Driver Gene Identification for Prostate Cancer Using an Integrative Probabilistic Graphical Model

Raquel Aoki
Simon Fraser University
Vancouver, BC
raoki@sfu.ca

Oliver Snow
Simon Fraser University
Vancouver, BC
osnow@sfu.ca

ABSTRACT

The availability of large-scale genomic data combined with the recent success of targeted therapies has motivated researchers to characterize the genes that drive cancer progression. Despite strong efforts using both experimental and computational approaches, uncovering true driver genes has remained challenging. The most successful methods have integrated multiple sources of genomic data, combining several weak signals to identify drivers. Here we present a probabilistic graphical model, which incorporates different types of prostate cancer data into one score to discern driver genes from passenger genes. Our final model finds 21 bona-fide prostate cancer driver genes, demonstrating the flexibility of this approach to be applied to different cancer types. Future work would aim to improve the model to expand this list of identified drivers in the hopes of finding novel drug targets.

KEYWORDS

Probabilistic graphical model, prostate cancer, driver gene

1 INTRODUCTION

Recent advancements in next generation sequencing technology has enabled the collection of large genomic datasets from cancer patients through projects such as The Cancer Genome Atlas (TCGA) [1] and the International Cancer Genome Consortium (ICGC) [2]. These datasets have in turn allowed large scale analysis of the underlying genetic causes of cancer and identification of drug targets. Despite this progression, however, it has remained challenging to uncover true driver genes that promote cell growth and ultimately cause cancer. [3]

Furthermore, driver gene discovery has largely focused on point mutations as these are easier to discover than other genomic aberrations. Some popular tools include Mutsig, OncodriveClust, OncodriveFM, MuSiC, and ActiveDriver. [4] These methods work by identifying different signals of positive selection (high frequency or clusters of mutations) and they have been reasonably successful in identifying causal genes. Unfortunately, even the most common point mutations, such as in the tumour suppressor gene TP53, are only found in approximately 20% of patients, thus making any targeted treatment only effective for a small subset of patients. [5] Copy number alterations (CNA) on the other hand can be found in much larger groups of patients (sometimes 70%) but the lack of a definitive causal gene within an altered region makes it hard to identify the underlying driver. [6] CNAs therefore present a great potential for the discovery of cancer genes and to develop effective targeted treatments if the challenge of pinpointing specific genes in a region can be overcome.

A study by Sanchez-Garcia et al. in 2014 [7] focused on CNAs for discovering driver genes in breast cancer by integrating data from TCGA with small interfering RNA (siRNA) functional screens. siRNA screens have been a valuable tool for uncovering driver genes by selectively knocking down a single gene in a single cell and then observing which genes are crucial for cell growth. However, these screens often suffer from experimental noise and off-target effects, decreasing confidence in the driver gene predictions drawn from them. [8]

Sanchez-Garcia et al.'s algorithm, which they named Helios, first identified areas of the genome that had altered copy numbers with a method they called Identification of Significantly Altered Regions (ISAR). These regions were then used as input to a probabilistic graphical model that incorporated information from multiple data types including gene expression, somatic mutations, and siRNA knockdown effects, in order to determine the most likely driver in each region. The strength of their approach was that the combination of multiple data types avoided the drawbacks or bias of any one source of data. This enabled them to identify driver genes that would normally be missed by previous methods that focused on only one signal of positive selection. The researchers were able to discover 29 high-confidence driver genes, doubling previous lists. Many of these genes were bona-fide breast cancer drivers and 10 previously unidentified genes were validated in an anchorage-independent growth assay, demonstrating the strength of their approach.

In this work, we develop a similar algorithm as Helios, incorporating multiple types of open-source genomic data from TCGA with siRNA data from Project Achilles [9] in order to identify driver genes in prostate cancer rather than breast cancer, which the previous work focused on. While

breast cancer is the most common malignancy in females in North America, prostate cancer is the most common for males, affecting approximately 1 in 9 men. Prostate cancer also shares many molecular similarities with breast cancer as both cancers are characterized by alterations in steroid hormones. [10] Therefore, prostate cancer is the most logical next step for applying this integrative approach in order to discover driver genes. Our final method, modelled after Helios but with some significant differences, identifies 239 driver genes, of which 21 are found in the cancer gene census list (33 in total), thus confirming the potential of this approach. Future work should aim to apply this method to other cancer types or integrate more sources of data to expand the list of known drivers and ultimately find viable drug targets.

2 METHODS

2.1 Data preprocessing

Genomic data including copy number variation, mRNA expression, and somatic mutation was downloaded using the Firebrowse tool from the Broad Institute as well as siRNA data from the Project Achilles database as mentioned above. [11] CNA data was first analyzed with GISTIC, resulting in a g-score indicating the significance of alteration for different genomic regions. [12] Genes were then mapped to regions giving a g-score for each gene and with multiple genes mapping the same region. GISTIC computes this g-score by computing a null distribution of CNA across the whole genome, which is different than the previously mentioned ISAR method that uses a sliding window to compute a local significance. We did not have access to the ISAR method so we chose to use GISTIC as it is freely available and the results of the two algorithms should be similar.

Expression data used was log2 normalized RNA-seq by Expectation Maximization (RSEM) scores, as this is a commonly used and accurate measure of gene expression. [13] In order to create a single expression value for each gene to be input into our model, we took the average RSEM score across all 550 samples. Somatic mutation p-values for each gene were obtained from Mutsig and then log-transformed to scale the values. [14] The functional siRNA screen data is the inferred gene knockdown effect in primary prostate tumor cell lines. Z-scores were obtained from the DEMETER algorithm, which models and accounts for off-target effects of siRNAs.

The expression, mutation, and siRNA datasets were then combined into a single data frame resulting in a total of 21,257 genes as rows and the three data types as columns. Not all genes were measured in each assay, which means there were numerous missing values. Figure 1 shows their distributions in their original scales. Then, before inputting the features

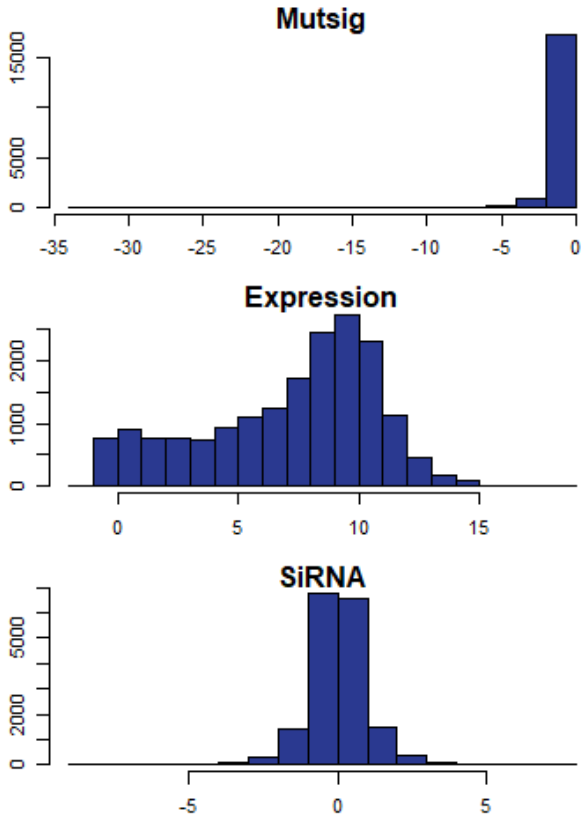

Figure 1: Feature distributions in their original scales

into the model, the values were normalized. Lastly, the signs of the mutation data and siRNA data were flipped (multiplied by -1) so that large positive values meant more likely to be a driver gene.

2.2 Model

The problem of this project can be summarized as a classification problem, where each gene with a significant number of amplifications is classified either as driver gene or passenger gene.

One of the challenges is to create a model that can combine different sources of data in an independent way. This is necessary because part of our data comes from a combination of different data sources that are fed through a logistic regression model while the g-score information is added considering that driver and passengers genes have two different normal distributions, as shown in Figure 2. The general idea is that driver genes have a larger g-score than passenger genes.

The most suitable model for addressing this challenge of incorporating multiple data types is a Probabilistic Graphical

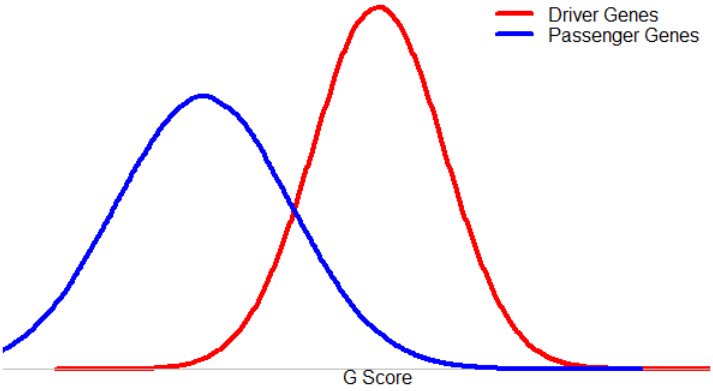

Figure 2: Theoretical hypothesis about the G-score distribution of driver and passenger genes

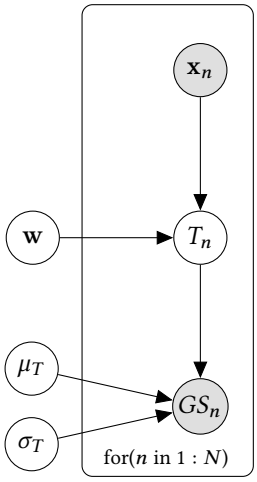

Figure 3: Complete Model - N is the number of genes

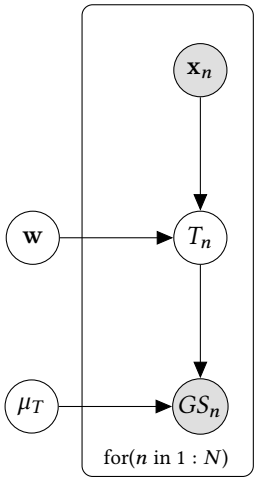

Figure 4: Simplified model - N is the number of genes

Model (PGM) due to its flexibility. The structure of the PGM used in this project is shown in Figure 3.

The T_n node represents the class of each gene from the set of N genes considered in the model. Here it is important to point out that this work does not consider all human genes, but only the genes with a large g-score. This filter is necessary because we only want to classify genes as drivers or passengers; if we had included all genes, we would need to add another label/class for the unaltered genes.

The x_n node represents the feature values of $gene_n$ and is an observed value. $\mathbf{w}$ represents the weights of these features, which are unknown, but will be learned by the PGM. The idea is to connect the features and weights with the T_n node using a logistic regression model, as show in Equation 1.

$$P(T = 1|x, w) = Logit(wx) = \frac{1}{1 + exp(-xw)} \quad (1)$$

The second part of the graphical model aims to approximate two different normal distributions, one for driver genes and the other for passenger genes, as shown in Equation 2. For this task, we use the node GS_n, that represents the observed value of the g-score. This node depends on T_n, the previous classification of that gene (1 if driver and 0 otherwise) as well as μ_T and σ_T that are the mean and variance of the normal distribution. The values μ_T and σ_T, for $T = 0, 1$, are unobserved values and they need to be learned by the model.

$$P(GScore|\mu_T, \sigma_T) = Normal(\mu_T, \sigma_T) \quad (2)$$

Priori distributions were defined for unknown parameters in the Probabilistic Graphical Model. The Equation 3 shows the distributions used in this work. The $\mathbf{w}$ parameter represents the logistic regression weights, and it was set as a normal distribution with mean 0 and variance equal to 0.25. The μ_0(passenger genes average G-score) and μ_1(driver genes average G-score) prioris were also defined as normal distributions, but μ_1 has a bigger g-score mean and a lower variance. Finally, the variances σ_0 and σ_1 were defined as Uniform distributions. Initially, a Gamma Distribution was used because $\sigma_i, \forall 0, 1$ can not assume negative values and this distribution is commonly used as the priori distribution of unknown variance parameters in PGM. However, the variance of our problem is very small (about 0.0004) and the Gamma distribution is not very well defined for small shape parameters, which thus caused convergence problems in our model. In order to continue working with σ and learn its value on the model, we changed from a Gamma distribution to a Uniform distribution.

$$
\begin{aligned}
&Prioris \\
&\quad P(\mathbf{w}) = Normal(0, 0.25) \\
&\quad P(\mu_0) = Normal(0.04, 0.003) \\
&\quad P(\mu_1) = Normal(0.07, 0.001) \\
&\quad P(\sigma_0^2) = Uniform() \\
&\quad P(\sigma_1^2) = Uniform()
\end{aligned} \tag{3}
$$

Like most PGMs, our posterior distribution does not have an exact form. But using the approximation $Posterior \propto Priori \times Likelihood$, we can obtain a sample from the Posterior Distribution with Markov Chain Monte Carlo (MCMC) sampling, as it is described in Section 2.3.

We also work with a simpler version of our model, as shown on Figure 4. In this second model, the variance was fixed and only the mean and weight were estimated using the PGM. This new approach was used to overcome the initial convergence problems in the model.

The gene labels on node T are updated each time a new proposed value is accepted (the proposal of new values and acceptance are discussed in more details at Section 2.3). This update is made using the logistic regression output that is the probability p_n of being a driver gene (belongs to class 1) times the probability of that gene have the observed g-score given its class. Its probability to be classified as passenger($T = 0$) or driver($T = 1$) is then normalized to sum up 1. Besides, in this model, instead of using a threshold as is normally used ($p_n > 0.5$ is class 1 and 0 otherwise), we use a probabilistic approach, by sampling the T_n label from a $Bernoulli(p_n)$. We save all T labels during the MCMC sampling, and in the end we use them to estimate a P_n that defines the final classification of T_n using a $Bernoulli(P_n)$. The P_n is calculated after removing the burn-in period by summing up all iterations where the gene n was classified as a driver and divided by the iteration number.

A challenge faced in this project was the lack of a good training set. In other words, we do not have a reliable T classification set to train our model. As made by [7], we define some driver gene seeds as the genes with highest amplification score, however, there is some issues by using this approach as we discuss at Section 4.

2.3 MCMC sampling

In this section we present a short description about the MCMC algorithm used as a sampling method.

The first step during the simulation process is to propose new parameter values that are compared with the current parameter values. The comparison is made using the ratio of the current parameters over the new values proposed by the proportional approximated posterior distribution. This new set of parameters is accepted according to this ratio r. If r is bigger than 1, the new values are accepted; if r is on the interval $[0, 1]$, then the new set of values are accepted with probability equal to r. This process is repeated K times and the first 20% of samples are used as burn-in period. A MCMC pseudocode is shown below to illustrate how the posterior sample is obtained.

```
#----#----#----#----#----#----#----#
#k = MCMC iterations
#c_coef = current values
#p_coef = proposed values
#p() = posterior distribution
c_coef = initial_values
for (k in 1:K)
        p_coef = proposal(c_coef, sd)
        r = p(p_coef)/ p(c_coef)
        if(bernoulli(r)==1)
            chain[k+1] = prop_coef
            curr_coef = prop_coef
        else
            chain[k+1] = curr_coef
#----#----#----#----#----#----#----#
```

The model convergence is evaluated using graphics with the chain of sampled values of the unknown parameters and the acceptance rate. The convergence of the model is very related to the quality of the values proposed to be part of the sample, as it usually reflects directly on the acceptance rate. To propose new values we use a proposed distribution set as a Normal curve. After defining this distribution, the challenge is to define its variance. While the mean used is the current chain value, the variance is defined as a hyperparameter. A large variance can imply large steps on the random walk and a low acceptance rate, as only a few values will be accepted; a small variance, on the other hand, can lead to a high acceptance rate, because all values proposed are going to be accepted. Although at first glance a high acceptance rate looks like a good result, in practice the model can be stuck in a local minimum and/or does not explore all possibilities.

3 RESULTS

For both models (complete and simplified) we simulated 8000 samples and used a burn-in period of 1600. The parameter convergence of the complete model is shown on Figure 5. As we can see, the parameters only oscillate around an average value and its acceptance rate was around 28.4%, a reasonable value compatible with most works present in the literature[15]. These results indicate that the model achieves convergence.

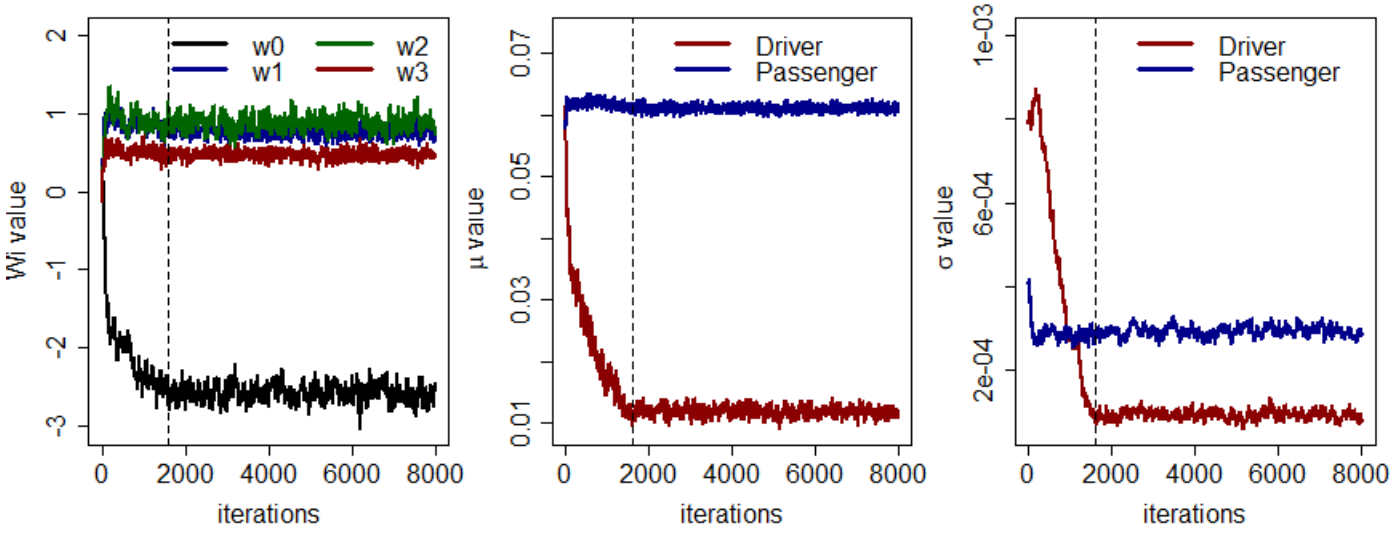

Figure 5: Evaluation of parameter convergence in the complete model. The vertical line shows the end of the burn-in period.

The simplified model from Figure 4 also achieves convergence, as Figure 6 shows. The main difference between these two models is the σ parameter. While in the complete model we try to learn σ, in the simplified model we work with a fixed σ. The simplified model has a acceptance rate of 32.2%, which is also a reasonable value.

Besides the final classification obtained from the PGM as explained in the Section 2.2, we also select the top 100 genes ranked by their probability P_n of being a driver gene. Table 1 shows the number of driver genes found by each model and its comparison with the golden standard list. This driver gene golden standard list is the results found by other reliable methods that we use to evaluate the quality of the driver genes found by our method. Due the presence of missing data or because they have a small g-score value, only 33 known prostate cancer genes from a list of 109 are present in our final dataset.

According to the results shown in Table 1, the set of driver genes obtained from the complete model top100 has the best precision between the four approaches (12%). Considering the recovery rate, the simplified model genes and complete model genes with a final classification using the probabilistic approach have the highest values (66.6% and 63.6%). It is possible to notice, however, that the simplified model has the lowest precision because it tends to classify more genes as driver genes in comparison with the complete model. Finally, although the two models classify a different number of genes as driver, 237 of the 239 genes classified by the complete model as drivers are also classified as drivers by the simplified model.

Table 1: Summary of number of driver genes (DG) found by each model and comparison with the 33 genes from Golden Standard list (GS)

Model	DG	DG+GS	Precision	Recovery
Complete Prob	239	21	7.2%	63.6%
Complete Top100	100	12	12.0%	36.3%
Simplified Prob	312	22	7.0%	66.6%
Simplified Top100	100	7	7.0%	2.1%

As most driver gene discovery methods, our four models have a good recovery rate but a small precision. This precision, however, is based on the golden standard lists that are known to be limited and biased towards the most common mutations. For a more complete evaluation of the quality of our driver genes list, it is necessary to do more detailed research about each gene classified as driver and conduct possible lab experiments, analyses beyond the scope of this work.

After all considerations about limitations of the the evaluation set and comparisons between the four different approaches, the complete model with a probabilistic final classification was considered the best one.

4 DISCUSSION

As with many other fields in medicine, cancer research has been moving towards more targeted therapies as opposed to general cytotoxic chemotherapies. This personalized approach to cancer medicine, precision oncology, aims to find

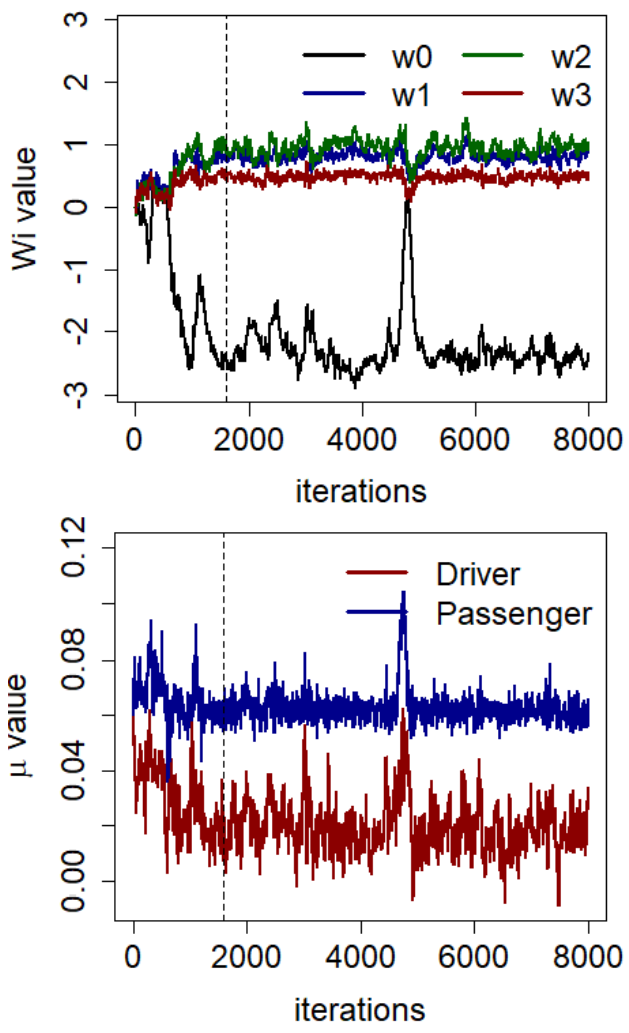

Figure 6: Evaluation of parameter convergence in the simplified model. The vertical line shows the end of the burn-in period.

driver genes that truly cause the tumor, which can then be targeted with highly specific drugs, killing the tumor but with minimal side effects to the patient. There are a number of effective targeted drugs currently on the market and their success has driven efforts to find and understand the landscape of driver genes causing cancer. [16]

However, several factors have made it difficult for driver gene discovery and the development of new precision drugs. One factor is that of tumor heterogeneity, meaning there is both large variation between cancer patients' genomic profiles and large variation even within the same patient. The implications of this is that many currently identified driver genes are only found in a small fraction of patients, making any developed drug not widely useful. The hope is that patients can subtyped into reasonably sized clusters based on their genomic profiles and that drugs could then be developed to treat each cluster, but this is a challenging problem itself. Another factor limiting driver gene discovery is that only a small proportion of genes are druggable; only the genes that are present in the tumor and that are able to be inhibited will be suitable drug targets.

It is important to highlight here the distinction between tumor suppressors and oncogenes. Tumor suppressors can be thought of as the 'brakes', genes whose normal function is to halt cell growth and when eliminated or altered allow uncontrolled cell proliferation. Oncogenes, on the other hand, can be thought of as the 'gas' and are genes that promote growth or evasion of the body's natural control mechanisms. While tumor suppressor genes can be considered driver genes, they are often deleted and thus not present in the tumor or restoring their function with a drug would be too complicated making them rarely suitable as drug targets. Therefore, oncogenes serve as the most suitable drug targets because their activity directly causes the tumor and inhibiting that activity with an antagonistic drug is likely to suppress tumor growth.

In light of the current challenges, one of the most promising ways of finding common driver genes is by targeting CNAs as they are found in a much larger fraction of patients. Here we present an modified application of a previous method developed by Sanchez-Garcia et al. which focuses on CNAs and then pinpoints driver genes within those regions. This approach is flexible and takes advantage of the freely available, diverse genomic datasets from TCGA and other initiatives.

The two models presented in this work found two lists of driver genes for prostate cancer. As most other known methods to identify driver genes, they have a good recovery rate but a poor precision. However, as discussed on Section 3, precision isn't considered a good metric to measure the quality driver genes lists, because it is well known that golden standard lists are limited and many driver genes are yet to be discovered and tested.

As examples of known driver genes identified by our complete model, we present:

- MECOM (also known as EVI1) is a transcription factor and well documented oncogene that, when overexpressed, is associated with cell proliferation and evasion of apoptosis. [17, 18] While its role in acute and chronic myeloid leukemia has been studied extensively, it has only recently been implicated as a driver of prostate cancer. [19]
- TBL1XR1 has also been identified as a prostate cancer driver although its role is less clear. Overexpression of TBL1XR1 is associated with poor prognosis and

increased risk of metastasis in cervical cancer [20] whereas, in prostate cancer, it acts as a tumor suppressor and underexpression or mutation of the gene correlates with cancer progression. [21]

- NDRG1, normally controls cell growth, but is a well known driver of multiple cancer types including prostate cancer. Mutation or underexpression of NDRG1 has been shown to promote metastasis and tumor growth. [22]

Taken together, these three examples demonstrate the multitude of mechanisms by which genes may cause cancer making the task of driver gene discovery complex and difficult. However, the combination of multiple sources of data into a single model can potentially overcome this, as evidenced by our model's ability to identify very distinct drivers through different signals in the genomic data. Despite our complete model not having very high sensitivity (found only 21 out of 33 known drivers), it is able to incorporate multiple datatypes to discern drivers from passengers. With some further tweaking of the model parameters and datasets, the number of identified drivers could be expanded without incurring many false positives.

Future work would aim to address some of the limitations of our implementation and also test some alternative approaches. One of the major limitations of our method is that GISTIC may miss some regions of significant alteration that the ISAR method does not by computing its significance locally. Additionally, genes within the same region are given the same g-score meaning that there is no way to differentiate between genes in a region. Implementing ISAR instead of GISTIC could bring significant improvements to our model. Another improvement to our model could be to work with more features for the logistic regression and also look for more complete data as many genes were excluded from our analysis due to missing values in the genomic features. Another possibility would be to explore different sets of initial driver genes used as input to the model. Finally, this model can also be expanded to other cancer types types for which data is publicly available on the TCGA.

The complete source code to models and graphics can be found here.

REFERENCES

[1] Cancer Genome Atlas Research Network et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061, 2008.

[2] International Cancer Genome Consortium et al. International network of cancer genome projects. *Nature*, 464(7291):993, 2010.

[3] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[4] David Tamborero, Abel Gonzalez-Perez, Christian Perez-Llamas, Jordi Deu-Pons, Cyriac Kandoth, Jüri Reimand, Michael S Lawrence, Gad Getz, Gary D Bader, Li Ding, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific reports*, 3:2650, 2013.

[5] Philip J Stephens, Patrick S Tarpey, Helen Davies, Peter Van Loo, Chris Greenman, David C Wedge, Serena Nik-Zainal, Sancha Martin, Ignacio Varela, Graham R Bignell, et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400, 2012.

[6] Xiguo Yuan, Junying Zhang, Shengli Zhang, Guoqiang Yu, and Yue Wang. Comparative analysis of methods for identifying recurrent copy number alterations in cancer. *PloS one*, 7(12):e52516, 2012.

[7] Félix Sanchez-Garcia, Patricia Villagrasa, Junji Matsui, Dylan Kotliar, Verónica Castro, Uri-David Akavia, Bo-Juen Chen, Laura Saucedo-Cuevas, Ruth Rodriguez Barrueco, David Llobet-Navas, et al. Integration of genomic data enables selective discovery of breast cancer drivers. *Cell*, 159(6):1461–1475, 2014.

[8] William G Kaelin. Use and abuse of rnai to study mammalian gene function. *Science*, 337(6093):421–422, 2012.

[9] Aviad Tsherniak, Francisca Vazquez, Phil G Montgomery, Barbara A Weir, Gregory Kryukov, Glenn S Cowley, Stanley Gill, William F Harrington, Sasha Pantel, John M Krill-Burger, et al. Defining a cancer dependency map. *Cell*, 170(3):564–576, 2017.

[10] Carlos López-Ón and Eleftherios P Diamandis. Breast and prostate cancer: an analysis of common epidemiological, genetic, and biochemical features. *Endocrine reviews*, 19(4):365–396, 1998.

[11] Broad Institute TCGA Genome Data Analysis Center. Analysis-ready standardized tcga data from broad gdac firehose 2016_01_28 run. 2016.

[12] Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology*, 12(4):R41, 2011.

[13] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

[14] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214, 2013.

[15] Chris Sherlock and Gareth Roberts. Optimal scaling of the random walk metropolis on elliptically symmetric unimodal targets. *Bernoulli*, pages 774–798, 2009.

[16] Min Huang, Aijun Shen, Jian Ding, and Meiyu Geng. Molecularly targeted cancer therapy: some lessons from the past decade. *Trends in pharmacological sciences*, 35(1):41–50, 2014.

[17] Kazuhiro Morishita, Diana S Parker, Michael L Mucenski, Nancy A Jenkins, Neal G Copeland, and James N Ihle. Retroviral activation of a novel gene encoding a zinc finger protein in il-3-dependent myeloid leukemia cell lines. *Cell*, 54(6):831–840, 1988.

[18] Daniel J White, Richard D Unwin, Eric Bindels, Andrew Pierce, Hsiang-Ying Teng, Joanne Muter, Brigit Greystoke, Tim D Somerville, John Griffiths, Simon Lovell, et al. Phosphorylation of the leukemic oncoprotein evi1 on serine 196 modulates dna binding, transcriptional repression and transforming ability. *PloS one*, 8(6):e66510, 2013.

[19] A Queisser, S Hagedorn, H Wang, T Schaefer, M Konantz, S Alavi, M Deng, W Vogel, A von Mässenhausen, G Kristiansen, et al. Ecotropic viral integration site 1, a novel oncogene in prostate cancer. *Oncogene*, 36(11):1573, 2017.

[20] J Wang, J Ou, Y Guo, T Dai, X Li, J Liu, M Xia, L Liu, and M He. Tblr1 is a novel prognostic marker and promotes epithelial–mesenchymal transition in cervical cancer. *British journal of cancer*, 111(1):112, 2014.

[21] Garrett Daniels, Xinmin Zhang, Xuelin Zhong, Larion Santiago, Ling Hang Wang, Xinyu Wu, Jack Y Zhang, Fengxia Liang, Xin Li, Thomas A Neubert, et al. Cytoplasmic, full length and novel cleaved

variant, tblr1 reduces apoptosis in prostate cancer under androgen deprivation. *Oncotarget*, 7(26):39556, 2016.

[22] Anup Sharma, Janet Mendonca, James Ying, Hea-Soo Kim, James E Verdone, Jelani C Zarif, Michael Carducci, Hans Hammers, Kenneth J Pienta, and Sushant Kachhap. The prostate metastasis suppressor gene ndrg1 differentially regulates cell motility and invasion. *Molecular oncology*, 11(6):655–669, 2017.