Integrating Mathematical Modeling with Machine Learning to Identify Cancer Driver Genes

Seo-Yeon (Karen) Chung

Irvine, California

Abstract

The identification of cancer driver genes is a critical component of precision oncology. Given the large feature space of The Cancer Genome Atlas (TCGA), which catalogs millions of somatic mutations observed in human tumors, machine learning techniques are ideally suited to driver gene identification. In existing models, however, the objective assessment of such machine learners is complicated by unexplained errors in the mutational data used to train the algorithms and by the absence of a perfect drivers list. This study employs mathematical modeling in tandem with machine learning processes to construct an objective and accurate classifier that identifies cancer driver genes. A set of in-silico mutational data is generated by the stochastic simulation of a differential equations model of feedback-controlled cancer population dynamics. The synthetic dataset, validated through the assessment of mutational pattern distributions, trains a selected machine learning algorithm, producing a driver gene classifier. The gene classifier prioritizes high-impact driver genes in four cancer types with ~85% accuracy and outperforms existing methods in recall. Additionally, it prioritizes genes that are heavily cited in association with carcinogenesis and pivotal gene pathways. Notably, top colorectal cancer driver genes from the classifier hold key roles in the PI3K-AKT and Wnt pathways, which have well-documented implications in carcinogenesis. The interdisciplinary methodology developed here produces an efficient and unbiased cancer driver gene classifier that can be utilized to identify henceforth unknown driver genes, providing insight essential for targeted cancer screening and treatment.

Introduction

Cancer is caused by the accumulation of driver mutations in cells. Cells acquire about one mutation in every 30 million base pairs during cell division; while most of these are harmless passenger mutations, mutations that induce the gain or loss of cellular function drive the formation of cancer and are therefore classified as *driver mutations* [1]. Subsequently, *driver genes* are genes with the potential to harbor driver mutations. Driver genes represent a minority in the mutations that a malignant tumor harbors: while there are a median of 33 to 66 mutations in most solid tumors, between 5 to 8 of them are drivers [2]. Driver genes are classified as either oncogenes (activation of cellular function) or tumor suppressor genes (loss of cancer prevention function) [3]. The identification of genes that drive tumorigenesis is an important first step in the development of targeted screenings and therapies [1].

While driver genes can be distinguished through assessment of functional impact in biological experiments and clinical studies, the impracticality of performing functional testing in a high-throughput manner necessiates the development of relevant bioinformatic methods [3]. Recent advances in Next Generation Sequencing have helped identify millions of somatic mutations in tumors from thousands of cancer patients, most notably in The Cancer Genome Atlas (TCGA) [4]. These have encouraged the use of computational and machine learning approaches in driver gene classification, intended to statistically prioritize genes to be functionally evaluated in an in-vivo experiment setting.

Machine learning-based driver gene prioritization methods identify frequency- or function-related mutation patterns observed in known driver genes but not in passenger genes. For example, the 20/20 rule [1], OncodriveClust [5], and TUSON [6] detect mutation frequency and clustering patterns, while ActiveDriver [7] predicts functional impact based on mutations' proximity to phosphorylation sites. Methods such as 20/20+ [8] extend these paradigms, integrating multiple features of positive selection.

However, the accurate identification of cancer driver genes versus the more frequent but incidental passenger genes nevertheless remains a challenging task; the nascent genomics-and-machine-learning paradigm currently do not address critical shortcomings. First, mutation datasets (e.g. TCGA) that are large, complex, and prone to unknown sequencing errors (1e-3 is considered the 'best' achievable rate) that confound the detection of low-frequency variants [9]. Similarly, The Cancer Gene Consensus (CGC) [10], a compilation of 478 generally accepted

cancer genes, is not completely experimentally verified, nor is it a comprehensive list. Existing methods rely on these two datasets in the machine learning process, rendering training inaccurate and evaluation unobjective.

This study generates an alternative training set ("synthetic dataset" hereafter) to train and validate a machine learning model. A simulation-generated synthetic genome's mutation data and predetermined list of driver genes bypasses the aforementioned problems of biological sequencing errors and incomprehensive training data (passenger vs. driver) labels. Furthermore, the use of synthetic data allows more in-vivo data to be reserved for testing instead of being partitioned. Similarly, training with an external dataset preempts data overfitting, guaranteeing the generalizability of classifier performance.

Goals

In developing a machine learning classifier for cancer driver genes identification, the study aims to achieve two specific functions: retrieval and prioritization. Those objectives are defined as follows.

1) The classifier is trained with synthetic data produced by mathematical modeling, allowing objective classification of biological data without input of a priori information.
2) The classifier, as indicated by robust statistical measures, effectively discriminates between driver and passenger genes, *retrieving* a useful list of putative driver genes.
3) The classifier, as indicated by biological analysis, effectively ranks genes according to driver-like characteristics, *prioritizing* driver genes with high functional impact.

All objectives were met in this study.

Methods

I. Mathematical model of the cancer environment

The study's first goal is to mathematically characterize the cancer environment, in which tumors evolve by acquiring a series of driver genes over time. The mathematical model considers two homogenous populations of cells: normal cells, N(t), and driver-mutated cancer cells, M(t). The kinetics of either population are expressed in *growth rate – death rate* form:

$$\frac{dN}{dt} = v_N N(t) L(t)\big((p_n - p_s) - d_N\big) - D(t)N(t)$$

$$\frac{dM}{dt} = (v_M - d_M)M(t)L(t) + v_N N(t)(2p_s + p_a)L(t)) - D(t)M(t)$$

The probability of mutation during cell division is denoted by $\mu = \mu_t + \mu_o$, where μ_t is the probability of a tumor suppressor gene mutation and μ_o is the probability of an oncogenic mutation. Then, the probability of one normal cell dividing into two normal cells (p_n), asymmetrically into one normal and one cancer cell (p_a), and two cancer cells (p_s), is expressed as the following [15]:

$$p_n = (1-\mu)^2 \qquad p_a = 2\mu(1-\mu) \qquad p_s = \mu^2$$

Addressing the importance of the "tumor microenvironment" in carcinogenesis [11], the model, via the logistic growth term L(t), directly relates cellular growth rates to the presence of resources necessary for growth . Considering the pivotal role of oxygen in normal and carcinogenic cellular processes [12], $O_2(t)$ is chosen to symbolize temporal concentration fluctuations of all such resources. For example, L(t) reaches its maximum of 1 when $O_2(t) = O_{2_{max}}$, thus maximizing $\frac{dN}{dt}, \frac{dM}{dt}$. Conversely, when $O_2(t)$ is low, the rate of necrosis, D(t), increases exponentially. $O_2(t)$ itself is formulated to "favor" tumor growth beyond a threshold; given $c_1 > c_2$, the denominator of $O_2(t)$ decreases with cancer cell population growth M(t), indicating a switch to more conservative consumption of resources. This is analogous to cancer cells' favoring of anaerobic glycolysis under the Warburg Effect [13], or to similar selective advantages that foster the evolution of cancer cell lineages [11].

$$L(t) = 1 - \frac{O_{2_{max}} - O_2(t)}{O_{2_{thresh}}} \qquad D(t) = \frac{D_{max}}{1 + \left(\frac{O_2(t)}{O_{2_{max}} - O_{2_{thresh}}}\right)^5} \qquad O_2(t) = \frac{v_o}{a_o + \big(M(t)+N(t)\big)\left(c_1 - \frac{c_1 - c_2}{1 + \left(\frac{5000}{M(t)}\right)^{10}}\right)}$$

Indeed, cancerous cell growth within the model is additionally influenced by feedback loops representing the "Hallmarks of Cancer"—proliferative signaling, evasion of growth suppressors, resistance of cell death, and replicative immortality [11]. The model regulates cancer cells' growth rate (v_M) with a positive feedback loop, where accumulations of oncogenic mutations upregulate cellular division. Similarly, cancer cells' apoptosis rate (d_M) is regulated with a negative feedback loop, with cells tumor-suppressing mutations repressing apoptosis. α_v, α_d are feedback strength parameters.

$$v_M = v_{M0} + \left(\frac{\alpha_v}{1 + \left(\frac{\mu_o}{\mu}\right)/M(t)} \right) \qquad d_M = d_{M0} - d_M \left(\frac{\alpha_d}{1 + \left(\frac{\mu_t}{\mu}\right)/M(t)} \right)$$

Ultimately, the mathematical model dictates the temporal evolution of wild-type and driver-mutated (cancerous) cells. When graphically expressed (Fig. 1), it satisfies the common definition of carcinogenesis as "cellular proliferation [following] successful adaptation to varying environmental constraints". Due to the absence of in-vivo experimental data that may motivate nonlinear best-fit optimization of parameters, model parameters are chosen randomly within acceptable ranges; for example, with human somatic mutation rates experimentally reported as ranging from 3.5e-9/bp/division to 1.6e-7/bp/division [14], μ is assigned a value of 3e-8.

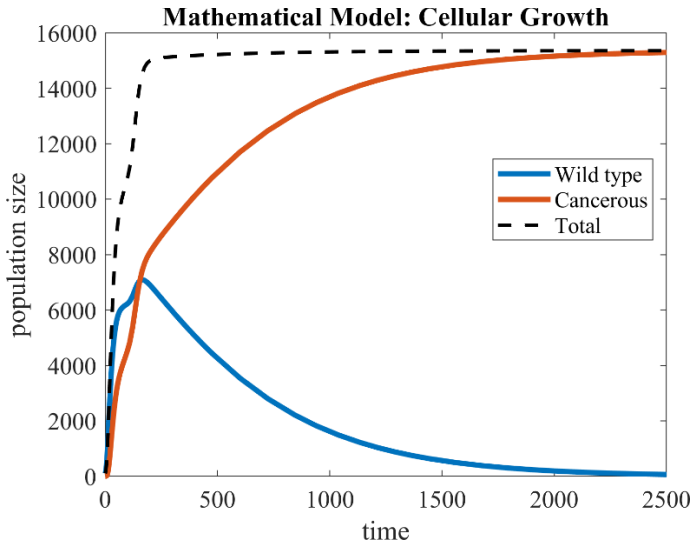


Figure 1. A mathematical model of cell population growth; tumor cells override wild-type cells over time

II. Synthetic Biological Data Generation

Thus far, the mathematical model has provided a deterministic representation of cellular population dynamics. However, the true purpose of the model is to guide the stochastic, in-silico simulation of carcinogenesis, from which the mutational profiles of each simulated cell is harvested. It is important to note that, when establishing the synthetic cells' genome, its driver genes are predetermined. Data harvested from the simulation can therefore capture insight to the differential mutation patterns between passenger and driver genes; this insight, then, is acquired by the machine learning process.

The aforementioned population-based interpretation of tumor proliferation is converted into an individual-cells-based interpretation via the use of the Gillespie Stochastic Simulation Algorithm. Within each small time interval ($t + \tau$, $t + \tau + dt$), Gillespie SSA associates with each reaction R_j, from a set of possible chemical reactions $R_1, \ldots, R_M$ a reaction propensity [15]. Thus, at each time interval of the cancer environment simulation, a cell may undergo none or one of its feasible reactions: division with or without mutation, or apoptosis (Fig. 2).

The mathematical model described in the previous section serves to define the reactions or to update their propensities (e.g. feedback loops that update a driver-gene-mutated cell's division rate) at each time step. The simulation instantiates and tracks cells down the lineage of 500,000 cells, each harboring 10,000 genes with unique nucleotide sequences with an average length of 100 base pairs. At the end of seventy-five simulated years, 500 randomly selected cells from each of the cancer and wild-type populations are post-processed. Their genomes are read, and sequences are recorded into a FASTQ. The somatic mutations are found and consolidated into a Mutation Annotation Format (MAF) file.

III. Feature computation

a. Feature definitions

A machine learning classifier requires the input of training data in the form of a feature vector. To quantify the mutational patterns of genes documented in the synthetic data, the raw, nonnumerical information from the MAF and FASTA file are converted into standard data arrays.

Then, thirteen mutational features are formulated and calculated for every gene. The features can be divided into five categories: (1) the frequency of a mutation type (silent, missense, and nonsense) in all documented mutations of the gene, (2) the ratio between occurrences of different mutation types (nonsense to missense, missense to silent, silent to nonsilent), (3) the

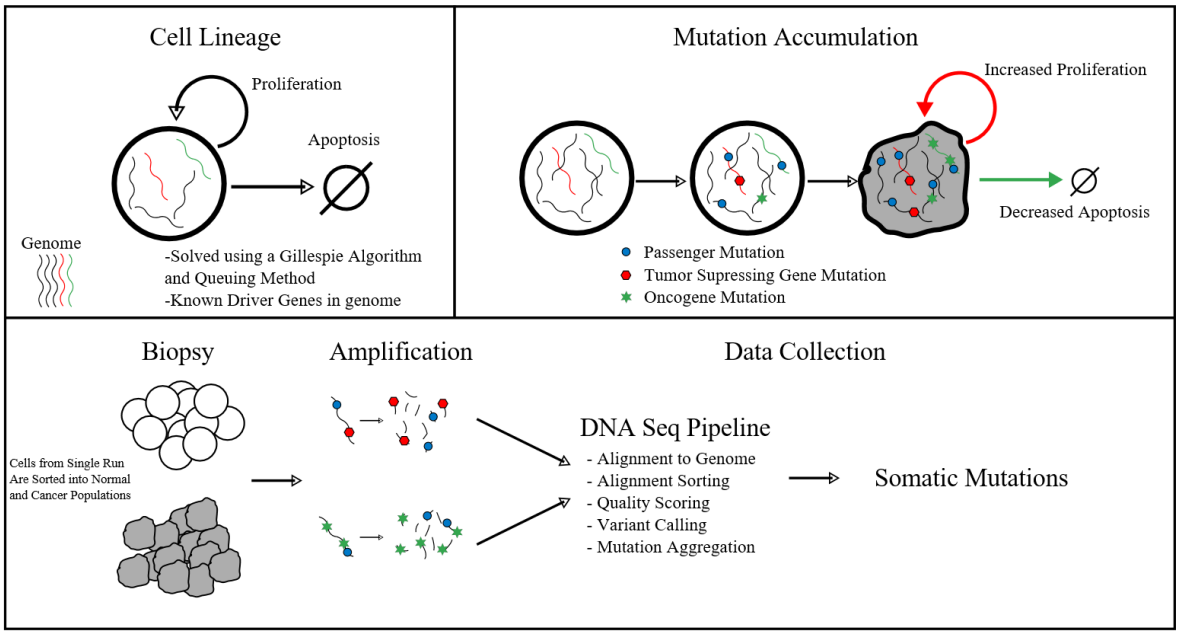


Figure 2. Workflow: generation of the in-silico, synthetic data set

clustering, or recurrence of mutations at hotspots (missense mutations, or encompassing all types), and (4) the statistical significance of harboring a mutation (missense, nonsense, or silent) at its particular location. The remaining two features are the number of samples the gene is found mutated and gene length, respectively (Fig. 3).

Mutational clustering is equivalent to the normalized Shannon's entropy value of $\Sigma_i p_i log_2 p_i / log_2 k$, where k = # of missense or all types of mutations, and p_i = # of missense or any mutations in the gene's ith codon. Driver genes are expected to have higher occurrences of hotspots, thus lower entropy values, than passenger genes.

The significance of a specific type of mutation's specific genetic location is measured by computing its "default probability", or probability of occurrence by chance. For this, Monte Carlo simulations are used to permute the dataset. Specifically, for each gene, each mutation is shuffled to a random spot in the gene with the same nucleotide as in the observed spot (the total mutation count remained fixed). Then, with a rereading of the gene's codon, each relocated mutation is labeled as either a missense, nonsense, or silent mutation. When eight such permuted datasets are created, the (missense, nonsense, silent) default probability is calculated as the fraction of simulated datasets with (missense, nonsense, silent) mutation counts greater than or equal to that of the original dataset. Again, driver genes are expected to be more statistically significant (lower default probability) on average.

b. Biological data retrieval

This study's goal is to identify actual human driver genes using a machine learning classifier that distinguishes between mutational patterns. Cancer genomic mutations data, consisting of 729,205 small somatic variants encompassing 7,916 distinct samples across 34 cancer types, is retrieved from a 2016 study by Tokheim et al. [16]. The data merges results from multiple published genome sequencing studies as well as from The Cancer Genome Atlas. To contrast with the in-silico methodology of synthetic biological data production, this merged biological data will be referred to as the "In-vivo Dataset". Because the synthetic biological data contains only missense, nonsense, and silent mutations, mutation entries in the In-vivo Dataset that are not of these types are discarded. Feature vector generation process is then repeated on the In-vivo Dataset.

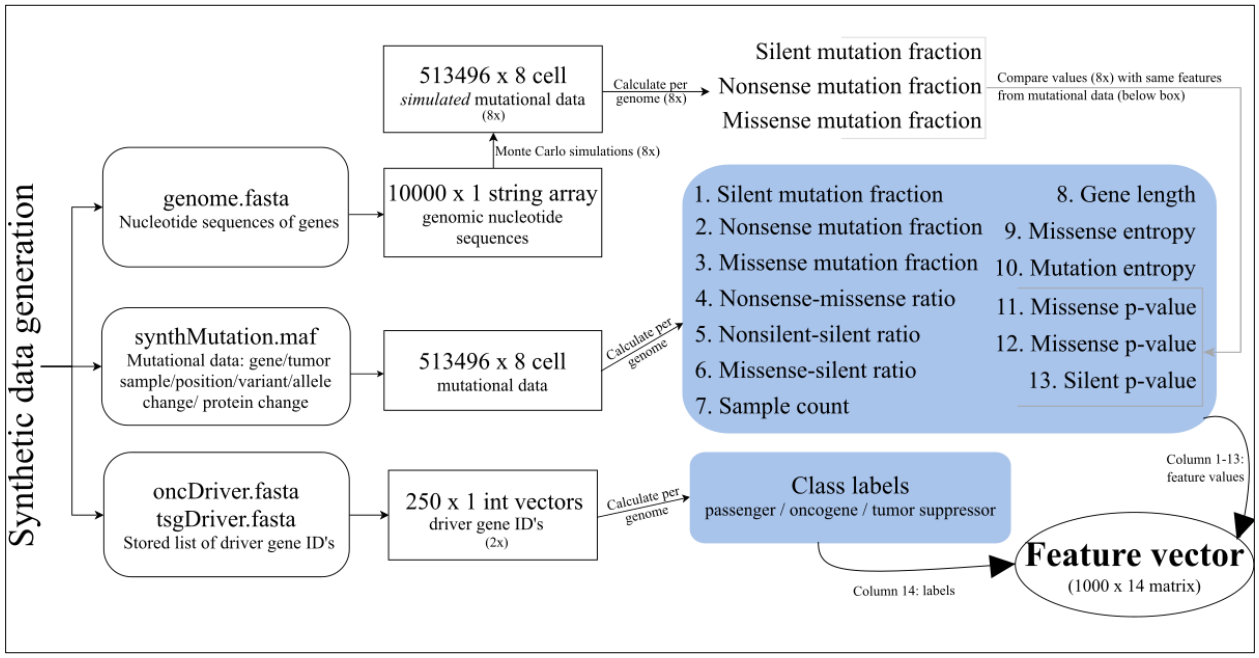


Figure 3. Processing of the synthetic data into a feature vector, the input to machine learning algorithms.

Results

I. Comparison of feature value distributions across datasets

Before the feature vector derived from synthetic data is used to train a machine learning algorithm, the similarity of our mathematical model-derived synthetic dataset to the TCGA dataset is verified. The distributions of feature values derived from the two datasets are qualitatively compared. For all four features categories, distributions appear generally similar in shape, center, and spread (Fig. 4). Thus, it is reasonable to proceed with training a driver gene classifier using synthetic data. Of course, it is difficult to quantify the accuracy of the synthetic data in capturing the differential *combinations* of feature values in driver genes versus passenger genes at this stage. The performance of the machine learning classifier, trained with synthetic data, in identifying TCGA driver genes may be interpreted as a more meaningful measure of the synthetic data's "realness".

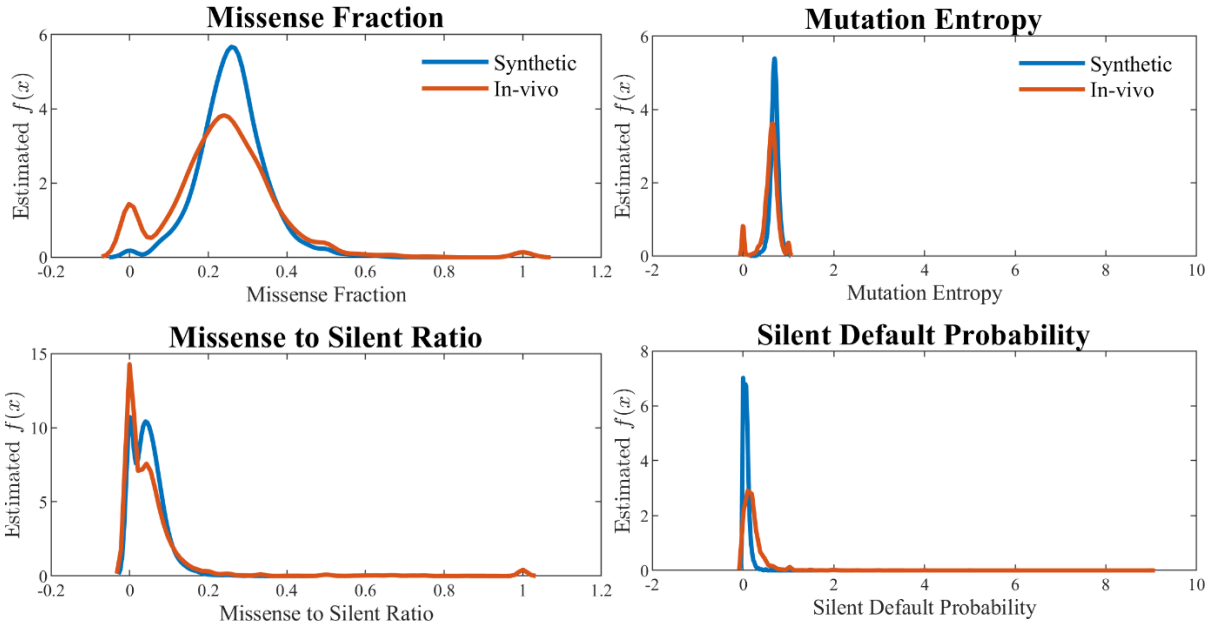


Figure 4. Comparison of feature value distributions across datasets. Each plot illustrates one gene feature category. (A): mutations type; (B): mutations type ratios; (C): mutations clustering; (D) mutation significance.

II. Machine Learning on Synthetic Data

As aforementioned in the Introduction section, this study defines the task of cancer driver genes identification as *retrieval* and *prioritization*. These objectives guide the selection of statistical and biological analysis measures in training and testing the model.

a. Training and validation with synthetic data: boosted ensemble algorithm outperforms other methods

The feature vector derived from synthetic data is partitioned into a training set (85%) and a validation set (15%) for hold-out validation.

The validation results of five machine learning algorithms are compared: logistic regression, decision tree, cubic polynomial SVM, random forest, and RUSBoosted random forest. Note that two variants of random forest, an ensemble algorithm that aggregates results from multiple decision trees, are selected for evaluation. Given that overfitting to synthetic data would prove fatal to the classifier's performance on external datasets, ensemble methods' ability to control variance [17] is useful. Furthermore, the RUSBoost meta-algorithm, which reduces bias through an iterative sampling scheme [18], is applied to account for the data's class imbalance between passenger and drivers.

Two retrieval-related statistics are reported from 15% hold-out validation on the five methods (Table 1). One is recall, quantifying the proportion of driver genes in the validation set that is correctly identified by the classifier. High recall does not alone indicate a good classifier, however. The hypergeometric test indicates the significance of the recall value in the context of possible chance selection. Then, among the four algorithms (excluding logistic regression, which predicted every validation set gene as a driver) testing as significant at $p < 0.01$, Random forest and RUSBoosted random forest perform with the highest recall.

	Logistic Reg.	**Decision Tree**	**Cubic SVM**	**Random Forest**	**RUSBoosted RF**
Recall	1.000	0.4533	0.4400	0.4800	**0.6795**
p-value	1.000	5.585e-10	1.695e-11	3.475e-17	5.761e-9

Table 1. Validation performance of five classification algorithms.

To evaluate the strength of classifier prioritization, gene ranks are examined. All algorithms tested attribute numeric scores to each observation's likelihood of classified as either a passenger or driver, allowing for genes to be ranked by those measures. Then, the cumulative distribution function (CDF) of the rank of true driver genes, i.e., the number of predetermined synthetic-data drivers that are ranked in the top *k* genes of the list as a function of *k*, is plotted for each method (Fig. 5a). Note that there are 1500 genes in the validation set, 75 of which are predetermined synthetic-data drivers. RUSBoosted random forest outperforms other methods in driver-gene prioritization, consistently recovering drivers at a better rank. Zooming into the CDF function, 30% of all driver genes are tied at the top rank by RUSBoosted random forest (Fig. 5b).

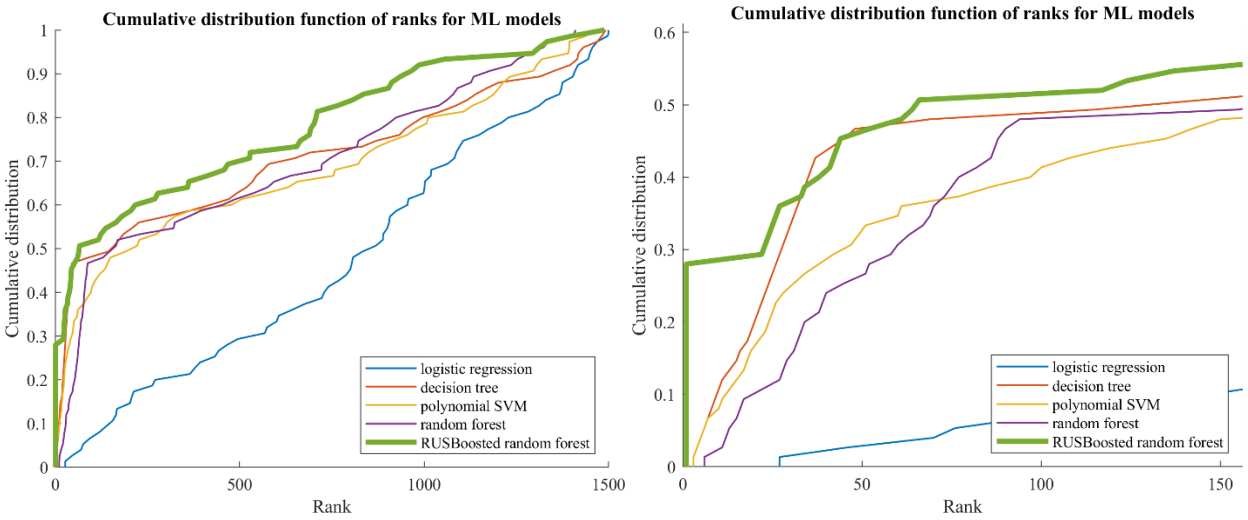


Figure 5a (left panel) and 5b (right panel). Cumulative distribution function of the rank for local methods, in the validation stage. Fig. 5a: Global curve, Fig. 5b: Zoom on the beginning of the curve.

II. Driver gene identification on the In-vivo Dataset

The trained RUSBoosted random forest classifier is now a *driver gene classifier*. The true objective of this classifier, to identify driver genes in biological data, is evaluated by using the feature vector derived from the In-vivo Dataset as testing data. With the rich biological insights affiliated with the dataset's genes, criteria for *prioritization* can extend beyond the driver-or-passenger labeling of top-ranked genes as in the preceding section.

a. Retrieval: Synthetic-data-trained classifier outperforms other driver gene prediction methods

Driver gene retrieval can be benchmarked by overlap with the Cancer Gene Census (CGC) [10], a manually curated list of likely but not necessarily validated drivers. Of 19319 genes in the In-vivo Dataset, CGC labels 478 as likely drivers. Considering CGC as a hypothetical "true class label", two retrieval-related statistics of recall and p-value are again calculated (Table 2).

	Accuracy	**Recall (CGC enrichment)**	**p-value**
Value	0.8490	0.1715 (82 CGC genes retrieved)	1.197e-5

Table 2. Validation performance of five classification algorithms.

Because the above recall statistics is necessarily an underestimate, it is helpful to additionally benchmark the above reported performances through comparison with previous studies' results. Tokheim et al. reports the driver score predictions by existing methods—including TUSON [6], OncodriveClust [5], ActiveDriver [7], and the 20/20 Rule [1]—for all genes within the In-vivo Dataset.

Then, the CGC enrichment of top ranked genes for each method is compared (Fig. 6). This study's driver gene classifier outperforms all other methods at driver gene retrieval when the top 10 (80% enrichment), 30 (37%), and 50 (32%) genes are considered. Note that this performance comparison understates the relative retrieval power of the classifier against other methods. The In-vivo Data merges mutation data, either from independent sequencing experiments, TCGA, or COSMIC, that were used to train TUSON [6], OncodriveClust [5], and ActiveDriver [7]; therefore, these methods by default can recognize a big subset of CGC-predicted genes as drivers. Meanwhile, this study's classifier had no *a priori* insight on the In-vivo Dataset, a feature indicating the generalizability of its stronger performance.

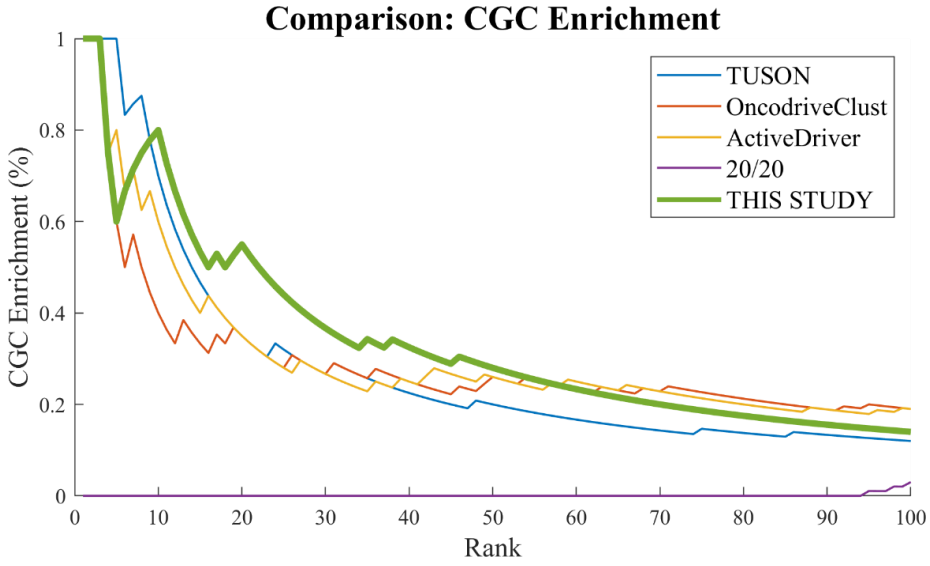


Figure 6. Comparison of classifier with previous methods on top-ranked genes' CGC enrichment.

The enrichment analysis also suggests that the classifier's predictive power decreases down its rank, confirmed by the classifier's low overall AUROC of 0.55. This phenomena helps narrow down the list of classifier-predicted driver genes that warrant functional analysis in and out of this study's scope. Based on the genes ranked in the top 1000 (top 5% of ~20000 genes), the classifier yields an AUROC of 0.71, indicating a more efficient ranking (i.e. likelihood that rank of randomly selected driver gene higher than randomly selected passenger [19]) of genes.

b. Prioritization: Top putative drivers play key roles in carcinogenesis.

Therefore, the focus in evaluating the classifier's ability to *prioritize* highly plausible driver genes is on top ranked (~top 100) putative drivers. The analysis is cancer-type specific. From the raw In-vivo Dataset, which encompasses 34 cancer types, four type-specific datasets

are extracted: colorectal, pancreatic, breast, and melanoma. Each raw data is converted into a feature vector, from which the classifier yields predicted driver genes.

The biological significance of top genes is validated by referencing NCBI through Gene-Valorization [20], an automatic bibliography search tool. Across four cancer types, the number of publication hits in NCBI relevant to a query cancer type and a query gene are in the hundreds or thousands for topmost genes (Fig. 6).

Further insight is mined through biological pathway analysis, using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis tool [21]. Putative driver genes across the four cancer types are active in pathways involved in cell development and apoptosis, including PI3K-AKT and Wnt/β-Catenin. (Full pathway included in the Appendix.)

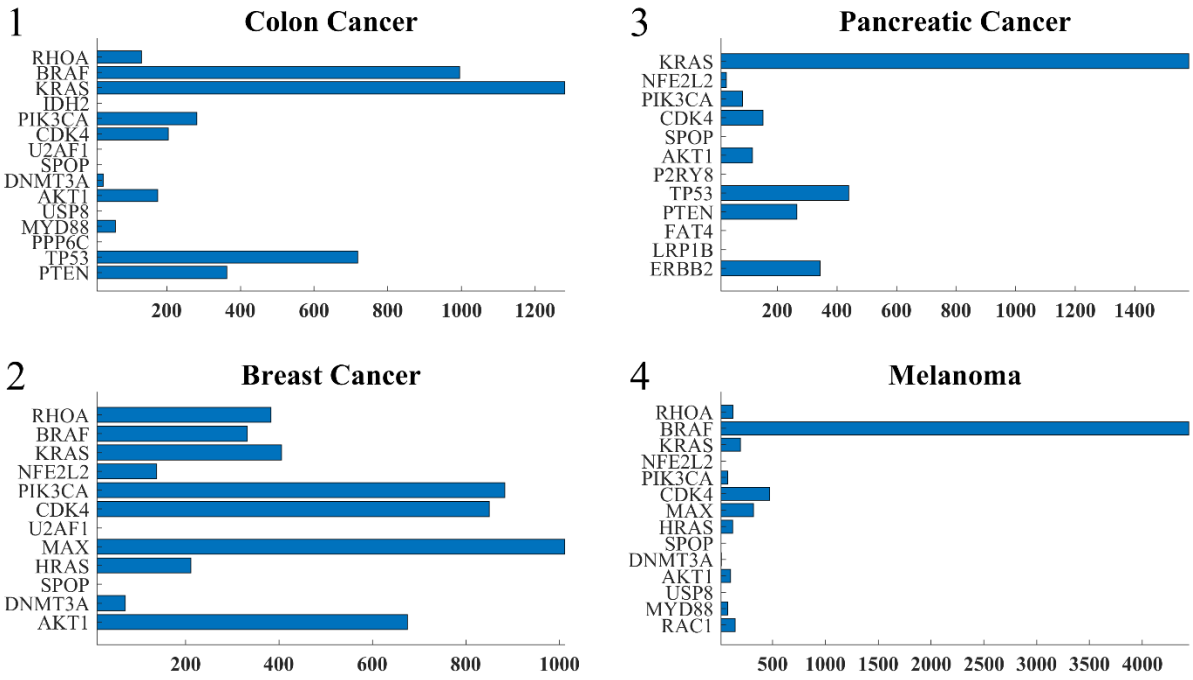


Figure 6. NCBI publication hits of top predicted genes across four cancer types.

Phosphatidyl-inositol 3-kinase (PI3K) activity is associated with the transformation of viral oncoproteins; PI3K (encoded by PIK3CA, etc.) is central in several cellular processes, such as proliferation, growth, apoptosis, and cytoskeletal rearrangement [22]. The primary consequence of PI3K activation is the generation of the second messenger PIP_3, which activates downstream pathways that involve proteins involving AKT and others. The phosphatase PTEN terminates the PI3K-AKT signaling pathway by dephosphorylating PIP_3 to PIP_2. Mutations on genes in the pathway — PIK3CA, AKT1, and PTEN — lead to aberrant pathway activation, and

often, carcinogenesis [23]. The classifier lists of top ten driver genes for each of the four cancer types include at least two of these three PI3K pathway genes.

Another signaling pathway associated with the classifier's putative driver genes is the Wnt pathway. Signaling by the Wnt proteins via the transcription co-activator β-catenin controls embryonic development and adult homeostasis [24]. β-catenin (encoded by CTNNB1) is regulated and destroyed by the beta-catenin destruction complex, and in particular by the adenomatous polyposis coli (APC) protein, encoded by the tumor-suppressing APC gene. The loss of negative Wnt signaling regulation by APC is responsible for around 15% of colorectal cancer cases [25].

A comprehensive signaling scheme of colorectal cancer carcinogenesis can be constructed based upon the list of predicted genes active in either the Wnt or PI3K pathway. Figure 7 is retrieved from KEGG as is a partial summary of the pathway. Products of the KRAS (driver score rank #3), PIK3CA (rank #5), AKT1 (rank #10), APC (rank #25), and CTNNB1 (rank #43) genes are depicted (Appendix).

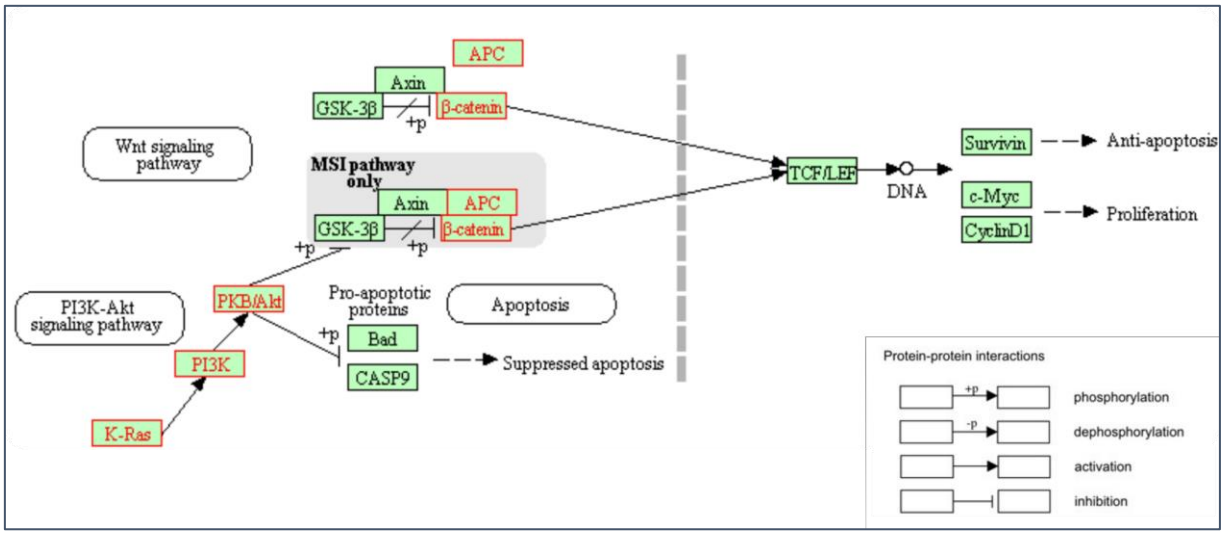


Figure 7. Interaction pathway of select colorectal cancer driver genes identified by classifier. The proteins typed in red are affiliated with genes identified correctly as a driver.

Discussion and Conclusion

This research introduces a new approach to harnessing machine learning to statistically discriminate between likely passenger and driver genes in cancer. A deterministic mathematical model of the cancer environment motivated a stochastic simulation of carcinogenesis, yielding a

synthetic dataset that compiles simulated cells' mutational profiles, formatted in parallel with standard biological data formats (MAF, FASTQ). Then, mutational patterns of every synthetic gene were extracted, forming a feature vector that trains and validates the driver gene classifier. The algorithm of choice was RUSBoosted Random Forest, for its evident immunity to high bias and variability in dealing with inherently noisy and skewed biological data.

The driver gene classifier achieves ~85% accuracy. It retrieves 82 genes listed in the Cancer Genome Census, a statistically significant ($p < 0.02$) level of enrichment that also outperforms existing state-of-the-art machine learning approaches at the top 100 genes level [1], [5]–[7]. This top-gene prioritization scheme gives way to useful biological insight: in addition to being heavily cited in association with carcinogenesis, top genes are strongly implicated in pivotal gene pathways (PI3K-AKT, Wnt) that directly regulate the cell cycle and other key metabolic processes.

This study's main limitation originates from the difficulty of quantifying the validity of model-generated synthetic data. The current, qualitative validation scheme in the Results section may be improved by executing multiple simulations with multiple biological parameters, and then picking the best combination of parameters that yields feature vectors with the lowest Kullback-Leibler divergence when compared with in-vivo-data-derived feature vectors.

Nevertheless, the classifier's performance results suggest the usefulness of integrating mathematical modeling with machine learning in bioinformatics. The use of synthetic data as training data is beneficial because it (1) allows for entire in-vivo datasets to be tested without partitioning into training-validating sets, (2) eliminates the uncertainty of training a machine learning model with data containing unknown sequencing errors, and (3) bypasses the inaccuracy of training a model with in-vivo data containing undiscovered driver genes falsely labeled as passenger genes.

Identifying the genetic basis of human diseases is central in predicting and diagnosing disease onset, characterizing evolving disease phenotypes, and most importantly, developing cures. The cancer driver genes and pathways highlighted in this study represent potential subjects of further investigation. The methodology developed in this study is broadly applicable to the investigation of other gene-driven diseases.

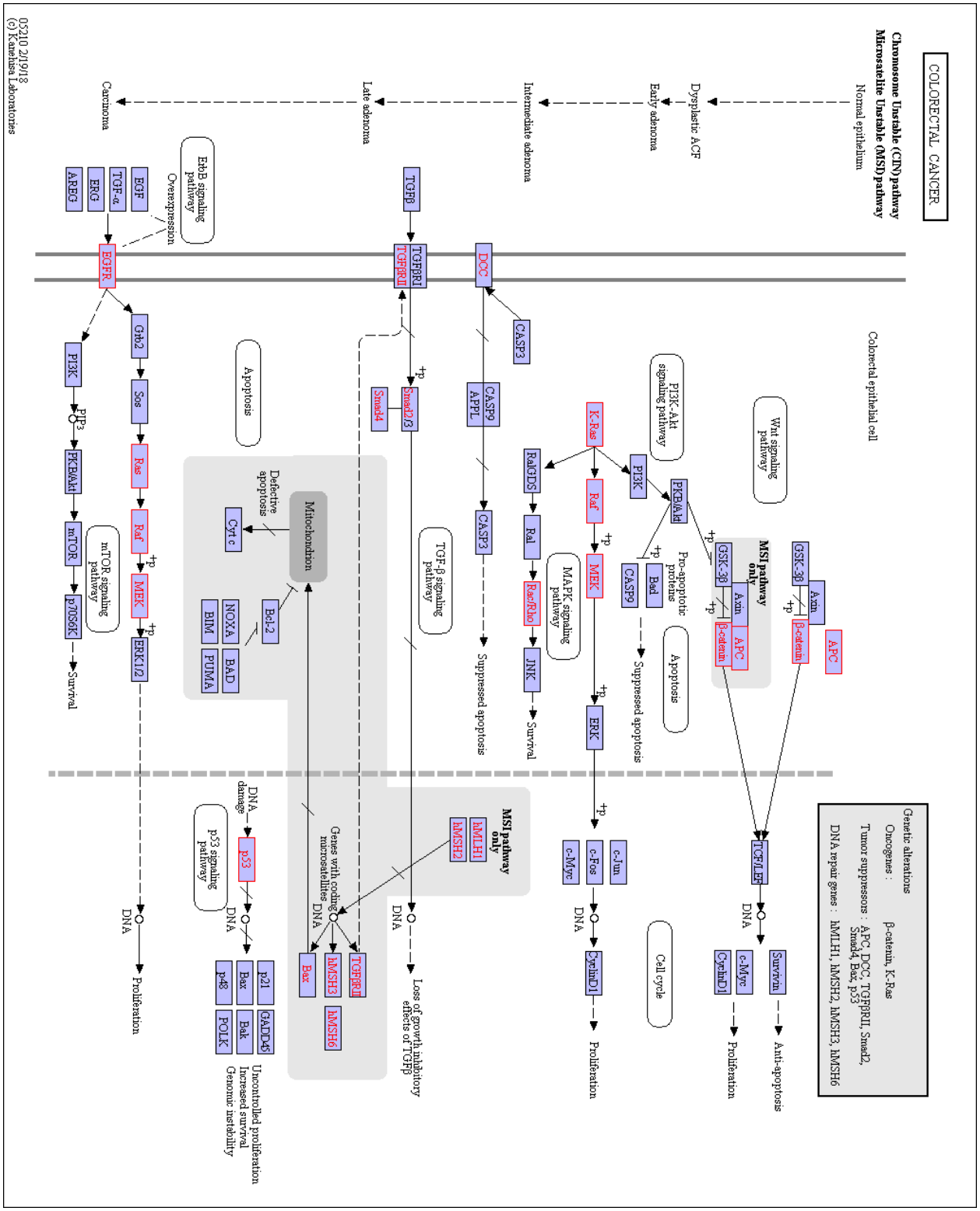
Appendix

I. CGC-enriched classifier-predicted driver genes, ranked.

#	Colorectal	Pancreatic	Breast	Melanoma
1	RHOA	KRAS	RHOA	RHOA
2	BRAF	NFE2L2	BRAF	BRAF
3	KRAS	PIK3CA	KRAS	KRAS
4	IDH2	CDK4	NFE2L2	NFE2L2
5	PIK3CA	SPOP	PIK3CA	PIK3CA
6	CDK4	AKT1	CDK4	CDK4
7	U2AF1	P2RY8	U2AF1	MAX
8	SPOP	TP53	MAX	HRAS
9	DNMT3A	PTEN	HRAS	SPOP
10	AKT1	FAT4	SPOP	DNMT3A
11	USP8	LRP1B	DNMT3A	AKT1
12	MYD88	ERBB2	AKT1	USP8
13	PPP6C	PTPRT	USP8	MYD88
14	TP53	EP300	MYD88	RAC1
15	PTEN	NCOR1	P2RY8	P2RY8
16	FAT4	ARID2	TP53	PPP6C
17	MAP2K1	APC	PTEN	TP53
18	LRP1B	ARID1A	NPM1	PTEN
19	ERBB2	CUX1	FAT4	NPM1
20	B2M	FAT1	MAP2K1	FAT4
21	PTPRT	KMT2C	LRP1B	MAP2K1
22	EP300	KMT2D	ERBB2	LRP1B
23	NCOR1	IDH1	B2M	ERBB2
24	ARID2	EBF1	PTPRT	B2M
25	APC	PTCH1	HIST1H3B	PTPRT
26	ARID1A	MAPK1	EP300	HIST1H3B
27	CUX1	XPO1	NCOR1	EP300
28	PPP2R1A	FGFR2	ARID2	NCOR1
29	FAT1	AR	APC	ARID2
30	KMT2C	FOXA1	ARID1A	APC
31	KMT2D	MYCN	CUX1	ARID1A
32	IDH1	CTNNB1	PPP2R1A	CUX1
33	EBF1	KDM6A	FAT1	PPP2R1A
34	EGFR	POLE	KMT2C	FAT1
35	PTCH1	ATRX	KMT2D	KMT2C
36	MAPK1	PTPRC	IDH1	KMT2D
37	CD79B	RAF1	EBF1	IDH1
38	NR4A3	JUN	EGFR	EBF1
39	FGFR2	RB1	PTCH1	EGFR
40	NRAS	SETD2	MAPK1	PTCH1
41	AR	SETBP1	CD79B	MAPK1
42	FOXA1	IL7R	NR4A3	CD79B
43	CTNNB1	WHSC1	XPO1	VHL
44	KDM6A	BRCA2	FGFR2	NR4A3
45	POLE	PAX3	CDK6	XPO1
46	ATRX	KEAP1	NRAS	FGFR2

47	PTPRC	DDX3X	AR	CDK6
48	GNA11	CCNB1IP1	FOXA1	NRAS
49	RAF1	ATR	MYCN	AR
50	TLX3	SLC34A2	CTNNB1	FOXA1
51	CDH1		KDM6A	MYCN
52	RB1		POLE	CTNNB1
53	SETD2		ATRX	KDM6A
54	SETBP1		PTPRC	NKX2-1
55	IL7R		RAF1	POLE
56	WHSC1		JUN	ATRX
57	BRCA2		CDH1	PTPRC
58	PAX3		RB1	GNA11
59	KEAP1		SETD2	RAF1
60	BCORL1		SETBP1	TLX3
61	PDGFRB		IL7R	JUN
62	TFEB		WHSC1	CDH1
63	ATR		BRCA2	RB1
64	SLC34A2		PAX3	SETD2
65	ACVR1		KEAP1	SETBP1
66			BCORL1	IL7R
67			PDGFRB	WHSC1
68			DDX3X	BRCA2
69			CCNB1IP1	PAX3
70			TFEB	KEAP1
71			ATR	BCORL1
72			TAL1	PDGFRB
73			SLC34A2	DDX3X
74			ACVR1	CCNB1IP1
75				TFEB
76				ATR
77				TAL1
78				SLC34A2
79				ACVR1

II. Full colorectal cancer gene pathway scheme, retrieved from KEGG.



Works Cited

[1] M. H. Bailey *et al.*, “Comprehensive Characterization of Cancer Driver Genes and Mutations,” *Cell*, vol. 173, no. 2, pp. 371-385.e18, 2018.

[2] B. Vogelstein, N. Papadopoulos, and V. E. Velculescu, “S1D_Cancer Genome Landscapes,” *Science (80-.).*, no. March, 2013.

[3] J. R. Pon and M. A. Marra, “Driver and Passenger Mutations in Cancer,” *Annu. Rev. Pathol. Mech. Dis.*, vol. 10, no. 1, pp. 25–50, 2015.

[4] J. N. Weinstein *et al.*, “The Cancer Genome Atlas Pan-Cancer Analysis Project,” *Nat Genet. 2013 Oct. ; 45(10)*, vol. 45, no. 10, pp. 1113–1120, 2014.

[5] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, “OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes,” *Bioinformatics*, vol. 29, no. 18, pp. 2238–2244, 2013.

[6] T. Davoli *et al.*, “NIH Public Access Aneuploidy Patterns to Shape the Cancer Genome,” vol. 155, no. 4, pp. 948–962, 2014.

[7] J. Reimand, O. Wagih, and G. D. Bader, “The mutational landscape of phosphorylation signaling in cancer,” *Sci. Rep.*, vol. 3, 2013.

[8] C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, and R. Karchin, “Evaluating the evaluation of cancer driver genes,” *Proc. Natl. Acad. Sci.*, vol. 113, no. 50, pp. 14330–14335, 2016.

[9] X. Ma *et al.*, “Analysis of error profiles in deep next-generation sequencing data,” *Genome Biol.*, vol. 20, no. 1, pp. 1–15, 2019.

[10] P. A. Futreal *et al.*, “A CENSUS OF HUMAN CANCER GENES,” vol. 4, no. 3, pp. 177–183, 2009.

[11] D. Hanahan and R. A. Weinberg, “Hallmarks of Cancer: The Next Generation,” *Cell*, vol. 100, 2000.

[12] P. P. Hsu and D. M. Sabatini, “Cancer cell metabolism: Warburg and beyond,” *Cell*, vol. 134, no. 5, pp. 703–707, 2008.

[13] M. Vander Heiden, L. Cantley, and C. Thompson, “Understanding the Warburg effect: The metabolic Requiremetns of cell proliferation,” *Science (80-.).*, vol. 324, no. 5930, pp. 1029–1033,

2009.

[14] B. Werner and A. Sottoriva, “Variation of mutational burden in healthy human tissues suggests non-random strand segregation and allows measuring somatic mutation rates,” *PLoS Comput. Biol.*, vol. 14, no. 6, pp. 1–12, 2018.

[15] D. T. Gillespie, “Stochastic Simulation of Chemical Kinetics,” *Annu. Rev. Phys. Chem.*, vol. 58, no. 1, pp. 35–55, 2007.

[16] C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, and R. Karchin, “Evaluating the evaluation of cancer driver genes,” *Proc. Natl. Acad. Sci.*, vol. 113, no. 50, pp. 14330–14335, 2016.

[17] A. L. and M. Wiener, “Classification and Regression by randomForest. R News 2,” vol. 3, no. December 2002, pp. 18–22, 2003.

[18] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, “RUSBoost : A Hybrid Approach to Alleviating Class Imbalance,” vol. 40, no. 1, pp. 185–197, 2010.

[19] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[20] B. Brancotte, A. Biton, I. Bernard-Pierrot, F. Radvanyi, F. Reyal, and S. Cohen-Boulakia, “Gene list significance at-a-glance with GeneValorization,” *Bioinformatics*, vol. 27, no. 8, pp. 1187–1189, 2011.

[21] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, “KEGG: Kyoto encyclopedia of genes and genomes,” *Nucleic Acids Res.*, vol. 27, no. 1, pp. 29–34, 1999.

[22] K. D. Courtney, R. B. Corcoran, and J. A. Engelman, “The PI3K pathway as drug target in human cancer,” *J. Clin. Oncol.*, vol. 28, no. 6, pp. 1075–1083, 2010.

[23] I. Vivanco and C. L. Sawyers, “The phosphatidylinositol 3-kinase-AKT pathway in humancancer,” *Nat. Rev. Cancer*, vol. 2, no. 7, pp. 489–501, 2002.

[24] B. MacDonald, K. Tamai, and X. He, “Wnt/β-catenin signaling: components, mechanisms, and diseases,” *Dev Cell.*, vol. 17, no. 1, pp. 9–26, 2010.

[25] T. Zhan, N. Rindtorff, and M. Boutros, “Wnt signaling in cancer,” *Nat. Publ. Gr.*, vol. 36, no. 11, pp. 1461–1473, 2016.