January 2020

Text Mining Of Variant-Genotype-Phenotype Associations From Biomedical Literature

Nafiseh Saberian
Wayne State University

Recommended Citation

Saberian, Nafiseh, "Text Mining Of Variant-Genotype-Phenotype Associations From Biomedical Literature" (2020). *Wayne State University Theses*. 798.
https://digitalcommons.wayne.edu/oa_theses/798

TEXT MINING OF VARIANT-GENOTYPE-PHENOTYPE ASSOCIATIONS FROM BIOMEDICAL LITERATURE

by

NAFISEH SABERIAN

THESIS

Submitted to the Graduate School,

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

2020

MAJOR: COMPUTER SCIENCE

Approved By:

____

Advisor Date

DEDICATION

To my parents who have always been a constant source of love, care, and inspiration.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Sorin Draghici, my dissertation committee members, Dr. Loren Schwiebert and Dr. Suzan Arslanturk and my colleagues at the Intelligent Systems and Bioinformatics Laboratory.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

One crucial step in understanding the biological mechanism underlying a disease condition is to capture the relationship between the variants and the disease risk [96]. There are several publicly available databases contain the disease-associated variants such as UMD [15], Swiss-Prot [16], SNPedia [25], COSMIC [34], OMIM [44], Clinvar [63], InSiGHT [78], dbSNP [93], MutDB [95], HGMD [99], HGVbaseG2P [104], PharmGKB [105], BioMuta [115], etc. All these databases are manually curated by human experts. While this manual curation ensures a high quality of the annotations, the manual extraction of this type of information from the biomedical literature takes an enormous amount of time and effort. The current rate with which new variants are published is simply too high for any manual annotation process. As an additional challenge, despite the HGVS (Human Genome Variation Society) standard recommendations for the description of the variants, many variants are still reported in the literature in non-standard formats. A number of automatic mutation indexing tools have been developed. Such tools process biomedical literature and produce a list of mutations that appear in these papers. These include MutationMiner [9], MutationFinder [23], EMU [29], MuteXt [49], Mutation Grab [67], MEMA [82], etc. The most recent such tool, tmVar 2.0 [114] extracts variants from an article and normalizes them to their unique dbSNP identifiers. The next step is to develop software tools to extract variants-disease associations from the biomedical literature. Several methods have been proposed for this purpose such as OSIRIS [17], MuGeX [32], EnzyMiner [118], the methods proposed by Singhal *et al.* [96, 97], etc. All these methods have been applied to only the title and the abstract section of biomedical articles.

However, a comprehensive study showed that a significant number of genetic variants are only included in the full text and the supplementary materials of the articles [55]. These will be missed if the variants are only extracted from titles and abstracts. Doughty *et al.* [29] also proposed a tool named EMU for extracting the disease-associated mutations from biomedical literature. Although this tool automatically extracts the mutations and their corresponding genes from an article, it still requires human curation to discover the mutation-disease associations.

Here we propose an autoMated pipeline for inferring vAriant-driven Gene PanEls from the full-length biomedical Literature (MAGPEL) [89]. As the first step, the proposed framework employs word cloud analysis to identify the variant-relevant articles. The variant-gene-disease associations are then extracted from these articles. An evidence-based variant-driven gene panel is then generated based on the mined triplet information. A comprehensive validation procedure illustrates the capabilities of the proposed framework. We validate the proposed variant-driven gene panel by showing their abilities to predict the patients' clinical conditions (healthy vs. disease) on multiple independent validation datasets.

This document is organized as follows. In chapter 2, we first present a survey on the current publicly available databases and resources for disease-associated variants. Then, we provide an overview of the existing variant indexing tools that are able to extract variant entities from biomedical text. Chapter 3 focuses on our proposed automated pipeline for extracting variants from full-length biomedical literature. The detailed explanation of each step of the proposed pipeline and also the proposed validation analysis are presented in this chapter. Chapter 4 presents the results and the discussion section is provided in chapter 5. Finally, the conclusion and future work are discussed in chapter 6.

CHAPTER 2 BACKGROUND AND RELATED WORK

A genetic variant refers to the presence of alterations in the DNA sequences among individuals within a population. The disease-associated variants and the genetic polymorphisms (not disease-associated variants) are the two main categories of the genetic variants based on their frequencies within a population. Single-nucleotide polymorphism (SNP) is the most common type of polymorphism that occurs when one single nucleotide is replaced with another nucleotide. A disease-associated variant on the other hand refers to the rare type of variant that increases the risk of developing diseases. There are several forms of genetic variants depending on the changes in the reference sequence such as insertions, deletions, duplications, etc [101]. The detailed explanations and examples for each type of variants described on both the DNA level and the protein level are provided in Tables 1 and 2, respectively.

Table 1: Different types of variants described at the DNA level based on HGVS (Human Genome Variation Society) standard recommendations (`https://varnomen.hgvs.org/recommendations/DNA/`).

Type of variant	**Definition**	**Example**	**Chromosome Position**	**Reference nucleotide(s)**	**Alternative nucleotide(s)**
Insertion	A sequence change when one or multiple nucleotides are inserted.	c.104**ins**T	104	NA	T
Deletion	A sequence change when one or multiple nucleotides are deleted.	c.104**del**T	104	T	NA
Substitution	A sequence change when one nucleotide is replaced with another nucleotide.	c.435C>G	435	C	G
Duplication	A sequence change when copy of one or multiple nucleotides are inserted.	c.64_65**dup**TT	64_65	NA	TT
Deletion-insertion	A sequence change when one or multiple nucleotides are replaced by one or multiple other nucleotides.	c.145_147**delins**TGG	145_147	NA	TGG
Inversion	A sequence change when multiple nucleotides are replaced by the reverse complement of the original sequence.	c.5657_5660**inv**	5657_5660	TCAG	CTGA
Conversion	A sequence change when multiple nucleotides are replaced by multiple nucleotides copied from different positions in the sequence.	c.732_749**con**818_835	732_749	NA	NA

2.1 Current databases for disease-associated variants

With rapidly evolving sequencing technologies, the number of articles studying genomic variants and their associations with human diseases is dramatically increased [24,

Table 2: Different types of variants described at the protein level based on HGVS (Human Genome Variation Society) standard recommendations (`https://varnomen.hgvs.org/recommendations/protein/`).

Type of variant	Definition	Example	Chromosome Position	Reference amino acid(s)	Alternative amino acid(s)
Insertion	A sequence change when one or multiple amino acids are inserted.	p.His4_Gln5**ins**Ala	4_5	NA	Ala
Deletion	A sequence change when one or multiple amino acids are deleted.	p.Trp4**del**	4	Trp	NA
Substitution	A sequence change when one amino acid is replaced with another amino acid.	p.Trp24Cys	24	Trp	Cys
Duplication	A sequence change when copy of one or multiple amino acids are inserted.	p.Ala3**dup**	3	NA	Ala
Deletion-insertion	A sequence change when one or multiple amino acids are replaced by one or multiple other amino acids.	p.Cys28**delins**TrpVal	28	NA	TrpVal
Frame shift	A sequence change because of translation shift into another reading frame.	p.Arg97Pro**fs**Ter23	97	Arg	Pro

124]. Publicly available databases such as SNPedia [25], Clinvar [63], dbSNP [93], TopoSNP [100], etc. have been developed to aggregate and provide easy access to the results of these studies. In this section, we provide an overview of such open-access repositories designed specifically for the genomic variants (see Table 3).

Table 3: Summary of the current open-source warehouses providing information about variants, genes, and disease phenotypes.

Database	Description	URL
dbSNP [93]	Catalog of SNPs	https://www.ncbi.nlm.nih.gov/snp/
PharmGKB [47]	Catalog of human variations and drug responses	https://www.pharmgkb.org/
Ensembl [50]	Catalog of vertebrate genomes	https://ensembl.org
TopoSNP [100]	Catalog of SNPs	http://sts.bioe.uic.edu/toposnp/
COSMIC [34]	Catalog of cancer-associated somatic mutations	https://cancer.sanger.ac.uk/cosmic
SNPedia [25]	Catalog of disease-associated SNPs	https://snpedia.com
SwissVar [119]	Catalog of mutations present in UniProt	https://swissvar.expasy.org
ICGC [123]	Catalog of cancer-associated variants	https://dcc.icgc.org
HGVbaseG2P [104]	Catalog of disease-associated variants	https://www.gwascentral.org
1000 Genomes [74]	Catalog of human variations	https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/
MoKCa [85]	Catalog of cancer-associated mutations	http://strubiol.icr.ac.uk/extra/mokca/
OMIM [3]	Catalog of disease-associated mendelian mutations	https://omim.org
Clinvar [63]	Catalog of disease-associated variants	https://www.ncbi.nlm.nih.gov/clinvar/
IntOGen-mutations [41]	Catalog of cancer-associated mutations	https://www.intogen.org
BioMuta [28, 115]	Catalog of cancer-associated SNPs	https://hive.biochemistry.gwu.edu/biomuta
CIViC [42]	Catalog of cancer-associated variants	https://civicdb.org/home
LitVar [2]	Catalog of variants and associated genes, diseases and drugs	https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/LitVar/

Established by the National Center for Biotechnology Information (NCBI), dbSNP is the largest database providing information for the identified single nucleotide variants (SNPs) [93]. The latest version of dbSNP (build 154) which was released in April 2020 contains over two billion submitted SNP and 729,491,867 reference SNP records. For each SNP record, dbSNP provides a wide range of information described as follows:

- The *clinical significance* tab provides a list of diseases known to be associated with the queried SNP derived from ClinVar [63].
- The *frequency* tab displays a table of the reference and the alternative allele frequencies for the queried SNP obtained from biomedical articles.
- The *aliases* tab displays all the different HGVS entries such as DNA and protein level HGVS format of the queried SNP.
- The *submissions* tab shows a list of variants originally were submitted to dbSNP and now support the queried SNP.
- The *history* tab displays all the associated RefSNPs published in the previous dbSNP versions.
- The *publications* tab displays all the PubMed articles that mention the queried SNP.

The Human Genome Variation database of Genotype to Phenotype (HGVbaseG2P) is a website providing information for the identified SNPs and their related diseases [104]. For each SNP record, this database provides the general genomic information as well as the corresponding hyperlinks to OMIM [3], SNPedia [25], and dbSNP [93] databases for further information.

The disease-SNP association database named SNPedia [25] provides a summary of the existing knowledge about the disease-associated SNPs through a user-friendly web-based tool. For each queried SNP, in addition to the basic genomic information such as chromosome number, position, alleles, etc. this database provides hyperlinks to the external databases such as dbSNP [93], ClinGen [83], etc. The list of articles that have cited the queried SNP is also available through SNPedia.

ClinVar [63] is one of the largest publicly available web-based tools for human genetic variants. This database was launched in 2013 by the National Center for Biotechnology Information, National Institutes of Health (NIH). The variants are submitted to ClinVar by the research and clinical laboratories and expert groups. For each variant record, in addition to the basic genomic properties, the corresponding ClinVar web page provides the following information:

- The *Conditions* tab provides information and evidence regarding the diseases known to be associated with the queried variant.
- The *Gene(s)* tab shows the region overview of the variant's corresponding gene as well as a hyperlink to the gene's corresponding page in the OMIM [3] database.

Several databases have been designed and implemented specifically for cancer-associated variants. These includes BioMuta [28, 115], COSMIC [34], CIViC [42] and ICGC [122].

The Catalogue Of Somatic Mutations In Cancer (COSMIC [34]) is the largest database contains cancer-related somatic mutations. The two main resources feed into this database are *i)* manually curation of the scientific literature and *ii)* the Cancer Genome Project (CGP) at the Sanger Institute UK. For each mutation record, COSMIC provides the following information:

- The *Overview* tab provides a summary of the general genomic properties of the queried mutation such as chromosome, position, reference, and alternative alleles and the corresponding gene.

- The *Tissue distribution* tab shows the top 5 tissue types with the highest number of identified mutated samples.
- The *Samples* tab displays all the available information for the mutated samples such as tissue, histology, zygosity, and also reference articles.
- The *Pathway affected* tab shows a list of pathways known to be affected by the queried mutation.
- The *References* tab shows a list of publications providing evidence and information for the queried mutation.

The International Cancer Genome Consortium (ICGC) data portal [122] is an advanced web-based tool providing comprehensive information for the mutations identified in several major cancer types. This database is a collection of over 81 million cancer-associated mutations collected from 86 different projects. For each mutation record, ICGC provides a wide range of information such as:

- The *Summary* tab summarizes all the available information for the queried mutations such as genomic properties, cancer distribution, etc.
- The *Clinical evidence* tab shows a table with all the available clinical studies related to the queried mutation obtained from the Clinical Interpretation of Variants in Cancer (CIViC) database [42].

- The *Protein* tab shows the distribution of the identified mutations along with the corresponding protein sequences.

- The *Genome viewer* tab provides the region overview of the corresponding gene.

BioMuta [28, 115] is another web-based tool designed specifically for cancer-associated SNPs. This tool collects data from different resources such as UniProt [8], COSMIC [34], IntOGen [41], ClinVar [63], TCGA [103] and ICGC [122]. BioMuta provides the list of the most common variants identified for each cancer type. The general page layout provided by BioMuta for a queried gene contains information for all the identified Nonsynonymous single-nucleotide variations (nsSNVs) such as the genomic coordinates, the identified cancer types, and the supporting articles.

Clinical Interpretation of Variants in Cancer (CIViC [42]) is another open access resource for the cancer-associated variants. For each queried gene, this database provides detailed information about all the identified cancer-related somatic mutations. These include the variant genomic coordinates and the corresponding hyperlinks to the external databases such as COSMIC [34], ClinVar [63], and dbSNP [93].

Finally, PharmGKB [47] provides association information regarding human genetic variations and drug responses. The main goal of this database is to integrate available knowledge regarding human genetic variations and their effects on drug responses. For each queried gene, this database summarizes all the genomic variants associated with the queried gene, the drug it interacts with as well as hyperlinks to the corresponding evidence, studies, and articles.

2.2 Current automatic variant indexing tools

The current database curators are not able to keep track of all the new annotated variants because the current rate with which new variants are published is too high. In order to keep up with the new variants being published in the literature, a number of automation tools for indexing mutations from the biomedical text have been developed [53]. These tools use different computational algorithms such as conditional random field (CRF), regular expressions (RegExp), machine learning, and graph theory to identify variant-genotype-phenotype associations from biomedical literature. In the following, we review some of these methods focusing on the underlying concept used, primary features, availability, and key advantages (see Table 4).

Table 4: List of existing text-mining variant extraction methods based on the criteria related to the computational models and implementations.

Method	**Concept used**	**Availability**	**Language**	**Year**
VTag [73]	Conditional random field	N/A	N/A	2004
MuteXt [49]	Regular expressions	Standalone	Python	2004
MEMA [82]	Regular expressions	N/A	N/A	2004
MutationMiner [9]	Regular expressions	N/A	N/A	2006
YIP [119]	Regular expressions	N/A	N/A	2007
MuGeX [32]	Regular expressions	N/A	N/A	2007
Mutation GraB [67]	Regular expressions	N/A	N/A	2007
MutationFinder [23]	Regular expressions	Standalone	Python, Perl, Java	2007
EMU [29]	Regular expressions	Standalone	Perl	2011
tmVar [112]	Conditional random field	Standalone	Java	2013
tmVar 2.0 [114]	Conditional random field	Standalone	Java	2017

A number of methods have been proposed for the variant entity extraction from text using a machine learning technique named conditional random fields (CRF) [62]. The main characteristics of these methods are summarized in Table 5.

Table 5: Summary of the important characteristics of the reviewed conditional random field (CRF) approaches for variant extraction from biomedical text.

Method	Type of mutation	Gene/protein identification	Disease identification	RSID normalization
VTag [73]	Protein, DNA	✗	✗	✓
tmVar [114]	Protein, DNA, SNP	✓	✗	✗
tmVar 2.0 [114]	Protein, DNA, SNP	✓	✗	✓

McDonald *et al.* [73] proposed an automated variant extraction tool named VTag in 2004. The proposed CRF-based method extracts the sequence variations mentioned in the cancer-related articles and further maps them to their corresponding dbSNP identifiers. On a corpus of 105 cancer-related abstracts, the method achieves 79% recall, 85% precision, and 82% F-measure score.

Wei *et al.* [112] proposed another CRF-based model named tmVar to extract the mentioned variants from biomedical articles. The proposed model considers each component of a variant entity as one label and the variant itself as a sequence of labels. For example, tmVar retrieves each component of `c.607_608insACA` mutation separately (eg. "ins" as the mutation type, "CAA" as the alternative sequence and "`607_608`" as the position). Identification of a wide range of mutation types (DNA, protein, and SNP) is one of the key advantages of tmVar. In 2017, the second version of this tool named tmVar 2.0 [114] was proposed. This tool first extracts the variant entities using the same algorithm as tmVar and further normalizes them to their unique dbSNP identifiers.

Several methods have been proposed to extract variants from biomedical literature using the standard regular expressions algorithm (RegExp). Here, we review several of these methods which are listed in Table 6.

Horn *et al.* [49] introduced MuteXt for extracting point mutations from biomedical

Table 6: Summary of the important characteristics of the reviewed regular expressions (RegExp) approaches for variant extraction from biomedical text.

Method	Type of mutation	Gene/protein identification	Disease identification	RSID normalization
MuteXt [49]	Protein, DNA	✓	✗	✗
MEMA [82]	Protein, DNA	✓	✗	✗
MutationMiner [9]	Protein	✗	✗	✗
MuGeX [32]	Protein	✓	✓	✗
MutationFinder [23]	Protein	✗	✗	✗
Mutation GraB [67]	Protein	✓	✗	✗
EMU [29]	Protein, DNA	✓	✗	✗

literature in 2004. MuteXt uses regular expressions (RegExp) with a pattern starts with one amino acid that can be one- or three-letter term, followed by a number, and ends with another amino acid followed by the format of the first one (e.g. *G12D* or *Gly12Asp*). MuteXt is also able to extract protein names and species names from an article. The identified mutation-protein pairs are then validated in two different ways: *i) sequence filtering* and *ii) distance filtering*. The sequence filtering checks whether the reference amino acid in the mutation position is matched with the amino acid in the corresponding protein sequence. The distance filtering refers to the co-occurrence of the mutation, the protein name, and the organism type in the text. The pairs with the shortest distance (word counts) are considered as relevant. One limitation of MuteXt is that it is only trained to retrieve mutations for GPCR and NR protein superfamilies.

MEMA, proposed by Rebholz *et al.* [82], is another regular expression (RegExp) based mutation extraction tool that was only applied to the Medline abstracts. The proposed method has three main steps: *i) gene name identification*, *ii) mutation identification* and *iii) disambiguation module*. MEMA uses regular expression patterns for both gene and mutation identification. For gene identification, the method simply searches for any gene name that matches with the list of genes obtained from the Human Genome Organization (HUGO) gene database [79]. A set of 30 different patterns is used for the mutation identification. These include *Arg506 to Gln, Ile15 to Thr15, 1166 A/C, C282Y,* etc. Then, for each identified mutation, the method follows a set of certain rules in order to identify the corresponding gene:

1. If there is only one gene mentioned in the abstract, that gene would be considered as the corresponding gene.

2. If the abstract contains multiple genes, the corresponding gene would be the one mentioned in the same sentence as the mutation.

3. If there is more than one gene in the same sentence as the mutation, then the closest mentioned gene (word counts) to the mutation would be considered as the corresponding gene.

On a sample of 100 abstracts, MEMA achieves 67% recall, and 96% precision rate on the mutation extraction, and 35% recall, and precision of 93% on the mutation–gene pairs identification.

Erdogmus *et al.* [32] proposed MuGeX (Mutation Gene eXtractor) for mutation identification from the Medline abstracts. MuGeX uses a set of 20 different patterns for mutation extraction such as *G12D, Gly-12-Asp, Gly12 to Asp, Substitution of Glycine for Aspartic Acid at position 12,* etc. On a set 231 Medline abstracts the MuGeX mutation detection method achieves 85.9% and 95.9% recall and precision, respectively. For the mutation–gene pairs identification, the estimated recall, and precision is 91.3% and 88.9%, respectively. One drawback of this method is its inability to identify correct mutation-gene pairs when multiple mutations and genes mentioned in the text.

MutationFinder is another mutation extraction tool proposed by Caporaso *et al.* [23] in 2007. The proposed method uses a modified version of the regular expression method proposed by Erdogmus *et al.* [32]. These modifications include the following six new rules:

- The numeric position of the one-letter abbreviations mutation format should be greater than a certain number.
- The one-letter allele of the mutation should be presented in the upper-case format.
- The reference and alternative alleles should not be the same.
- Unlike MuteXt [49], the proposed method is able to identify mutations with the non-alphanumeric characters as well.
- MutationFinder is also able to identify mutations described in the human natural language in addition to the abbreviated formatted mutations (e.g. *Substitution of Glycine for Aspartic Acid at position 12*).
- The regular expression patterns are applied to each sentence separately.

Overall, MutationFinder had better performance (both recall and precision) compared to MuteXt [49].

Lee *et al.* [67] proposed Mutation Graph Bigram (Mutation GraB) method for extracting point mutations from biomedical literature. Similar to the previous methods in this category, mutations and gene names are identified using the pre-defined regular expression patterns. As the next step, the proposed method uses the graph-based bigram traversal method to associate an identified mutation with a corresponding protein. In particular, for each mutation entity, the method searches for the corresponding gene using the shortest path distance algorithm. The identified mutation-protein pairs are then verified based on the Swiss-Prot [16] database.

In 2010, Doughty *et al.* [29] introduced EMU, a semi-automated method for mutation-genotype-phenotype identification from biomedical literature. The proposed method follows the same regular expression patterns proposed by Garten *et al.* [39] for the mutation identification. EMU uses the HUman Genome Organization (HUGO) gene database [79] as a dictionary containing the list of human genes to extract any gene names or their synonyms from a text. Same as MuteXt [49], EMU also uses sequence filtering to validate the extracted mutation-gene pairs.

Singhal *et al.* [96] implemented a machine-learning-based method to extract and identify the disease-related mutations from biomedical literature. The proposed method uses tmVar [112] and DNorm [64] to extract mutation and disease entities, respectively. For a target disease, the proposed method uses the following 6 different features to determine whether the identified mutation *G* from an input article is related to the target disease *D*.

1. The number of times the target disease *D* is mentioned as the closest (based on word counts) disease to the identified mutation *G*.

2. The number of times the target disease *D* is mentioned in the input article.

3. The number of times the next most frequently mentioned disease other than *D* is mentioned in the input article.

4. Whether the target disease D and the mutation G are mentioned in the same sentence in the input article (binary score).
5. The sentiment score of the text between the mutation G and its nearest mention of the target disease D.
6. The subjectivity of the sentiment score was calculated in step 5.

The authors used two benchmark datasets provided by EMU [29] as the training datasets. As the next step, they used Weka3.6 tool [43] to build a machine learning classifier based on the training datasets and the developed features set. The results showed the outperformance of the proposed method compared to EMU [29].

In another work, Singhal *et al.* [97] proposed an automatic framework for extracting mutation-genotype-phenotype triplet associations from biomedical literature. The main steps of the proposed work can be summarized as follows:

1. Disease, gene, and mutation identification from an input article using DNorm [64], GNormPlus [113], and tmVar [112], respectively.
2. Disease-mutation association identification using their previous proposed method [96].
3. Gene-mutation association identification using PubMed Rank, Bing Rank, and sequence filtering methods. In particular, for an identified gene G and an identified mutation M, these scores are calculated as follows:
   - PubMed Rank: the frequency of appearances of the gene G in the abstract section of the articles that are known to be related to the mutation M.

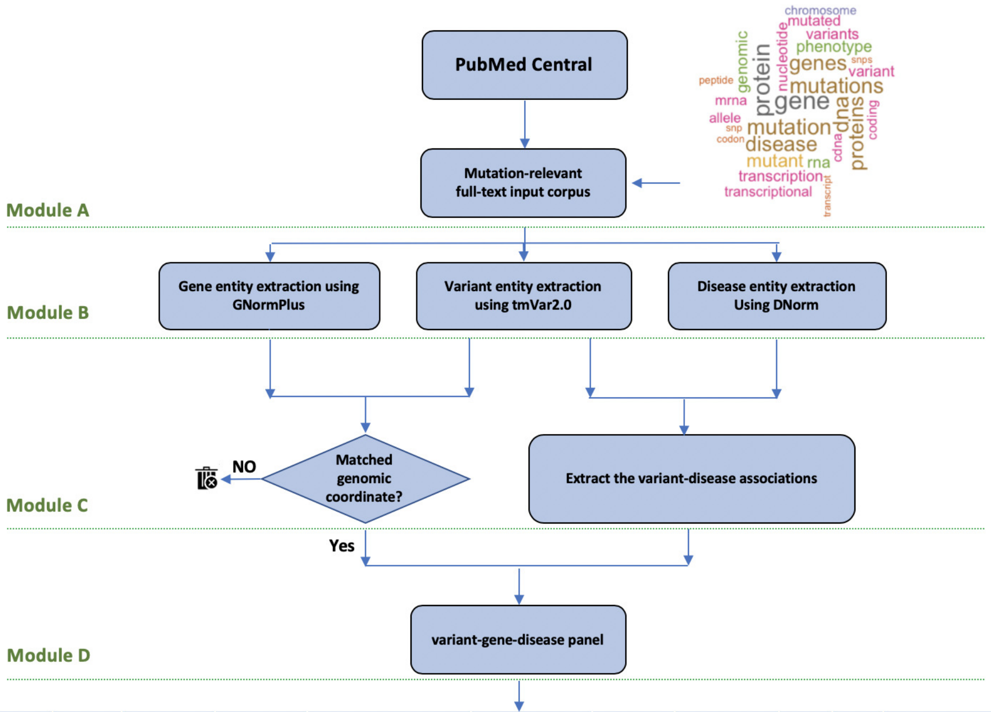
- Bing Rank: the frequency of appearances of the gene *G* in the top 20 Bing search results when searching for the mutation *M*.

- Sequence filtering: similar to the validation process proposed by Doughty *et al.* [29], the sequence filtering process checks whether the reference amino acid in the mutation position is matched with the amino acid in the associated gene's sequence.

CHAPTER 3 PROPOSED METHOD

In this thesis, we propose an automated framework to extract disease-associated variants from the full-length biomedical literature and design a variant-driven gene signature for a given disease phenotype. The process of extracting variants from a full-length article is challenging because any chemical formulae, figure numbers, etc. that are represented in a “character-number-character” format could potentially be a variant [114]. One solution to address this challenge is to mine only the variant-relevant articles. As the first step, the proposed framework employs word cloud analysis to identify such articles. The variant-gene-disease associations are then extracted from these articles using the entity recognition tools. An evidence-based variant-driven gene signature is then generated based on the mined triplet information. We use a comprehensive validation procedure to illustrate the capabilities of the proposed framework. We compare the proposed panels with other variant-driven gene panels obtained from Clinvar [63], Mastermind [40], and others from the literature [29, 96], as well as with a panel identified with a classical differentially expressed genes (DEGs) approach. The proposed variant-driven gene signatures are then validated by showing their abilities to predict the patients’ clinical conditions (healthy vs. disease) on multiple independent validation datasets.

Figure 1 illustrates the proposed framework that consists of the following four major modules: (1) obtain the full-length variant-relevant articles; (2) extract all the variant, gene and disease entities from each input article; (3) identify the variant-gene, and the variant-disease associations in each input article; (4) design a variant-driven gene panel for a given phenotype. The detailed descriptions of each step are provided in the following sections.



Variant	Gene	TYPE	RS id	Disease	PMC id	CHROM	POS	REF	ALT	Mentioned in abstract?	Mentioned in full-text?
p.Thr241Met	XRCC3	Protein Mutation	rs861539	breast cancer	PMC4269000	chr14	103699416	G	A	NO	YES
c.530A>T	DNMT3A	DNA Mutation	NA	acute myelogenous leukemia	PMC4306257	chr2	25275050	T	A	NO	YES

Figure 1: Framework overview. Module (A) obtains all the publicly available full-length articles from the PubMed Central (PMC) database. Then it uses the word cloud analysis and generate a weighted list of variant-relevant keywords. The variant-relevant articles are then selected based on the presence of this list in their full text (section 2.1). Module (B) uses GNormPlus [113], tmVar 2.0 [114] and DNorm [64] tools to extract the gene, variant, and disease phenotype entities, respectively (section 2.2). Module (C) extracts the gene-variant associations from each input article (section 2.3). This module also uses a set of features to discover the disease-variant associations (section 2.4). Module (D) generates a panel consists of the variant-gene-disease associations.

3.1 Variant-relevant input corpus

The input of the proposed framework consists of 3,322,746 full-length articles downloaded from the PMC database in January 2020. The variant indexing procedure from a full-length article is challenging because any chemical formulae, figure numbers, etc. that are represented in "Character-Number-Character" format could be identified as a variant [114]. One solution to address this challenge is to mine only the variant-relevant articles. We compare the performances of two different approaches for detecting the variant-relevant articles. The first approach considers only the articles that mention any disease or gene or any of their synonyms in the title and abstract sections [40]. In the second approach, we employ the word cloud analysis and generate a weighted list of variant-relevant keywords. In particular, we first generate a weighted list of words (referred to as variant-relevant keywords) that appear frequently in the full-body text of 10,000 random articles with at least one mentioned variant (using tmVar 2.0). Subsequently, an article is considered to be relevant to variants if at least 10% of these keywords appear in the full-body of the article. We apply both approaches on a new set of 10,000 random full-length articles. Figure 2 shows the identified variant overlaps and differences between the two approaches.

The number of papers with at least one mentioned variant overlapped between the two approaches is 836 and the number of overlapped variants is 5,476. The number of variants that are only found by the first approach is 284 from 91 papers, in which a manual validation process revealed that 97% of them are false positive (extracted entity is not a variant and it is wrongly identified as a variant.). The number of variants that are only

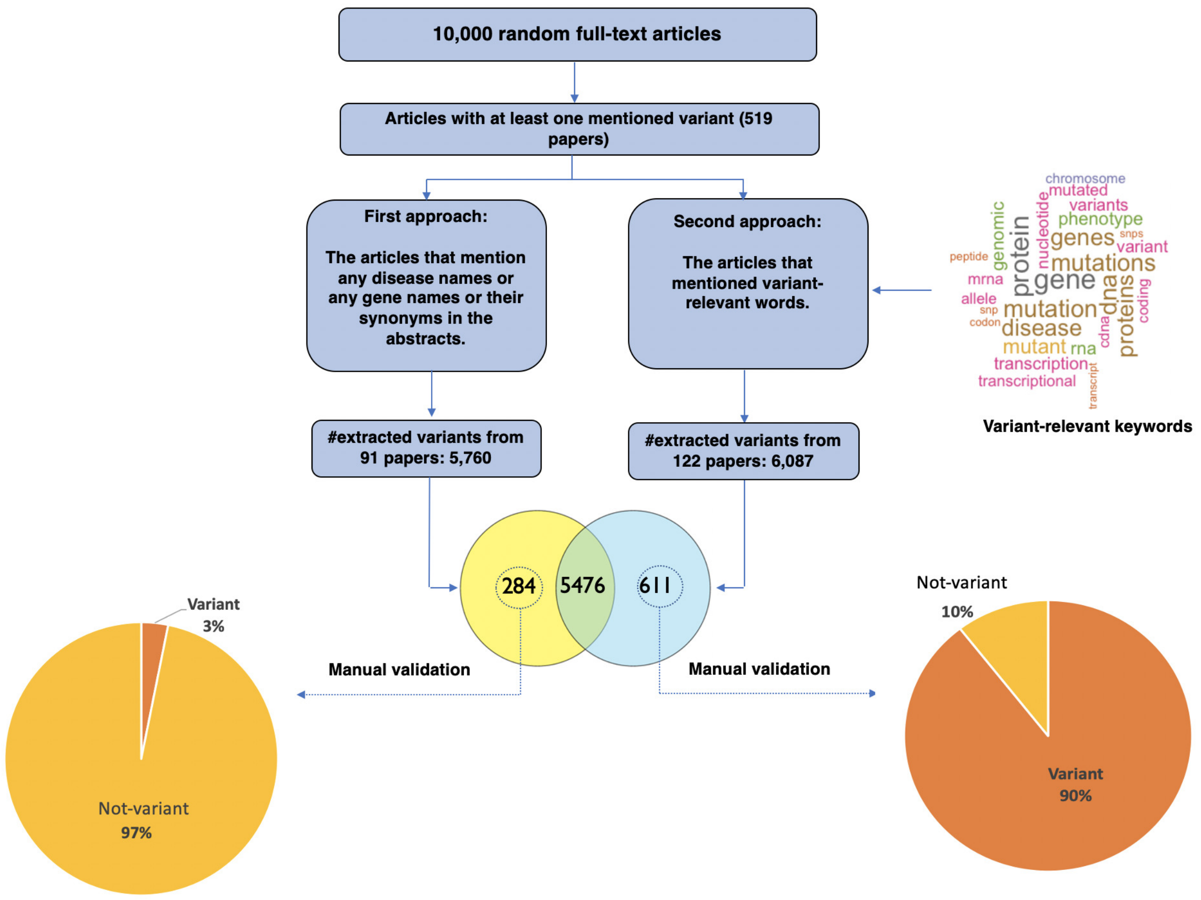

10,000 random full-text articles
Articles with at least one mentioned variant (519 papers)
First approach:
The articles that mention any disease names or any gene names or their synonyms in the abstracts.
Second approach:
The articles that mentioned variant-relevant words.
chromosome
mutated
variants
phenotype
genomic
protein
nucleotide
genes
snps
variant
peptide
mutations
mrna
gene
dna
proteins
coding
allele
snp
mutation
codon
disease
cdna
mutant
rna
transcription
transcriptional
transcript
Variant-relevant keywords
#extracted variants from 91 papers: 5,760
#extracted variants from 122 papers: 6,087
284
5476
611
Manual validation
Manual validation
Variant
3%
Not-variant
97%
Not-variant
10%
Variant
90%

Figure 2: Among the 10,000 random articles, the articles with at least one mentioned mutation are selected (using tmVar 2.0). We compare the performances of two different approaches for detecting the variant-relevant articles. The first approach identifies articles that mention any disease or gene or any of their synonym in their titles and abstracts [40]. In the second approach, we only search for the articles that mention the variant-relevant keywords in their full-body text. The variant-relevant keywords list is a weighted list of the words that appear frequently in a set of 10,000 random articles with at least one mentioned variants (using tmVar 2.0). Subsequently, an article is considered to be relevant to variants if at least 10% of these variant-relevant keywords are appearing in the full-body text. The number of variants that are found in the articles selected by the first approach and the second approach is 5,760 and 6,087, respectively. The number of variants identified by both approaches is 5,476. The number of variants that are only found by the first approach is 284, of which 97% are false positive (extracted entity is not a variant and it is wrongly identified as a variant.). The number of variants that are only found by the second approach is 611, of which only 10% are false positive. These results show that the second approach which is based on the variant-relevant keywords outperforms the first approach.

found by the second approach is 611 from 122 papers, in which only 10% of them are false positive. The manual validation of the extracted variants to listed in Appendix Table 33. These results show that the second approach which is based on the variant-relevant keywords outperforms the first approach. This leads us to the conclusion that the second approach performs better in terms of the ability to index the variant-relevant articles. This approach results in a list of 1,274,775 full-length articles that contain genomic variants.

3.2 Extract the variant, gene and disease entities

We use the publicly available and well-known entity recognition tools to extract the variant, gene, and disease phenotype from each input article. In particular, we use GNormPlus [113] to identify the appropriate genes. The tmVar 2.0 [114] is the tool we employ for extracting the variants and normalizing those which are included in dbSNP to their unique identifiers (dbSNP RSIDs). We use DNorm [64] to identify all the disease phenotypes mentioned in an article.

3.3 Extract the variant-gene associations

Once a variant is extracted from an input article, we follow the steps provided by Wei *et al.* [114] to find the associated gene. Then, we map each retrieved variant-gene pair to the corresponding genomic coordinates (chromosome number, position, reference and alternative alleles) using the Variant Recoder [50] tool. Variant Recoder provides translation between the different formats of a variant. This tool supports HGVS annotations as well as dbSNP, Clinvar [63], and PharmGKB [47]. We eliminate the variant-gene associations with no matched genomic coordinates (referred to as false positive pairs).

3.4 Extract the variant-disease associations

We use a set of features to capture the variant-disease associations from an input article adapted from tmVar [96]. Let $C = \{V, D_1, D_2, ..., D_{\mathrm{k}}\}$ be a collection of appearances of the variant *V* and the closest (based on the word counts) mentioned diseases in an article, where *k* is the number of times this variant is mentioned in that article. The disease association score is calculated for each appearance of variant *V* and the closest mentioned disease D_{i}, where 1≤i≤k. This score is the summation of the following set of scores:

- The Same Sentence Occurrence (SSO) is a binary score which is 1 when the variant *V* and the disease D_{i} are mentioned in the same sentence and 0 otherwise.
- The Same Paragraph Occurrence (SPO) is a binary score which is 1 when the variant *V* and the disease D_{i} are mentioned in the same paragraph and 0 otherwise.
- The sentiment score (SS) calculates the polarity sentiment value for the text mentioned between the variant *V* and the disease D_{i}. We use the R package "sentimentr" [86] for this analysis.

The variant *V* is considered to be associated with disease D_{i} that has the highest disease association score.

We also performed an experiment to compare the performance of the proposed scoring method for extracting the variant-disease associations with the simple sentence co-occurrence scoring method. In this experiment, we used two manually curated benchmark datasets provided by Doughty *et al.* [29]. These datasets contain variant-disease pairs extracted from 29 and 129 PubMed articles for prostate cancer and breast cancer, respec-

Table 7: Comparison of the proposed variant-disease association scoring method with the baseline approach (Co-occurrence only) on the benchmark datasets. These datasets are provided by Doughty *et al.* [29]. The proposed approach performs better compare to the baseline approach.

Corpus	Evaluation metrics	Proposed method	Baseline method
Breact cancer	Precison	**0.90385**	0.31731
	Recall	**0.85455**	0.30000
	F1 measure	**0.87850**	0.30841
Prostate cancer	Precison	**0.91111**	0.37778
	Recall	**0.85417**	0.35417
	F1 measure	**0.88172**	0.36559

tively. We used these datasets and reported the standard evaluation metrics (precision, recall and F1-measure) for the proposed scoring approach compared to the sentence co-occurrence scoring approach. As shown in Table 7, the proposed method outperforms the baseline method which is only based on the sentence co-occurrence appearance of the variant-disease pairs. The complete list of mined variant-disease pairs for this experience is listed in Appendix Table 32.

3.5 Variant-driven gene panel design

In this step, we first generate a variant-gene-disease panel which includes all the associations between the gene, variant, and disease entities extracted from the input corpus (Module D in Figure 1). This panel includes 18,254 genes with 313,780 variants discovered to be associated with 5,202 unique diseases. For a given disease, we then generate the variant-driven gene panel which includes all the genes with at least one mentioned variant discovered to be associated with the given disease.

3.6 Validation method

In this section, we describe two experiments performed to assess the diagnostic value of the proposed variant-driven gene panel.

In the first experiment, we use the genes present in the proposed panel to predict the patients' clinical condition (healthy vs. disease) from several independent patient cohorts. The hypothesis is that a better gene panel will yield better classification results. For this purpose, we use disease gene expression datasets and machine-learning classification techniques. A disease gene expression dataset is a matrix in which the rows represent the measured genes and the columns represent the samples (healthy or disease individual). The value in each cell is the expression level of a gene in a particular sample. We use cross-validation method for this analysis. In particular, in each round of sampling, we use one of the gene expression datasets as the training dataset and we use the rest as the testing datasets. We use the genes present in the proposed variant-driven gene panel along with their expression values from the training dataset to build a random forest classifier [20]. Then, we apply the trained classifier on each of the testing datasets in order to predict the patients' clinical outcomes. We use the area under the curve (AUC) of the receiver-operator characteristic (ROC) to assess the performance of the classifier. We repeat this procedure n times (where n is the number of available gene expression datasets). An average of the AUCs is calculated over the n rounds of sampling. This procedure is used to compare the diagnostic quality of the proposed gene panel with the current available variant-relevant gene panels.

In the second experiment, we assess the relevance of the proposed gene panel to a given

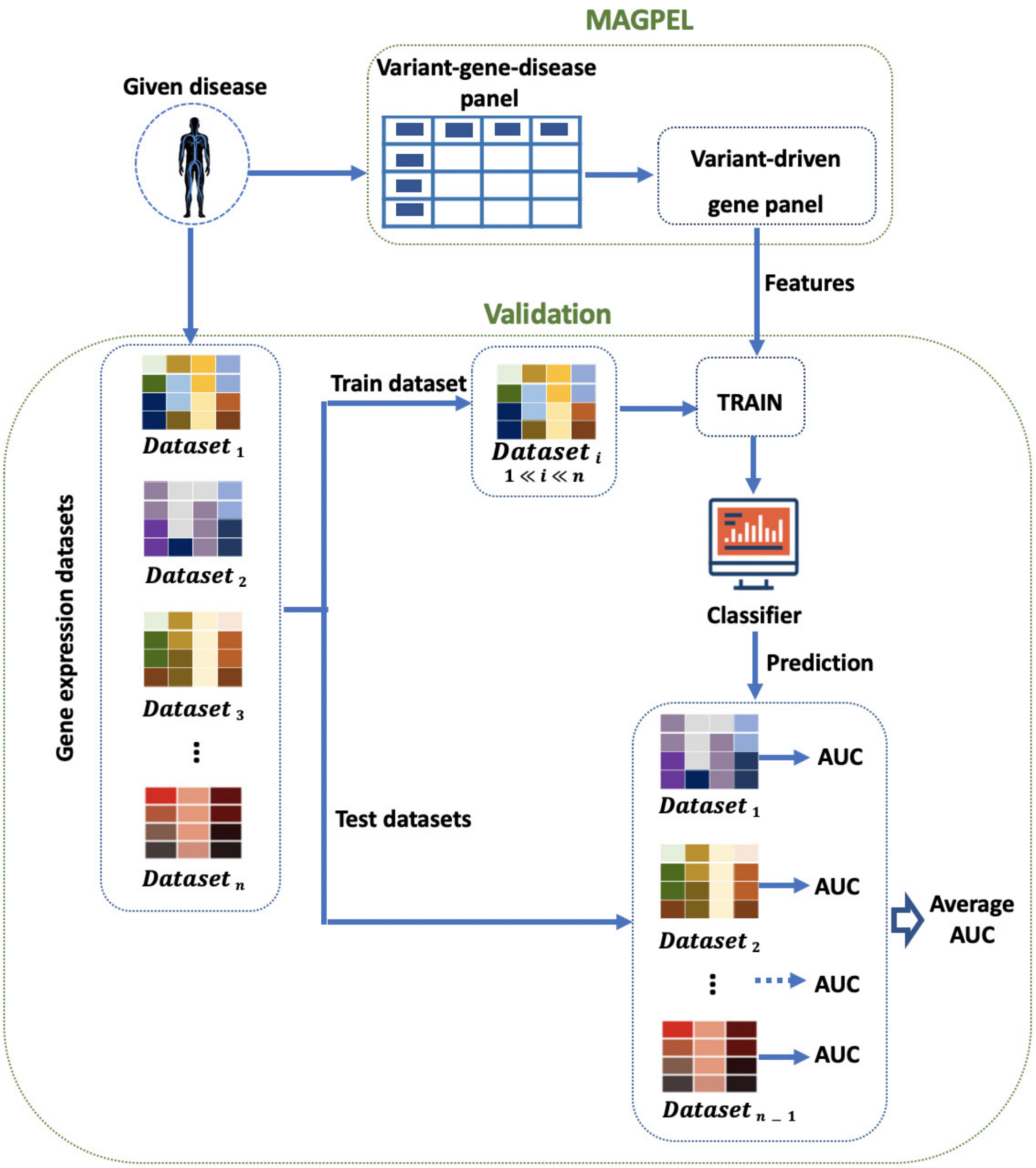


Figure 3: Validation framework overview. Module (A) identifies all the genes with at least one variant discovered to be associated with the given disease by the proposed framework. We refer to this list of genes as the proposed variant-driven gene panel. Module (B) first analyzes several independent gene expression datasets studying the given phenotype. We use a cross-validation method. In each round of sampling, we use one of the gene expression datasets as the training dataset and we use the rest as the testing datasets. We use the expression values of the genes included in the proposed gene panel as the features to build a classifier. Then, we apply the trained classifier on each of the testing datasets in order to predict the patients' clinical outcomes in each testing dataset. We use the area under the curve (AUC) of the receiver-operator characteristic to assess the performance of the classifier. We repeat this procedure n times (where n is the number of gene expression datasets). An average of AUCs is calculated over the n rounds of sampling. This procedure is used to compare the diagnostic quality of the proposed variant-driven gene panel with the current available variant-relevant gene panels.

disease based on the rank of the target pathway when an enrichment pathway analysis is performed. A “target pathway” refers to the pathway that was created to explain the mechanism of the given disease (e.g. the acute myeloid leukemia KEGG pathway (hsa05221) is the target pathway for acute myeloid leukemia).

A signaling pathway refers to a graph in which nodes represent genes/proteins, and edges represent existing interactions between such genes or proteins. In general, the main goal of the pathway analysis methods is the correct identification of the pathways that are significantly impacted when comparing two phenotypes (*e.g.* healthy vs. disease) [30, 58]. Many pathway analysis methods have been proposed [75, 76, 58]. A very recent extensive benchmarking of the existing pathway analysis methods are provided by Nguyen *et al.* [77].

In this thesis, we use the enrichment pathway analysis method called over-representation analysis (ORA) [57]. The goal of this method is to find the pathways that are enriched within a list of genes. In particular, this method calculates the probability of finding a center number of gene overlaps between the proposed gene panel and the presented genes in each pathway just by chance. For a pathway P, this probability is calculated as follows:

$$p\text{-}value = 1 - \sum_{i=0}^{k\text{-}1} \frac{\binom{M}{i}\binom{N\text{ - }M}{n\text{ - }i}}{\binom{N}{n}} \tag{3.1}$$

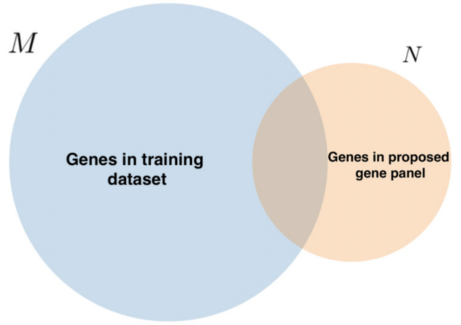
In this equation, N is the total number of genes in the genome that have been annotated, n is the total number of genes in the proposed gene panel, k is the total number of gene overlaps between the proposed gene panel and the pathway P, and M is the total number of genes included in the pathway P.

These probability values are calculated for all pathways. Subsequently, they should be adjusted for multiple comparisons with an approach such as the false discovery rate correction (FDR) [13, 14]. For each pathway, if the FDR-corrected *p*-value is less than a certain threshold (usually less than 0.05), then the pathway is considered to be significantly involved in the experiment. The list of significant pathways is then ranked from the one with the lowest FDR-corrected *p*-value (most significant) to the one with the highest *p*-value (least significant). For this analysis, we use the R package "clusterProfiler v3.12.0" [120]. The expectation here is that a gene panel that is relevant to the given disease would rank the target pathway at the very top of the ranked list of pathways. This validation method was widely adopted by others, such as [5, 51, 69, 71, 75, 76, 77, 91, 102]. We also provide the top 10 significantly enriched pathways and the references explaining the association of the respective pathways to the disease case study for each gene panel.

CHAPTER 4 RESULTS

As representative examples, we present the results for acute myeloid leukemia (AML), breast cancer, and prostate cancer. The resulted gene panel proposed for each case study is included in the Appendix. All the gene expression datasets used in this manuscript for the classification analysis are obtained from GEO [12].

For each disease case study, we also calculate the percentage of the genes in the proposed gene panel that overlap with the genes in each gene expression dataset. We performed the following experiment as a quality check to ensure that the majority of the genes in the proposed gene panel are contributing to the validation analysis. In order to do this, we calculated the percentage of the genes in the proposed gene panel that overlap with the genes in the training dataset as follows:



$$P = \frac{|N \cap M|}{|N|}$$

In this equation, *N* represents the genes in the proposed gene panel and *M* represents the genes in the training gene expression dataset. For each case study, the average of this percentage across all the gene expression datasets is more than 80% (Tables 8 to 10).

Table 8: The percentage of the genes in each AML gene panel that overlap with the genes in each GEO gene expression dataset.

Dataset	MAGPEL (proposed)	Mastermind	Clinvar [63]	Singhal *et al.* [97]
GSE15061	96.94	97.44	88.68	90.24
GSE17054	96.51	97.12	88.68	90.24
GSE34577	97.38	98.4	88.68	90.24
GSE35008	92.58	95.53	86.79	87.8
GSE37307	90.39	95.21	79.25	84.15
GSE42140	96.51	97.12	88.68	90.24
GSE9476	90.39	95.21	79.25	84.15
GSE982	90.39	95.21	79.25	84.15
Average	93.88	96.40	84.90	87.65

Table 9: The percentage of the genes in each prostate cancer gene panel that overlap with the genes in each GEO gene expression dataset.

Dataset	MAGPEL (proposed)	EMU [29]	Clinvar [63]	Singhal *et al.* [97]
GSE12348	86.28	100.00	68.01	85.87
GSE17906	97.18	100.00	89.08	94.35
GSE17951	97.18	100.00	89.08	94.35
GSE32448	97.18	100.00	89.08	94.35
GSE46602	97.18	100.00	89.08	94.35
GSE55945	97.18	100.00	89.08	94.35
GSE68882	75.56	82.35	50.00	72.44
GSE6956	86.28	100.00	68.01	85.87
GSE70768	98.31	100.00	94.44	94.35
Average	92.48	98.04	80.65	90.03

Table 10: The percentage of the genes in each breast cancer gene panel that overlap with the genes in each GEO gene expression dataset.

Dataset	MAGPEL (proposed)	EMU [29]	Clinvar [63]	Singhal *et al.* [97]
GSE10780	97.66	100.00	85.64	86.90
GSE10810	75.05	72.73	50.21	61.90
GSE20086	97.66	100.00	85.64	86.90
GSE29431	97.66	100.00	85.64	86.90
GSE36295	94.74	95.45	85.47	85.69
GSE42568	97.66	100.00	85.64	86.90
GSE54002	97.66	100.00	85.64	86.90
GSE61304	97.66	100.00	85.64	86.90
GSE86374	94.74	95.45	85.47	85.69
GSE8977	92.01	95.45	73.83	79.44
Average	94.25	95.91	80.88	83.41

4.1 Acute myeloid leukemia

First, we extract all the genes with at least one mentioned variant discovered to be associated with AML by the proposed framework (Table 29). The top 10 genes that have the highest number of variants are TP53, FLT3, KIT, DNMT3A, IDH1, COX8A, RUNX1, TYMS, NPM1, and SLC29A1. These genes play significant roles in the underlying mechanisms of AML. For instance, Kadia *et al.* [56] demonstrated that AML patients with TP53 alterations have a lower response rate to intensive chemotherapy and therefore have an inferior survival rate. FLT3 and C-KIT are known to be associated with poor AML prognosis discovered by Pratz *et al.* [80] and Yang *et al.* [117], respectively. Ley *et al.* [68] investigated the role of DNMT3A and found that there is a direct link between the presence of mutations in this gene and the intermediate risk of AML. Chaturvedi *et al.* [27] also reported the therapeutic role of mutant IDH1 in AML. Gaidzik *et al.* [37] have shown that therapy-resistance and inferior outcomes are the main genetic characteristics of AML patients with RUNX1 mutations. The presence of mutations in TYMS and NPM1 is also discovered in AML patients [38, 70]. SLC29A1 mutations are found to be associated with poor therapy outcome in AML patients [59].

We assess the utility of the proposed gene panel on independent gene expression datasets studying AML obtained from GEO [12]. Dataset summaries are described in Table 11.

The other variant-driven gene panels which are available for AML are obtained from Clinvar [63], Mastermind [40], and the panel proposed by Singhal *et al.* [97] Clinvar is a repository for mutations and their associated disease phenotypes which are man-

Table 11: Summary of the datasets used for the AML case study.

Dataset	**Title**	**#Disease samples**	**#Control samples**
GSE15061	Gene array prediction of AML transformation in MDS	202	69
GSE17054	Dysregulated gene expression networks in human acute myelogenous leukemia stem cells	9	4
GSE2191	pediatric AML and normal bone marrow	54	4
GSE34577	Routine use of microarray-based gene expression profiling to identify patients with low cytogenetic risk acute myeloid leukemia	21	8
GSE35008	Expression data from human hematopoietic stem and progenitor compartments from patients with acute myeloid leukemia with normal karyotype and healthy controls	12	16
GSE37307	Aberrant expressed genes in AML	30	19
GSE42140	Gene expression in signaling subsets of AML blasts induced by G-CSF	33	7
GSE9476	Abnormal expression changes in AML	26	38
GSE982	Gene Expression-Based High Throughput Screening: HL-60 Cell Treatment with Candidate Compounds	9	6

ually curated from the biomedical literature. The Mastermind search engine provides literature-based variant-genotype-phenotype association information. We also include the results when using only the differentially expressed genes (FDR-corrected p-value$<$0.05 and $|\log_2(\text{fold change})|>=$1.5) as a gene panel. Figure 4 illustrates the performance comparison of these gene panels. The results show that the classification based on the proposed gene panel achieves the best result (the highest median AUC value) and outperforms the classification based on all the other published panels.

The results for the pathway enrichment analysis are summarized in Table 12. The proposed gene panel has better performance than the other available panels and ranked the AML target pathway as the top-ranked pathway. The top 10 significantly enriched pathways and the references explaining the association of the respective pathways to AML for each gene panel are summarized in (Tables 13 to 16).

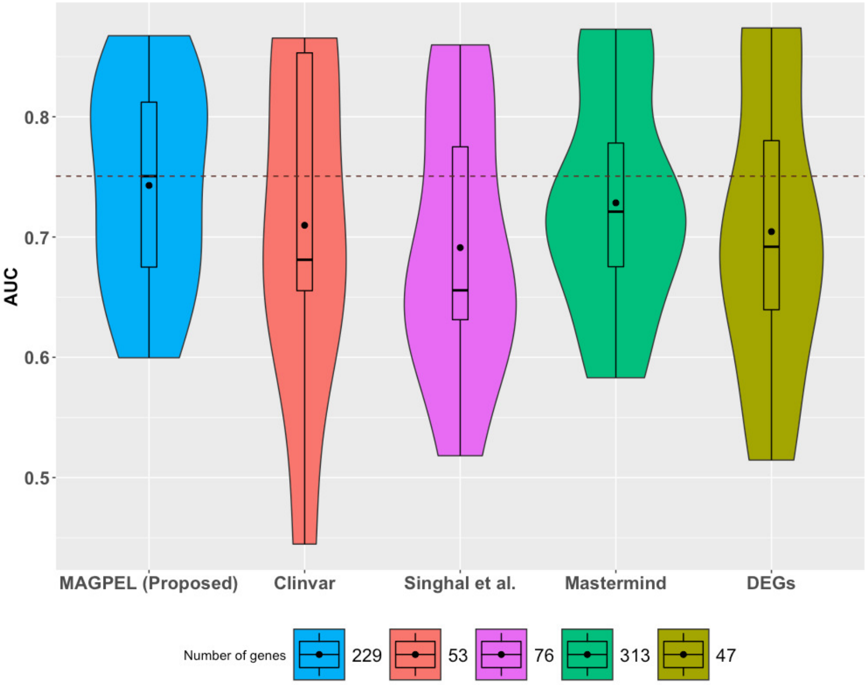


Figure 4: The diagnostic performances of the random forest classifier based on five different gene panels. In this figure, the proposed panel (blue panel) performs better than the ones obtained from Clinvar (red panel), Mastermind [40] (purple panel), the panel proposed by Singhal *et al.* [97] (green panel), and the differentially expressed genes (FDR-corrected p-value$<$0.05 and $|\log_2$(fold change)$|>=$1.5) (DEGs) (olive-tone panel) in terms of the ability to distinguish between healthy volunteer and the AML patients. In this figure, the black dot inside each box plot represents the mean AUC value and the dashed line represents the highest median AUC value.

Table 12: The results of the pathway enrichment analysis based on four different gene panels for AML. The comparison is based on the rank of the acute myeloid leukemia KEGG pathway (hsa05221). The proposed panel performs better in terms of the ability to highly rank the target pathway.

Panel	Number of genes	Rank of target pathway	*p*-value (FDR)
MAGPEL (proposed)	229	**1**	1.57E-15
Clinvar [63]	53	2	8.36E-07
Singhal *et al.* [97]	76	3	1.62E-14
Mastermind [40]	313	9	5.50E-26

Table 13: The top 10 significantly enriched pathways identified by the enrichment pathway analysis based on the proposed gene panel (MAGPEL) for AML case study. Rows with green background indicate the target pathway for AML. Rows with blue background indicate pathways for which we found indication of their association to AML.

Rank	Pathway name	*p*-value (FDR)
1	Acute myeloid leukemia	1.57E-15
2	PI3K-Akt signaling pathway [1, 61]	2.42E-11
3	Transcriptional misregulation in cancer [88]	1.44E-10
4	Prostate cancer	1.44E-10
5	ErbB signaling pathway [109]	1.44E-10
6	Chronic myeloid leukemia [90]	2.26E-10
7	PD-L1 expression and PD-1 checkpoint pathway in cancer	2.26E-10
8	JAK-STAT signaling pathway [33]	1.26E-09
9	Hepatitis B	1.26E-09
10	EGFR tyrosine kinase inhibitor resistance	3.18E-09

Table 14: The top 10 significantly enriched pathways identified by the enrichment pathway analysis based on the Clinvar [63] gene panel for AML case study. Rows with green background indicate the target pathway for AML. Rows with blue background indicate pathways for which we found indication of their association to AML.

Rank	Pathway name	*p*-value (FDR)
1	PI3K-Akt signaling pathway [1]	2.53E-07
2	Acute myeloid leukemia	8.36E-07
3	Central carbon metabolism in cancer [18]	8.36E-07
4	PD-L1 expression and PD-1 checkpoint pathway in cancer	3.76E-06
5	Thyroid cancer	1.54E-05
6	Bladder cancer	1.89E-05
7	Chronic myeloid leukemia [90]	1.89E-05
8	EGFR tyrosine kinase inhibitor resistance	2.08E-05
9	Endometrial cancer	8.42E-05
10	Non-small cell lung cancer	0.00014

Table 15: The top 10 significantly enriched pathways identified by the enrichment pathway analysis based on the Singhal *et al.* [97] gene panel for AML case study. Rows with green background indicate the target pathway for AML. Rows with blue background indicate pathways for which we found indication of their association to AML.

Rank	Pathway name	*p*-value (FDR)
1	JAK-STAT signaling pathway [33]	2.13E-17
2	Chronic myeloid leukemia [90]	4.70E-15
3	Acute myeloid leukemia	1.62E-14
4	Human T-cell leukemia virus 1 infection	1.44E-10
5	Transcriptional misregulation in cancer [88]	1.53E-09
6	Hepatitis B	2.39E-09
7	PI3K-Akt signaling pathway [1]	2.39E-09
8	Non-small cell lung cancer	4.48E-09
9	Central carbon metabolism in cancer [18]	6.29E-09
10	Pancreatic cancer	1.51E-08

Table 16: The top 10 significantly enriched pathways identified by the enrichment pathway analysis based on the Mastermind [40] gene panel for AML case study. Rows with green background indicate the target pathway for AML. Rows with blue background indicate pathways for which we found indication of their association to AML.

Rank	Pathway name	*p*-value (FDR)
1	Hematopoietic cell lineage	6.98E-37
2	JAK-STAT signaling pathway [33]	1.11E-34
3	PI3K-Akt signaling pathway [1]	6.78E-34
4	Transcriptional misregulation in cancer[88]	1.97E-30
5	Hepatitis B	2.60E-30
6	Human T-cell leukemia virus 1 infection	8.14E-30
7	Epstein-Barr virus infection	7.66E-29
8	Kaposi sarcoma-associated herpesvirus infection	1.53E-27
9	Acute myeloid leukemia	5.50E-26
10	Prostate cancer	6.83E-26

4.2 Prostate cancer

In this case study, we discover 532 genes with variants associated with prostate cancer (Table 30). The proposed prostate cancer variant-driven gene panel contains several genes known to be involved in prostate cancer development and progression. For instance, the androgen receptor (AR) plays important role in prostate cancer cell proliferation as demonstrated by Balk *et al.* [10] The mutated BRCA2, TP53, KLK3, and RNASEL genes are directly associated with the risk of developing prostate cancer [108, 31, 60, 26]. SPOP is the most frequent mutated gene in primary prostate cancer [11, 19].

The gene expression dataset summaries are described in the Table 17.

Table 17: Summary of the datasets used for prostate cancer case study.

Dataset	**Title**	**#Disease samples**	**#Control samples**
GSE12348	Prostate cancer cell lines and normal prostate epithelial and stromal cells in primary culture	6	3
GSE17906	Gene expression down-regulation in prostate tumor-associated stromal cells involves organ-specific genes	10	10
GSE17951	Gene expression analysis of prostate cancer samples using Affymetrix U133Plus2 array	68	13
GSE32448	CPDR tumor-benign 80 genechip dataset	40	40
GSE46602	Expression data from prostate cancer and benign prostate glands	36	14
GSE55945	Gene expression profiling of prostate benign and malignant tissue	13	8
GSE68882	Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease	23	3
GSE6956	Tumor immunobiological differences in prostate cancer between african-american and european-american men	69	18
GSE70768	Prostate cancer stratification using molecular profiles	125	74

The classification results also demonstrate that the proposed gene panel outperforms the other available gene panels [63, 97, 29] in terms of the ability to predict the patients' clinical outcome on several independent validation cohorts (Figure 5).

The results for the pathway enrichment analysis are summarized in Table 18. The

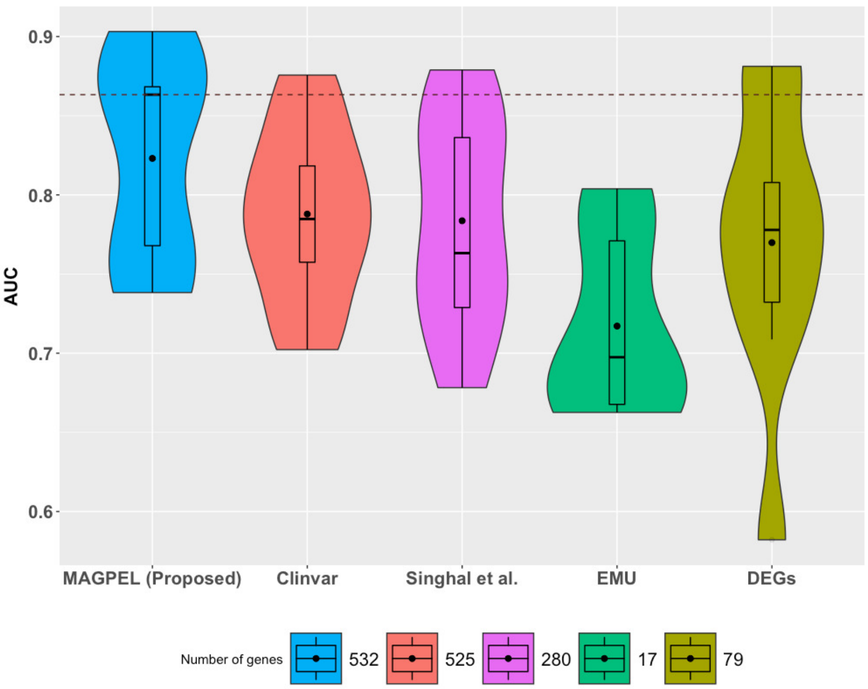


Figure 5: The diagnostic performances of the random forest classifier based on five different gene panels. In this figure, the proposed panel (blue panel) performs better than the ones obtained from Clinvar (red panel), the panels proposed by Singhal *et al.* [97] (purple panel), EMU [29] (green panel), and also the differentially expressed genes (FDR-corrected p-value$<$0.05 and $|\log_2(\text{fold change})|>=$1.5) (DEGs) (olive-tone panel) in terms of the ability to distinguish between healthy volunteer and the breast cancer patients. In this figure, the black dot inside each box plot represents the mean AUC value and the dashed line represents the highest median AUC value.

top 10 significantly enriched pathways and the references explaining the association of the respective pathways to prostate cancer for each gene panel are also summarized in (Tables 19 to 22).

Table 18: The results of the enrichment pathway analysis based on different gene panels obtained for prostate cancer. The comparison is based on the rank of the prostate cancer KEGG pathway (hsa05215).

Panel	Number of genes	Rank of target pathway	*p*-value (FDR)
MAGPEL (Proposed)	532	**1**	5.49E-28
Clinvar [63]	525	7	5.10E-05
Singhal *et al.* [97]	280	**1**	8.12E-12
EMU [29]	17	2	5.83E-07

Table 19: The top 10 significantly enriched pathways identified by the enrichment pathway analysis based on the proposed gene panel (MAGPEL) for prostate cancer case study. Rows with green background indicate the target pathway for prostate cancer. Rows with blue background indicate pathways for which we found indication of their association to prostate cancer.

Rank	Pathway name	*p*-value (FDR)
1	Prostate cancer	5.49E-28
2	FoxO signaling pathway [92]	3.13E-25
3	Endocrine resistance	2.29E-21
4	Colorectal cancer	2.29E-21
5	Pancreatic cancer	3.54E-21
6	PI3K-Akt signaling pathway [94]	2.35E-20
7	AGE-RAGE signaling pathway in diabetic complications	4.05E-19
8	Endometrial cancer	3.96E-18
9	Chronic myeloid leukemia	6.04E-18
10	Bladder cancer	1.51E-17

Table 20: The top 10 significantly enriched pathways identified by the enrichment pathway analysis based on the Clinvar [63] for prostate cancer case study. Rows with green background indicate the target pathway for prostate cancer. Rows with blue background indicate pathways for which we found indication of their association to prostate cancer.

Rank	Pathway name	*p*-value (FDR)
1	Pancreatic cancer	1.93E-06
2	Endometrial cancer	1.96E-06
3	Melanoma	1.69E-05
4	Endocrine resistance	1.69E-05
5	Breast cancer	2.71E-05
6	Non-small cell lung cancer	2.74E-05
7	Prostate cancer	5.10E-05
8	Colorectal cancer	6.29E-05
9	Glioma	7.57E-05
10	Bladder cancer	8.38E-05

Table 21: The top 10 significantly enriched pathways identified by the enrichment pathway analysis based on the EMU [29] for prostate cancer case study. Rows with green background indicate the target pathway for prostate cancer. Rows with blue background indicate pathways for which we found indication of their association to prostate cancer.

Rank	Pathway name	*p*-value (FDR)
1	Endometrial cancer	4.96E-08
2	Prostate cancer	5.83E-07
3	Gastric cancer	5.13E-06
4	Colorectal cancer	8.28E-06
5	Platinum drug resistance	0.00015237
6	Hepatocellular carcinoma	0.00015268
7	Thyroid cancer	0.00054001
8	Bladder cancer	0.00064535
9	Breast cancer	0.00133837
10	Hepatitis C	0.00147899

Table 22: The top 10 significantly enriched pathways identified by the enrichment pathway analysis based on the Singhal *et al.* [97]for prostate cancer case study. Rows with green background indicate the target pathway for prostate cancer. Rows with blue background indicate pathways for which we found indication of their association to prostate cancer.

Rank	Pathway name	*p*-value (FDR)
1	Prostate cancer	8.12E-12
2	Hepatitis B	1.82E-08
3	Platinum drug resistance	2.85E-07
4	Bladder cancer	2.85E-07
5	FoxO signaling pathway [92]	3.12E-07
6	Steroid hormone biosynthesis	1.66E-06
7	Pancreatic cancer	1.95E-06
8	Transcriptional misregulation in cancer	1.95E-06
9	PI3K-Akt signaling pathway [94]	3.08E-06
10	Endometrial cancer	5.58E-06

4.3 Breast cancer

The resulted panel for breast cancer includes 513 genes. This panel contains several genes that are known to play crucial roles in the underlying mechanisms of breast cancer. For instance, BRCA1, BRCA2, TP53, ESR1, PIK3CA, ERBB2, and PALB2 are among the genes with a high number of variants associated to breast cancer. The mutations in BRCA1, BRCA2, and TP53 are well-known to be associated with a high breast cancer risk [35, 111]. ESR1 mutations are involved in hormone-resistant metastatic breast cancer [87, 107, 48, 36, 54]. PIK3CA is an oncogene in breast cancer [22, 7, 98, 52] and ERBB2 is shown to be up-regulated in several breast tumors [45, 110, 116, 84]. PALB2 is also reported as one of the breast cancer susceptibility genes [81, 6, 106, 121].

The gene expression dataset summaries are described in Table 23.

Table 23: Summary of the datasets used for breast cancer case study.

Dataset	**Title**	**#Disease samples**	**#Control samples**
GSE10780	Proliferative genes dominate malignancy-risk gene panel in histologically-normal breast tissue	42	143
GSE10810	Gene expression panels in breast cancer distinguish phenotype charact., histological subtypes, and tumor invasivness	31	27
GSE20086	Heterogeneity of gene expression in stromal fibroblasts of human breast carcinomas and normal breast	6	6
GSE29431	Identifying breast cancer biomarkers	25	12
GSE36295	Transcriptomic analysis of breast cancer	45	5
GSE42568	Breast cancer gene expression analysis	67	17
GSE54002	Gene expression profiling of LCM captured breast cancer cells	417	16
GSE61304	Novel bio-marker discovery for stratification and prognosis of breast cancer patients	56	4
GSE86374	Analysis of somatic DNA copy number alterations and frequency of breast cancer intrinsic subtypes from Mexican women	50	36
GSE8977	Bone-marrow-derived mesenchymal stem cells promote breast cancer metastasis	7	15

We compare our panels with several other previously proposed variant-driven breast cancer gene panels as follows: i) Clinvar [63], ii) Singhal *et al.* [97], iii) Doughty *et al.* [29] and iv) the classical DEGs. The classification results demonstrate that the gene panel proposed here performs better than the other gene panels in terms of the ability to predict the patients' clinical outcome on several independent validation datasets (Figure 6).

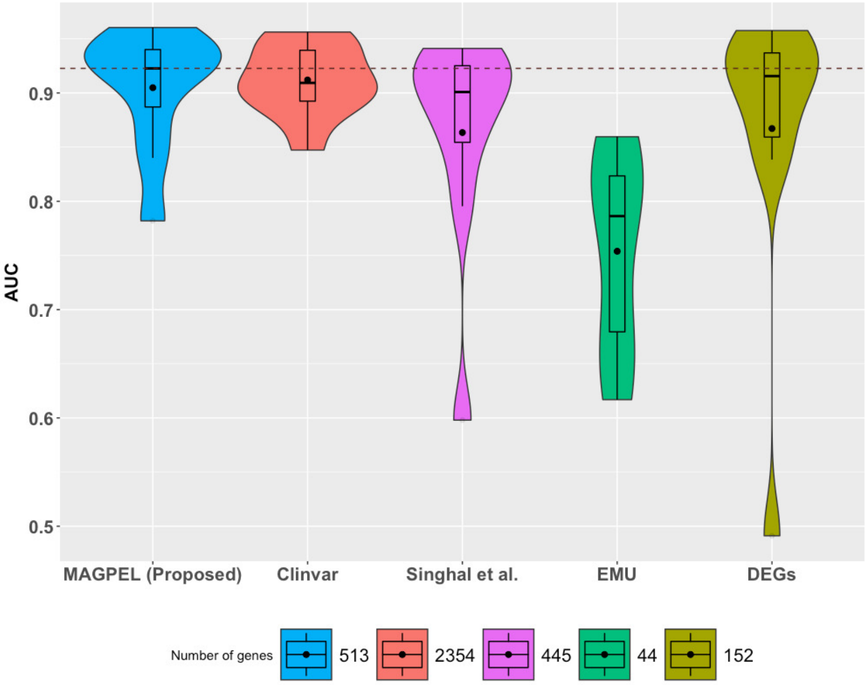


Figure 6: The diagnostic performances of the random forest classifier based on five different gene panels. In this figure, the proposed panel (blue panel) performs better than the ones obtained from Clinvar (red panel), the panels proposed by Singhal *et al.* [97] (purple panel) and Doughty *et al.* [29] (green panel), and also the differentially expressed genes (FDR-corrected p-value<0.05 and $|\log_2(\text{fold change})|>=1.5$) (DEGs) (olive-tone panel) in terms of the ability to distinguish between healthy volunteer and the breast cancer patients. In this figure, the black dot inside each box plot represents the mean AUC value and the dashed line represents the highest median AUC value.

The results for the pathway enrichment analysis are summarized in Table 24. The top 10 significantly enriched pathways and the references explaining the association of the respective pathways to breast cancer for each gene panel are summarized in (Tables 25 to 28).

Table 24: The results of the enrichment pathway analysis based on different gene panels obtained for breast cancer. The comparison is based on the rank of the breast cancer KEGG pathway (hsa05224).

Panel	Number of genes	Rank of target pathway	*p*-value (FDR)
MAGPEL (Proposed)	513	15	5.21E-16
Clinvar [63]	445	22	1.45E-01
Singhal *et al.* [97]	100	152	4.96E-15
EMU [29]	44	**6**	1.46E-09

Table 25: The top 10 significantly enriched pathways identified by the enrichment pathway analysis based on the proposed gene panel (MAGPEL) for breast cancer case study. Rows with green background indicate the target pathway for breast cancer. Rows with blue background indicate pathways for which we found indication of their association to breast cancer.

Rank	Pathway name	*p*-value (FDR)
1	Proteoglycans in cancer	2.54E-27
2	Pancreatic cancer	2.41E-24
3	ErbB signaling pathway [46]	1.14E-23
4	Colorectal cancer	1.41E-23
5	Endocrine resistance	2.51E-21
6	Chronic myeloid leukemia	3.95E-21
7	Endometrial cancer	1.67E-20
8	Hepatitis B	1.46E-18
9	Prostate cancer	1.57E-18
10	EGFR tyrosine kinase inhibitor resistance	1.94E-18

Table 26: The top 10 significantly enriched pathways identified by the enrichment pathway analysis based on the Clinvar [63] gene panel for breast cancer case study. Rows with green background indicate the target pathway for breast cancer. Rows with blue background indicate pathways for which we found indication of their association to breast cancer.

Rank	Pathway name	*p*-value (FDR)
1	Herpes simplex virus 1 infection	2.67E-21
2	Taste transduction	0.000787
3	Natural killer cell mediated cytotoxicity [4]	0.000787
4	Antigen processing and presentation	0.00196105
5	Fanconi anemia pathway	0.00196105
6	B cell receptor signaling pathway	0.00317
7	Human papillomavirus infection	0.00691236
8	Graft-versus-host disease	0.00691236
9	Pancreatic secretion	0.02699356
10	Aldosterone-regulated sodium reabsorption	0.02707948

Table 27: The top 10 significantly enriched pathways identified by the enrichment pathway analysis based on the EMU [29] gene panel for breast cancer case study. Rows with green background indicate the target pathway for breast cancer. Rows with blue background indicate pathways for which we found indication of their association to breast cancer.

Rank	Pathway name	*p*-value (FDR)
1	Prostate cancer	8.68E-11
2	Endometrial cancer	3.85E-10
3	Homologous recombination [21]	6.27E-10
4	Melanoma	1.33E-09
5	Platinum drug resistance	1.33E-09
6	Breast cancer	1.46E-09
7	Hepatocellular carcinoma	5.35E-09
8	Bladder cancer	1.38E-08
9	Gastric cancer	2.44E-08
10	Glioma	2.74E-08

Table 28: The top 10 significantly enriched pathways identified by the enrichment pathway analysis based on the Singhal *et al.* [97] gene panel for breast cancer case study. Rows with green background indicate the target pathway for breast cancer. Rows with blue background indicate pathways for which we found indication of their association to breast cancer.

Rank	Pathway name	*p*-value (FDR)
1	Proteoglycans in cancer	1.01E-20
2	Colorectal cancer	2.49E-20
3	Pancreatic cancer	4.76E-20
4	Prostate cancer	7.23E-20
5	Chronic myeloid leukemia	4.46E-19
6	Gastric cancer	3.30E-18
7	ErbB signaling pathway [46]	1.22E-17
8	Hepatocellular carcinoma	2.89E-17
9	Endometrial cancer	6.50E-17
10	Endocrine resistance	6.76E-17

CHAPTER 5 DISCUSSION

We investigate the novelty of our identified genes by checking their overlap with other available variant-driven gene panels for AML (Figure 7). Although 58% of the proposed genes are not included in the other panels, the classification and pathway analysis based on these genes achieve the best results. The gene differences between the proposed panel and Clinvar could arise from the fact that Clinvar is a manually curated database. In principle, manual curation is expected to yield very accurate but possibly incomplete annotations, which is consistent with the smaller number of genes included in the Clinvar panel. The consideration of only the title and abstract of the articles for extracting the variants by Singhal *et al.* [97], could be the reason for the gene differences between these two panels. The corresponding figures for prostate cancer and breast cancer are shown in Figures 8 and Figures 9, respectively.

We also investigate the percentage of the identified AML-related variants which are mentioned in the title and abstract sections of the articles and compared them with those that are mentioned in the full body of the articles but not in the title and the abstract. Figure 10 visualizes the variant overlaps and differences between these sections. As the figure shows, about 89% of the variants mentioned in an article do not appear in the title and the abstract section, which emphasizes the need to analyze the entire text of the articles. This represents a significant limitation of the existing methods that use only the title and abstract sections of an article. The Venn diagrams for prostate cancer and breast cancer are shown in Figures 11, 12, respectively.

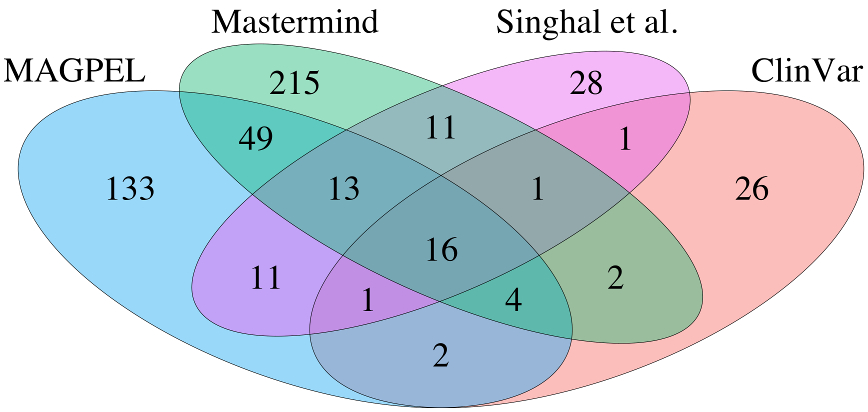


Figure 7: An overview of the gene overlaps and differences between the variant-driven gene panels. The proposed gene panel (MAGPEL) consists of 229 genes. The AML-related gene panel obtained from Clinvar and Mastermind includes 53 and 313 genes, respectively, and the one proposed by Singhal *et al.* [97] includes 76 genes.

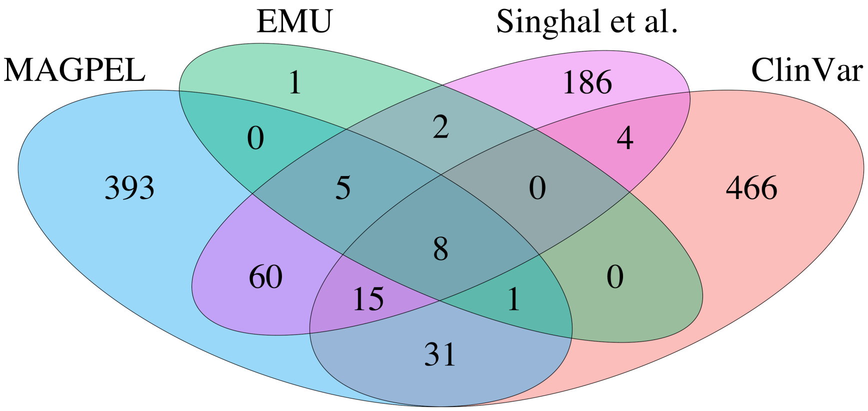


Figure 8: An overview of the gene overlaps and differences between the variant-driven gene panels for prostate cancer. The proposed gene panel (MAGPEL) consists of 532 genes. The prostate cancer-related gene panel obtained from Clinvar and EMU includes 525 and 17 genes, respectively, and the one proposed by Singhal *et al.* [97] includes 280 genes.

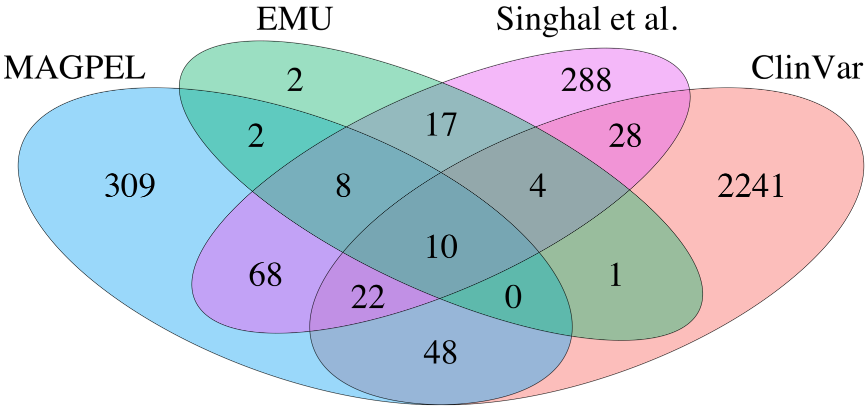


Figure 9: An overview of the gene overlaps and differences between the variant-driven gene panels for breast cancer. The proposed gene panel (MAGPEL) consists of 513 genes. The breast cancer-related gene panel obtained from Clinvar and EMU includes 2,354 and 44 genes, respectively, and the one proposed by Singhal *et al.* [97] includes 445 genes.

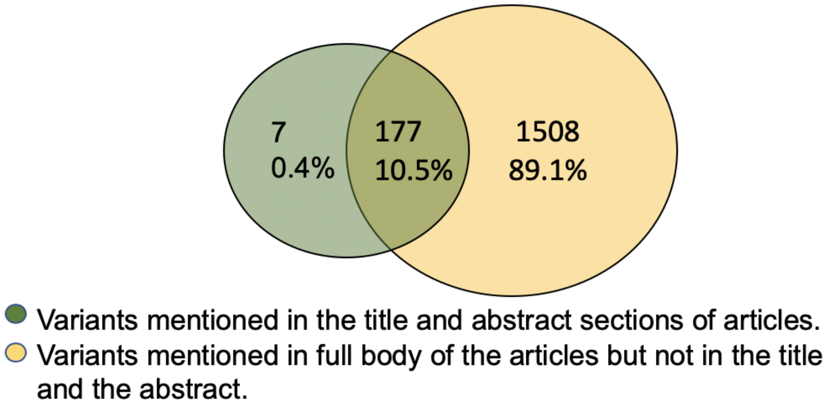


Figure 10: An overview of the overlap and differences between the variants mentioned in the title and abstract sections of the articles (green) and those that appear in the full body of the articles but not in the title and abstract section (gold) in AML case study.

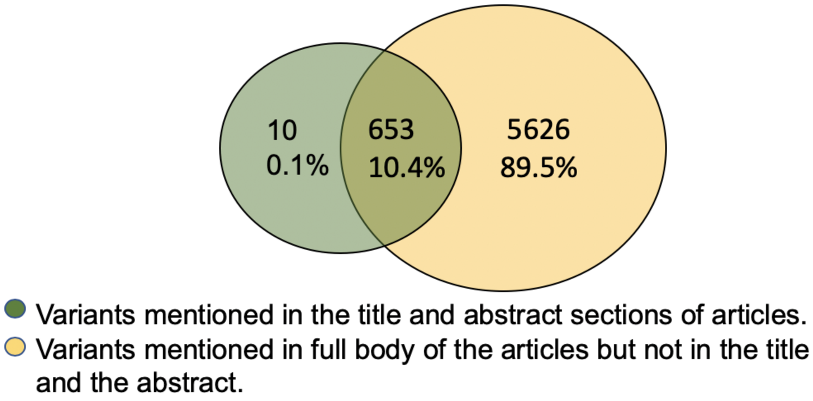


Figure 11: An overview of the overlap and differences between the variants mentioned in the title and abstract sections of the articles (green) and those that appear in the full body of the articles but not in the title and abstract section (gold) in prostate cancer case study.

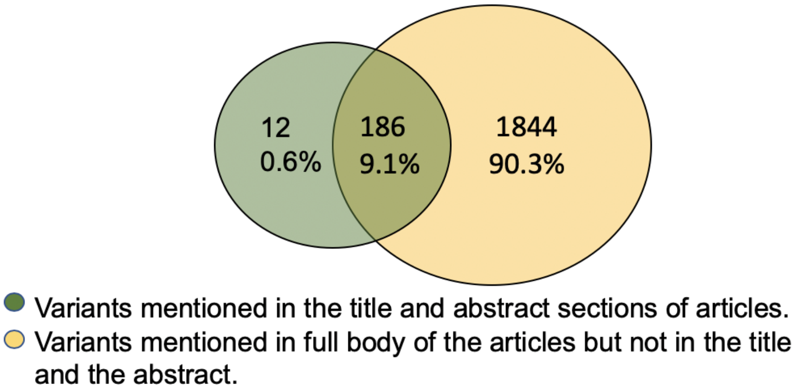


Figure 12: An overview of the overlap and differences between the variants mentioned in the title and abstract sections of the articles (green) and those that appear in the full body of the articles but not in the title and abstract section (gold) in breast cancer case study.

CHAPTER 6 CONCLUSION

6.1 Summary of contributions

The number of published articles describing the disease-related variants had a dramatic rise because of the recent advance sequencing technologies. This highlights the pressing need for the development of automated tools that are able to extract the variant-disease associations from literature. The manual extraction of this type of information from the biomedical literature takes an enormous amount of time and effort. Several automatic variant indexing tools have been developed to assist this manual curation. Correctly retrieving the disease-associated variants from biomedical texts remains a challenge mainly because of the complexity of the natural language processing and inconsistent use of standard recommendations for variant description.

Here, we present an automated framework to design an evidence-based variant-driven gene signature for a given disease phenotype. The identification of the variant-relevant articles using the word cloud analysis and the consideration of the full-length articles are the main contributions of the proposed framework. We illustrate the diagnostic value of the proposed gene signatures in capturing the mechanism involved in acute myeloid leukemia (AML), breast cancer, and prostate cancer using 29 independent gene expression datasets containing a total of 2,203 patients. We compare our signatures with several other available gene signatures as follows: i) Clinvar [63], ii) Mastermind [40], iii) Singhal *et al.* [97], iv) Doughty *et al.* [29] and v) the classical differentially expressed genes. The results show that the signatures obtained by the proposed framework yield better results than the other signatures currently available for these phenotypes.

We believe the proposed framework has significant advantages since it could be used to identify the gene biomarkers that describe the key biological phenomena for a given disease. The proposed framework is expected to be of interest to researchers from the computational biology and machine learning community.

6.2 Future work

In this thesis, we proposed an automated framework to extract the variant-gene-disease associations from biomedical literature. Studies have been shown the role of genomic mutations in improving the patient survival rate, personalizing medicine, and also reducing the risks of different therapies and drug responses [72]. However, the same as variants, this information is also buried in the scientific literature. Future work involves the identification and extraction of associations between genomic variants and drug responses from literature. For this purpose, first, we will use tmChem [65] to extract any chemical and drug names from biomedical text. Then, we will use a set of features to identify the associations between the mentioned mutations and drug responses. Similar to the method proposed by Mahmood *et al.* [72], our main interest is to capture the association between mutation existence and the drug responses and the treatment outcome. We will use a gold standard benchmark dataset named BRONCO [66] to validate the extracted variant-drug response associations. Lee *et al.* [66] generated BRONCO which contains associations between variant, gene, disease, drug, and cell line entities extracted from 108 full-text biomedical articles. We will use the BRONCO dataset and will report the standard evaluation metrics (precision, recall, and F1 score) for the proposed framework.

APPENDIX A

Table 29: The list of variant-driven genes obtained by the proposed framework for AML.

FLT3	ATM	STAT3	CD38	GSK3A	GSK3B	STAT5B	ERG	RPS6KB1
PSMD5	GFI1	CD8A	SAMHD1	U2AF1	BCR	IDH1	RPS10	NPM1
IDH2	FBXW7	SKI	SPI1	CSF3R	CEBPA	JAK2	MYB	HCFC1
DNMT3A	LUC7L2	ZNF672	GH1	ARHGAP35	MYBPC2	RUNX1	PGK2	TNS3
LYZL4	MPL	LYPLA1	EZH2	KIT	TP53	ASXL1	EDNRB	MYD88
CD274	CDCA7L	PRPF4B	TRIB1	CYP3A5	EIF4EBP1	EIF4B	WT1	EEF1A2
NUP98	CBFB	CD34	GLI1	PDGFRA	ABL1	ELANE	PTPA	GATA2
MYC	MAPK8	TNFRSF11A	PIP4K2A	PTPN11	CDK6	PML	PTMS	ZNF221
FOSL2	INO80B	MTHFR	PPP2R1A	ERBB2	CALR	TET2	GATA1	ZNF274
PPP2R2A	LYST	NCR2	FTO	ALKBH1	PIK3CA	FOXP3	FANCB	KIR2DL4
MTOR	CD33	PLXNB1	STAT1	KMT2B	SHH	MAP2K7	NUP62	RUNX2
IPO9	NT5C2	CBL	KMT2A	ZBTB7A	MDM2	DDX41	HDAC1	NRAS
EGFR	CYP2C19	ADAR	APOBEC3A	RUNX1T1	TERT	IL3	FOS	SMARCA5
RLF	GRM1	LRP11	CREB1	MAPT	TAPBPL	CDKN1B	ETV6	NOD2
SMC1A	NCOR1	MALT1	SF3B1	TYK2	SETD2	ASXL3	RUNX3	HHEX
CYR61	FYN	ABCB1	BCOR	SMYD2	TYMS	COX8A	FSTL4	KRAS
CYP1B1	MAF	ABCG2	HLA-G	ARID5B	SLC29A1	HLA-C	HPSE	MIR204
IL17A	UNG	HDAC9	CTNNB1	JAK3	EIF4E	NR3C1	CTD	HOXA9
ERCC2	MARCO	WRN	NAPRT	IL10	CRBN	GSTP1	NAT2	HFE
ASPG	NQO1	FASLG	HAMP	CXCR4	RAD51	PPP2R1B	RMI1	ANKRD26
SERPINA1	BRCA2	MECP2	CYP2E1	CDA	CYP4F2	CXCR6	XPC	POU1F1
MPO	SETBP1	XRCC1	CYP3A4	REST	TOP3A	CXXC5	ZNF763	CYTB
SLC24A3	FKBP5	MYBPC3	CYP1A1	TLR4	HSPD1	DCK	FANCA	SH2B3
ZHX2	BRCA1	PDCD1	SLU7	SRSF2	CREBBP	CRP	BCL2	ARIH2
IL17F	PDE9A	KLF1	DNMT3B	RAD52	PIM1	FPGS	DHX15	ALK
MAP4	ESR1	LSD1						

Table 30: The list of variant-driven genes obtained by the proposed framework for prostate cancer.

SNCA
MYC
AKT1
CYR61
ARF6
BCHE
CDK5
RNF19A
IRS1
KIT
ACTR1B
SMAD4
TGFBR2
FAM83G
PRKAA2
LIN28B
USP39
NSD2
ERG
WDR77
ABI1
F2RL1
CTCF
PIK3CA
TNFSF10
DNAH8
PPP6R2
ETS1
BRAF
RHOD
MID1
ROCK2
RNASEH2A
APOE
PAK6
RIT2
SDK1
GPX1
HSD17B4
ERCC1
SHBG
PKD1L3
OAS2
BCL11A
IL10
NOD2
MC4R
PPARG
ABCC4
CCL5
GP6
NSD1
CX3CR1
ZBTB10
IL21
PCSK9
ALDH2
SETD7
CBR1

STAT3
MAPKAPK2
TRPV6
SQSTM1
SYTL1
RNASEL
BDNF
ATM
KLK4
PKN1
RNF41
KMT2D
HUWE1
USP44
MAGEA11
GZMM
GSTK1
USF2
RASSF1
MFSD2A
KEAP1
RAB11A
EWSR1
PIK3R1
CASP8
BRIP1
TK1
USP2
KDM4C
NFKB1
PDE4D
SH3KBP1
HSD3B1
TRAM1
FASN
ABCB4
LILRA3
EPHX1
HTR3B
DUT
EPHB2
GCKR
CCHCR1
NPHP1
FTO
CASR
TNF
RAD51
SHMT1
FGF10
MMP2
BAG6
CDON
AHR
SOD2
BIK
PRDM9
MARCHF8
HMGA2

AR
PTEN
CHUK
PLK1
RAB8A
FOXA1
ABCB1
PRKDC
KLK14
HOTAIR
ELAC2
NOTCH4
MYO1C
CHD1
HIST2H2AA3
CD40LG
E2F1
GSK3B
SLC19A1
CTSA
TREX2
GHRL
EFNA5
EPHA5
ESR1
TSC2
PCSK1
CDK4
CYP1A1
CXCR4
RPS6KB2
CCND1
SNRNP70
KRT6A
CDKN2A
REPS2
ALS2CR12
IL10RB
HNF1B
LINC00673
ABL1
ITGA6
ERCC2
NFAT5
CYP24A1
ADIPOQ
XRCC1
MC1R
FOXP4
FOXO3
MRTFB
MSH5
P2RX7
HAPLN1
CTBP2
OPRD1
MMAB
JAK2
NCOR2

YAP1
MDM2
RPS6KB1
USP7
TBC1D10A
ADORA2A
C3
ATR
BUB1B
KLK2
AKT3
SUMO4
HIST2H3C
CDC42
KRT79
MAGEC3
HIF1A
C9orf3
MRNIP
ATP4A
IL17RA
FSD1
UXT
ABL2
CD82
FKBP4
MAP3K8
STK11
OTUB1
MAPK1
SERPINA1
SIAH2
CYP17A1
BTG2
KMT2C
AMPH
TMPRSS2
ITGA2
EPCAM
ORAI1
IDH1
IRX4
KLF12
THADA
CDH1
PRMT6
CCR2
LIG4
CYP2E1
IL10RA
VGLL3
POLB
FMN1
TLR10
TLR4
FUT2
SIX1
PEX2
HLA-DRB1

CDKN1B
RALA
PARP1
EEF1A2
LCN2
LPAR2
TRIM27
ERCC3
KRAS
CNOT1
PSMD4
SENP2
CASP3
SUMO2
PAK2
EP300
FOLH1
ID4
BAZ2A
DAB2IP
AGAP2
IGFBP2
MLPH
TRRAP
SATB2
IGF1R
TARDBP
INTS6
NCOR1
IL31RA
PRPF31
SH3RF1
GPRC6A
EHMT2
KMT2A
SGK1
APOB
CYP2R1
AXIN2
EFNB2
CDKAL1
TERC
LDAH
DKK3
MGMT
GOLPH3L
MLH1
CA4
IL2
G6PC2
EHBP1
DHODH
CYP2D6
GRIK1
BCL2
PEX14
ZNF652
AMZ2
NOL10

PIM1
SH3BP1
PSMD5
CDKN1A
EIF4E
BMPR2
LAMP1
FANCA
BRCA1
CYB5A
ULK1
TP63
DHX15
SENP6
HAT1
HEY1
MCM3
RAB5A
RAD50
SKP2
MED25
CUL3
S100A10
MSH2
GRK3
ATF3
APOBEC3G
NR2E1
HSPA4
DTYMK
DEFB109C
CYP1B1
BGLAP
VDR
NANOG
COX18
MSH6
OAS1
FBN1
IGFBP3
TCF7L2
SORT1
CHEK2
ABCG2
GSTP1
MAPT
NOS3
XPC
TLR5
IL1RN
ERAP1
CCND2
MUC1
COMT
LCT
FANCI
ADIPOR1
NAALADL2
CTDSPL

FOXO1
RAC1
PRL
IFI27
PALB2
TRA
ZFHX3
HDAC2
HDAC1
USP10
TMEM37
PRKD1
PDPK1
ALOX12
CXCL8
MLX
MTOR
WDR35
MSR1
IGF1
HCFC1
SPDEF
IFT81
ELK1
CREB1
CHRM3
CHD4
KRT14
NCOA2
CDK1
FRMD6
PLAT
ACPP
PARD6A
ARF1
TERT
TMEM38B
MSMB
LEPR
PHLPP2
ADH1B
PPFIBP2
AARS2
CCDC78
CYP19A1
POR
TGFB1
KDR
TNFRSF11B
TBX1
XAGE3
LPL
SLCO1B1
WFS1
SRD5A2
NR5A2
RMST
JMJD1C
MYO6

EGFR
PDE8B
STAT5A
UGDH
TP53
HOXB13
KLK3
MLH3
TET1
AMD1
BRCA2
MARK2
EZH2
KLF6
MAP2K1
NCOA3
PCA3
ETV6
SEMA6B
VEGFA
ST8SIA4
NFKBIA
FCF1
ETS2
GAPDH
HRAS
FAT1
INF2
TBP
SS18L1
MAPK8
ERBB3
PSMD9
ANXA2
TNK2
SPTY2D1
JAZF1
SCARB1
IL1B
NAT2
GPRC5B
OCA2
TACSTD2
LRSAM1
EPAS1
CDKN2B-AS1
NQO1
TLR9
COL1A1
PSCA
CRP
RFX6
TNFRSF1A
NEDD9
SLC22A3
LMTK2
IRS2
TTC9
FRK

ESR2
NR3C1
SRC
8-Mar
MSN
IL6
RHOA
NBN
IL4I1
RPTOR
CDK11B
RET
SPOP
CTNNB1
CDK5RAP2
CTC1
DBN1
ETV7
PIAS2
CGB5
XPR1
REV3L
FGFR4
PAWR
HSP90AA1
BRD2
ERBB2
DAPK2
PLXNB1
COL18A1
ITGA9
SLC1A2
FGF9
XPO1
MAPK14
CCL2
PKHD1
FGFR2
NDUFS2
MDM4
PDLIM5
HMGCR
ALDH9A1
NKX3-1
APC
SLC41A1
MTHFR
UGT2B15
PROM1
SPINK5
NRIP1
MUTYH
IRAK4
TLR1
TNS3
SLC9C2
APOC3
ASNA1
UNC5D

Table 31: The list of variant-driven genes obtained by the proposed framework for breast cancer.

RPS6KB1	CHEK1	PIK3CA	STAT3	TGM2	PARP1	EGFR	CTCF	TRIM28
ESR1	HRAS	CAV1	RAC1	TAZ	PRL	EIF2S1	TP53	FXYD1
PRDX1	USP1	AKT1	PTEN	SRC	PIK3CG	ANO1	MAPK8	PSMD5
IMMT	GPR132	EPCAM	RHOA	SKP2	ABL1	ATM	CTNNB1	MST1
BRCA1	MIR10B	MTHFR	TLN2	KEAP1	HIF1A	PGR	EIF2AK4	EHMT2
SYK	GTF2B	ELK1	MMP14	SOD2	ERBB2	AR	BRCA2	PARD3
TRIM25	HPR	CD44	CASP8	RELA	IL1B	ZHX2	RALGDS	NCK1
NFKBIA	RAB11A	RHOC	MAPK14	FGFR2	PTPN11	PIN1	STAT5A	GSK3B
PPM1G	PLD2	PIK3CD	STIM1	ADAM17	TOX3	NOS3	PRKAA1	GUCY2F
NOX5	DECR1	ADARB1	PRNP	SCN2A	BST2	SLC9A1	ARF1	PICK1
POU5F1	SELENBP1	SEPHS1	TYRP1	BMPR1B	ARHGEF1	BCR	BRK1	TRIM62
MIR200C	DHX58	CDH1	NUTF2	RAN	MAP2K6	PKD1	CACNA1B	CHEK2
PIK3CB	KCNQ2	CD36	MDM2	CYB5A	IKBKE	NME1	RAB5A	PDP1
GRIN1	AKT2	POLG	HFM1	FHIT	TGFB1	PAK4	TF	MASTL
VAV1	FBXO28	RAF1	GPNMB	EZH2	KMT2C	PA2G4	NUB1	RAD51
LMNA	ARF6	PTPN18	PSTPIP1	NR4A2	MAP2K5	MAP3K1	MAP2K4	PAK2
PAK1	CDC42	TRPV4	KCNJ11	GAS7	DICER1	MIR34A	SELENOW	PRKCE
CSF1	AKIP1	FLT4	RAB27A	ARRB2	RAB24	SLC5A5	MST1R	HSF1
PTPRJ	ADCY1	LDLRAP1	ADCY2	SMAD3	HSP90AA1	MTOR	CREB1	SH2B1
POT1	EEA1	PAFAH1B1	BCL2	BAX	CCND1	MME	BARD1	ZNF135
CSF2RA	FBXW7	JAK2	MYC	HSP90AB1	RRM1	FOXP3	LCK	FGFR4
FOXA1	GJA1	PTK2	RIN1	PKD2	MSN	EZR	UMOD	GRHL1
GRHL3	TP73	CFL1	GRIA1	SH3BP4	ACAP1	ACAP2	ASAP1	KRAS
PRKAB1	PXN	SIRT6	FOXO3	LOX	RAB7B	PCNA	AZIN1	CA1
CASP2	BECN1	OCRL	MIEN1	SMN2	MED12	CDKN2A	PLD1	PPP2R1A
FOXL2	CDK4	BDKRB1	YAP1	SFTPC	TRIO	TIAM1	FGF14	L3MBTL3
BAIAP2	TRPV1	ATF1	LMAN1	UQCRHL	ADAM10	ADAMTS15	UIMC1	AMOTL1
MR1	PPARG	ARSB	RBBP8	RNF213	ABCE1	DNAH8	FANCI	NEDD9
BCAR1	LAMB1	STARD8	RDX	APP	GAPDH	IL4R	SPHK1	CFTR
TLK2	GHR	TUBB3	TSC22D3	LGALS4	APC2	AXIN1	COPS6	DNMT3A
WEE1	AGO1	TDP1	NEDD8	NAE1	PRKDC	ZNF217	ZNF516	CTBP2
HDAC1	AMELX	ITCH	NEDD4L	TOP1	MTA1	MTA2	POLR3K	CKLF
GJB1	RYR1	NSF	PLEK	FGF2	AKAP10	RIOK2	FANCD2	CA2
IRF8	SCRIB	DLC1	APAF1	RND1	S100A4	GPI	ABCG2	HIVEP3
MAP1LC3A	GNAS	PLP1	CD24	STARD13	TRPM7	RIC8A	PIK3C3	RAB33B
RAB1A	RAB6A	RAB2A	AURKA	INS	HEBP1	CRH	CRHR1	PTPRF
RPL27A	PADI4	MUC5AC	NQO1	RAB34	FMR1	DHX16	RAB7A	HEY1
NOTCH1	CAMK2G	DDX23	RAB22A	ZAR1	CSNK1G1	TALDO1	ERN1	RAB4A
YIPF5	RABAC1	SRGAP2	PDCD10	TERT	GRB2	DRG1	SYVN1	GALNT6
CALR	MAPK8IP1	KCNJ2	NEDD4	FANCA	TOE1	ESS2	P2RY2	VIM
TIPRL	GLS	PMP22	MAPK8IP3	CMTR1	DHX15	GAB1	KRT15	RNASE1
DCN	CSF1R	MUT	MAPK1	RAC3	NCK2	ATL3	RHOQ	LINC00310
WWOX	CSNK2A1	CCNB1	SIRT1	PRRX1	KLF4	PLK4	COIL	TET1
MED14	PRPF4B	EPO	SMAD4	RNPEP	ACE	ATAT1	CDK1	PRKCA
ADA2	CDK9	EIF2AK3	PTK2B	TPT1	RAB35	PON1	ARL4A	CHD1
VRK3	NRP1	PRRT2	MIB1	KAT5	ESRP2	SPRY2	CTTN	SMURF2
MAPKAPK2	USF1	CPT1A	CDKN1A	TGFBR1	PARD6A	BMP15	ERBB3	STAT2
SENP1	NOTCH3	PSEN1	FAAH	MTA3	NUMB	NANOG	RAB23	FIP1L1
RAB8A	RAB10	ARF4	HSPB1	SEC24D	SVIL	RUNX2	ARL8B	BABAM2
ABCB1	POU1F1	GATA3	USF3	CYP2D6	SLX4	SDHB	SDHD	LEPR
ERCC1	NAT2	ACTN4	NRAS	STK11	FGFR1	KCNQ1	ERCC2	ERBB4
APOB	STAT1	MUTYH	CLDN1	NBN	GALNT12	ABCB6	MCM8	ABCC11
TCF4	XRCC1	XPA	CASR	PALB2	ATR	PDGFRA	VPS35	CYP2C8
MLH1	KCNE1	NOD2	CYP2E1	MEN1	SEMA3F	SLCO1B1	CTLA4	MMP2
RAD51B	GFAP	KLHDC7A	XRCC2	PHLDA3	ABCC1	U2AF1	SCN1A	PITX2
UGT1A1	APOE	COMT	CYP2C19	AGTR1	SF3B1	FAM20A	BRIP1	CELSR2
MSH2	LDLR	WDR43	CUX1	ETV6	VDR	RET	COL1A1	NA

Table 32: The complete list of variant-disease pairs identified by the proposed method and the baseline method from the gold standard databases [29].

PMID	Mutation	Gold_standard	Proposed_method	Co-occurrence method
12023985	p|SUB|R|188|H	breast neoplasms	breast cancer	breast cancer
12100746	p|SUB|V|89|L	breast neoplasms	breast cancer	breast cancer
12516098	p|SUB|P|1315|L	breast neoplasms	breast cancer	breast cancer
12602915	p|SUB|R|726|L	breast neoplasms	prostate cancer	prostate cancer
12628588	p|SUB|Q|253|H	breast neoplasms	breast cancer	breast cancer
12702523	p|SUB|R|72|P	breast neoplasms	breast cancer	breast cancer
12786840	p|SUB|R|156|G	breast neoplasms	breast carcinoma	breast carcinoma
12810666	p|SUB|L|1420|F	breast neoplasms, ovarian neoplasms	breast cancer	breast cancer
12872252	p|SUB|Q|540|L	breast neoplasms, ovarian neoplasms	breast cancer	breast cancer
12917204	p|SUB|L|546|V	breast neoplasms	breast cancer	breast cancer
14683420	p|SUB|R|72|P	breast neoplasms	breast cancer	breast cancer
15059511	p|SUB|P|359|L	breast neoplasms	breast cancer	breast cancer
15170666	p|SUB|E|233|G	breast neoplasms	breast cancer	breast cancer
16061562	p|SUB|C|645|R	breast neoplasms, ovarian neoplasms	ovarian tumor	ovarian tumor
16168123	p|SUB|S|384|F	breast neoplasms	breast cancer	breast cancer
16333312	p|SUB|V|507|M	breast neoplasms, ovarian neoplasms	breast cancer	breast cancer
16652348	c|SUB|C|146|G	breast neoplasms	breast cancer	breast cancer
16760288	p|SUB|F|486|L	breast neoplasms, ovarian neoplasms	breast cancer	breast cancer
16760288	p|SUB|N|550|H	breast neoplasms, ovarian neoplasms	breast cancer	breast cancer
16760288	p|SUB|Y|179|C	breast neoplasms, ovarian neoplasms	breast cancer	breast cancer
16822847	p|SUB|G|388|R	breast neoplasms	breast cancer	breast cancer
16825437	p|SUB|S|558|P	breast neoplasms	breast cancer	breast cancer
17001622	p|SUB|V|2424|G	breast neoplasms	breast cancer	breast cancer
17217814	p|SUB|V|158|M	breast neoplasms	breast cancer	breast cancer
17427234	p|SUB|R|248|W	neuroblastoma, li-fraumeni syndrome	neuroblastoma	neuroblastoma
17541742	p|SUB|R|213|Q	breast neoplasms, ovarian neoplasms	li-fraumeni syndrome	li-fraumeni syndrome
17541742	p|SUB|R|290|H	breast neoplasms, ovarian neoplasms	li-fraumeni-like	li-fraumeni-like
17553133	p|SUB|K|303|R	breast neoplasms	breast cancer	breast cancer
17574969	p|SUB|R|1699|W	phyllodes tumor	tumor of the breast	tumor of the breast
17848578	p|SUB|Q|564|H	breast neoplasms, ovarian neoplasms, endometrial tumors	endometrial tumors	endometrial tumors
17848578	p|SUB|V|695|L	breast neoplasms, ovarian neoplasms, endometrial tumors	endometrial tumors	endometrial tumors
18241035	p|SUB|D|301|H	breast neoplasms	breast cancers	breast cancers
18241035	p|SUB|G|479|E	breast neoplasms	breast cancers	breast cancers
18241035	p|SUB|L|792|F	breast neoplasms	breast cancers	breast cancers

18565893	p|SUB|S|707|P	breast neoplasms, thyroid neoplasms, ENDOCRINE GLAND NEOPLASMS	breast cancer	breast cancer
10477429	p|SUB|M|133|T	breast neoplasms, li-fraumeni syndrome	breast sarcoma	NA
10485478	p|SUB|G|1449|V	breast neoplasms, hepatocellular carcinoma	breast cancers	NA
10485478	p|SUB|G|1464|E	breast neoplasms, hepatocellular carcinoma	breast cancers	NA
10485478	p|SUB|I|1572|T	breast neoplasms, hepatocellular carcinoma	breast cancers	NA
10485478	p|SUB|Q|1445|H	breast neoplasms, hepatocellular carcinoma	breast cancers	NA
10534763	p|SUB|G|2765|S	breast neoplasms	breast cancer	NA
10547570	p|SUB|R|273|C	phyllodes tumor	phyllodes tumours	NA
11212236	p|SUB|L|452|M	breast neoplasms	breast tumor	NA
11212236	p|SUB|N|435|S	breast neoplasms	breast tumor	NA
11212236	p|SUB|V|387|M	breast neoplasms	breast tumor	NA
11212236	p|SUB|V|447|A	breast neoplasms	breast tumor	NA
12100746	p|SUB|A|49|T	breast neoplasms	breast cancer	NA
12645254	p|SUB|M|1652|I	breast neoplasms, ovarian neoplasms	ovarian cancer	NA
12645254	p|SUB|S|1613|G	breast neoplasms, ovarian neoplasms	ovarian cancer	NA
12645254	p|SUB|W|1837|R	breast neoplasms, ovarian neoplasms	ovarian cancer	NA
12649339	p|SUB|S|215|I	breast neoplasms	breast cancer	NA
12668615	p|SUB|D|213|N	breast neoplasms	breast cancer	NA
12872252	p|SUB|V|524|I	breast neoplasms, ovarian neoplasms	breast cancer	NA
15059511	p|SUB|N|289|H	breast neoplasms	breast cancer	NA
15059511	p|SUB|N|371|H	breast neoplasms	breast cancer	NA
15059511	p|SUB|N|991|D	breast neoplasms	breast cancer	NA
15101044	c|SUB|T|2572|C	breast neoplasms	breast cancer	NA
15101044	p|SUB|P|1054|R	breast neoplasms	breast cancer	NA
15235021	p|SUB|R|732|Q	stomach neoplasms, breast neoplasms	breast cancers	NA
15235021	p|SUB|W|409|R	stomach neoplasms, breast neoplasms	breast cancers	NA
15649950	p|SUB|P|85|L	breast neoplasms	breast cancer	NA
15665273	p|SUB|S|148|A	breast neoplasms	breast cancer	NA
15665273	p|SUB|S|251|A	breast neoplasms	breast cancer	NA
15665273	p|SUB|S|288|A	breast neoplasms	breast cancer	NA
15870154	p|SUB|T|461|D	breast neoplasms	breast cancer	NA
16061562	p|SUB|C|557|S	ovarian neoplasms	ovarian tumour	NA
16061562	p|SUB|I|738|V	breast neoplasms	ovarian tumour	NA
16061562	p|SUB|S|761|N	breast neoplasms, uterine neoplasms	ovarian tumours	NA

16123141	p|SUB|T|135|E	breast neoplasms	breast cancer	NA
16280053	p|SUB|P|47|A	breast neoplasms	breast cancer	NA
16503999	p|SUB|C|282|Y	breast neoplasms	breast cancer	NA
16563154	c|SUB|T|309|G	breast neoplasms	breast cancer	NA
16652348	p|SUB|F|858|L	breast neoplasms	breast cancer	NA
16969499	p|SUB|S|1841|N	breast neoplasms, ovarian neoplasms	breast tumorigenesis	NA
17130833	p|SUB|V|143|A	breast neoplasms	breast cancer	NA
17317153	p|SUB|P|871|L	breast neoplasms	breast cancers	NA
17493881	p|SUB|V|1833|M	breast neoplasms, ovarian neoplasms	hereditary breast/ovarian cancer	NA
17531442	p|SUB|S|1143|A	breast neoplasms	breast cancer	NA
17531442	p|SUB|S|1280|A	breast neoplasms	breast cancer	NA
17550384	p|SUB|P|350|R	breast carcinoma, colorectal neoplasms, non-small-cell lung carcinoma, gastric carcinoma	breast carcinoma	NA
17550384	p|SUB|R|389|C	breast carcinoma, colorectal neoplasms, non-small-cell lung carcinoma, gastric carcinoma	breast carcinoma	NA
17889706	p|SUB|S|255|R	breast neoplasms	breast tumors	NA
18036263	p|SUB|A|1708|V	breast neoplasms	breast cancer	NA
18036263	p|SUB|G|1738|R	breast neoplasms	breast cancer	NA
18036263	p|SUB|R|1699|Q	breast neoplasms	breast cancer	NA
18083510	p|SUB|A|238|V	breast neoplasms	ovarian cancer	NA
18083510	p|SUB|R|259|H	breast neoplasms	ovarian cancer	NA
18083510	p|SUB|S|313|G	breast neoplasms	ovarian cancer	NA
18186519	p|SUB|G|12|S	breast neoplasms, colon neoplasms	colon and breast cancer	NA
18186519	p|SUB|G|12|V	breast neoplasms, colon neoplasms	colon and breast cancer	NA
18186519	p|SUB|V|600|E	breast neoplasms, colon neoplasms	colon and breast cancer	NA
18307025	p|SUB|Y|220|C	osteosarcoma, breast neoplasms, colon neoplasms, malignant fibrous histiocytoma, lung neoplasms	malignant tumors	NA
18332865	p|SUB|C|124|S	breast neoplasms	breast cancer	NA
18332865	p|SUB|G|129|E	breast neoplasms	breast cancer	NA
18372405	p|SUB|A|111|D	breast neoplasms	breast cancer	NA
18372405	p|SUB|G|160|R	breast neoplasms	breast cancer	NA
18375489	p|SUB|E|542|K	breast neoplasms, colorectal neoplasms, lung neoplasms, melanoma	colorectal cancer	NA
18431743	p|SUB|F|31|I	ovarian neoplasms	breast cancer	NA
18431743	p|SUB|N|372|H	ovarian neoplasms	breast cancer	NA

<table>
<tr><th></th><th></th><th></th><th></th><th></th></tr>
<tr><td>18558293</td><td>p|SUB|A|39|P</td><td>multiple hamartoma syndrome, breast neoplasms, thyroid neoplasms, lymphoma</td><td>gastric malignant lymphoma</td><td>NA</td></tr>
<tr><td>9407971</td><td>p|SUB|R|175|H</td><td>breast neoplasms</td><td>breast cancer</td><td>NA</td></tr>
<tr><td>9407971</td><td>p|SUB|R|249|S</td><td>breast neoplasms</td><td>breast cancer</td><td>NA</td></tr>
<tr><td>9407971</td><td>p|SUB|R|273|H</td><td>breast neoplasms</td><td>breast cancer</td><td>NA</td></tr>
<tr><td>9806478</td><td>p|SUB|A|148|T</td><td>melanoma</td><td>melanoma</td><td>NA</td></tr>
</table>

Table 33: The complete list of variants automatically extracted from 10,000 random articles and also manually reviewed and validated. False positive means the extracted entity is not a variant and it is wrongly identified as a variant.

<table>
<thead>
<tr><th>PMCID</th><th>NORMALIZED_FORM</th><th>TYPE</th><th>MENTIONED</th><th>FALSE POSITIVE?</th></tr>
</thead>
<tbody>
<tr><td>PMC4502233</td><td>p|SUB|R|4810|K</td><td>ProteinMutation</td><td>p.R4810K</td><td>NO</td></tr>
<tr><td>PMC4565919</td><td>p|SUB|L|90|M</td><td>ProteinMutation</td><td>L90M</td><td>NO</td></tr>
<tr><td>PMC4876505</td><td>c|SUB|G|93|A</td><td>DNAMutation</td><td>G93A</td><td>NO</td></tr>
<tr><td>PMC2684265</td><td>p|SUB|P|504|S</td><td>ProteinMutation</td><td>P504S</td><td>NO</td></tr>
<tr><td>PMC3666908</td><td>p|SUB|V|158|M</td><td>ProteinMutation</td><td>Val158Met</td><td>NO</td></tr>
<tr><td>PMC4718276</td><td>p|SUB|V|600|E</td><td>ProteinMutation</td><td>V600E</td><td>NO</td></tr>
<tr><td>PMC4962770</td><td>p|SUB|V|66|M</td><td>ProteinMutation</td><td>Val66Met</td><td>NO</td></tr>
<tr><td>PMC4991467</td><td>rs6295</td><td>SNP</td><td>rs6295</td><td>NO</td></tr>
<tr><td>PMC5012569</td><td>p|SUB|P|301|S</td><td>ProteinMutation</td><td>P301S</td><td>NO</td></tr>
<tr><td>PMC1247523</td><td>c|SUB|G|1800|A</td><td>DNAMutation</td><td>G1800A</td><td>YES</td></tr>
<tr><td>PMC2188802</td><td>c|SUB|C|200|T</td><td>DNAMutation</td><td>C/T200</td><td>YES</td></tr>
<tr><td>PMC2584175</td><td>c|SUB|A|1|C</td><td>DNAMutation</td><td>A1C</td><td>YES</td></tr>
<tr><td>PMC2720913</td><td>p|SUB|E||T</td><td>ProteinMutation</td><td>E/T</td><td>YES</td></tr>
<tr><td>PMC2829413</td><td>c|SUB|A|1|C</td><td>DNAMutation</td><td>A1C</td><td>YES</td></tr>
<tr><td>PMC2841238</td><td>c|SUB|G|1311|A</td><td>DNAMutation</td><td>G1311A</td><td>YES</td></tr>
<tr><td>PMC2841238</td><td>c|SUB|G|1329|A</td><td>DNAMutation</td><td>G1329A</td><td>YES</td></tr>
<tr><td>PMC2841238</td><td>c|SUB|G|1379|A</td><td>DNAMutation</td><td>G1379A</td><td>YES</td></tr>
<tr><td>PMC2841238</td><td>c|SUB|G|1316|A</td><td>DNAMutation</td><td>G1316A</td><td>YES</td></tr>
<tr><td>PMC2875450</td><td>c|SUB|A|1|C</td><td>DNAMutation</td><td>A1C</td><td>YES</td></tr>
<tr><td>PMC2889782</td><td>c|SUB|A|1|C</td><td>DNAMutation</td><td>A1C</td><td>YES</td></tr>
<tr><td>PMC2915039</td><td>p|SUB|H|14|A</td><td>ProteinMutation</td><td>H14A</td><td>YES</td></tr>
<tr><td>PMC2915039</td><td>p|SUB|H|14|C</td><td>ProteinMutation</td><td>H14C</td><td>YES</td></tr>
<tr><td>PMC2959805</td><td>p|SUB|H|11|A</td><td>ProteinMutation</td><td>H11A</td><td>YES</td></tr>
<tr><td>PMC2968464</td><td>p|SUB|H|11|A</td><td>ProteinMutation</td><td>H11A</td><td>YES</td></tr>
<tr><td>PMC2968464</td><td>p|SUB|H|11|C</td><td>ProteinMutation</td><td>H11C</td><td>YES</td></tr>
<tr><td>PMC2968464</td><td>p|SUB|H|12|A</td><td>ProteinMutation</td><td>H12A</td><td>YES</td></tr>
<tr><td>PMC2968464</td><td>p|SUB|H|12|C</td><td>ProteinMutation</td><td>H12C</td><td>YES</td></tr>
<tr><td>PMC2968899</td><td>p|SUB|H|10|A</td><td>ProteinMutation</td><td>H10A</td><td>YES</td></tr>
<tr><td>PMC2969907</td><td>p|SUB|H|23|A</td><td>ProteinMutation</td><td>H23A</td><td>YES</td></tr>
<tr><td>PMC2969907</td><td>p|SUB|H|14|A</td><td>ProteinMutation</td><td>H14A</td><td>YES</td></tr>
<tr><td>PMC2969907</td><td>p|SUB|H|20|A</td><td>ProteinMutation</td><td>H20A</td><td>YES</td></tr>
<tr><td>PMC2969907</td><td>p|SUB|H|26|A</td><td>ProteinMutation</td><td>H26A</td><td>YES</td></tr>
<tr><td>PMC2969907</td><td>p|SUB|H|18|A</td><td>ProteinMutation</td><td>H18A</td><td>YES</td></tr>
<tr><td>PMC2969907</td><td>p|SUB|H|29|A</td><td>ProteinMutation</td><td>H29A</td><td>YES</td></tr>
<tr><td>PMC2969907</td><td>p|SUB|H|28|A</td><td>ProteinMutation</td><td>H28A</td><td>YES</td></tr>
</tbody>
</table>

PMC2969907	p|SUB|H|19|A	ProteinMutation	H19A	YES
PMC2969907	p|SUB|H|22|A	ProteinMutation	H22A	YES
PMC2969907	p|SUB|H|22|C	ProteinMutation	H22C	YES
PMC2969907	p|SUB|H|27|A	ProteinMutation	H27A	YES
PMC2969907	p|SUB|H|13|A	ProteinMutation	H13A	YES
PMC2969907	p|SUB|H|11|A	ProteinMutation	H11A	YES
PMC2969907	p|SUB|H|12|A	ProteinMutation	H12A	YES
PMC2969907	p|SUB|H|31|A	ProteinMutation	H31A	YES
PMC2969907	p|SUB|H|31|C	ProteinMutation	H31C	YES
PMC2979928	p|SUB|H|11|A	ProteinMutation	H11A	YES
PMC2979928	p|SUB|H|14|A	ProteinMutation	H14A	YES
PMC2979928	p|SUB|H|14|C	ProteinMutation	H14C	YES
PMC2979928	p|SUB|H|15|A	ProteinMutation	H15A	YES
PMC2979928	p|SUB|H|15|C	ProteinMutation	H15C	YES
PMC2983608	p|SUB|H|20|A	ProteinMutation	H20A	YES
PMC2983608	p|SUB|H|20|C	ProteinMutation	H20C	YES
PMC2992198	c|SUB|A|1|C	DNAMutation	A1C	YES
PMC3005448	c|SUB|A|1|C	DNAMutation	A1C	YES
PMC3005479	c|SUB|A|1|C	DNAMutation	A1C	YES
PMC3007219	p|SUB|H|13|A	ProteinMutation	H13A	YES
PMC3007219	p|SUB|H|13|C	ProteinMutation	H13C	YES
PMC3007494	p|SUB|H|23|A	ProteinMutation	H23A	YES
PMC3007494	p|SUB|H|24|A	ProteinMutation	H24A	YES
PMC3007494	p|SUB|H|24|C	ProteinMutation	H24C	YES
PMC3007494	p|SUB|H|23|C	ProteinMutation	H23C	YES
PMC3007494	p|SUB|H|23|D	ProteinMutation	H23D	YES
PMC3007494	p|SUB|H|24|D	ProteinMutation	H24D	YES
PMC3007494	p|SUB|H|24|E	ProteinMutation	H24E	YES
PMC3007494	p|SUB|H|24|F	ProteinMutation	H24F	YES
PMC3008083	p|SUB|H|16|A	ProteinMutation	H16A	YES
PMC3008083	p|SUB|H|16|C	ProteinMutation	H16C	YES
PMC3008083	p|SUB|H|17|A	ProteinMutation	H17A	YES
PMC3008083	p|SUB|H|17|C	ProteinMutation	H17C	YES
PMC3008083	p|SUB|H|32|A	ProteinMutation	H32A	YES
PMC3008083	p|SUB|H|32|C	ProteinMutation	H32C	YES
PMC3009229	p|SUB|H|10|A	ProteinMutation	H10A	YES
PMC3009229	p|SUB|H|11|A	ProteinMutation	H11A	YES
PMC3009229	p|SUB|H|11|C	ProteinMutation	H11C	YES
PMC3009229	p|SUB|H|12|A	ProteinMutation	H12A	YES
PMC3009229	p|SUB|H|12|C	ProteinMutation	H12C	YES

PMC3009229	p|SUB|H|13|A	ProteinMutation	H13A	YES
PMC3009229	p|SUB|H|14|A	ProteinMutation	H14A	YES
PMC3009229	p|SUB|H|14|C	ProteinMutation	H14C	YES
PMC3009229	p|SUB|H|15|A	ProteinMutation	H15A	YES
PMC3009229	p|SUB|H|15|C	ProteinMutation	H15C	YES
PMC3011503	p|SUB|H|10|A	ProteinMutation	H10A	YES
PMC3011503	p|SUB|H|10|C	ProteinMutation	H10C	YES
PMC3011503	p|SUB|H|11|A	ProteinMutation	H11A	YES
PMC3011503	p|SUB|H|11|C	ProteinMutation	H11C	YES
PMC3011503	p|SUB|H|13|A	ProteinMutation	H13A	YES
PMC3011503	p|SUB|H|13|C	ProteinMutation	H13C	YES
PMC3011503	p|SUB|H|14|A	ProteinMutation	H14A	YES
PMC3011503	p|SUB|H|14|C	ProteinMutation	H14C	YES
PMC3011503	p|SUB|H|15|A	ProteinMutation	H15A	YES
PMC3011503	p|SUB|H|15|C	ProteinMutation	H15C	YES
PMC3011503	p|SUB|H|30|A	ProteinMutation	H30A	YES
PMC3011503	p|SUB|H|30|C	ProteinMutation	H30C	YES
PMC3011503	p|SUB|H|31|A	ProteinMutation	H31A	YES
PMC3011503	p|SUB|H|31|C	ProteinMutation	H31C	YES
PMC3011503	p|SUB|H|32|A	ProteinMutation	H32A	YES
PMC3011503	p|SUB|H|32|C	ProteinMutation	H32C	YES
PMC3011503	p|SUB|H|34|A	ProteinMutation	H34A	YES
PMC3011503	p|SUB|H|34|C	ProteinMutation	H34C	YES
PMC3011503	p|SUB|H|35|A	ProteinMutation	H35A	YES
PMC3011503	p|SUB|H|35|C	ProteinMutation	H35C	YES
PMC3011503	p|SUB|H|36|A	ProteinMutation	H36A	YES
PMC3011503	p|SUB|H|36|C	ProteinMutation	H36C	YES
PMC3012188	c|SUB|A|1|C	DNAMutation	A1C	YES
PMC3030744	c|SUB|G|1575|A	DNAMutation	G1575A	YES
PMC3051471	p|SUB|H|31|A	ProteinMutation	H31A	YES
PMC3051471	p|SUB|H|32|A	ProteinMutation	H32A	YES
PMC3051471	p|SUB|H|33|A	ProteinMutation	H33A	YES
PMC3051471	p|SUB|H|34|A	ProteinMutation	H34A	YES
PMC3051517	p|SUB|H|14|A	ProteinMutation	H14A	YES
PMC3051517	p|SUB|H|13|A	ProteinMutation	H13A	YES
PMC3051517	p|SUB|H|12|A	ProteinMutation	H12A	YES
PMC3088031	p|SUB|T|9|S	ProteinMutation	T9S	YES
PMC3122477	c|DEL||	DNAMutation	DELTA	YES
PMC3139590	RS800	SNP	RS800	YES
PMC3164897	5bins	DNAMutation	5 bins	YES

<table>
<tr><td>PMC3164897</td><td>3bins</td><td>DNAMutation</td><td>3 bins</td><td>YES</td></tr>
<tr><td>PMC3196552</td><td>c|SUB|S|-2600|H</td><td>DNAMutation</td><td>S-2600H</td><td>YES</td></tr>
<tr><td>PMC3202145</td><td>c|SUB|C|31|G</td><td>DNAMutation</td><td>C31G</td><td>YES</td></tr>
<tr><td>PMC3225014</td><td>c|SUB|GGCA|6|C</td><td>DNAMutation</td><td>GGCA6C</td><td>YES</td></tr>
<tr><td>PMC3225014</td><td>c|SUB|CCGT|6|G</td><td>DNAMutation</td><td>CCGT6G</td><td>YES</td></tr>
<tr><td>PMC3275195</td><td>p|SUB|H|10|A</td><td>ProteinMutation</td><td>H10A</td><td>YES</td></tr>
<tr><td>PMC3275195</td><td>p|SUB|H|11|A</td><td>ProteinMutation</td><td>H11A</td><td>YES</td></tr>
<tr><td>PMC3275195</td><td>p|SUB|H|12|A</td><td>ProteinMutation</td><td>H12A</td><td>YES</td></tr>
<tr><td>PMC3275195</td><td>p|SUB|H|14|A</td><td>ProteinMutation</td><td>H14A</td><td>YES</td></tr>
<tr><td>PMC3275195</td><td>p|SUB|H|16|A</td><td>ProteinMutation</td><td>H16A</td><td>YES</td></tr>
<tr><td>PMC3275195</td><td>p|SUB|H|17|A</td><td>ProteinMutation</td><td>H17A</td><td>YES</td></tr>
<tr><td>PMC3275195</td><td>p|SUB|H|18|A</td><td>ProteinMutation</td><td>H18A</td><td>YES</td></tr>
<tr><td>PMC3297250</td><td>p|SUB|H|13|A</td><td>ProteinMutation</td><td>H13A</td><td>YES</td></tr>
<tr><td>PMC3297250</td><td>p|SUB|H|14|A</td><td>ProteinMutation</td><td>H14A</td><td>YES</td></tr>
<tr><td>PMC3353114</td><td>p|SUB|D|125|I</td><td>ProteinMutation</td><td>D 125I</td><td>YES</td></tr>
<tr><td>PMC3537756</td><td>c|SUB|G|3|A</td><td>DNAMutation</td><td>G3A</td><td>YES</td></tr>
<tr><td>PMC3551769</td><td>p|SUB|V|36|G</td><td>ProteinMutation</td><td>V36G</td><td>YES</td></tr>
<tr><td>PMC3588815</td><td>p|SUB|H|11|A</td><td>ProteinMutation</td><td>H11A</td><td>YES</td></tr>
<tr><td>PMC3588815</td><td>p|SUB|H|12|A</td><td>ProteinMutation</td><td>H12A</td><td>YES</td></tr>
<tr><td>PMC3588815</td><td>p|SUB|H|19|A</td><td>ProteinMutation</td><td>H19A</td><td>YES</td></tr>
<tr><td>PMC3588815</td><td>p|SUB|H|19|C</td><td>ProteinMutation</td><td>H19C</td><td>YES</td></tr>
<tr><td>PMC3588815</td><td>p|SUB|H|20|A</td><td>ProteinMutation</td><td>H20A</td><td>YES</td></tr>
<tr><td>PMC3588815</td><td>p|SUB|H|20|C</td><td>ProteinMutation</td><td>H20C</td><td>YES</td></tr>
<tr><td>PMC3588815</td><td>p|SUB|H|21|A</td><td>ProteinMutation</td><td>H21A</td><td>YES</td></tr>
<tr><td>PMC3588815</td><td>p|SUB|H|21|C</td><td>ProteinMutation</td><td>H21C</td><td>YES</td></tr>
<tr><td>PMC3598566</td><td>p|SUB|V|150|T</td><td>ProteinMutation</td><td>150 V/T</td><td>YES</td></tr>
<tr><td>PMC3721223</td><td>g|SUB|A|80915|G</td><td>DNAMutation</td><td>A80915G</td><td>YES</td></tr>
<tr><td>PMC3841629</td><td>c|SUB|A|1|C</td><td>DNAMutation</td><td>A1C</td><td>YES</td></tr>
<tr><td>PMC3868173</td><td>MOT>LUM</td><td>ProteinMutation</td><td>MOT >LUM</td><td>YES</td></tr>
<tr><td>PMC3881066</td><td>p|SUB|A|30|P</td><td>ProteinMutation</td><td>A30P</td><td>YES</td></tr>
<tr><td>PMC3927678</td><td>p|SUB|H|379|UF</td><td>ProteinMutation</td><td>H 379 UF</td><td>YES</td></tr>
<tr><td>PMC3949696</td><td>c|SUB|A|1|C</td><td>DNAMutation</td><td>A1C</td><td>YES</td></tr>
<tr><td>PMC3998480</td><td>p|SUB|H|12|A</td><td>ProteinMutation</td><td>H12A</td><td>YES</td></tr>
<tr><td>PMC4013063</td><td>p|SUB|M|20|A</td><td>ProteinMutation</td><td>M20A</td><td>YES</td></tr>
<tr><td>PMC4013537</td><td>c|SUB|A|1|C</td><td>DNAMutation</td><td>A1C</td><td>YES</td></tr>
<tr><td>PMC4023268</td><td>c|SUB|A|86|C</td><td>DNAMutation</td><td>A86C</td><td>YES</td></tr>
<tr><td>PMC4051029</td><td>p|SUB|H|12|A</td><td>ProteinMutation</td><td>H12A</td><td>YES</td></tr>
<tr><td>PMC4051064</td><td>p|SUB|H|10|A</td><td>ProteinMutation</td><td>H10A</td><td>YES</td></tr>
<tr><td>PMC4051064</td><td>p|SUB|H|10|C</td><td>ProteinMutation</td><td>H10C</td><td>YES</td></tr>
<tr><td>PMC4051064</td><td>p|SUB|H|11|A</td><td>ProteinMutation</td><td>H11A</td><td>YES</td></tr>
</table>

PMC4051064	p|SUB|H|11|C	ProteinMutation	H11C	YES
PMC4051064	p|SUB|H|12|A	ProteinMutation	H12A	YES
PMC4051064	p|SUB|H|13|A	ProteinMutation	H13A	YES
PMC4051064	p|SUB|H|16|A	ProteinMutation	H16A	YES
PMC4051064	p|SUB|H|16|C	ProteinMutation	H16C	YES
PMC4051064	p|SUB|H|17|A	ProteinMutation	H17A	YES
PMC4051064	p|SUB|H|17|C	ProteinMutation	H17C	YES
PMC4051064	p|SUB|H|18|A	ProteinMutation	H18A	YES
PMC4051064	p|SUB|H|18|C	ProteinMutation	H18C	YES
PMC4051074	c|SUB|C|28|A	DNAMutation	C28A	YES
PMC4051074	p|SUB|H|10|A	ProteinMutation	H10A	YES
PMC4051074	c|SUB|C|11|A	DNAMutation	C11A	YES
PMC4051074	p|SUB|H|26|A	ProteinMutation	H26A	YES
PMC4051074	p|SUB|H|27|A	ProteinMutation	H27A	YES
PMC4051074	c|SUB|C|29|A	DNAMutation	C29A	YES
PMC4051074	p|SUB|H|29|A	ProteinMutation	H29A	YES
PMC4051074	c|SUB|C|30|A	DNAMutation	C30A	YES
PMC4051074	p|SUB|H|30|A	ProteinMutation	H30A	YES
PMC4051074	c|SUB|C|31|A	DNAMutation	C31A	YES
PMC4051074	p|SUB|H|31|A	ProteinMutation	H31A	YES
PMC4051074	c|SUB|C|32|A	DNAMutation	C32A	YES
PMC4051074	p|SUB|H|32|A	ProteinMutation	H32A	YES
PMC4051074	p|SUB|H|33|A	ProteinMutation	H33A	YES
PMC4139185	p|SUB|C|13|N	ProteinMutation	C13 N	YES
PMC4153077	p|SUB|A|16|S	ProteinMutation	A 16S	YES
PMC4257264	p|SUB|N|11|C	ProteinMutation	N11C	YES
PMC4257264	c|SUB|C|11|A	DNAMutation	C11A	YES
PMC4257264	c|SUB|C|5|A	DNAMutation	C5A	YES
PMC4257264	p|SUB|N|11|D	ProteinMutation	N11D	YES
PMC4257264	c|SUB|C|6|A	DNAMutation	C6A	YES
PMC4257264	p|SUB|H|11|A	ProteinMutation	H11A	YES
PMC4257264	p|SUB|H|11|C	ProteinMutation	H11C	YES
PMC4257264	p|SUB|H|12|C	ProteinMutation	H12C	YES
PMC4257264	p|SUB|H|13|C	ProteinMutation	H13C	YES
PMC4257264	p|SUB|H|21|C	ProteinMutation	H21C	YES
PMC4257264	p|SUB|H|31|C	ProteinMutation	H31C	YES
PMC4257264	p|SUB|H|32|C	ProteinMutation	H32C	YES
PMC4257264	p|SUB|H|41|C	ProteinMutation	H41C	YES
PMC4257264	p|SUB|H|42|C	ProteinMutation	H42C	YES
PMC4257264	p|SUB|H|51|C	ProteinMutation	H51C	YES

PMC4257264	p|SUB|H|52|C	ProteinMutation	H52C	YES
PMC4257264	p|SUB|H|12|D	ProteinMutation	H12D	YES
PMC4257264	p|SUB|H|11|D	ProteinMutation	H11D	YES
PMC4257264	p|SUB|H|21|D	ProteinMutation	H21D	YES
PMC4257264	p|SUB|H|13|D	ProteinMutation	H13D	YES
PMC4257264	p|SUB|H|14|D	ProteinMutation	H14D	YES
PMC4257264	p|SUB|H|31|D	ProteinMutation	H31D	YES
PMC4257264	p|SUB|H|32|D	ProteinMutation	H32D	YES
PMC4257264	p|SUB|H|41|D	ProteinMutation	H41D	YES
PMC4257264	p|SUB|H|42|D	ProteinMutation	H42D	YES
PMC4257264	p|SUB|H|51|D	ProteinMutation	H51D	YES
PMC4257264	p|SUB|H|52|D	ProteinMutation	H52D	YES
PMC4320108	p|SUB|T|17|N	ProteinMutation	T17N	YES
PMC4327586	p|SUB|S|010111|C	ProteinMutation	S010111C	YES
PMC4329618	c|SUB|G|1322|A	DNAMutation	G1322A	YES
PMC4329618	c|SUB|G|1312|A	DNAMutation	G1312A	YES
PMC4329618	c|SUB|G|1367|C	DNAMutation	G1367C	YES
PMC4329618	c|SUB|G|1316|A	DNAMutation	G1316A	YES
PMC4370234	p.136]	ProteinMutation	p. 136]	YES
PMC4372839	p|SUB|E|200|V	ProteinMutation	E200V	YES
PMC4378971	c|SUB|T||C	DNAMutation	T/C	YES
PMC4384578	p|SUB|H|10|A	ProteinMutation	H10A	YES
PMC4384578	p|SUB|H|10|C	ProteinMutation	H10C	YES
PMC4384578	p|SUB|H|11|A	ProteinMutation	H11A	YES
PMC4384578	p|SUB|H|11|C	ProteinMutation	H11C	YES
PMC4384578	p|SUB|H|12|A	ProteinMutation	H12A	YES
PMC4384578	p|SUB|H|12|C	ProteinMutation	H12C	YES
PMC4384578	p|SUB|H|13|C	ProteinMutation	H13C	YES
PMC4384578	p|SUB|H|13|D	ProteinMutation	H13D	YES
PMC4384578	p|SUB|H|13|E	ProteinMutation	H13E	YES
PMC4384578	p|SUB|H|13|F	ProteinMutation	H13F	YES
PMC4384578	p|SUB|H|13|A	ProteinMutation	H13A	YES
PMC4433076	c|SUB|A|1|C	DNAMutation	A1C	YES
PMC4439532	c|SUB|A|1|C	DNAMutation	A1C	YES
PMC4462250	p|SUB|L||T	ProteinMutation	L/T	YES
PMC4465688	c|SUB|C|3|G	DNAMutation	C3G	YES
PMC4513483	p|SUB|A|10|V	ProteinMutation	A10 V	YES
PMC4517832	c|SUB|C|-46|A	DNAMutation	C-46A	YES
PMC4530960	c|SUB|A|7|T	DNAMutation	A 7T	YES
PMC4592593	c|SUB|GC|-9|A	DNAMutation	GC-9A	YES

PMC4631937	p|SUB|P|2714|H	ProteinMutation	P2714H	YES
PMC4644923	p|SUB|G|12|L	ProteinMutation	G12L	YES
PMC4644923	p|SUB|G|12|H	ProteinMutation	G12H	YES
PMC4644923	p|SUB|G|30|L	ProteinMutation	G30L	YES
PMC4644923	p|SUB|H|23|R	ProteinMutation	H23R	YES
PMC4682137	p|SUB|R|6|G	ProteinMutation	R6G	YES
PMC4719921	p|SUB|H|21|A	ProteinMutation	H21A	YES
PMC4719921	p|SUB|H|21|C	ProteinMutation	H21C	YES
PMC4719921	p|SUB|H|22|A	ProteinMutation	H22A	YES
PMC4719921	p|SUB|H|22|C	ProteinMutation	H22C	YES
PMC4719921	p|SUB|H|43|A	ProteinMutation	H43A	YES
PMC4719921	p|SUB|H|43|C	ProteinMutation	H43C	YES
PMC4719921	p|SUB|H|44|A	ProteinMutation	H44A	YES
PMC4719921	p|SUB|H|44|C	ProteinMutation	H44C	YES
PMC4745523	p|SUB|G|25|N	ProteinMutation	G25N	YES
PMC4772241	c|SUB|T|2|C	DNAMutation	T2C	YES
PMC4840265	c|SUB|TC|-202|A	DNAMutation	TC-202A	YES
PMC4851292	p|SUB|S|25|N	ProteinMutation	S25N	YES
PMC4971855	p|SUB|H|13|A	ProteinMutation	H13A	YES
PMC4971855	p|SUB|H|14|A	ProteinMutation	H14A	YES
PMC4971855	p|SUB|H|31|A	ProteinMutation	H31A	YES
PMC4971855	p|SUB|H|31|C	ProteinMutation	H31C	YES
PMC5055046	c|SUB|C|10|AA	DNAMutation	C10AA	YES
PMC5055046	c|SUB|C|09|A	DNAMutation	C09A	YES
PMC5055046	c|SUB|A|1|C	DNAMutation	A1C	YES
PMC5059018	p|SUB|V|103|C	ProteinMutation	V103C	YES
PMC5088442	c|SUB|A|41|G	DNAMutation	A 41G	YES
PMC5118020	p|SUB|T|40|S	ProteinMutation	T40S	YES
PMC5119779	c|DUP|[19||[22][51][52][53][54][55]	DNAMutation	[19] [22] [51] [52] [53] [54] [55]	YES
PMC5146877	p|SUB|Q|1000|P	ProteinMutation	Q1000P	YES
PMC5244540	c|SUB|A|2|C	DNAMutation	A2 to C	YES
PMC5290578	p|SUB|H|10|A	ProteinMutation	H10A	YES
PMC5290578	p|SUB|S|2|C	ProteinMutation	S2C	YES
PMC5290578	p|SUB|H|11|A	ProteinMutation	H11A	YES
PMC5290578	p|SUB|H|12|A	ProteinMutation	H12A	YES
PMC5290578	p|SUB|H|13|A	ProteinMutation	H13A	YES
PMC5290578	p|SUB|H|13|C	ProteinMutation	H13C	YES
PMC5319737	c|SUB|CT|-90|A	DNAMutation	CT-90A	YES
PMC5319737	c|SUB|CT|90|A	DNAMutation	CT90A	YES

<table>
<tr><td>PMC5357070</td><td>c|SUB|A|1|C</td><td>DNAMutation</td><td>A1C</td><td>YES</td></tr>
<tr><td>PMC5409166</td><td>c|SUB|A|1|C</td><td>DNAMutation</td><td>A1C</td><td>YES</td></tr>
<tr><td>PMC5466104</td><td>p|SUB|K|550|X</td><td>ProteinMutation</td><td>K550X</td><td>YES</td></tr>
<tr><td>PMC5523872</td><td>p|SUB|T|28|N</td><td>ProteinMutation</td><td>T28N</td><td>YES</td></tr>
<tr><td>PMC5523872</td><td>p|SUB|T|27|N</td><td>ProteinMutation</td><td>T27N</td><td>YES</td></tr>
<tr><td>PMC5523872</td><td>p|SUB|T|37|N</td><td>ProteinMutation</td><td>T37N</td><td>YES</td></tr>
<tr><td>PMC5603897</td><td>p|SUB|M|062|X</td><td>ProteinMutation</td><td>M062X</td><td>YES</td></tr>
<tr><td>PMC5631406</td><td>c|SUB|C||T</td><td>DNAMutation</td><td>C/T</td><td>YES</td></tr>
<tr><td>PMC5657054</td><td>|SUB|DIS|2001|SEP</td><td>ProteinMutation</td><td>Dis 2001 Sep</td><td>YES</td></tr>
<tr><td>PMC56607</td><td>|[||76</td><td>DNAMutation</td><td>[76]</td><td>YES</td></tr>
</table>

REFERENCES

[1] Y. A. Adi, F. Adi-Kusumo, L. Aryati, and M. S. Hardianti. A dynamic model of pi3k/akt pathways in acute myeloid leukemia. *Journal of Applied Mathematics*, 2018, 2018.

[2] A. Allot, Y. Peng, C.-H. Wei, K. Lee, L. Phan, and Z. Lu. Litvar: a semantic search engine for linking genomic variant data in pubmed and pmc. *Nucleic Acids Research*, 46(W1):W530–W536, 2018.

[3] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh. Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 43(D1):D789–D798, 2014.

[4] E. Ames, W. Hallett, and W. J. Murphy. Sensitization of human breast cancer cells to natural killer cell-mediated cytotoxicity by proteasome inhibition. *Clinical & Experimental Immunology*, 155(3):504–513, 2009.

[5] S. Ansari, M. Donato, N. Saberian, and S. Draghici. An approach to infer putative disease-specific mechanisms using neighboring gene networks. *Bioinformatics*, 33(13):1987–1994, 2017.

[6] A. C. Antoniou, S. Casadei, T. Heikkinen, D. Barrowdale, K. Pylkäs, J. Roberts, A. Lee, D. Subramanian, K. De Leeneer, F. Fostira, et al. Breast-cancer risk in families with mutations in palb2. *New England Journal of Medicine*, 371(6):497–506, 2014.

[7] K. E. Bachman, P. Argani, Y. Samuels, N. Silliman, J. Ptak, S. Szabo, H. Konishi, B. Karakas, B. G. Blair, C. Lin, et al. The pik3ca gene is mutated with high frequency in human breast cancers. *Cancer biology & therapy*, 3(8):772–775, 2004.

[8] A. Bairoch, L. Bougueleret, S. Altairac, V. Amendolia, A. Auchincloss, G. Argoud-Puy, K. Axelsen, D. Baratin, M. C. Blatter, B. Boeckmann, J. Bolleman, L. Bollondi, E. Boutet, S. B. Quintaje, L. Breuza, A. Bridge, E. deCastro, L. Ciapina, D. Coral, E. Coudert, I. Cusin, G. Delbard, D. Dornevil, P. D. Roggli, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, S. Gehant, N. Farriol-Mathis, S. Ferro, E. Gasteiger, A. Gateau, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, N. Hulo, J. James, S. Jimenez, F. Jungo, V. Junker, T. Kappler, G. Keller, C. Lachaize, L. Lane-Guermonprez, P. Langendijk-Genevaux, V. Lara, P. Lemercier, V. Le Saux, D. Lieberherr, T. d. e. O. Lima, V. Mangold, X. Martin, P. Masson, K. Michoud, M. Moinat, A. Morgat, A. Mottaz, S. Paesano, I. Pedruzzi, I. Phan, S. Pilbout, V. Pillet, S. Poux, M. Pozzato, N. Redaschi, S. Reynaud, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. L. Veuthey, L. Yip, L. Zuletta, R. Apweiler, Y. Alam-Faruque, R. Antunes, D. Barrell, D. Binns, L. Bower, P. Browne, W. M. Chan, E. Dimmer, R. Eberhardt, A. Fedotov, R. Foulger, J. Garavelli, R. Golin, A. Horne, R. Huntley, J. Jacobsen, M. Kleen, P. Kersey, K. Laiho, R. Leinonen, D. Legge, Q. Lin, M. Magrane, M. J. Martin, C. O'Donovan, S. Orchard, J. O'Rourke, S. Patient, M. Pruess, A. Sitnov, E. Stanley, M. Corbett, G. di Martino, M. Donnelly, J. Luo, P. van Rensburg, C. Wu, C. Arighi, L. Arminski, W. Barker, Y. Chen, Z. Z. Hu, H. K. Hua, H. Huang, R. Mazumder, P. McGarvey, D. A. Natale, A. Nikolskaya, N. Petrova, B. E. Suzek, S. Vasudevan, C. R. Vinayaka, L. S. Yeh, and J. Zhang. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Research*, 37:D169–174, Jan 2009.

[9] C. J. Baker and R. Witte. Mutation mining—a prospector's tale. *Information Systems*

Frontiers, 8(1):47–57, 2006.

[10] S. P. Balk and K. E. Knudsen. Ar, the cell cycle, and prostate cancer. *Nuclear Receptor Signaling*, 6(1):nrs–06001, 2008.

[11] C. E. Barbieri, S. C. Baca, M. S. Lawrence, F. Demichelis, M. Blattner, J.-P. Theurillat, T. A. White, P. Stojanov, E. Van Allen, N. Stransky, et al. Exome sequencing identifies recurrent spop, foxa1 and med12 mutations in prostate cancer. *Nature Genetics*, 44(6):685, 2012.

[12] T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W. C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar. NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Research*, 33(Database Issue):D562–6, 2005.

[13] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of The Royal Statistical Society B*, 57(1):289–300, 1995.

[14] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, August 2001.

[15] C. Béroud, G. Collod-Béroud, C. Boileau, T. Soussi, and C. Junien. Umd (universal mutation database): a generic software to build and analyze locus-specific databases. *Human Mutation*, 15(1):86–94, 2000.

[16] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'donovan, I. Phan, et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Research*, 31(1):365–370, 2003.

[17] J. Bonis, L. I. Furlong, and F. Sanz. Osiris: a tool for retrieving literature about sequence variants. *Bioinformatics*, 22(20):2567–2569, 2006.

[18] D. Bonnet and J. E. Dick. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature Medicine*, 3(7):730–737, 1997.

[19] G. Boysen, C. E. Barbieri, D. Prandi, M. Blattner, S.-S. Chae, A. Dahija, S. Nataraj, D. Huang, C. Marotz, L. Xu, et al. Spop mutation leads to genomic instability in prostate cancer. *Elife*, 4:e09207, 2015.

[20] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[21] R. Buisson, A.-M. Dion-Côté, Y. Coulombe, H. Launay, H. Cai, A. Z. Stasiak, A. Stasiak, B. Xia, and J.-Y. Masson. Cooperation of breast cancer proteins palb2 and piccolo brca2 in stimulating homologous recombination. *Nature Structural & Molecular Biology*, 17(10):1247, 2010.

[22] I. G. Campbell, S. E. Russell, D. Y. Choong, K. G. Montgomery, M. L. Ciavarella, C. S. Hooi, B. E. Cristiano, R. B. Pearson, and W. A. Phillips. Mutation of the pik3ca gene in ovarian and breast cancer. *Cancer Research*, 64(21):7678–7681, 2004.

[23] J. G. Caporaso, W. A. Baumgartner Jr, D. A. Randolph, K. B. Cohen, and L. Hunter. Mutationfinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23(14):1862–1865, 2007.

[24] E. Capriotti, N. L. Nehrt, M. G. Kann, and Y. Bromberg. Bioinformatics for personal genome interpretation. *Briefings in Bioinformatics*, 13(4):495–512, 2012.

[25] M. Cariaso and G. Lennon. Snpedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Research*, 40(D1):D1308–D1312, 2011.

[26] G. Casey, P. J. Neville, S. J. Plummer, Y. Xiang, L. M. Krumroy, E. A. Klein, W. J. Catalona, N. Nupponen, J. D. Carpten, J. M. Trent, et al. Rnasel arg462gln variant is implicated in up to 13% of prostate cancer cases. *Nature Genetics*, 32(4):581, 2002.

[27] A. Chaturvedi, M. M. A. Cruz, N. Jyotsana, A. Sharma, H. Yun, K. Görlich, M. Wichmann, A. Schwarzer, M. Preller, F. Thol, et al. Mutant idh1 promotes leukemogenesis in vivo and can be specifically targeted in human aml. *Blood*, 122(16):2877–2887, 2013.

[28] H. M. Dingerdissen, J. Torcivia-Rodriguez, Y. Hu, T.-C. Chang, R. Mazumder, and R. Kahsay. Biomuta and bioxpress: mutation and expression knowledgebases for cancer biomarker discovery. *Nucleic Acids Research*, 46(D1):D1128–D1136, 2018.

[29] E. Doughty, A. Kertesz-Farkas, O. Bodenreider, G. Thompson, A. Adadey, T. Peterson, and M. G. Kann. Toward an automatic method for extracting cancer-and other disease-related point mutations from the biomedical literature. *Bioinformatics*, 27(3):408–415, 2010.

[30] S. Draghici. *Statistics and Data Analysis for Microarrays using R and Bioconductor*. Chapman and Hall/CRC Press, 2011.

[31] T. H. Ecke, H. H. Schlechte, K. Schiemenz, M. D. Sachs, S. V. Lenk, B. D. Rudolph, and S. A. Loening. Tp53 gene mutations in prostate cancer progression. *Anticancer Research*, 30(5):1579–1586, 2010.

[32] M. Erdogmus and O. U. Sezerman. Application of automatic mutation–gene pair extraction to diseases. *Journal of Bioinformatics and Computational Biology*, 5(06):1261–1275, 2007.

[33] S. Faderl, A. Ferrajoli, D. Harris, Q. Van, H. M. Kantarjian, and Z. Estrov. Atiprimod blocks phosphorylation of jak-stat and inhibits proliferation of acute myeloid leukemia (aml) cells. *Leukemia Research*, 31(1):91–95, 2007.

[34] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, et al. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 39(suppl_1):D945–D950, 2010.

[35] D. Ford, D. Easton, M. Stratton, S. Narod, D. Goldgar, P. Devilee, D. Bishop, B. Weber, G. Lenoir, J. Chang-Claude, et al. Genetic heterogeneity and penetrance analysis of the brca1 and brca2 genes in breast cancer families. *The American Journal of Human Genetics*, 62(3):676–689, 1998.

[36] C. Fribbens, B. O'Leary, L. Kilburn, S. Hrebien, I. Garcia-Murillas, M. Beaney, M. Cristofanilli, F. Andre, S. Loi, S. Loibl, et al. Plasma esr1 mutations and the treatment of estrogen receptor-positive advanced breast cancer. *Journal of Clinical Oncology*, 2016.

[37] V. I. Gaidzik, L. Bullinger, R. F. Schlenk, A. S. Zimmermann, J. Röck, P. Paschka, A. Corbacioglu, J. Krauter, B. Schlegelberger, A. Ganser, et al. Runx1 mutations in acute myeloid leukemia: results from a comprehensive genetic and clinical analysis from the aml study group. *Journal of Clinical Oncology*, 29(10):1364–1372, 2011.

[38] V. I. Gaidzik, P. Paschka, D. Spath, M. Habdank, C. Kohne, U. Germing, M. von Lilienfeld-Toal, G. Held, H.-A. Horst, D. Haase, et al. Tet2 mutations in acute myeloid leukemia (aml): results from a comprehensive genetic and clinical analysis of the aml study group. *Journal of Clinical Oncology*, 30(12):1350–1357, 2012.

[39] Y. Garten and R. B. Altman. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. In *BMC Bioinformatics*, volume 10, page S6. Springer, 2009.

[40] Genomenon. Mastermind: Automated gene panel design mobilizing evidence from the medical literature. https://mastermind.genomenon.com, 2017.

[41] A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, D. Tamborero, M. P. Schroeder, A. Jene-Sanz, A. Santos, and N. Lopez-Bigas. Intogen-mutations identifies cancer drivers across tumor types. *Nature Methods*, 10(11):1081–1082, 2013.

[42] M. Griffith, N. C. Spies, K. Krysiak, J. F. McMichael, A. C. Coffman, A. M. Danos, B. J. Ainscough, C. A. Ramirez, D. T. Rieke, L. Kujan, et al. Civic is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature Genetics*, 49(2):170, 2017.

[43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[44] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl_1):D514–D517, 2005.

[45] D. Harari and Y. Yarden. Molecular mechanisms underlying erbb2/her2 action in breast cancer. *Oncogene*, 19(53):6102, 2000.

[46] K. M. Hardy, B. W. Booth, M. J. Hendrix, D. S. Salomon, and L. Strizzi. Erbb/egf signaling and emt in mammary development and breast cancer. *Journal of Mammary Gland Biology and Neoplasia*, 15(2):191–199, 2010.

[47] M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman, and T. E. Klein. Pharmgkb: the pharmacogenetics knowledge base. *Nucleic Acids Research*, 30(1):163–165, 2002.

[48] F. Holst, P. R. Stahl, C. Ruiz, O. Hellwinkel, Z. Jehan, M. Wendland, A. Lebeau, L. Terracciano, K. Al-Kuraya, F. Jänicke, et al. Estrogen receptor alpha (esr1) gene amplification is frequent in breast cancer. *Nature Genetics*, 39(5):655, 2007.

[49] F. Horn, A. L. Lau, and F. E. Cohen. Automated extraction of mutation data from the literature: application of mutext to g protein-coupled receptors and nuclear hormone receptors. *Bioinformatics*, 20(4):557–568, 2004.

[50] S. E. Hunt, W. McLaren, L. Gil, A. Thormann, H. Schuilenburg, D. Sheppard, A. Parton, I. M. Armean, S. J. Trevanion, P. Flicek, et al. Ensembl variation resources. *Database*, 2018, 2018.

[51] I. Ihnatova, V. Popovici, and E. Budinska. A critical comparison of topology-based pathway analysis methods. *PloS One*, 13(1):e0191154, 2018.

[52] S. J. Isakoff, J. A. Engelman, H. Y. Irie, J. Luo, S. M. Brachmann, R. V. Pearline, L. C. Cantley, and J. S. Brugge. Breast cancer–associated PIK3CA mutations are oncogenic in mammary epithelial cells. *Cancer Research*, 65(23):10992–11000, 2005.

[53] J. M. Izarzugaza, M. Krallinger, and A. Valencia. Interpretation of the consequences of mutations in protein kinases: combined use of bioinformatics and text mining. *Frontiers in Physiology*, 3:323, 2012.

[54] R. Jeselsohn, G. Buchwalter, C. De Angelis, M. Brown, and R. Schiff. Esr1 mutations—a mechanism for acquired endocrine resistance in breast cancer. *Nature Reviews Clinical Oncology*, 12(10):573, 2015.

[55] A. Jimeno Yepes and K. Verspoor. Literature mining of genetic variants for curation: quantifying the importance of supplementary material. *Database*, 2014, 2014.

[56] T. M. Kadia, P. Jain, F. Ravandi, G. Garcia-Manero, M. Andreef, K. Takahashi, G. Borthakur, E. Jabbour, M. Konopleva, N. G. Daver, et al. Tp53 mutations in newly diagnosed acute myeloid leukemia: clinicomolecular characteristics, response to therapy, and outcomes. *Cancer*, 122(22):3484–3491, 2016.

[57] P. Khatri and S. Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005.

[58] P. Khatri, M. Sirota, and A. J. Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLOS Computational Biology*, 8(2):e1002375, 2012.

[59] J.-H. Kim, C. Lee, H. S. Cheong, Y. Koh, K.-S. Ahn, H.-L. Kim, H. D. Shin, and S.-S. Yoon. Slc29a1 (ent1) polymorphisms and outcome of complete remission in acute myeloid leukemia. *Cancer Chemotherapy and Pharmacology*, 78(3):533–540, 2016.

[60] Z. Kote-Jarai, A. A. Al Olama, D. Leongamornlert, M. Tymrakiewicz, E. Saunders, M. Guy, G. Giles, G. Severi, M. Southey, J. Hopper, et al. Identification of a novel prostate cancer susceptibility variant in the klk3 gene transcript. *Human Genetics*, 129(6):687, 2011.

[61] Y. Kubota, H. Ohnishi, A. Kitanaka, T. Ishida, and T. Tanaka. Constitutive activation of pi3k is involved in the spontaneous proliferation of primary acute myeloid leukemia cells: direct evidence of pi3k activation. *Leukemia*, 18(8):1438–1440, 2004.

[62] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[63] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1):D980–D985, 2013.

[64] R. Leaman, R. Islamaj Doğan, and Z. Lu. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917, 2013.

[65] R. Leaman, C.-H. Wei, and Z. Lu. tmchem: a high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, 7(S1):S3, 2015.

[66] K. Lee, S. Lee, S. Park, S. Kim, S. Kim, K. Choi, A. C. Tan, and J. Kang. Bronco: Biomedical entity relation oncology corpus for extracting gene-variant-disease-drug relations. *Database*, 2016, 2016.

[67] L. C. Lee, F. Horn, and F. E. Cohen. Automatic extraction of protein point mutations using a graph bigram association. *PLoS Computational Biology*, 3(2):e16, 2007.

[68] T. J. Ley, L. Ding, M. J. Walter, M. D. McLellan, T. Lamprecht, D. E. Larson, C. Kandoth, J. E. Payton, J. Baty, J. Welch, et al. Dnmt3a mutations in acute myeloid leukemia. *New England Journal of Medicine*, 363(25):2424–2433, 2010.

[69] M. Liu, A. Liberzon, S. W. Kong, W. R. Lai, P. J. Park, I. S. Kohane, and S. Kasif. Network-based analysis of affected biological processes in type 2 diabetes models. *PLOS Genetics*, 3(6):e96, 2007.

[70] M. R. Luskin, A. O. Huen, S. A. Brooks, C. Stewart, C. D. Watt, J. J. Morrissette, D. B. Lieberman, A. Bagg, M. Rosenbach, and A. E. Perl. Npm1 mutation is associated with leukemia cutis in acute myeloid leukemia with monocytic features. *Haematologica*, 100(10):e412, 2015.

[71] J. Ma, A. Shojaie, and G. Michailidis. A comparative study of topology-based pathway enrichment analysis methods. *BMC bioinformatics*, 20(1):546, 2019.

[72] A. Mahmood. *Text mining of mutations and their impact from biomedical literature*. PhD thesis, University of Delaware, 2018.

[73] R. T. McDonald, R. S. Winters, M. Mandel, Y. Jin, P. S. White, and F. Pereira. An entity tagger for recognizing acquired genomic variations in cancer literature. *Bioinformatics*, 20(17):3249–3251, 2004.

[74] R. E. Mills, K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. Lam, J. Leng, R. Li, Y. Li, C.-Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stütz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, J. O. Korbel, and 1000 Genomes Project. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, 2011.

[75] C. Mitrea, Z. Taghavi, B. Bokanizad, S. Hanoudi, R. Tagett, M. Donato, C. Voichiţa, and S. Draghici. Methods and approaches in the topology-based analysis of biolog-

ical pathways. *Frontiers in Physiology*, 4:278, 2013.

[76] T. Nguyen, C. Mitrea, and S. Draghici. Network-based approaches for pathway level analysis. *Current Protocols in Bioinformatics*, 61(1):8–25, 2018.

[77] T.-M. Nguyen, A. Shafi, T. Nguyen, and S. Draghici. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biology*, 20(1):1–15, 2019.

[78] J.-P. Plazzer, R. H. Sijmons, M. O. Woods, P. Peltomäki, B. Thompson, J. T. Den Dunnen, and F. Macrae. The insight database: utilizing 100 years of insights into lynch syndrome. *Familial Cancer*, 12(2):175–180, 2013.

[79] S. Povey, R. Lovering, E. Bruford, M. Wright, M. Lush, and H. Wain. The hugo gene nomenclature committee (hgnc). *Human Genetics*, 109(6):678–680, 2001.

[80] K. W. Pratz, T. Sato, K. M. Murphy, A. Stine, T. Rajkhowa, and M. Levis. Flt3-mutant allelic burden and clinical status are predictive of response to flt3 inhibitors in aml. *Blood*, 115(7):1425–1432, 2010.

[81] N. Rahman, S. Seal, D. Thompson, P. Kelly, A. Renwick, A. Elliott, S. Reid, K. Spanova, R. Barfoot, T. Chagtai, H. Jayatilake, L. McGuffog, S. Hanks, G. Evans, and Eccles. Palb2, which encodes a brca2-interacting protein, is a breast cancer susceptibility gene. *Nature Genetics*, 39(2):165, 2007.

[82] D. Rebholz-Schuhmann, S. Marcel, S. Albert, R. Tolle, G. Casari, and H. Kirsch. Automatic extraction of mutations from medline and cross-validation with omim. *Nucleic Acids Research*, 32(1):135–142, 2004.

[83] H. L. Rehm, J. S. Berg, and S. E. Plon. Clingen and clinvar–enabling genomics in precision medicine. *Human Mutation*, 39(11):1473–1475, 2018.

[84] F. Revillion, J. Bonneterre, and J. Peyrat. Erbb2 oncogene in human breast cancer and its clinical significance. *European Journal of Cancer*, 34(6):791–808, 1998.

[85] C. J. Richardson, Q. Gao, C. Mitsopoulous, M. Zvelebil, L. H. Pearl, and F. M. Pearl. Mokca database—mutations of kinases in cancer. *Nucleic Acids Research*, 37(suppl_1):D824–D831, 2009.

[86] T. W. Rinker. *sentimentr: Calculate Text Polarity Sentiment*. Buffalo, New York, 2018. version 2.3.2.

[87] D. R. Robinson, Y.-M. Wu, P. Vats, F. Su, R. J. Lonigro, X. Cao, S. Kalyana-Sundaram, R. Wang, Y. Ning, L. Hodges, et al. Activating esr1 mutations in hormone-resistant metastatic breast cancer. *Nature Genetics*, 45(12):1446, 2013.

[88] R. Roushangar and G. I. Mias. Multi-study reanalysis of 2,213 acute myeloid leukemia patients reveals age-and sex-dependent gene expression signatures. *Scientific Reports*, 9(1):1–17, 2019.

[89] N. Saberian, A. Shafi, A. Peyvandipour, and S. Draghici. MAGPEL: an automated pipeline for inferring variant-driven gene panels from the full-length biomedical literature. *Scientific Reports*, 10(1):1–11, 2020.

[90] C. L. Sawyers. Chronic myeloid leukemia. *New England Journal of Medicine*, 340(17):1330–1340, 1999.

[91] A. Shafi, T. Nguyen, A. Peyvandipour, and S. Draghici. GSMA: an approach to identify robust global and test Gene Signatures using Meta-Analysis. *Bioinformatics*, 36(2):487–495, 2020.

[92] Z. Shan, Y. Li, S. Yu, J. Wu, C. Zhang, Y. Ma, G. Zhuang, J. Wang, Z. Gao, and D. Liu. Ctcf regulates the foxo signaling pathway to affect the progression of prostate can-

cer. *Journal of Cellular and Molecular Medicine*, 23(5):3130–3139, 2019.

[93] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 2001.

[94] S. Shukla, G. T. MacLennan, D. J. Hartman, P. Fu, M. I. Resnick, and S. Gupta. Activation of pi3k-akt signaling pathway promotes prostate cancer cell invasion. *International Journal of Cancer*, 121(7):1424–1432, 2007.

[95] A. Singh, A. Olowoyeye, P. H. Baenziger, J. Dantzer, M. G. Kann, P. Radivojac, R. Heiland, and S. D. Mooney. Mutdb: update on development of tools for the biochemical analysis of genetic variation. *Nucleic Acids Research*, 36(suppl_1):D815–D819, 2007.

[96] A. Singhal, M. Simmons, and Z. Lu. Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature. *Journal of the American Medical Informatics Association*, 23(4):766–772, 2016.

[97] A. Singhal, M. Simmons, and Z. Lu. Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. *PLoS Computational Biology*, 12(11):e1005017, 2016.

[98] K. Stemke-Hale, A. M. Gonzalez-Angulo, A. Lluch, R. M. Neve, W.-L. Kuo, M. Davies, M. Carey, Z. Hu, Y. Guan, A. Sahin, et al. An integrative genomic and proteomic analysis of pik3ca, pten, and akt mutations in breast cancer. *Cancer Research*, 68(15):6084–6091, 2008.

[99] P. D. Stenson, M. Mort, E. V. Ball, K. Howells, A. D. Phillips, N. S. Thomas, and D. N. Cooper. The human gene mutation database: 2008 update. *Genome Medicine*,

1(1):13, 2009.

[100] N. O. Stitziel, T. A. Binkowski, Y. Y. Tseng, S. Kasif, and J. Liang. toposnp: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Research*, 32(suppl_1):D520–D522, 2004.

[101] B. A. Talseth-Palmer and R. J. Scott. Genetic variation and its role in malignancy. *International Journal of Biomedical Science: IJBS*, 7(3):158, 2011.

[102] A. L. Tarca, S. Draghici, G. Bhatti, and R. Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13(1):136, 2012.

[103] TCGA Research Network. The Cancer Genome Atlas Research Network. `http://cancergenome.nih.gov/`.

[104] G. A. Thorisson, O. Lancaster, R. C. Free, R. K. Hastings, P. Sarmah, D. Dash, S. K. Brahmachari, and A. J. Brookes. Hgvbaseg2p: a central genetic association database. *Nucleic Acids Research*, 37(suppl_1):D797–D802, 2008.

[105] C. F. Thorn, T. E. Klein, and R. B. Altman. Pharmgkb: the pharmacogenomics knowledge base. In *Pharmacogenomics*, pages 311–320. Springer, 2013.

[106] M. Tischkowitz, B. Xia, N. Sabbaghian, J. S. Reis-Filho, N. Hamel, G. Li, E. H. Van Beers, L. Li, T. Khalil, and Quenneville. Analysis of palb2/fancn-associated breast cancer families. *Proceedings of the National Academy of Sciences*, 104(16):6788–6793, 2007.

[107] W. Toy, Y. Shen, H. Won, B. Green, R. A. Sakr, M. Will, Z. Li, K. Gala, S. Fanning, T. A. King, et al. Esr1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nature Genetics*, 45(12):1439, 2013.

[108] L. Tryggvadóttir, L. Vidarsdóttir, T. Thorgeirsson, J. G. Jonasson, E. J. Ólafsdóttir, G. H. Ólafsdóttir, T. Rafnar, S. Thorlacius, E. Jonsson, J. E. Eyfjord, et al. Prostate cancer progression and survival in brca2 mutation carriers. *Journal of the National Cancer Institute*, 99(12):929–935, 2007.

[109] M. L. Ufkin, S. Peterson, X. Yang, H. Driscoll, C. Duarte, and P. Sathyanarayana. mir-125a regulates cell cycle, proliferation, and apoptosis by targeting the erbb pathway in acute myeloid leukemia. *Leukemia Research*, 38(3):402–410, 2014.

[110] J. Ursini-Siegel, B. Schade, R. D. Cardiff, and W. J. Muller. Insights from transgenic mouse models of erbb2-induced breast cancer. *Nature Reviews Cancer*, 7(5):389, 2007.

[111] T. Walsh, S. Casadei, K. H. Coats, E. Swisher, S. M. Stray, J. Higgins, K. C. Roach, J. Mandell, M. K. Lee, S. Ciernikova, et al. Spectrum of mutations in brca1, brca2, chek2, and tp53 in families at high risk of breast cancer. *Jama*, 295(12):1379–1388, 2006.

[112] C.-H. Wei, B. R. Harris, H.-Y. Kao, and Z. Lu. tmvar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, 29(11):1433–1439, 2013.

[113] C.-H. Wei, H.-Y. Kao, and Z. Lu. Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed Research International*, 2015, 2015.

[114] C.-H. Wei, L. Phan, J. Feltz, R. Maiti, T. Hefferon, and Z. Lu. tmvar 2.0: integrating genomic variant information from literature with dbsnp and clinvar for precision medicine. *Bioinformatics*, 34(1):80–87, 2017.

[115] T.-J. Wu, A. Shamsaddini, Y. Pan, K. Smith, D. J. Crichton, V. Simonyan, and R. Mazumder. A framework for organizing cancer-related variations from existing databases, publications and ngs data using a high-performance integrated virtual environment (hive). *Database*, 2014, 2014.

[116] W. Xia, C. M. Gerard, L. Liu, N. M. Baudson, T. L. Ory, and N. L. Spector. Combining lapatinib (gw572016), a small molecule inhibitor of erbb1 and erbb2 tyrosine kinases, with therapeutic anti-erbb2 antibodies enhances apoptosis of erbb2-overexpressing breast cancer cells. *Oncogene*, 24(41):6213, 2005.

[117] Y. Yang, Q. Huang, Y. Lu, X. Li, and S. Huang. Reactivating pp2a by fty720 as a novel therapy for aml with c-kit tyrosine kinase domain mutation. *Journal of Cellular Biochemistry*, 113(4):1314–1322, 2012.

[118] S. Yeniterzi and U. Sezerman. Enzyminer: automatic identification of protein level mutations and their impact on target enzymes from pubmed abstracts. *BMC Bioinformatics*, 10(8):S2, 2009.

[119] Y. L. Yip, N. Lachenal, V. Pillet, and A.-L. Veuthey. Retrieving mutation-specific information for human proteins in uniprot/swiss-prot knowledgebase. *Journal of Bioinformatics and Computational Biology*, 5(06):1215–1231, 2007.

[120] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. Clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a Journal of Integrative Biology*, 16(5):284–287, 2012.

[121] F. Zhang, Q. Fan, K. Ren, and P. R. Andreassen. Palb2 functionally connects the breast cancer susceptibility proteins brca1 and brca2. *Molecular Cancer Research*, 7(7):1110–1118, 2009.

[122] J. Zhang, R. Bajari, D. Andric, F. Gerthoffert, A. Lepsa, H. Nahal-Bose, L. D. Stein, and V. Ferretti. The international cancer genome consortium data portal. *Nature Biotechnology*, 37(4):367–369, 2019.

[123] J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, et al. International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database*, 2011, 2011.

[124] J. Zhang, R. Chiodini, A. Badr, and G. Zhang. The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, 38(3):95–109, 2011.

ABSTRACT

TEXT MINING OF VARIANT-GENOTYPE-PHENOTYPE ASSOCIATIONS FROM BIOMEDICAL LITERATURE

by

NAFISEH SABERIAN

December 2020

Advisor: Dr. Sorin Draghici

Major: Computer Science

Degree: Master of Science

In spite of the efforts in developing and maintaining accurate variant databases, a large number of disease-associated variants are still hidden in the biomedical literature. Curation of the biomedical literature in an effort to extract this information is a challenging task due to i) the complexity of natural language processing, ii) inconsistent use of standard recommendations for variant description, and iii) the lack of clarity and consistency in describing the variant-genotype-phenotype associations in the biomedical literature. In this article, we employ text mining and word cloud analysis techniques to address these challenges. The proposed framework extracts the variant-gene-disease associations from the full-length biomedical literature and designs an evidence-based variant-driven gene panel for a given condition. We validate the identified genes by showing their diagnostic abilities to predict the patients' clinical outcomes on several independent validation cohorts. As representative examples, we present our results for acute myeloid leukemia (AML), breast cancer, and prostate cancer. We compare these panels with other variant-driven gene panels obtained from Clinvar, Mastermind, and others from literature, as well

as with a panel identified with a classical differentially expressed genes (DEGs) approach. The results show that the panels obtained by the proposed framework yield better results than the other gene panels currently available in the literature.

AUTOBIOGRAPHICAL STATEMENT

NAFISEH SABERIAN

EDUCATION

- M.S. Computer Science, Wayne State University, Detroit, MI, USA, 2020.
- B.S. Software Engineering, Ferdowsi University of Mashhad, Iran, 2012.

PUBLICATIONS

- N. Saberian, A. Shafi, A. Peyvandipour, and S. Draghici. MAGPEL: an automated pipeline for inferring variant-driven gene panels from the full-length biomedical literature. Scientific Reports, 10(1):1–11, 2020.
- N. Saberian, A. Peyvandipour, M. Donato, S. Ansari, and S. Draghici. A new computational drug repurposing method using established disease–drug pair knowledge. Bioinformatics, 35(19):3672–3678, 2019.
- A. Peyvandipour, A. Shafi, N. Saberian, and S. Draghici. Identification of cell types from single cell data using stable clustering. Scientific reports, 10(1):1–12, 2020.
- A. Peyvandipour, N. Saberian, A. Shafi, M. Donato, and S. Draghici. A novel computational approach for drug repurposing using systems biology. Bioinformatics, 34(16):2817–2825, 2018.
- S. Ansari, M. Donato, N. Saberian, and S. Draghici. An approach to infer putative disease-specific mechanisms using neighboring gene networks. Bioinformatics, 33(13):1987–1994, 2017.