






Article

Network-Based Analysis to Identify Drivers of Metastatic Prostate Cancer Using GoNetic

Louise de Schaetzen van Brienen ^{1,2}, Giles Miclotte ^{1,2}, Maarten Larmuseau ^{1,2}, Jimmy Van den Eynden ³
and Kathleen Marchal ^{1,2,*}

¹ Department of Plant Biotechnology and Bioinformatics, Faculty of Sciences, Ghent University, 9052 Ghent, Belgium; louise.deschaetzenvanbrienen@ugent.be (L.d.S.v.B.); giles.miclotte@ugent.be (G.M.); maarten.larmuseau@ugent.be (M.L.)

² Department of Information Technology, Faculty of Engineering and Architecture, Ghent University-IMEC, 9052 Ghent, Belgium

³ Department of Human Structure and Repair, Faculty of Medicine and Health Sciences, Ghent University, 9000 Ghent, Belgium; jimmy.vandeneinden@ugent.be

* Correspondence: kathleen.marchal@ugent.be

Simple Summary: The identification of cancer driver genes is, for statistical reasons, often biased toward genes that are altered frequently in a cohort. However, genes that are less frequently mutated can also alter cancer hallmarks. To detect such rarely mutated genes involved in driving metastatic prostate cancer, we analyzed the Hartwig Medical Foundation metastatic prostate cancer cohort. Hereto, we developed GoNetic, a novel network-based method that can detect genes with a lower mutational rate as members of recurrently mutated sets of genes connected on a prior interaction network. In contrast to state-of-the-art network-based driver identification methods, GoNetic retains information on sample-specific mutations and uses more properties of the prior interaction network. When applied to the Hartwig Medical Foundation cohort, GoNetic successfully prioritized both known drivers and rarely mutated driver candidates of metastatic prostate cancer. Comprehensive validation with other public data sets further supported the driver potential of these novel candidates.

Abstract: Most known driver genes of metastatic prostate cancer are frequently mutated. To dig into the long tail of rarely mutated drivers, we performed network-based driver identification on the Hartwig Medical Foundation metastatic prostate cancer data set (HMF cohort). Hereto, we developed GoNetic, a method based on probabilistic pathfinding, to identify recurrently mutated subnetworks. In contrast to most state-of-the-art network-based methods, GoNetic can leverage sample-specific mutational information and the weights of the underlying prior network. When applied to the HMF cohort, GoNetic successfully recovered known primary and metastatic drivers of prostate cancer that are frequently mutated in the HMF cohort (*TP53*, *RB1*, and *CTNNB1*). In addition, the identified subnetworks contain frequently mutated genes, reflect processes related to metastatic prostate cancer, and contain rarely mutated driver candidates. To further validate these rarely mutated genes, we assessed whether the identified genes were more mutated in metastatic than in primary samples using an independent cohort. Then we evaluated their association with tumor evolution and with the lymph node status of the patients. This resulted in forwarding several novel putative driver genes for metastatic prostate cancer, some of which might be prognostic for disease evolution.

Keywords: network-based cancer data analysis; driver identification; metastatic prostate cancer; somatic mutations



Citation: de Schaetzen van Brienen, L.; Miclotte, G.; Larmuseau, M.; Van den Eynden, J.; Marchal, K. Network-Based Analysis to Identify Drivers of Metastatic Prostate Cancer Using GoNetic. *Cancers* **2021**, *13*, 5291. <https://doi.org/10.3390/cancers13215291>

Academic Editors: Andrei Zinovyev, Laurence Calzone and Inna Kuperstein

Received: 27 August 2021
Accepted: 19 October 2021
Published: 21 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cancer driver identification depends on the genomic analysis of large cohorts of clinically well-characterized tumor samples. While gene-centric methods are ideal for identifying frequently mutated drivers in a cancer cohort, they often lack the power to

identify more rarely mutated genes. Several studies showed that driver genes follow a long tail distribution, where their mutation rate can be lower than 3% [1–3]. To allow digging in the long tail of rarely mutated genes, network- or pathway-based methods search for recurrently mutated pathways or gene sets. These methods indirectly prioritize rare drivers by making use of the connectivity they display with more frequently mutated drivers [2]. Network-based methods use a network model to identify recurrently mutated subnetworks as proxies of recurrently mutated pathways. This network model is derived from a network prior in which the nodes are genes, and the edges represent interactions between genes [4].

When seeking to identify recurrently mutated subnetworks, several existing network-based methods propagate a gene-centric signal over the network prior. In this approach, gene-specific mutational information is mapped to nodes in the network prior by assigning a single score to each gene. This score is typically derived from a gene-centric driver identification method (e.g., ActiveDriver scores [5] or frequency scores) that is propagated over one single network (e.g., HotNet2 [3]). However, not all mutations in a given gene have the same driver potential. The use of a single score for each gene thus results in the loss of sample-specific information (i.e., HotNet2 [3], Hierarchical HotNet [6], NBS [7], MUNDIS [8]). Several tools exist to determine the mutation deleteriousness (e.g., CADD scores [9], FATHMM scores [10]) and provide sample-specific gene scores. In addition, only a few network-based driver identification methods can use the information embedded in the network model, such as the edge weights or directionalities [4].

To cope with these issues, we developed GoNetic, a novel network-based method that relies on probabilistic pathfinding. GoNetic takes as input a weighted interaction network and, for all samples in the cohort, a list of somatic mutations and their functional impact scores. Genes that are mutated in at least one sample are mapped on the interaction network, and the paths (i.e., series of consecutive edges in the network) between genes mutated in different samples are identified. Using this set of paths between somatic mutations, a subnetwork that connects as many pairs of relevant somatic mutations as possible is inferred. GoNetic is designed to efficiently handle data sets of up to thousands of cancer patients. Benchmarking GoNetic with other driver identification methods shows its performance and robustness are on par with what can be expected from the state-of-the-art.

Because drivers of metastatic prostate cancer are known to follow a long tail distribution [1], we applied GoNetic to identify novel drivers with a lower mutational rate of metastatic prostate cancer (mPCa). Metastasis is the leading cause of mortality among men with prostate cancer (PCa) [11]. Despite an initial benefit with androgen deprivation therapy, newer hormonal treatments, or cytotoxic chemotherapies, patients with metastatic disease eventually develop drug resistance and progress to lethal castration-resistant disease [11]. Prevention of metastatic disease depends on a more comprehensive identification of early drivers of mPCa. To discover novel mPCa drivers, we analyzed the HMF cohort, one of the largest cohorts on metastatic cancer [12]. GoNetic identified frequently mutated subnetworks that contain well-known primary and metastatic drivers as well as several rarely mutated driver candidates.

By exploiting the specific properties of complementary data sets, we could provide further support for the role of some of these more rarely mutated genes in driving mPCa. Some of the genes that were predicted as metastatic drivers by GoNetic were indeed found to be more frequently mutated in metastatic than in primary samples. Using matching primary and metastatic samples [13,14] allowed associating some of the predicted drivers with the early onset of metastatic disease. Using a cohort of primary samples with annotated lymph node status [1,15] allowed the identification of predicted drivers associated with a positive lymph node status. This further confirms their role in metastatic disease.

2. Materials and Methods

2.1. Discovery and Validation Cohorts

The discovery cohort is obtained from the HMF and consists of 352 tumor samples from patients with mPCa [12]. This cohort is further referred to as the HMF cohort. Samples were selected based on the following criteria: the primary tumor originated in the prostate, the sex of the patient is male, and the biopsy site is not from the primary tumor. This discovery cohort was used as a case study for GoNetic, after which we validated our findings using five independent publicly available data sets.

To determine which drivers prioritized by GoNetic were more significantly mutated in metastatic than in primary samples, we used data obtained from primary tumor samples from Armenia et al. [1]. We selected samples from patients that had a low risk of metastasizing at the time of the resection of their primary tumor. More specifically, we focused on the patients that: (1) did not show a new tumor event, (2) were free from tumor material after resection, and (3) were annotated as lymph node-negative according to the clinical information downloaded from TCGA. As such, 287 samples of the Armenia data set were selected, hereafter referred to as the TCGA cohort. We also used somatic mutation data from metastatic samples from Armenia et al. [1] as an independent metastatic cohort, hereafter referred to as the ARM cohort. This cohort originally consisted of 317 mPCa samples, of which nine with outlying mutation rates were removed.

To associate the potential driver genes with metastatic evolution, we used the data sets of Gundem et al. [13] and Kumar et al. [14]. Samples from the Gundem data set are whole-genome sequences (WGS), while those from the Kumar data set are whole-exome sequences (WES). Those data sets both contain multiple metastatic samples from the same patient for 26 patients. Additionally, for 16 patients, they also contained the matched-primary sample. Patients with only one metastatic sample were not considered in our analysis.

Finally, we tested whether some of the putative driver genes reported by GoNetic could be associated with positive lymph node (LN+) status. Hereto, we used the 69 primary samples from TCGA and 6 samples of the MPC project [15] with positive lymph node status. This cohort is further referred to as the LN+ cohort and compared to the TCGA cohort.

2.2. Novel Network-Based Driver Identification Method: GoNetic

GoNetic is a network-based driver identification tool, conceptually based on our seminal work [16–18] but redesigned to efficiently handle large cancer data sets. In general, the algorithm searches for short paths between a set Y of genes of interest, in our case: between genes containing somatic variants in different samples. It then heuristically selects a subnetwork $S = (V, E)$ that maximizes the scoring function:

$$\sum P(\text{path}(g, Y) | S) - p | E | \quad (1)$$

Here, the sum is over all genes g in Y , and $P(\text{path}(g, Y) | S)$ is the probability that a non-empty path between gene g and the other genes in the set Y exists in S , i.e., the existence of paths in the network contributes positively to the score. The term $p | E |$ is an edge penalty p for each of the $| E |$ edges in S . The penalty enforces the subnetwork to be parsimonious in the number of edges. This constraint leads to the preferential selection of paths that contain overlapping edges, and hence the algorithm finds frequently mutated subnetworks.

To determine the probability that a path exists between two nodes, edge probabilities are required. If the input network already includes edge probabilities, then these are used as initial weights; otherwise, all edges are initialized with a weight of 1. The edges are then reweighted by GoNetic based on the topology of the network model. From the distribution of out-degrees of the nodes in the interaction network, a power-law distribution is estimated. Then, all edges with start node i are reweighted with a sigmoidal factor:

$$s(i) = (1 + \exp(3 * (\text{out_degree}(i)/\text{cdf}_{90} - 1))) \quad (2)$$

where cdf_{90} is the out-degree corresponding with the 90th percentile of the power-law distribution. The new edge weights are then recomputed as:

$$w(i, j) = 1 - (1 - 1/s(i)) * (1 - w(i, j)) \quad (3)$$

In this manner, edges connecting to hubs are downweighted, but they can still contribute to the final resulting network.

The score of a path is the product of the weights of its edges and can hence again be interpreted as a probability. These path probabilities are then further downweighted with sample-specific information derived from the functional impact score, the VAF, and a correction factor for samples with significantly higher mutation rates (mutation rate Z-score > 3.5). For a path from mutation i to mutation j , the path score is multiplied by the relevance scores and correction factors of the terminal nodes i and j .

GoNetic maps all mutations from the input data on the network and looks for paths between mutations in different samples. As such, for each mutated gene in the data set, a branch-and-bound algorithm constructs a path tree incrementally. In this path tree, the root corresponds with the mutated gene, and each path from root to leaf corresponds with a path in the network. For each node in the path tree, two scores are retained: the score of the current path and the maximal score that can be obtained by extending this path further, taking into account the weights of the next edges and the sample-specific weight factors of potential terminal nodes. Using these scores, a priority queue determines the next leaf from which the path tree should be extended.

To keep the computational complexity of the subnetwork selection in check, not all paths are retained. As such, the following filters are applied to the pathfinding procedure: only the paths with the highest scores are retained (default: max 25 paths), all paths with a path probability below a given threshold value (default: 0.2) are rejected. Additionally, the path length is limited to 4 edges. To avoid connecting mutations that are far removed from the network, longer paths are less likely to be biologically relevant [19,20]. The priority queue guarantees that high scoring paths are found early on, and as such, these filters are integrated into the branch-and-bound algorithm, allowing the early termination of the pathfinding procedure.

The subnetwork selection is then modeled as a decision-theoretic problem in the same way as in PheNetic [16–18]: first, for each mutation, all paths originating from this mutation are encoded as a conjunctive normal form and then compiled to a deterministic decomposable negation normal form (d-DNNF) with c2d [21]. Using these d-DNNFs, the probability of the existence of at least one path originating from this gene in a given subnetwork can be efficiently computed. To allow this approach to scale to the large cancer data sets, the subnetwork selection algorithm was parallelized and extended with a score cache. This cache encodes for each d-DNNF the interactions in the subnetwork that are relevant to this d-DNNF in a bit-vector, which is used as the key in the score map. Multiple distinct subnetworks can have the same set of relevant interactions for a given d-DNNF; as such, this cache allows computed probabilities to be reused for the scoring of multiple subnetworks.

The subnetwork selection is performed for a range of edge penalties. For each edge penalty, multiple runs are performed, and the resulting networks are compared based on score, resulting in network stability and size metrics. Edge penalties that result in either: (1) a negative score, (2) an average Jaccard index between the subnetworks smaller than 0.5, or (3) subnetworks containing more than 80 interactions are rejected. The union of the highest-scoring subnetworks of all the remaining edge penalties then forms the final resulting network of GoNetic. The genes in this network that are also mutated in the input data are compiled in a gene list, and the genes in this list are assigned a rank based on the highest edge penalty for which they appear in the resulting subnetworks. Genes that appear in a subnetwork with a higher edge penalty require more or stronger connections with other mutated genes through paths in the network, else the score of that subnetwork

would be too low. As such, this approach prioritizes genes that occur in highly connected subnetworks while ranking more weakly connected genes lower.

We can distinguish two groups of genes in this resulting subnetwork: (1) mutated genes, i.e., the start and end points of the paths of interest in this subnetwork; and (2) connector genes, i.e., genes that are not themselves in the mutation data, but that are required to connect the mutated genes on the network.

2.3. GoNetic Benchmarking

We compared GoNetic with other driver identification methods using a well-established evaluation framework based on the study of Tokheim et al. [22]. GoNetic was compared with 2020+ (<https://github.com/KarchinLab/2020plus>, accessed on 25 May 2021), TUSON [23], OncodriveFML [24], MutsigCV [25], OncodriveClust [26], MuSiC [27], ActiveDriver [5], and OncodriveFM [28] using the benchmark result described in Tokheim et al. [22] study.

On the Tokheim data set, GoNetic was run with the default parameter settings, and Reactome (version 2018) [29], consisting of 207,668 interactions among 10,123 genes, was used as a network model. The mutations in the Tokheim et al. [22] benchmark data set were assigned functional scores using the combined annotation-dependent depletion (CADD [9,30]). CADD scores were used because they were already available for the genome build used in Tokheim et al. [22] study. To convert the CADD scores to probabilities, representing the impact of the mutation on the path scoring, a sigmoid function was used:

$$p = (1 + 1.04^{(-x+10)})^{-1} \quad (4)$$

where x is the CADD score of a somatic mutation. To reduce the computational cost, only the mutations with a score over 70% were used to run GoNetic on the benchmark, resulting in a total of 58,567 mutations used as input.

Next to GoNetic, Hierarchical HotNet, another network-based driver identification method, was added to the benchmark results as part of the current study [6]. The Hierarchical HotNet results are poorer than expected, see Table 1. However, as this method is also a network-based approach that does not assign a gene-based p -value, it is not entirely clear how a threshold on significance should be set. The results of GoNetic and Hierarchical HotNet included in the benchmark analysis can be obtained in Table S1.

Table 1. CGC benchmark of different driver identification methods. #CGC: number of identified drivers that are in the gold standard CGC, #significant: number of driver genes prioritized by each method on the Tokheim data set. % CGC: positive predictive value or the number of true positive predictions (overlap with CGC) on the total number of predictions. Panel A. Considering, for each method, all the significantly identified drivers. Panel B. Considering, for each method, the top 87 predicted genes.

	Panel A			Panel B		
	#CGC	# Significant	% CGC	# CGC	# Significant	% CGC
TUSON	115	243	0.4733	71	87	0.8161
2020+	104	208	0.5000	63	87	0.7241
OncodriveFML	107	679	0.1576	50	87	0.5747
OncodriveFM	143	2600	0.0550	41	87	0.4713
GoNetic	81	331	0.2447	37	87	0.4253
MutSigCV	71	158	0.4494	37	87	0.4253
OncodriveClust	59	586	0.1007	22	87	0.2529
MuSiC	173	1975	0.0876	19	87	0.2184
ActiveDriver	34	417	0.0815	12	87	0.1379
Hierarchical HotNet	41	455	0.0901			

To compare results, the Tokheim benchmark assumes that the output of the driver identification method is a gene list, with p -values assigned to each gene. Network-based driver identification methods typically identify modules or subnetworks of putative drivers rather than individual drivers, resulting in several driver genes receiving the same rank or p -value. To make GoNetic fit the Tokheim framework, we extract from the output network the nodes that had at least one mutation in the input data set and ranked them based on the highest edge penalty for which they were included in the final network, where a higher edge penalty corresponds to a lower p -value in other methods.

For each method, the performance was assessed by calculating the true positive ratio. This corresponds to the number of true positive predictions on the total number of predictions, where true positive predictions correspond to those driver genes that were already reported in the Cancer Gene Census (CGC). We used this metric rather than the AUC because predictions that are not present in CGC are not necessarily false positives. Indeed, metrics, such as AUC, that consider all the non-CGC predictions as false positives are biased toward methods that rely on prior information (e.g., supervised methods or methods that use a network prior). To calculate the true positive ratio for the methods that report a q -value, we set a threshold of 0.1 to determine the significantly selected genes [22]. For GoNetic and Hierarchical HotNet, the entire resulting mutated gene lists were used since these methods do not compute a gene-level q -value. To have a more comparable number of genes between the methods, we redid the analysis using for all methods the same number of genes, corresponding to the number of genes that received the highest rank with GoNetic (87 genes).

2.4. OnCompare: Procedure to Compare the Mutation Rate of Driver Genes between Sample Groups

To compute the statistical significance of the degree to which genes were more mutated in metastatic than in primary samples, we developed OnCompare. This statistical procedure copes with confounding factors such as a tumor mutational burden (TMB) and the tumor purity of a sample. This is important because these factors can cause differences in the number of mutations between groups of samples.

In general, OnCompare is a statistical testing procedure that can be used to compare the number of mutations in a gene, or in sets of genes, between two different groups of samples. Contrarily to commonly used tests, such as Fisher's exact test or the binomial test, it relies on a Poisson binomial distribution, which accommodates for the presence of sample-specific confounding factors. The testing procedure consists of two components: (i) the assignment of sample-specific probabilities that compensate for confounders and (ii) the statistical test. To assign the probabilities of a gene being mutated in a sample, we rely on logistic regression to model the relation between a confounder in a sample and the probability that the gene is mutated in that sample. Typically, there exists a proportionality between confounders and mutation rate that is naturally captured by linear models. For instance, samples with a higher TMB are assigned a higher probability of containing a mutation in a predefined gene. Once the logistic regression model is fit to the data, it is used to assign to each sample a background probability. The test operates under the hypothesis that the probability of a sample having a gene mutated under the null hypothesis is independent of the group it belongs to. Here, this means that the null hypothesis assumes no intrinsic difference in mutation rate for gene i between the different sample groups. Denoting the probability of a sample harboring a mutation in gene i by $P(S_i)$, the group by G , and possible confounders by θ , we can write this as:

$$P(S_i | G, \theta) = P(S_i | \theta) \quad (5)$$

where $P(S_i | \theta)$ is modeled using logistic regression, fit to all samples from both groups. When no confounders θ are known or measured, we have under the null hypothesis:

$$P(S_i) = (m_1 + m_2)/(n_1 + n_2) \quad (6)$$

where n_i denotes the number of samples in group i and m_i the number of samples with a mutation in group i . Given the sample-specific probabilities, the number of mutations in both groups, M_1 and M_2 , each follow a Poisson binomial distribution. Because the mutation status in one group is independent of the other group, we can write the joint probability mass function as:

$$P(M_1, M_2) = P(M_1) \times P(M_2). \quad (7)$$

To obtain a right-sided probability or p -value, we sum over all states that result in an equal or higher ratio than the observed ratio, m_1/m_2 , of mutations in group 1 to group 2:

$$\sum_{(v_1, v_2) \in V} P(v_1)P(v_2), \quad V = \{(v_1, v_2) \mid v_1/v_2 \geq m_1/m_2\}. \quad (8)$$

Conversely, a left-sided p -value can be obtained by summing over all states:

$$\sum_{(v_1, v_2) \in V} P(v_1)P(v_2), \quad V = \{(v_1, v_2) \mid v_1/v_2 \leq m_1/m_2\}. \quad (9)$$

OnCompare can also be applied to compare the mutation rates in sets of genes. To do so, we fit a logistic model to each of the genes separately and then determine a single probability that a sample contains a mutation in at least one of the genes in the set. Let U be a set of genes $\{g_1, \dots, g_{|U|}\}$ and denote the probability of a sample being mutated in g_i by $P(S_i | \theta)$, then the probability of a sample containing at least one mutation in U is given by:

$$P(U) = 1 - \prod_{i=1}^{|U|} (1 - P(S_i | \theta)). \quad (10)$$

The resulting p -value then expresses the probability of the observed ratio of the number of samples in both groups that have at least one mutation in the set of genes U , under the null hypothesis that the intrinsic mutation rate is equal in both groups.

Without confounders taken into account, we have demonstrated that the test shows a suitable correspondence to Fisher's exact test when comparing the difference in mutation rate between, respectively, the primary and HMF and ARM metastatic cohort (Figure S1).

2.5. Analysis of the Discovery Cohort

As a case study, we applied GoNetic to identify novel drivers of metastatic prostate cancer (mPCa) on the HMF cohort. Prior to applying GoNetic on the HMF cohort, several pre-processing steps were performed. Somatic variants obtained with the processing and analysis pipeline of the HMF were downloaded from the official HMF website. Somatic variants that did not pass all criteria to be called somatic were removed. We removed somatic variants with coverage below five reads and variant allele frequency (VAF) below 0.1.

To annotate somatic variants, we used the dbNSFP variant effect predictor (VEP) plugin [31]. We used the 'pick' option to ensure we prioritize per variant the most relevant annotation only (see details in VEP documentation). Silent mutations, mutations in introns, intergenic, upstream, downstream, 3'UTR and 5'UTR regions were discarded from the analysis.

Functional impact scores of the variants used to define the relevance scores in GoNetic were derived from FATHMM-MKL, which assigns pathogenicity scores to single nucleotide variants [10], and FATHMM-INDEL, which predicts the functional effect of small insertions and deletions (indels) [32]. Both tools assign a value between 0 and 1 to somatic variants where values closer to 1 correspond to a stronger functional impact of the mutation on the gene. Somatic variants that did not receive a FATHMM score were discarded from our analysis.

To identify samples with a deviating number of mutations, the distribution of the number of mutations per sample was log-transformed to obtain a normal-like distribution.

Outlier samples were subsequently defined as the samples that fall below the boxplot's lower whisker ($Q1 - 1.5 \times IQR$) or above the boxplot's upper whisker ($Q3 + 1.5 \times IQR$). For the HMF cohort, 325 samples remained after filtering outliers.

Table S2 summarizes the number, the type, and annotations of somatic variants retained after filtering outliers. In total, we retained 325 samples from the HMF cohort containing 18,318 somatic mutations, of which 16,261 SNP and 2057 indels (596 insertions and 1461 deletions).

GoNetic was run on a human-specific interaction network, Reactome [29], and 18,318 somatic mutations of the HMF cohort. The frequency of the somatic variants was corrected for purity and ploidy by HMF, and the functional impact scores were determined with FATHMM predictions.

2.6. *SomInaClust Analysis on the HMF Data Set*

To identify mutational hotspots in the genes prioritized with GoNetic, we used SomInaClust [33]. As in the original publication, we ran SomInaClust with COSMIC as reference (v92) to obtain prior information on gene hotspots and background mutation rates. During the reference step, 7378 mutation hotspots were identified in 429 genes from the CGC list (v92) [34]. When searching for mutational patterns with SomInaClust, only frameshift and in-frame deletions or insertions, missense, nonsense, splice site, and silent mutations were considered [33]. SomInaClust determines, based on the cohort's mutation data, gene-specific oncogene (OG) and tumor suppressor gene (TSG) random mutation probabilities (pOG and pTSG) [33]. A multiple testing correction of the pOG and pTSG is performed using the Benjamini–Hochberg method. The driver gene probability is calculated based on the product of the corrected p -values (qOG and qTSG), and genes with $qDG \leq 0.05$ are defined as putative driver genes. In addition, to classify the putative driver genes as putative OGs or TSGs, the OG and TSG scores are defined. The OG score is the proportion of clustered OG mutations to the total number of OG mutations. In contrast, the TSG score is the proportion of TSG mutations on the total number of mutations. Genes are classified by SomInaClust as OGs when their OG score is above 20% and as TSGs when their TSG score is above 20%. When both the OG and the TSG score are above 20%, the driver is classified as a TSG [35].

2.7. *Gene Ontology Enrichment Analysis*

Gene ontology (GO) enrichment analysis was performed using the analysis option of the STRING v11 platform [36], focusing on the GO biological process terms. GO enrichment analysis was performed on the genes belonging to each subnetwork identified by GoNetic. Results are reported in Appendix A.1. Biological processes enriched with a false discovery rate (FDR)-adjusted p -value below 0.05 were considered as significant.

2.8. *Validation of the Other Cohorts*

To determine which drivers prioritized by GoNetic were more significantly mutated in metastatic than in primary samples, the mutational enrichment between, respectively, the two metastatic cohorts (HMF and ARM) and the primary cohort (TCGA) could be assessed (Figure 1, panel 1). Mutational enrichment comparisons were performed using OnCompare.

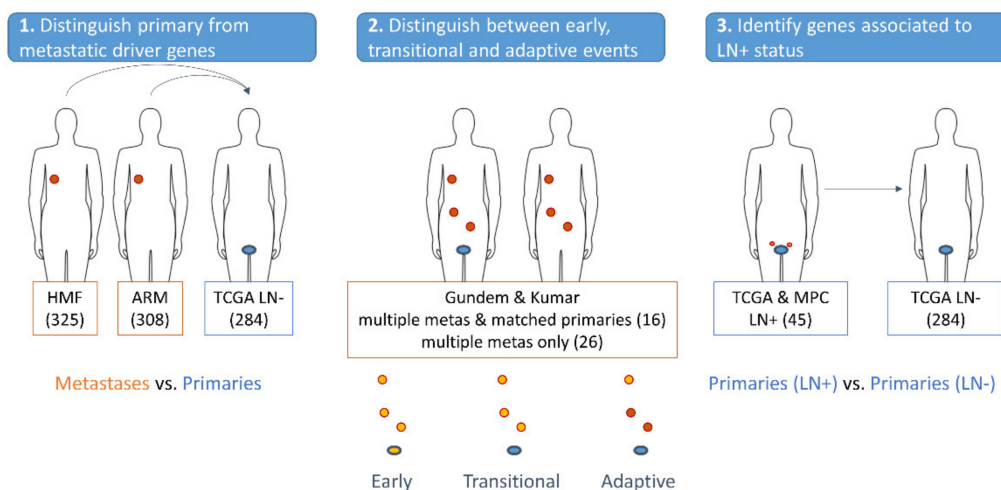


Figure 1. Validation of genes identified by GoNetic. **Panel 1.** First, we compared the mutation rate of genes prioritized by GoNetic between each of the metastatic cohorts (HMF and ARM) and the primary cohort (TCGA). Mutational enrichment signal in metastatic samples is divided into three categories: strong, intermediate, and weak. **Panel 2.** Secondly, we assessed whether GoNetic drivers could be associated with tumor evolution using cohorts of matching primary and metastatic samples. Genes reported by GoNetic were defined as primary/early drivers, transitional drivers, and late/adaptive drivers. **Panel 3.** Lastly, we assessed whether mutations in any of the genes prioritized by GoNetic could be associated with the lymph node-positive status of the patients.

To maximize the comparability between, respectively, the WGS samples from HMF and the WES samples from ARM and TCGA, we performed the analysis only on somatic mutations detected in the HMF cohort that mapped to regions that were subjected to sequencing in the Armenia et al. [1] samples. Two percent of the somatic variants in coding regions originally detected in the HMF cohort could not be mapped to regions covered by the study of Armenia et al. [1]. For only one gene identified by GoNetic (i.e., MUC19), the mutations detected in the HMF cohort fell outside the regions covered by the Armenia et al. [1] study.

Figure S2 shows that the TMB largely differed between the primary and both metastatic cohorts, but the tumor purity did not. Therefore, only TMB was considered as a confounder in our analysis. TMB was evaluated prior to applying any filtering.

Table S3 presents, for each gene reported by GoNetic, the OnCompare mutational enrichment *p*-values obtained with and without correcting for TMB for both the HMF and ARM metastatic cohorts.

All genes identified by GoNetic that were significantly enriched in mutations in at least one metastatic cohort with and without correction for TMB were kept for further analysis. Mutational enrichment signal in metastatic samples is divided into three categories. First, the strong signal corresponding to genes significantly enriched in both HMF and ARM metastatic cohorts even after TMB correction. Second, the intermediate signal corresponds to genes enriched in mutations in both the HMF and ARM metastatic cohorts without TMB correction and only in one after TMB correction. Third, the weak signal corresponds to genes enriched in mutations with and without TMB correction in only one cohort. In addition, the most likely primary drivers (i.e., genes that were not more mutated in the metastatic than in the primary samples but that were frequently mutated across the cohorts) were also retained. A distinction between genes that were never enriched in mutations in both metastatic cohorts and genes that were enriched in mutations in metastatic cohorts only before TMB correction was made. Three genes (i.e., GLI3, GAK, and SCN9A) were significantly enriched in mutations in both the HMF and ARM metastatic cohorts, but only without TMB correction. Although those genes were not mutated in the primary cohort, their low mutational frequency in the metastatic cohorts did not allow finding significant

mutational enrichment signals in either of the metastatic cohorts. Therefore, these genes were no longer considered in the subsequent analysis.

To assess whether mutations in the retained genes prioritized by GoNetic could be associated with the evolution from primary to metastatic disease, we analyzed the Gundem and Kumar data sets. We searched for mutations in genes prioritized by GoNetic that occurred in different samples from the same patient (Figure 1, panel 2).

Genes mutated in the primary (if available) and in all matching metastatic samples for the majority of the patients (>50% of the patients containing a mutation in the gene under consideration in all of their metastatic samples) were considered as involved in the early onset of the metastatic disease or primary drivers. Genes for which no mutation was detected in the matching primary if it was available, but that contained the same mutation in the majority (>50%) of the metastatic samples (then referred to as truncal mutation) in most patients (>50%) were considered transitional drivers, as they must originate at the transition between the primary to the metastatic stage. Genes that were never found in the matching primary samples if available and not found truncal in most of the patients (so the mutation occurred in less than 50% of the matching metastatic lesions) were considered late drivers.

To test whether some of the putative driver genes reported by GoNetic could be associated with positive lymph node (LN+) status (Figure 1, panel 3), we used OnCompare without correcting for differences in TMB because the average TMB was comparable between primary samples of patients with positive and negative lymph node (LN−) status (Figure S3).

3. Results

3.1. GoNetic: A Flexible Network-Based Driver Identification Method

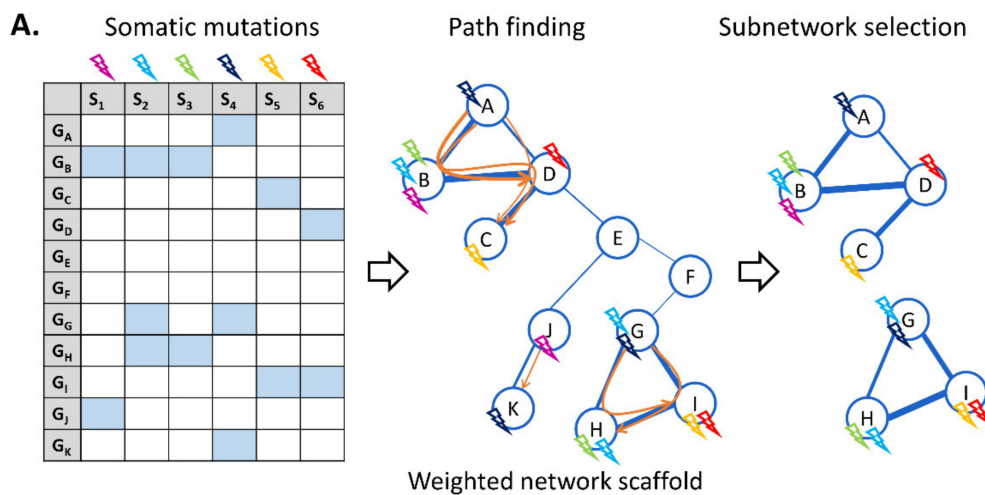
We developed GoNetic, a network-based method based on probabilistic pathfinding, designed to handle large cancer cohort sizes using large network priors.

The method takes as input a set of somatic mutations and a network prior to driving its analysis (Figure 2, panel A). In such a network prior, nodes represent genes, and edges represent interactions between genes. Edge weights can reflect any desired network property, e.g., the degree of belief one assigns to the edges. In our case, we opted for a topology-based weighting in order to reduce the bias toward highly connected genes. Genes in the network are annotated with sample-specific mutational information obtained from the input cohort. Subsequently, paths between mutated genes in different samples are enumerated, where a path is defined as a list of consecutive edges in the interaction network. The source and target nodes of a path thus correspond with two different sample-specific mutations. Paths between sources and targets that are mutated in the same samples are excluded, hereby assuming that multiple mutations in the same pathway in a single tumor are less informative than mutations in that pathway occurring in independent samples.

Once all paths between sources and targets have been derived, an optimization step infers a subnetwork that contains paths connecting as many pairs of sample-specific mutations as possible. However, the paths connecting these mutations should consist of a minimal number of edges. By imposing this parsimony in the number of edges, the algorithm searches for paths that contain as many overlapping edges as possible and hence detects subnetworks that are consistently mutated in independent samples in the cohort. These subnetworks are considered proxies of driver pathways.

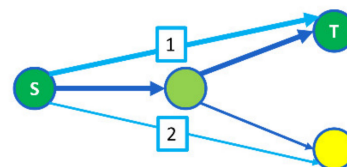
During pathfinding and optimization, each path is assigned a probability, which reflects the degree of belief that the path is associated with the carcinogenic phenotype under study. This probability takes into account the weights of the edges composing the path and properties that reflect the driver potential of the connected source and target mutations, such as sample-specific functional impact scores and VAF (Figure 2, panel B). To avoid that samples with a higher mutation rate would skew the search for paths in the pathfinding step toward these populations with higher mutation rates, the scores are corrected based on the mutation rate of the sample in which they occurred. Because the

probabilities of the paths are accounted for, GoNetic does not only prioritize relevant genes but also extracts the most relevant edges. In this way, GoNetic allows extracting the most important network context in which the potential driver mutations occur.



B. Path probabilities ~ nodes and edges weights

$$\begin{aligned}
 & \text{probability}(S, n, E, m)_{S \neq E, n \neq m} \\
 &= \prod_{(i,j) \in P} (\text{weight}_{(i,j)}) * \text{relevance}(S, n) * \text{relevance}(E, m)
 \end{aligned}$$



- Node weights = sample-specific
- Edge weights = network topology-based

Figure 2. GoNetic workflow. **Panel A.** GoNetic uses as input somatic mutations of a cohort of tumor samples and a network prior. Somatic mutations occurring in each of the samples are mapped to the genes in the network (mutations occurring in the same sample are represented in the same color). During pathfinding, GoNetic identifies paths between genes that are mutated in different samples, displayed in orange. During the optimization step, GoNetic selects subnetworks that connect as many genes as possible with highly probable paths while using a minimal number of edges. **Panel B.** To each path, a probability is assigned that is a function of the node and edge weights. Nodes are weighted based on sample-specific information (i.e., the variant allele frequency (VAF), the functional impact score, and the sample mutation rate). Nodes with high weights are indicated in green, and those with lower weights in yellow. Edges are weighted according to the connectivity of the network prior. The line width reflects the edge weight. The example graph illustrates how path 1 will be assigned a higher probability than path 2 because it connects a source (S) to a higher weighted target (T) node and because the edges that make up the paths also have a higher weight than the one used to constitute path 2.

3.2. GoNetic Benchmark with the State-of-the-Art

To compare the performance of our method with that of some of the well-known state-of-the-art driver identification methods, we performed a benchmark using the evaluation framework of Tokheim et al. [22] (Materials and Methods).

The performance of each method was assessed by the true positive ratio. This corresponds to the number of true positive predictions on the total number of predictions, where true positive predictions correspond to those driver genes that were already reported in the CGC list. The higher the true positive ratio, the more the predictions made by a method are enriched in known cancer genes. For each method, Table 1 shows the overlap between their identified drivers and the CGC list; a larger overlap suggests that the predictions made by a method are more enriched in known cancer genes. Table 1 (panel A) considers for all methods that report a *q*-value, the true positive ratio for the predictions obtained

with a q -value < 0.1 . The network-based methods GoNetic and Hierarchical HotNet prioritize subnetworks containing several genes at once and hence do not provide a gene-level q -value. For these methods, the entire prioritized gene lists were used.

However, as with a certain q -value, the number of prioritized genes differed largely between the different methods. Table 1 (panel B) shows the true positive ratio considering this time for all methods the same number of genes, corresponding to the number of genes that received the highest rank with GoNetic (87 genes). Because Hierarchical HotNet does not provide a gene ranking, it could not be included in this comparison. This restriction to the top 87 genes resulted in a larger overlap ratio with CGC, indicating that all tools, including GoNetic, correctly prioritize known drivers.

Table 1 and Figure 3 show that the obtained true positive ratio depends on the properties of the methods: supervised driver identification methods (i.e., 2020+, TUSON), which use next to sequence-derived also pathway-based information, show a relatively higher true positive ratio than unsupervised methods that use merely sequence-derived information (i.e., MuSiC, MutSigCV, OncoDriveFM, etc.).

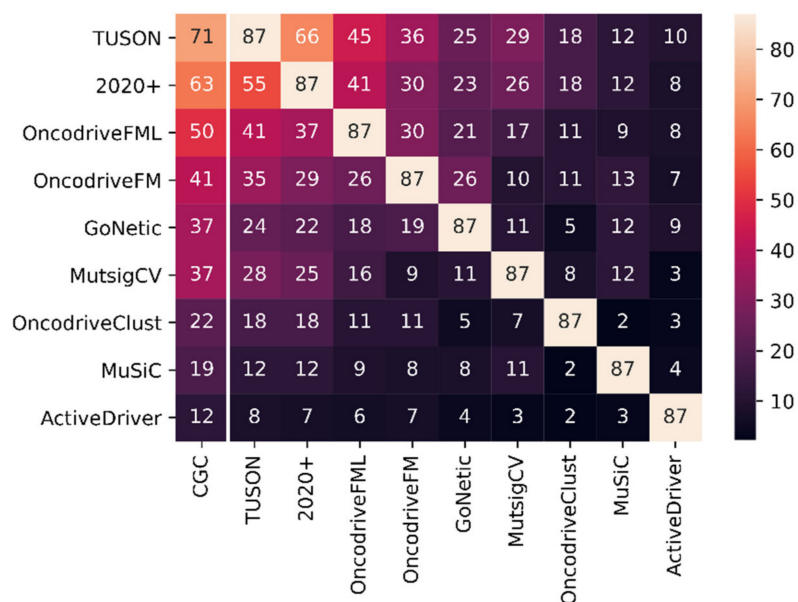


Figure 3. CGC benchmark heatmap between methods. Left-most column: intersection between the top 87 genes identified by each method and CGC. Upper triangle: intersection of the top 87 genes between methods. Lower triangle: intersection between the CGC genes in the top 87 genes between methods.

Expectedly, GoNetic has a true positive ratio falling between the true positive ratio of the supervised and the unsupervised methods that rely on frequency only (e.g., MuSiC) as it is not biased by training data but still uses a prior network to drive its analysis. In general, the performance of network-based driver identification methods is expected to be underestimated on gold standards such as CGC since network methods prioritize subnetworks containing several genes at once. Some of these genes are frequently mutated and likely already described in CGC, but the infrequently mutated ones are mostly unknown. Unknowns are considered false positive in the benchmark, even though they might be true drivers. However, Figure 3 also shows that most methods, including GoNetic, agree at least to some extent on the strong signals in the data. Indeed, mutual comparison between the predictions made by each of the different methods shows that the overlapping predictions (Figure 3, upper triangle) mostly correspond to the predictions that also overlap with CGC genes (Figure 3, lower triangle). Methods that are more similar in their underlying methods and in the information they use (e.g., TUSON and 2020+) tend to overlap more. For instance, GoNetic mostly overlaps with OncodriveFM, OncodriveFML, and the supervised

methods (i.e., TUSON and 2020+). These methods and GoNetic all use the frequency of the mutations and their functional impact scores.

3.3. Application of GoNetic to the HMF Data Set

To discover mPCa drivers, we analyzed the HMF cohort, one of the largest cohorts on metastatic cancer. GoNetic identified 14 recurrently mutated subnetworks, consisting of 75 interactions and 87 genes, of which five were connector genes (genes that were not mutated themselves). The vast majority of mPCa samples from the HMF cohort contained at least one mutation in any of those 82 putative driver genes (306/325, 94.15%). GoNetic assigns a rank to its identified drivers based on their contribution to the subnetwork selection (see Materials and Methods). Several of the top prioritized genes (i.e., *TP53*, *AR*, *SYNE1*, *MUC16*, *FOXA1*, *APC*, *RYR2*, *KMT2D*, *KMT2C*, *RB1*, *BRCA2*, and *SPOP*) correspond to the more frequently mutated genes in the HMF cohort (see genes in green in Table S4 and in blue in Figure S4). Moreover, the large overlap detected between the top-ranked driver genes prioritized by GoNetic and the ones identified by previous studies on the HMF cohort further confirms that GoNetic is able to capture the most clear signals in the data (Appendix A.2).

In addition to identifying the most significant signals in the data, GoNetic can dig into the long tail of rare drivers: 18 of its prioritized drivers were part of recurrently mutated subnetworks driven by one of the more frequently mutated drivers but were infrequently mutated themselves (mutated in less than 1% of the samples, highlighted in red in Table S4). Subnetworks identified by GoNetic were ranked based on the number of samples with at least one mutation in the subnetwork. Somatic mutations in subnetworks 1 and 2 cover a large number of samples in the HMF cohort, respectively 76% for subnetwork 1 and 30% for subnetwork 2. These two subnetworks also appeared to be centered around frequently mutated and well-known drivers of mPCa, respectively *TP53* and *AR* [1,12,14,37,38] (see left panel of Figure 4 for the mutational plot of subnetwork 1, for the other subnetworks, see Figure S5 panels A to N).

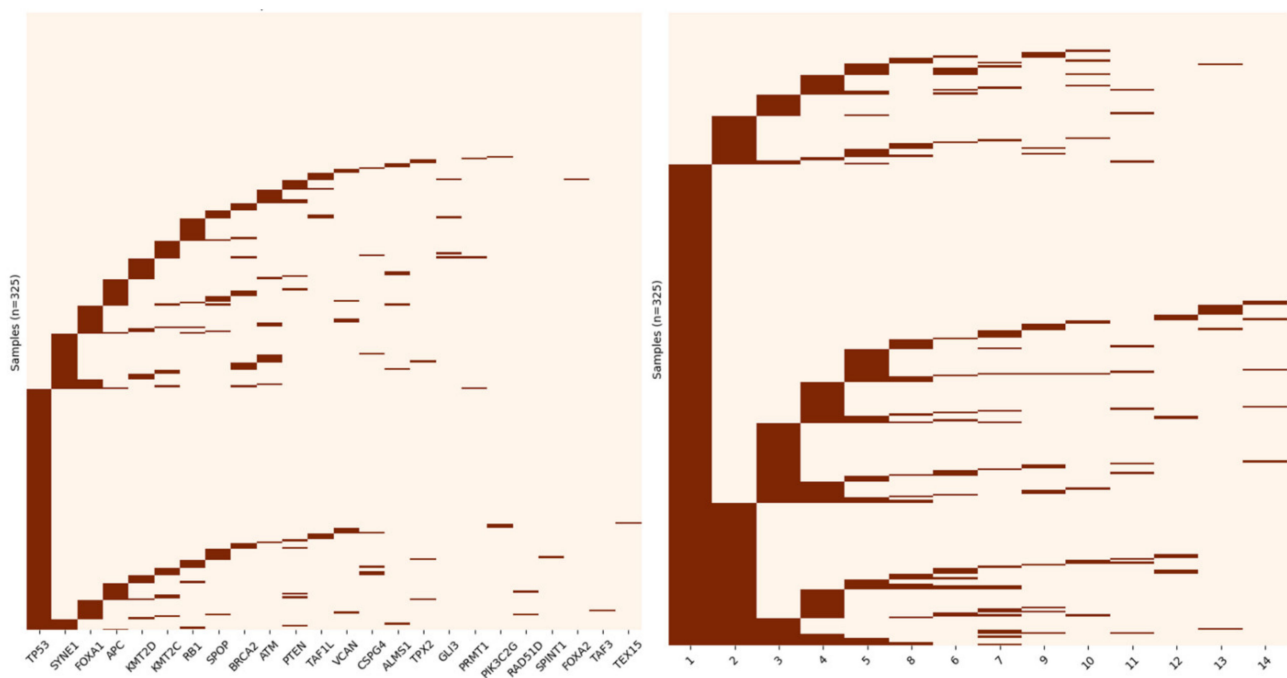


Figure 4. Left panel. Distribution of mutations in the HMF cohort in the first subnetwork identified by GoNetic. Right panel. Distribution of the samples (Y-axis) that carry mutations in any of the 14 subnetworks identified by GoNetic (X-axis). Samples that have at least one mutation in any of the 14 subnetworks are displayed.

Most of the samples have mutations in multiple subnetworks (207/306 samples with mutations in genes prioritized by GoNetic, 67.65%, (Figure 4, right panel)), suggesting that hitting multiple subnetworks is a prerequisite to acquiring carcinogenic properties.

We also assessed whether genes within the same subnetwork were mutated in a specific metastatic tumor site (i.e., lung, liver, or bone). Indeed, this would indicate that the pathway reflected by the subnetwork is acting as a driver pathway in different sites but is affected in each site through a different gene or set of genes (Tables S5 and S6). Overall, 18 genes were at least once reported as significantly differentially mutated between different metastatic sites (Table S6).

Except for three subnetworks (subnetworks 9, 11, and 14, which each contained two genes only), all other subnetworks were found to be enriched in several GO biological processes, including processes playing a crucial role in either cancer progression or metastasis (Table S5 and Figure S5, panels A to N).

From the 82 prioritized genes, SomInaClust identified mutational hotspots in only 12 of them (Appendix A.3, Figure S7). The fact that only a restricted number of genes prioritized by GoNetic were identified to contain mutational hotspots is expected, given their low mutational frequency in the HMF cohort.

To further validate the role of these prioritized genes in mPCa, we designed a strategy based on the analysis of complementary cohort data.

3.4. Distinguishing Primary from Metastatic Drivers Using Mutational Enrichment Analysis

First, we assessed the degree to which the identified genes were more mutated in metastatic than in primary samples. This allowed distinguishing the primary from the metastatic drivers. Because no matching primary samples for the metastatic tumors were available in HMF, we used primary samples reported in the study of Armenia et al. [1] originating from TCGA (Figure 1, panel 1). We hereby focused on primary samples that were the least likely to already contain metastatic drivers (see Materials and Methods). We performed the same enrichment analysis using an independent metastatic cohort described in Armenia et al. [1] and compared it with the same TCGA samples as a primary reference set. When assessing the difference in mutation frequency of the driver genes between the primary and metastatic cohorts, differences in tumor mutational burden (TMB) was accounted for (see Materials and Methods). The use of TMB as a cofactor affects the *p*-values (Figure S6), which is to be expected given the much higher TMB in metastatic than in primary samples.

According to the mutational enrichment analysis, the genes *TP53*, *AR*, *RB1*, *CTNNB1*, *CSPG4*, and *APC* exhibit a strong mutational enrichment signal in metastatic samples, both in the HMF and in the ARM metastatic cohorts, even after correcting for TMB. Except for *CSPG4*, these genes were also the most frequently mutated genes in the cohort and contain the well-known drivers of mPCa (i.e., *TP53*, *AR*, *CTNNB1*, and *RB1*) [1,12,14,37,38] (Table 2, 1st column).

For the remainder of the drivers prioritized by GoNetic, the lower mutational frequency limits the significance level for mutational enrichment that can be reached with the current sizes of the ARM and HMF cohorts, especially when also correcting for TMB. Table 2 (columns) summarizes how driver genes were categorized according to their level of significant enrichment between metastatic and primary samples.

Genes that were significantly more mutated in metastatic than in primary samples in both metastatic cohorts without TMB correction and in one metastatic cohort after TMB correction were labeled as genes with an intermediate mutational enrichment signal in metastatic samples. To this category belong, next to the genes that are mutated in at least 5% of the cohort (*BRCA2*, *MUC16*, *FAT4*, and *DCHS2*), several more rarely mutated genes (*MUC4*, *VCAN*, *TAF1L*, *CACNA1H*, *IGFR2*, *MUC2*, and *SHANK1*). Interestingly, *BRCA2*, a previously described metastatic driver of PCa, belonged to this category. *BRCA2* affects the DNA repair system and hence is associated with samples that have a high TMB [39]. Hence, accounting for TMB as a confounding factor might dilute the signal too much.

Table 2. Classification of the driver genes identified by GoNetic according to the mutational enrichment in metastatic cohorts (columns) and their association with tumor evolution (rows). **Columns:** Mutational enrichment signal in metastatic samples is divided into five categories (gray background): strong = genes significantly enriched in mutations in the HMF and ARM cohorts even after TMB correction, intermediate = genes significantly enriched in mutations in both the HMF and ARM cohorts without TMB correction but in one cohort after TMB correction, weak = genes significantly enriched in mutations with TMB correction in one metastatic cohort only. The two last categories correspond to frequently mutated genes that i) were enriched in mutations in both metastatic cohorts but only without TMB correction (no sign. enrichment with TMB correction) and ii) for which no enrichment at all could be detected, even not without correcting for TMB (no sign. enrichment). **Rows:** The primary/early metastatic category corresponds to genes for which mutations were present in the primary tumor (when available) and truncal to the matching metastases in most of the patients. The **transitional** category corresponds to genes for which mutations were truncal in the majority of the patients but not present in the primary tumor. The late category corresponds to genes for which mutations were non-truncal to metastatic lesions and never detected in the matching primary tumor.

	Drivers with Mutational Overrepresentation in Metastatic Samples			Primary Drivers or Genes Potentiating Metastasis	
	Strong	Intermediate	Weak	No Sign. Enrichment with TMB Correction	No Sign. Enrichment
Primary/early metastatic drivers	<i>TP53</i> ** <i>RB1</i> <i>CTNNB1</i>	<i>DCHS2</i> * <i>MUC2</i> *		<i>KMT2D</i> <i>KMT2C</i>	<i>FOXA1</i> <i>SPOP</i> <i>ATM</i> <i>PTEN</i>
Transitional metastatic drivers		<i>SYNE1</i> * <i>MUC16</i> <i>TAF1L</i> ** <i>VCAN</i>			
Late metastatic drivers	<i>AR</i>	<i>FAT4</i> <i>CACNA1H</i>	<i>LAMA2</i>		
Not present in Gudem nor in Kumar	<i>APC</i> <i>CSPG4</i> *	<i>BRCA2</i> ** <i>IGF2R</i> * <i>MUC4</i> <i>SHANK1</i> *	<i>MYH11</i> ** <i>LRRC7</i> <i>AMPH</i>		

* = genes for which mutations occur significantly more frequently in primary tumors of patients with positive lymph node status than in primary tumor samples of patients with negative lymph node status. ** significance p -value < 0.05. * p -value < 0.1.

Genes that were significantly more mutated in metastatic samples in one of the metastatic cohorts after correction for TMB but that could not be confirmed by the other metastatic cohort with or without correction for TMB were categorized as genes with a weaker mutational enrichment signal in metastatic samples. To this category belong also rarely mutated prioritized drivers *LAMA2*, *MYH11*, *AMPH*, and *LRRC7*.

At the other end of the spectrum, we have genes that were clearly not more mutated in the metastatic than in the primary samples. To this category belong the frequently mutated genes, i.e., mutated in more than 4% of the samples over all cohorts. Here, a statistical signal could have been detected, but no mutational enrichment signal was found, not even after omitting the correction for TMB. Therefore, these genes are most likely primary drivers. Indeed, the genes prioritized by GoNetic that fall in this category correspond to previously described primary drivers, i.e., *SPOP*, *FOXA1*, and *PTEN* [40,41].

The last category is genes that are relatively frequently mutated in both metastatic cohorts and for which an enrichment in mutations in metastatic samples can be detected in both the HMF and ARM metastatic cohorts, but only without correcting for TMB (*KMT2D* (22), *KMT2C* (20)). These genes are frequently mutated but not significantly enriched in mutations in metastatic samples. This suggests that they are likely primary or potentiating drivers of metastasis that are already frequently present in the primary tumor.

Finally, we applied OnCompare at the subnetwork level rather than at the gene level. Five subnetworks (1, 2, 5, 7, and 12) were significantly more mutated in metastatic samples than in primary samples in both the HMF and ARM metastatic cohorts (Table S7). For each

of these subnetworks, a lower p -value was obtained compared to p -values of any gene in the subnetwork. These results highlight that the genes in the subnetworks each affect a different set of metastatic samples and act as independent drivers. For subnetworks 1 and 2, this is to be expected since they contain several genes that are themselves significantly differentially mutated between the primary and the metastatic samples (*TP53*, *APC*, *RB1*, and *CSPG4* for subnetwork 1, and *AR* and *CTNNB1* for subnetwork 2). In addition, subnetworks 5 and 12 illustrate that rarely mutated genes that lack statistical power at the gene level can still be significant at the subnetwork level.

3.5. Association of the Identified Drivers with Tumor Evolution

Secondly, we used the data of Gundem et al. [13] and Kumar et al. [14] to assess whether the drivers identified by GoNetic could be associated with the evolution from primary toward metastatic disease. Both studies provide data for multiple metastatic samples in the same patients and sometimes also for the matching primary samples. This setting allows distinguishing drivers involved in the early stages of metastatic disease from those relevant in later disease stages.

For only a few genes prioritized by GoNetic, mutations detected in a metastatic sample could also be recovered in the matching primary sample. This can be explained by the small size of the set of patients with matching primary samples (16 patients). In addition, none of the patients carried in their primary lesion a mutation in the primary-specific drivers identified by GoNetic (*SPOP*, *FOXA1*, and *ATM*, Table 2). Moreover, due to heterogeneity of the primary tumor, the clone that seeded metastasis might have been absent in the primary lesion that was selected for sequencing. Because only a few mutations were recovered in the primary samples, we also considered whether mutations were found to be truncal to metastatic lesions of the same patient. Indeed, these can serve as an additional indication of their involvement in primary disease or of potentiating later metastatic disease.

So, all genes for which mutation data were available were categorized in one of the following three classes: primary or early metastatic drivers, transitional metastatic drivers, and late metastatic drivers. This classification is based on the occurrence of their mutations in the primary and metastatic lesions (see Materials and Methods).

Genes belonging to these categories and those derived in Section 3.4 are summarized in Table 2.

Primary or early metastatic drivers (Table 2, rows) are driver candidates derived from the Gundem and Kumar data sets. These drivers are detected in the primary sample (if available) and truncal to the metastases in the majority of the patients. Expectedly, to this category belong the well-known primary drivers *FOXA1*, *PTEN* [40,41], and *ATM*. Interestingly, mutations in *KMT2D* and *KMT2C* were detected in primary samples but not recovered in all metastatic samples of the same patient. This suggests that during metastatic spread, mutations in those genes were not selected. The following genes with mutational overrepresentation in metastatic samples also qualified as early metastatic drivers: *TP53*, *RB1*, *CTNNB1*, and *DCHS2*. These genes might be suitable markers for disease progression, as they are likely metastatic drivers and they are already detectable in the primary tumor. In contrast, *MUC2* was not recovered in all matching metastatic samples, suggesting this gene might not be as important for metastatic spread.

Mutations in genes prioritized by GoNetic that were not detected in the primary sample, but found to be truncal to the metastases in the majority of the patients, were considered transitional metastatic events (i.e., *SYNE1*, *MUC16*, *TAF1L*, and *VCAN*). Because of their subclonal nature, they might remain undetected in the primary sample, or they might play a role in the transition toward metastasis. To this category belong the genes with an intermediate signal of metastatic overrepresentation.

A representative example of a late metastatic driver is *AR*, which is very rarely found truncal to matching metastatic tumors of the same patient and is never detected in the matching primary tumors. *LAMA2* shows a similar mutational profile as *AR*: it is enriched in mutations in metastatic samples but non-truncal. Mutations in *FAT4* and *CACNA1H*

were found non-truncal in half of the patients. All these drivers are more likely involved in conferring adaptive phenotypes in later disease stages (such as disease resistance).

APC and *BRCA2* are never observed in any of the Gudem or Kumar metastatic samples, despite their rather high mutational frequency in both the HMF and ARM metastatic cohorts. Similarly, the potential metastatic drivers *MUC4*, *SHANK1*, *CSPG4*, *IGF2R*, *PCDHA8*, *AMPH*, *LRR7*, and *MYH11* were not mutated in any of the Gudem or Kumar samples. These genes are harder to detect in the validation data sets, because of their small mutation rate and because of the small size of the validation data sets.

3.6. Identifying Drivers Prognostic of Disease Staging

Lastly, we assessed whether the mutations in the genes identified by GoNetic were associated with the lymph node status of the patient. Patients with lymph node metastases at the time of diagnosis (LN+ patients) usually have a worse prognosis than patients without lymph node metastases (LN− patients). An association with lymph node status, therefore, also indicates a potential role in driving metastatic disease. Table 2 shows with an asterisk the genes that were significantly more mutated in LN+ patients than in LN− patients. Table S3 shows the OnCompare *p*-values for all genes identified by GoNetic. Despite the relatively small size of the lymph node-positive data set and the highly unbalanced number of samples (75 LN+ versus 284 LN− samples), the observed enrichment of mutations in LN+ patients was significant for several genes (i.e., *TP53*, *TAF1L*, *BRCA2*, and *MYH11*, *p*-value < 0.05). Even though the size of the data set lacks power, some of these genes, i.e., *TAF1L* and *MYH11* (*p*-value < 0.05) and *SHANK1*, *MUC2*, *IGF2R*, and *CSPG4* (*p*-value < 0.1), were never observed mutated in the TCGA cohort (LN−). In contrast, known and likely primary drivers (i.e., *SPOP*, *FOXA1*, *PTEN*, and *ATM*) show the least significant *p*-values. With the exception of *FAT4*, none of the late metastatic drivers were more mutated in primary tumors of LN+ than in primary tumors of LN− patients.

4. Discussion

In this work, we present GoNetic, a flexible network-based driver identification method based on probabilistic pathfinding. GoNetic does not only prioritize driver genes. It also selects the edges from the network prior that are most relevant under the conditions investigated. Hereto, it integrates the network structure, the edge weights, and the sample-specific node weights. GoNetic can handle large tumor cohorts and performs in line with the state-of-the-art.

Applying GoNetic on the HMF cohort showed how the method successfully recovered both known PCa drivers (*SPOP*, *FOXA1*, and *PTEN* [40,41]) and known mPCa drivers (*TP53*, *AR*, *RBI*, *APC*, and *CTNNB1* [1,12,14,37,38]). These genes carry a strong frequency-based driver signal in the HMF cohort. Most were also picked up using SomInaClust and reported by van Dessel et al. [12]. Besides agreeing with previous studies on the statistically strong signals in the data, GoNetic also identified more rarely mutated driver candidates (genes mutated in less than 5% of the samples, i.e., in fewer than 16 samples out of 325 samples). These genes were identified as members of 14 subnetworks, several of which reflect processes related to mPCa disease. The first subnetwork, centered around *TP53*, is enriched in the cell cycle, DNA damage response, and apoptosis. The second network, centered around *AR* and *CTNNB1*, is associated with, among others, the androgen receptor signaling pathway and cell-cell adhesion. Interestingly, the third subnetwork, which is enriched in calcium-dependent cell-cell adhesion processes, mainly consists of protocadherins and cadherins. Cadherin disruption is known to impact tumor progression, cancer cell invasion, and metastasis [42,43]. The fourth subnetwork, centered around *RYR2*, is related to the regulation of the membrane potential, regulation of calcium concentration, and calcium transmembrane transport. Dysregulation of calcium ion signaling and transport remodeling is known to promote cancer cell proliferation [44,45]. In addition, in PCa, the inhibition of *RYR2* has been shown to result in the release of calcium ions and the protection of the cells against apoptosis [46]. The fifth subnetwork is enriched in O-glycan

processing and cell adhesion. Aberrant glycosylation is a cancer hallmark and influences cancer progression [47,48]. The sixth subnetwork is enriched in processes that relate to extracellular matrix (ECM) organization and cell adhesion. The seventh subnetwork is enriched in membrane organization and endocytosis, which has recently been reported as a potential regulator of tumor metastasis [49]. The other subnetworks were not enriched in any relevant processes.

SomInaClust mutational hotspot analysis could confirm one novel tumor suppressor not yet identified in GCG: *SYCP1*. Aberrations in *SYCP1* have been associated with progression toward metastatic disease in castrate-resistant PCa [50].

To further validate the role of these prioritized genes in mPCa, we performed three additional analyses. Firstly, we assessed the degree to which the identified genes were more mutated in metastatic than in primary samples. Secondly, we assessed whether they could be associated with tumor evolution. Finally, we evaluated whether genes could be associated with the lymph node status of the patients.

Based on these analyses, prioritized drivers were subdivided into primary versus metastatic drivers and classified according to their putative time of origin during tumor evolution. Using this scheme, known primary and metastatic drivers were classified as expected, validating the results of the categorization scheme. Indeed, known primary drivers did not show a signal of mutational enrichment in metastases, whereas known metastatic drivers did. Moreover, known primary drivers were found to originate early during tumor evolution, whereas the time of origin of known metastatic drivers was in line with what was reported in the literature, i.e., either being early metastatic (*TP53*, *RB1*, and *CTNNB1*) or late adaptive (*AR*). In addition, *TP53* was found to be significantly associated with a positive lymph node status (p -value $< 10^{-5}$) and hence could be used as an early prognostic factor.

Given the low mutation rate of the rarely mutated genes prioritized by GoNetic, the size of most available validation cohorts limits the power of detecting significant validation signals. Despite this, using these meta-analyses, we found for most of the genes a further validation of their role in metastatic disease in at least one of the validation cohorts. For instance, *TAF1L* is identified as a strong metastatic driver originating during the transition from primary to metastatic disease. In addition, it was found to be significantly associated with a positive lymph node status and hence might be prognostic for advanced disease. Increased expression of the homolog of *TAF1L* (*TAF1*) has been associated with the progression of human PCa to the lethal castration-resistant state [51]. Given the close homology between *TAF1* and *TAF1L*, it was hypothesized that *TAF1L* may have similar regulatory functions in cancer [52].

Despite not being present in the Gundem or Kumar data sets, *CSPG4* was strongly overrepresented in metastatic samples in both the HMF and ARM metastatic cohorts. In addition, it was associated with the positive lymph node status of PCa patients (p -value < 0.1), even though this association is weak. Previous studies have shown that *CSPG4* plays an important role in tumor cell proliferation and migration, as well as with poor prognosis and relapse in breast cancers [53].

MUC16 and *SYNE1* are relatively frequently mutated in the HMF cohort, showing a mutational enrichment signal even after TMB correction in the cohort. They are classified as transitional drivers and *SYNE1* associated with the patient's positive lymph node status (p -value < 0.05). Mutations in both genes have previously been associated with PCa in young men, supporting their putative role in advanced disease [54].

For most of the rarely mutated genes, their role in metastatic disease was supported by only one of the meta-analyses. For instance, *MYH11* was not detected in the Gundem or Kumar data sets used for the association with tumor evolution. Furthermore, it was not detected as enriched in mutations in metastatic samples using the correction for TMB. In contrast, it was found to be significantly associated with the positive lymph node status of the patients. *MYH11* has a role in cell migration and interacts with cell adhesion proteins; additionally, mutations in *MYH11* have been associated with several cancer types [55].

Surprisingly, *APC* and *BRCA2* are not observed in any of the Gundem or Kumar metastatic samples, despite their rather high mutational overrepresentation in the HMF and ARM metastatic samples. This indicates that they might be representative of a subset of mPCa patients that are underrepresented in the Gundem and Kumar data sets but prevalent in the HMF and ARM metastatic cohorts.

5. Conclusions

In this study, we present GoNetic, a network-based analysis framework to perform cancer cohort analysis. GoNetic allows exploiting large cohorts of sample-specific mutational information and properties of the network prior to identifying driver candidates. Being a flexible framework, GoNetic is easily extendable to other data sets (e.g., to combine mutational information with and expression data).

Analysis of the HMF mPCa cohort illustrates the potential of GoNetic in identifying novel drivers. Further validation of those driver candidates through meta-analyses resulted in forwarding several novel putative driver genes for mPCa, some of which might be prognostic for disease evolution.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/cancers13215291/s1>, Figure S1. Correlation between the OnCompare and the Fisher's exact test *p*-values, for the comparison between the primary cohort and the HMF metastatic cohort and for the comparison between the primary cohort and the ARM metastatic cohort. No prior correction for TMB or tumor purity was performed, nor were these factors taken as covariates. Figure S2. Comparison of the tumor purity (left) and tumor mutational burden (right) between the primary cohort (TCGA) and the two metastatic cohorts (HMF and ARM). TMB is evaluated considering all somatic variants that mapped to the regions sequenced by Armenia et al. prior to applying any filtering. Figure S3. Comparison of the tumor mutational burden (right) between the primary cohort (TCGA) with LN⁻ and LN⁺ status. Figure S4. Histogram of the number of mutations per gene. Genes mutated in more than 5% of the samples (left to the red vertical dashed line) are shown in the boxes. In blue are genes identified by GoNetic, and in gray are all the remaining genes mutated in the HMF cohort. Figure S5. Subnetworks identified by GoNetic (left panel: Oncoprint showing the degree to which mutations in genes belonging to the same subnetwork display ME pattern, right panel: subnetwork representation). Figure S6. Correlation between OnCompare *p*-value with and without TMB correction when comparing the two metastatic cohorts to the primary cohort. Here, for each method, the top 87 predicted genes were selected. Figure S7. Results of SomInaClust analysis on the genes prioritized by GoNetic on the HMF cohort. Table S1. GoNetic and Hierarchical HotNet Tokheim benchmarking results. Table S2. Overview of the number and type of variants and samples selected for analysis in the two mPCa cohorts (ARM and HMF) and the primary PCa cohort (TCGA). Table S3. Summary table providing the mutational enrichment *p*-values obtained with and without correcting for TMB and this for, respectively, the HMF and ARM metastatic. Table S4. Genes prioritized by GoNetic on the HMF data set. For each gene, the prioritization rank and the subnetwork to which they belong and their mutational frequency in the HMF cohort are given. Legend: Green = genes mutated in more than 5% of the samples, red = genes mutated in less than 1% of the samples. Table S5. GO enrichment of each GoNetic subnetwork and site-specific characteristics. Table S6. Genes that are significantly more (red)/less (blue) mutated in one metastatic site versus others metastatic sites, LN: lymph node; B: bone; L: lung. Table S7. OnCompare significance level (prior and after TMB correction) at the subnetwork level for the HMF and ARM data sets (bold = subnetwork significantly enriched in mutations in metastatic samples in both metastatic cohorts). Table S8. SomInaClust results of mutations in genes identified by GoNetic in the HMF cohort.

Author Contributions: Conceptualization, K.M. and L.d.S.v.B.; methodology, L.d.S.v.B., G.M., and M.L.; software, G.M. and M.L.; validation, L.d.S.v.B. and J.V.d.E.; formal analysis, L.d.S.v.B.; investigation, L.d.S.v.B.; resources, L.d.S.v.B.; data curation, L.d.S.v.B.; writing—original draft preparation, L.d.S.v.B.; writing—review and editing, L.d.S.v.B., K.M., G.M., M.L., and J.V.d.E.; visualization, L.d.S.v.B.; supervision, K.M.; project administration, K.M.; funding acquisition, L.d.S.v.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by grants of the Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) (G046318, G.0371.06, and 3G045620) and the UGent Bijzonder Onderzoeksfonds. L.S. and M.L. hold a personal FWO grant.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study through the HMF foundation.

Data Availability Statement: Data were obtained from the Hartwig Medical Foundation at <https://www.hartwigmedicalfoundation.nl/en/applying-for-data/> (accessed on 20 January 2020) with the permission of the Hartwig Medical Foundation. Data from Armenia et al. [1] are available at http://www.cbioportal.org/study?id=prad_p1000 (accessed on 11 September 2020). The data from Gundem et al. [13] are available at https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-020-2581-5/MediaObjects/41586_2020_2581_MOESM1_ESM.xlsx (accessed on 18 February 2021), and Kumar et al. [14] are available at https://www.cbioportal.org/study/summary?cancer_study_id=prad_fhcr (accessed on 23 March 2021). The results here include the use of data from The Metastatic Prostate Cancer Project (<https://mpcproject.org/>, accessed on 17 April 2021), a project of Count Me In (<https://joincountmein.org/>, accessed on 17 April 2021). GoNetic software is available at <https://github.com/gmiclotte/gonetic> (accessed on 17 April 2021) for non-commercial academic use.

Acknowledgments: This publication and the underlying research are partly facilitated by Hartwig Medical Foundation and the Center for Personalized Cancer Treatment (CPCT), which have generated, analyzed, and made available data for this research under controlled access. We would like to thank Dries De Maeyer, Bram Weytjens, and Luc De Raedt for the initial developments of PheNetic, on which GoNetic is based. In addition, we thank Dries Van Daele for his help with the CGC benchmark.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Appendix A.1. GO Enrichment Analysis of the Subnetworks Prioritized by GoNetic

Of the subnetworks composed of two genes only, subnetwork eight consisting of *FAT4* and *DCHS2* relates to subnetwork 3 and is enriched in homophilic cell adhesion via plasma membrane adhesion molecules. Subnetworks 10 (*FBXO41*, *RNF213*) and 13 (*HERC5*, *HUWE1*) were both enriched in protein poly-ubiquitination, with a GO enrichment FDR adjusted *p*-value below the 0.01 threshold when both subnetworks were merged (*q*-value = 3.96×10^{-6}). Defects in protein poly-ubiquitination might have an influence on invasion and metastasis [56]. In addition, *RNF213* could be used to monitor response to therapy (i.e., chemo- and radiotherapy) in cervical cancer [57]. In addition, *RNF213* was identified as more frequently in early-stage lung cancer compared to benign nodules and highly expressed, suggesting it could be used as a biomarker for the early diagnosis of lung cancer in pulmonary nodules [58].

The first subnetwork centered around *TP53*, for instance, is significantly enriched in the cell cycle (*TEX15*, *APC*, *RB1*, *TP53*, *ATM*, *TPX2*, *BRCA2*, *PRMT1*, *TAF1L*, *RAD51D*, and *ALMS1*), regulation of signal transduction by p53 class mediators (*TP53*, *ATM*, *TPX2*, *TAF3*, and *TAF1L*), regulation of signaling (*FOXA1*, *APC*, *RB1*, *TP53*, *ATM*, *TPX2*, *KMT2D*, *CSPG4*, *TAF3*, *PTEN*, *SPOP*, *GLI3*, *FOXA2*, *PRMT1*, and *TAF1L*), DNA repair (*TEX15*, *TP53*, *ATM*, *BRCA2*, and *RAD51D*), DNA damage response (*TP53*, *ATM*, *BRCA2*, and *PRMT1*), regulation of EMT transition (*FOXA1*, *PTEN*, and *FOXA2*) and apoptosis (*RB1*, *TP53*, *ATM*, *TPX2*, *PTEN*, *BRCA2*, and *GLI3*).

The second subnetwork, driven by mutations in *AR* and *CTNNB1*, is significantly associated with the androgen receptor signaling pathway (*AR*, *CTNNB1*, and *BRCA1*), which is key to the acquisition of therapy resistance in PCa. In addition, processes related to cell-cell adhesion (*VCL*, *CTNNB1*, *ITGAL*, and *ITGA5*) and extracellular matrix organization (*VWF*, *ITGA5*, *ITGAL*, *MYH11*, and *LAMA2*) are significantly enriched.

The third subnetwork mainly consists of protocadherins and cadherins. Cadherins are calcium-dependent transmembrane proteins that play a role in cell-cell adhesion and regulate cell growth and differentiation. The subnetwork is enriched in homophilic cell adhesion (*PCDHB14*, *FAT2*, *DCHS1*, *PCDH11X*, *PCDH15*, *PCDHA12*, *CDH23*, *PCDHA2*, *FAT3*, and *PCDHA8*), calcium-dependent cell-cell adhesion (*PCDHB14*, *DCHS1*, and *CDH23*), and cell-cell signaling (*PCDHB14*, *PCDH15*, *PCDHA12*, *PCDHA2*, and *PCDHA8*). Cadherin disruption is known to impact tumor formation, progression, drug resistance, cancer cell invasion, metastasis, and angiogenesis [42,43].

The fourth subnetwork is related to the regulation of the membrane potential (*GRIN2C*, *SHANK1*, *CACNA1H*, *RYR2*, and *GRIA1*), glutamate receptor signaling (*GRIN2C*, *GRM1*, and *GRIA1*), regulation of the calcium concentration (*GRIN2C*, *GRM1*, *RYR2*, and *GRIA1*), calcium transmembrane transport (*GRIN2C*, *CACNA1H*, and *RYR2*) and cellular homeostasis (*GRIN2C*, *GRM1*, *RYR2*, *GRIA1*, and *NOS1*). Calcium ions are involved in the regulation of cell migration, proliferation, and cell fate: dysregulation of calcium ion signaling and transport remodeling helps to sustain cancer hallmarks and promote cancer cell proliferation [44,45]. Cancer cells may escape apoptosis by a decrease in calcium ion concentration influx in the cytoplasm [44], and calcium ions released from bones during bone metastasis in breast cancer and abnormal calcium homeostasis can aggravate tumor progression [59]. In addition, in PCa, the inhibition of *RYR2* has been shown to result in the release of calcium ions and the protection of the cells against apoptosis [46].

The fifth subnetwork is enriched in stimulatory C-type lectin receptor signaling pathway (*CLEC10A*, *MUC17*, *MUC16*, and *MUC4*), O-glycan processing (*MUC17*, *GALNT12*, *MUC16*, and *MUC4*), and cell adhesion (*MUC16* and *MUC4*). C-type lectins contribute to the metastatic spread of cancer, especially in lymph node and liver metastasis [47]. In addition, aberrant glycosylation is a cancer hallmark and influences cancer progression, and *MYC* is known to regulate abnormal O-glycosylation resulting in downstream signaling changes promoting castration resistance in PCa and metastasis emergence [48].

The sixth subnetwork is enriched in processes that relate to the extracellular matrix (ECM) organization (*ITGB6*, *COL11A1*, *COL5A1*, and *ITGAX*), cell adhesion (*ITGB6*, *COL5A1*, and *ITGAX*), and collagen fibril organization (*COL11A1* and *COL5A1*), but also in the integrin-mediated signaling pathway (*ITGB6* and *ITGAX*). Cancerous cells are known to interact with ECM components and collagens by using cell surface receptors such as integrin [60]. In addition, integrin beta receptors have already been identified as potential targets to inhibit signaling pathways involved in PCa progression [61,62]. In addition, *COL11A1* promotes the proliferation of cancer cells and enables apoptotic evasion to promote resistance to chemotherapy in pancreatic cancer [63].

The seventh subnetwork was significantly enriched in membrane organization (*GAK*, *SYNJ2*, *AMPH*, and *IGF2R*) and endocytosis (*GAK*, *AMPH*, and *IGF2R*). Endocytosis has recently been reported as a potential regulator of tumor metastasis: multiple endocytic proteins are dysregulated in cancer and impact tumor migration and invasion [49].

Of the subnetworks being composed of two genes only, subnetwork eight consisting of *FAT4* and *DCHS2* relates to subnetwork 3 and is enriched in homophilic cell adhesion via plasma membrane adhesion molecules. Subnetworks 10 (*FBXO41*, *RNF213*) and 13 (*HERC5*, *HUWE1*) were both enriched in protein poly-ubiquitination, with a GO enrichment FDR adjusted *p*-value below the 0.01 threshold when both subnetworks were merged (*q*-value = 3.96×10^{-6}). Defects in protein poly-ubiquitination might have an influence on invasion and metastasis [56]. In addition, *RNF213* could be used to monitor response to therapy (i.e., chemo- and radiotherapy) in cervical cancer [57]. In addition, *RNF213* was identified as more frequently in early-stage lung cancer compared to benign nodules and

highly expressed, suggesting it could be used as a biomarker for the early diagnosis of lung cancer in pulmonary nodules [58]. In the ninth subnetwork, no GO enrichment term was reported as significant.

Appendix A.2. Comparison of the Results of GoNetic with Those Obtained by Previous Studies

Applying GoNetic on the HMF mPCa cohort showed how the method successfully recovered already known primary (*SPOP*, *FOXA1*, and *PTEN* [40,41]) and metastatic drivers of PCa (*TP53*, *AR*, *RB1*, *APC* and *CTNNB1* [1,12,14,37,38]). These genes carry a strong frequency-based driver signal in the HMF cohort as they were also picked up by previous studies that used the same cohort data.

van Dessel et al. [12] identified, using a previous version of the HMF data (DR-011), *AR*, *TP53*, *ZMYM3*, *APC*, *RB1*, *CDK12*, *ERF*, and *ZFP36L2* as genes that were differentially altered between the metastatic and the primary cohorts. Of those genes, *CDK12*, *ERF*, and *ZFP36L2* were not recovered by GoNetic but were also not found to be significantly more enriched in mutations in the HMF cohort compared to the TCGA cohort based on our statistical test, even not after omitting the TMB correction. *ZMYM3*, a gene that passed our mutational enrichment test, was missed GoNetic as it was not present in Reactome.

Armenia et al. [1] reported nine genes (i.e., *TP53*, *AR*, *PTEN*, *RB1*, *FOXA1*, *APC*, *BRCA2*, *KMT2C*, and *KMT2D*) in their cohort that were, according to their test statistic, more frequently mutated in metastatic than in primary samples. All of these genes were also reported by GoNetic, but only *TP53*, *AR*, *RB1*, *BRCA2*, and *APC* met the stringent mutational enrichment criteria used in our study to qualify as being mutationally enriched in metastases.

Appendix A.3. SomInaClust on GoNetic Candidate Drivers

SomInaClust was used to identify whether the drivers prioritized by GoNetic show mutational patterns reminiscent of oncogenes (OG) or tumor suppressor genes (TSG), represented by, respectively, positional clustering of the aberrations or enrichment of inactivating mutations in a gene.

Of the 82 drivers prioritized by GoNetic in the HMF cohort, 12 contained mutational hotspots (14.63%, Figure S7, and Table S8). Of those, all but one (i.e., *SYCP1*) were already reported in CGC. *SYCP1* is required for the assembly of synaptonemal complexes, for centromere pairing during meiosis, and for meiotic chromosome synapsis during oocyte and spermatocyte development. *SYCP1* is identified by GoNetic in subnetwork 11 together with *DIDO1*, *DIDO1* expression has been shown to have prognostic value in PCa [64]. Aberrations in *SYCP1* have been associated with progression toward metastatic disease in castrate-resistant PCa [50]. Except for *PTEN* (14/325), *BRCA1* (7/325), and *SYCP1* (5/325), which were rather infrequently mutated in the HMF cohort, all other genes prioritized by GoNetic that were shown to contain mutational hotspots were mutated in more than 5% of the samples in the HMF cohort.

References

1. Armenia, J.; Wankowicz, S.A.M.; Liu, D.; Gao, J.; Kundra, R.; Reznik, E.; Chatila, W.K.; Chakravarty, D.; Han, G.C.; Coleman, I.; et al. The long tail of oncogenic drivers in prostate cancer. *Nat. Genet.* **2018**, *50*, 645–651. [CrossRef] [PubMed]
2. Reyna, M.A.; Haan, D.; Paczkowska, M.; Verbeke, L.P.C.; Vazquez, M.; Kahraman, A.; Pulido-Tamayo, S.; Barenboim, J.; Wadi, L.; Dhingra, P.; et al. Pathway and network analysis of more than 2500 whole cancer genomes. *Nat. Commun.* **2020**, *11*, 1–17. [CrossRef]
3. Leiserson, M.D.M.; Vandin, F.; Wu, H.-T.; Dobson, J.R.; Eldridge, J.V.; Thomas, J.L.; Papoutsaki, A.; Kim, Y.; Niu, B.; McLellan, M.; et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **2015**, *47*, 106–114. [CrossRef] [PubMed]
4. Dimitrakopoulos, C.; Hindupur, S.K.; Häfliger, L.; Behr, J.; Montazeri, H.; Hall, M.N.; Beerenwinkel, N. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* **2018**, *34*, 2441–2448. [CrossRef] [PubMed]
5. Reimand, J.; Bader, G.D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **2013**, *9*, 637. [CrossRef] [PubMed]

6. Reyna, M.A.; Leiserson, M.D.M.; Raphael, B.J. Hierarchical HotNet: Identifying hierarchies of altered subnetworks. *Bioinformatics* **2018**, *34*, i972–i980. [[CrossRef](#)] [[PubMed](#)]
7. Hofree, M.; Shen, J.P.; Carter, H.; Gross, A.; Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **2013**, *10*, 1108–1115. [[CrossRef](#)]
8. Verbeke, L.P.C.; Van den Eynden, J.; Fierro, A.C.; Demeester, P.; Fostier, J.; Marchal, K. Pathway Relevance Ranking for Tumor Samples through Network-Based Data Integration. *PLoS ONE* **2015**, *10*, e0133503. [[CrossRef](#)] [[PubMed](#)]
9. Rentzsch, P.; Witten, D.; Cooper, G.M.; Shendure, J.; Kircher, M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **2019**, *47*, D886–D894. [[CrossRef](#)]
10. Shihab, H.A.; Rogers, M.F.; Gough, J.; Mort, M.; Cooper, D.N.; Day, I.N.M.; Gaunt, T.R.; Campbell, C. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **2015**, *31*, 1536–1543. [[CrossRef](#)]
11. Helgstrand, J.T.; Røder, M.A.; Klemann, N.; Toft, B.G.; Lichtensztajn, D.Y.; Brooks, J.D.; Brasso, K.; Vainer, B.; Iversen, P. Trends in incidence and 5-year mortality in men with newly diagnosed, metastatic prostate cancer—A population-based analysis of 2 national cohorts. *Cancer* **2018**, *124*, 2931–2938. [[CrossRef](#)]
12. Van Dessel, L.F.; van Riet, J.; Smits, M.; Zhu, Y.; Hamberg, P.; van der Heijden, M.S.; Bergman, A.M.; van Oort, I.M.; de Wit, R.; Voest, E.E.; et al. The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact. *Nat. Commun.* **2019**, *10*, 5251. [[CrossRef](#)]
13. Gundem, G.; Van Loo, P.; Kremeyer, B.; Alexandrov, L.B.; Tubio, J.M.C.; Papaemmanuil, E.; Brewer, D.S.; Kallio, H.M.L.; Högnäs, G.; Annala, M.; et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* **2015**, *520*, 353–357. [[CrossRef](#)]
14. Kumar, A.; Coleman, I.; Morrissey, C.; Zhang, X.; True, L.D.; Gulati, R.; Etzioni, R.; Bolouri, H.; Montgomery, B.; White, T.; et al. Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nat. Med.* **2016**, *22*, 369–378. [[CrossRef](#)]
15. Crowdis, J.; Balch, S.; Sterlin, L.; Thomas, B.S.; Camp, S.Y.; Dunphy, M.; Anastasio, E.; Shah, S.; Damon, A.L.; Ramos, R.; et al. A patient-driven clinicogenomic partnership through the Metastatic Prostate Cancer Project. *bioRxiv* **2021**. [[CrossRef](#)]
16. De Maeyer, D.; Renkens, J.; Cloots, L.; De Raedt, L.; Marchal, K. PheNetic: Network-based interpretation of unstructured gene lists in *E. coli*. *Mol. Biosyst.* **2013**, *9*, 1594–1603. [[CrossRef](#)]
17. De Maeyer, D.; Weytjens, B.; Renkens, J.; De Raedt, L.; Marchal, K. PheNetic: Network-based interpretation of molecular profiling data. *Nucleic Acids Res.* **2015**, *43*, W244–W250. [[CrossRef](#)]
18. Perez-Romero, C.A.; Weytjens, B.; Decap, D.; Swings, T.; Michiels, J.; De Maeyer, D.; Marchal, K. IAMBEE: A web-service for the identification of adaptive pathways from parallel evolved clonal populations. *Nucleic Acids Res.* **2019**, *47*, W151–W157. [[CrossRef](#)] [[PubMed](#)]
19. Gitter, A.; Klein-Seetharaman, J.; Gupta, A.; Bar-Joseph, Z. Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Res.* **2011**, *39*, e22. [[CrossRef](#)] [[PubMed](#)]
20. Navlakha, S.; Gitter, A.; Bar-Joseph, Z. A Network-based Approach for Predicting Missing Pathway Interactions. *PLOS Comput. Biol.* **2012**, *8*, 1–13. [[CrossRef](#)]
21. Darwiche, A. New advances in compiling CNF to decomposable negation normal form. In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004), Valencia, Spain, 22–27 August 2004; pp. 328–332.
22. Tokheim, C.J.; Papadopoulos, N.; Kinzler, K.W.; Vogelstein, B.; Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 14330–14335. [[CrossRef](#)] [[PubMed](#)]
23. Davoli, T.; Xu, A.W.; Mengwasser, K.E.; Sack, L.M.; Yoon, J.C.; Park, P.J.; Elledge, S.J. XCumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **2013**, *155*, 948. [[CrossRef](#)] [[PubMed](#)]
24. Mularoni, L.; Sabarinathan, R.; Deu-Pons, J.; Gonzalez-Perez, A.; López-Bigas, N. OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **2016**, *17*, 128. [[CrossRef](#)] [[PubMed](#)]
25. Lawrence, M.S.; Stojanov, P.; Mermel, C.H.; Robinson, J.T.; Garraway, L.A.; Golub, T.R.; Meyerson, M.; Gabriel, S.B.; Lander, E.S.; Getz, G. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **2014**, *505*, 495–501. [[CrossRef](#)]
26. Tamborero, D.; Gonzalez-Perez, A.; Lopez-Bigas, N. OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **2013**, *29*, 2238–2244. [[CrossRef](#)] [[PubMed](#)]
27. Dees, N.D.; Zhang, Q.; Kandoth, C.; Wendl, M.C.; Schierding, W.; Koboldt, D.C.; Mooney, T.B.; Callaway, M.B.; Dooling, D.; Mardis, E.R.; et al. MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **2012**, *22*, 1589–1598. [[CrossRef](#)] [[PubMed](#)]
28. Gonzalez-Perez, A.; Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **2012**, *40*, 1–10. [[CrossRef](#)]
29. Wu, G.; Haw, R. Functional interaction network construction and analysis for disease discovery. In *Protein Bioinformatics*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 235–253.
30. Kircher, M.; Witten, D.M.; Jain, P.; O’Roak, B.J.; Cooper, G.M.; Shendure, J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **2014**, *46*, 310–315. [[CrossRef](#)]
31. Liu, X.; Jian, X.; Boerwinkle, E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* **2011**, *32*, 894–899. [[CrossRef](#)]
32. Ferlaino, M.; Rogers, M.F.; Shihab, H.A.; Mort, M.; Cooper, D.N.; Gaunt, T.R.; Campbell, C. An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome. *BMC Bioinform.* **2017**, *18*, 1–8. [[CrossRef](#)]

33. Van den Eynden, J.; Fierro, A.C.; Verbeke, L.P.C.; Marchal, K. SomInaClust: Detection of cancer genes based on somatic mutation patterns of inactivation and clustering. *BMC Bioinform.* **2015**, *16*, 125. [[CrossRef](#)]
34. Sondka, Z.; Bamford, S.; Cole, C.G.; Ward, S.A.; Dunham, I.; Forbes, S.A. The COSMIC Cancer Gene Census: Describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **2018**, *18*, 696–705. [[CrossRef](#)]
35. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A.J.; Kinzler, K.W. Cancer genome landscapes. *Science* **2013**, *339*, 1546–1558. [[CrossRef](#)]
36. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613. [[CrossRef](#)]
37. Grasso, C.S.; Wu, Y.-M.; Robinson, D.R.; Cao, X.; Dhanasekaran, S.M.; Khan, A.P.; Quist, M.J.; Jing, X.; Lonigro, R.J.; Brenner, J.C.; et al. The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **2012**, *487*, 239–243. [[CrossRef](#)]
38. Robinson, D.; Van Allen, E.M.; Wu, Y.-M.; Schultz, N.; Lonigro, R.J.; Mosquera, J.-M.; Montgomery, B.; Taplin, M.-E.; Pritchard, C.C.; Attard, G.; et al. Integrative clinical genomics of advanced prostate cancer. *Cell* **2015**, *161*, 1215–1228. [[CrossRef](#)] [[PubMed](#)]
39. Roy, R.; Chun, J.; Powell, S.N. BRCA1 and BRCA2: Different roles in a common pathway of genome protection. *Nat. Rev. Cancer* **2012**, *12*, 68–78. [[CrossRef](#)]
40. Barbieri, C.E.; Baca, S.C.; Lawrence, M.S.; Demichelis, F.; Blattner, M.; Theurillat, J.-P.; White, T.A.; Stojanov, P.; Van Allen, E.; Stransky, N.; et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.* **2012**, *44*, 685–689. [[CrossRef](#)] [[PubMed](#)]
41. Berger, M.F.; Lawrence, M.S.; Demichelis, F.; Drier, Y.; Cibulskis, K.; Sivachenko, A.Y.; Sboner, A.; Esgueva, R.; Pflueger, D.; Sougnez, C.; et al. The genomic complexity of primary human prostate cancer. *Nature* **2011**, *470*, 214–220. [[CrossRef](#)] [[PubMed](#)]
42. Yu, W.; Yang, L.; Li, T.; Zhang, Y. Cadherin signaling in cancer: Its functions and role as a therapeutic target. *Front. Oncol.* **2019**, *9*, 989. [[CrossRef](#)] [[PubMed](#)]
43. Berx, G.; Van Roy, F. Involvement of members of the cadherin superfamily in cancer. *Cold Spring Harb. Perspect. Biol.* **2009**, *1*, a003129. [[CrossRef](#)]
44. Prevarskaya, N.; Ouadid-Ahidouch, H.; Skryma, R.; Shuba, Y. Remodelling of Ca²⁺ transport in cancer: How it contributes to cancer hallmarks? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2014**, *369*, 20130097. [[CrossRef](#)]
45. Prevarskaya, N.; Skryma, R.; Shuba, Y. Calcium in tumour metastasis: New roles for known actors. *Nat. Rev. Cancer* **2011**, *11*, 609–618. [[CrossRef](#)] [[PubMed](#)]
46. Mariot, P.; Prevarskaya, N.; Roudbaraki, M.M.; Le Bourhis, X.; Van Coppenolle, F.; Vanoverberghe, K.; Skryma, R. Evidence of functional ryanodine receptor involved in apoptosis of prostate cancer (LNCaP) cells. *Prostate* **2000**, *43*, 205–214. [[CrossRef](#)]
47. Ding, D.; Yao, Y.; Zhang, S.; Su, C.; Zhang, Y. C-type lectins facilitate tumor metastasis. *Oncol. Lett.* **2017**, *13*, 13–21. [[CrossRef](#)]
48. Tzeng, S.-F.; Tsai, C.-H.; Chao, T.-K.; Chou, Y.-C.; Yang, Y.-C.; Tsai, M.-H.; Cha, T.-L.; Hsiao, P.-W. O-Glycosylation-mediated signaling circuit drives metastatic castration-resistant prostate cancer. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **2018**, fj201800687. [[CrossRef](#)] [[PubMed](#)]
49. Khan, I.; Steeg, P.S. Endocytosis: A pivotal pathway for regulating metastasis. *Br. J. Cancer* **2021**, *124*, 66–75. [[CrossRef](#)]
50. Nickerson, M.L.; Im, K.M.; Misner, K.J.; Tan, W.; Lou, H.; Gold, B.; Wells, D.W.; Bravo, H.C.; Fredrikson, K.M.; Harkins, T.T.; et al. Somatic alterations contributing to metastasis of a castration-resistant prostate cancer. *Hum. Mutat.* **2013**, *34*, 1231–1241. [[CrossRef](#)]
51. Tavassoli, P.; Wafa, L.A.; Cheng, H.; Zoubeydi, A.; Fazli, L.; Gleave, M.; Snoek, R.; Rennie, P.S. TAF1 differentially enhances androgen receptor transcriptional activity via its N-terminal kinase and ubiquitin-activating and -conjugating domains. *Mol. Endocrinol.* **2010**, *24*, 696–708. [[CrossRef](#)]
52. Zhong, S.; Yan, H.; Chen, Z.; Li, Y.; Shen, Y.; Wang, Y.; Li, L.; Sheng, S.; Wang, Y. Overexpression of TAF1L promotes cell proliferation, migration and invasion in esophageal squamous cell carcinoma. *J. Cancer* **2019**, *10*, 979. [[CrossRef](#)]
53. Ilieva, K.M.; Cheung, A.; Mele, S.; Chiaruttini, G.; Crescioli, S.; Griffin, M.; Nakamura, M.; Spicer, J.F.; Tsoka, S.; Lacy, K.E.; et al. Chondroitin sulfate proteoglycan 4 and its potential as an antibody immunotherapy target across different tumor types. *Front. Immunol.* **2018**, *8*, 1911. [[CrossRef](#)]
54. Maistro, S.; dos Santos Xavier, C.; Serio, P.A.M.P.; Katayama, M.L.H.; Roela, R.A.; Koike Folgueira, M.A.A. Cancer driver genes in prostate cancer from young men. *J. Clin. Oncol.* **2019**, *37*, e16586. [[CrossRef](#)]
55. Nie, M.; Pan, X.; Tao, H.; Xu, M.; Liu, S.; Sun, W.; Wu, J.; Zou, X. Clinical and prognostic significance of MYH11 in lung cancer. *Oncol Lett* **2020**, *19*, 3899–3906. [[CrossRef](#)] [[PubMed](#)]
56. Song, Y.; Xu, Y.; Pan, C.; Yan, L.; Wang, Z.; Zhu, X. The emerging role of SPOP protein in tumorigenesis and cancer therapy. *Mol. Cancer* **2020**, *19*, 2. [[CrossRef](#)] [[PubMed](#)]
57. Lee, S.-Y.; Chae, D.-K.; Lee, S.-H.; Lim, Y.; An, J.; Chae, C.H.; Kim, B.C.; Bhak, J.; Bolser, D.; Cho, D.-H. Efficient mutation screening for cervical cancers from circulating tumor DNA in blood. *BMC Cancer* **2020**, *20*, 694. [[CrossRef](#)] [[PubMed](#)]
58. Jiang, N.; Zhou, J.; Zhang, W.; Li, P.; Liu, Y.; Shi, H.; Zhang, C.; Wang, Y.; Zhou, C.; Peng, C.; et al. RNF213 gene mutation in circulating tumor DNA detected by targeted next-generation sequencing in the assisted discrimination of early-stage lung cancer from pulmonary nodules. *Thorac. Cancer* **2021**, *12*, 181–193. [[CrossRef](#)]
59. Yang, Z.; Yue, Z.; Ma, X.; Xu, Z. Calcium homeostasis: A potential vicious cycle of bone metastasis in breast cancers. *Front. Oncol.* **2020**, *10*, 293. [[CrossRef](#)]

60. Bourgot, I.; Primac, I.; Louis, T.; Noël, A.; Maquoi, E. Reciprocal Interplay Between Fibrillar Collagens and Collagen-Binding Integrins: Implications in Cancer Progression and Metastasis. *Front. Oncol.* **2020**, *10*, 1488. [[CrossRef](#)] [[PubMed](#)]
61. Juan-Rivera, M.C.; Martínez-Ferrer, M. Integrin inhibitors in prostate cancer. *Cancers* **2018**, *10*, 44. [[CrossRef](#)] [[PubMed](#)]
62. Quaglia, F.; Krishn, S.R.; Wang, Y.; Goodrich, D.W.; McCue, P.; Kossenkov, A.V.; Mandigo, A.C.; Knudsen, K.E.; Weinreb, P.H.; Corey, E.; et al. Differential expression of $\alpha V\beta 3$ and $\alpha V\beta 6$ integrins in prostate cancer progression. *PLoS ONE* **2021**, *16*, e0244985. [[CrossRef](#)]
63. Wang, H.; Ren, R.; Yang, Z.; Cai, J.; Du, S.; Shen, X. The COL11A1/Akt/CREB signaling axis enables mitochondrial-mediated apoptotic evasion to promote chemoresistance in pancreatic cancer cells through modulating BAX/BCL-2 function. *J. Cancer* **2021**, *12*, 1406–1420. [[CrossRef](#)]
64. Lyu, P.; Zhang, S.-D.; Yuen, H.-F.; McCrudden, C.M.; Wen, Q.; Chan, K.-W.; Kwok, H.F. Identification of TWIST-interacting genes in prostate cancer. *Sci. China Life Sci.* **2017**, *60*, 386–396. [[CrossRef](#)]