L., Mahmoody, A., & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome biology*, 14(7), 1-21 (2013). https://doi.org/10.1186/gb-2013-14-7-r80

42. Chen, M., Gunel, M., & Zhao, H. SomatiCA: identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data. *PloS one,* 8(11), e78143 (2013). https://doi.org/10.1371/journal.pone.0078143

43. Ghahramani, Z., Jordan, M., & Adams, R. P. Tree-structured stick breaking for hierarchical data. *Advances in neural information processing systems*, 23 (2010).

44. Nik-Zainal, S., et al. Breast Cancer Working Group of the International Cancer Genome Consortium. The life history of 21 breast cancers. *Cell,* 149(5), 994–1007 (2012). https://doi.org/10.1016/j.cell.2012.04.023

6. Supplementary Material

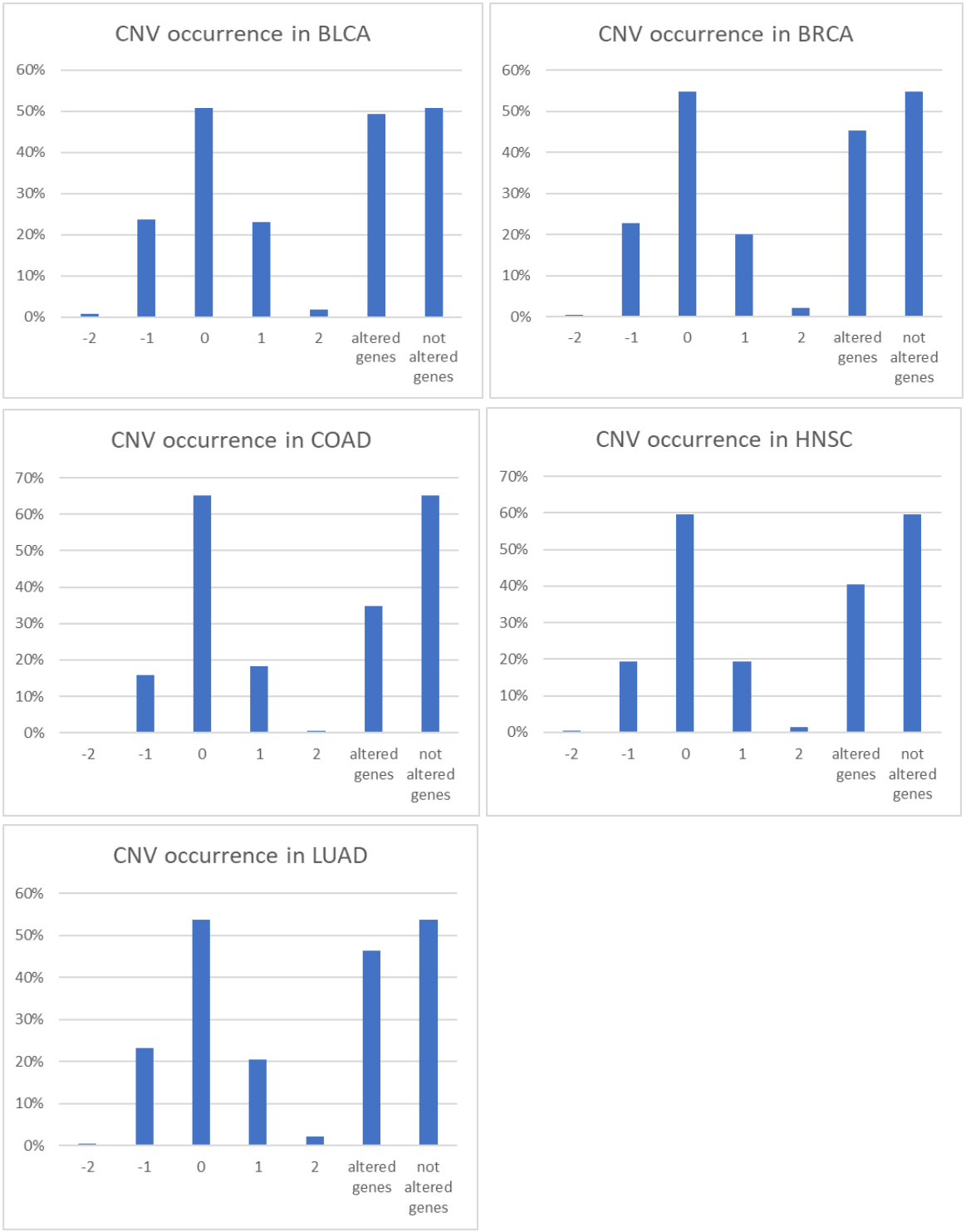


Figure s1: Percentage of genes that underwent different types of copy number variations in each of five TCGA cohorts: *BLCA* - bladder urothelial carcinoma, *BRCA* - breast invasive carcinoma, *COAD* - colon adenocarcinoma, *HNSC* - head and neck squamous cell carcinoma, *LUAD* - lung adenocarcinoma. Copy number Variations: *-2* - homozygous deletion, *-1* - single copy deletion, *0* - normal copy number, *1* - low-level amplification, *2* - high-level amplification. Altered genes are those with any CNV, i.e., type ≠ 0.

Distribution of lengths of Copy Number Altered Segments

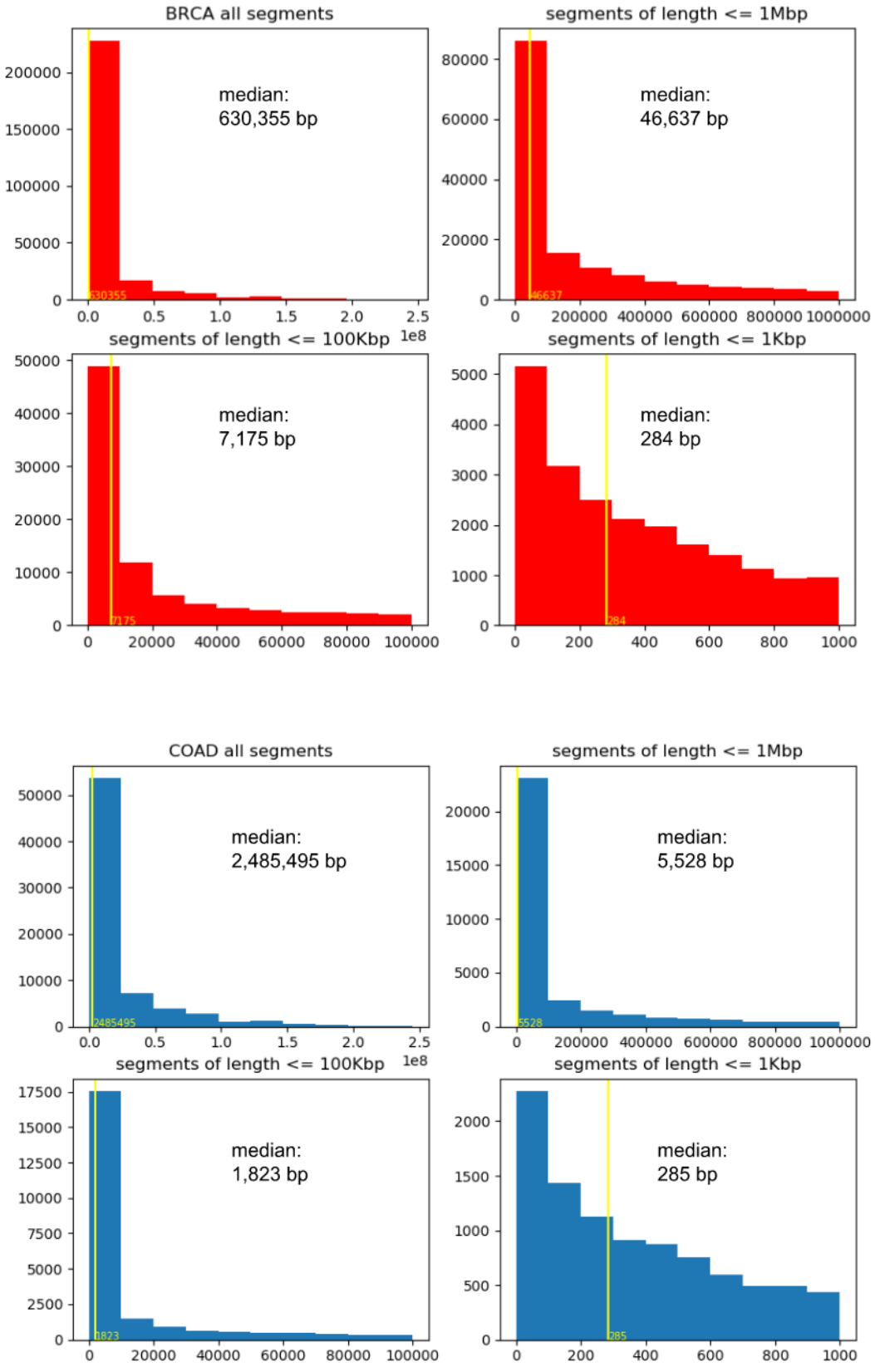


Figure s2, part 1 (description in next page)

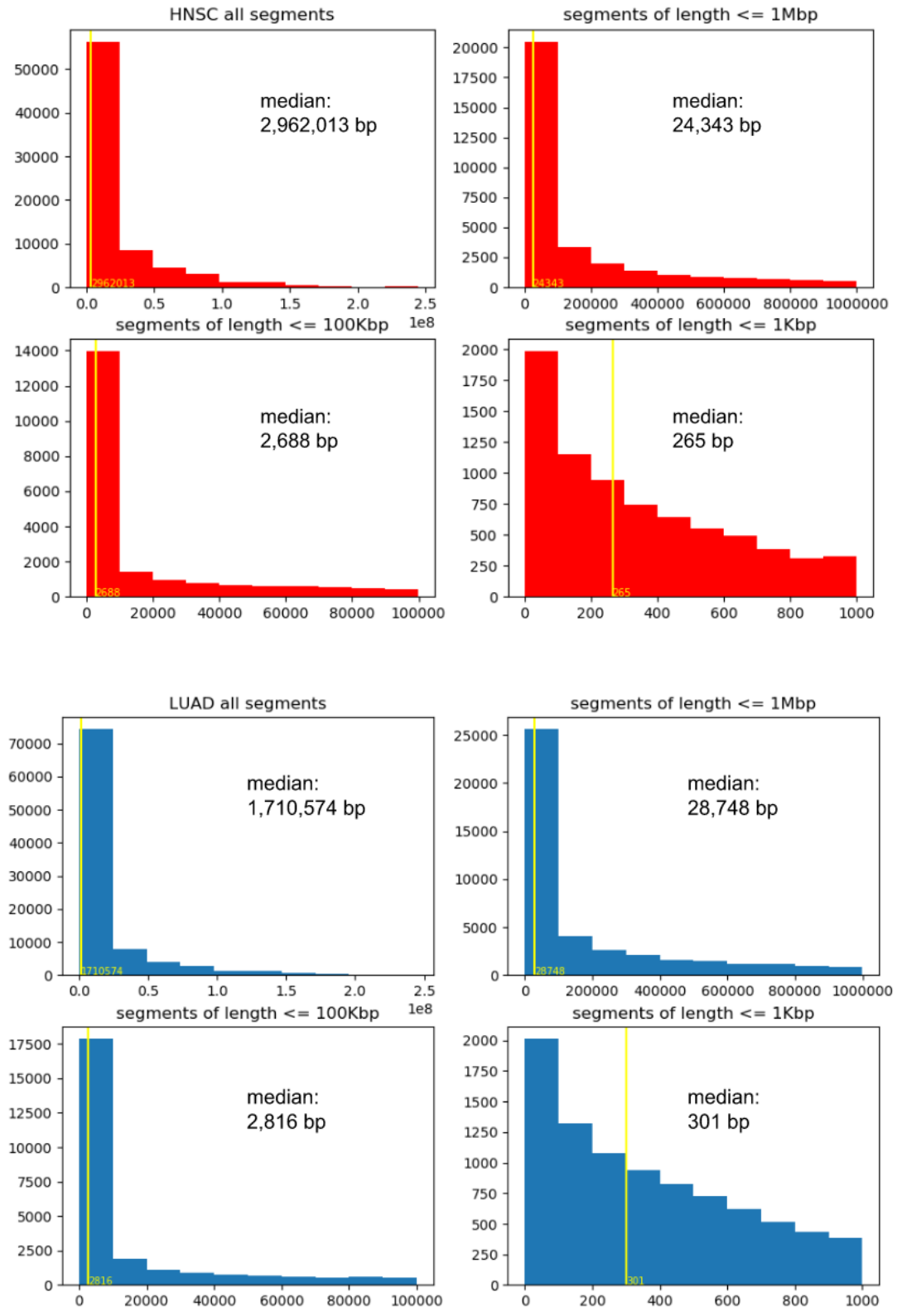


Figure s2: Distribution of lengths of copy number altered segments in basepairs of four cancer cohorts (BLCA histograms are in **Figure 13**). Four histograms are produced for each cohort: all segment lengths, segments of length ≤ 1Mbp, segments of length ≤ 100Kbp and segments of length ≤ 1Kbp. Median lengths are inlaid.

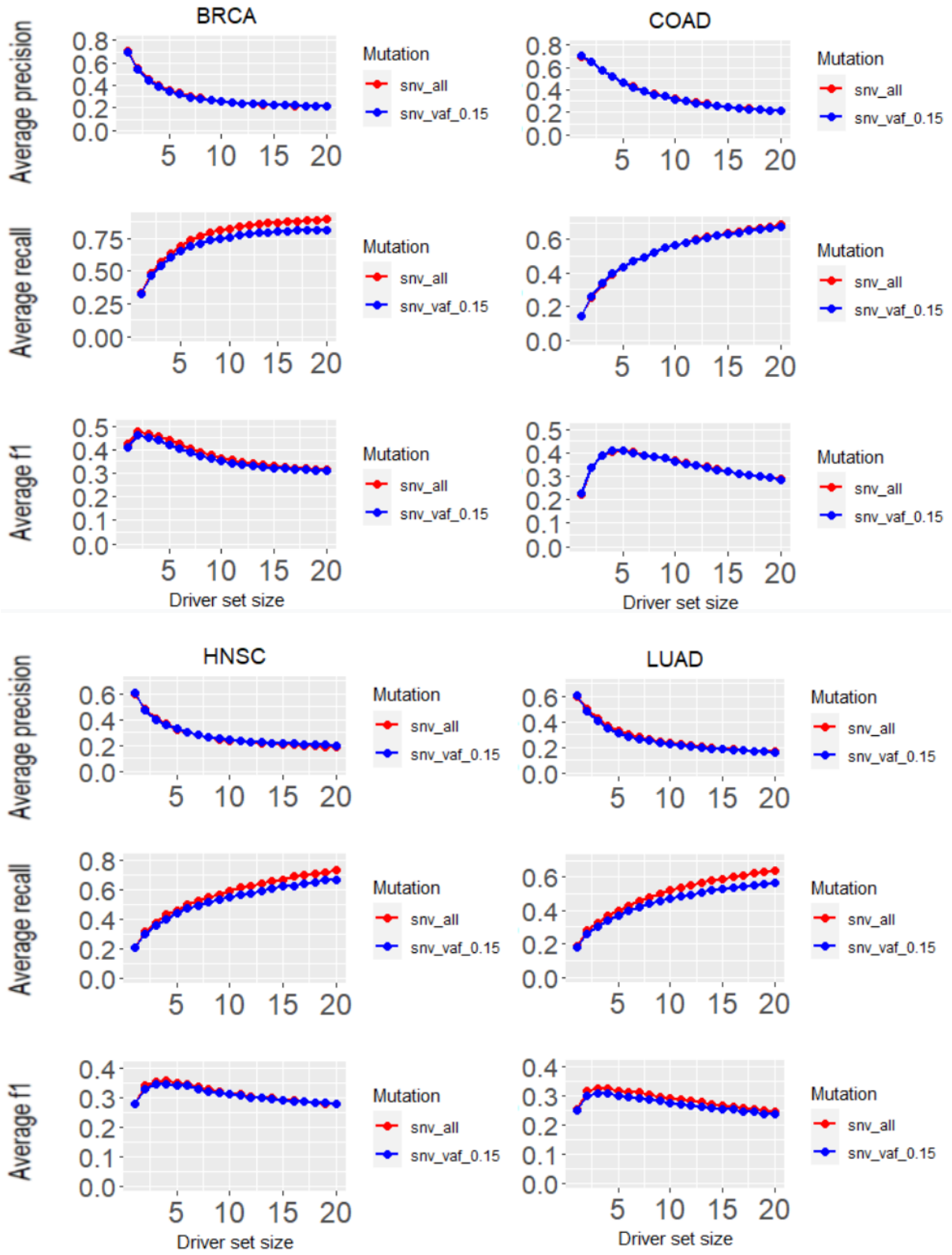


Figure 3s: PRODIGY performance when the set of candidate drivers is composed of SNV mutated genes with VAF ≥ 0.15 in comparison to all SNV mutated genes. Average precision, recall and f1 were measured for x top-scored detected genes in four cancer cohorts, for 1 ≤ x ≤ 20 (BLCA performance is shown in **Figure 18**).

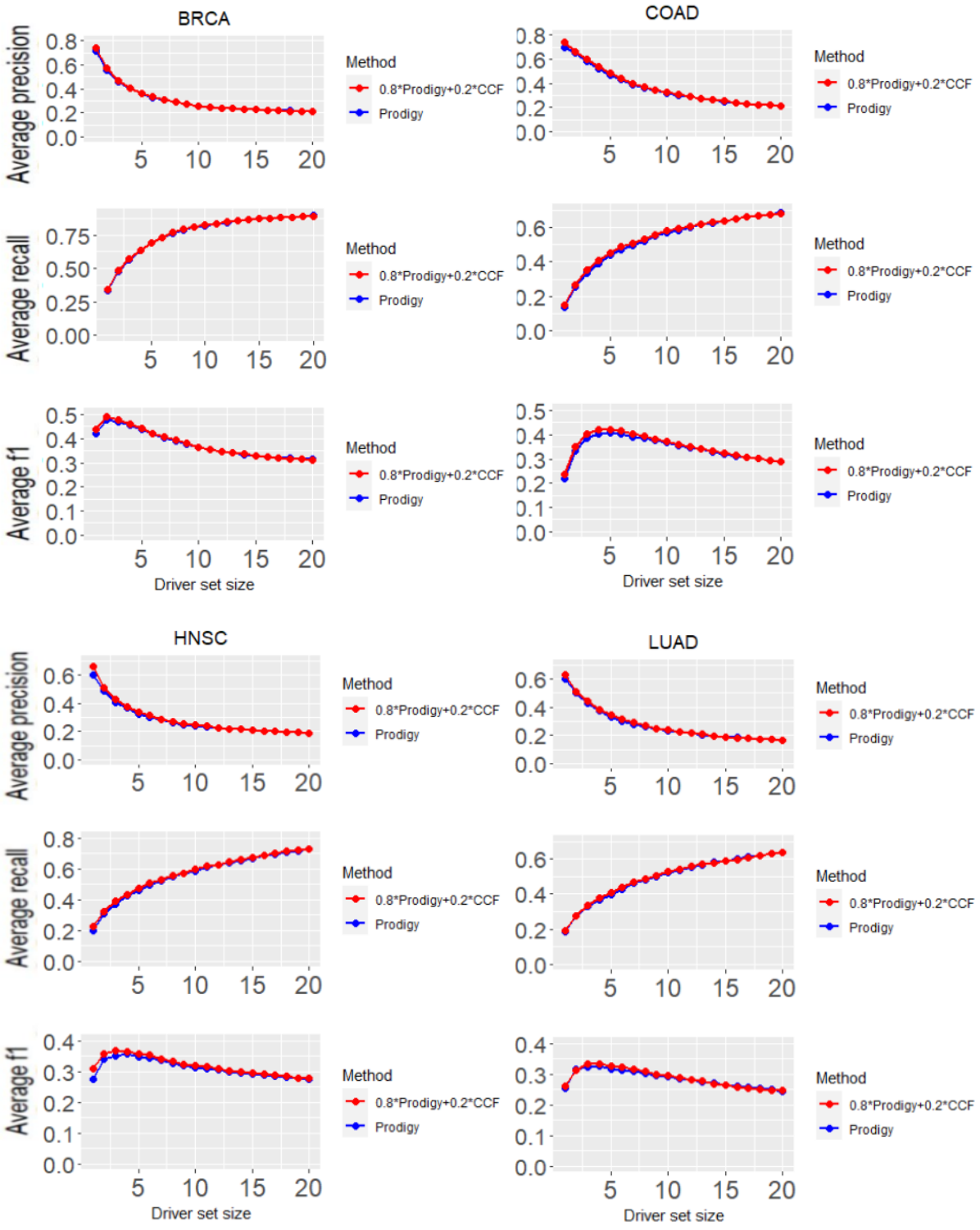


Figure 4s: Driver ranking performance when driver scores are a combination of PRODIGY scores and CCF values in comparison to plain PRODIGY scores. F1 was measured for x top-scored detected genes in four TCGA cohorts, for 1 ≤ x ≤ 20. Gold standards are the full CGC validated collection (BLCA performance and type-specific gold standards results are shown in **Figure 21**).

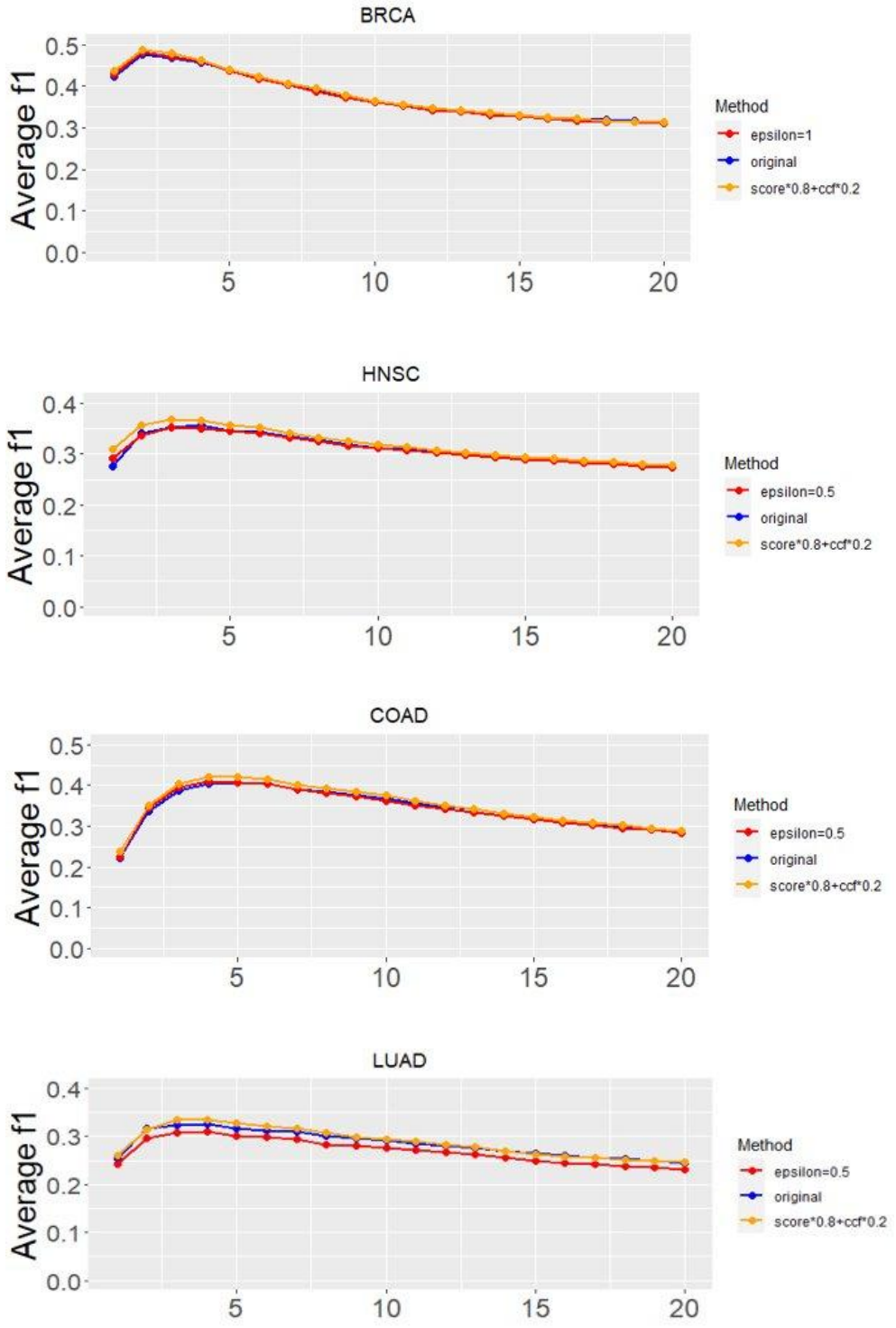


Figure 5s: Driver ranking performance with the double layered ranking method and $\varepsilon = 0.5$, in comparison to the original PRODIGY ranking and the combined scores of CCF values and PRODIGY, in four TCGA cohorts (BLCA performance is shown in **Figure 23**).

תקציר

רקע

מחלת הסרטן היא מהנפוצות בעולם ואחד מגורמי המוות השכיחים ביותר. המחלה נגרמת ע"י מוטציות בחומר התורשתי (ד.נ.א.) של תאים סומטיים, אשר גורמות לשיבוש תהליכים ביולוגיים שקשורים בשרידות התא, בגורל התא ובתחזוקת הגנום. כתוצאה מכך, התאים משתכפלים ללא שליטה ויוצרים גידולים אשר פולשים לתוך רקמות סמוכות ועלולים לשלוח גרורות לעבר איברים מרוחקים. הסוגים העיקריים של מוטציות כוללים החלפות נקודתיות של בסיסים, מחיקות או הכפלות של מקטעים גנומיים, הגורמות לשינוי במספר העותקים שלהם בתא, והתקות של מקטעים גנומיים מנקודה אחת אל נקודה אחרת. בגנום, רב המוטציות אינן מזיקות, אך מיעוטן מעודדות התפתחות של סרטן ע"י שיבוש מנגנונים תאיים כאמור לעיל. המוטציות המזיקות נקראות מוטציות נהג (driver mutations). זיהוין הוא מרכיב מפתח לטיפול נכון ומותאם אישית במחלה.

שיטות ותוצאות

העבודה הזו מתארת שיטות חישוביות לזיהוי גנים שעברו מוטציות נהג בהתאם למידת ההשפעה שלהם על התהליך הסרטני בגנום של חולה מסוים מתוך כלל הגנים שעברו מוטציות, ודירוגם. בחינת השיטות נעשתה בדרך כלל ע״י יצירה של קבוצות מסוימות של גנים המועמדים להיות נהגים ובחינה שלהם בעזרת האלגוריתם PRODIGY. בנוסף, נעשה שימוש בשחזור הפילוגנזה של מוטציות סרטניות, אשר מודלו למבנה של עץ אבולוציוני. שילוב זה התברר כמוצלח והוביל לשיפור ביחס לתוצאות שהושגו בשיטות קודמות.

הקבוצה הראשונה של ניסויים שהרצנו בחנה מוטציות מסוג של הכפלות ומחיקות של מקטעים גנומיים. תחילה, קבוצת הגנים המועמדים כללה את כל הגנים שממוקמים בתוך המקטעים שהשתנו. הביצועים היו נמוכים בהשוואה לקבוצה שכללה רק החלפות בסיסים. מכאן הסקנו שישנה כמות גדולה של רעש בנתונים ושעלינו לזקק את קבוצת המועמדים. בניסוי הבא צמצמנו את רשימת המועמדים כך שתכיל מחיקות או הכפלות בכמויות משמעותיות סטטיסטית בלבד. הביצועים הראו שיפור רב ביחס לבחינת כלל הגנים, אך היו נמוכים ביחס לשימוש בהחלפות בסיסים בלבד. בניסוי נוסף, בדקנו את ההתפלגות של כמויות הגנים אשר משתייכים יחד למקטעים שעברו הכפלות או מחיקות. מתברר שמבין המקטעים שעברו את השינויים המשמעותיים ביותר, 70% הכילו גן בודד. לפיכך יצרנו קבוצת מועמדים מהגנים הבודדים הללו יחד עם גנים שעברו החלפות בסיסים, וקיבלנו תוצאות דומות לאלה שהושגו עבור הקבוצה של החלפות הבסיסים בלבד.

מכאן ניתן להסיק שניתן להרחיב את הקלט של האלגוריתם, ולשלב בו את הגנים הבודדים להשגת רשימה מפורטת יותר של גנים נהגים באופן מדויק.

בניסויים נוספים עשינו שימוש נרחב בעקרונות הפילוגנזה. השתמשנו בערכים של VAF ו-CCF המחושבים על בסיס תדירות החלפות בסיס, הכפלות ומחיקות, שהם הערכות של אחוז האללים המוטנטים מתוך כלל האללים או של אחוז התאים הסרטניים בדגימה. יצרנו ציונים חדשים לגנים מוטנטים, אשר משלבים בתוכם את הציונים של PRODIGY ואת ערכי ה-VAF או ה-CCF. הציונים המאוחדים גרמו לדירוג מחודש של הגנים הנהגים והובילו לשיפור בתוצאות, במיוחד באוכלוסיית סרטן שלפוחית השתן. מגמת השיפור התגלתה בצורה ברורה אף יותר כאשר השווינו את רשימת הגנים שמצאנו לגנים נהגים שמקושרים ספציפית לסוג הסרטן הנבדק, ולא לרשימה כללית של גנים הידועים כנהגים. התוצאות מעידות על הקשר החשוב שבין זמן היווצרות המוטציה לבין האפקט שהיא מותירה בתאים הסרטניים. ככל שמוטציה קורית בשלב מוקדם יותר בהתפחות הגידול, כך סביר יותר שהיא מוטציית נהג.

אוניברסיטת תל אביב

הפקולטה למדעים מדוייקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ע"ש בלווטניק

תעדוף מותאם אישית מונחה-פילוגנזה של גנים נהגים על בסיס מוטציות נקודתיות ושינויים כמותיים בגנום הסרטני

חיבור זה הוגש כעבודת גמר לתואר "מוסמך אוניברסיטה" בבית הספר

למדעי המחשב

על ידי

נעמה קדוש

בהנחיית

פרופ' רון שמיר

סיוון תשפ"ב