Computational prediction of the pathogenic status of cancer-specific somatic variants

By:

Nikta Feizi

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Biochemistry and Medical Genetics

University of Manitoba

Winnipeg, Manitoba, Canada

Abstract

The ever-increasing shift in cancer genetic testing is accompanied by a demand for interpretation of the results. In-silico interpretation approaches are shown to be promising in increasing the clinical utilization of genetic tests by classifying variants into well-defined pathogenic and non-pathogenic clinical groups. Current prediction tools are mostly trained based on the characteristics of germ-line variants and utilized for classifying both gremline and somatic variants, which may result in biased predictions in the event of classifying somatic variants. Considering the critical role of somatic variants in cancer occurrence and progression establishing cancer specific prediction tools which are designed solely based on the characteristics of somatic variants is of high importance.

To this aim, we established a gold standard dataset exclusively for cancer somatic single nucleotide variants (SNVs) collected from the catalogue of somatic mutations in cancer (COSMIC). To label the pathogenic (positive) variants we introduced a bi-dimensional recurrence score spanning both the frequency of a given mutation across a dataset and the number of cancer types the mutation affects. To label the non-pathogenic (negative) variants we collected the somatic SNVs with the minor allele frequency of equal or greater than 1% in at least one of 26 populations from 1000 Genomes project that are also defined in COSMIC. We portrayed the genomic characteristics of the somatic variants located in coding and non-coding regions of the genome by defining 80 and 65 different genomic features in two distinct gold standards, respectively. Using a support vector machine (SVM) classification method we designed two distinct models achieving the AUC (area under the ROC curve which is a measure of classification capability of a given model) of 0.94 and 0.89 in classifying the pathogenic status of somatic variants located in coding and non-coding regions of the genome, respectively. We compared the

performance of our models with two well-known classification tools including FATHMM-FX and CScape which are not originally designed for classifying cancer somatic variants. Our models outperformed both tools in classifying variants located in both coding and non-coding regions of the genome. Furthermore, we applied our models to predict the pathogenic status of somatic variants identified in young patients (under 45 years of age) with breast cancer from METABRIC and TCGA-BRCA studies. The results indicated that using the classification threshold of 0.8 our "coding" model predicted 1853 positive SNVs (out of 6910) from TCGA-BRCA dataset, and 500 positive SNVs (out of 1882) from METABRIC dataset. Interestingly, through comparative survival analysis of the positive predictions from our models, we identified a young-specific pathogenic somatic variant potential for the prognosis of early onset of breast cancer in young women.

In conclusion, the computational models designed originally based on the characteristics of cancer somatic variants are proven to have high discriminative power for classifying the variants into pathogenic and non-pathogenic groups. In addition, the potential application of the models in revealing the functional and clinical impacts of cancer somatic variants is strongly suggested through our study.

Acknowledgement

I would like to express my sincere thanks to my supervisor Dr.Pingzhao Hu. He gave me the opportunity to work on a project that provided me with the educations I needed to follow my future goals. I am cordially thankful for all his supports and guidance throughout the entire time I spent in the University of Manitoba.

It is an honor for me to express my deep gratitude to my supervisory committee members Dr. Leigh Murphy and Dr. Jean-Eric Ghia for their encouragements, constructive comments and insightful suggestions which kept me on track during my research.

I would like to thank all my colleagues specially Mehrdad Hossein Zadeh and Svetlana Frankel, as well as all my friends for their friendships, helps and supports.

I would like to acknowledge the faculty and staff members in the Department of Biochemistry and Medical Genetics specially Dr. Jeffrey Wigle who chaired all my committee meetings.

Finally, I would like to dedicate this thesis to my parents, whose constant encouragement, understanding and love have helped me through the ups and downs, and have contributed immeasurably to everything I have achieved.

Table of contents

List of Tables

List of Figures

Table of abbreviations

Abbreviation	Description
DNA	Deoxyribonucleic acid
SNV	Single Nucleotide variant
SNP	Single nucleotide polymorphism
WGS	whole genome sequencing
GWAS	Genome-wide association studies
ACMG	American College of Medical Genetics and Genomics
MSA	Multiple alignment sequence
SIFT	Sorting Intolerant From Tolerant
POLYOHEN2	Polymorphism Phenotyping version2
CADD	Combined Annotation–Dependent Depletion
VEP	Variant Effect Predictor
HGMD	Human Gene Mutation Database
COSMIC	Catalogue of somatic mutations in cancer
SVM	Support vector machine
LASSO	Least absolute shrinkage and selection operator
CGP	Cancer Genome Project
DIP	Deletion insertion polymorphisms
STR	Short tandem repeats
GERP	Genomic Evolutionary Rate Profiling
MICE	Multivariate imputation by chained equations
ROC	Receiver operating characteristic
AUC	Area under the curve
CV	Cross validation
FPR	False positive rate
TPR	TPR
METEBRIC	Molecular Taxonomy of Breast Cancer International Consortium
TCGA	The Cancer Genome Atlas
ncRNA	Noncoding RNA
ER	Estrogen receptor
DFS	Disease free survival
OS	Overall survival
DSS	Disease specific survival
GSEA	Gene set enrichment analysis

Chapter 1 : Background and Introduction

1.1. Introduction to genetic variants

Permanent changes in the nucleotide sequence are known as mutations (Richards et al., 2015) occurring in both normal and neoplastic proliferating cells. Errors during DNA replication as well as exposure to endogenous or exogenous mutagens are addressed as the major etiology of mutations (Greenman et al., 2007).

Depending on whether the changes are happened in the genome of somatic or germinal cells, they are referred to as somatic mutations and germline mutations, respectively (Griffiths AJF, 2000). Germinal mutations only influence the progenitors of germ line cells which are sex cells passing on to next generation. Individuals of perfectly normal phenotype can harbor abnormal mutations in their sex cells which can result in susceptibility to complex diseases such as cancer (J. Thusberg, A. Olatubosun, & M. Vihinen, 2011). Somatic mutations also can arise susceptibility to cancer by being transmitted asexually to the descendants of the mutated cells. The potential phenotypic outcome of a somatic mutation will only be observable in the carrier individuals (Griffiths AJF, 2000).

Regardless of the tissue origin, if the population frequency of a nucleotide change is above 1%, it is known as a polymorphism (Richards et al., 2015). Single nucleotide polymorphism (SNP) is the most represented variation in the human genome, many of which underlie the phenotypic differences between individuals (J. Thusberg et al., 2011). To avoid the confusion resulting from the incorrect assumption of benign and pathogenic effects of “polymorphisms” and “mutations”, the most recent guideline by American College of Medical Genetics and Genomics (ACMG) has

recommended the term “variant” or “single nucleotide variant (SNV)” to be used as an alternative to both terms(Richards et al., 2015).

Susceptibility to diseases such as cancer is associated with SNVs in genes regulating cell cycle process, DNA mismatch repair, metabolism and immunity (Landau et al., 2015; Oldridge et al., 2015). The effects of alterations in gene expressions on cancer susceptibility depend on the location of SNVs. The promoter region SNVs may change promoter activity, transcription factor binding status, histone modification and DNA methylation (Schirmer et al., 2016; H. Wu et al., 2014). The exonic SNVs might either suppress or escalate transcription and translation (Griseri et al., 2011). Intronic SNVs can alter the function and binding of long non-coding RNAs by generating transcript splice variants (Xiong et al., 2015). SNVs residing in 5’-UTR can affect translation process, while SNVs in 3’-UTR can affect microRNA binding and functions (Dunna et al., 2014). SNVs locating in non-coding regions distant from the actual genes can alter transcription by affecting long-range cis-regulatory elements (He et al., 2015). Considering the role of SNVs in genetic and epigenetic changes leading to cancer susceptibility, the potential usage of SNVs as cancer biomarkers is suggested (Hubner & Houlston, 2017).

1.2. Benign and pathogenic mutations

High-throughput sequencing platforms such as whole genome sequencing (WGS) and targeted sequencing analysis have disclosed many variants associated with human diseases defined as pathogenic variants (F. Zhang & Lupski, 2015). The potential pathogenic role of non-synonymous variants such as frame shift, nonsense, missense and other types of variants affecting protein coding genes and amino acids have been investigated assiduously. An amino acid substitution can lead to production of dysfunctional proteins by altering their structure, folding and stability (Futreal et al., 2004).

Furthermore, a great amount of evidence suggests that variants residing in non-coding regions are either associated with functional consequences on nearby genes or are linked with causal coding variants (J. Thusberg et al., 2011). Interestingly, genome-wide association studies (GWAS) which investigate the statistical association between variants and complex diseases suggest that most of the significant SNVs associated with complex diseases are mapped to non-coding regions and are considerably enriched in the functional non-coding elements including enhancers, chromatin marks and DNase hypersensitivity regions (Ahonen et al., 2009; Degner et al., 2012; Trynka et al., 2013).

Our understanding of the pathogenicity of any given variant falls into a spectrum between almost certainly pathogenic to almost certainly benign for a disease. Benign variants are defined as genomic alterations without any noticeable clinical significance. ACMG has proposed five modifiers to address the pathogenic status of a variant, including pathogenic, likely pathogenic, unknown significance, likely benign, and benign (Richards et al., 2015).

Compared to germ-line variants a different terminology is often used for addressing the category of somatic variants found in tumor cells (Richards et al., 2015). Pathogenic mutations conferring fitness advantages to the cell in which they occur are called “drivers” (Hodis et al., 2012). Unlike “passenger” mutations that have merely been present in the progenitor cell of a cancer clone, driver mutations are positively selected for their tumor progression advantages. Notably, passenger mutations are biologically natural and never confer a growth/survival advantage (Greenman et al., 2007).

1.3. Cancer and somatic variants

Cancer has been related to an evolutionary process in which genomic mutations grant tumor cells with growth and survival advantages over their neighboring cells (A. Gonzalez-Perez et al., 2013; Stratton, Campbell, & Futreal, 2009). Through the mutations, normal cells are reprogramed in a way that acquires abilities such as resistance to apoptosis, deregulated cell division, and inappropriate or failed responses to external signals such as ligand mediated signaling (Hanahan & Weinberg, 2011). Cells present in a tumor tissue possess both somatic and germline mutations.

A simulation study by Caballero et al (Caballero, Tenesa, & Keightley, 2015) suggested that variants with moderate impact on development of cancer are exposed to a weaker selection compared to variants with higher impacts. Accordingly, these variants might have moderate frequency in population preventing the thorough identification of these functional regions through germline studies. The mentioned difference in selection pressure has made it possible to define cancer specific somatic mutation landscapes (Alexandrov et al., 2013).

In a similar way that it is possible to detect the negative or positive effects of non-fixed or fixed germline mutations through evolution, somatic cells with growth and survival fitness for tumor cells are also detectable through their effects on protein sequence (Abel Gonzalez-Perez et al., 2013). Genomic sequences with lower rates of somatic mutations can be significantly helpful in disclosing the impact of somatic mutations in tumor development. This is especially reliable if the mutation leads to functional changes such as gain or loss (Khurana et al., 2016).

A principal aim of cancer research has been to identify the mutations affecting the genes with causal roles in cancer susceptibility. Following the report of the first somatic mutation identified in a human oncogene (Reddy, Reynolds, Santos, & Barbacid, 1982; Tabin et al., 1982)

a substantial number of oncogenes and their relevant somatic mutations have been detected (Futreal et al., 2004). These mutations can be either pathogenic driver variants or passenger benign variants addressing the biggest challenge of all systemic mutation screenings which aim to distinguish the two groups of variants.

1.4. Importance of classifying variants

Clinical molecular laboratories are continually testing patient specimens to identify novel sequence variants potential for being associated with the etiology of complex diseases including cancer. By adopting and leveraging high-throughput technologies an ever-increasing list of genes, exomes, transcriptomes and genomes are sequenced. The challenge is to interpret the variants identified in the sequence data, which helps choose the most efficient therapy as well as predict responses to the therapy and outcomes such as overall survival and tumor recurrence-free survival (Richards et al., 2015). The central aim of clinical laboratory genetic testing is to identify or confirm the cause of diseases and help health-care providers make personalized treatment decisions. As health-care providers use genomic sequencing tests more than ever the need for accurate interpretation of the sequence data is assumed drastically important.

The American College of Medical Genetics and Genomics (ACMG) has recently published a guideline specifically for interpreting sequence variants associated with Mendelian disorders (Richards et al., 2015) classifying variants into five major tiers including i) Pathogenic, ii) Likely pathogenic, iii) Benign, iv) Likely benign, v) Uncertain significance. ACMG guideline also introduces a range of qualitative evidence for classifying genetic mutations into the mentioned groups. The evidence spans from very strongly valid for classifying a variant into a specific group to strongly, moderately and supportively valid for conducting the classification. Recommendations by ACMG guideline can be beneficial in classification of only the variants associated with

Mendelian disorders as the guideline is not intended to be used for classifying somatic, pharmacogenomic or any types of variants associated with non-Mendelian complex diseases, such as cancers.

In an attempt to standardize the interpretation of somatic sequence variants identified in sequence based cancer testing a consensus classification guideline was convened by the American College of Medical Genetics and Genomics, American Society of Clinical Oncology, and College of American Pathologists (Mark F Rogers, Hashem A Shihab, Tom R Gaunt, & Colin Campbell, 2017). The guideline introduces a four-tiered evidence based system for classifying cancer somatic variants based on their clinical significance. The tiers i) variants with strong clinical significance; ii) variants with potential clinical significance; iii) variants of unknown clinical significance; iv) variants deemed benign or likely benign (Mark F Rogers et al., 2017) were proposed based on four levels of clinical and experimental evidences spanning from level A to level D.

Level A evidence involves biomarkers that are either included in professional therapeutic/diagnostic guidelines or known to result in response/resistance to FDA approved therapies for a particular type of tumor. Level B evidence involves biomarkers that are either introduced in well-powered therapeutic/diagnostic studies or known to result in response/resistance to therapies based on well-powered studies for a particular type of tumor. Level C evidence involves biomarkers that are either included in professional therapeutic/diagnostic guidelines or known to result in response/resistance to FDA approved therapies for different types of tumors. Level D evidence involves biomarkers that are either introduced in preclinical therapeutic/diagnostic studies or known to result in response/resistance to therapies based on multiple small preclinical studies for different types of tumors (Mark F Rogers et al., 2017). To assign a somatic variant into one of the four levels, a few qualitative

measures were proposed to be met (Mark F Rogers et al., 2017). Among them, evidence of the pathogenic status of the somatic variant is critical in deciding the clinical significance tier suggested by the guideline. Accordingly, clarifying the pathogenic status of the variants is a prerequisite for the classification based on the guidelines recommendations.

1.5. Benefits of in-silico classification

Traditional laboratory-based variant classification requires a vast number of trial and error in-vivo/in-vitro experiments applied to animals and harvested tissues, respectively (Trisilowati & Mallet, 2012). These approaches are tremendously time consuming and also costly in terms of expenses such as laboratory setup, space, equipment, materials and the time spent on conducting the hands-on experiments. It is notable that the costs are multiplied by the number of repetitions of the experimental works. In-silico experimentation on the other hand, combines computing technologies and mathematical strategies with expert opinion and biological data (such as theoretical characterization of variants biology) to build models of biology (Trisilowati & Mallet, 2012) in an efficient and economic manner. These computational approaches have the potential to allow a great number of experiments to be conducted simultaneously, be observed and controlled at any level of detail , be repeated as many times as desired, while saving the laboratory costs. Many experts in the closely related areas of theoretical and computational biology have shared a view that in-silico experiments can be used as a pioneer or in accompanied with experimental studies (Trisilowati & Mallet, 2012). The principal goal of in-silico models is to predict the clinical outputs of its input data based on validated experimental data and expert opinion. The predictions ought to advise clinical trials while increasing the efficiency and reducing the costs.

As an example, in terms of classifying the clinical significance of genetic variants, in-silico strategies will provide primary information about the pathogenic effects of a given variant which

can be used to provide guidance for a wide range of consecutive preclinical and clinical steps from developing a new hypothesis and designing new experimental strategies to discovering new therapy targets and choosing the most effective treatment intervention.

1.6. Available in-silico classification tools

The availability of a great amount of genomic data provided by high throughput technologies means nothing without converting the data into useful information (Trisilowati & Mallet, 2012). In terms of interpreting novel genetic variants identified from NGS analysis, numerous computational models have been developed. The models differ based on the training variant dataset they take into account, as well as the training classification method employed for prediction. Some models are specialized for studying a specific type of variants (e.g cancer variants) (Forbes et al., 2008) or a specific mechanism associated with variants, while others are general models designed to predict whether a variant of any type is benign or pathogenic. The ultimate purpose of these tools is to prioritize for subsequent experiments that can demonstrate their actual role in the advent of the disease (Abel Gonzalez-Perez et al., 2013). The following section will briefly discuss five widely used variant classification methods including SIFT, POLYPHEN2, CADD, FATHMM-XF and CScape.

SIFT (Sorting Intolerant From Tolerant) (Pauline C Ng & Henikoff, 2003) is a computational classification tool that measures the effects of amino acid substitutions on protein function and classifies them into two groups of tolerated and not-tolerated (deleterious). To this aim, SIFT does not rely on knowledge of protein function or structure (which gives it the advantage to be applicable on uncharacterized proteins as well), but merely relies on information from multiple alignment sequence (MSA) data (P. C. Ng & Henikoff, 2001). It calculates the probability of an amino acid substitution being tolerated based on its frequency at each position in the MSA.

Given an SNV leading to amino acid substitution, SIFT assumes the most frequent amino acid at the given position in the MSA to be tolerated (J. Thusberg et al., 2011). Accordingly, SIFT uses mathematical operations to infer the sequence similarity associated with conservation features. Strongly conserved regions are not expected to tolerate most substitutions, while weakly conserved regions show more tolerance (P. C. Ng & Henikoff, 2001). Obviously, SIFT application is limited to predicting the phenotypic outcome of variants located on coding regions of the genome.

POLYOHEN2 (Polymorphism Phenotyping version2) is a variant classification tool that estimates the probability of a given variant to be damaging to human proteins function and stability (I. Adzhubei, D. M. Jordan, & S. R. Sunyaev, 2013). Its pipeline includes functional annotation of variants, mapping the coding variants into their gene transcripts, finding structural attributes and annotations of the related proteins, generating annotation profiles, estimating the effect of mutations on proteins using naïve Bayesian classifier which is a supervised machine-learning classification method. Polyphen2 utilizes two groups of features including sequence-based and structure-derived features (I. Adzhubei et al., 2013). The sequence-based features involve scores from Position-Specific Independent Counts (PSIC) software (Sunyaev et al., 1999), MSA properties, and mutation position regarding its relation to domain boundaries characterized by Pfam (Finn et al., 2010). The structure-derived features involve solvent accessibility and its changes for buried residues, as well as crystallographic B-factor (J. Thusberg et al., 2011). Similar to SIFT, POLYOHEN2 is also limited to classification of variants located in coding regions of the genome.

CADD (Combined Annotation–Dependent Depletion) is one of the leading variant scoring tools that classifies both coding and non-coding variants based on their deleteriousness which considerably correlates with pathogenicity and molecular functionality (Kimura, 1991). The

principle theory behind CADD designation is that it assumes the mutations that have not been fixed or nearly fixed in populations to be deleterious, since they reduce organismal fitness and therefore depleted by natural selection. CADD is trained to differentiate between high-frequency human-derived alleles and simulated variants using support vector machine (SVM) methods (M. Kircher et al., 2014). It is obvious that simulated variants generated by computer are not exposed to natural selection forces and therefore include deleterious variants. The features used for training CADD includes conservation metrics such as phylop (Pollard, Hubisz, Rosenbloom, & Siepel, 2010), phastCons (Siepel et al., 2005), GERP (Cooper et al., 2005); regulatory features (E. P. Consortium, 2012) such as DNaseI hypersensitivity (Boyle et al., 2008) and transcription binding sites (Johnson, Mortazavi, Myers, & Wold, 2007); transcript information such as expression levels in well characterized cell lines (E. P. Consortium, 2012) ; and protein level features such as scores form SIFT and PolyPhen. It is notable that CADD is capable of classifying both SNVs and small insertion/deletions.

Despite its advantages CADD has a number of limitations which restrict its usefulness in certain areas. The accuracy of CADD score is affected by phenomena such as biased gene conservation, background selection, local mutation rate, etc. Accordingly, CADD scores are suggested to be used only as one piece of evidence for pathogenicity and not a substitution for genetic information.

FATHMM-XF is a SNV-specific classifier for predicting pathogenic mutations applicable to both coding and non-coding variants in human genome (Rogers et al., 2018). The features used for characterizing SNVs in FATHMM-XF training data set are annotations from ENCODE (E. P. Consortium, 2012), NIH Roadmap Epigenomics (Bernstein et al., 2010), Variant Effect Predictor (VEP) (W. McLaren et al., 2016), genes models, scores from conservation tools, and information

from DNA sequence itself. FATHMM-XF is trained by gradient boosting algorithm which is a supervised machine learning method. It is notable that all the SNVs in FATHMM-XF training data set are germline mutations from Human Gene Mutation Database (HGMD) (Stenson et al., 2017) in terms of pathogenic positive examples, and non-pathogenic negative examples are a mixture of germline-somatic mutations from 1,000 Genomes Project (Abecasis et al., 2012) with minor allele frequency of at least 1%. FATHMM-XF is regarded as one of the state-of-art available scoring tools. Its previous version FATHMM-MKL (H. A. Shihab et al., 2015) is used as scoring criteria for addressing the pathogenicity of mutations in COSMIC (Forbes et al., 2008). FATHMM-MKL is trained by the same datasets used for training FATHMM-XF (mixture of germline-somatic variants from 1000 genome project as negative samples and germline variants from HGMD as positive samples). However, FATHMM-XF is trained by a set of extra features (XF stands for extra features) that distinguishes this tool from FATHMM-MKL. Since all the mutations reported in COSMIC (Forbes et al., 2008) are somatic variants from cancer patients using FATHMM-MKL as scoring criteria for classifying these variants can be biased since the tool is not originally designed for classifying cancer somatic variants. Similarly, since FATHMM-XF is also originally trained based on the characteristic of germline variants from all types of diseases its performance is drastically limited in the event of classifying somatic variants from cancers.

CScape is a classification tool intended for predicting oncogenic somatic SNVs in both coding and non-coding regions of DNA (M. F. Rogers, H. A. Shihab, T. R. Gaunt, & C. Campbell, 2017). The features used for annotating CScape training dataset are mostly similar to FATHMM-MKL and FATHMM-XF which include genomic features such as sequence spectra (Leslie, Eskin, & Noble, 2002), GC content, measures of region uniqueness, evolutionary features such as scores from phylop (Pollard et al., 2010),and phastCons (Siepel et al., 2005). Consequence features,

which represent the effect of a mutation on amino acids within its associated transcripts, are annotations from VEP (W. McLaren et al., 2016) exclusively used for characterizing variants located in coding regions of the genome. CScape has improved its performance by defining a new group of features specifically for the variants located in non-coding regions of the genome. These features fall into two main groups; sequence features, detecting effects of the disruption due to a mutation in the sequence, and genomic context features that includes the genomic features such as known start/stop codons or splice sites within a mutation's vicinity. CScape training dataset includes somatic point mutations from COSMIC, as pathogenic positive examples, and a mixture of somatic-germline SNVs from 1000 Genomes Project (Abecasis et al., 2012) as benign negative examples. The fact that CScape is specified for classifying a certain type of variants makes it a powerful candidate for classifying cancer somatic variants. However, there exist a couple of factors affecting its classification power.

Positive examples used for training CScape are collected based on the recurrence level of the mutations across COSMIC dataset. The recurrence threshold of 5 and 3 are used for collecting the variants located in coding and non-coding regions of the genome, respectively (M. F. Rogers et al., 2017) .

Relying merely on the recurrence level of a given mutation can limit the classifiers performance by introducing a potential bias to the training dataset. A mutation that is frequently reported in one cancer type is not necessarily important in occurrence of other cancer types. Accordingly, considering the number of cancer types a mutation is identified in can be beneficial in generating a pan-cancer classification tool applicable for classifying the mutations identified in all cancer types. Furthermore, the presence of germline negative variants in CScape training dataset can limit its prediction accuracy in terms of classifying cancer somatic variants. **Table 1.1**

represents a list of key properties of the tools discussed above including their strength, limitations, methods and materials.

Table 1.1 Summary of most popular available classification methods

Tool	Training Method	Positive samples		Negative samples		Strength	Limitation
		Training dataset	Mutation type	Training dataset	Mutation type		
SIFT	Alignment scores	-	-	-	-	Relatively high prediction accuracy	Limited to coding regions
POLYPHEN2	Bayesian classifica-tion	Swiss-Prot	Germline-somatic mixture	Neutral pseudo-mutation	-	Relatively high prediction accuracy	Limited to coding regions
CADD	SVM	Simulated de novo mutations	-	1000 Genome project	Germline-somatic mixture	No limitation in scope (genome wide)	Limited accuracy due to biased gene conservation, background selection, etc.
FATHMM-MKL	SVM	HGMD	Germline	1000 Genome project	Germline-somatic mixture	No limitation in scope (genome wide)	Limited accuracy for classifying specific type of variants (e.g somatic variants)
FATHMM-XF	Gradient boosting	HGMD	Germline	1000 Genome project	Germline-somatic mixture	No limitation in scope (genome wide)	Limited accuracy for classifying specific type of variants (e.g somatic variants) Biased filtering threshold for collecting positive samples
CScape	Gradient boosting	COSMIC	Somatic	1000 Genome project	Germline-somatic mixture	Specified for cancer somatic variants	Performance affected by presence of germline cancer variants in the negative training dataset. Biased filtering threshold for collecting positives

1.7. Breast cancer in young patients

Despite the low occurrence rate of breast cancer in individuals under the age of 40 (7% in developed world and 25% in developing world) these patients suffer from more severe representations of the disease as well as lower survival rate and higher risks of relapse compared

to their older counterparts (Azim et al., 2012; El Saghir et al., 2007). Several factors including accelerated mitotic pace and lymphovascular invasion, higher tumor grade at the time of diagnosis, increased expression of HER2, decreased expression of estrogen and progesterone receptors (Anders, Hsu, et al., 2008; Bharat, Aft, Gao, & Margenthaler, 2009) , and large tumor size are suggested to be associated with poor prognosis in young breast cancer patients. It is notable that even with the same disease stage and tumor size younger patients still possess a higher chances of tumor relapse and disease related death (Adami, Malker, Holmberg, Persson, & Stone, 1986; Fredholm et al., 2009; Kollias, Elston, Ellis, Robertson, & Blamey, 1997). Breast cancer's key biomarkers include Her2 and endocrine receptors as well as proliferation markers, representing different expression patterns in young patients. It is demonstrated that more aggressive subtypes (basal and HER2-enriched) are more common in young patients compared to their older counterparts (Azim & Partridge, 2014).

Using next generation sequencing various studies have examined the pattern of somatic mutations in breast cancer patients (Stephens et al., 2012). According to these studies TP53 and PIK3CA are reported to account for hosting about 25% of somatic point mutations in breast cancer (Azim & Partridge, 2014). However, there exists a lack of evidence about the landscape of somatic mutations in young patients. Stephen et al (Stephens et al., 2012) studied the occurrence of somatic mutations through whole genome sequencing of 100 breast tumors but did not find any correlation between the total number of somatic mutations and the age of the patients with ER-positive and ER-negative tumors.

With respect to the critical role of somatic mutations in cancer occurrence and the poor knowledge about the patterns of somatic mutations in particularly early onset of breast cancer,

there is an urgent demand for discovering potential age-specific biomarkers and genomic signatures benefiting younger patients' survival and prognosis.

Chapter 2 : Motivation, Hypothesis and Research Objectives

2.1. Motivation

Most of the computational tools including the approaches discussed above have been benchmarked on an appropriate set of positive and negative examples (Janita Thusberg, Ayodeji Olatubosun, & Mauno Vihinen, 2011). It is notable that an "appropriate" set represents the types of variants that the classification tool is originally designed for. The results from these benchmark studies have shown that the examined tools have the classification accuracy of about 80% which underscores the need for new classification tools with higher discrimination power.

It is noteworthy that most classification tools have been widely used for classifying a type of variants that is not same as the types of variants included in their original training dataset. In addition, there is no evidence available about the performance of these tools in classifying a type of variants other than their original training dataset (Abel Gonzalez-Perez et al., 2013). A good example of this circumstance is the employment of FATHMM-MKL (designed based on the characteristics of germline non-cancer variants) for predicting the pathogenic status of cancer somatic mutations in COSMIC dataset. Most likely, a classification tool that is designed originally based on the characteristics of cancer somatic variants can dramatically increase the classification accuracy.

Considering the importance of both coding and non-coding somatic point mutations in cancer occurrence and progression , as well as the increasing emergence of cancer sequence databases such as the international Cancer Genome Consortium (J. Zhang et al., 2011), Cancer Genome Atlas (Weinstein et al., 2013), Genomics England (100,000 genomes) Projects (Samuel & Farsides, 2017) there is a strong demand for development of interpretation tools specified for

classifying cancer somatic variants. In this study we intend to address this problem and develop a computational classification tool exclusively for cancer somatic variants or more specifically cancer somatic single nucleotide variants since point mutations are reported to be one of the most frequent mutations in human genome (Hubner & Houlston, 2017). Our proposed method classifies both coding and non-coding SNVs into two distinct groups of pathogenic, driver mutations resulting in cancer occurrence or progression, and non-pathogenic, passenger mutations that do not provide any fitness advantage for cancer cells.

2.2. Hypothesis

We assume if a computational tool is originally designed for classifying a specific type of genomic variants, in terms of assigning pathogenicity to that specific type of variant, it would have a higher discrimination power compared to a model not designed based on the characteristics of that specific type of variants. Accordingly, we hypothesize that in terms of in-silico classification of cancer somatic SNVs, the classification tool would be more robust if the classification algorithm is trained only based on the characteristics of cancer somatic SNVs rather than germ-line cancer SNVs or a mixture of gremline-somatic cancer SNVs.

2.3. Research Aims

Our study follows three general research aims listed below.

Aim 1: Developing computational models based on genomic features of cancer somatic SNVs, capable of classifying the variants into two groups of pathogenic and non-pathogenic. Comparing the discrimination power of the developed models, in terms of classifying cancer somatic variants, with the available classification tools which are designed based on the characteristics of germ line variants or a mixture of gremline-somatic variants.

Aim 2: Applying the developed models for predicting the pathogenic status of a group of cancer somatic SNVs from the young patients with breast cancer whose pathogenicity has not been assigned before.

Aim 3: Characterizing the clinical and functional significance of the predicted pathogenic positive cancer somatic SNVs.

Chapter 3 : Materials and Methods

A detailed workflow of the methods and materials used in this study is delineated in

Figure 3.1. The materials include: 1) A list of cancer somatic SNVs constituting the gold standard dataset, 2) A list of cancer somatic SNVs whose pathogenicity is not assigned constituting the prediction dataset, and 3) Feature sets that characterize the cancer somatic SNVs in gold standard. All the material will be discussed in detail in the following sections.

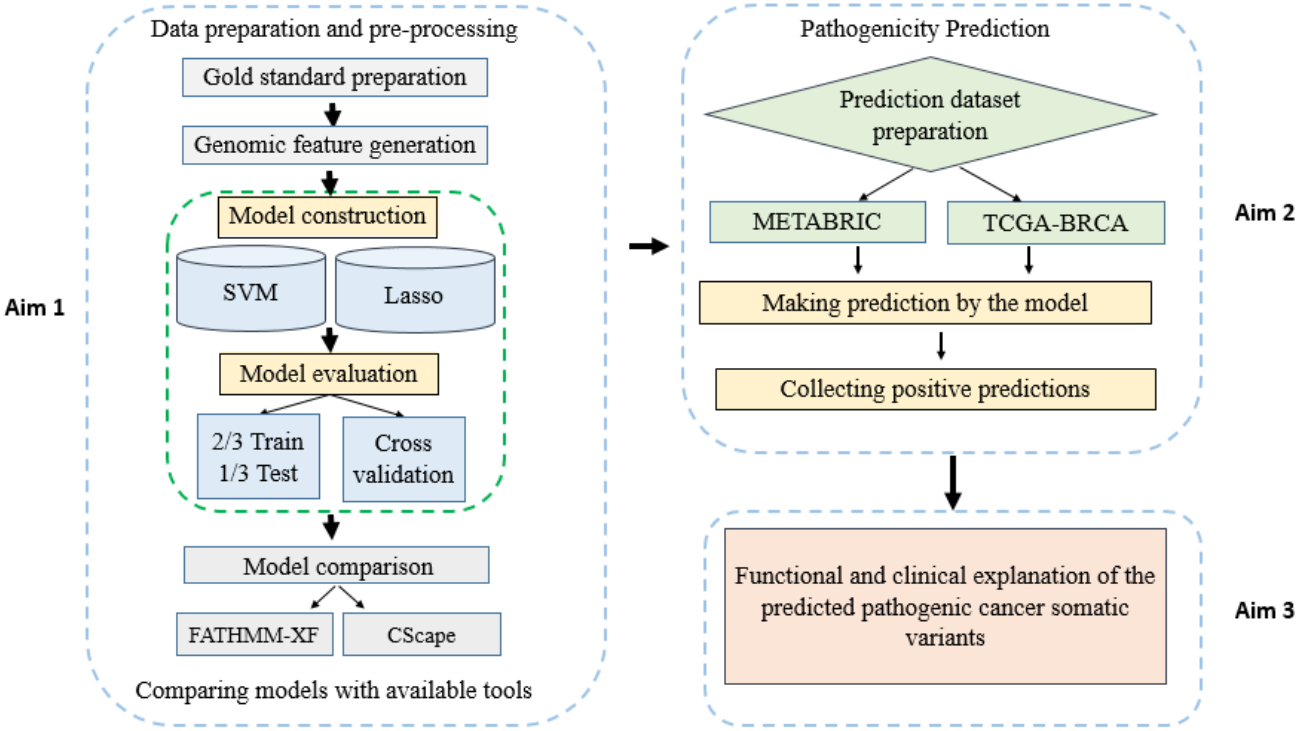

Figure 3.1 A flowchart overview of the project based on supervised classification methods

3.1. Gold standard

Gold standard is usually defined as the best referable criteria available under reasonable condition (Versi, 1992). In this study, our gold standard involves a set of somatic SNVs identified in cancer patients and healthy individuals from both coding and non-coding regions of the DNA.

As the pathogenic status assigned to the variants in gold standard provides the basis of subsequent steps of the model designation process, it is of critical importance that maximum precautions be taken in defining the pathogenic (positive) and non-pathogenic (negative) examples. It is worthwhile to emphasize that our potential models will be trained based on the pathogenic status of the somatic variants present in the gold standard.

Generally, mutations occurring in coding and non-coding regions of the genome share the same basic characteristics such as the innate changes they introduce into the DNA sequence. However, coding mutations can be studied from additional aspects in terms of the gene, transcript, and protein they deal with. Accordingly we need to split our data into coding and non-coding subsets to be able to train each set based on their specific set of features separately. To do so we followed the definitions of coding and non-coding variants provided by COSMIC (Forbes et al., 2008). Based on COSMIC description, coding variants are mostly defined by whole exome sequencing and occur in exonic regions of the genome while non-coding variants are generally identified through whole genome sequencing and occur in either inter-genic or intronic regions of the genome. It is important to mention in COSMIC database coding and non-coding variants are already separated and available in different files.

Eventually we generated two separate gold standards for coding and non-coding variants each including both positive and negative examples.

3.1.1. Labeling positive examples

Our positive dataset was built using cancer somatic point mutations reported in COSMIC (Forbes et al., 2011). COSMIC is initially designed to unite and organize thorough information about cancer somatic mutations. It integrates information about cancer mutation data from the Cancer Genome Project (CGP) (I. C. G. Consortium, 2010) at the Sanger institute, and the scientific literature. The genes from Cancer Gene Census (Sondka et al., 2018) affected by at least one somatic mutation are also reported in COSMIC. Non-coding mutations as well as structural rearrangements identified through genome wide sequencing of tumor samples have also been recently added to COSMIC branding it as the most all-inclusive available cancer somatic mutation dataset.

Among the COSMIC database annotations it is indicated whether a given mutation is also found in healthy individuals. As our first step, we excluded all the SNVs identified in non-patient individuals. However, even after applying this filter we could not assume all the remaining SNVs to be necessarily pathogenic driver mutations. Rogers et.al addressed this issue by defining a recurrence threshold to increase the likelihood that a positive example extracted from COSMIC is truly positive (M. F. Rogers et al., 2017). The score shows the number of repetitions of a given mutation across the whole dataset. They based their choice of the best threshold on the size of the positive dataset after filtering out the mutations with a frequency smaller than a given threshold. As expected, increasing the threshold decreases the number of available samples, while an extremely high threshold may result in a potential bias by limiting the samples to only a set of relevant genes. They suggested that the best threshold would provide the classifier with a sufficient number of training examples while introducing the minimum bias (M. F. Rogers et al., 2017). Following their work we considered “recurrence score” as an essential step in preparing our

positive dataset. However, to be more conservative and to ascertain that our positive samples are representatives of pathogenic somatic SNVs with high confidence, we defined a bi-dimensional recurrence score reflecting both the number of cancer types a given SNV is found in as well as the number of samples the mutation is identified in (frequency across original dataset). The novelty of this score is that it considers the occurrence of a given mutation in more than one cancer type, since traditionally an obvious step in defining the clinical implication of mutations is to determine those that have been identified in other cancers or have been involved in other disorders (Abel Gonzalez-Perez et al., 2013; Mark F Rogers et al., 2017).

3.1.2. Labeling Negative examples

We assembled our negative dataset through integrating information from two distinct datasets including dbSNP (Sherry et al., 2001) and 1000 Genome Project (G. P. Consortium, 2012), with variants from COSMIC.

The Single Nucleotide Polymorphism database (dbSNP) was initially designed to support research in many biological studies such as evolutionary studies, association studies, pharmacogenomics, functional analysis and physical genomic mapping (Sherry et al., 2001). It is a publically available archive containing an extensive collection of simple genetic polymorphisms such as single nucleotide polymorphisms (SNPs), deletion insertion polymorphisms (DIPs), short tandem repeats (STR) which are microsatellite repeat variations, and retroposable element insertions. dbSNP entries are submissions from the literature accompanied by information such as population or individual frequency of the polymorphism, sequence context of the polymorphism, and the experimental methods used to evaluate the variation (Sherry et al., 2001). Obviously, as the term “polymorphism” indicates, the contents of dbSNP are assumed as benign variations spread through populations.

The 1000 Genomes Project (G. P. Consortium, 2012) is the largest publically available directory of human genetic variations and genotype data. The initial goal of 1000 Genome Project was to identify as many human genetic variations possible with a population frequency of 1% or more. To this aim, it took advantage of reduced costs of sequencing provided by developments of high throughput technologies. The final database includes data from 2,504 individuals belonging to 26 distinct populations. All the participants declared themselves as healthy individuals at the time of sampling (G. P. Consortium, 2012).

To construct our negative dataset which includes non-pathogenic passenger somatic SNVs, we extracted the SNVs with the minor allele frequency of equal or greater than 1% in at least one 1000 Genomes population reported in dbSNP. This assured us that the collected SNVs are confidently non-pathogenic as reported by two validated resources for genetic variations in healthy individuals. At this point, our negative dataset was a mixture of germline and somatic variants. To extract somatic SNVs from this mixture, we first collected all the SNVs annotated as "found in healthy individuals" from COSMIC, providing us with somatic variants, and subsequently identified the mutual SNVs between these mutations and our mixture. Eventually, we were confident each SNV in our negative dataset is a somatic benign variant found in healthy individuals. **Figure 3.2** designates our negative dataset through a schematic Venn diagram. Rogers et al (Rogers et al., 2018) assumed a chance of potential bias introduced to negative variants located remotely from positive samples (Mark F Rogers et al., 2017), specifically under conditions such as positive examples being mostly located in critical spots such as transcription start sites, while negative samples being sporadically distributed throughout the genome (Martin Kircher et al., 2014; Hashem A Shihab et al., 2015). In their study, they filter out the negative samples located outside of the 10,000 and 1000 nucleotide window distance from the coding and non-coding

positive samples, respectively. Considering the sufficient size of our negative dataset as well as the all-inclusive identity of our potential model we found it unnecessary to apply the mentioned filter.

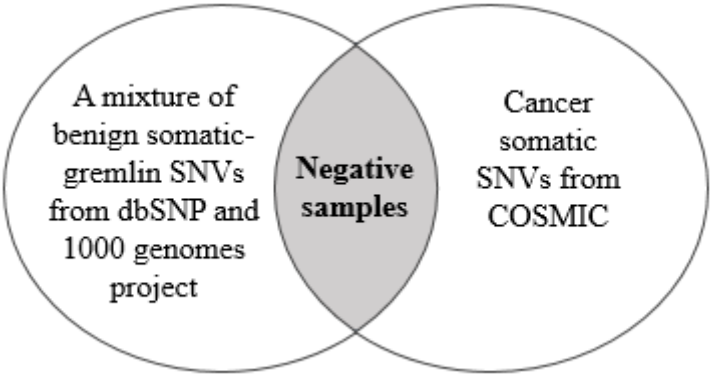

Figure 3.2 Schematic venn diagram of negative examples. Negative dataset includes those SNVs common to dbSNP, 1000 genome project and COSMIC.

3.2. Genomic features

One goal of our classification models is to learn the discrepancies and similarities between pathogenic and non-pathogenic cancer somatic SNVs to be able to determine the likely functional consequences of the mutations. To this aim, we rely on genomic features that characterize the mutations through a wide range of criteria spanning sequence characteristics, genomic content of the mutation sites and functional and structural consequences of the mutations. Genomic features defined in this project are mostly annotations from different projects such as ENCODE (E. P. Consortium, 2012), CADD (Rentzsch, Witten, Cooper, Shendure, & Kircher, 2018), and ENSEMBL variant effect predictor (VEP)(William McLaren et al., 2016), scores from pathogenic SNV predictors such as POLYPHEN (Ivan Adzhubei, Daniel M Jordan, & Shamil R Sunyaev, 2013) and SIFT (Pauline C Ng & Henikoff, 2003), as well as information from variant browsers such as BRAVO (bravo.sph.umich.edu/freeze5/hg38). We have grouped the features into

four major subsets, each portraying the mutations from a specific aspect including I) structural and genomic context features, II) epigenetic features, III) genomic distance features, IV) genomic conservation features.

Most of our defined features (about75%) from all of the four major groups are applicable to SNVs from both coding and non-coding regions of the genome. However, as mentioned earlier, we were obliged to define two distinct set of features for coding and non-coding variants since some features are coding-specific features characterizing specific attributes of genes or transcripts affected by the SNVs. For example, "Distance of the affected amino acid from start codon" is a coding-specific feature not definable for non-coding SNVs.

It is notable that since to date only a limited genomic annotation sources are available, almost all the features defined in this project, except for a handful of self-defined features (such as Transversion), overlap with the features defined by existing classification tools including FATHMM-XF, CADD and CScape.

Overall, we defined 80 genomic features for characterizing coding variants and 65 genomic features for non-coding variants. A brief description of each genomic feature along with the major category it belongs to is provided in **Table 3.1**.

Table 3.1 A summary of the features defined for the variants in coding and non-coding gold standards. Coding specific features are assigned as "Yes" in the last column

#	Feature	Description	Feature category	Coding specific
1	Transversion	Type of nucleotide change (Trans version/sition)	Structural & genomic context	No
2	Consequence S	Deleterious score assigned to potential impact of mutations based on VEP consequences	Structural & genomic context	No
3	Consequence category	Impact category assigned to mutations based on VEP consequences	Structural & genomic context	No
4	GC %	Percentage of GC in a +/-75 bp window	Structural & genomic context	No
5	CpG %	Percentage of CpG in a +/-75 bp window	Structural & genomic context	No
6	cDNA Pos	Distance from transcription start	Distance feature	Yes
7	Rel cDNA Pos	Relative distance from transcription start	Distance feature	Yes
8	CD start	Distance from coding start site	Distance feature	Yes
9	Rel CD start	Relative distance from coding start site	Distance feature	Yes
10	Prot Pos	Amino acid distance from coding start site	Distance feature	Yes
11	Rel Prot Pos	Relative amino acid distance from coding start site	Distance feature	Yes
12	Min dist TSS	Minimum distance to Transcribed Sequence Start(TSS)	Distance feature	No
13	Min dist TSE	Minimum distance to Transcribed Sequence End (TSE)	Distance feature	No
14	SIFT	SIFT score	Structural & genomic context	Yes
15	PolyPhen	PolyPhen score	Structural & genomic context	Yes
16	PhastCons-pri	PhastCons Primate score	Conservation	No
17	PhastCons-mam	PhastCons Mammalian score	Conservation	No
18	PhastCons-ver	PhastCons Vertebrate score	Conservation	No
19	Phylop-Pri	Phylop Primate score	Conservation	No
20	Phylop-mam	Phylop Mammalian score	Conservation	No
21	Phylop-ver	Phylop Vertebrate score	Conservation	No
22	Background S	Background selection score	Conservation	No
23	Gerp RS	Gerp RS score	Conservation	Yes
24	Gerp RS pval	Gerp RS score p-Value	Conservation	Yes
25	Gerp N	Neutral evolution score from GERP++	Conservation	No
26	Gerp S	Rejected substitution score from GERP++	Conservation	No
27	H3K4me1	Maximum H3K4 methylation level from Encode	Epigenetic feature	No
28	H3K4me2	Maximum H3K4 dimethylation level from Encode	Epigenetic feature	No
29	H3K4me3	Maximum H3K4 trimethylation level from Encode	Epigenetic feature	No
30	H3K9ac	Maximum H3K9 acetylation level from Encode	Epigenetic feature	No
31	H3K9me3	Maximum H3K9 trimethylation level from Encode	Epigenetic feature	No
32	H3K27ac	Maximum H3K27 acetylation level from Encode	Epigenetic feature	No
33	H3K27me3	Maximum H3K27 trimethylation level from Encode	Epigenetic feature	No
34	H3K36me3	Maximum H3K36 trimethylation level from Encode	Epigenetic feature	No
35	H3K79me2	Maximum H3K79 dimethylation level from Encode	Epigenetic feature	No
36	H4K20me1	Maximum H4K20 methylation level from Encode	Epigenetic feature	No
37	H2AFZ	Maximum level of H2A Histone Family Member Z	Epigenetic feature	No
38	DNase	Maximum DNase-seq level form Encode	Epigenetic feature	No
39	RNA	Maximum total RNA-seq level form Encode	Epigenetic feature	No
40	Grantham	Grantham score	Conservation feature	Yes
41	PHRED	CADD PHRED score	Structural & genomic context	No
42	cHMM-E1	Number of 48 cell types in chromosome state 1 from chromHMM	Epigenetic feature	No
43	cHMM-E2	Number of 48 cell types in chromosome state 2 from chromHMM	Epigenetic feature	No

44	cHMM-E3	Number of 48 cell types in chromosome state 3 from chromHMM	Epigenetic feature	No
45	cHMM-E4	Number of 48 cell types in chromosome state 4 from chromHMM	Epigenetic feature	No
46	cHMM-E5	Number of 48 cell types in chromosome state 5 from chromHMM	Epigenetic feature	No
47	cHMM-E6	Number of 48 cell types in chromosome state 6 from chromHMM	Epigenetic feature	No
48	cHMM-E7	Number of 48 cell types in chromosome state 7 from chromHMM	Epigenetic feature	No
49	cHMM-E8	Number of 48 cell types in chromosome state 8 from chromHMM	Epigenetic feature	No
50	cHMM-E9	Number of 48 cell types in chromosome state 9 from chromHMM	Epigenetic feature	No
51	cHMM-E10	Number of 48 cell types in chromosome state 10 from chromHMM	Epigenetic feature	No
52	cHMM-E11	Number of 48 cell types in chromosome state 11 from chromHMM	Epigenetic feature	No
53	cHMM-E12	Number of 48 cell types in chromosome state 12 from chromHMM	Epigenetic feature	No
54	cHMM-E13	Number of 48 cell types in chromosome state 13 from chromHMM	Epigenetic feature	No
55	cHMM-E14	Number of 48 cell types in chromosome state 14 from chromHMM	Epigenetic feature	No
56	cHMM-E15	Number of 48 cell types in chromosome state 15 from chromHMM	Epigenetic feature	No
57	cHMM-E16	Number of 48 cell types in chromosome state 16 from chromHMM	Epigenetic feature	No
58	cHMM-E17	Number of 48 cell types in chromosome state 17 from chromHMM	Epigenetic feature	No
59	cHMM-E18	Number of 48 cell types in chromosome state 18 from chromHMM	Epigenetic feature	No
60	cHMM-E19	Number of 48 cell types in chromosome state 19 from chromHMM	Epigenetic feature	No
61	cHMM-E20	Number of 48 cell types in chromosome state 20 from chromHMM	Epigenetic feature	No
62	cHMM-E21	Number of 48 cell types in chromosome state 21 from chromHMM	Epigenetic feature	No
63	cHMM-E22	Number of 48 cell types in chromosome state 22 from chromHMM	Epigenetic feature	No
64	cHMM-E23	Number of 48 cell types in chromosome state 23 from chromHMM	Epigenetic feature	No
65	cHMM-E24	Number of 48 cell types in chromosome state 24 from chromHMM	Epigenetic feature	No
66	cHMM-E25	Number of 48 cell types in chromosome state 25 from chromHMM	Epigenetic feature	No
67	Intron-Exon	Binary variable, if the SNV is located in intron its score is 0 and if its located in exon the score is 1	Structural & genomic context	Yes
68	Domain-VEP	Dummy variable based on Domain annotations from VEP (e.g. if the variants is located in a sigp)	Structural & genomic context	Yes
69	BRAVO-Freq100bp	Count of frequent SNVs (MAF>0.05) in 100 bp window distance from the mutation base on BRAVO	Structural & genomic context	No
70	BRAVO-Rare100bp	Count of rare SNVs (MAF<0.05) in 100 bp window distance from the mutation base on BRAVO	Structural & genomic context	No

71	BRAVO-Sngl100bp	Count of single occurrence of the SNVs (MAF<0.05) in 100 bp window distance base on BRAVO	Structural & genomic context	No
72	BRAVO-Freq1000bp	Count of frequent SNVs (MAF>0.05) in 1000 bp window distance from the mutation base on BRAVO	Structural & genomic context	No
73	BRAVO-Rare1000bp	Count of rare SNVs (MAF<0.05) in 1000 bp window distance from the mutation base on BRAVO	Structural & genomic context	No
74	BRAVO-Sngl1000bp	Count of single occurrence of the SNVs (MAF<0.05) in 100 bp window distance base on BRAVO	Structural & genomic context	No
75	BRAVO-Freq10000bp	Count of frequent SNVs (MAF>0.05) in 10000 bp window distance from the mutation base on BRAVO	Structural & genomic context	No
76	BRAVO-Rare10000bp	Count of rare SNVs (MAF<0.05) in 1000 bp window distance from the mutation base on BRAVO	Structural & genomic context	No
77	BRAVO-Sngl10000bp	Count of single occurrence of the SNVs (MAF<0.05) in 100 bp window distance base on BRAVO	Structural & genomic context	No
78	BRAVO-dist-mutation	Distance between closest up and down BRAVO SNVs	Distance feature	No
79	RemapOverlap TF	Number of different transcription factor binding affected by the mutation from Remap	Structural and genomic context	Yes
80	RemapOverlap CL	Number of different transcription factor binding – cell line combination affected by the mutation from Remap	Structural and genomic context	Yes

3.2.1. Structural and genomic context features

Structural and genomic context features are related to the genomic features characterizing sequence attributes of the mutations location. These features estimate the disruption in the mutations surrounding sequence both in coding and non-coding regions. This information is mostly obtained from quantifying the annotations from VEP (William McLaren et al., 2016) and CADD (Rentzsch et al., 2018). For instance, “Transversion” is a self-defined feature belonging to this group; it is a binary variable taking the value of “0” if both the reference and the alternative nucleotides are either purine (Double carbon-nitrogen ring and four nitrogen atoms) or Pyrimidine

(Single carbon-nitrogen ring and two nitrogen atoms). “Transversion” will be equal to “1” if one of the reference or alternative nucleotides is purine while the other is pyrimidine.

For coding regions, genomic context features also assess the potential impact of a given SNV on amino acids and its associated transcripts and proteins. For instance, “Consequence score” is a feature extracted from VEP annotations assigned based on the potential deleterious impact of the mutations. Scores form predictors of pathogenic variants in coding regions such as POLYPHEN and SIFT, which quantify the potential impact of a variant on proteins structure or function, also belong to this category. It is noteworthy that using the final consequence labels such as pathogenicity impact levels (“High”, “Low”) from VEP, or pathogenic status of mutations (such as “Deleterious”, “Benign”) from prediction tools will introduce a bias to our final classification models (Rogers et al., 2018). To avoid the potential bias we did not include any final consequence impact predicted by any of the classification tools, instead we allow our models to learn these consequences through the training process.

A summary of all the features belonging to “structural and genomic context” category is provided in **Table 3.1**.

3.2.2. Epigenetic features

Epigenetic marks such as modifications in DNA methylation, histones and other chromatin associated proteins play an important role in regulating gene expression patterns and chromatin architecture (Sharma, Kelly, & Jones, 2009). It is suggested that failure in appropriate maintenance of epigenetic features may result in deregulation of various signaling pathways leading to complex diseases including cancer (Egger, Liang, Aparicio, & Jones, 2004; Jones & Baylin, 2002). Advances in cancer research have demonstrated that the interaction between genetic changes (such

as mutations) and epigenetic alterations is observable at all stages of cancer development promoting cancer progression (Jones & Baylin, 2002, 2007; Jones & Laird, 1999). Importance of epigenetic features has got to the point that recent studies have considered epigenetic alterations as the potential initiator of some forms of cancers (Feinberg, Ohlsson, & Henikoff, 2006).

In this study we have defined a number of epigenetic features describing histone modifications and methylation alterations. Histone modification features are extracted from ENCODE project reporting 14 histone modifications across 45 cell lines using CHIP-Seq peak calls (Y. Zhang et al., 2008), and methylation alterations include 25 chromosome states from ENCODE (E. P. Consortium, 2012) and NIH Roadmap Epigenetics (Bernstein et al., 2010) cell lines, prepared based on whole genome bisulfite sequencing (WGBS) (Vargas-Landin, Pfluger, & Lister, 2018).

A summary of all epigenetic features category is provided in **Table 3.1**.

3.2.3. Genomic distance features

Genomic distance features measure the distance between a given SNV and critical functional and structural genomic elements such as transcription start and end sites. These features are specifically proved to be important in classifying non-coding SNVs (Rogers et al., 2018). For instance, identifying the SNVs closer to exon boundaries, splice sites, or promoter regions might help the training models learn the relationships between non-coding mutations and important gene elements (Rogers et al., 2018). In addition, it is suggested that distance features can be determinants of evolutionary conservation and cis-regulatory motifs (Vardhanabhuti, Wang, & Hannenhalli, 2007). In our project we have defined both regular distance features (based on actual base pair distance between elements) as well as relative distance features which normalize the

distances based on their gene lengths. A summary of all genomic distance features category is provided in **Table 3.1**.

3.2.4. Genomic conservation features

Conserved regions of the genome that are remained unchanged over long evolutionary periods are high likely to be functional (J. W. Thomas et al., 2003). Genomic conservation features measure the evolutionary conservation at the mutation alignment sites in an effort to help the training models learn the relationships between the measurements and pathogenicity of the SNVs. It is noteworthy that conserved non-coding regions mostly involve regulatory elements (J. Thomas et al., 2003). We have obtained conservation scores from different alignment-based conservation evaluation tools including PhastCons (Siepel et al., 2005; Siepel & Haussler, 2004) which measures the probability of an individual genomic site belonging to a conserved element, as well as Phylop (Pollard et al., 2010) measuring the –log(p-value) under the null hypothesis of neutral evolution. Using the mentioned tools, we have measured conservation from vertebrates, mammals and primates. In addition, we have used scores from Genomic Evolutionary Rate Profiling (GERP) software (Cooper et al., 2005) that identifies constrained genomic elements reflecting the strength of past selection forces through multiple alignment.

A summary of all genomic conservation features category is provided in **Table 3.1**.

3.3. Handling missing data

Missing data, known as the values not collected for an observation of interest (H. Kang, 2013), is a common occurrence in almost all data collection-based research studies. Missing data can affect the results of analysis by reducing the statistical power of the tests, causing bias in the estimation of feature parameters, reducing the representativeness of the sample data, and eventually complicating the final interpretation of the analysis (H. Kang, 2013). Each of the mentioned effects can invalidate the conclusions by questioning the validity of model training processes. Accordingly, many studies have focused on developing missing data handling methods to avoid or minimize their following complications (O'neill & Temple, 2012). These methods are classified into two major categories including i) deleting all the missing data, ii) imputing all the missing data.

3.3.1. Missing data deletion

The most frequently used approach in face of missing data is to simply delete all the variables with even one missing data and analyze the complete set of remaining data. Deleting missing data is popular to the level that it is set as the default option in many data analyzing software packages (H. Kang, 2013). However, it can introduce bias in estimating the parameters specifically if the data are not missed at random (Donner, 1982). Furthermore, since deleting missing data can decrease the size of sample data set significantly, it is not recommended if the shrinkage of the data size is more than 5 percent (Wulff & Ejlskov, 2017).

3.3.2. Missing data imputation

Imputation is the process of filling the missing values with hypothetical estimations. Unlike "missing data deletion" imputation preserves all the data by using information from other available values. Single imputation methods such as substituting the missing data by "mean" or "median" might lead to underestimation of errors (Malhotra, 1987) or inconsistent bias in the event of variation in the number of missing data for different variables. On the other hand, multiple imputation methods such as regression imputation can address the statistical uncertainty resulting from single imputation, and also reduce the standard error while stabilizing the standard deviation or the shape of data distribution (H. Kang, 2013).

In our study we used a popular regression imputation technique called MICE (Multivariate imputation by chained equations) which is applicable to datasets with missing values in multiple variables (Wulff & Ejlskov, 2017), and is also adapted to handle different types of data (e.g. continuous or binary) (Buuren & Groothuis-Oudshoorn, 2010). MICE uses the distribution of the available data for each variable (starting from the variable with least missing values) to estimate a set of candidate values for replacing the missing value. It repeats the estimation procedure M times (number of missing data for a variable) each time including a random component to address the statistical uncertainty (Wulff & Ejlskov, 2017).

3.4. Data Normalization

Data normalization or scaling is known to be beneficial in improving the performance of some classifiers such as SVM (Minaxi Arora, 2014). To investigate the effect of data normalization on our models we used python Scale package from sklearn library (Pedregosa et al., 2011) to normalize the feature values in our gold standard datasets. Generally two major points are counted for data normalization (Minaxi Arora, 2014): first, to avoid large values in wider numeric scales which can cause numerical problems; second, to simplify the classification calculation process.
To identify the best data preparation strategy we trained our models on both the normalized and non-normalized data.

3.5. Classification methods

In machine learning, classification is a process in which the computer learns from its input data by configuring the patterns of similarities and differences between different classes and eventually uses the learning knowledge to classify new observations. There are many machine learning classification methods available but considering the high-throughput nature of genomic data one important aspect in choosing the best method is the computational efficiency and classification power of the algorithm given thousands of variants (Abel Gonzalez-Perez et al., 2013). In this study we have used two of the most popular classification methods with demonstrated capability of dealing with big data (e.g. cancer genomic data) as well as conducting feature selection.

3.5.1. Lasso (least absolute shrinkage and selection operator) regression model

Lasso model is a type of linear regression that sums up penalty scores equal to the absolute value of the coefficients resulting in elimination of features with large penalty scores (Tibshirani, 1996).

The standard logistic regression model with the general equation can be expressed as

$$\text{Log}\,(\text{P}(Y_i|X_i)) = \beta_0 + \sum_{j=1}^{m} \beta_j \; X_{ij}$$

where Y is the pathogenic status of example *i* and X is the feature j of the given example *i*, j=1,2,...,80 for the modeling of coding regions and *j*=1,2,…,65 for the modeling of noncoding regions. It has been shown that the standard logistic regression model often results in inaccuracies once used for big data analysis such as genetic studies. However, using penalty methods such as Lasso can shrink the coefficients in the regression models and control the effects of over-fitting and model instabilities (Fontanarosa & Dai, 2011). To do so, lasso selects the coefficients (βs) based on the maximization of $\text{Log}\,(\beta|\,\text{Y, X}) - \lambda\sum_{j=1}^{m} |\,\beta_j|$ Where log (β| Y, X) is the logistic log odds of a mutation to be pathogenic for one unit change in the value of the feature X (while the values of other Xs do not change), and λ is the shrinkage parameter.

The final coefficients estimated by Lasso regression indicate the contribution of each feature in predicting the outcome value. Lasso conducts feature selection by setting the coefficients of non-discriminative features to zero. This is especially suitable for models with high levels of multicollinearity (Farrar & Glauber, 1967) which occurs when there is a high correlation between two or more features in the model.

In our study since we have defined different features that characterize the same implications (such as the effects of mutations on conservation), it is not far from expectations if

the features from the same subgroup correlate with each other. On the other hand, not all the defined features equally contribute to the model's discrimination power. Accordingly, Lasso's capability in handling multicollinearity and feature selection perfectly matches with our data in a way that increases the prediction accuracy and interpretability of the final regression model (Tibshirani, 1996).

3.5.2. Support vector machine (SVM) model

SVM is a powerful classification method capable of predicting labels of two classes based on their defined features (S. Huang et al., 2018). SVM discriminates the two groups by creating a decision boundary called the hyperplane which is oriented in a way that keeps the largest possible distance from the closest data points of each class known as support vectors shown in **Figure 3.3**.

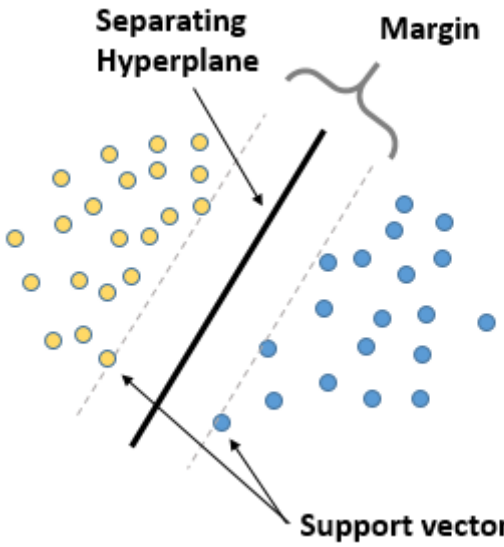

Figure 3.3 SVM classification algorithm. SVM separates the members of two classes using support vectors and a hyper-plane

In addition to linear classification, SVM is supplied with kernel method which facilitates certain calculations needed for high dimensional space non-linear classification (S. Huang et al., 2018). Among other parameters choice of a kernel function can affect SVM classification power enormously. However, there is no certain way of choosing the best kernel without conducting trial and error practices starting from a simple SVM model.

In this study we have used SVM as our classifier regarding its strength in detecting a classification pattern from hard-to-discern (S. Huang et al., 2018) data sets. Since genomic data (e.g. cancer data) are large, noisy and complex SVM is a well-suited classification method in many cancer studies (Furey et al., 2000; Golub et al., 1999; Moler, Chow, & Mian, 2000).

3.6. Model evaluation

Model building is a trial-and-error process in search for the best model parameters fitting the data. Generally, a proportion of the main data is used for training the model and the remaining proportion is used for both testing the models performance and detection of critical situations such as over-fitting (when model performance on training data is significantly better that testing data) (Van der Aalst et al., 2010). In this study we tried two data selection approaches for defining the training and testing subsets: i) Training-test strategy, ii) Cross-validation strategy

In addition, in order to evaluate our models performance we used “area under the receiver operating characteristic (ROC) curve” (AUC) as our measuring criteria. AUC of a ROC curve is the most well-established estimation metrics which is also helpful in visualizing the classification performance of a model.

3.6.1 Training-test strategy

Collecting a proportion of data that can be used for training and testing a computational model can be as simple as randomly choosing a bigger subset of the data for training and the remaining smaller subset for testing. Traditionally, two-third and one-third of the main data is used for training and testing, respectively. This is a reliable method as long as the dataset is thoroughly homogenous and properly shuffled. Otherwise, the estimation of the model performance might be deluded by selection bias (Cawley, 2010). For example, in the event of a non-discriminative model that classifies any input into just one class, if the testing dataset mostly contains members of that specific group, based on the testing results the model would be mistaken for being high-discriminative. To avoid such a similar circumstance, cross-validation strategy is proved to be helpful.

3.6.2. Cross-validation strategy

Cross-validation is a resampling procedure with only a single parameter called k indicating the number of groups the dataset is to be split into. Accordingly the procedure is often called k-fold cross-validation (Kohavi, 1995). Through cross-validation the data is divided into k groups of roughly equal size. The first fold (group) is treated as testing dataset and the remaining k-1 is used for training. Once the model is fit on the training data the procedure will be repeated using the second fold as testing dataset and other k-1 folds as training dataset. The loop will continue until each fold is used as testing dataset. It is important that the k value be chosen carefully to avoid overestimation of the model performance resulting from a small testing dataset. In this study we have chosen K=10 which is a recommended value in the field of applied machine learning (P. Zhang, 1993).

3.6.3. Receiver Operating Characteristic (ROC) curve

ROC curve evaluates the performance of a model by plotting the false positive rate (FPR) (1-specificity) against the true positive rate (TPR) (also known as sensitivity) of a classification task performed by the model using various decision thresholds. Area under the curve (AUC) is the measure of the models discrimination power demonstrating how well the model can distinguish between the members of the two groups. The AUC ranges from 0.5 (indicating a non-discriminative model) to a theoretical maximum of 1 (indicating a perfect discrimination) (Cook, 2007). AUC is equivalent to the probability that the measure of a random positive sample is higher than a random negative sample based on the models estimations (Cook, 2007; Fawcett, 2006).

Considering the accuracy and popularity of AUC measurements in evaluating machine learning classification models (Cook, 2007), in this study we have based our evaluation assessments on calculations from ROC curves.

3.7. Applying the models to breast cancer cohort studies

Using our trained models we intended to study the genomic profile of young breast cancer patients in terms of occurrence of somatic point mutations and the potential impacts of the mutations on disease treatment and prognosis.

To this aim, we applied our trained models on a set of cancer somatic variants whose pathogenicity was not clinically assigned yet and also the variants were unseen to our models (not included in the training dataset).

We used the information from two cohort studies investigates the genomics of breast cancer tumors, which include METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) (Pereira et al., 2016) and TCGA (The Cancer Genome Atlas)-BRCA (Grossman et

al., 2016). With respect to the higher severity of breast cancer in young patients (Azim et al., 2012) we excluded the information from older patients and focused on applying our models on SNVs identified in patients younger than 45 years of age.

3.7.1. METABRIC

METABRIC is a vigorous population-based project integrating data from more than 2,000 fresh-frozen breast cancer tissue samples from tumor banks in Canada and UK (Curtis et al., 2012). The stage-one of the study have generated genomic and transcriptional profiles of the specimens through microarray analysis of extracted DNA and RNA samples using Affymetrix SNP 6.0 and Illumina HT-12 v3 platforms, respectively (Curtis et al., 2012). The stage-two of the METABRIC study preformed whole exome sequencing of 173 genes from the 2,433 primary breast cancer samples used in the stage-one. The 173 genes include those that had been observed to be mutated in at least 2 samples from 5 sequencing-based studies (Banerji et al., 2012; Ellis et al., 2012; Network, 2012; Shah et al., 2012; Stephens et al., 2012). METABRIC also includes the long-term follow-up clinical data for majority of the patients (2,319 patients).

From a total of 32,476 SNVs identified by the stage-two of the METABRIC study we were able to define the genomic features for 13,942 somatic SNVs from all the patients (including patients older and younger than 45 years of age). For the interest of our study, we filtered out the mutations belonging to patients older than 45 years of age. The age filtering process shrank the number of eligible somatic mutations to 1882 from 326 young patients. In order to be able to apply our model to the collected mutations we had to generate all the coding features for each mutation in the dataset. Obviously, we applied our “coding model” to the data since all the mutations were

identified by exome sequencing techniques. Furthermore, we used the clinical follow-up data for survival analysis explained later in this chapter.

We downloaded the data from (www.cbioportal.org/study/summary?id=brca_metabric) which is freely available to the public. It is notable that METABRIC study uses the information from 19th version of the human reference genome (GRCh37) (Rhead et al., 2009). In this study we have upgraded all GRCh37 genome coordinates to the GRCh38 (Casper et al., 2017) to be able to benefit from the updated annotations provided in the latest reference genome assembly (GRCh38) aiding feature generating process. To this aim we have used LiftOver (Kent et al., 2002) which is a computational tool by USCS (Kent et al., 2002) that converts genome coordinates between different genomic assemblies.

3.7.2. TCGA-BRCA

The Cancer Genome Atlas (TCGA) is a cooperative cancer genomic program between the National Human Genome Research Institute and the National Cancer Institute portraying 33 cancer types through molecular characterization of approximately 20,000 tumor specimens and their matched normal tissues (Grossman et al., 2016). TCGA-BRCA is part of the TCGA program characterizing human breast tumors using five molecular assessment platforms involving Affymetrix SNP arrays, Illumina Infinium DNA methylation chips, Agilent mRNA expression microarrays, whole exome sequencing and microRNA sequencing. Overall, we downloaded TCGA-BRCA data from http://download.cbioportal.org/brca_tcga.tar.gz encompassing 80,227 somatic SNVs from 976 patients (including patients older and younger than 45 years of age). For the interest of our study we only focused on the data from patients younger than 45 years of age. Eventually, we were able to define genomic features for a total of 8,647 somatic SNVs from 142

young patients (<45 years old). It is noteworthy that regarding the data collection approaches followed by TCGA-BRCA study we were able to identify somatic SNVs from both coding (e.g. provided by whole exome sequencing) and non-coding regions (e.g. provided by microRNA sequencing) of the genome. Accordingly, we extracted 6910 and 1737 somatic SNVs from coding and noncoding regions of the genome, respectively.

Similar to METABRIC data, here again we used LiftOver to up-grade the GRCh37 genomic coordinates of somatic SNVs to GRCh38.

3.8. Survival analysis

Survival analysis is defined as a set of statistical approaches for estimating probability of the time for occurrence of an event of interest such as patients death or disease relapse measured in days, months, years, etc. Ali et al (Ali, Chlon, Pharoah, Markowetz, & Caldas, 2016) suggested the following hierarchy for conducting different survival analysis once more than one type of follow-up data is available; 1) disease free survival (DFS, time until the relapse of the disease), 2) distant metastasis-free survival, 3) disease specific survival (DSS, death from the disease of interest), 4) overall survival (OS, death from any cause). In this study, we have the follow-up data for conducting DSS and OS analysis from METABRIC data, as well as the follow-up data for performing DFS and OS analysis from TCGA-BRCA. We used Kaplan Meier methods (Rich et al., 2010) to estimate and graph both young and old patients' survival probabilities.

3.8.1. SNV-level survival analysis

SNV-level survival analysis regards the occurrence of a given mutation in a given patient as the "event of interest". On other words, it assesses whether the survival time of the patients who

harbor a specific mutation in their genome is significantly different from the patients who do not harbor the mutation.

We identified the mutations that significantly (p-value <0.05) affected the survival time of young patients and compared the results with outcomes from SNV-level survival analysis of old patients. We were particularly interested in the mutations that affected the survival experience of young patients but did not have a significant effect on the survival time of the older patients. It is noteworthy that we conducted the SNV-survival analysis for the somatic SNVs that were predicted to be pathogenic based on our trained models used for predicting the somatic variants from METABRIC and TCGA-BRCA studies.

3.8.2. Gene-level survival analysis

In addition to SNV-level survival analysis mentioned above we also conducted gene-level survival analysis. In gene-level survival analysis the fact that whether a given gene is mutated or not is considered as the "event of interest". Again, we identified the genes that once being mutated, significantly (p-value < 0.05) affected the survival time of the patients. In consistent with the purpose of our study we were only interested in the genes where their mutation status was significantly associated with the survival time of young patients while not having a significant effect on the survival time of the older patients. Here we conducted the gene-level survival analysis for a set of genes that were affected by the somatic SNVs labeled as pathogenic based on our models predictions on METABRIC and TCGA-BRCA data. To conduct gene-level survival analysis, we defined a gene as mutated in a given sample if it is affected by at least one predicted pathogenic positive SNV in the sample. We focused our survival analysis on the genes that harbor a higher number of predicted pathogenic positive SNVs. To be consistence with the frequency

threshold we defined for labeling pathogenic positive examples in our gold standards at the very first steps of model building process, we selected the genes harboring at least four and two mutations in coding (METABRIC and TCGA-BRCA coding) and non-coding (TCGA-BRCA) datasets, respectively.

3.9. Gene set enrichment analysis (GSEA)

Gene set enrichment analysis also referred to as functional enrichment analysis is an analytical method that determines whether the members of a given list of genes are over-represented in a priori known set of genes or proteins (Subramanian et al., 2005). GSEA also helps in investigating the association between expressions of a given list of genes with disease phenotypes.

In this study we used Enrichr software (Chen et al., 2013; Kuleshov et al., 2016) to conduct enrichment analysis investigating whether the genes that are affected by the somatic SNVs predicted as pathogenic by our models are over-represented in any interesting cellular pathway or function. Enrichr is an integrative software encompassing information from over 150 up-to-date gene libraries from publications about genes co-expression and cellular pathways.

Chapter 4 : Results and Discussions

4.1. Gold standard dataset

We defined two separate gold standards for cancer somatic variants located in coding and non-coding regions of the DNA as coding variants possess additional features regarding their direct association with transcripts and proteins. **Table 4.1** shows the overall number of positive (pathogenic) and negative (non-pathogenic) cancer somatic variants belonging to coding and non-coding gold standards. Detailed explanations of how we got the number of positive and negative examples for each gold standard, mentioned in **Table 4.1**, is provided in the following sections. Obviously, we are certain that there is no overlap between the positive and negative examples which means each SNV in our gold standard has a unique label of either positive or negative.

Table 4.1 The total number of positive and negative samples in coding and non-coding regions

Data type	Coding	Non-coding
Positive	12,313	28,993
Negative	16,594	58,995
Total	28,907	87,998

4.1.1. Positive examples

Following our bi-dimensional scoring approach for labeling positive cancer somatic SNVs from COSMIC, we counted the number of SNVs belonging to different thresholds of each dimension of our scoring method shown in **Table 4.2** and **Table 4.3** for coding and non-coding datasets, respectively.

As an example, the highlighted cells in **Table 4.2** indicate 122,054 SNVs are repeated at least four times across the whole dataset, and 206,349 SNVs are reported to be identified in at least two cancer types (e.g. skin and breast cancers). As mentioned earlier, the best threshold provides

a classifier with a sufficient number of training examples while introducing the minimum bias. For the cancer somatic SNVs located in coding regions of the genome, we found the thresholds of four and two to properly meet the desired conditions. Accordingly, the numbers "four" and "two" constitute the dimensions of our final bi-dimensional score for positive examples in coding gold standard, indicating each SNV in the positive coding dataset has the frequency of four or more identified in at least two cancer types. The same explanation applies to the highlighted cells in **Table 4.3**, indicating that each of the positive SNVs in our non-coding dataset has the frequency of three or more identified in at least two cancer types. Similarly, these thresholds (three and two) are chosen with respect to the "sufficient examples with minimum bias" condition mentioned above.

Table 4.2 Number of positive examples per different thresholds in coding regions. The highlighted cells show the chosen threshold for each category. Our gold-standard includes the SNVs passing both filters

Threshold	Number of positive samples across whole dataset	Number of positive samples based on number of cancer types
1	3,569,793	3,318,128
2	798,856	206,349
3	261,833	30,294
4	122,054	6,999
5	67,028	2,601
6	42,328	1,484
7	31,437	977
8	25,624	674
9	21,739	494
10	19,238	372
11	17,299	309
12	15,816	249
13	14,671	191
14	13,868	141

Table 4.3 Number of positive examples per different thresholds in non-coding regions. The highlighted cells show the chosen threshold for each category. Our gold-standard includes the SNVs passing both filters

Threshold	Number of positive samples across whole dataset	Number of positive samples based on number of cancer types
1	16,966,056	16,796,255
2	1,405,383	156,816
3	266,704	11,117
4	115,861	1,507
5	58,561	266
6	31,454	49
7	20,039	9
8	13,897	4
9	8,760	2
10	3,875	2
11	2,464	2
12	1,701	2
13	1,198	2
14	905	2

4.1.1. Negative examples

Negative dataset includes the mutual SNVs between somatic point mutations with the minor allele frequency of equal or greater than 1% in at least one of 26 populations studied in 1000 Genomes project reported in dbSNP and the somatic SNVs reported in COSMIC database.

SNVs extracted from dbSNP guaranteed all the mutations involved in the final negative dataset were non-pathogenic and the SNVs extracted from COSMIC guaranteed that all the mutations in the final negative dataset were somatic. For the latter, this has not been examined in the negative examples used in previous tools (M. F. Rogers et al., 2017; Rogers et al., 2018; H. A. Shihab et al., 2015). We collected a total of 27,367,652 point mutations from dbSNP which was an integration of coding and non-coding variants. 16,594 of the SNVs from dbSNP were also

reported in COSMIC coding dataset and 58,995 SNVs were reported in COSMIC noncoding dataset (**Table 4.1**).

4.2. Genomic features

We characterized the mutations in our gold standard datasets by defining four major groups of genomic features summarized in **Table 3.1**. Subsequently we tried to extract the measurement of each feature for each mutation in both coding and non-coding gold-standards. Similar to almost any data collection-based study missing data were inevitable in our study. We were not able to find the measurements of all the defined features for all the SNVs in our gold standard datasets. We calculated the proportion of missing values for each feature in our coding and non-coding data sets (**Figure 4.1** and **Figure 4.2)**, concluding our sample size would shrink significantly (more than 5 %) in the event of deleting any SNVs that have any missing values. Accordingly, we decided that removing the missing data would not be an appropriate approach in handling missing data in our study. Therefore, imputation approaches were used for estimating the potential values of the missing data. However, to be more conservative we trained our prediction models both with and without inclusion of features with high range (more than 35 percent) of missing values. As shown in **Figure 4.1** and **Figure 4.2** the features with high proportion of missing values include "SIFTval","PolyPhenVal","Grantham","GerpRS", "GerpRSpval" in coding dataset and "GerpN","GerpS","EncodetotalRNA-max" in non-coding dataset.

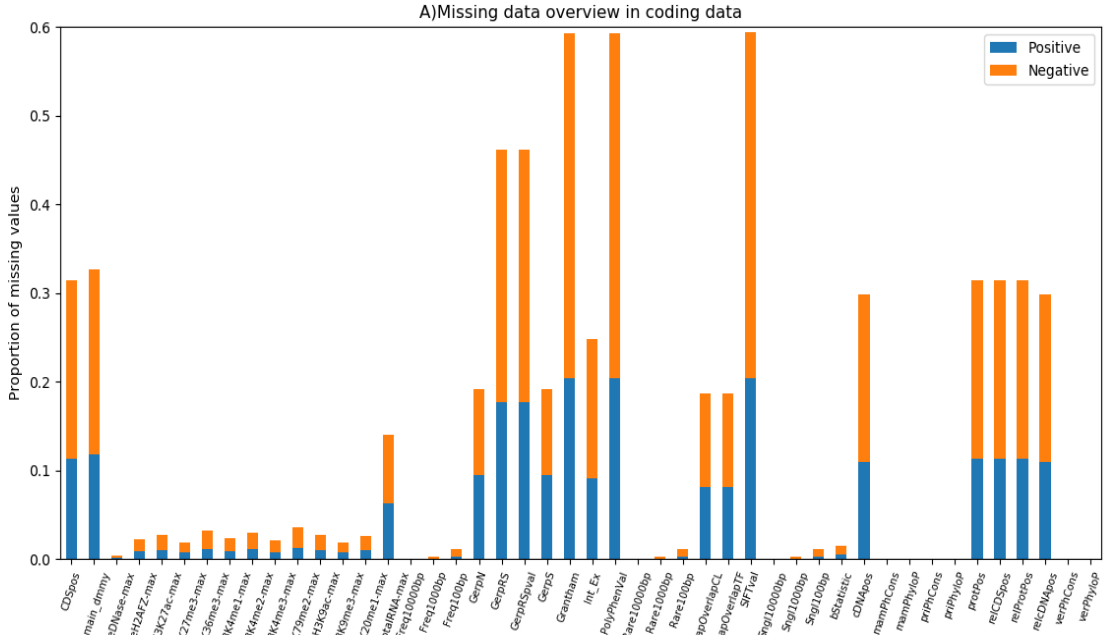

Figure 4.1 Proportion of missing data for different features in coding data from the gold standard

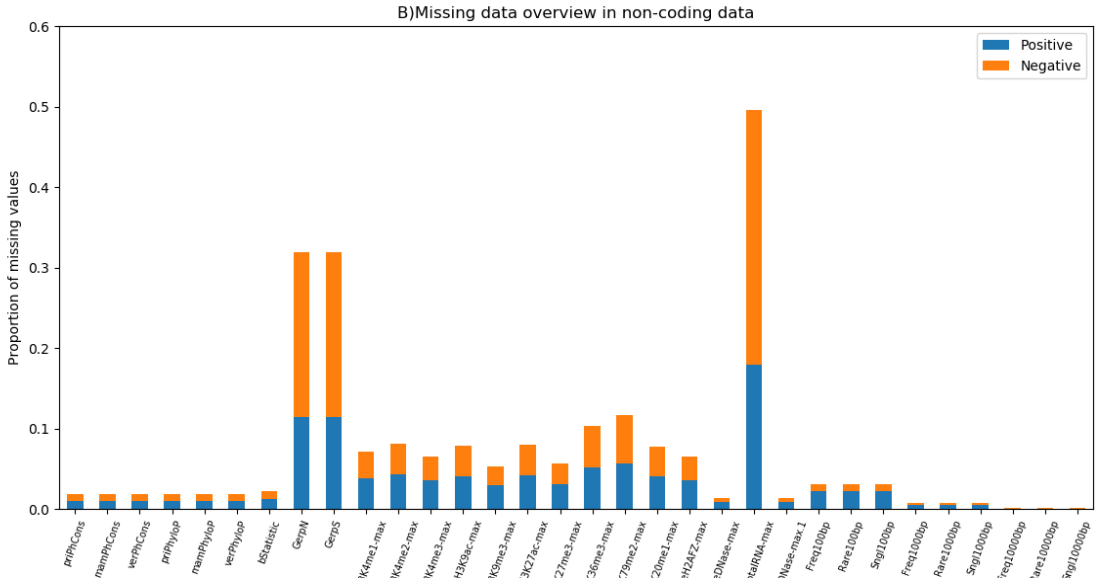

Figure 4.2 Proportion of missing data for different features in non-coding data from the gold standard

4.2.1. Most discriminative features

Identifying the features that contribute mostly to the discrimination power of a model is important in evaluating the model from biological aspects. These features that are chosen through feature selection process play the most critical role in distinguishing between different classes. In Lasso models contribution of each feature to the final classification outcome is reflected through the absolute value of the coefficients. The bigger the absolute value of the coefficient of a feature the more discriminative the feature. Accordingly, the coefficient of zero effectively implies that the feature is discarded through feature selection process and has not been used for model training.

Table 4.4 and **Table 4.5** represent the coefficients belonging to each feature from Lasso models built for coding and non-coding regions, respectively. The features are sorted vertically by the absolute value of their coefficient.

Table 4.4 indicates that structural and genomic context features including percentage of CpG islands and GC percentage in a +/- 75 window from a given mutation, and the number of single occurrence of SNVs (MAF<0.05) in a +/- 100 window from the given mutation are the most discriminative features as they have the biggest coefficient absolute values. Genomic distance features such as relative distance of a SNV from transcription start site, genomic conservation features such as scores form PhastCons and epigenetic features such as methylation modifications to protein histone H3 (H3K9me3) at 9^{th} lysine residue are the next discriminative feature groups respectively. As shown in **Table 4.4** two features including number of amino acid distance from coding start site and the p-values from GerpRS evolution scoring tool have not been used in the coding model as they have a coefficient of 0. It is noteworthy that both features are coding-specific features only defined for coding variants.

Table 4.4 Coefficient of features from the Lasso model of the coding regions. The absolute value of the coefficient denotes the importance of each feature in distinguishing between pathogenic and non-pathogenic groups. The coefficient of 0 implies that the feature is not selected in the feature selection process.

Feature	Coef	Group	Feature	Coef	Group	Feature	Coef	Group
CpG	5.430427368	Structural	cHmm_E17	0.02981	Epigenetic	Sngl100bp	-0.00439	Structural
GC	-2.152262409	Structural	cHmm_E22	0.028782	Epigenetic	cHmm_E14	-0.00437	Epigenetic
Freq100bp	-2.083440449	Structural	Freq1000bp	0.023307	Structural	H3K4me1	-0.00391	Epigenetic
Rare100bp	-0.700953376	Structural	cHmm_E19	-0.02256	Epigenetic	cHmm_E8	-0.00375	Epigenetic
relcDNApos	-0.482798812	Distance	H4K20me1	0.019239	Epigenetic	Domain_dmm	0.003591	Structural
Transversion	-0.452752695	Structural	cHmm_E11	0.018627	Epigenetic	H3K27me3	0.003292	Epigenetic
SIFTval	-0.444132285	Structural	DNase	0.018349	Epigenetic	Sngl1000bp	0.002419	Structural
verPhCons	-0.375352798	Conservation	cHmm_E21	0.017092	Epigenetic	cHmm_E15	-0.002	Epigenetic
priPhyloP	-0.346568505	Conservation	cHmm_E10	-0.017	Epigenetic	Grantham	-0.00193	conservation
Int_Ex	-0.260106506	Structural	cHmm_E25	-0.01388	Epigenetic	cHmm_E7	0.001807	Epigenetic
mamPhCons	-0.22078794	Conservation	H3K79me2	0.013789	Epigenetic	totalRNA	0.00168	Epigenetic
priPhCons	-0.17883215	Conservation	cHmm_E16	0.012791	Epigenetic	cHmm_E24	-0.00142	Epigenetic
Cons_dummy	-0.124538094	Structural	H2AFZ	0.011427	Epigenetic	Rare1000bp	-0.00058	Structural
relCDSpos	0.115313109	Distance	GerpS	-0.01135	Conservation	Rare10000bp	-0.0005	Structural
PHRED	0.112139281	Structural	cHmm_E6	-0.01035	Epigenetic	OverlapCL	0.000386	Structural
PolyPhenVal	0.110474063	Structural	cHmm_E9	-0.01029	Epigenetic	Sngl10000bp	0.000109	Structural
ConsScore	0.107390987	Structural	H3K27ac	0.009979	Epigenetic	bStatistic	-5.89E-05	Conservation
H3K9me3	0.092988449	Epigenetic	GerpN	-0.00941	Conservation	CDSpos	-3.06E-05	Distance
verPhyloP	0.073078108	Conservation	H3K9ac	-0.00799	Epigenetic	GerpRS	2.97E-05	Conservation
mamPhyloP	-0.062205422	Conservation	H3K4me2	-0.00697	Epigenetic	cDNApos	1.99E-05	Distance
cHmm_E23	-0.053494173	Epigenetic	H3K4me3	-0.00644	Epigenetic	protPos	-7.00E-06	Distance
cHmm_E1	0.04944372	Epigenetic	Freq10000bp	-0.00613	Structural	Dist2Mutation	5.21E-06	Distance
cHmm_E13	0.046496532	Epigenetic	cHmm_E12	0.005691	Epigenetic	minDistTSS	9.78E-07	Distance
cHmm_E18	-0.040574404	Epigenetic	cHmm_E2	-0.00531	Epigenetic	minDistTSE	-4.95E-07	Distance
cHmm_E3	0.037666804	Epigenetic	OverlapTF	-0.00487	Structural	relProtPos	0	Distance
cHmm_E5	0.035345748	Epigenetic	H3K36me3	0.00477	Epigenetic	GerpRSpval	0	Conservation
cHmm_E4	0.031469897	Epigenetic	cHmm_E20	-0.00445	Epigenetic			

Regarding **Table 4.5** in terms of possessing the most discriminative feature members almost the same trend as the coding model is observed among the feature groups in the non-coding model. As indicated in **Table 4.5** four of five of the top features belong to "structural and genomic context" feature group, including percentage of CpG islands in a +/- 75 window from the mutation, number of single occurrence of the SNVs (MAF<0.05) in +/- 100 window from the mutation, transverion/transition identity of the nucleotide change, and GC percentage in a +/- 75 window from the mutation. Conservation features such as scores from Phylop and epigenetic features such as methylation modifications to protein histone H3 (H3K36me3) at 36th lysine residue were ranked

as the second and third for possessing most discriminative feature members. Genomic distance features such as distance of the SNV from transcribed sequence start/end start are shown to have the least contribution to the discrimination power of the non-coding model. Interestingly, unlike the coding model, there are no features assigned with the coefficient of zero in the non-coding model.

Table 4.5 Coefficient of features from the Lasso model in the non-coding regions. The absolute value of the coefficient denotes the importance of each feature in distinguishing between pathogenic and non-pathogenic groups.

Feature	Coef	Group	Feature	Coef	Group	Feature	Coef	Group
CpG	4.061880549	Structural	cHmm_E2	0.019548838	Epigenetics	cHmm_E14	0.005800379	Epigenetics
Freq100bp	1.690500049	Structural	H3K4me1	0.017862322	Epigenetics	H3K79me2	0.005246062	Epigenetics
Transversion	0.640151714	Structural	H3K4me3	0.017224207	Epigenetics	cHmm_E23	0.005154772	Epigenetics
GC	0.597857517	Structural	cHmm_E25	0.016014631	Epigenetics	GerpS	0.004004753	Conservation
verPhCons	0.469939117	Conservation	PHRED	0.015900759	Structural	totalRNA-max	0.003576802	Epigenetics
Rare100bp	0.423388778	Structural	cHmm_E10	0.015729031	Epigenetics	cHmm_E8	0.002611941	Epigenetics
ConsScore	0.310816596	Structural	cHmm_E12	0.014548169	Epigenetics	cHmm_E7	0.00194307	Epigenetics
priPhCons	0.224622944	Conservation	Freq1000bp	0.01333975	Structural	H3K27ac	0.001560174	Epigenetics
mamPhCons	0.200537783	Conservation	cHmm_E21	0.012206939	Epigenetics	Sngl1000bp	0.001515173	Structural
priPhyloP	0.187248332	Conservation	cHmm_E9	0.011978767	Epigenetics	cHmm_E16	0.001471679	Epigenetics
Cons_dummy	0.179209493	Structural	cHmm_E22	0.011843811	Epigenetics	cHmm_E3	0.001347666	Epigenetics
cHmm_E1	0.071626125	Epigenetics	cHmm_E15	0.011703168	Epigenetics	Rare10000bp	0.001048594	Structural
H3K36me3	0.070601559	Epigenetics	cHmm_E11	0.010543743	Epigenetics	DNase	0.000579038	Epigenetics
mamPhyloP	0.066234151	Conservation	cHmm_E24	0.010133424	Epigenetics	H3K4me2	0.000502246	Epigenetics
verPhyloP	0.056137415	Conservation	H3K9ac	0.009987905	Epigenetics	cHmm_E5	0.00043036	Epigenetics
H3K9me3	0.045100854	Epigenetics	cHmm_E19	0.009336748	Epigenetics	cHmm_E4	0.000227864	Epigenetics
H2AFZ	0.03795365	Epigenetics	cHmm_E20	0.008293349	Epigenetics	Sngl10000bp	0.000141943	Structural
cHmm_E13	0.024148465	Epigenetics	Rare1000bp	0.007715412	Structural	bStatistic	0.0000823	Conservation
cHmm_E18	0.023008619	Epigenetics	cHmm_E17	0.007030198	Epigenetics	Dist2Mutation	0.0000225	Distance
H3K27me3	0.022805297	Epigenetics	GerpN	0.006913753	Conservation	minDistTSS	0.000000379	Distance
cHmm_E6	0.021209513	Epigenetics	Freq10000bp	0.00635071	Structural	minDistTSE	0.000000304	Distance
H4K20me1	0.020029121	Epigenetics	Sngl100bp	0.005881123	Structural			

4.3. Classification models

4.3.1. Model selection and visualization

Determining the model parameters as well as data preparation approaches that best exploit the features to explain the gold standard data is a fundamental task in model building process. We applied our two model building methods (LASSO and SVM) to four data preparation conditions including 1) normalized data including all imputed features, and 2) normalized data excluding features with high proportion (>35%) of missing data, 3) non-normalized data including all imputed features, and 4) non-normalized data excluding features with high proportion (> 35%) of missing data. All the conditions are tested through both Training-test (2/3 training, 1/3 testing) and 10-fold cross-validation (10F CV) settings. In addition, regarding the SVM models we tried "sigmoid" and "Gaussian" (rbf) kernels to find the best kernel fitting our data. The results of the model evaluations are provided in **Table 4.6** and **Table 4.7** for coding and non-coding data sets, respectively.

Table 4.6: Summary of the model performance in the coding regions in different data preparation settings and inclusion/exclusion of features with more than 35% of missing data. In the Data-proportion column "All" implies the inclusion of all the defined features in model training process and "dropped" implies the removal of the features with more than 35% of missing data. "Norm" column indicates whether the analysis were conducted for normalized data or not ("Yes" stands for normalized and "No" stands for non-normalized data). Train/Test column shows the model evaluation strategy (10- F CV stands for 10-fold cross-validation and 2/3 stands for Training (2/3) – test (1/3) setting).

Coding					
Model	Data-portion	Norm	Training/Test	AUC	Accuracy
Lasso	All	No	10F CV	0.88	0.81
Lasso	All	No	2/3	0.88	0.80
Lasso	All	Yes	10F CV	0.88	0.81
Lasso	All	Yes	2/3	0.89	0.81
Lasso	Dropped	No	10F CV	0.88	0.81
Lasso	Dropped	No	2/3	0.88	0..80
Lasso	Dropped	Yes	10F CV	0.88	0.81
Lasso	Dropped	Yes	2/3	0.87	0.80
SVM-"rbf"	All	No	10F CV	0.54	0.68
SVM-"rbf"	All	No	2/3	0.53	0.68
SVM-"rbf"	All	Yes	10F CV	0.94	0.88
SVM-"rbf"	All	Yes	2/3	0.93	0.87
SVM-"rbf"	Dropped	No	10F CV	0.54	0.68
SVM-"rbf"	Dropped	No	2/3	0.53	0.68
SVM-"rbf"	Dropped	Yes	10F CV	0.94	0.88
SVM-"rbf"	Dropped	Yes	2/3	0.94	0.88
SVM-"sig"	All	No	10F CV	0.5	0.67
SVM-"sig"	All	No	2/3	0.5	0.69
SVM-"sig"	All	Yes	10F CV	0.66	0.67
SVM-"sig"	All	Yes	2/3	0.66	0.67
SVM-"sig"	Dropped	No	10F CV	0.5	0.67
SVM-"sig"	Dropped	No	2/3	0.5	0.67
SVM-"sig"	Dropped	Yes	10F CV	0.66	0.67
SVM-"sig"	Dropped	Yes	2/3	0.66	0.67

Table 4.7: Summary of the model performance in the noncoding regions in different data preparation settings and inclusion/exclusion of features with more than 35% of missing data. In the Data-proportion column "All" implies the inclusion of all the defined features in model training process and "dropped" implies the removal of the features with more than 35% of missing data. "Norm" column indicates whether the analysis were conducted for normalized data or not ("Yes" stands for normalized and "No" stands for non-normalized data). Train/Test column shows the model evaluation strategy (10- F CV stands for 10-fold cross-validation and 2/3 stands for Training (2/3) – test (1/3) setting).

Non-Coding					
Model	Data-portion	Norm	Train/Test	AUC	Accuracy
Lasso	All	No	10F CV	0.83	0.78
Lasso	All	No	2/3	0.84	0.78
Lasso	All	Yes	10F CV	0.83	0.78
Lasso	All	Yes	2/3	0.83	0.78
Lasso	Dropped	No	10F CV	0.83	0.78
Lasso	Dropped	No	2/3	0.83	0.78
Lasso	Dropped	Yes	10F CV	0.83	0.78
Lasso	Dropped	Yes	2/3	0.83	0.78
SVM-"rbf"	All	No	10F CV	0.54	0.68
SVM-"rbf"	All	No	2/3	0.53	0.68
SVM-"rbf"	All	Yes	10F CV	0.89	0.88
SVM-"rbf"	All	Yes	2/3	0.89	0.87
SVM-"rbf"	Dropped	No	10F CV	0.54	0.68
SVM-"rbf"	Dropped	No	2/3	0.53	0.68
SVM-"rbf"	Dropped	Yes	10F CV	0.89	0.88
SVM-"rbf"	Dropped	Yes	2/3	0.89	0.88
SVM-"sig"	All	No	10F CV	0.5	0.67
SVM-"sig"	All	No	2/3	0.5	0.67
SVM-"sig"	All	Yes	10F CV	0.66	0.67
SVM-"sig"	All	Yes	2/3	0.66	0.67
SVM-"sig"	Dropped	No	10F CV	0.5	0.67
SVM-"sig"	Dropped	No	2/3	0.5	0.67
SVM-"sig"	Dropped	Yes	10F CV	0.66	0.67
SVM-"sig"	Dropped	Yes	2/3	0.66	0.67

As indicated in **Table 4.6** and **Table 4.7** Lasso models show the similar accuracy for both normalized and non-normalized data while SVM models have a considerably better performance in the event of normalized data. Furthermore, for both Lasso and SVM models deletion of the features with high proportion of missing values (>35%) does not seem to have a noticeable effect

on models performance. This can be explained by small impact of these features in models classification in addition to the accurate imputation of the missing data.

Similarity of the models performance under 10-fold cross-validation and training-test settings implies the homogeneity of the data. Based on the models accuracy and AUC results indicated in **Table 4.6** and **Table 4.7** SVM models with "Gaussian" kernel applied to the normalized data yields the highest AUC (0.94 for the coding and 0.89 for the noncoding models) and accuracy (0.88 for the coding and for the noncoding 0.87 models) among all the model training conditions for the both coding and non-coding regions. We narrowed down our subsequent model analysis to Lasso and SVM methods with "Gaussian" kernel for normalized data.

To visualize the model performance, we plotted the ROC curves of the final candidate Lasso and SVM models using the normalized data. **Figure 4.3** and **Figure 4.4** indicate the ROC curves under 10-fold cross-validation for the coding and non-coding regions, respectively.

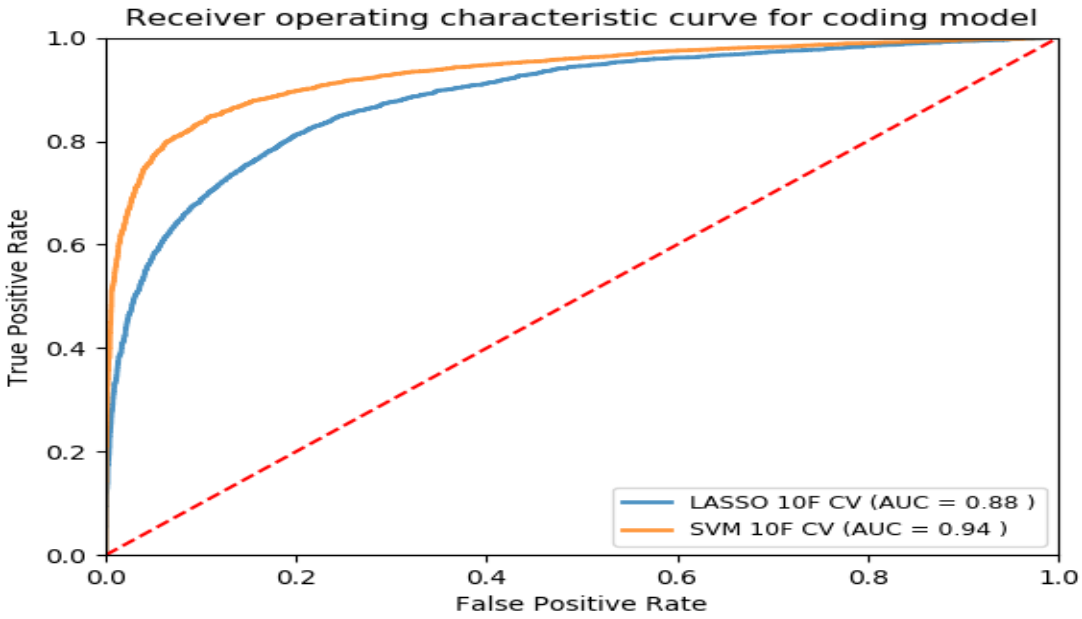

Figure 4.3 ROC curves of the models designed for classifying cancer somatic variants from coding regions of the genome.

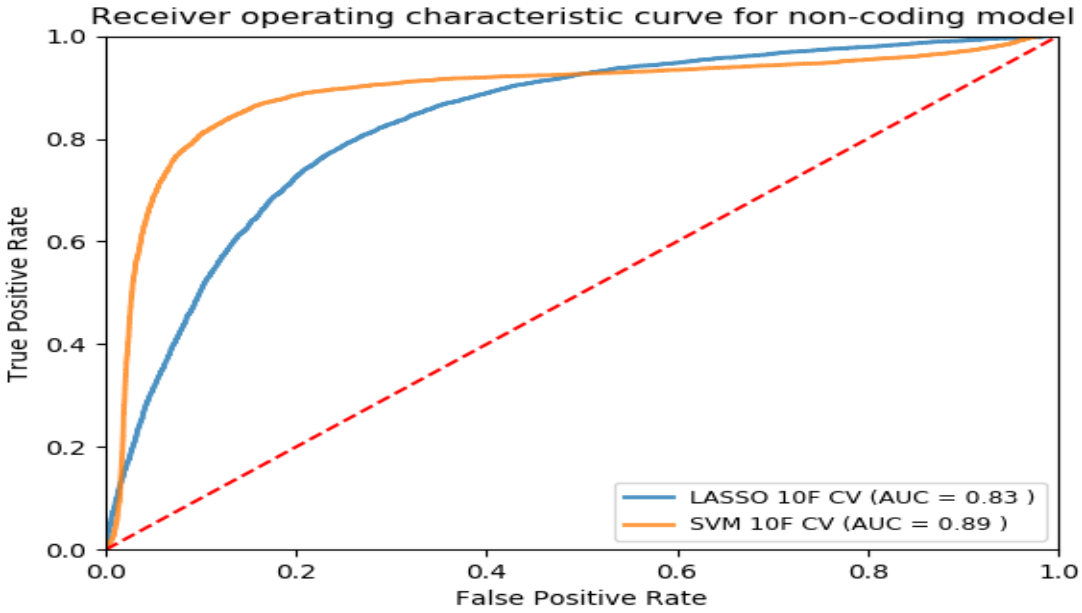

Figure 4.4 ROC curves of the models designed for classifying cancer somatic variants from non-coding regions of the genome.

As shown in **Figure 4.3** and **Figure 4.4** SVM models yield a higher AUC values (0.94 and 0.89 for the coding and non-coding regions, respectively) using10-fold cross-validation. Therefore, we chose the SVM method with Gaussian kernel as the final training approach in the model building process.

Classification models return a continuous probability value (from 0 to 1) which needs to be mapped to a binary category (e.g. pathogenic or non-pathogenic) using a classification threshold (also known as decision threshold) for the coding and noncoding regions, respectively. ROC curves can be beneficial in identifying the best classification threshold (also called optimum threshold) yielding the highest true positive while lowest false positive results. **Table 4.8** and **Table 4.9** represent the values of the true positive rate (tpr) and the false positive rate (fpr) per different classification thresholds for our coding and noncoding region-based models respectively. We chose the classification threshold of 0.55 as the optimum threshold for our coding region-based model, reaching the tpr of 0.80 and fpr of 0.06 (**Table 4.8**). To acquire the same tpr by our noncoding region-based model we chose the optimum threshold of 0.41, reaching the tpr of 0.80 and fpr of 0.09 (**Table 4.9**).

Table 4.8 True positive and false positive rates of the coding region-based models at different prediction thresholds. The selected threshold yielding the highest true positive rate (0.80) is highlighted in the table.

Threshold	True positive rate (TPR)	False positive rate (FPR)
0.55	0.80	0.06
0.60	0.78	0.05
0.65	0.75	0.04
0.70	0.73	0.04
0.75	0.70	0.03
0.80	0.65	0.02
0.85	0.60	0.01

Table 4.9 True positive and false positive rates of the noncoding region-based models at different prediction thresholds. The selected threshold yielding the highest true positive rate (0.80) is highlighted in the table.

Threshold	True positive rate (TPR)	False positive rate (FPR)
0.41	0.80	0.09
0.45	0.77	0.08
0.50	0.75	0.07
0.55	0.71	0.06
0.60	0.67	0.05
0.65	0.62	0.04
0.70	0.57	0.03

4.3.2. Model comparison

We compared the performance of our final SVM models in terms of their discrimination power in the task of classifying cancer somatic variants with two of the leading classification models including FATHMM-XF and Cscape. To be consistent with our model comparison, we applied the same test set used in evaluating our SVM models for coding and noncoding regions to these two models. Using FATHMM web server (http://fathmm.biocompute.org.uk/) we were able to apply FATHMM-XF and Cscape to the test set. As demanded by FATHMM web server, we were required to first convert the genomic coordinates of the selected dataset from GRCh38 to GRCh37 and subsequently upload a VCF format (Danecek et al., 2011) of the variants to the classification tools. We compared the prediction results from the two tools with the actual labels and investigated the rate of the true positive predictions against false positive predictions by plotting ROC curves shown in **Figure 4.5** and **Figure 4.6** for the coding and noncoding regions, respectively.

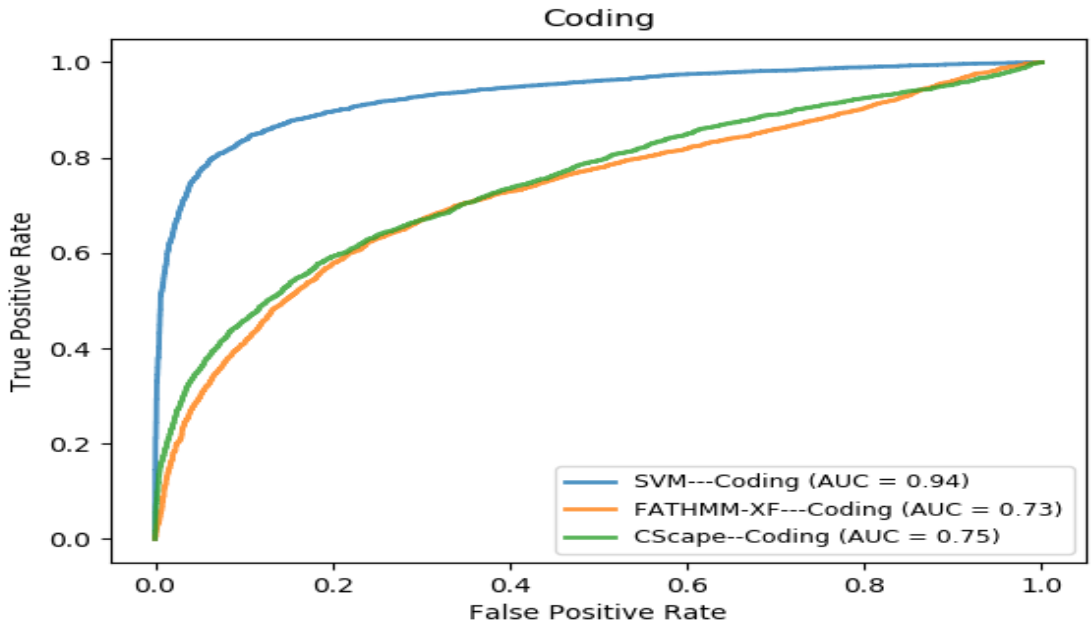

Figure 4.5 ROC curves comparing the performance of our model (SVM) with FATHMM-XF and CScape for somatic cancer variants in coding regions of the genome

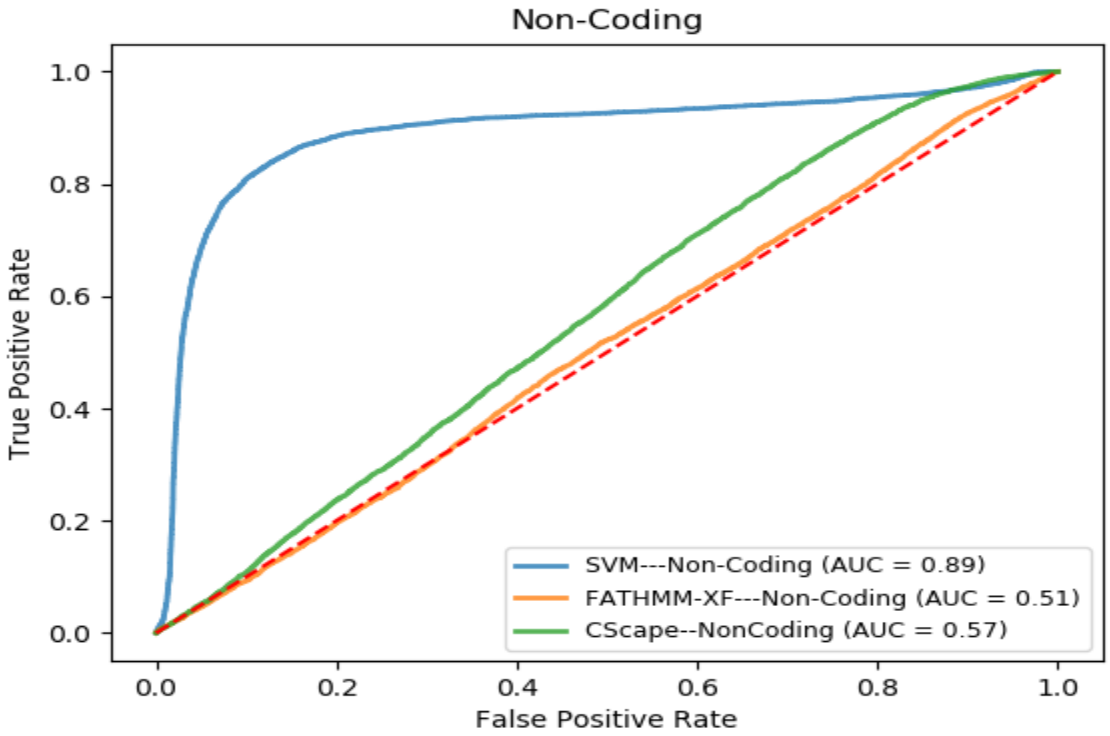

Figure 4.6 ROC curves comparing the performance of our model (SVM) with FATHMM-XF and CScape for somatic cancer variants in Noncoding regions of the genome

The dramatically higher AUCs of our SVM models shown in **Figure 4.5** and **Figure 4.6** suggest that these models outperform the competitors for both coding and noncoding regions. It is notable that in consistent with our hypothesis, in the event of classifying cancer somatic variants, Cscape, which is designed based on the characteristics of a mixture of somatic-germline cancer variants, yielded a better performance compared to FATHMM-XF which is originally designed for germline variants.

4.4. Making predictions by models

Predicting the pathogenic status of somatic point mutations identified in breast cancer patients under 45 years of age was done by applying our models to the set of SNVs extracted from both METABRIC and TCGA-BRCA studies.

4.4.1 METABRIC

Excluding the SNVs from patients over 45 years of age our METABRIC data consisted of 1882 cancer somatic SNVs from 326 patients. Regarding the data extraction methods used in METABRIC study (whole exome sequencing of the 173 candidate genes – see Methods) these SNVs are all "coding mutations" located in coding regions of the genome. The results from applying our SVM model designed for coding region to the METABRIC data are shown in **Table 4.10**.

Table 4.10 Number of pathogenic positive SNVs predicted by the SVM model for the coding regions with the number of their associated genes per different cut-offs and recurrence frequency for METABRIC dataset. The highlighted row shows the selected cut-off for subsequent analysis.

Prediction Cut-off	No. of pathogenic positive predictions	No. of affected genes	The frequency of pathogenic SNVs >=2		The frequency of pathogenic SNVs >=3		The frequency of pathogenic SNVs >=4	
			No. of SNVs	No. affected of Genes	No. of SNVs	No. of affected Genes	No. of SNVs	No. of affected Genes
0.55	959	154	52	18	17	5	12	3
0.60	880	147	49	18	14	4	10	3
0.65	795	143	48	18	14	4	10	3
0.70	702	138	43	15	14	4	10	3
0.75	605	130	40	15	14	4	10	3
0.80	500	116	32	11	12	4	9	3
0.85	365	105	17	7	6	3	3	2

Table 4.10 shows an overview of the number of SNVs predicted as pathogenic as well as the number of genes affected by these SNVs per different classification thresholds from METABRIC dataset. Using the selected optimum threshold for our coding model (0.55) 959 SNVs are predicted as pathogenic affecting 154 genes of which 18 are affected by 52 SNVs with the frequency of equal or more than two across the entire data set. The number of affected genes by SNVs with recurrence frequency of bigger than three and four is also indicated in **Table 4.10**. To narrow down the number of pathogenic SNVs to a reasonable number suitable for conducting subsequent analysis we extracted the 500 SNVs labeled as pathogenic using threshold of 0.8.

4.4.2 TCGA-BRCA

Considering the data collection approaches used in TCGA-BRCA project (including whole genome and microRNA sequencing) we were able to identify somatic SNVs from both coding and non-coding regions of the genome.

Our final TCGA-BRCA dataset involved a total of 8,647 somatic SNVs from 142 young patients (<45 years old) including 6910 and 1737 somatic SNVs from coding and noncoding regions of the genome, respectively. The results from applying our models to somatic SNVs form TCGA-BRCA data from coding and noncoding regions of the genome is presented in **Table 4.11** and **Table 4.12**, respectively.

Table 4.11 Number of pathogenic positive SNVs predicted by the SVM model for the coding regions with the number of their associated genes per different cut-offs and recurrence frequency for TCGA coding dataset. The highlighted row shows the selected cut-off for subsequent analysis.

Prediction Cut-off	No. of pathogenic positive predictions	No. of affected genes	The frequency of pathogenic SNVs >=2		The frequency of pathogenic SNVs >=3		The frequency of pathogenic SNVs >=4	
			No. of SNVs	No. affected of Genes	No. of SNVs	No. of affected Genes	No. of SNVs	No. of affected Genes
0.55	3510	2537	232	184	6	2	4	2
0.60	3213	2354	211	165	6	2	4	2
0.65	2915	2179	191	150	6	2	4	2
0.70	2599	1970	168	133	6	2	4	2
0.75	2274	1745	146	116	6	2	4	2
0.80	1858	1473	118	94	6	2	4	2
0.85	1394	1136	88	75	4	2	2	1

As indicated in **Table 4.11** using the optimum threshold of 0.55 our "coding model" predicted 3510 pathogenic somatic SNVs which affect 2537 genes. The count of SNVs with higher recurrence frequencies and their related genes are also presented in **Table 4.11**. To be consistence with our filtering approach regarding METABRIC data we used the threshold of 0.8 to narrow down our subsequent analysis of TCGA-BRCA data to 1858 SNVs predicted as pathogenic by this threshold.

Table 4.12 Number of pathogenic positive SNVs predicted by the SVM model for the coding regions with the number of their associated genes per different cut-offs and recurrence frequency for TCGA noncoding dataset. The highlighted row shows the selected cut-off for subsequent analysis.

Prediction Cut-off	No. of pathogenic positive predictions	No. of affected genes	The frequency of pathogenic SNVs >=2		The frequency of pathogenic SNVs >=3		The frequency of pathogenic SNVs >=4	
			No. of SNVs	No. affected of Genes	No. of SNVs	No. of affected Genes	No. of SNVs	No. of affected Genes
0.41	943	331	58	25	0	0	0	0
0.45	865	316	53	23	0	0	0	0
0.50	744	291	45	19	0	0	0	0
0.55	633	272	40	19	0	0	0	0
0.60	532	259	31	18	0	0	0	0
0.65	435	238	23	14	0	0	0	0
0.70	343	227	20	14	0	0	0	0

As indicated in **Table 4.12** using the threshold of 0.41 which is the optimum threshold for the "non-coding" model (designed based on the features of non-coding mutations) 943 somatic SNVs were labeled as pathogenic. Even though these SNVs are located in non-coding regions of the genome they are commonly located in regulatory elements potential for affecting the protein coding regions of the genome. In **Table 4.12** we have indicated the number of genes likely to be affected by somatic SNVs predicted as pathogenic from non-coding regions of the genome. Again we selected a smaller proportion of SNVs labeled as pathogenic using the threshold of 0.65 which provides us with a sufficient number of SNVs and genes facilitating our subsequent analysis.

4.5. Potential prognostic genes for breast cancer

To identify the potential prognostic genes for breast cancer, we investigated the genes affected by the SNVs predicted as pathogenic positive by our models from the prediction datasets. The genes were either affected by the mutations changing their nucleotide sequence or affected by the mutations in their regulatory elements (e.g. mutations in noncoding regions of the genome). To collect the mentioned genes, we first counted the recurrence frequency of pathogenic positive SNVs across the whole dataset (reported in **Table 4.10**, **Table 4.11**, **Table 4.12** for different prediction datasets) and subsequently collected the genes that were affected by recurrent pathogenic positive SNVs. For instance, if gene A is mutated by mutation X, which is repeated more than once across the whole dataset, we regarded gene A as being affected by a recurrent mutation. **Table 4.13** lists the genes that were affected by pathogenic positive mutations with recurrence frequency of four or more in coding datasets (METABRIC and TCGA-BRCA coding) as well as the genes affected by the pathogenic positive mutations with recurrence frequency of two across the TCGA-BRCA noncoding dataset.

Table 4.13 An overview of the genes harboring the recurrent pathogenic positive SNVs predicted by our models. The "SNV ID" column shows the ID of the recurrent SNV that affects the gene mentioned in "Gene" column. "Ref" column shows the nucleotide in the reference genome sequence and "Alt" column shows in the alternative nucleotide that is substituted for the reference nucleotide. The highlighted row shows the SNV appeared as significant through our subsequent survival analysis.

Cohort	Gene	SNV ID	Ref	Alt	SNV Position (Chr: Position (base pair: GRCh38))	SNV consequence from VEP Ensembl
METABRIC	AKT1	14:104780214_C>T	C	T	14:104780214	Missense
METABRIC	PIK3CA	3:179203765_T>A	T	A	3:179203765	Missense
METABRIC	PIK3CA	3:179218294_G>A	G	A	3:179218294	Missense
METABRIC	PIK3CA	3:179218303_G>A	G	A	3:179218303	Missense
METABRIC	PIK3CA	3:179234297_A>T	A	T	3:179234297	Missense
METABRIC	TP53	17:7673802_C>T	C	T	17:7673802	Missense
METABRIC	TP53	17:7674220_C>T	C	T	17:7674220	Missense
METABRIC	TP53	17:7674221_G>A	G	A	17:7674221	Missense
METABRIC	TP53	17:7675088_C>T	C	T	17:7675088	Missense
TCGA-CD	PIK3CA	3:179203765_T>A	T	A	3:179203765	Missense
TCGA-CD	PIK3CA	3:179218294_G>A	T	A	3:179218294	Missense
TCGA-CD	PIK3CA	3:179218303_G>A	G	A	3:179218303	Missense
TCGA-CD	TP53	17:7675088_C>T	C	T	17:7675088	Missense
TCGA-NC	ZFP30	19:37613150_G>A	G	A	19:37613150	Missense
TCGA-NC	CLIC3	9:136993900_A>C	A	C	9:136993900	Regulatory_region_SNV
TCGA-NC	AC211476.2	7:72926895_G>C	G	C	7:72926895	Missense
TCGA-NC	ZNF512	2:27578227_C>T	C	T	2:27578227	Missense
TCGA-NC	KRTAP19-11P	21:30541689_G>A	G	A	21:30541689	Missense
TCGA-NC	AL034345.1	6:38924007_C>G	C	G	6:38924007	Missense
TCGA-NC	PGAM1P6	2:23869699_C>A	C	A	2:23869699	Missense
TCGA-NC	ZDHHC11B	5:711218_G>C	G	C	5:711218	Noncoding_exon_SNV
TCGA-NC	AC120498.10	16:1220974_G>A	G	A	16:1220974	Missense
TCGA-NC	RF00092	1:37880149_C>G	C	G	1:37880149	Missense
TCGA-NC	MIR519A2	19:53761153_G>A	G	A	19:53761153	Mature miRNA variant
TCGA-NC	AL049555.1	6:54941625_C>T	C	T	6:54941625	Missense
TCGA-NC	PLIN5	19:4538646_C>T	C	T	19:4538646	Missense
TCGA-NC	CDC27P1	2:132257729_T>G	T	G	2:132257729	Noncoding_exon_SNV

To validate our models' performances we investigated the potential role of the genes reported in **Table 4.13** in breast cancer occurrence. It is noteworthy that as indicated in **Table 4.12** in TCGA-BRCA noncoding dataset 14 genes were affected by pathogenic positive SNVs with the recurrence frequency of two across the dataset. These genes mostly belong to three major

categories including pseudogenes (KRTAP19-11P, AL034345.1, PGAM1P6, CDC27P1), RNA genes (AC211476.2, AC120498.10, RF00092, MIR519A2, AL049555.1), and members of zinc finger gene family (ZNF512, ZFP30, ZDHHC11B).

TP53 (Tumor protein p53)

TP53 is a well-known tumor suppressor that responds to various stress signals by regulating the specific cellular responses that are associated with tumor suppression (Bieging, Mello, & Attardi, 2014). TP53 (also known as P53), is a member of TP53 gene family, known as the most commonly affected tumor suppressor gene in human cancer (Béroud & Soussi, 2003). Conditions such as hypoxia, DNA damage or aberrant oncogene expression leads to activation of P53 gene which promotes cell-cycle checkpoints, cellular senescence, DNA repair and apoptosis (Fridman & Lowe, 2003) which is the self-destruction of cells in the event of severe damage or once they are no longer needed (Elmore, 2007). According to recent studies P53 also plays a role in communication within the tumor microenvironment as well as stem cell maintenance, invasion and metastasis (Bieging et al., 2014). It is demonstrated that disruption of P53 function leads to cell-cycle arrest defects, genome instability, cellular immortalization and continued proliferation and evolution of damaged cells (Fridman & Lowe, 2003).

Mutations in TP53 (both germline and somatic) are demonstrated to be associated with the advent of various cancers including breast, brain, adrenal cortical and etc.(Huszno & Grzybowska, 2018). TP53 mutations are regarded as negative prognostic factors in breast cancer (Varna, Bousquet, Plassa, Bertheau, & Janin, 2011). Breast tumors with mutations in TP53 have been shown to be more aggressive types (e.g. triple negative and HER-2 positive) (Langerød et al.,

2007; Y. Wang et al., 2004) with poor response to radiotherapy and chemotherapy (Chae et al., 2009; Wattel et al., 1994). Furthermore, it is reported that there is an association between early-onset of breast cancer and TP53 germline mutations (Srivastava, Zou, Pirollo, Blattner, & Chang, 1990).

AKT1 (RAC-alpha serine/threonine-protein kinase)

The gene AKT1 is a member of the AGC Serine/Threonine protein kinase family belonging to a class known as oncogenes whose mutation has the potential to convert a normal cell into a malignant cancerous cell (Riggio et al., 2017). AKT1 regulates many cellular process including cell proliferation, growth and survival as well as the process of angiogenesis. AKT1 also helps in controlling apoptosis (Riggio et al., 2017).

It is frequently reported that over activation of AKT1 results in increased mammary tumor growth (Dillon et al., 2009; María Laura Polo et al., 2015; Riggio et al., 2011). Furthermore, PI3K/AKT/mTOR is known as the most commonly deregulated pathway in many solid tumors including breast cancer (Cheng, Lindsley, Cheng, Yang, & Nicosia, 2005; Kreisberg et al., 2004). It is shown that over-activation of the pathway is associated with growth factor independent cell proliferation (Toker, 2012), endocrine receptor deregulation (Maria Laura Polo et al., 2010; Riggio et al., 2011), cell invasion (Vandermoere et al., 2007) and resistance to therapy (Brown & Toker, 2015). Accordingly, mutations in AKT and PI3K can lead to deregulation of the pathway however the mechanism and downstream signals that regulate each step of tumor development are not fully understood.

PIK3CA (PHOSPHATIDYLINOSITOL 3-KINASE, CATALYTIC, ALPHA)

PIK3CA which is a member of PI3/PI4-kinase family produces the 110 (kDA) alpha protein, the catalytic subunit of PIK3 enzyme. PIK3 is a key element of PI3K/AKT/mTOR signaling pathway regulating cell growth and proliferation (S. Kang, Bader, & Vogt, 2005). PIK3CA responds to various growth factors by activating signaling cascades that are involved in cell survival, growth, proliferation, morphology and motility. Accordingly, mutations in PIK3CA can induce oncogenic transformation of the normal cells (Elkabets et al., 2013; Riggio et al., 2017). Somatic mutations in PIK3CA are reported to be frequently observed in different human primary tumors specially colon, breast and stomach cancers (Samuels et al., 2004). Interestingly, Anders et al suggested that based on their study (Anders, Acharya, et al., 2008) there is higher chances of PI3K pathway deregulation in young breast cancer patients compared to their older counterparts.

Zinc finger genes

ZNF512, ZNF30 and ZDHHC11B all encode proteins belonging to the zinc finger protein family which is the largest transcription factors family in the human genome (Jen & Wang, 2016). Transcription factors coordinate numerous cellular processes such as differentiation, development, apoptosis, etc. (Arenzana, Schjerven, & Smale, 2015; Krebs, Schultz, & Robins, 2012; Ma et al., 2016). It is frequently reported that mutations resulting in aberrant expression of zinc finger proteins contribute to different types of cancers such as breast, ovarian, colorectal, etc. (Aslan et al., 2015; Horiguchi et al., 2012; Serra, Fang, Park, Hutchinson, & Green, 2014). In consistent with our results Heiser et al have reported ZNF30 as one of the genes with aberrant expression in breast cancer cell lines relative to other cell lines (Heiser et al., 2012). In addition, Bao et al

identified ZNF512 as one of the nine genes that are significantly up regulated in lung adenocarcinoma (Bao et al., 2016).

Pseudogenes

Pseudogenes have been generally considered as the imperfect copy of coding genes that have lost their functionality and protein coding ability (L. Poliseno, Marranci, & Pandolfi, 2015). Recently, it has been demonstrated that many pseudogenes are also translated and transcribed as well (Djebali et al., 2012; M.-S. Kim et al., 2014). Several studies (D'Errico, Gadaleta, & Saccone, 2004; Laura Poliseno, 2012; Vihinen, 2014) have suggested the importance of pseudogenes in the context of cancer development. Poliseno et al suggested a causal association between cancer initiation and altered expression of pseudogenes (L. Poliseno et al., 2015). Kalyana-Sundaram et al (Kalyana-Sundaram et al., 2012) identified cancer type-specific pseudogenes only expressed in one cancer type (e.g. breast cancer). Interestingly, Han et al (Han et al., 2014) revealed subtype-specific expression of pseudogenes through RNA-seq analysis of different breast cancer subtypes. In conclusion, pseudogenes are considered as potential diagnostic cancer biomarkers contributing to coding-dependent and coding-independent regulatory networks (L. Poliseno et al., 2015)

RNA genes

RNA genes encode noncoding RNA (ncRNA) molecules that are not translated to proteins but can strongly contribute to different diseases, particularly to cancer (T. Huang, Alvarez, Hu, & Cheng, 2013). Some ncRNAs (e.g. long noncoding RNAs, micro RNAs, and etc.) play an important role in crucial cellular processes including transcription, post-transcription modifications, and translation (Cech & Steitz, 2014). It has been demonstrated that ncRNAs regulate intercellular and intracellular signaling in breast cancer (Klinge, 2018). The intracellular

activities of ncRNAs in breast cancer expands a wide range of cellular processes from regulating the activity and level of estrogen receptor α (ERα) to controlling cell division, stemness (Z.-Y. Wang & Yin, 2015), migration, invasion (Cai, He, & Zhang, 2015) and apoptosis (Vendramin, Marine, & Leucci, 2017). Furthermore, ncRNAs participate in intracellular communication by being packaged into extracellular vesicles and transferred to recipient cells (Sun et al., 2018). Interestingly, in consistent with our study, ncRNAs are frequently referred to as promising biomarkers for detecting early and advanced breast cancer (Asaga et al., 2011; Heneghan, Miller, Kelly, Newell, & Kerin, 2010; Lo et al., 2016; Schwarzenbach, Milde-Langosch, Steinbach, Müller, & Pantel, 2012).

Chloride intracellular channel protein 3 (CLIC3)

CLIC3 encodes chloride intracellular channel 3 protein which is located in the nucleus and activates chloride ion channels. Chloride channels participate in stabilizing cell membrane potential, maintaining intracellular pH and regulating trans-epithelial transport and cell volume (Money et al., 2007). It is demonstrated that CLIC3 promotes breast cancer cells invasiveness by increasing the delivery of MTI-MMP and internalised α5β1 integrin to the plasma membrane (Dozynkiewicz et al., 2012; Macpherson et al., 2014). Furthermore, Macpherson et al showed that overexpression of CLIC3 is associated with greater risk of death in breast cancer patients.

PLIN5

PLIN5 gene encodes PLIN5 which is a member of perilipins family. Perilipins are important structural proteins that modulate lipid storage and are known to play a crucial role in diseases associated with lipid manifestations (Kimmel & Sztalryd, 2016; Sztalryd & Brasaemle, 2017). PLIN5 regulates storage and release of fatty acid in order to maintain lipid homeostasis and

also participates in mediating mitochondrial functions (Dalen et al., 2007; Mason & Watt, 2015). The role of perilipins in cancer is not well studied yet. However, it is suggested that they can contribute to cancer by regulating the metabolism of cancer cells (Asimakopoulou et al., 2019) as their proliferation demands a high levels of energy. PLIN1 (another member of perilipins family) is deregulated in breast cancer (S. Kim, Lee, & Koo, 2015). PLIN5 is suggested to play a role in lipo-carcinoma, renal carcinoma, rhabdomyosarcoma and liver cancer. As our study implies, PLIN5 might play a role in breast cancer too. However, to our understanding expression of PLIN5 in breast cancer has not been investigated yet.

4.6. Survival analysis

4.6.1. SNV-level survival analysis

We evaluated the association between occurrence of the positive SNVs based on our model predictions and survival outcome of the patients from prediction sets. We performed the analysis in young patient (Age<45) and old patient (Age>=45) groups, respectively. To this aim we used the highlighted set of SNVs in **Table 4.10, Table 4.11, Table 4.12**. Following our survival analysis we found the occurrence of "17:7674220_C>T" SNV (highlighted in **Table 4.13**) to be significantly associated with survival outcome of the young patients but not the old patients from METABRIC study. **Figure 4.7** and **Figure 4.8** indicate the results from disease specific survival analysis and overall survival analysis in both old and young patients. However, we did not replicate this finding in the TCGA-BRCA study.

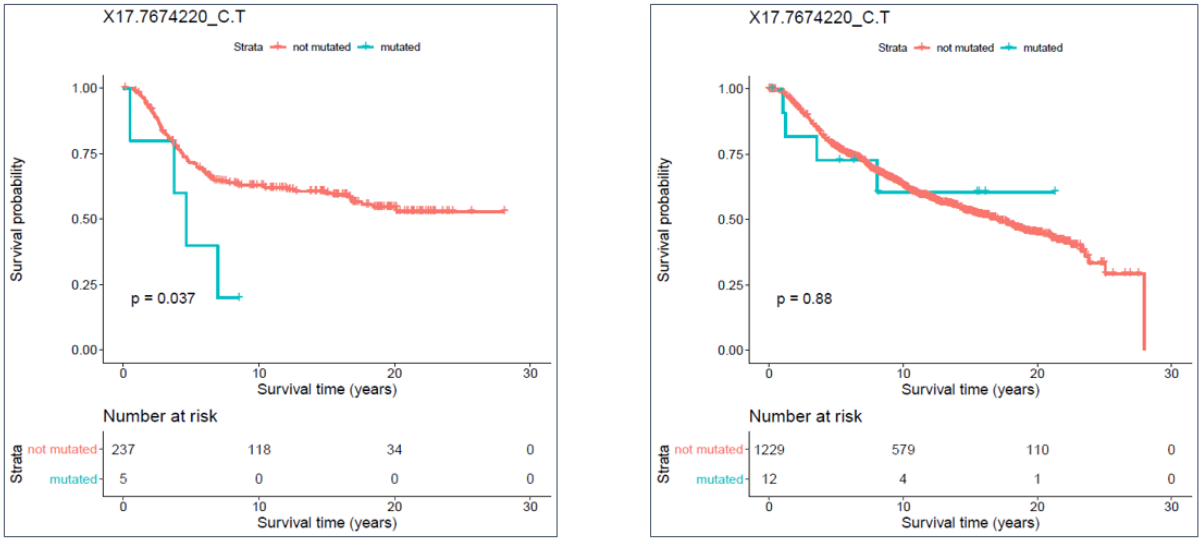

Figure 4.7 Results from disease specific survival (DSS) analysis comparing the survival time of breast cancer patients with and without the mutation. The difference between the two groups of patients (with and without the mutation) is significant among young (under 45 years of age) individuals (p-value=0.037), but not significant in older patients (p-value=0.88)

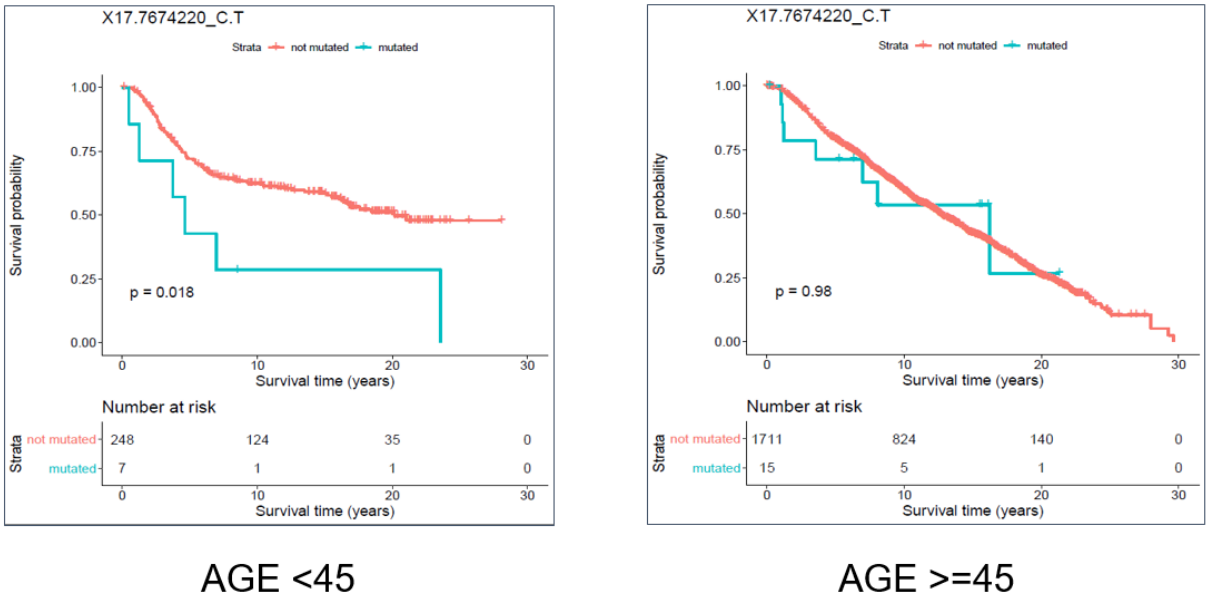

Figure 4.8 Results from overall survival (OS) analysis comparing the survival time of breast cancer patients with and without the mutation. The difference between the two groups of patients (with and

without the mutation) is significant among young (under 45 years of age) individuals (p-value=0.018), but not significant in older patients (p-value=0.98)

Following our subsequent investigations we figured out that the mentioned mutation is one of the most frequently reported mutations in TP53 gene. The mutation which is regarded as a "hotspot mutation" is a single nucleotide substitution resulting in arginine to glutamine substitution at 248th residue of TP53 (Shajani-Yi, de Abreu, Peterson, & Tsongalis, 2018). It is suggested that the mutation disrupts the tumor suppressive activity of TP53 by hindering the binding of the TP53 product to DNA (Bullock & Fersht, 2001; Saha, Kar, & Sa, 2015; Soussi & Wiman, 2015).

Furthermore, we investigated the biochemical properties of arginine and glutamine to unravel the impact of their substitution on proteins structure and function. The positively charged guanidinium group in arginine makes it a hydrophilic amino-acid appropriate for being located on the surface of proteins in an aquatic environment (Borders Jr et al., 1994). Arginine has an important role in binding of a protein's active site to negatively charges cofactors, effectors, substrates (Riordan, McElvany, & Borders, 1977). Arginine also participates in formation of salt bridges which stabilize the tertiary and quaternary structure of proteins (Borders Jr et al., 1994). Glutamine on the other hand is a polar neutral amino acid which is found either as the N-terminus of proteins or in the free form. Accordingly, being substituted for arginine, glutamine is not capable of compensating for arginine role in maintaining the proteins structure and binding activity.

Furthermore, we also investigated the clinical data of the patients included in our SNV-level survival analysis harboring "17:7674220_C>T" SNV in their genome. About half of the patients (4 out of 7) had a HER2 positive subtype and the rest of the patients (3 out of 7) had a triple negative subtype. All the patients suffered from a grade 3 cancer.

4.6.1. Gene-level survival analysis

To select the candidate genes for conducting gene-level survival analysis we extracted the genes that harbor a higher number of positive SNVs. To do so, we counted the number of positive SNVs affecting each gene in the prediction datasets. **Table 4.14** indicates the number of genes affected by different number of positive somatic point mutations. For example, **Table 4.14** shows that in METABRIC dataset 29 genes are affected by at least 7 positive SNVs.

Table 4.14 Number of genes affected per different thresholds. The thresholds indicate the number of positive somatic point mutations each gene is harboring.

Frequency Threshold	1	2	3	4	5	6	7	8	9
Count of genes-METABRIC	154	106	74	55	44	36	29	27	23
Count of genes-TCGA-coding	2539	412	92	29	11	7	3	2	2
Count of genes-TCGA- non-coding	330	22	3	0	0	0	0	0	0

To be consistent with the frequency threshold we defined for labeling positive examples in our gold standards at the very first steps of model building process, we selected the genes harboring at least four and two pathogenic positive SNVs in coding (METABRIC and TCGA-BRCA coding) and noncoding (TCGA-BRCA) datasets, respectively. It should be noted that these mutations may be not recurrent. Regarding **Table 4.14** the selected thresholds provided us with 55 genes from METABRIC, 29 genes from TCGA-coding, and 22 genes from TCGA non-coding.

Through gene-level survival analysis for the selected genes we investigated whether the fact that a given gene is mutated or not, can significantly affect the survival experience of the patients. We conducted the gene-level survival analysis for both young (<45 years old) and old (>=45 years old) groups of patients and compared the results from the two groups. The gene-level survival analysis of METABRIC data revealed that mutated “MUC16” gene significantly (p-

value<0.05) affected the survival experience of the young patients but did not have a significant (p-value>0.05) impact on the survival time of the older patients. **Figure 4.9** and **Figure 4.10** show the Kaplan Maier survival plots from DFS and OS analysis for "Muc 16" gene, respectively.

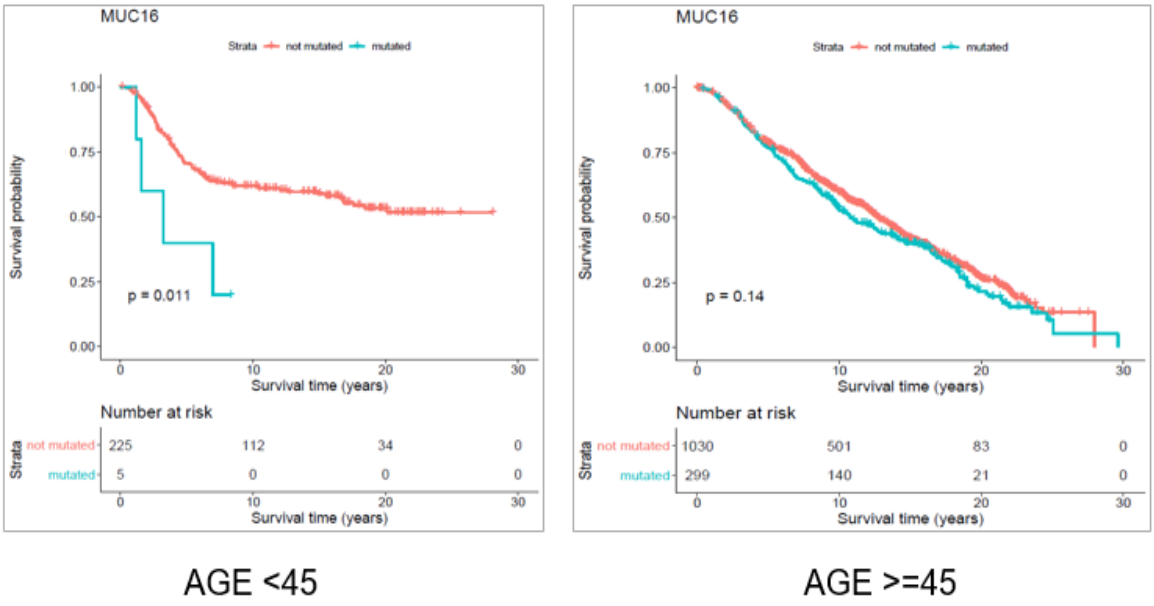

Figure 4.9 Results from disease free survival (DFS) analysis comparing the survival time of breast cancer patients possessing a mutated or un-mutated Muc16 gene. The difference between the two groups (with and without mutations in Muc16 gene) is significant among young (under 45 years of age) individuals (p-value=0.011), while not significant in older patients (p-value=0.14)

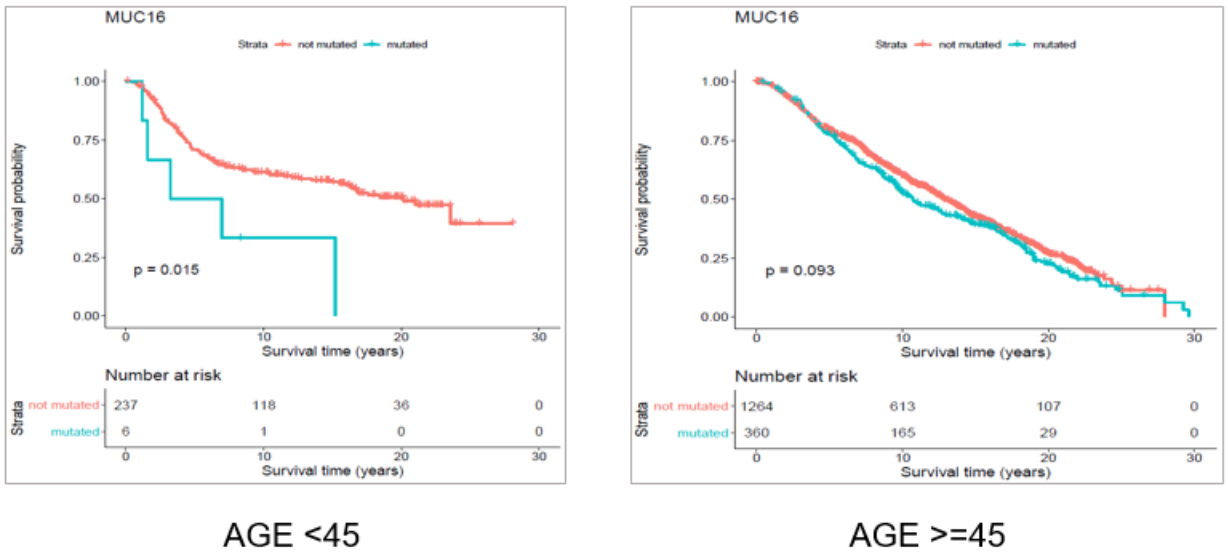

Figure 4.10 Results from overall survival (OS) analysis comparing the survival time of breast cancer patients possessing a mutated or un-mutated Muc16 gene. The difference between the two groups (with and without mutations in Muc16 gene) is significant among young (under 45 years of age) individuals (p-value=0.011), while not significant in older patients (p-value=0.14)

MUC16 gene encodes MUC16 or CA125 protein belonging to Mucin family glycoproteins. Mucins are synthesized by epithelial cells and participate in formation of chemical barriers, cell signalling and lubrication (Lakshmanan et al., 2012). It is demonstrated that there is an association between aberrant expression of mucins and cell growth, development, transformation, invasion, adhesion and immune survival (Kufe, 2009). Muc16 is regarded as a blood biomarker mainly used for monitoring the progression of ovarian cancer (Yin, Dnistrian, & Lloyd, 2002). However, Muc16 expression is also demonstrated to be associated with the development of other cancer types including pancreatic (Y.-M. Wu, Nowack, Omenn, & Haab, 2009) and breast cancers (Moritani et al., 2008). Lakshmanan et al (Lakshmanan et al., 2012) suggested that Muc16 contributes to breast cancer progression by increasing cell's proliferation through interaction with Janus kinase (JAK2) (Firmbach-Kraft, Byers, Shows, Dalla-Favera, & Krolewski, 1990) as well as inhibiting cell apoptosis by downregulating of TRAIL (LeBlanc & Ashkenazi, 2003).

Interestingly, in consistent with our findings Muc16 overexpression in epithelial breast cancer tissues is demonstrated to be positively associated with the stage of the disease (Lakshmanan et al., 2012). In addition, Norum et al (Norum, Erikstein, & Nustad, 2001) regarded the elevated expression of Muc16 in breast cancer as a sign of advanced disease. They demonstrated that increased expression of Muc16 is associated with metastasis and poor prognosis in stage IV of the disease.

To our understanding to date there is no evidence concerning the association between Muc16 deregulation and early onset of breast cancer. However, association between the stage of breast cancer and elevated Muc16 expression as well as incidence of higher stages of the disease in younger patients drastically supports our findings.

4.7. Gene set enrichment analysis (GSEA)

We performed GSEA for the genes that were affected by the somatic pathogenic positive SNVs predicted by our models. We conducted the analysis for the genes from both coding (METABRIC and TCGA coding) and noncoding datasets (TCGA noncoding). However, no significant results (adjusted p-value>0.05) were obtained for the genes from TCGA noncoding dataset. **Table 4.15** represents the results from GSEA for the genes from coding datasets.

As shown in **Table 4.15** we organized the genes from coding datasets into four gene lists to cover all the possible combination of the genes from the coding datasets. The four gene lists include A) genes from METABRIC dataset only (highlighted in **Table 4.10** and **Table 4.14**), B) genes from TCGA coding dataset only (highlighted in **Table 4.11** and **Table 4.14**), C) a combination of genes harbouring any pathogenic positive SNVs predicted by our models from METABRIC and TCGA coding datasets (highlighted in **Table 4.10** and **Table 4.11**), D) a

combination of genes affected by four or more pathogenic positive SNVs predicted by the models from METABRIC and TCGA coding datasets (highlighted in **Table 4.14**).

Table 4.15 Significant (adjusted P-value<0.05) gene sets showing an overrepresentation of our candidate gene lists. For each library we have only reported the 5 top significant gene sets.

Gene list	Reactome 2016	Panther 2016	KEGG 2019 Human	GO Biological Process 2018	GO Molecular Function 2018	ChEA 2016
A	Chromatin modifying enzymes_Homo sapiens_R-HSA-3247509	EGF receptor signaling pathway_Homo sapiens_P00018	Endometrial cancer	positive regulation of nucleic acid-templated transcription	protein kinase activity	AR_22383394_ChIP-Seq_PROSTATE_CANCER_Human
A	Chromatin organization_Homo sapiens_R-HSA-4839726	p53 pathway feedback loops 2_Homo sapiens_P04398	Hepatocellular carcinoma	positive regulation of gene expression	protein kinase binding	STAT3_23295773_ChIP-Seq_U87_Human
A	Diseases of signal transduction_Homo sapiens_R-HSA-5663202	Angiogenesis_Homo sapiens_P00005	Pathways in cancer	positive regulation of transcription, DNA-templated	transcription coactivator activity	SMAD4_21799915_ChIP-Seq_A2780_Human
A	PI-3K cascade:FGFR1_Homo sapiens_R-HSA-5654689	Insulin/IGF pathway-protein kinase B signaling cascade_Homo sapiens_P00033	Human papillomavirus infection	phosphatidylinositol 3-kinase signaling	ubiquitin protein ligase binding	ZNF217_24962896_ChIP-Seq_MCF-7_Human
A	PI-3K cascade:FGFR2_Homo sapiens_R-HSA-5654695	Apoptosis signaling pathway_Homo sapiens_P00006	Breast cancer	chromatin disassembly	ubiquitin-like protein ligase binding	DROSHA_22980978_ChIP-Seq_HELA_Human
B	Neuronal System_Homo sapiens_R-HSA-112316	Endothelin signaling pathway_Homo sapiens_P00019	Endometrial cancer	calcium ion import	calcium ion transmembrane transporter activity	STAT3_23295773_ChIP-Seq_U87_Human
B	Transmission across Chemical Synapses_Homo sapiens_R-HSA-112315	p53 pathway feedback loops 2_Homo sapiens_P04398	PI3K-Akt signaling pathway	axonogenesis	ATPase activity (	TCF4_23295773_ChIP-Seq_U87_Human
B	PI-3K cascade:FGFR1_Homo sapiens_R-HSA-5654689	p53 pathway_Homo sapiens_P00059	Pathways in cancer	calcium ion transmembrane transport	calcium channel activity	SMAD4_21799915_ChIP-Seq_A2780_Human
B	PI-3K cascade:FGFR2_Homo sapiens_R-HSA-5654695	Ionotropic glutamate receptor pathway_Homo sapiens_P00037	Breast cancer	protein phosphorylation	motor activity	AR_22383394_ChIP-Seq_PROSTATE_CANCER_Human

B	PI-3K cascade:FGFR3_Homo sapiens_R-HSA-5654710	Wnt signaling pathway_Homo sapiens_P00057	Pathways in cancer	calcium ion transport	voltage-gated cation channel activity	PAX3-FKHR_20663909_ChIP-Seq_RHABDOMYOSARCOMA_Human
C	PI-3K cascade:FGFR1_Homo sapiens_R-HSA-5654689	p53 pathway feedback loops 2_Homo sapiens_P04398	Endometrial cancer	protein phosphorylation (GO:0006468)	MAP kinase kinase kinase activity	STAT3_23295773_ChIP-Seq_U87_Human
C	PI-3K cascade:FGFR2_Homo sapiens_R-HSA-5654695	EGF receptor signaling pathway_Homo sapiens_P00018	Gastric cancer	protein autophosphorylation (GO:0046777)	calcium ion transmembrane transporter activity	TCF4_23295773_ChIP-Seq_U87_Human
C	PI-3K cascade:FGFR3_Homo sapiens_R-HSA-5654710	Endothelin signaling pathway_Homo sapiens_P00019	Thyroid hormone signaling pathway	calcium ion import (GO:0070509)	protein kinase activity (GO:0004672)	SMAD4_21799915_ChIP-Seq_A2780_Human
C	PI-3K cascade:FGFR4_Homo sapiens_R-HSA-5654720	p53 pathway_Homo sapiens_P00059	Central carbon metabolism in cancer	peptidyl-serine phosphorylation (GO:0018105)	ATPase activity (GO:0016887)	AR_22383394_ChIP-Seq_PROSTATE_CANCER_Human
C	PI3K events in ERBB4 signaling_Homo sapiens_R-HSA-1250342	Wnt signaling pathway_Homo sapiens_P00057	Breast cancer	phosphorylation (GO:0016310)	ATP-dependent microtubule motor activity, minus-end-directed (GO:0008569)	DROSHA_22980978_ChIP-Seq_HELA_Human
D	Chromatin modifying enzymes_Homo sapiens_R-HSA-3247509	CCKR signaling map ST_Homo sapiens_P06959	Endometrial cancer	regulation of megakaryocyte differentiation (GO:0045652)	ATP-dependent microtubule motor activity, minus-end-directed (GO:0008569)	AR_19668381_ChIP-Seq_PC3_Human
D	Chromatin organization_Homo sapiens_R-HSA-4839726	Wnt signaling pathway_Homo sapiens_P00057	Human papillomavirus infection	regulation of myeloid cell differentiation (GO:0045637)	ATP-dependent microtubule motor activity (GO:1990939)	TCF4_23295773_ChIP-Seq_U87_Human
D	Developmental Biology_Homo sapiens_R-HSA-1266738	Huntington disease_Homo sapiens_P00029	Hepatocellular carcinoma	cellular response to caffeine (GO:0071313)	ligand-gated calcium channel activity (GO:0099604)	SMAD4_21799915_ChIP-Seq_A2780_Human

D	PI3K/AKT Signaling in Cancer_Homo sapiens_R-HSA-2219528	p53 pathway_Homo sapiens_P00059	Lysine degradation	response to caffeine (GO:0031000)	protein kinase binding (GO:0019901)	STAT3_23295773_ChIP-Seq_U87_Human
D	PKMTs methylate histone lysines_Homo sapiens_R-HSA-3214841	Beta1 adrenergic receptor signaling	Huntington disease	regulation of cardiac muscle cell contraction (GO:0086004)	ATPase activity (GO:0016887)	ZNF217_24962896_ChIP-Seq_MCF-7_Human

In **Table 4.15** we have reported the significant results (adjusted p-value <0.05) from the following databases:

i) Reactome 2016, including information about genes, proteins, vaccines, etc. that participate in a network of biological interactions and pathways (Fabregat et al., 2018; Fabregat et al., 2017).

ii) Panther 2016, including information from over 177 primary signalling pathways along with the subfamilies and the protein sequences belonging to each pathway (Mi, Muruganujan, Ebert, Huang, & Thomas, 2018; Mi & Thomas, 2009).

iii) KEGG 2019 Human, including information from a collection of cellular pathways contributing to metabolism, human diseases, organismal systems, genetic information processing etc.(Kanehisa, Furumichi, Tanabe, Sato, & Morishima, 2016; Kanehisa & Goto, 2000; Kanehisa, Sato, Furumichi, Morishima, & Tanabe, 2018).

iv) GO Biological Process 2018, including information from Gene Ontology (GO) consortium which is the largest available source of information about the functions of genes. “Go biological process” database contains the informations from pathways and biological processes in which product of a given gene is involved (Ashburner et al., 2000; G. O. Consortium, 2018).

v) GO Molecular Function 2018, including information from GO spanning the molecular activities of the products of genes (Ashburner et al., 2000; G. O. Consortium, 2018).

vi) ChEA 2016, including information from Chip-seq studies of transcription factors extracted from various publications (Lachmann et al., 2010).

Each cell in **Table 4.15** represents a gene set (from the above databases) in which our input gene list is significantly over represented. As inferred from **Table 4.15** many of the significant gene sets are associated with pathways/biologicals functions contributing to breast cancer. For instance, gene sets related to pathways such as PIK3 cascade ("PI-3K cascade: FGFR1", "PI-3K cascade: FGFR2", "PI-3K cascade: FGFR3", "PI3K events in ERBB4", "PI3K/AKT Signaling in Cancer") and TP53 pathway ("p53 pathway feedback loops 2", "p53 pathway") are frequently appeared among the most significant results for different gene lists. As reported earlier in this study TP53 and PI3K are among the most important genes with causal effect in breast cancer. Similarly, gene sets related to zinc finger protein family (e.g. "ZNF217_24962896_ChIP-Seq_MCF-7_Human ") also appeared as one of the significant results from "ChEA 2016" database. Interestingly, in consistent with the purpose of our analysis, the results from "KEGG 2019 database regard breast cancer as one of the cancers most significantly associated with the input gene lists. In conclusion, the GSEA fully supports the association between the input gene lists (genes identified as significant through our analysis) and breast cancer development.

4.8. Significance and conclusion

Computational models for classifying the pathogenic status of cancer somatic variants located in coding and noncoding regions of the genome are developed in this study. The significant novelty of the proposed models is development of a bi-dimensional threshold that considers the recurrent frequency of a given SNV across the whole dataset as well as the number of the cancer types the SNV is identified in. Furthermore, both the pathogenic positive SNVs and benign negative SNVs included in the gold standard datasets are exclusively cancer somatic variants distinguishing our models from the available classifiers.

The developed models outperform the most powerful available classification tools which is evidence enough for robustness of the discrimination capabilities of our models in terms of classifying cancer somatic variants.

The high classification accuracy of the developed models are promising in terms of predicting the pathogenic status of a set of cancer somatic variants whose pathogenicity has not been assigned before.

The potential application of the computational models in identifying novel candidate target genes and biomarkers is also suggested through our study. Our survival analysis on pathogenic positive cancer somatic SNVs (predicted by our models) and their related genes highlighted the age-specific significance of a SNV and a gene impacting the survival time of young (<45 years old) breast cancer patients more than their older counterparts.

We hope that the models can be helpful in predicting the potential biological impact of novel cancer somatic SNVs identified through next generation sequencing of cancer tissues every day.

Chapter 5 : Limitations and Future Directions

Our classification models are designed based on a robust labeling process defined for the first time in this study. However, labels from clinical wet lab experiments are assumed more reliable. To date there is no major dataset available including the clinical significance of cancer somatic variants based on the validated results from wet lab experiments.

A higher number of features can usually positively affect the classification power of a computational model. The genomic features we used for training our models are annotations from the limited available annotating tools. A greater number of annotation tools potential for generating more genomic features can considerably elevate the discrimination power of computational models.

SVM models are proved to be effective classification methods. However, generating models using more complicated model building approaches such as deep neural network methods can be beneficial in increasing the models classification accuracy.

As future directions, conducting further analysis on the prediction results from applying our models to somatic mutations from breast cancer datasets can lead to identification of novel therapeutic targets and biomarkers.

Furthermore, experimental validation of the results from our model predictions can provide a strong proof of the classification performance of our computational models.

Bibliography

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., . . . McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65. doi:10.1038/nature11632

Adami, H.-O., Malker, B., Holmberg, L., Persson, I., & Stone, B. (1986). The relation between survival and age at diagnosis in breast cancer. *New England Journal of Medicine, 315*(9), 559-563.

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet, Chapter 7*, Unit7 20. doi:10.1002/0471142905.hg0720s76

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics, 76*(1), 7.20. 21-27.20. 41.

Ahonen, T., Saltevo, J., Laakso, M., Kautiainen, H., Kumpusalo, E., & Vanhala, M. (2009). Gender differences relating to metabolic syndrome and proinflammation in Finnish subjects with elevated blood pressure. *Mediators Inflamm, 2009*, 959281. doi:10.1155/2009/959281

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., . . . Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature, 500*(7463), 415-421. doi:10.1038/nature12477

Ali, H. R., Chlon, L., Pharoah, P. D., Markowetz, F., & Caldas, C. (2016). Patterns of immune infiltration in breast cancer and their clinical implications: a gene-expression-based retrospective study. *PLoS medicine, 13*(12), e1002194.

Anders, C. K., Acharya, C. R., Hsu, D. S., Broadwater, G., Garman, K., Foekens, J. A., . . . Marks, J. R. (2008). Age-specific differences in oncogenic pathway deregulation seen in human breast tumors. *PloS one, 3*(1), e1373.

Anders, C. K., Hsu, D. S., Broadwater, G., Acharya, C. R., Foekens, J. A., Zhang, Y., . . . Febbo, P. G. (2008). Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. *Journal of clinical oncology, 26*(20), 3324-3330.

Arenzana, T. L., Schjerven, H., & Smale, S. T. (2015). Regulation of gene expression dynamics during developmental transitions by the Ikaros transcription factor. *Genes & development, 29*(17), 1801-1816.

Asaga, S., Kuo, C., Nguyen, T., Terpenning, M., Giuliano, A. E., & Hoon, D. S. (2011). Direct serum assay for microRNA-21 concentrations in early and advanced breast cancer. *Clinical chemistry, 57*(1), 84-91.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Eppig, J. T. (2000). Gene ontology: tool for the unification of biology. *Nature genetics, 25*(1), 25.

Asimakopoulou, A., Vucur, M., Luedde, T., Schneiders, S., Kalampoka, S., Weiss, T. S., & Weiskirchen, R. (2019). Perilipin 5 and Lipocalin 2 Expression in Hepatocellular Carcinoma. *Cancers, 11*(3), 385.

Aslan, B., Monroig, P., Hsu, M.-C., Pena, G. A., Rodriguez-Aguayo, C., Gonzalez-Villasana, V., . . . Han, H.-D. (2015). The ZNF304-integrin axis protects against anoikis in cancer. *Nature communications, 6*, 7351.

Azim, H. A., Michiels, S., Bedard, P. L., Singhal, S. K., Criscitiello, C., Ignatiadis, M., . . . Loi, S. (2012). Elucidating prognosis and biology of breast cancer arising in young women using gene expression profiling. *Clinical cancer research, 18*(5), 1341-1351.

Azim, H. A., & Partridge, A. H. (2014). Biology of breast cancer in young women. *Breast cancer research, 16*(4), 427.

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K. K., Carter, S. L., Frederick, A. M., . . . Zou, L. (2012). Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature, 486*(7403), 405.

Bao, L., Zhang, Y., Wang, J., Wang, H., Dong, N., Su, X., . . . Wang, X. (2016). Variations of chromosome 2 gene expressions among patients with lung cancer or non-cancer. *Cell biology and toxicology, 32*(5), 419-435.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., . . . Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol, 28*(10), 1045-1048. doi:10.1038/nbt1010-1045

Béroud, C., & Soussi, T. (2003). The UMD-p53 database: new mutations and analysis tools. *Human mutation, 21*(3), 176-181.

Bharat, A., Aft, R. L., Gao, F., & Margenthaler, J. A. (2009). Patient and tumor characteristics associated with increased mortality in young women (≤ 40 years) with breast cancer. *Journal of surgical oncology, 100*(3), 248-251.

Bieging, K. T., Mello, S. S., & Attardi, L. D. (2014). Unravelling mechanisms of p53-mediated tumour suppression. *Nature Reviews Cancer, 14*(5), 359.

Borders Jr, C., Broadwater, J. A., Bekeny, P. A., Salmon, J. E., Lee, A. S., Eldridge, A. M., & Pett, V. B. (1994). A structural role for arginine in proteins: multiple hydrogen bonds to backbone carbonyl oxygens. *Protein Science, 3*(4), 541-548.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., . . . Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell, 132*(2), 311-322. doi:10.1016/j.cell.2007.12.014

Brown, K. K., & Toker, A. (2015). The phosphoinositide 3-kinase pathway and therapy resistance in cancer. *F1000prime reports, 7*.

Bullock, A. N., & Fersht, A. R. (2001). Rescuing the function of mutant p53. *Nature Reviews Cancer, 1*(1), 68.

Buuren, S. v., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1-68.

Caballero, A., Tenesa, A., & Keightley, P. D. (2015). The Nature of Genetic Variation for Complex Traits Revealed by GWAS and Regional Heritability Mapping Analyses. *Genetics, 201*(4), 1601-1613. doi:10.1534/genetics.115.177220

Cai, Y., He, J., & Zhang, D. (2015). Long noncoding RNA CCAT2 promotes breast tumor growth by regulating the Wnt signaling pathway. *OncoTargets and therapy, 8*, 2657.

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., . . . Kent, W J. (2017). The UCSC Genome Browser database: 2018 update. *Nucleic Acids Research, 46*(D1), D762-D769. doi:10.1093/nar/gkx1020

Cawley, G. C. T., Nicola L. C. . (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research, 2079–2107*(11).

Cech, T. R., & Steitz, J. A. (2014). The noncoding RNA revolution—trashing old rules to forge new ones. *Cell, 157*(1), 77-94.

Chae, B. J., Bae, J. S., Lee, A., Park, W. C., Seo, Y. J., Song, B. J., . . . Jung, S. S. (2009). p53 as a specific prognostic factor in triple-negative breast cancer. *Japanese Journal of Clinical Oncology, 39*(4), 217-224.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., . . . Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics, 14*(1), 128.

Cheng, J. Q., Lindsley, C. W., Cheng, G. Z., Yang, H., & Nicosia, S. V. (2005). The Akt/PKB pathway: molecular target for cancer drug discovery. *Oncogene, 24*(50), 7482.

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57.

Consortium, G. O. (2018). The Gene Ontology resource: 20 years and still GOing strong. *Nucleic acids research, 47*(D1), D330-D338.

Consortium, G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56.

Consortium, I. C. G. (2010). International network of cancer genome projects. *Nature, 464*(7291), 993.

Cook, N. R. (2007). Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation, 115*(7), 928-935.

Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., & Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res, 15*(7), 901-913. doi:10.1101/gr.3577405

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., . . . Yuan, Y. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature, 486*(7403), 346.

D'Errico, I., Gadaleta, G., & Saccone, C. (2004). Pseudogenes in metazoa: origin and features. *Briefings in Functional Genomics, 3*(2), 157-167.

Dalen, K. T., Dahl, T., Holter, E., Arntsen, B., Londos, C., Sztalryd, C., & Nebb, H. I. (2007). LSDP5 is a PAT protein specifically expressed in fatty acid oxidizing tissues. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 1771*(2), 210-227.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Sherry, S. T. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158.

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J. B., Gaffney, D. J., Pickrell, J. K., . . . Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature, 482*(7385), 390-394. doi:10.1038/nature10808

Dillon, R. L., Marcotte, R., Hennessy, B. T., Woodgett, J. R., Mills, G. B., & Muller, W. J. (2009). Akt1 and akt2 play distinct roles in the initiation and metastatic phases of mammary tumor progression. *Cancer research, 69*(12), 5057-5064.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., . . . Schlesinger, F. (2012). Landscape of transcription in human cells. *Nature, 489*(7414), 101.

Donner, A. (1982). The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *The American Statistician, 36*(4), 378-381.

Dozynkiewicz, M. A., Jamieson, N. B., MacPherson, I., Grindlay, J., van den Berghe, P. V., von Thun, A., . . . Nixon, C. (2012). Rab25 and CLIC3 collaborate to promote integrin recycling from late endosomes/lysosomes and drive cancer progression. *Developmental cell, 22*(1), 131-145.

Dunna, N. R., Naushad, S. M., Vuree, S., Anuradha, C., Sailaja, K., Surekha, D., . . . Vishnupriya, S. (2014). Association of Thymidylate Synthase 5'-UTR 28bp Tandem Repeat and Serine Hydroxymethyltransfarase C1420T Polymorphisms with Susceptibility to Acute Leukemia. *Asian Pacific Journal of Cancer Prevention, 15*(4), 1719-1723. doi:10.7314/apjcp.2014.15.4.1719

Egger, G., Liang, G., Aparicio, A., & Jones, P. A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature, 429*(6990), 457-463. doi:10.1038/nature02625

El Saghir, N. S., Khalil, M. K., Eid, T., El Kinge, A. R., Charafeddine, M., Geara, F., . . . Shamseddine, A. I. (2007). Trends in epidemiology and management of breast cancer in developing Arab countries: a literature and registry analysis. *International journal of surgery, 5*(4), 225-233.

Elkabets, M., Vora, S., Juric, D., Morse, N., Mino-Kenudson, M., Muranen, T., . . . Ibrahim, Y. H. (2013). mTORC1 inhibition is required for sensitivity to PI3K p110α inhibitors in PIK3CA-mutant breast cancer. *Science translational medicine, 5*(196), 196ra199-196ra199.

Ellis, M. J., Ding, L., Shen, D., Luo, J., Suman, V. J., Wallis, J. W., . . . Goldstein, T. C. (2012). Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature, 486*(7403), 353.

Elmore, S. (2007). Apoptosis: a review of programmed cell death. *Toxicologic pathology, 35*(4), 495-516.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., . . . D'Eustachio, P. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Res, 46*(D1), D649-d655. doi:10.1093/nar/gkx1132

Fabregat, A., Sidiropoulos, K., Viteri, G., Forner, O., Marin-Garcia, P., Arnau, V., . . . Hermjakob, H. (2017). Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics, 18*(1), 142. doi:10.1186/s12859-017-1559-2

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 92-107.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874.

Feinberg, A. P., Ohlsson, R., & Henikoff, S. (2006). The epigenetic progenitor origin of human cancer. *Nat Rev Genet, 7*(1), 21-33. doi:10.1038/nrg1748

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., . . . Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Res, 38*(Database issue), D211-222. doi:10.1093/nar/gkp985

Firmbach-Kraft, I., Byers, M., Shows, T., Dalla-Favera, R., & Krolewski, J. (1990). tyk2, prototype of a novel class of non-receptor tyrosine kinase genes. *Oncogene, 5*(9), 1329-1336.

Fontanarosa, J. B., & Dai, Y. (2011). *Using LASSO regression to detect predictive aggregate effects in genetic studies.* Paper presented at the BMC proceedings.

Forbes, S. A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., . . . Stratton, M. R. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Hum Genet, Chapter 10*, Unit 10.11. doi:10.1002/0471142905.hg1011s57

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., . . . Futreal, P. A. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res, 39*(Database issue), D945-950. doi:10.1093/nar/gkq929

Fredholm, H., Eaker, S., Frisell, J., Holmberg, L., Fredriksson, I., & Lindman, H. (2009). Breast cancer in young women: poor survival despite intensive treatment. *PloS one, 4*(11), e7695.

Fridman, J. S., & Lowe, S. W. (2003). Control of apoptosis by p53. *Oncogene, 22*(56), 9030.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics, 16*(10), 906-914.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., . . . Stratton, M. R. (2004). A census of human cancer genes. *Nat Rev Cancer, 4*(3), 177-183. doi:10.1038/nrc1299

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., . . . Caligiuri, M. A. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science, 286*(5439), 531-537.

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R., Creixell, P., Karchin, R., . . . Pearson, J. V. (2013). Computational approaches to identify functional genetic variants in cancer genomes. *Nature methods, 10*(8), 723.

Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R., Creixell, P., Karchin, R., . . . Consequences Subgroup of the Bioinformatics Analyses Working, G. (2013). Computational approaches to identify functional genetic variants in cancer genomes. *Nat Methods, 10*(8), 723-729. doi:10.1038/nmeth.2562

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., . . . Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature, 446*(7132), 153-158. doi:10.1038/nature05610

Griffiths AJF, M. J., Suzuki DT, et al. (2000). Somatic versus germinal mutation. In W. H. Freeman (Ed.), *An Introduction to Genetic Analysis* (7 edition ed.). New York.

Griseri, P., Bourcier, C., Hieblot, C., Essafi-Benkhadir, K., Chamorey, E., Touriol, C., & Pages, G. (2011). A synonymous polymorphism of the Tristetraprolin (TTP) gene, an AU-rich mRNA-binding protein, affects translation efficiency and response to Herceptin treatment in breast cancer patients. *Hum Mol Genet, 20*(23), 4556-4568. doi:10.1093/hmg/ddr390

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a shared vision for cancer genomic data. *New England Journal of Medicine, 375*(12), 1109-1112.

Han, L., Yuan, Y., Zheng, S., Yang, Y., Li, J., Edgerton, M. E., . . . Liang, H. (2014). The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nature communications, 5*, 3963.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell, 144*(5), 646-674.

He, H., Li, W., Liyanarachchi, S., Srinivas, M., Wang, Y., Akagi, K., . . . de la Chapelle, A. (2015). Multiple functional variants in long-range enhancer elements contribute to the risk of SNP rs965513 in thyroid cancer. *Proc Natl Acad Sci U S A, 112*(19), 6128-6133. doi:10.1073/pnas.1506255112

Heiser, L. M., Sadanandam, A., Kuo, W.-L., Benz, S. C., Goldstein, T. C., Ng, S., . . . Tong, F. (2012). Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proceedings of the National Academy of Sciences, 109*(8), 2724-2729.

Heneghan, H. M., Miller, N., Kelly, R., Newell, J., & Kerin, M. J. (2010). Systemic miRNA-195 differentiates breast cancer from other malignancies and is a potential biomarker for detecting noninvasive and early stage disease. *The oncologist, 15*(7), 673-682.

Hodis, E., Watson, I. R., Kryukov, G. V., Arold, S. T., Imielinski, M., Theurillat, J. P., . . . Chin, L. (2012). A landscape of driver mutations in melanoma. *Cell, 150*(2), 251-263. doi:10.1016/j.cell.2012.06.024

Horiguchi, K., Sakamoto, K., Koinuma, D., Semba, K., Inoue, A., Inoue, S., . . . Miyazono, K. (2012). TGF-β drives epithelial-mesenchymal transition through δEF1-mediated downregulation of ESRP. *Oncogene, 31*(26), 3190.

Huang, S., Cai, N., Pacheco, P. P., NARRANDES, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics-Proteomics, 15*(1), 41-51.

Huang, T., Alvarez, A., Hu, B., & Cheng, S.-Y. (2013). Noncoding RNAs in cancer and cancer stem cells. *Chinese journal of cancer, 32*(11), 582.

Hubner, R. A., & Houlston, R. S. (2017). Single nucleotide polymorphisms and cancer susceptibility *The Molecular Basis of Human Cancer* (pp. 231-239): Springer.

Huszno, J., & Grzybowska, E. (2018). TP53 mutations and SNPs as prognostic and predictive factors in patients with breast cancer. *Oncology letters, 16*(1), 34-40.

Jen, J., & Wang, Y.-C. (2016). Zinc finger proteins in cancer progression. *Journal of biomedical science, 23*(1), 53.

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science, 316*(5830), 1497-1502. doi:10.1126/science.1141319

Jones, P. A., & Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nat Rev Genet, 3*(6), 415-428. doi:10.1038/nrg816

Jones, P. A., & Baylin, S. B. (2007). The epigenomics of cancer. *Cell, 128*(4), 683-692. doi:10.1016/j.cell.2007.01.029

Jones, P. A., & Laird, P. W. (1999). Cancer epigenetics comes of age. *Nat Genet, 21*(2), 163-167. doi:10.1038/5947

Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D. R., Wu, Y.-M., Cao, X., . . . Lonigro, R. J. (2012). Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell, 149*(7), 1622-1634.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research, 45*(D1), D353-D361.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research, 28*(1), 27-30.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., & Tanabe, M. (2018). New approach for understanding genome variations in KEGG. *Nucleic acids research, 47*(D1), D590-D595.

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology, 64*(5), 402.

Kang, S., Bader, A. G., & Vogt, P. K. (2005). Phosphatidylinositol 3-kinase mutations identified in human cancer are oncogenic. *Proceedings of the National Academy of Sciences, 102*(3), 802-807.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome research, 12*(6), 996-1006.

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., & Gerstein, M. (2016). Role of non-coding sequence variants in cancer. *Nat Rev Genet, 17*(2), 93-108. doi:10.1038/nrg.2015.17

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., . . . Jain, S. (2014). A draft map of the human proteome. *Nature, 509*(7502), 575.

Kim, S., Lee, Y., & Koo, J. S. (2015). Differential expression of lipid metabolism-related proteins in different breast cancer subtypes. *PloS one, 10*(3), e0119473.

Kimmel, A. R., & Sztalryd, C. (2016). The perilipins: Major cytosolic lipid droplet–associated proteins and their roles in cellular lipid storage, mobilization, and systemic homeostasis. *Annual review of nutrition, 36*, 471-509.

Kimura, M. (1991). The neutral theory of molecular evolution: a review of recent evidence. *Jpn J Genet, 66*(4), 367-386.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet, 46*(3), 310-315. doi:10.1038/ng.2892

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics, 46*(3), 310.

Klinge, C. (2018). Non-coding RNAs in breast cancer: intracellular and intercellular communication. *Non-coding RNA, 4*(4), 40.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the Ijcai.

Kollias, J., Elston, C., Ellis, I., Robertson, J., & Blamey, R. (1997). Early-onset breast cancer–histopathological and prognostic considerations. *British journal of cancer, 75*(9), 1318.

Krebs, C. J., Schultz, D. C., & Robins, D. M. (2012). The KRAB zinc finger protein RSL1 regulates sex-and tissue-specific promoter methylation and dynamic hormone-responsive chromatin configuration. *Molecular and cellular biology, 32*(18), 3732-3742.

Kreisberg, J. I., Malik, S. N., Prihoda, T. J., Bedolla, R. G., Troyer, D. A., Kreisberg, S., & Ghosh, P. M. (2004). Phosphorylation of Akt (Ser473) is an excellent predictor of poor clinical outcome in prostate cancer. *Cancer research, 64*(15), 5232-5236.

Kufe, D. W. (2009). Mucins in cancer: function, prognosis and therapy. *Nature Reviews Cancer, 9*(12), 874.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., . . . Lachmann, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research, 44*(W1), W90-W97.

Lachmann, A., Xu, H., Krishnan, J., Berger, S. I., Mazloom, A. R., & Ma'ayan, A. (2010). ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics, 26*(19), 2438-2444.

Lakshmanan, I., Ponnusamy, M. P., Das, S., Chakraborty, S., Haridas, D., Mukhopadhyay, P., . . . Batra, S. K. (2012). MUC16 induced rapid G2/M transition via interactions with JAK2 for increased proliferation and anti-apoptosis in breast cancer cells. *Oncogene, 31*(7), 805.

Landau, D. A., Tausch, E., Taylor-Weiner, A. N., Stewart, C., Reiter, J. G., Bahlo, J., . . . Wu, C. J. (2015). Mutations driving CLL and their evolution in progression and relapse. *Nature, 526*(7574), 525-530. doi:10.1038/nature15395

Langerød, A., Zhao, H., Borgan, Ø., Nesland, J. M., Bukholm, I. R., Ikdahl, T., . . . Jeffrey, S. S. (2007). TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast cancer research, 9*(3), R30.

LeBlanc, H., & Ashkenazi, A. (2003). Apo2L/TRAIL and its death and decoy receptors. *Cell death and differentiation, 10*(1), 66.

Leslie, C., Eskin, E., & Noble, W. S. (2002). The spectrum kernel: a string kernel for SVM protein classification. *Pac Symp Biocomput*, 564-575.

Lo, P. K., Wolfson, B., Zhou, X., Duru, N., Gernapudi, R., & Zhou, Q. (2016). Noncoding RNAs in breast cancer. *Brief Funct Genomics, 15*(3), 200-221. doi:10.1093/bfgp/elv055

Ma, X., Huang, M., Wang, Z., Liu, B., Zhu, Z., & Li, C. (2016). ZHX1 inhibits gastric cancer cell growth through inducing cell-cycle arrest and apoptosis. *Journal of Cancer, 7*(1), 60.

Macpherson, I. R., Rainero, E., Mitchell, L. E., van den Berghe, P. V., Speirs, C., Dozynkiewicz, M. A., . . . Timpson, P. (2014). CLIC3 controls recycling of late endosomal MT1-MMP and dictates invasion and metastasis in breast cancer. *J Cell Sci, 127*(18), 3893-3901.

Malhotra, N. K. (1987). Analyzing marketing research data with incomplete information on the dependent variable. *Journal of Marketing Research, 24*(1), 74-84.

Mason, R. R., & Watt, M. J. (2015). Unraveling the roles of PLIN5: linking cell biology to physiology. *Trends in Endocrinology & Metabolism, 26*(3), 144-152.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., . . . Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol, 17*(1), 122. doi:10.1186/s13059-016-0974-4

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., . . . Cunningham, F. (2016). The ensembl variant effect predictor. *Genome biology, 17*(1), 122.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2018). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic acids research, 47*(D1), D419-D426.

Mi, H., & Thomas, P. (2009). PANTHER pathway: an ontology-based pathway database coupled with data analysis tools *Protein Networks and Pathway Analysis* (pp. 123-140): Springer.

Minaxi Arora, L. B., M.Tech Scholar. (2014). Role of Scaling in Data Classification Using SVM. *International Journal of Advanced Research in Computer Science and Software Engineering, 4*(10).

Moler, E., Chow, M., & Mian, I. (2000). Analysis of molecular profile data using generative and discriminative methods. *Physiological genomics, 4*(2), 109-126.

Money, T., King, R., Wong, M., Stevenson, J., Kalionis, B., Erwich, J., . . . Desoye, G. (2007). Expression and cellular localisation of chloride intracellular channel 3 in human placenta and fetal membranes. *Placenta, 28*(5-6), 429-436.

Moritani, S., Ichihara, S., Hasegawa, M., Endo, T., Oiwa, M., Yoshikawa, K., . . . Kushima, R. (2008). Serous papillary adenocarcinoma of the female genital organs and invasive micropapillary carcinoma of the breast. Are WT1, CA125, and GCDFP-15 useful in differential diagnosis? *Human pathology, 39*(5), 666-671.

Network, C. G. A. (2012). Comprehensive molecular portraits of human breast tumours. *Nature, 490*(7418), 61.

Ng, P. C., & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res, 11*(5), 863-874. doi:10.1101/gr.176601

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research, 31*(13), 3812-3814.

Norum, L. F., Erikstein, B., & Nustad, K. (2001). Elevated CA 125 in breast cancer–a sign of advanced disease. *Tumor biology, 22*(4), 223-228.

O'neill, R., & Temple, R. (2012). The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. *Clinical Pharmacology & Therapeutics, 91*(3), 550-554.

Oldridge, D. A., Wood, A. C., Weichert-Leahey, N., Crimmins, I., Sussman, R., Winter, C., . . . Maris, J. M. (2015). Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. *Nature, 528*(7582), 418-421. doi:10.1038/nature15540

Pedregosa, F., Ga, #235, Varoquaux, l., Gramfort, A., Michel, V., . . . Duchesnay, d. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res., 12*, 2825-2830.

Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., . . . Sammut, S.-J. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature communications, 7*, 11479.

Poliseno, L. (2012). Pseudogenes: newly discovered players in human cancer. *Sci. Signal., 5*(242), re5-re5.

Poliseno, L., Marranci, A., & Pandolfi, P. P. (2015). Pseudogenes in Human Cancer. *Front Med (Lausanne), 2*, 68. doi:10.3389/fmed.2015.00068

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res, 20*(1), 110-121. doi:10.1101/gr.097857.109

Polo, M. L., Arnoni, M. V., Riggio, M., Wargon, V., Lanari, C., & Novaro, V. (2010). Responsiveness to PI3K and MEK inhibitors in breast cancer. Use of a 3D culture system to study pathways related to hormone independence in mice. *PLoS One, 5*(5), e10786.

Polo, M. L., Riggio, M., May, M., Rodríguez, M. J., Perrone, M. C., Stallings-Mann, M., . . . Boughey, J. (2015). Activation of PI3K/Akt/mTOR signaling in the tumor stroma drives endocrine therapy-dependent breast tumor regression. *Oncotarget, 6*(26), 22081.

Reddy, E. P., Reynolds, R. K., Santos, E., & Barbacid, M. (1982). A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature, 300*(5888), 149-152. doi:10.1038/300149a0

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2018). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research, 47*(D1), D886-D894.

Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., . . . Raney, B. J. (2009). The UCSC genome browser database: update 2010. *Nucleic acids research, 38*(suppl_1), D613-D619.

Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. *Otolaryngology—Head and Neck Surgery, 143*(3), 331-336.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Committee, A. L. Q. A. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus

recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med, 17*(5), 405-424. doi:10.1038/gim.2015.30

Riggio, M., Perrone, M. C., Polo, M. L., Rodriguez, M. J., May, M., Abba, M., . . . Novaro, V. (2017). AKT1 and AKT2 isoforms play distinct roles during breast cancer progression through the regulation of specific downstream proteins. *Scientific reports, 7*, 44244.

Riggio, M., Polo, M. L., Blaustein, M., Colman-Lerner, A., Lüthy, I., Lanari, C., & Novaro, V. (2011). PI3K/AKT pathway regulates phosphorylation of steroid receptors, hormone independence and tumor differentiation in breast cancer. *Carcinogenesis, 33*(3), 509-518.

Riordan, J., McElvany, K., & Borders, C. (1977). Arginyl residues: anion recognition sites in enzymes. *Science, 195*(4281), 884-886.

Rogers, M. F., Shihab, H. A., Gaunt, T. R., & Campbell, C. (2017). CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Sci Rep, 7*(1), 11597. doi:10.1038/s41598-017-11746-4

Rogers, M. F., Shihab, H. A., Gaunt, T. R., & Campbell, C. (2017). CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Scientific reports, 7*(1), 11597.

Rogers, M. F., Shihab, H. A., Mort, M., Cooper, D. N., Gaunt, T. R., & Campbell, C. (2018). FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics, 34*(3), 511-513. doi:10.1093/bioinformatics/btx536

Saha, T., Kar, R. K., & Sa, G. (2015). Structural and sequential context of p53: A review of experimental and theoretical evidence. *Progress in biophysics and molecular biology, 117*(2-3), 250-263.

Samuel, G. N., & Farsides, B. (2017). The UK's 100,000 Genomes Project: manifesting policymakers' expectations. *New Genet Soc, 36*(4), 336-353. doi:10.1080/14636778.2017.1370671

Samuels, Y., Wang, Z., Bardelli, A., Silliman, N., Ptak, J., Szabo, S., . . . Riggins, G. J. (2004). High frequency of mutations of the PIK3CA gene in human cancers. *Science, 304*(5670), 554-554.

Schirmer, M. A., Luske, C. M., Roppel, S., Schaudinn, A., Zimmer, C., Pfluger, R., . . . Ghadimi, B. M. (2016). Relevance of Sp Binding Site Polymorphism in WWOX for Treatment Outcome in Pancreatic Cancer. *J Natl Cancer Inst, 108*(5). doi:10.1093/jnci/djv387

Schwarzenbach, H., Milde-Langosch, K., Steinbach, B., Müller, V., & Pantel, K. (2012). Diagnostic potential of PTEN-targeting miR-214 in the blood of breast cancer patients. *Breast cancer research and treatment, 134*(3), 933-941.

Serra, R. W., Fang, M., Park, S. M., Hutchinson, L., & Green, M. R. (2014). A KRAS-directed transcriptional silencing pathway that mediates the CpG island methylator phenotype. *Elife, 3*, e02313.

Shah, S. P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., . . . Haffari, G. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature, 486*(7403), 395.

Shajani-Yi, Z., de Abreu, F. B., Peterson, J. D., & Tsongalis, G. J. (2018). Frequency of Somatic TP53 Mutations in Combination with Known Pathogenic Mutations in Colon Adenocarcinoma, Non–Small Cell Lung Carcinoma, and Gliomas as Identified by Next-Generation Sequencing. *Neoplasia, 20*(3), 256-262.

Sharma, S., Kelly, T. K., & Jones, P. A. (2009). Epigenetics in cancer. *Carcinogenesis, 31*(1), 27-36. doi:10.1093/carcin/bgp220

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research, 29*(1), 308-311.

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N., . . . Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics, 31*(10), 1536-1543. doi:10.1093/bioinformatics/btv009

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N., . . . Campbell, C. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics, 31*(10), 1536-1543.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., . . . Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res, 15*(8), 1034-1050. doi:10.1101/gr.3715005

Siepel, A., & Haussler, D. (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol, 11*(2-3), 413-428. doi:10.1089/1066527041410472

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., & Forbes, S. A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 1.

Soussi, T., & Wiman, K. (2015). TP53: an oncogene in disguise. *Cell death and differentiation, 22*(8), 1239.

Srivastava, S., Zou, Z., Pirollo, K., Blattner, W., & Chang, E. H. (1990). Germ-line transmission of a mutated p53 gene in a cancer-prone family with Li–Fraumeni syndrome. *Nature, 348*(6303), 747.

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., . . . Cooper, D. N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet, 136*(6), 665-677. doi:10.1007/s00439-017-1779-6

Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., . . . Bignell, G. R. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature, 486*(7403), 400.

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

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Lander, E. S. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences, 102*(43), 15545-15550.

Sun, Z., Yang, S., Zhou, Q., Wang, G., Song, J., Li, Z., . . . Chang, Y. (2018). Emerging role of exosome-derived long non-coding RNAs in tumor microenvironment. *Molecular cancer, 17*(1), 82.

Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G., & Kuznetsov, E. N. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Engineering, Design and Selection, 12*(5), 387-394. doi:10.1093/protein/12.5.387

Sztalryd, C., & Brasaemle, D. L. (2017). The perilipin family of lipid droplet proteins: Gatekeepers of intracellular lipolysis. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 1862*(10), 1221-1232.

Tabin, C. J., Bradley, S. M., Bargmann, C. I., Weinberg, R. A., Papageorge, A. G., Scolnick, E. M., . . . Chang, E. H. (1982). Mechanism of activation of a human oncogene. *Nature, 300*(5888), 143-149.

Thomas, J., Touchman, J., Blakesley, R., Bouffard, G., Beckstrom-Sternberg, S. M., Margulies, E., . . . McDowell, J. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature, 424*(6950), 788.

Thomas, J. W., Touchman, J. W., Blakesley, R. W., Bouffard, G. G., Beckstrom-Sternberg, S. M., Margulies, E. H., . . . Green, E. D. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature, 424*(6950), 788-793. doi:10.1038/nature01858

Thusberg, J., Olatubosun, A., & Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Human mutation, 32*(4), 358-368.

Thusberg, J., Olatubosun, A., & Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat, 32*(4), 358-368. doi:10.1002/humu.21445

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.

Toker, A. (2012). Achieving specificity in Akt signaling in cancer. *Advances in biological regulation, 52*(1), 78.

Trisilowati, & Mallet, D. G. (2012). In silico experimental modeling of cancer treatment. *ISRN Oncol, 2012*, 828701. doi:10.5402/2012/828701

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., & Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet, 45*(2), 124-130. doi:10.1038/ng.2504

Van der Aalst, W. M., Rubin, V., Verbeek, H., van Dongen, B. F., Kindler, E., & Günther, C. W. (2010). Process mining: a two-step approach to balance between underfitting and overfitting. *Software & Systems Modeling, 9*(1), 87.

Vandermoere, F., El Yazidi-Belkoura, I., Demont, Y., Slomianny, C., Antol, J., Lemoine, J., & Hondermarck, H. (2007). Proteomics exploration reveals that actin is a signaling target of the kinase Akt. *Molecular & Cellular Proteomics, 6*(1), 114-124.

Vardhanabhuti, S., Wang, J., & Hannenhalli, S. (2007). Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res, 35*(10), 3203-3213. doi:10.1093/nar/gkm201

Vargas-Landin, D. B., Pfluger, J., & Lister, R. (2018). Generation of Whole Genome Bisulfite Sequencing Libraries for Comprehensive DNA Methylome Analysis. *Methods Mol Biol, 1767*, 291-298. doi:10.1007/978-1-4939-7774-1_16

Varna, M., Bousquet, G., Plassa, L.-F., Bertheau, P., & Janin, A. (2011). TP53 status and response to treatment in breast cancers. *BioMed Research International, 2011*.

Vendramin, R., Marine, J. C., & Leucci, E. (2017). Non-coding RNAs: the dark side of nuclear–mitochondrial communication. *The EMBO journal, 36*(9), 1123-1133.

Versi, E. (1992). "Gold standard" is an appropriate term. *Bmj, 305*(6846), 187. doi:10.1136/bmj.305.6846.187-b

Vihinen, M. (2014). Contribution of pseudogenes to sequence diversity *Pseudogenes* (pp. 15-24): Springer.

Wang, Y., Helland, Å., Holm, R., Skomedal, H., Abeler, V., Danielsen, H., . . . Kristensen, G. (2004). TP53 mutations in early-stage ovarian carcinoma, relation to long-term survival. *British journal of cancer, 90*(3), 678.

Wang, Z.-Y., & Yin, L. (2015). Estrogen receptor alpha-36 (ER-α36): A new player in human breast cancer. *Molecular and cellular endocrinology, 418*, 193-206.

Wattel, E., Preudhomme, C., Hecquet, B., Vanrumbeke, M., Quesnel, B., Dervite, I., . . . Fenaux, P. (1994). p53 mutations are associated with resistance to chemotherapy and short survival in hematologic malignancies. *Blood, 84*(9), 3148-3157.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., . . . Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet, 45*(10), 1113-1120. doi:10.1038/ng.2764

Wu, H., Zhang, K., Gong, P., Qiao, F., Wang, L., Cui, H., . . . Fan, H. (2014). A novel functional TagSNP Rs7560488 in the DNMT3A1 promoter is associated with susceptibility to gastric cancer by modulating promoter activity. *PLoS One, 9*(3), e92911. doi:10.1371/journal.pone.0092911

Wu, Y.-M., Nowack, D. D., Omenn, G. S., & Haab, B. B. (2009). Mucin glycosylation is altered by pro-inflammatory signaling in pancreatic-cancer cells. *Journal of proteome research, 8*(4), 1876-1886.

Wulff, J. N., & Ejlskov, L. (2017). Multiple Imputation by Chained Equations in Praxis: Guidelines and Review. *Electronic Journal of Business Research Methods, 15*(1).

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K., . . . Frey, B. J. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science, 347*(6218), 1254806. doi:10.1126/science.1254806

Yin, B. W., Dnistrian, A., & Lloyd, K. O. (2002). Ovarian cancer antigen CA125 is encoded by the MUC16 mucin gene. *International journal of cancer, 98*(5), 737-740.

Zhang, F., & Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Hum Mol Genet, 24*(R1), R102-110. doi:10.1093/hmg/ddv259

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., . . . Kasprzyk, A. (2011). International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database (Oxford), 2011*, bar026. doi:10.1093/database/bar026

Zhang, P. (1993). Model selection via multifold cross validation. *The Annals of Statistics*, 299-313.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., . . . Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol, 9*(9), R137. doi:10.1186/gb-2008-9-9-r137

Appendix

METBARIC

Genes affected by somatic SNVs predicted as pathogenic positive from METABRIC dataset:

'AC004160.1', 'AC004160.2', 'AC005066.1', 'AC008575.1', 'AC009095.1', 'AC016576.1', 'AC026954.2', 'AFDN', 'AGMO', 'AHNAK', 'AHNAK2', 'AKAP9', 'AKT1', 'AKT2', 'AL132671.2', 'ALK', 'APC', 'ARID1A', 'ARID1B', 'ARID2', 'ARID5B', 'ATR', 'BAP1', 'BCAS3', 'BIRC6', 'BRCA1', 'BRCA2', 'BRIP1', 'CACNA2D3', 'CASP8', 'CBFB', 'CBR4', 'CDH1', 'CHD1', 'CHEK2', 'COL12A1', 'COL22A1', 'COL6A3', 'CTCF', 'CTNNA3', 'DCAF4L2', 'DNAH11', 'DNAH2', 'DNAH5', 'EGFR', 'ERBB2', 'ERBB3', 'FLT3', 'FOXO1', 'FOXO3', 'GATA3', 'GLDC', 'GPR32', 'HDAC9', 'HERC2', 'HSD3B7', 'JAK1', 'KMT2C', 'KMT2D', 'LAMA2', 'LAMB3', 'LARGE1', 'LIPI', 'MAP2K4', 'MAP3K1', 'MEN1', 'MIR4673', 'MUC16', 'MYH9', 'MYO1A', 'MYO3A', 'NCOA3', 'NCOR1', 'NCOR2', 'NEK1', 'NF1', 'NF2', 'NOTCH1', 'NPNT', 'NRG3', 'PALLD', 'PBRM1', 'PIK3CA', 'PIK3R1', 'PIP5K1B', 'PPP2CB', 'PRKCQ', 'PRKG1-AS1', 'PTPN22', 'PTPRD', 'RB1', 'ROS1', 'RPL29P2', 'RUNX1', 'RYR2', 'SBNO1', 'SETD1A', 'SETD2', 'SETDB1', 'SHANK2', 'SIK2', 'SMARCC1', 'SMARCC2', 'SMARCD1', 'STAB2', 'SYNE1', 'TAF4B', 'TBX3', 'TG', 'THADA', 'THSD7A', 'TP53', 'TUBB8P1', 'UBR5', 'USH2A', 'UTRN'

TCGA-BRCA coding

Genes affected by somatic SNVs predicted as pathogenic positive from TCGA_BRCA coding dataset:

'10-Mar', '14-Sep', '8-Sep', 'A2M', 'A2ML1', 'AADACL4', 'AASS', 'ABCA10', 'ABCA12', 'ABCA13', 'ABCA7', 'ABCB1', 'ABCC11', 'ABCF1', 'ABCG4', 'ABHD13', 'ABL1', 'ABL2', 'ABLIM1', 'AC002070.1', 'AC002429.2', 'AC005005.3', 'AC007540.1', 'AC007610.4', 'AC008655.1', 'AC008937.2', 'AC009095.1', 'AC010185.1', 'AC010336.6', 'AC010605.1', 'AC010680.4', 'AC013472.3', 'AC022098.1', 'AC024598.1', 'AC025576.2', 'AC073896.3', 'AC074091.1', 'AC083809.1', 'AC084866.1', 'AC091167.7', 'AC093668.3', 'AC093827.5', 'AC099314.1', 'AC104389.4', 'AC107016.1', 'AC117503.2', 'AC117834.1', 'AC118549.1', 'ACACA', 'ACCSL', 'ACSF3', 'ACSM4', 'ACTL9', 'ADAM18', 'ADAMTS12', 'ADAMTS16', 'ADAMTS20', 'ADCY2', 'ADCY5', 'ADCY8', 'ADD2', 'ADGRB3', 'ADGRE2', 'ADGRE3', 'ADGRF5', 'ADGRG7', 'ADGRL2', 'ADGRL3', 'ADGRL4', 'ADGRV1', 'ADH7', 'ADPRH', 'AFDN', 'AFF3', 'AFG3L2', 'AFMID', 'AGBL1', 'AGO2', 'AGO4', 'AGPAT4', 'AHCTF1', 'AHCYL2', 'AK8', 'AKAP11', 'AKAP13', 'AKAP9', 'AKIRIN1', 'AKR1C1', 'AKT3', 'AKTIP', 'AL049779.1', 'AL121928.1', 'AL136295.2', 'AL138752.2', 'AL157786.1', 'AL157829.1', 'AL162741.1', 'AL354771.1', 'AL355385.1', 'AL392048.1', 'ALDH18A1', 'ALOXE3', 'ALS2', 'AMBRA1', 'AMIGO3', 'AMOTL1', 'ANK1', 'ANK3', 'ANKRD17', 'ANKRD19P', 'ANKRD20A4', 'ANKRD30A', 'ANKRD30B', 'ANKRD31', 'ANKRD36C', 'ANKRD44', 'ANO2', 'ANPEP', 'ANXA13', 'AP000781.1', 'AP000781.2', 'AP001092.1', 'AP002748.4', 'AP003086.2', 'AP1B1', 'AP2B1', 'AP4E1', 'AP4S1', 'APBA2', 'APBB1IP', 'APCDD1', 'APOB', 'APOL5', 'APPBP2', 'ARFGEF2', 'ARFGEF3', 'ARHGAP17', 'ARHGAP29', 'ARHGAP30', 'ARHGAP31', 'ARHGAP40', 'ARHGAP5', 'ARHGEF12', 'ARHGEF37', 'ARID1A', 'ARID2', 'ARID4B', 'ARL6', 'ARVCF', 'ASPHD2', 'ASPM', 'ASTN1', 'ASXL1', 'ASXL2', 'ASXL3', 'ASZ1', 'ATAD2B', 'ATF7IP', 'ATG2A', 'ATM', 'ATMIN', 'ATP10A', 'ATP10B', 'ATP13A4', 'ATP13A5', 'ATP2A1', 'ATP2B1', 'ATP2B2', 'ATP9A', 'ATRNL1', 'ATXN1', 'AURKB', 'AXIN1', 'AXL', 'BAG6', 'BAZ2A', 'BBS5', 'BCAS1', 'BCHE', 'BCL11A', 'BEND3', 'BEND6', 'BFSP1', 'BIRC6', 'BMP6',

'BMPR2', 'BMT2', 'BPTF', 'BRCA1', 'BRCA2', 'BRINP1', 'BRINP3', 'BRPF1', 'BRPF3', 'BRWD1', 'C10orf25', 'C11orf49', 'C11orf65', 'C2orf78', 'C3', 'C3orf20', 'C5', 'C6', 'C6orf118', 'C7', 'CABIN1', 'CACNA1A', 'CACNA1B', 'CACNA1D', 'CACNA1E', 'CACNA1I', 'CACNA2D3', 'CACNB3', 'CACNG3', 'CADM2', 'CADPS', 'CAMK2D', 'CAMKK1', 'CAMSAP2', 'CAPN13', 'CAPN2', 'CASP8', 'CASQ2', 'CASR', 'CBLB', 'CBLN2', 'CC2D1B', 'CCAR1', 'CCDC114', 'CCDC172', 'CCDC181', 'CCDC39', 'CCDC66', 'CCDC69', 'CCDC88C', 'CCNE1', 'CCNF', 'CCNY', 'CCR3', 'CCSER1', 'CD163L1', 'CD1B', 'CD300E', 'CD3E', 'CD53', 'CD86', 'CDC14A', 'CDC40', 'CDCA7L', 'CDH1', 'CDH11', 'CDH20', 'CDH22', 'CDH4', 'CDH9', 'CDK3', 'CDYL', 'CELSR1', 'CELSR2', 'CEP126', 'CEP128', 'CEP135', 'CEP164', 'CEP250', 'CEP350', 'CEP85', 'CFAP100', 'CFAP298', 'CFAP58', 'CFAP61', 'CFH', 'CFHR4', 'CFHR5', 'CFTR', 'CHADL', 'CHD3', 'CHD4', 'CHD5', 'CHD6', 'CHD7', 'CHD9', 'CHEK2', 'CHL1', 'CHRM3', 'CHRNB4', 'CHST9', 'CHSY3', 'CIITA', 'CILP', 'CKMT2', 'CLASP2', 'CLASRP', 'CLEC10A', 'CLIP2', 'CLIP3', 'CLK2', 'CMTR2', 'CNBD1', 'CNDP2', 'CNGA3', 'CNOT1', 'CNOT3', 'CNTN3', 'CNTN4-AS1', 'CNTN6', 'CNTNAP2', 'COG3', 'COL14A1', 'COL19A1', 'COL22A1', 'COL3A1', 'COL5A2', 'COL6A3', 'COL6A5', 'COX5A', 'COX7A2', 'CPA1', 'CPAMD8', 'CPNE4', 'CPSF6', 'CPXM1', 'CREB5', 'CREM', 'CRNKL1', 'CROCC', 'CROT', 'CRTC2', 'CRY1', 'CSDE1', 'CSF3R', 'CSMD1', 'CSMD2', 'CSMD3', 'CSNK1A1', 'CSPP1', 'CTBP2', 'CTCF', 'CTNNA1', 'CYB5D2', 'CYB5R4', 'CYFIP2', 'CYHR1', 'CYP4A11', 'CYP4F2', 'CYP4X1', 'CYP7A1', 'CYSLTR2', 'DAP', 'DAPK1', 'DAZL', 'DBT', 'DCAF1', 'DCAF4L2', 'DCAF6', 'DCC', 'DDA1', 'DDB1', 'DDI1', 'DDX20', 'DDX27', 'DDX60', 'DENND4C', 'DEPDC1B', 'DEPDC5', 'DET1', 'DGKA', 'DGKI', 'DHRS12', 'DHRS7C', 'DHX15', 'DHX8', 'DIS3L2', 'DISP1', 'DISP3', 'DLC1', 'DMXL1', 'DNAH10', 'DNAH6', 'DNAH8', 'DNAH9', 'DNAJB9', 'DNAJC13', 'DNASE1L2', 'DNHD1', 'DNMBP', 'DNMT3A', 'DOCK4', 'DOCK5', 'DOCK7', 'DOCK8', 'DOLPP1', 'DOP1A', 'DPP8', 'DPY19L3', 'DPYS', 'DRC7', 'DRGX', 'DROSHA', 'DSC1', 'DSC2', 'DSC3', 'DSCAM', 'DSG3', 'DSP', 'DSPP', 'DST', 'DTX1', 'DTX4', 'DUOX1', 'DUOX2', 'DUSP27', 'DYNC1H1', 'DYNC2H1', 'DYRK1A', 'DYSF', 'DZANK1', 'EDNRB', 'EED', 'EFCAB13', 'EFCAB6', 'EFNB3', 'EFTUD2', 'EHD3', 'EIF3A', 'EIF5A2P1', 'EIF5B', 'ELN', 'EMC1', 'EML5', 'ENAM', 'ENOX1', 'ENPP1', 'ENPP2', 'EP400', 'EPB41L1', 'EPB41L5', 'EPHA6', 'EPRS', 'EQTN', 'ERC1', 'ERC2', 'ERCC5', 'ERICH3-AS1', 'ERLEC1', 'ERV3-1', 'ESYT3', 'ETNPPL', 'ETV1', 'ETV5', 'EVC2', 'EXOC6B', 'EXPH5', 'EXT1', 'EYA2', 'EYA3', 'EZH2', 'F13A1', 'F13B', 'F5', 'FAM111A', 'FAM114A2', 'FAM117B', 'FAM120A', 'FAM124A', 'FAM135B', 'FAM13C', 'FAM149B1', 'FAM160B1', 'FAM166B', 'FAM171B', 'FAM180B', 'FAM200A', 'FAM208A', 'FAM234A', 'FAM49B', 'FAM78B', 'FAM83B', 'FANCC', 'FAP', 'FARS2', 'FAT3', 'FAT4', 'FBL', 'FBLN2', 'FBN1', 'FBN3', 'FBXL3', 'FBXO17', 'FBXO31', 'FBXO33', 'FCRL1', 'FCRL3', 'FGA', 'FGF18', 'FGFR2', 'FGFR4', 'FGR', 'FIGN', 'FILIP1', 'FIP1L1', 'FKTN', 'FLII', 'FLNB', 'FLRT3', 'FLVCR1', 'FOXJ3', 'FOXRED1', 'FPR2', 'FRAS1', 'FRG1', 'FRMD1', 'FRMD6', 'FRY', 'FSBP', 'FSHR', 'FTMT', 'FXR1', 'FYB1', 'FZD3', 'FZD4', 'FZD9', 'GAB2', 'GABRA2', 'GABRA5', 'GABRB3', 'GABRG1', 'GABRR3', 'GAL3ST1', 'GALNT7', 'GALNTL5', 'GAREM1', 'GATA4', 'GATAD2A', 'GATM', 'GBA3', 'GBF1', 'GBP5', 'GCN1', 'GDF9', 'GGT1', 'GHRHR', 'GIF', 'GIMAP8', 'GJA8', 'GK2', 'GKAP1', 'GLG1', 'GLI1', 'GLYR1', 'GMDS', 'GMFB', 'GOLGA2', 'GOSR1', 'GPR83', 'GPSM1', 'GRAMD1C', 'GRAMD4', 'GRIA4', 'GRID2', 'GRIK2', 'GRIK4', 'GRIK5', 'GRIN2A', 'GRIN2B', 'GRIN3A', 'GRM1', 'GRM3', 'GTF3C4', 'GUCY2D', 'GYS2', 'HAO2', 'HDLBP', 'HEATR1', 'HECTD4', 'HECW1', 'HELQ', 'HELZ', 'HEPN1', 'HERC2', 'HIC2', 'HID1', 'HIPK1', 'HIPK2', 'HIVEP1', 'HK2', 'HMCN1', 'HMGXB4', 'HNF1B', 'HOOK1', 'HOOK3', 'HPD', 'HPDL', 'HPSE', 'HRH2', 'HSD3B1', 'HSPA12A', 'HSPA14', 'HSPA4', 'HSPG2', 'HTR1E', 'HTR2A', 'HTT', 'HYDIN', 'IBSP', 'IDE', 'IFT81', 'IGF2R', 'IGSF10', 'IKBKE', 'IKZF1', 'IL17B', 'IL1RL1',

'IL22RA1', 'IL23R', 'IL27RA', 'INTS8', 'INTS9', 'INVS', 'IP6K3', 'IPO11', 'IQCH', 'IQGAP2', 'IQUB', 'IRF3', 'ITCH', 'ITFG2', 'ITGA8', 'ITGAX', 'ITGB1', 'ITGB2', 'ITIH2', 'ITK', 'JADE2', 'JAG1', 'JAK2', 'JAKMIP1', 'JHY', 'KANK1', 'KANSL3', 'KAT14', 'KAT6B', 'KCNAB1', 'KCNB1', 'KCNB2', 'KCNC2', 'KCND2', 'KCNG1', 'KCNH3', 'KCNJ2', 'KCNJ4', 'KCNJ6', 'KCNK2', 'KCNMB2', 'KCNQ3', 'KCNQ4', 'KCNS3', 'KCNT1', 'KCNT2', 'KCNU1', 'KCTD21', 'KDF1', 'KDM3B', 'KHDRBS2', 'KIAA0232', 'KIAA0319', 'KIAA0319L', 'KIAA0586', 'KIAA1109', 'KIAA1217', 'KIAA2026', 'KIF11', 'KIF13A', 'KIF13B', 'KIF19', 'KIF1A', 'KIF21B', 'KIF24', 'KIF2A', 'KIF3C', 'KIF4B', 'KIF5B', 'KIF7', 'KIFAP3', 'KIT', 'KLF7', 'KLHDC2', 'KLHL1', 'KLHL32', 'KMT2A', 'KMT2B', 'KMT2C', 'KMT2D', 'KMT2E', 'KPNA7', 'KRT14', 'KRT33B', 'KRT4', 'KRTAP21-2', 'KRTAP24-1', 'KTN1', 'L3MBTL2', 'L3MBTL4', 'LAMA1', 'LAMA2', 'LAMA3', 'LAMB3', 'LAMB4', 'LAMP5-AS1', 'LARP1', 'LARP4B', 'LATS2', 'LCK', 'LCMT1', 'LCP1', 'LEKR1', 'LINGO2', 'LMBRD2', 'LMX1A', 'LOXHD1', 'LPIN3', 'LPO', 'LRGUK', 'LRP1', 'LRP1B', 'LRP2', 'LRP4', 'LRP6', 'LRPPRC', 'LRRC28', 'LRRC66', 'LRRIQ1', 'LRRTM1', 'LTBP3', 'LYST', 'MAB21L3', 'MACF1', 'MACROD1', 'MAGI1', 'MANBA', 'MAP1A', 'MAP2K2', 'MAP2K4', 'MAP3K1', 'MAP3K10', 'MAP3K4', 'MAP3K9', 'MAPK1', 'MAPKAP1', 'MARF1', 'MAS1L', 'MASP2', 'MAST1', 'MAST3', 'MB21D2', 'MBD1', 'MBNL2', 'MC2R', 'MCCC1', 'MCM5', 'MCOLN1', 'MCOLN2', 'MCU', 'MDGA1', 'MDGA2', 'MDM2', 'MDN1', 'MECOM', 'MED13', 'MED13L', 'MED16', 'MED23', 'MEF2A', 'MEIOC', 'MEP1B', 'METTL15', 'MFAP1', 'MFN1', 'MFSD4A', 'MGA', 'MGAM', 'MINAR1', 'MINDY1', 'MIR1276', 'MIR4489', 'MIR488', 'MIR499A', 'MIR548D1', 'MIR6769B', 'MIR6802', 'MKNK1', 'MKRN1', 'MLH1', 'MMP16', 'MMP26', 'MOB1B', 'MOGAT3', 'MORC1', 'MOV10L1', 'MPHOSPH10', 'MPHOSPH8', 'MPV17L2', 'MRGPRX3', 'MROH1', 'MROH2B', 'MROH5', 'MROH8', 'MROH9', 'MRPS35', 'MRTFA', 'MTA2', 'MTF2', 'MTHFD1', 'MTX2', 'MUC12', 'MUC4', 'MUTYH', 'MXD1', 'MYBL2', 'MYBPC2', 'MYCBP2', 'MYCBPAP', 'MYH1', 'MYH10', 'MYH11', 'MYH15', 'MYH4', 'MYH7B', 'MYH9', 'MYL10', 'MYO18A', 'MYO1B', 'MYO1D', 'MYO6', 'MYOM3', 'MYT1L', 'NALCN', 'NALCN-AS1', 'NAMPT', 'NARS', 'NAT10', 'NAV3', 'NBEAL1', 'NBEAL2', 'NBR1', 'NCAPD2', 'NCKAP1L', 'NCKAP5', 'NCOA2', 'NCOA6', 'NCOR1', 'NDRG1', 'NDST1', 'NEB', 'NEBL', 'NECAB1', 'NECTIN3', 'NEDD1', 'NEDD4L', 'NELFCD', 'NELL1', 'NEMP2', 'NF2', 'NFIX', 'NFYB', 'NFYC', 'NID1', 'NIPBL', 'NKPD1', 'NLK', 'NMT1', 'NNT', 'NOS3', 'NPHP4', 'NPR2', 'NPVF', 'NR3C1', 'NR6A1', 'NRDC', 'NRDE2', 'NRXN2', 'NRXN3', 'NSD1', 'NT5DC3', 'NT5E', 'NTNG1', 'NTRK2', 'NTRK3', 'NTSR2', 'NUP155', 'NXPH1', 'ODF2', 'OR10G8', 'OR10Q1', 'OR13C2', 'OR13C8', 'OR2T11', 'OR2W3', 'OR4C12', 'OR4C13', 'OR4K13', 'OR4X1', 'OR5B17', 'OR5I1', 'OR5L1', 'OR5R1', 'OR5W2', 'OR6K2', 'OR6T1', 'OR6X1', 'OR7C2', 'OR7G1', 'OR8H2', 'OR8K1', 'OSBPL1A', 'OSBPL6', 'OSBPL8', 'OTOGL', 'OTUD4', 'OTUD7A', 'OTX2', 'OVCH1', 'PACS1', 'PACSIN2', 'PAK5', 'PALM2-AKAP2', 'PAMR1', 'PAN2', 'PAPPA', 'PAPPA2', 'PARD3B', 'PARPBP', 'PARVA', 'PAX3', 'PAXBP1', 'PCDH18', 'PCDHA12', 'PCDHB11', 'PCDHGA1', 'PCDHGA4', 'PCDHGB2', 'PCF11', 'PCLO', 'PCM1', 'PCNT', 'PCNX2', 'PCSK1', 'PCSK5', 'PDCD11', 'PDE3A', 'PDE3B', 'PDE4D', 'PDE9A', 'PDGFRB', 'PDP2', 'PDS5B', 'PDSS1', 'PDZRN3', 'PEG3', 'PEX5L', 'PGBD5', 'PGM2', 'PGM2L1', 'PHF21B', 'PHIP', 'PHKB', 'PHLDB1', 'PHLPP2', 'PHOSPHO2', 'PHYKPL', 'PI4KA', 'PIAS1', 'PIK3C2B', 'PIK3CA', 'PIK3CB', 'PIK3CD', 'PIK3R1', 'PIK3R2', 'PIK3R4', 'PIN4P1', 'PIP4K2C', 'PIP4P2', 'PIP5K1A', 'PITPNM3', 'PKHD1', 'PKHD1L1', 'PKLR', 'PKN2', 'PKNOX2', 'PLA2G4E-AS1', 'PLAG1', 'PLCB2', 'PLCB2-AS1', 'PLCB3', 'PLCE1', 'PLCH2', 'PLCL2', 'PLCXD3', 'PLCZ1', 'PLD5', 'PLEKHM1', 'PLEKHM2', 'PLG', 'PLXDC2', 'PMFBP1', 'PMS2', 'PNPLA6', 'POLE', 'POLH', 'POLI', 'POLQ', 'POLR2F', 'POLR3A', 'POM121', 'POU2F2', 'POU6F2', 'PPFIA2', 'PPM1M', 'PPP1R12A', 'PPP1R26', 'PPP1R9A', 'PPP2CA', 'PPP2R5C', 'PPP4R3B', 'PPP6R2', 'PRAMEF11',

'PRDM15', 'PRDM9', 'PREX2', 'PRKAA2', 'PRKCA', 'PRKCE', 'PRKCZ', 'PRKDC', 'PROX1', 'PRPF4', 'PRPF4B', 'PRPF6', 'PRPF8', 'PRR14', 'PRR16', 'PRRC2C', 'PRRX1', 'PSD2', 'PSMA5', 'PSMD1', 'PTCHD3', 'PTCHD4', 'PTEN', 'PTGES3', 'PTMS', 'PTPN22', 'PTPRCAP', 'PTPRD', 'PTPRE', 'PTPRF', 'PTPRK', 'PTPRN2', 'PTPRO', 'PTPRS', 'PTPRT', 'PTPRU', 'PTPRZ1', 'PUM1', 'PXDNL', 'PYHIN1', 'PYROXD1', 'QSER1', 'RAB14', 'RAF1', 'RAG1', 'RAP1GAP2', 'RAP1GDS1', 'RAPGEF2', 'RARS', 'RASGRF1', 'RBBP6', 'RBBP8', 'RBM19', 'RBM33', 'RBM46', 'RBM47', 'RBM5', 'REG3G', 'REL', 'RELCH', 'RELN', 'RERE', 'RF00402', 'RFK', 'RGS20', 'RGS4', 'RGS6', 'RGS7', 'RGS9', 'RGSL1', 'RHBDF2', 'RHOT1', 'RIF1', 'RIMBP2', 'RIMKLA', 'RIMKLB', 'RIMS3', 'RINT1', 'RN7SKP281', 'RN7SL835P', 'RNF111', 'RNF113B', 'RNF13', 'RNF157', 'RNF169', 'RNF20', 'RNF213', 'RNF34', 'RNU6-101P', 'RNU6-395P', 'RNU6-446P', 'RNU6-509P', 'ROBO1', 'ROBO2', 'ROR1', 'ROR2', 'RPAP3', 'RPL5', 'RPLP0', 'RPS6', 'RRP9', 'RSL1D1', 'RSPO2', 'RSRC2', 'RTCB', 'RTN4', 'RTTN', 'RUNX1', 'RUSC2', 'RXRG', 'RYR1', 'RYR2', 'RYR3', 'SACS', 'SALL1', 'SALL2', 'SALL4', 'SARS', 'SASH1', 'SCAMP5', 'SCAPER', 'SCEL', 'SCIN', 'SCN11A', 'SCN2A', 'SCN5A', 'SCP2', 'SCYL2', 'SDK2', 'SEC23A', 'SEC23B', 'SEC24B', 'SEC61A2', 'SEC63', 'SELENBP1', 'SELENOI', 'SEMA3A', 'SEMA3C', 'SEMA4G', 'SEMA6A', 'SEMA6C', 'SERBP1', 'SERPINB12', 'SERPINE1', 'SETD1A', 'SETDB1', 'SETX', 'SF3B1', 'SFXN4', 'SH3BP5', 'SH3GL2', 'SHANK2', 'SHISA6', 'SHISAL2B', 'SIM1', 'SIM2', 'SIPA1L1', 'SIPA1L2', 'SKA1', 'SLC12A2', 'SLC12A3', 'SLC12A5', 'SLC13A1', 'SLC15A5', 'SLC16A6', 'SLC17A1', 'SLC1A6', 'SLC22A8', 'SLC22A9', 'SLC25A11', 'SLC25A12', 'SLC25A13', 'SLC26A5', 'SLC27A4', 'SLC2A13', 'SLC2A14', 'SLC2A3', 'SLC2A9', 'SLC30A10', 'SLC30A4', 'SLC35B1', 'SLC35E3', 'SLC4A1', 'SLC4A10', 'SLC4A8', 'SLC5A12', 'SLC6A1', 'SLC9A2', 'SLC9A9', 'SLCO1B1', 'SLCO1C1', 'SLFN5', 'SLIT2', 'SMAD4', 'SMARCA2', 'SMARCA5', 'SMG5', 'SMURF2', 'SNAI1', 'SNAP91', 'SNAPC3', 'SNRNP200', 'SP140', 'SPART', 'SPAST', 'SPATA6', 'SPINT1', 'SPOCD1', 'SPOCK1', 'SPTA1', 'SPTBN4', 'SPX', 'SRPK1', 'SRPRA', 'SSR1', 'STAC', 'STAC2', 'STAT4', 'STAT5B', 'STK10', 'STK3', 'STK31', 'STK33', 'STON1', 'SUCO', 'SUPT16H', 'SUPT5H', 'SUZ12', 'SV2A', 'SVEP1', 'SVIL', 'SYCN', 'SYCP2', 'SYNE1', 'SYNE2', 'SYT14', 'SYT9', 'TAAR8', 'TAF5L', 'TANC1', 'TAOK1', 'TARS', 'TARS2', 'TAS2R31', 'TBC1D21', 'TBC1D23', 'TBC1D5', 'TBX15', 'TCF4', 'TCF7L1', 'TDRD1', 'TEAD4', 'TECTA', 'TEKT2', 'TENM3', 'TENM4', 'TENT2', 'TEP1', 'TET3', 'TEX2', 'TEX47', 'TFAP2A', 'TG', 'TGFBR1', 'TGFBR3', 'TGM3', 'TGM7', 'THAP3', 'THBS4', 'THEG', 'THEM5', 'THEMIS', 'THRAP3', 'THSD7A', 'THSD7B', 'TIAM1', 'TIFAB', 'TIPARP', 'TLK1', 'TM7SF3', 'TMCO5A', 'TMED2', 'TMEM132D', 'TMEM161B', 'TMEM170A', 'TMEM246-AS1', 'TMEM26', 'TMEM38A', 'TMEM63A', 'TMEM71', 'TMEM74', 'TMTC4', 'TNFRSF11B', 'TNFSF12', 'TNFSF4', 'TNN', 'TNNT3', 'TNPO2', 'TNRC6A', 'TNRC6B', 'TNRC6C', 'TOP3A', 'TP53', 'TP53BP1', 'TP63', 'TPR', 'TRA2B', 'TRAPPC9', 'TRIB2', 'TRIM49B', 'TRIM6', 'TRIM60', 'TRIM67', 'TRIML1', 'TRIOBP', 'TRPC3', 'TRPC4', 'TRPC7', 'TRPM6', 'TRPM8', 'TRPV4', 'TRPV6', 'TSC22D2', 'TSEN15', 'TSGA10', 'TSHZ2', 'TSHZ3', 'TSPAN2', 'TTC28', 'TTK', 'TTN', 'TTPAL', 'TUBB6', 'TULP3P1', 'TUSC3', 'TWISTNB', 'TXNDC16', 'U2SURP', 'UBAP1', 'UBE3C', 'UBN2', 'UBR3', 'UBR4', 'UBR5', 'UBTF', 'UBXN2B', 'UBXN4', 'UGGT2', 'UHRF1BP1', 'ULK2', 'UNC5C', 'UNC79', 'UNK', 'UPK2', 'UROC1', 'USE1', 'USH2A', 'USP12', 'USP24', 'USP33', 'USP34', 'USP42', 'USP45', 'USP7', 'VAC14', 'VANGL2', 'VAV1', 'VIPAS39', 'VIT', 'VLDLR', 'VPS13B', 'VPS13C', 'VPS13D', 'VPS4B', 'VRK2', 'VRTN', 'VSTM2A', 'VSTM2B', 'VTI1A', 'VWA3A', 'WAC', 'WASHC2A', 'WDFY4', 'WDR3', 'WDR33', 'WDR49', 'WDR93', 'WIF1', 'WISP1', 'WNT7A', 'WSCD1', 'WSCD2', 'XDH', 'XIRP2', 'XPNPEP1', 'XPO7', 'XPOT', 'XXYLT1', 'YES1', 'YLPM1', 'YPEL1', 'YTHDF3', 'YWHAG', 'Z82190.1', 'Z82190.2', 'ZBTB16', 'ZBTB18', 'ZBTB7C', 'ZC3H15', 'ZC3HC1', 'ZCWPW1', 'ZDBF2', 'ZDHHC18',

'ZDHHC2', 'ZFHX3', 'ZFHX4', 'ZFP30', 'ZFP91', 'ZFPM2-AS1', 'ZFYVE26', 'ZMYM2', 'ZNF101', 'ZNF107', 'ZNF142', 'ZNF181', 'ZNF207', 'ZNF208', 'ZNF215', 'ZNF232', 'ZNF3', 'ZNF335', 'ZNF35', 'ZNF362', 'ZNF43', 'ZNF439', 'ZNF461', 'ZNF462', 'ZNF484', 'ZNF516', 'ZNF518B', 'ZNF527', 'ZNF541', 'ZNF574', 'ZNF592', 'ZNF595', 'ZNF606', 'ZNF618', 'ZNF624', 'ZNF625', 'ZNF644', 'ZNF668', 'ZNF675', 'ZNF677', 'ZNF704', 'ZNF80', 'ZNF813', 'ZNF841', 'ZNRF3', 'ZP2', 'ZPLD1', 'ZPR1', 'ZRANB3'

TCGA-BRCA noncoding

Genes affected by somatic SNVs predicted as pathogenic positive from TCGA_BRCA noncoding dataset:

'AC117503.2', 'ADD2', 'AFDN', 'AL157786.1', 'ANKRD30B', 'ARFGEF2', 'ARHGAP17', 'ARHGAP31', 'BCHE', 'BPTF', 'BRWD1', 'CABIN1', 'CCDC172', 'CDYL', 'CEP128', 'CEP350', 'CMTR2', 'CNOT3', 'COL22A1', 'CPXM1', 'CSMD3', 'CYP4X1', 'DEPDC1B', 'DPY19L3', 'DSG3', 'DSP', 'DSPP', 'DTX1', 'DYRK1A', 'DZANK1', 'EFTUD2', 'FAM160B1', 'FBL', 'FCRL3', 'FGA', 'FLVCR1', 'FOXRED1', 'FRG1', 'GRM3', 'HOOK1', 'HYDIN', 'IGSF10', 'KCNJ2', 'KCNJ6', 'KIAA0319', 'KIF11', 'KLHL1', 'LAMA1', 'MED13', 'MGA', 'MYO1B', 'NAMPT', 'NAV3', 'NFYC', 'NTNG1', 'NTRK3', 'OR5R1', 'PCDHGB2', 'PCSK5', 'PDE9A', 'PDS5B', 'PIK3CA', 'PKHD1', 'PPP1R26', 'PRKCE', 'RELN', 'RNF169', 'RNF20', 'ROR1', 'RPL5', 'SIPA1L1', 'SLC35B1', 'SLC5A12', 'SNAI1', 'SUCO', 'SYNE1', 'TEX2', 'TP53', 'TSEN15', 'TSGA10', 'TULP3P1', 'UBR5', 'UBXN4', 'VAC14', 'VAV1', 'VPS4B', 'WAC', 'ZFYVE26', 'ZNF215', 'ZNF439', 'ZNF516', 'ZNF592', 'ZPLD1', 'ZRANB3'