12-18-2016

Integrative Analysis of Heterogeneous Genomics Data for Triple Negative Breast Cancer and High Grade Serous Ovarian Cancer

Abdelrahman Hosny Ibrahim
abdelrahman.hosny@ieee.org

Integrative Analysis of Heterogeneous Genomics Data for Triple Negative Breast Cancer and High Grade Serous Ovarian Cancer

Abdelrahman Hosny Mohammed Ibrahim

B.S., Computer Science, Assiut University, 2013

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

at the

University of Connecticut

2016

APPROVAL PAGE

Master of Science Thesis

Integrative Analysis of Heterogeneous Genomics Data for Triple-Negative Breast Cancer and High Grade Serous Ovarian Cancer

Presented by

Abdelrahman Hosny Mohammed Ibrahim, B.S.

Co-Major Advisor_____

Reda Ammar

Co-Major Advisor_____

Sheida Nabavi

Associate Advisor_____

Sanguthevar Rajasekaran

Associate Advisor_____

Yufeng Wu

University of Connecticut

December 2016

ACKNOWLEDGEMENTS

I cannot express enough thanks to my committee for their continued support and encouragement: Prof. Sheida Nabavi, Prof. Reda Ammar, Prof. Sanguthevar Rajasekaran and Prof. Yufeng Wu. I express my sincere appreciation for the learning opportunities provided by my committee. I would like also to thank my lab mates for their generous assistance in the completion of this thesis: Fatima Zare, Tianyu Wang, Dina Abdelhafiz and Nick Monteleone.

The completion of this thesis could not have been accomplished without the generous support of my sister, Eman – thank you for bridging biology and computer science. The countless times you helped me understand the language of biology will not be forgotten. Finally, for the never-ending encouragement, I dedicate this work to my parents.

Table of Contents

Table of Figures

Table of Tables

Abstract

The human body is made up of trillions of cells. Although all the human body cells contain the same DNA sequence inside their nuclei, each one carries out its own function. Normally, human cells grow and divide to form daughter cells as the body needs them. When cells grow old, or lose their ability to function properly, they die (in a very organized way called apoptosis or programmed cell death) and new cells take their role. Cancer is a disease that is caused by uncontrolled division of abnormal cells in some part of the body, breaking the natural process of growing. Old or damaged cells survive when they should die, and new (abnormal) cells form when they are not needed. Some types of cancer form solid tumors, which are masses of tissue. Others, such as leukemias, do not form solid tumors. It is widely believed that cancer is caused by the accumulation of detrimental variation in the genome over the course of a lifetime. Variations can take several forms. Single Nucleotide Polymorphism (SNP) is a mutation in a single base of the DNA. Indels describe insertions or deletions of bases in the genome. Copy Number Variation (CNV) represents multiplied and deleted segments in a genome. Most of the time, one type of mutation is not sufficient to induce cancer formation.

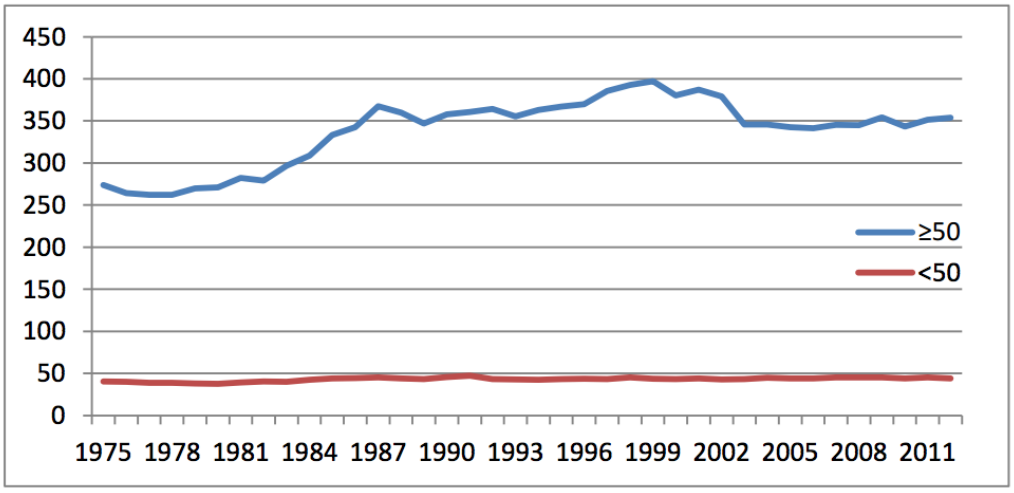
In this study, we have investigated genomic datasets of a phase-1 clinical trial on triple-negative breast cancer and ovarian cancer patients. The goal is to identify genes that drive drug resistance. We have developed data analysis pipelines to obtain genomics variations (somatic mutations and copy number variations) from the Whole Exome Sequencing (WES) raw data of 35 triple-negative breast cancer (TNBC) and ovarian cancer patients. In addition, we have analyzed the gene expression levels and gene fusion from the RNA-Seq raw reads data for a subset of 16 patients. This study is an effort toward optimizing the integrative analysis of genomic datasets under certain limitations. The main limitation is the small number of samples in the clinical trial (as is the case in most clinical trials). Another challenge is to find an abstract way to analyze the raw sequencing data given its large size and heterogeneity. The novelty of our work comes in following a data science approach in answering such research questions. The unbiased and data-driven approach was successful in identifying genes that are most likely related to the drug resistance. Our results will guide clinicians toward having an in-depth study of the driver genes.

Chapter 1: Introduction

The human body is made up of trillions of cells. Although all human body cells contain the same DNA sequence inside their nuclei, each one carries out its own function. Normally, human cells grow and divide to form daughter cells as the body needs them. When cells grow old, or lose their ability to function properly, they die (in a very organized way called apoptosis or programmed cell death) and new cells take their role. Cancer is a disease that is caused by uncontrolled division of abnormal cells in some part of the body, breaking the natural process of growing [1]. Old or damaged cells survive when they should die, and new (abnormal) cells form when they are not needed. Some types of cancer form solid tumors, which are masses of tissue. Others, such as leukemias, do not form solid tumors.

The Cancer

Cancer Research UK studies show that 8.2 million people died in 2012 worldwide because of cancer [2]. Worldwide, breast cancer accounts for nearly a quarter of all cancers in women and it is estimated that 1.7 million women are diagnosed with the disease annually. Although the incidence rates of invasive breast cancer have remained stable over the past several decades among women younger than 50, Figure 1 shows substantial changes in rates among women older than 50 years old. Moreover, Figure 2 shows the rate of metastatic breast cancer at initial diagnosis in the United States. There was no remarkable change in the rates despite the widespread use of mammography for early detection [3].

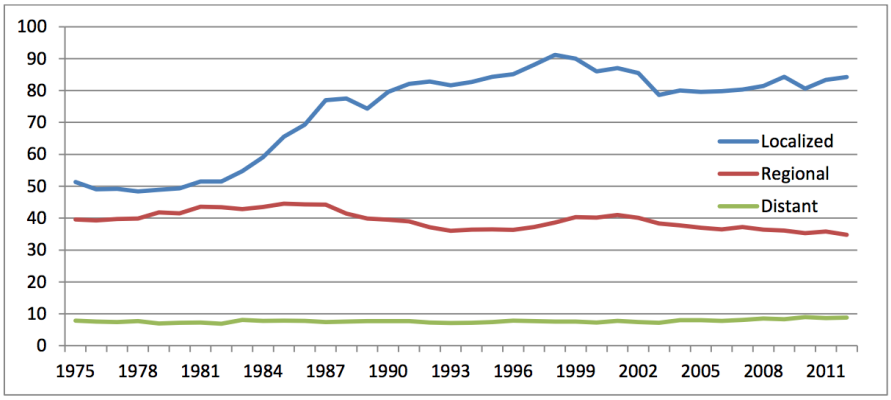


Source: SEER 9 registries, November 2014 data submission. Rates are per 100,000 and age-adjusted to the 2000 US Standard Population.

Figure 1 - U.S. incidence rates of invasive breast cancer among women <50 and ≥50, 1975-2012

There are two categories of genes that are directly related to cancer. The first category is called oncogenes, which boost the activity of a cell to grow and divide abnormally. The second category is tumor-suppressor genes, which kill cells that goes through abnormal division path [1]. Cancer cells either over-express oncogenes, under-express tumor-suppressor genes or both. It has been revealed that cancer is caused by an accumulation of detrimental variation in the genome over the course of a lifetime. Variations can take several forms. Single Nucleotide Polymorphism (SNP) is a variation in a single base

of the DNA. Indels describe insertions or deletions of bases in a genome. Most of the time, a single mutation is not sufficient to induce cancer formation. Larger regions of mutations in the DNA lead to the dysfunction (over-expression or under-expression) of genes (oncogenes or tumor-suppressor genes) that ultimately cause cancer.



Source: SEER 9 registries, November 2014 data submission. Rates are per 100,000 and age-adjusted to the 2000 US Standard Population. Localized – confined to the breast; regional – spread to regional lymph nodes; distant – metastatic disease.

Figure 2 - Female Breast Cancer Incidence Rates by Stage, US, 1975-2012

Triple-Negative Breast Cancer (TNBC) and High Grade Serous Ovarian Cancer

In this work, we focus our efforts on one type of breast cancer called Triple-Negative Breast Cancer (abbreviated TNBC) and its closely similar cancer, ovarian cancer [4]. TNBC refers to any type of breast cancer that does not express the genes for estrogen receptors (*ER*), progesterone receptors (*PR*) or *HER2/neu* receptors. *HER2* is a type of oncogenes that is overexpressed in some types of breast cancer (*ER* negative, *PR* negative and *HER2* negative). That's why it is called triple negative. This classification leads to therapeutic implications because patients cannot benefit from endocrine therapy as it lacks ER and PR. In addition, they cannot benefit from Anti-*HER2* agents since it is *HER2* negative. This makes it more difficult to treat as most chemotherapies target one of the three receptors, so it requires a combination of therapies [5].

According to [6], TNBC accounts for 12-17% of all breast cancers. It is more prevalent in younger women (< 50 years) and African and Hispanic descent. There is a huge enrichment for tumors of the triple negative phenotype in the subset of patients with *BRCA1* germline mutations. *BRCA1* is considered as a tumor-suppressor gene responsible for the DNA repair, and a mutation in it causes immature daughter cells at cell replication. The same study showed that 80%-85% of *BRCA1* breast cancers are triple negative and 10%-14% of TNBCs harbor *BRCA1* mutations. TNBCs span from low grade tumors to high grade tumors. Similarly, the study in [7] revealed that 1.5% of women will be diagnosed with ovarian cancer. A woman with a mutation in the BRCA1 gene has a 40–60% lifetime risk of developing ovarian cancer [8]. For BRCA2 the lifetime risk is lower at 10–30%. However, it is still around a 10-fold higher risk than for the general population.

The huge heterogeneity remains a challenge when it comes to the histological features of TNBCs and ovarian cancers. First generation prognostic signatures, such as molecular classification, invasiveness gene signatures or tumor size, are of no use for triple negative disease [9]. That is because these first-generation prognostic signatures are nothing but feature counters. They identify tumors as of prognosis on the basis of high-expression levels of proliferation-related genes. Given that 97% of TNBC and 80% of ovarian cancers have high-expression levels of proliferation-related genes [10,11], they offer no discriminatory power.

The Datasets

We have sequencing data from a Phase I clinical trial study of the Oral PI3kinase-inhibitor (NVP-BKM120) in combination with the Oral PARP-inhibitor (Olaparib) in patients with recurrent Triple Negative Breast Cancer or High Grade Serous Ovarian Cancer. Whole exome sequencing (WES) data of 35 patients from this study are available. There were a number of long-term responders to the regimen (> 6 months), raising the question as to what the genomic profile of the tumors that did respond was. RNAseq data of 8 responders and 8 non-responders of these 35 patients are also available. In summary, we have the following datasets:

- DNA sequencing data.
    - WES of 35 tumors and their match normal samples
- RNA sequencing data.
    - RNA-Seq of 16 tumors (8 responders and 8 non-responders) and 5 normal control
- Clinical data
    - Drug response of sensitive and resistant groups.

Source: Dr. Gerburg Wulf from BIDMC and Harvard Medical School provided WES and samples for RNA-Seq. RNA-Seq data have been generated by the CGI at UConn. All datasets are stored on UConn BBC HPC cluster [12].

Related Work

To understand the mechanism of cancer, biomedical researchers are interested in identifying genomic aberrations that promote cancer development, known as *drivers*. They are also interested in profiling gene expression data to understand disease pathogenesis. It is hypothesized that driver genes involved in resistance likely have aberrant copy numbers and/or mutations and that the expression patterns of these genes match the mutation and copy number patterns [13]. The genomic aberrations, incorporated with gene expression profiles in disease progression, has motivated several studies. Huang N. et. al. have shared lessons extracted from over ten years of performing integrative analysis on cancer data [14]. Several methods have been used to integrate gene expression and CNV data. In general, these methods are based on three main approaches: 1-regression, 2-correlatin, and 3-module network. Linear approaches, such as regression analysis or correlation analysis, do not work properly for heterogeneous data that their within-group variations are extremely high, such as ovarian cancer data. Further, in general, the relationship between gene expression and structural variations is not linear.

Kristensen V. et. al. surveyed the existing computational methodologies that are being used to assess cancer genomic data in [15]. In addition to the many tools surveyed, the review has assured that a more

fundamental understanding of the biological dynamics of cancer will enable us to better identify risk factors and refine cancer diagnosis. We can also predict therapeutic effects and prognosis, and identify new targets for therapy. The integrative analysis of genomic datasets introduced a paradigm shift from large randomized clinical trials towards treatment modalities that are tailored for stratified patient groups, down to N-of-1 trials, in which data from a single patient represents an entire trial.

It has been shown that the module network analysis, which is a non-linear approach, performs well in identifying driver genes in cancer [16]. In [17], Nabavi S. has employed the module network analysis upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. The module network analysis is a form of Bayesian network analysis. However, similarly behaving variables (genes) can be grouped into *modules* (a group of genes) and that the network can learn the same parents and parameters for each module, instead of each variable, as in a Bayesian network. The main motivation for using module network analysis instead of regular Bayesian network analysis is that biological systems, similar to all complex systems, have too many variables but not enough data to robustly learn networks. In biological systems, we have thousands of genes but few samples. In addition, large networks are difficult to interpret, especially in biological systems. Moreover, it is assumed that genes that are co-expressed are likely regulated in similar ways and might have the same drive. The study performed using module network analysis yielded a short list of aberrant genes that control the expression of their co-regulated genes. Although the method introduced was successful, we could not utilize it in our study since the number of samples we have is very small. In order to identify candidate biomarkers, we had to employ techniques that perform better in our situation.

Jennings E. et. al. performed an integrated analysis of heterogeneous genomic data using a hierarchical Bayesian analysis framework [18]. The framework incorporated the biological relationships among the different types of data to identify genes whose expression is related to clinical outcomes in cancer. This integrated approach lead to increased statistical power in finding these predictive genes, and further provided mechanistic information about the manner in which the gene affects the outcome. The study found 12 positive prognostic markers associated with nine genes and 13 negative prognostic markers associated with nine genes.

In [19], Wang W. et. al. proposed and implemented an integrative Bayesian analysis of genomics data (iBAG) framework for identifying important genes/biomarkers that are associated with clinical outcome. This framework also used the hierarchical modeling approach to combine the heterogeneous data obtained from multiple platforms into one model. The performance of our methods using several synthetic and real examples showed that the integrative methods have a higher power to detect disease-related genes than non-integrative methods. The methods were applied on the Cancer Genome Atlas glioblastoma dataset. The iBAG model integrated gene expression and methylation data to study their associations with patient survival. It discovered multiple methylation-regulated genes that are related to patient survival, most of which have important biological functions in other diseases but have not been previously studied in glioblastoma.

Although the aforementioned studies slightly differ in their approaches and implementation, they share a common integrative analysis characteristic. In our study, we present a data science approach toward achieving the same goal of identifying candidate biomarkers for clinical outcomes.

Chapter 2: Genomic Features

In this chapter, we cover important terminologies that we will use frequently in the subsequent chapters. We first start with basic definitions. Then, we discuss genomic features that will be the focus of our data analysis work.

Basic Definitions

The Body Cell

A cell is the smallest unit of a living organism. It is made up of many even smaller parts, each with its own function. Human cells vary in size, but all are small and cannot by seen by the naked eye. Figure 3 shows a breakdown of the human cell. The membrane holds the contents together. It also has receptors that bind to the signal molecules and communicate its presence inward into the cell. The receptors also react to substances produced in the body and to drugs taken into the body, selectively allowing these substances or drugs to trigger the cell functions. Reactions that take place at the receptors often alter or control a cell's functions. Within the cell membrane are two major components [20]:

- *The cytoplasm*: contains structures that consume and transform energy and perform the cell's functions.
- *The nucleus*: contains the cell's genetic material (DNA) and the structures that control cell division and reproduction.

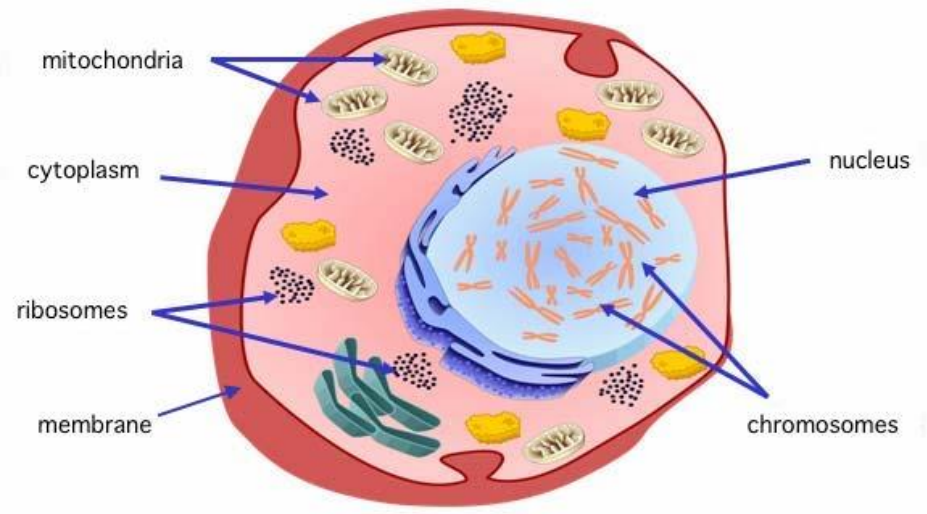


Figure 3 - Human Cell. Credit: edited from RuguSavay.com

The DNA

The DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. Every cell in a person's body has the same DNA. It is located in the cell nucleus where it is divided into chromosomes. In humans, the DNA is divided into 23 pairs of chromosomes. The information in DNA is stored as a code made up of four chemical bases: *adenine* (A), *cytosine* (C), *guanine* (G), and *thymine* (T). Human DNA consists of about 3 billion of these bases. The order, or sequence, of these bases determines the information available for building and maintaining an organism, similar to the way in

which letters of the alphabet appear in a certain order to form words and sentences. As shown in Figure 4, the DNA is a double helix formed by base pairs attached to a sugar-phosphate backbone [21].

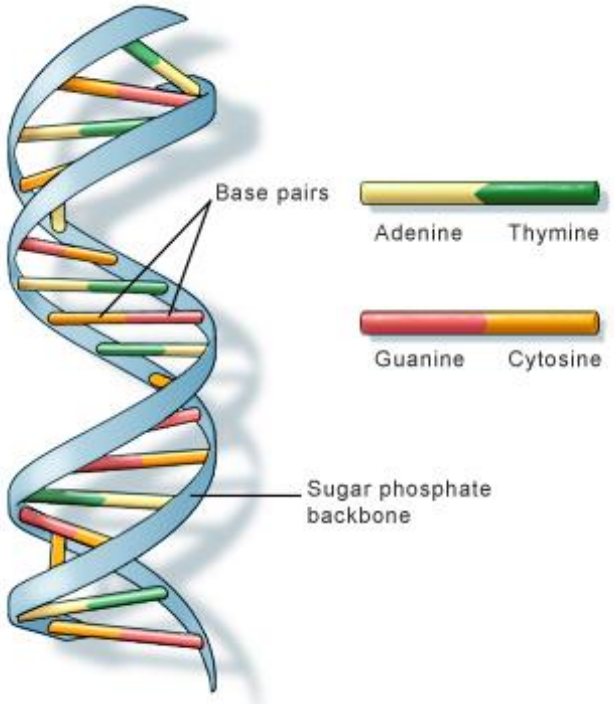


Figure 4 - DNA Structure. Credit: U.S National Library of Medicine

Genes

A gene is a region (or a set of non-overlapping regions) of the DNA that encodes for part of the cell functions. Each chromosome contains many genes. Genes act as instructions to make molecules called proteins. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. Each gene consists of coding regions (***exons***) and non-coding regions (***introns***). The Human Genome Project has estimated that humans have between 20,000 and 25,000 genes [22]. Every person has two copies of each gene, one inherited from each parent. Most genes are the same in all people, but a small number of genes (less than 1 percent of the total) are slightly different between people. These differences contribute to each person's unique physical features.

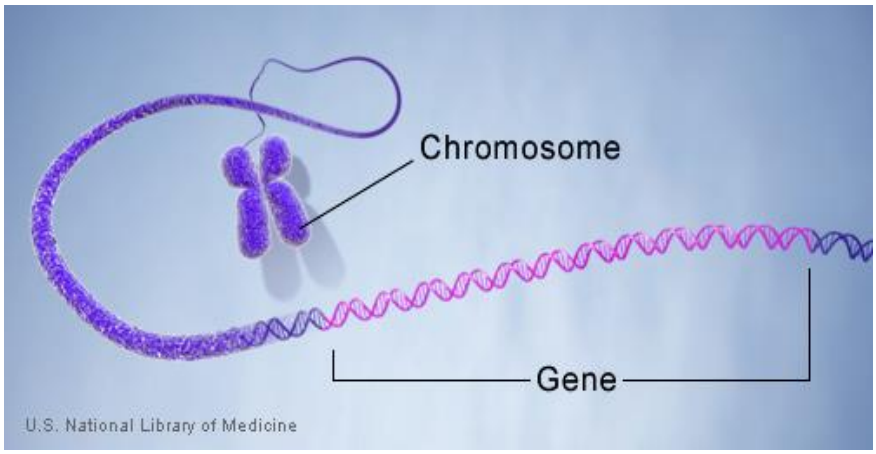


Figure 5 - Genes. Credit: U.S. National Library of Medicine

Gene Expression

The DNA nucleotides remain

inside the cell's nucleus. Some of them (*genes*) get copied to the cytoplasm in a process called *gene expression*. The ultimate product of gene expression are functional proteins, which can go on to perform essential functions as enzymes, hormones and receptors.

Gene expression starts with the *transcription* process, which copies gene regions from inside the nucleus to the outside. Transcription starts from a promoter region and ends at a stop codon site. It results in a pre-messenger RNA (ribonucleic acid). The RNA is represented the same as DNA with letters: A, C, G and U (uracil, instead of T). The next step involves removing introns from the copied regions in a process known as *splicing*. It results in a mature mRNA. After that, the mature mRNA is translated into functional proteins.

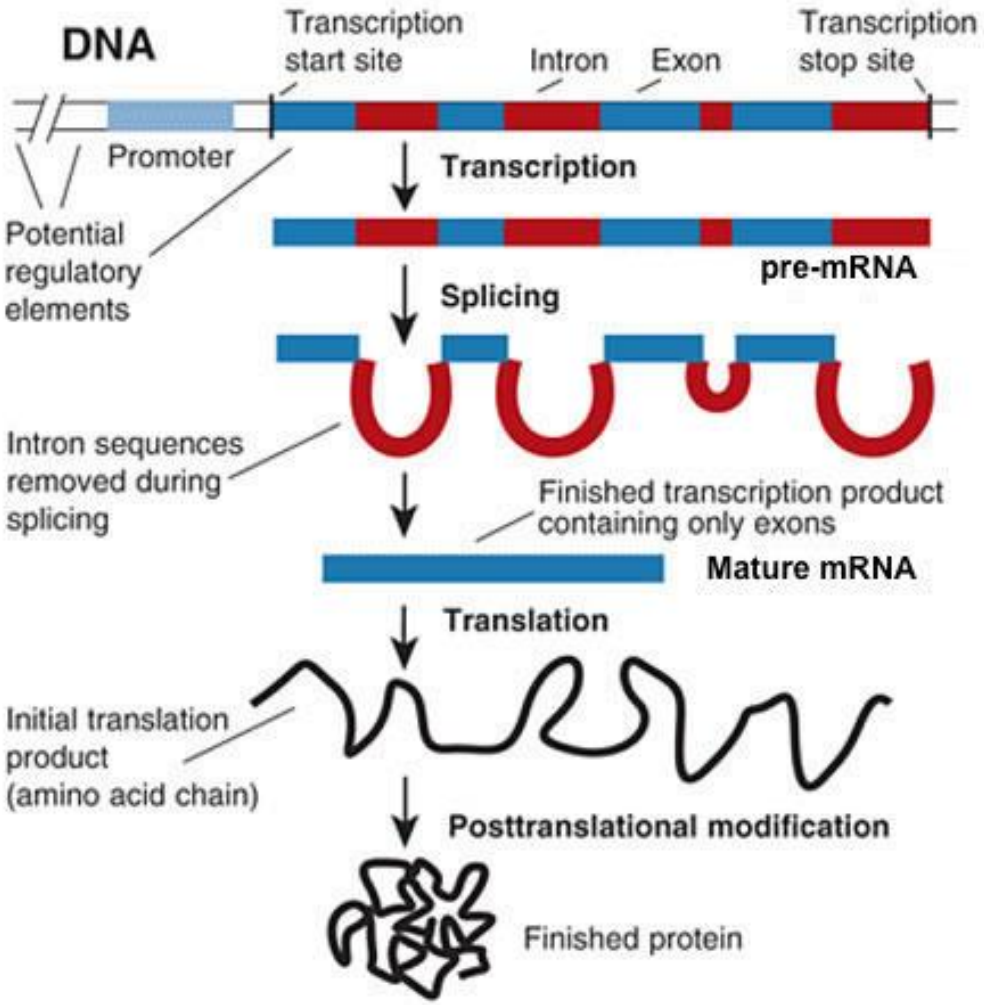


Figure 6 - Gene Expression. Credit: National Institute of Health

Proteins

Proteins are the ultimate product of the gene expression process. They do most of the work in cells and are required for the structure, function, and regulation of the body's tissues and organs. Proteins are made up of hundreds or thousands of smaller units called amino acids, which are attached to one another in long chains. There are 20 different types of amino acids that can be combined to make a protein. The sequence of amino acids determines each protein's unique 3-dimensional structure and its specific function. Table 1 shows the genetic code: rules by which information encoded by genetic material (DNA and RNA) is translated into amino acids (identified from 3-mer sequences) that combined form a protein.

Table 1 – Genetic Code

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC: Phe UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr **UAA: STOP** **UAG: STOP**/Pyl	UGU, UGC: Cys **UGA: STOP**/Sec UGG: Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile **AUG**: Met	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn AAA, AAG: Lys	AGU, AGC: Ser AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp GAA, GAG: Lys	GGU, GGC, GGA, GGG: Gly	U C A G

Gene Fusion

Due to abnormal cellular activities, two genes might fuse together at the transcription process forming a new gene as shown in Figure 7. This results a unique protein product at the end that might lead the cell to function improperly.

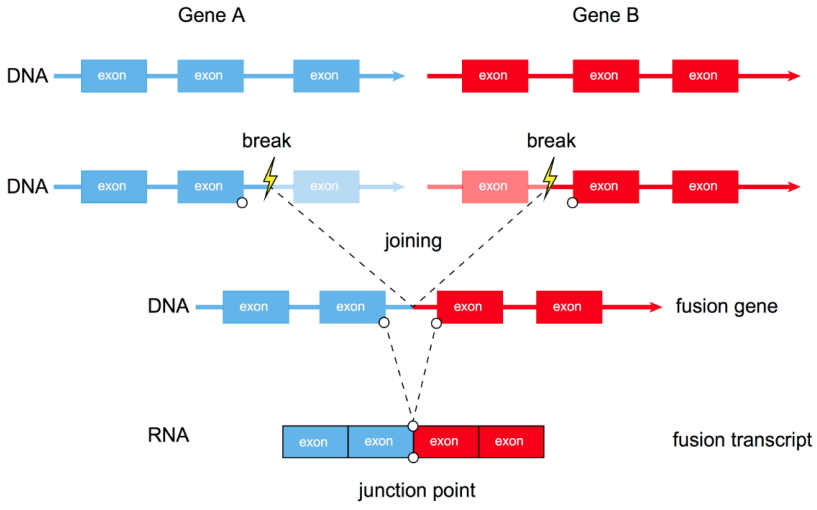


Figure 7 - Gene Fusion. Credit: EMC Galaxy Training

Mutations

Single Nucleotide Polymorphism (SNP)

SNPs are variations in a single base pair in a DNA sequence as shown in Figure 8(a).

Indel

Indels are the insertion or the deletion of bases in the DNA of an organism as shown in Figure 8(b).

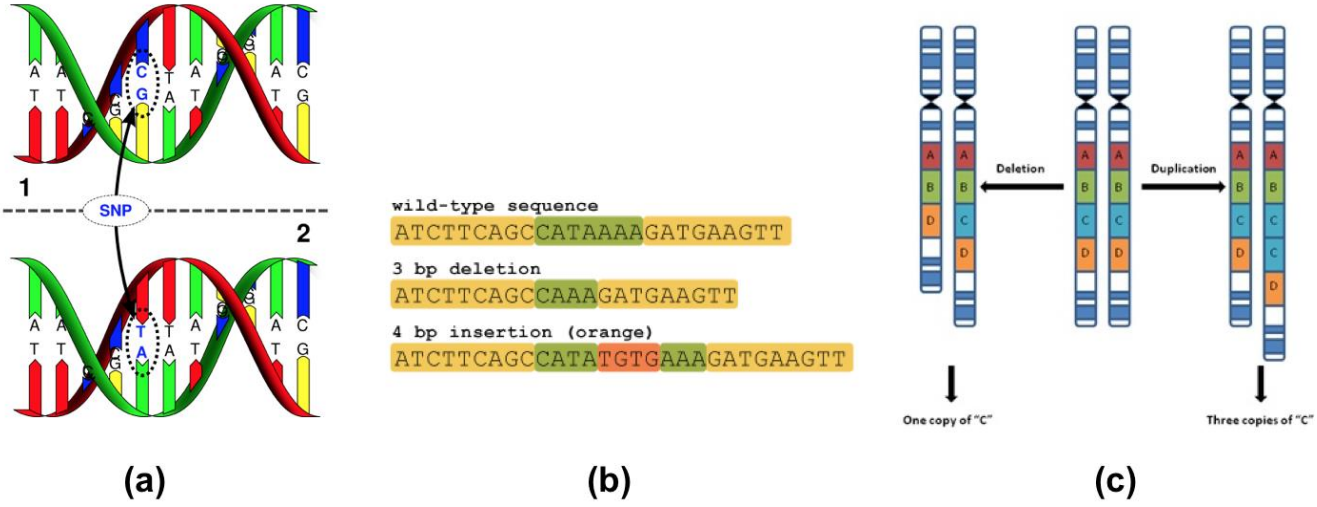


Figure 8 - Structural Variations. (a) SNP. Credit: International Society of Genetic Genealogy (b) Indels. Credit: Hackbright Academy. (c) CNVs. Credit: MindSpec.org

Synonymous vs. Non-synonymous Variations

In fact, not all mutations in the DNA are harmful and cause abnormal activity in the cell. Mutations can occur in non-coding regions which are not transcribed. Also, as we saw in the protein table previously, some sequences result in the same protein. Therefore, some variations are synonymous and result in the same ultimate protein sequence. In our study, we are interested only in non-synonymous variations. Figure 9shows a classification of variations.

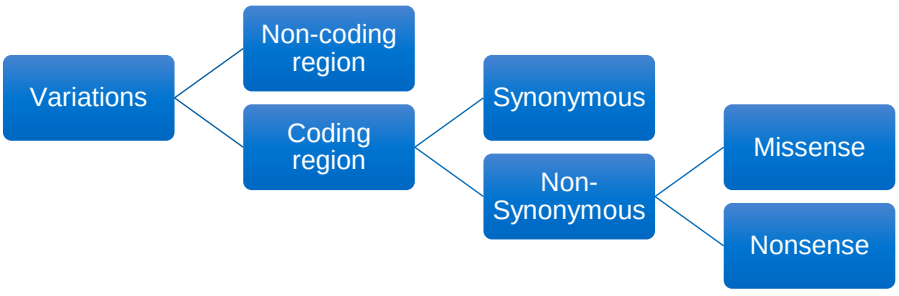


Figure 9 – Types of variations

Non-synonymous variations of the type *Missense* results in a change in the protein as in the following example.

```
DNA:  5' - AAC AGC CTG CGT ACG GCT CTC - 3'
mRNA: 5' - AAC AGC CUG CGU ACG GCU CUC - 3'
Protein:   Asn Ser Leu Arg Thr Ala Leu
```

- A mutation:

```
DNA:  5' - AAC AGC CTG CTT ACG GCT CTC - 3'
mRNA: 5' - AAC AGC CUG CUU ACG GCU CUC - 3'
Protein:   Asn Ser Leu Leu Thr Ala Leu
```

Non-synonymous variations of type *Nonsense* causes the transcription to in an earlier position than the mature stop position. The following is an example.

```
DNA:  5' - ATG ACT CAC CGA GCG CGA AGC TGA - 3'
mRNA: 5' - AUG ACU CAC CGA GCG CGA AGC UGA - 3'
Protein:   Met Thr His Arg Ala Arg Ser Stop
```

- A mutation:

```
DNA:  5' - ATG ACT CAC TGA GCG CGA AGC TGA - 3'
mRNA: 5' - AUG ACU CAC UGA GCG CGA AGC UGA - 3'
Protein:   Met Thr His Stop
```

Structural Variations

Some diseases, such as cancer, are caused by structural variations in the DNA [1]. Structural variations describe changes in the DNA sequence that might lead to a change in the proteins resulting from the gene expression process. We present the type of variations that we focused on during our study.

Copy Number Variation (CNV)

CNVs are a phenomenon in which parts of the genes are repeated or deleted. The number of repetition/deletion in the DNA varies between individuals in the human population. Figure 8(c) shows an example of a deletion on the left and an example of a duplication on the right.

Genetics of Cancer

Cancer is a genetic disease in most cases. It is caused by the abnormal cell division. Cells divide and grow in the presence of signals that normally inhibit abnormal cell growth [1]. As these cells grow, they develop new characteristics, including changes in DNA structure. Such changes allow the cancer cells to spread and invade other tissues, carrying with them their undesired characteristics. The abnormalities in cancer cells usually result from mutations in protein-encoding genes that regulate cell division. Over time more genes become mutated. This is mostly because the genes that make the proteins that normally repair DNA damage are themselves not functioning normally because they are also mutated. Consequently, mutations begin to increase in the cell, causing further abnormalities in that cell and the

daughter cells [1]. In cancer, somatic evolution is the accumulation of mutations in the cells of a body during a lifetime, and the effects of those mutations on the fitness of those cells. Somatic mutations can occur in any of the cells of the body except the germ cells (sperm and egg) and therefore are not passed on to children. These alterations can (but do not always) cause cancer or other diseases.

Recall from chapter 1 that our clinical trial investigates the drug response of a the Oral PI3kinase-inhibitor (NVP-BKM120) in combination with the Oral PARP-inhibitor (Olaparib) in patients with recurrent TNBC and ovarian cancer. In the next chapter, we discuss the data analysis pipelines we have developed for the whole exome sequencing (WES) and RNA Sequencing.

Chapter 3: Developing Pipelines for Analyzing High-throughput Data

In this chapter, we present the development stages of our data analysis pipeline. Recall that we are given whole exome sequencing data for 35 patients and RNA sequencing data for a subset of 16 patients. The next sections illustrate the pipelines in details.

Whole Exome Sequencing

Whole exome sequencing is a technique used for sequencing all the exonic regions of a genome. It starts by capturing only the regions of the DNA that encodes for proteins (i.e. exons). After that, instruments such as Illumina HiSeq are used to generate raw data files representing the DNA fragments (also known as short reads) [23]. Figure 10 shows the sequencing stages. For each patient, we obtained a sample from a normal tissue and another sample from the tumor tissue, which is the breast in our study. Then, using DNA templates in a flow cell, exons are extracted from the DNA. After that, a high-throughput sequencer is used to read and interpret the short fragments into representations of the base pairs: *A*, *C*, *G* and *T*.

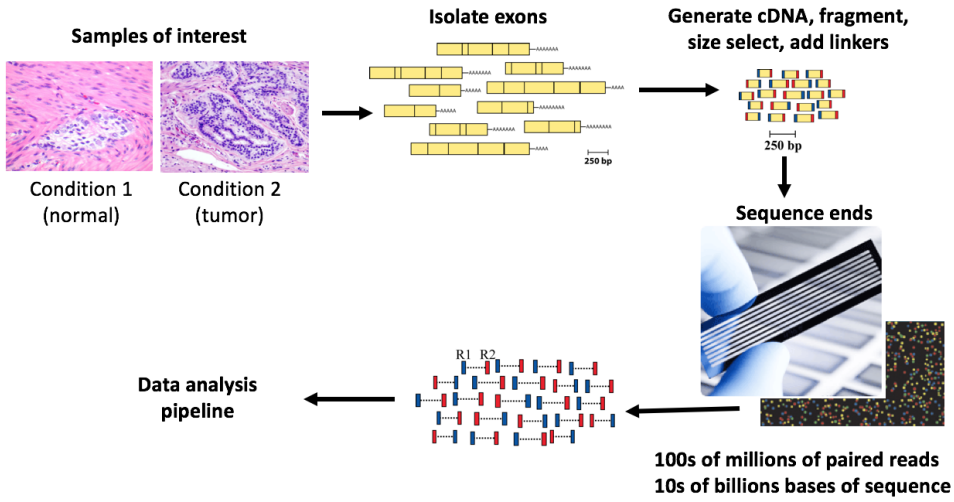


Figure 10 - Whole Exome Sequencing. Credit: Canadian Bioinformatics Workshops

Our study starts from the data analysis pipeline. We are given raw sequencing reads. Figure 11 shows the steps we have done in the analysis, followed by a detailed illustration of each step. The first two phases are considered as input to the pipeline. They have been performed under the supervision of Dr. Gerburg Wulf from BIDMC and Harvard Medical School.

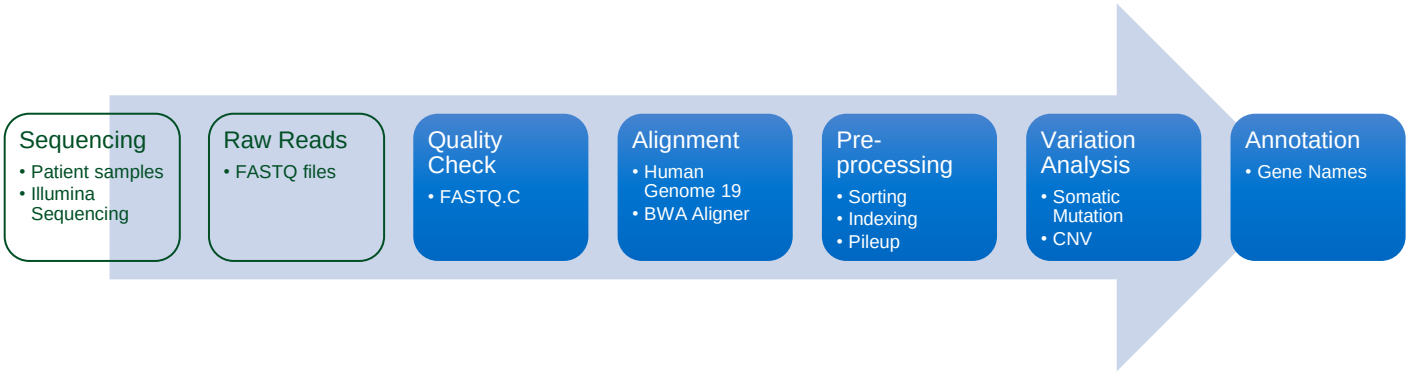


Figure 11 – Whole Exome Sequencing (WES) pipeline

Raw Reads

Data from sequencing comes in raw text files in FASTQ format [24]. A FASTQ file normally uses four lines per sequence. Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description. Line 2 is the raw sequence letters, which came in lengths of 100 base pairs in our study. Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again. Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence. For example, A FASTQ file containing a single sequence looks like this:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

The quality values are encoded using Phred quality scores. Phred quality scores are assigned to each nucleotide base call in automated sequencer traces. They have become widely accepted to characterize the quality of DNA sequences [24]. The quality scores become very important in the next step of the quality check.

For each patient, we have received two sets of FASTQ files: files for normal tissue and files for the tumor tissue. The total size of all FASTQ files for all 35 patients is **~700 GB** in a *gzip* compressed format. For the next pipeline phases, Linux scripts are provided for each task on the project's GitHub repository [25].

Quality Check

The current sequencing technologies produce errors in the read files. That's why the quality score line is included in the FASTQ file for each read. In order to guarantee reliable results by the end of the analysis, we checked the quality of all FASTQ files using FASTQC *v0.11.2* [26]. The summary reports showed no potential errors or warnings. Figure 12 shows an example of the report generated by the tool for one of the patients. Complete quality reports for all patients can be found on the project's GitHub repository under variation-analysis folder.

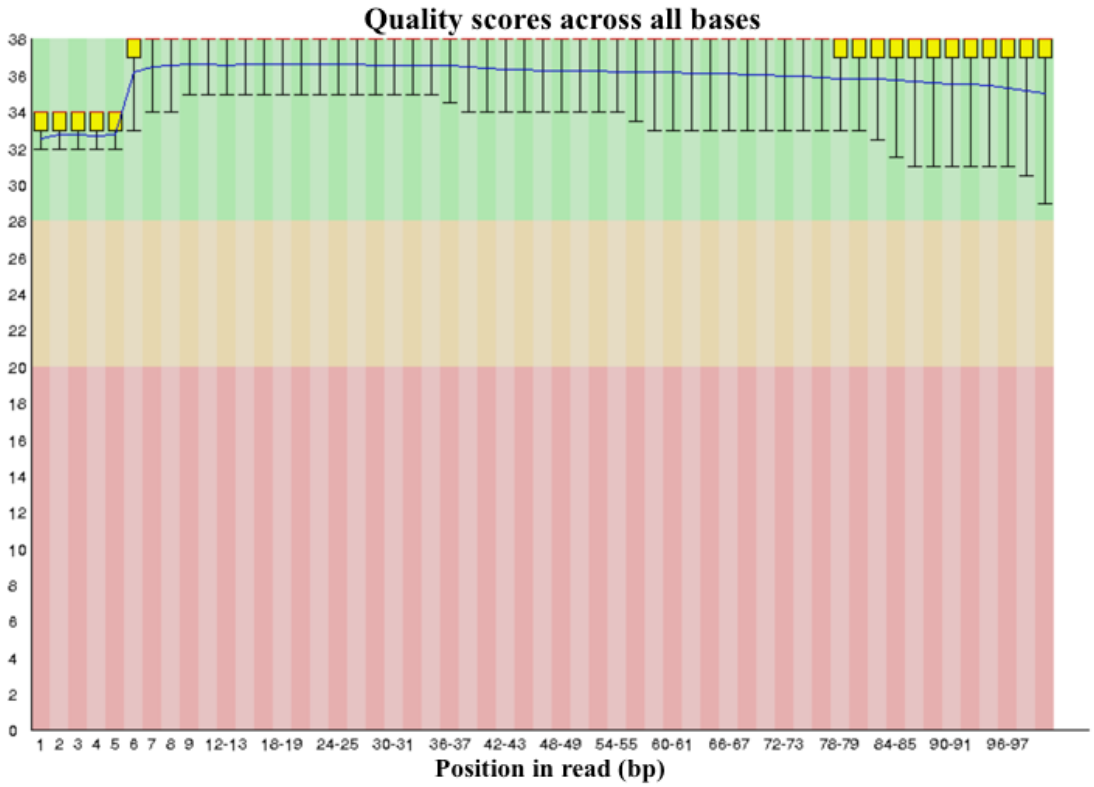


Figure 12 - Per base sequence quality for one of the samples

Alignment to the human genome

Since the short reads are not in a specific order in the file, the next step in the pipeline is to map them to the human genome. Figure 13 visualizes the reference genome in the top track and the short read mapping below it. The blue lines connect the pair-end reads. Red bars are the short reads. For the alignment, we have used the Burrows-Wheeler Aligner tool *v0.7.12-r1039* [27] and mapped the reads to the human genome reference (Hg19) downloaded from UCSC genome browser [28].

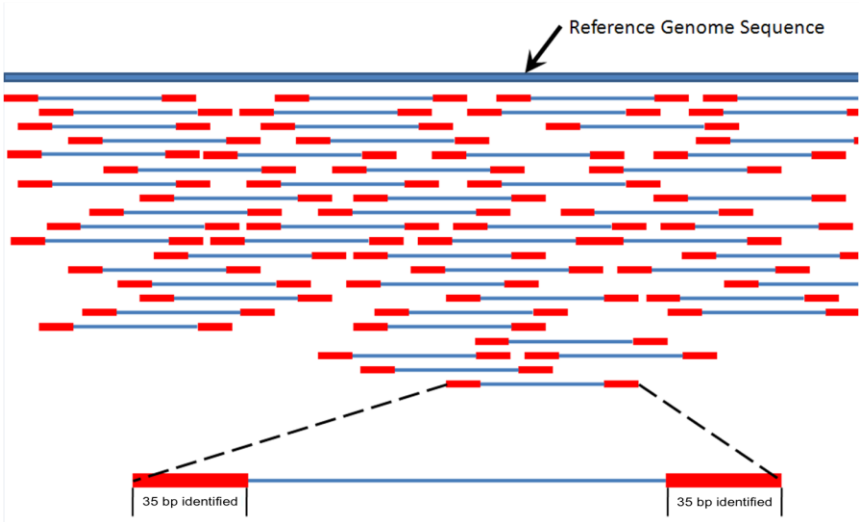


Figure 13 - Short read alignment to the human genome. Credit: Wikipedia

Aligning each patient's samples separately, we obtain two alignment files: one for normal and one for tumor. The alignment files are in SAM format (Sequence Alignment/Map) [29]. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to the alignments. Header lines start with '@', while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information. The total size of all SAM files for all 35 patients is **~3.5 TB**.

Preprocessing

Before starting to analyze the alignment files for variation analysis, it is of great benefit to pre-process the data for future operations to take less time. Preprocessing includes sorting and indexing the aligned reads and converting the results to fixed binary files for faster analysis. For this task, we have used Picard v2.3.0 included in the GATK framework [30] following the best practices described in [31] and [32]. This reduced the size of our files to **~2.6 TB** of processed alignment files along with indices.

Somatic Variation Analysis

In this step of the pipeline, we have analyzed the processed files in order to identify somatic variations. To achieve this, we perform a per-patient analysis comparing the patient's normal file with the tumor file. SNPs as well as Indels. In this analysis, variations are identified for each patient with relative to her normal sample (not with the reference human genome). For this phase, we have used VarScan2 *v2.4.2* [33] with some intermediate steps done using SAMtools *v1.2* and htslib *v1.2.1* [29].

This phase results in the variations for each patient separately in VCF format (Variant Call Format) [34]. The VCF format stores sequence variations. For example:

```
## header line 1
## header line 2
## rest of header lines
#CHROM  POS  ID  REF  ALT  QUAL  FILTER  INFO  FORMAT  Sample1  Sample2
2       4370 rs9 G    A    29    .       NS=.. DP:HQ.. 0|0:48.. 0|0:46..
:
:
```

Table 2 gives a brief description for each column:

Table 2 - Columns of the VCF file

	Name	Brief Description
1	CHROM	The name of the sequence (a chromosome) on which the variation is being called.
2	POS	The 1-based position of the variation on the given sequence.
3	ID	The identifier of the variation
4	REF	The reference base (or bases in the case of an Indel at the given position on the given reference sequence.
5	ALT	The list of alternative alleles at this position.
6	QUAL	A quality score associated with the inference of the given alleles.
7	FILTER	A flag indicating which of a given set of filters the variation has passed.

8	INFO	An extensible list of key-value pairs (fields) describing the variation.
9	FORMAT	An (optional) extensible list of fields for describing the samples.
10	Samples	(optional) For each sample, values are given for the fields listed in FORMAT.

Since we do not expect a large number of variations in each patient, VCF files are small files in the order of a few Megabytes. They serve one of our core analysis assets in the next chapter.

Copy Number Variation Analysis

As with the somatic variation analysis, we have analyzed the processed files in order to identify CNVs. This is also a per-patient analysis where we used VarScan2 as well. This step results in the CNVs for each patient separately in a TAB-delimited format. It stores information about the copy numbers.

For example:

```
chrom   start   stop   num_positions   normal_depth   tumor_depth   log2_ratio   gc_content
chr1    1489    1588   100             24.1           10.6          -1.182       41.0
:
:
```

Table 3 gives a brief description for each column.

Table 3 - Columns of the CNV output file

	Name	Brief Description
1	chrom	Chromosome or reference name.
2	start	Start position of a contiguous copy number region (1-based).
3	stop	Stop position of a contiguous copy number region (1-based).
4	num_positions	Length of the region.
5	normal_depth	Average sequence depth in the normal.
6	tumor_depth	Average sequence depth in the tumor.
7	log2_ratio	Log-base-2 ratio of the tumor/normal depth ratio.
8	gc_content	Proportion of GC bases in the region, between 0 and 100.

After that, we have processed the copy number data as following:

- Filter the copy number calls by minimum coverage and/or region size.
- Adjust raw copy number (log2) values for GC content.
- Classify each region as amplification (gain), deletion (loss), or neutral based on your preferred log2 thresholds.
- Re-center raw copy number data if neutral segments are not on the log2=0 axis.

VarScan2 provides an implementation for this filtering/adjustment. Therefore, we have integrated it in our pipeline. The file maps the repeated/deleted segments of the DNA to the corresponding chromosomes and marks the start/end positions along with the log ratio of the copy number value. The raw output from that phase of the pipeline is then smoothed and segmented using DNAcopy package in R [35]. The

package analyses the data to detect larger regions with abnormal copy numbers. Another filtering that we have applied is removing the copy number regions that have a log ratio in the range of $[-0.5, 0.5]$. This ensures that further analysis relies only on significant copy number values that denote a true duplication or deletion. After that, we map the CNV regions to the list of known genes in RefSeq database [36] with the goal of identifying the CNV genes.

Gene Annotation

The last step in this pipeline is to annotate the discovered mutations (SNPs and Indels) with regard to genes and proteins. The idea is to define what aberrations fall into what exons of the known genes and can change which protein. The obtain a comprehensive list of known genes along with their start/end positions, we have used RefSeq database [36] which we obtained from UCSC genome browser. We have used ANNOVAR *v2016-02-01* to perform the annotation step [37].

By the end of the first data analysis pipeline, we have obtained the following for each patient:

1. A list of genes that have somatic mutations.
2. A list of genes that have CNVs.

RNA Sequencing

Similar to WES, RNA-Seq aims at obtaining short reads for the genomic features. However, instead of capturing base pairs from the DNA samples, it operates on the RNA samples and quantifies presence of mRNA at a given moment in time [38]. RNA-Seq facilitates the study of changes at the gene expression in a given sample. Therefore, it is a tissue- and time-specific. Similar to the sequencing stages illustrated in Figure 10, RNA-Seq follows the same stages except that we isolate RNAs instead of the exons. We also study the gene expression only in the tumor tissue and compare it with the gene expression from the same tissue in normal population.

Figure 14 of the downstream data analysis is similar to the analysis we did in the WES pipeline. However, there is a number of differences due to the special nature of the RNA sequencing process. We explain them in the next subsections.

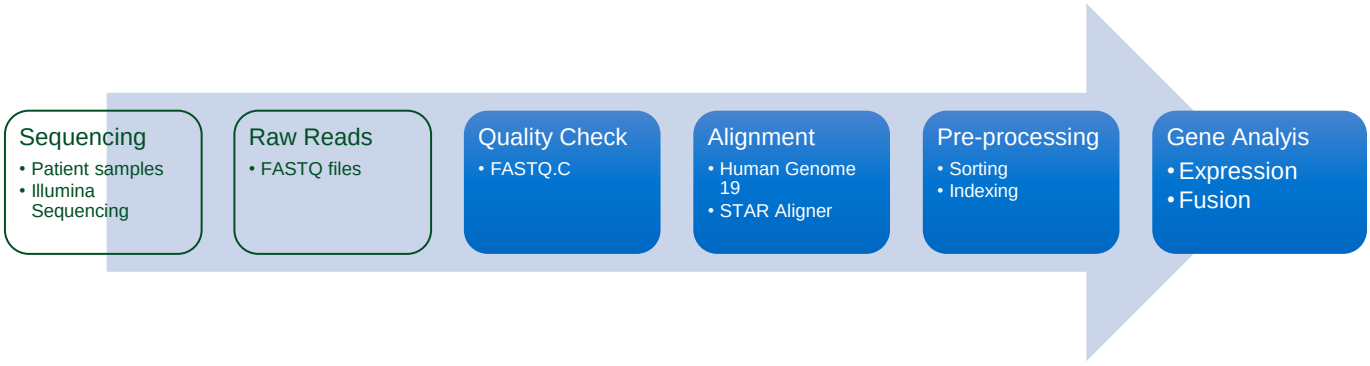


Figure 14 – RNA-Seq pipeline

Raw Reads and Quality Check

Similar to WES, data comes in raw text files in FASTQ format. For each patient, we have received a set of FASTQ files for sequenced tissue. Sequencing of 8 patients (out of 16) were carried out from the breast tissue. Sequencing for the other 8 patients were performed from the ovarian tissue. The reason for that is that breast and ovarian cancers share common features. In addition, we have sequencing data for 5 normal samples from the breast tissue of healthy women. The total size of all FASTQ files for all 21 patients is **~105 GB** in a *gzip* compressed format. For the next pipeline phases, Linux scripts are provided for each task on the project's GitHub repository [25].

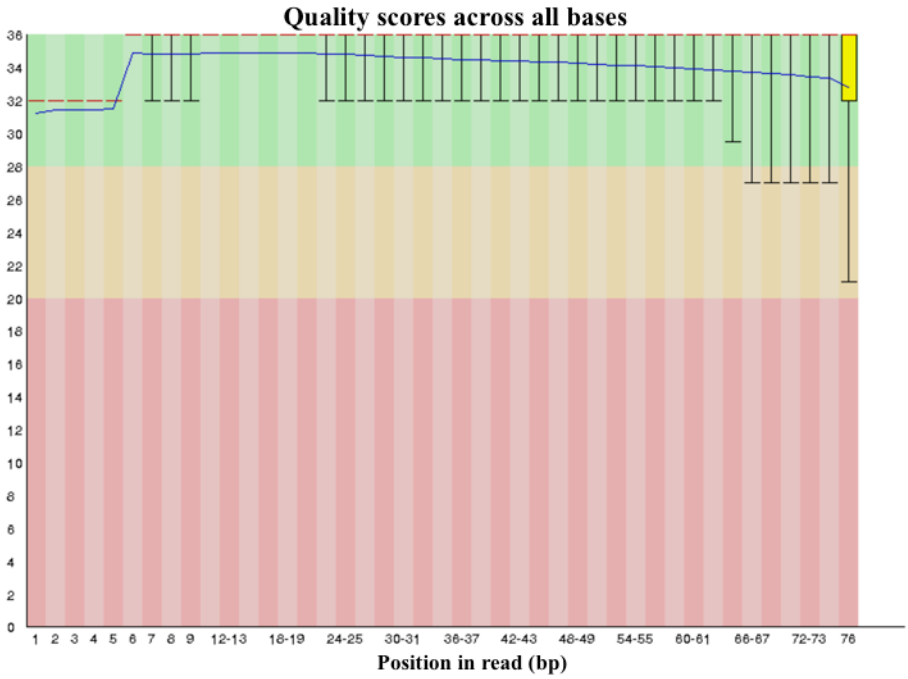


Figure 15 - Per base sequence quality for one RNA-Seq sample

The quality check summary reports showed no potential errors or warnings. Figure 15 shows an example of the report generated by FASTQC for one of the RNA-Seq samples. Complete quality reports for all patients can be found on the project's GitHub repository under gene-analysis folder.

Alignment and Pre-processing

The next step in the pipeline is to map the short reads to the human genome. There is a crucial difference from the WES reads alignment. Since RNA-Seq isolates RNAs instead of exons, generated reads come from different exons regions. In other words, one read might not be mapped as one fragment to the human genome; parts of the read may come from different exonic regions. Therefore, we have to use an aligner that is aware of the splices. Therefore, we have used the STAR Aligner tool *v020201* [39] and mapped the reads to the human genome reference (Hg19) downloaded from UCSC genome browser [28]. Figure 16 visualizes the spliced read alignment in RNA-Seq data.

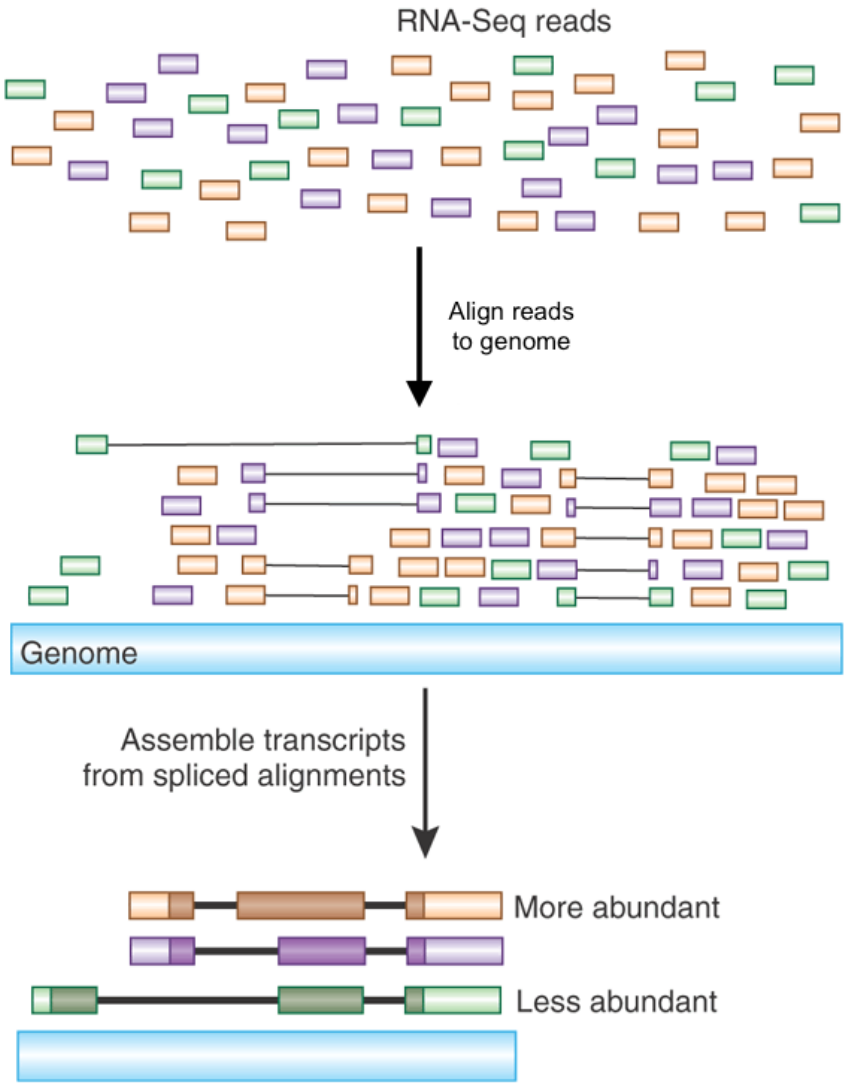


Figure 16 - Spliced read alignment to the human genome. Credit: edited from [40]

We have obtained the alignment file for each sample. The alignment files are in SAM format and the total size of all SAM files for all 21 samples is **~300 GB**. We have proceeded with the pre-processing of the alignment data to have a faster analysis in the subsequent pipeline steps.

Gene Expression Analysis

Gene expression is the most basic level at which the genotype gives rise to the phenotype. In other words, gene expression is affected by the genomic features at the DNA level and affects the observable results such as the drug response in our clinical trial. The goal of this phase is to quantify the expression levels of known genes in each of the given samples. To achieve this, we perform a per-sample analysis calculating the number of reads that lie in known gene regions. The more reads we find in a known gene, the more expression level this gene has. A number of normalization and statistical testing steps are

performed afterwards to remove biases. Fortunately, existing software packages carry out this task efficiently. We have used RSEM *v1.2.31* [41] with Ensembl gene annotation database [42].

This phase results in the gene expression level for each sample separately in TAB-delimited format. It stores information about the read counts for each gene. For example:

```
gene_id  transcript_id(s)  length  effective_length  expected_count  TPM  FPKM
ENSG003  ENST049,ENST0731  2201.90 2096.31           1803.00         9.67 20.07
:
:
```

Table 4 gives a brief description for each column:

Table 4 - Columns of the gene expression output file

	Name	Brief Description
1	gene_id	The Ensembl gene ID of a known gene.
2	transcript_id(s)	The Ensembl transcript ID (or IDs) corresponding to the gene ID above.
3	length	The transcript's sequence length.
4	effective_length	It counts only the positions that can generate a valid fragment.
5	expected_count[1]	The sum of the posterior probability of each read comes from this transcript over all reads.
6	TPM	It stands for Transcripts Per Million. It is a relative measure of transcript abundance. The sum of all transcripts' TPM is 1 million.
7	FPKM[2]	It stands for Fragments Per Kilobase of transcript per Million mapped reads. It is another relative measure of transcript abundance.

[1] Each read aligning to a transcript has a probability of being generated from background noise, that's why it takes into account the probabilities. RSEM may filter some aligned low quality reads, the sum of expected counts for all transcript are generally less than the total number of reads aligned.

[2] FPKM is calculated as following:

Let $\bar{l}$ be the mean transcript length in a sample, $\bar{l} = \sum_i TPM_i \frac{effective\_length_i}{10^6}$, where i is the transcript number. We get, $FPKM_i = \frac{10^3}{\bar{l}} \times TPM_i$

Gene Fusion Analysis

Recall from the previous chapter that two genes (or part) might fuse together forming a new gene that ends with an unknown transcript. Therefore, the gene expression analysis cannot capture the gene fusion phenomenon, since the new transcript maps to two different (and maybe distant) genes. This is different from spliced alignment that takes into account different exons of a single gene. Fusion occurs mostly due to genomic aberrations.

In this step of the pipeline, we have used Fusion Catcher *v0.99.6a beta* [43]. It searches for novel as well as known somatic fusion genes, translocations, and chimeras in RNA-Seq data from disease samples. translocation is a chromosome abnormality caused by rearrangement of parts between chromosomes.

The tool shows very good detection rate for finding candidate somatic fusion genes. It stores information about the candidate fused transcripts for each sample in a TAB-delimited file. Some of the columns are shown below:

```
gene_1  gene_2  fusion_description  spanning_pairs  fusion_sequence fused_transcript
ATAD2   NPM1    cancer_tissues     5               ..CTGA*GTCA..   ..TGAAG*ATGATG..
:
:
```

Table 5 shows some of the columns of the Fusion Catcher output. The original full list of columns can be found on the tool documentation files.

Table 5 - Columns of the gene fusion output file

	Name	Brief Description
1	gene_1	Gene symbol of the 5' end fusion partner.
2	gene_2	Gene symbol of the 3' end fusion partner.
3	fusion_description	Type of the fused gene referring to known database if it is a known fusion or uniquely identified.
4	spanning_pairs	Count of pair-end reads supporting the fusion.
5	fusion_sequence	The inferred fusion junction (the asterisk sign marks the junction point).
6	fused_transcript	All possible known fused transcripts

Summary

In this chapter, we have covered the development stages of the data analysis pipelines. We have developed and executed the necessary stages of obtaining genomic variation data as well as the gene profiling data for the patients covered in our study. We have utilized a number of existing software tools in different stages of the pipeline with the goal of optimizing the pipeline execution and obtaining reliable results.

In the next chapter, we proceed to the next step of the study in which we mine the output files from the developed pipelines to perform further analysis.

Chapter 4: Data Analysis

So far, we have been able to transfer the raw sequencing data into meaningful per-patient genomic features. We have obtained the somatic variation (SNP and Indel), CNV, gene expression and gene fusion data. In addition, we know that patients under study are treated with Oral PI3kinase-inhibitor (NVP-BKM120) in combination with the Oral PARP-inhibitor (Olaparib). Half of them responded to the treatment and the other half did not respond. We consider the availability of WES and RNA-Seq data for those patients. In this part of our study, we analyze the patients' data obtained from the pipelines illustrated in the previous chapter. The goal is to try answering the following question:

Can we identify candidate genomic biomarkers that drive drug resistance?

In other words, can we find aberrations in some genes of the non-responding patients that do not exist in the responders? The nature of this question implies the following challenges:

- Data Heterogeneity: as we have seen from the previous chapter, the datasets have different formats and represent different biological meanings.
- Feature selection: with the large number of features (genes) under study, we should use appropriate feature selection or dimensionality reduction methods to identify only the features that are biologically meaningful.
- Small sample size: the number of samples is relatively small. This imposes a challenge in applying machine learning and statistical methods.

Approach

The integrative analysis of genomic datasets comes in different flavors. We discuss the different approaches in the next chapter. In our study, we have followed a *data science* approach. The motivation is to extract knowledge from data. We aim at applying a wide range of techniques that guide our search for a smaller feature list.

The hypothesis we followed in shown in Figure 17. The hypothesis is that the drug response is most likely the result of the functional proteins produced by the tissue cells along with some other unknown factor θ_1. As we have shown in chapter 2, proteins are the ultimate results of gene expression. Therefore, they are affected by the transcribed mRNAs that represent genes. Fused genes also play a non-negligible role since they might lead to producing new proteins. Besides, the unknown factor θ_2 affects the functional proteins. Gene expression and gene fusion are a result of the DNA features. Therefore, aberrations in the DNA are considered the root cause of the drug response, without neglecting the third unknown factor θ_3.

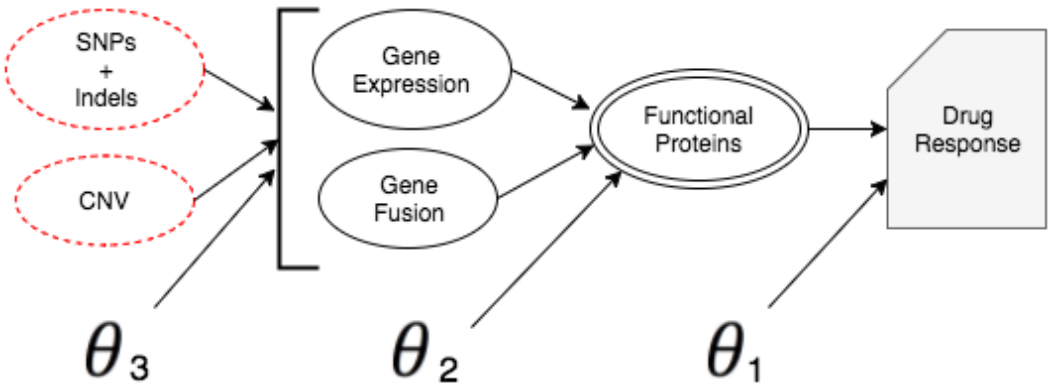


Figure 17 - Data Analysis Approach

Why did we include the unknown factors θs? In general, the living organisms have a very complex system of operation that cannot be modeled in a computation. Thus, being able to model many cellular processes such as the gene expression and aberrations does not mean that we have modeled all cellular processes. Without the loss of generality, Figure 17 captures all the data available in our study. In the next sections, we explain our data mining and analysis steps and drive conclusions afterwards.

Data Crunching

In the previous chapter, we have seen how heterogeneous the data resulted from the pipelines is. To perform any further analysis, we have to pre-process the data and construct meaningful data structures. Therefore, we have formulated four matrices; one for each of the genomic features. We present the process of obtaining them in the next subsections.

Gene Expression Matrix

The RNA-Seq data is tissue-specific. The data for the patients were sequenced from the breast for a number of patients and the ovarian from the rest of them. The belief is that breast cancer and ovarian cancer share many genotype features. Therefore, by mining the gene expression file for each patient, we have obtained two separate matrices for each type of RNA sequencing: $M1_{breast}$ for the breast samples and $M1_{ovarian}$ for the ovarian samples. We construct the matrix as following:

$$M1 = \begin{bmatrix} & p_1 & p_2 & \cdots & p_n \\ g_1 & & & & \\ g_2 & & & & \\ \vdots & \vdots & & \ddots & \vdots \\ g_k & & & & \\ r & & & \cdots & \end{bmatrix}$$

Where g stands for the gene, p stands for the patient and r stands for the drug response. The value at the r row is 1 if the patient responded to the drug and 0 if she belongs to the non-responders group. The values in each cell of the matrix represents the gene expression value as either *normalized read count* or *transcript per million (TPM)*.

***Normalized Read Count* Matrix**

The value in each one of these cells represents the normalized read count of gene g_i for patient p_j for all $i \in [1, 2, \ldots, k]$ and $j \in [1, 2, \ldots, n]$, where k is the number of genes and n is the number of samples.

***Transcript Per Million (TPM)* Matrix**

The value in each one of these cells represents the TPM value of gene g_i for patient p_j for all $i \in [1, 2, \ldots, k]$ and $j \in [1, 2, \ldots, n]$. The TPM is a metric used to normalize for sequencing depth and gene length. The software tool obtains it using the following operations:

1. Divide the read counts by the length of each gene in kilobases. This gives us reads per kilobase (RPK).
2. Count up all the RPK values in a sample and divide this number by $1{,}000{,}000$. This is the "per million" scaling factor.
3. Divide the RPK values by the "per million" scaling factor. This gives the TPM.

Gene Fusion Matrix

From the RNA-Seq data, we have obtained the genes that showed fused transcripts for each patient. By mining the gene fusion file for each patient, we construct the matrix $M2$ in the same way we constructed $M1$ as following:

$$M2 = \begin{bmatrix} & p_1 & p_2 & \cdots & p_n \\ g_1 & & & & \\ g_2 & & & & \\ \vdots & \vdots & & \ddots & \vdots \\ g_k & & & & \\ r & & & \cdots & \end{bmatrix}$$

The value in each one of the cells = $\begin{cases} 1 & if\ the\ gene\ is\ fused \\ 0 & otherwise \end{cases}$ for each gene g_i and patient p_j, where $i \in [1, 2, \ldots, k]$ and $j \in [1, 2, \ldots, n]$. We can notice that this matrix is very sparse containing a very large number of zeros, since each patient has a very few number of fused genes (less than 10).

Somatic Variation Matrix

For each patient, we have obtained the annotated somatic variation file. The file maps mutations to the corresponding genes from the RefSeq database [36]. We are only interested in variations that are nonsynonymous. Therefore, we have filtered each patient's file to opt out synonymous variations. Similar to $M1$ and $M2$, we construct the following matrix:

$$M3 = \begin{bmatrix} & p_1 & p_2 & \cdots & p_n \\ g_1 & & & & \\ g_2 & & & & \\ \vdots & \vdots & & \ddots & \vdots \\ g_k & & & & \\ r & & & \cdots & \end{bmatrix}$$

The value in each one of these cells is 1 if there is a nonsynonymous variation in gene g_i for patient p_j and 0 otherwise for all $i \in [1, 2, \ldots, k]$ and $j \in [1, 2, \ldots, n]$. This matrix also has a high level of sparsity.

CNV Matrix

For each patient, we have obtained the annotated CNV file. Although the filtering done in the pipeline seems firm, it resulted into thousands of CNV genes per patient. Of course, further analysis would fail if we keep this number as it is for the reason that the number of features will be too large to have a significant contribution to the drug response.

One further filtering step we have implemented is integrating the TPM value of a given gene for each patient with the CNV values. Figure 18 illustrates this filtering step. For each patient, we sample the values of genes TPM along the x axis. We will then investigate only the genes in the first and last quartile. We include the gene in our analysis only if:

- The gene lies in the first quartile and has a negative log ratio of the copy number, or
- The gene lies in the last quartile and has a positive log ratio of the copy number.

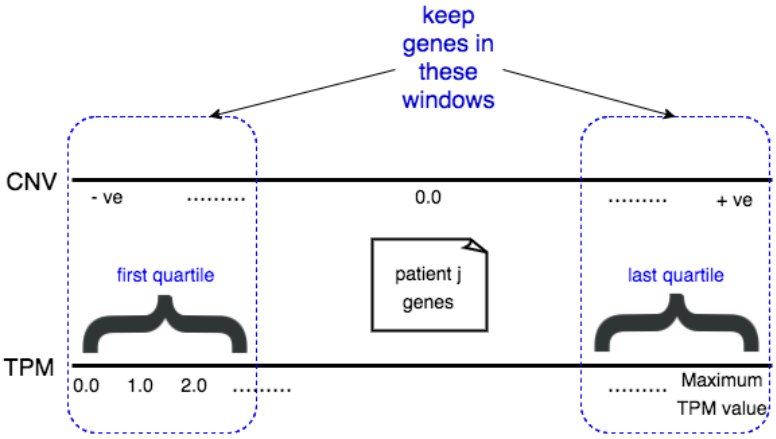


Figure 18 - Filtering CNV by TPM value

In a nutshell, this filtering keep only the genes that their CNV value is in the same direction of their transcripts expression value. The results is a reasonable number of CNV genes per patient that their copy numbers are aligned with their expression levels (in the order of 100s). Similar to $M1$, $M2$ and $M3$, we contruct the matrix $M4$ as following:

$$M4 = \begin{bmatrix} & p_1 & p_2 & \cdots & p_n \\ g_1 & & & & \\ g_2 & & & & \\ \vdots & \vdots & & \ddots & \vdots \\ g_k & & & & \\ r & & & \cdots & \end{bmatrix}$$

The value in each one of these cells represents the log ratio of the copy number in gene g_i for patient p_j for all $i \in [1, 2, \ldots, k]$ and $j \in [1, 2, \ldots, n]$. We set the value equals to 0 if the gene is opted out by previous filtering. Figure 19 presents overall CNV profiles of responders and non-responder samples. This plot shows accumulative sum of the CNV values for the genomic segments across all the chromosomes as the Segment Gain Or Loss (SGOL) versus the chromosomes in both responders and non-responders groups. The SGOL tells us patterns of copy number gain (positive) as green signal and patterns of copy number loss (negative) as red signal. We can notice that most regions share close SGOL values between the two groups. That means that copy number variations cannot be the sole cause for the drug

response. That's why an integrative analysis is required for this study. Plots are generated using cghMCR package in R [44].

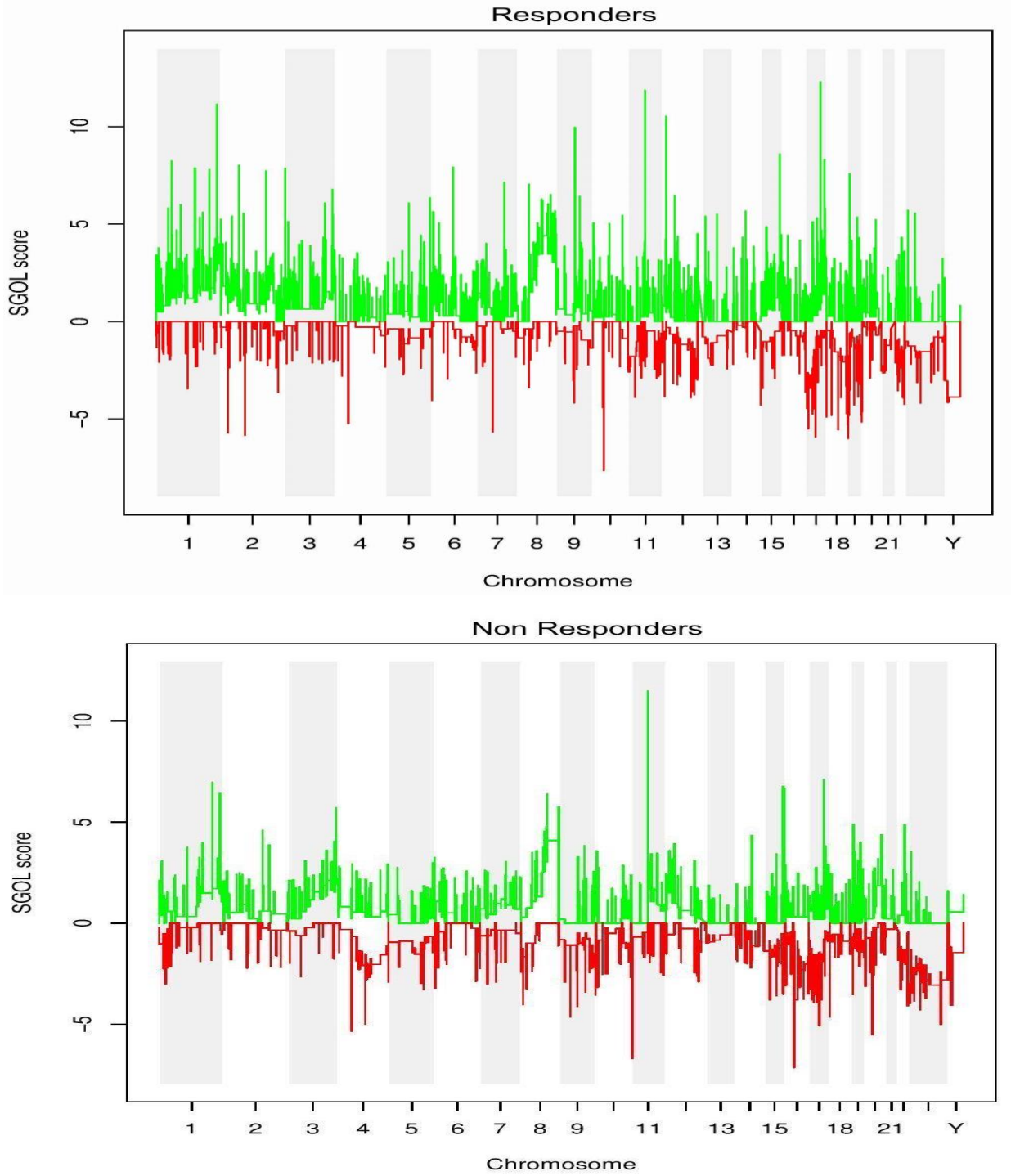


Figure 19 - CNV SGOL in responders and non-responders

Feature Extraction

Referring to Figure 17, our goal is to identify aberrant genes that affected the drug response as candidate biomarkers of drug resistance. The number of genes in each of the $M_{1..4}$ matrices is 19,963. This is a relatively large number compared to the number of samples we have – 16. Therefore, the first step is to narrow down our search to a smaller number of genes. Our strategy is to go backward from the drug response (phenotype) to the aberrant genes (genotype). First, we have extracted genes that their expression levels differ significantly among the responder and non-responder samples. This computation is known as the *differential expression analysis*. We hypothesize that genes that have the same (or close) expression levels between the two groups are transcribed into proteins at the same level. Also, genes that are differentially expressed are transcribed into proteins at different levels. This should be a major contributing factor to the drug response. Second, we try to find a smaller set of genes that can classify the two groups. In the next two subsections, we will see that it is not possible to find such a reduced dimension features. In the following sections, we will illustrate how the analysis is steered from that point.

Differential Expression Analysis

The idea of differential expression (DE) analysis is to identify transcripts and genes that have significantly higher or lower level of abundance in the responder group compared to the non-responder group. To perform DE analysis, we have used EBSeq package [45]. Filtering out with a false discovery rate (FDR) of 0.05, we found that **1188** genes are differentially expressed between the responder and non-responder groups in the breast sample. In the ovarian samples, we found that **657** genes are differentially expressed. It is important to note that:

- There are **50** differentially expressed genes in common between both tissues (breast and ovarian). $|DE\ genes_{breast}| \cap |DE\ genes_{ovarian}| = 50$
- The total number of differentially expressed genes is **1216**.
  $|DE\ genes_{breast}| \cup |DE\ genes_{ovarian}| = 1216$

The result from the differential expression analysis reduced the size of $M1_{breast}$ to have only 1188 genes (dimensions), and the size of $M1_{ovraian}$ to 657 genes (dimensions). However, this is still a large number for performing further regression or classification analysis. To further reduce the number of genes, we implemented a subsequent filtering step using the value of fold change.

Filtering with Fold Change

We removed the genes with low fold change (FC) from the two matrices ($M1_{breast}$ and $M1_{ovarian}$). Fold change is a measure describing how much a gene is over-expressed or down-regulated between the two groups. In order to decide the appropriate value of a threshold to use for filtering out low-concentration genes, we plotted the histograms of log2 of FC of all genes with FDR< 0.05 that is shown in Figure 20. We have noticed two small bumps around 1 and -1, which might be caused by an error. Therefore, we filtered out genes that lie in the range [-2, 2]. Since it is $\log_2$ ratio, this means that we are opting out genes with a fold change between 0.25 and 4 $(0.25 < \mathrm{FC} < 4)$. It turns out that there were 375 genes in this range in $M1_{breast}$ and 204 genes in this range in $M1_{ovarian}$. Therefore, the remaining genes that we will focus on are **813** genes in $M1_{breast}$ and **453** in the ovarian matrix $M1_{ovarian}$.

$$M1_{breast} = \begin{bmatrix} & p_1 & p_2 & \dots & p_7 \\ g_1 & & & & \\ g_2 & & & & \\ \vdots & \vdots & & \ddots & \vdots \\ g_{813} & & & \dots & \\ r & & & & \end{bmatrix}, \quad M1_{ovarian} = \begin{bmatrix} & p_1 & p_2 & \dots & p_9 \\ g_1 & & & & \\ g_2 & & & & \\ \vdots & \vdots & & \ddots & \vdots \\ g_{453} & & & \dots & \\ r & & & & \end{bmatrix}$$

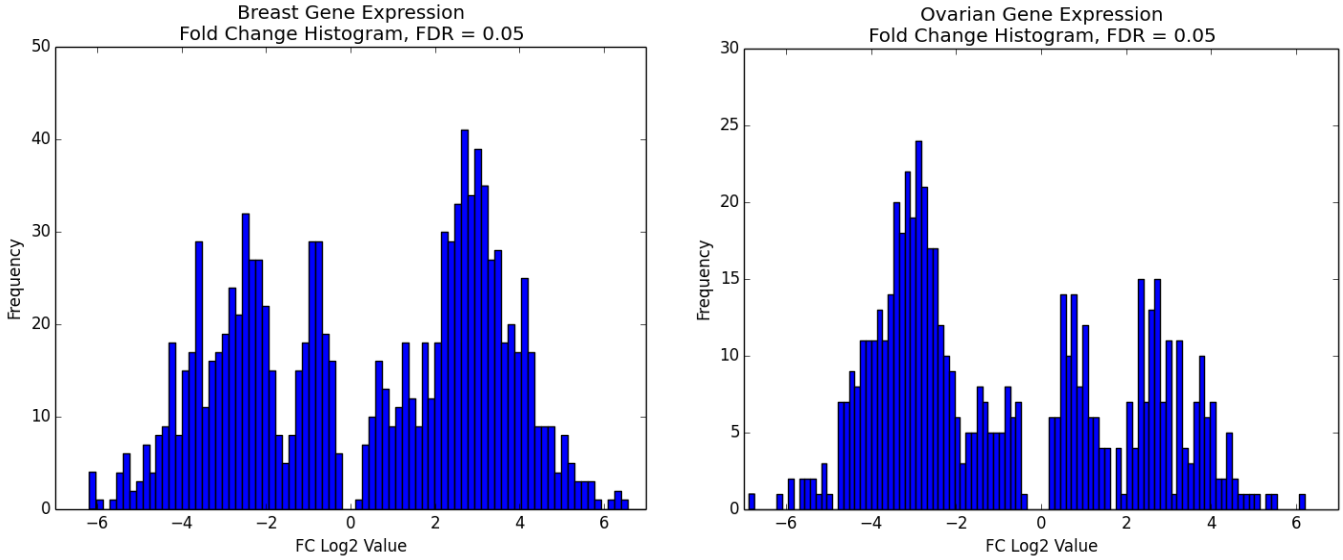


Figure 20 - Histograms of gene expression fold change values

Figure 21 shows the heat maps of the expression levels of the top 50 differentially expressed genes in both breast and ovarian samples. We can see significant over-expression in yellow in some genes of the non-responders group.

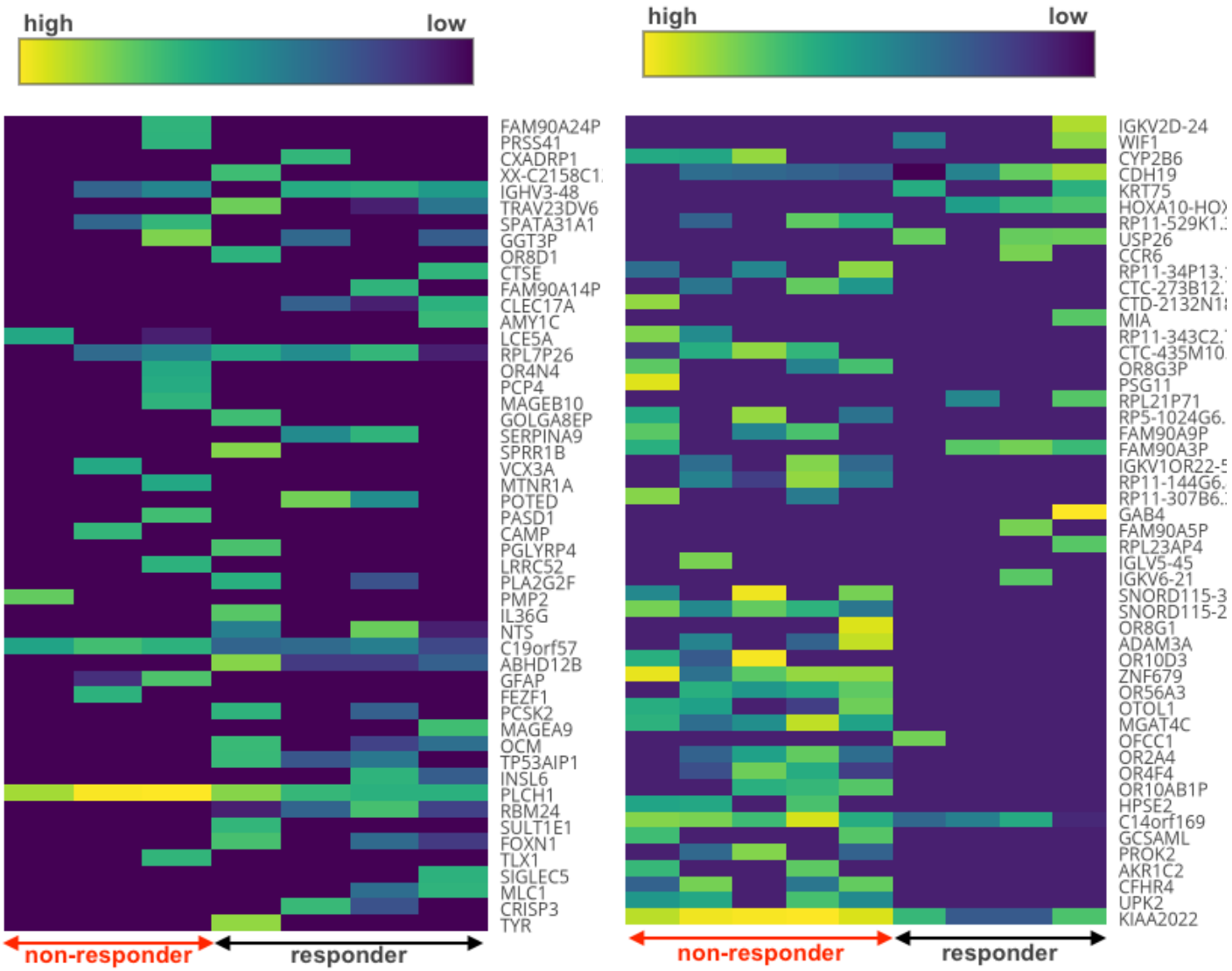


Figure 21 – Gene expression (log_2 TPM). Left: breast samples. Right: ovarian samples

Dimensionality Reduction

Given the differentially expressed genes $M1_{breast}$ and $M1_{ovarian}$, the question now becomes: can we classify the two groups and use them or some combination of them as a predictor of drug response? Having a naked-eye look at $M2$ and $M3$ after removing the non-differentially expressed genes, we did not see many aberrant genes among those DE genes as well. This motivated us to look for some combination of the genes to serve as predictors.

The classical Principle Component Analysis (PCA) comes in handy for this purpose [46]. Figure 23 shows a plot of principle component 1 versus 2 and principle component 1 versus 3 in the breast samples. Similarly, **Error! Reference source not found.** shows the same plots for the ovarian samples.

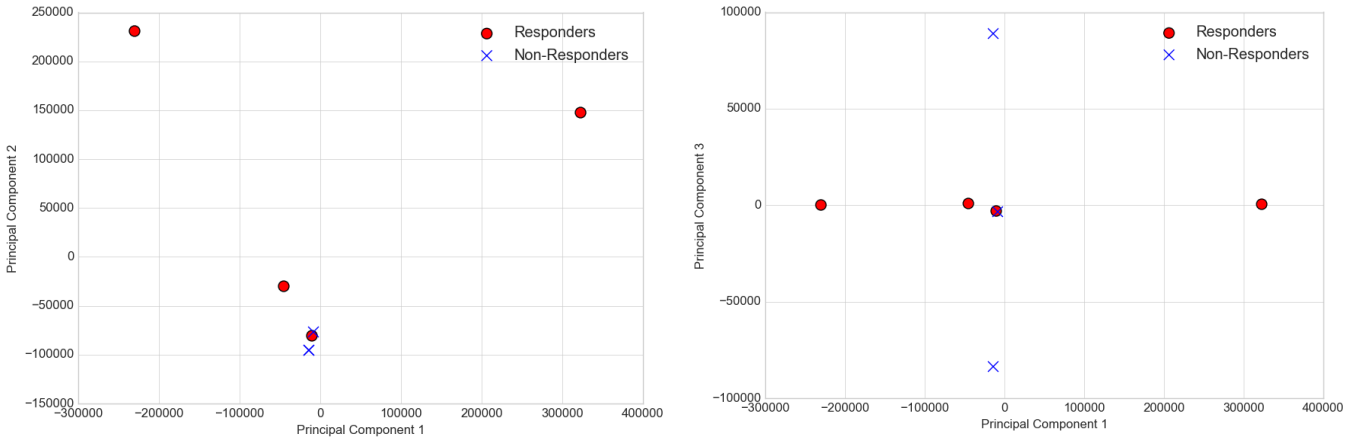


Figure 23 - PCA in breast samples

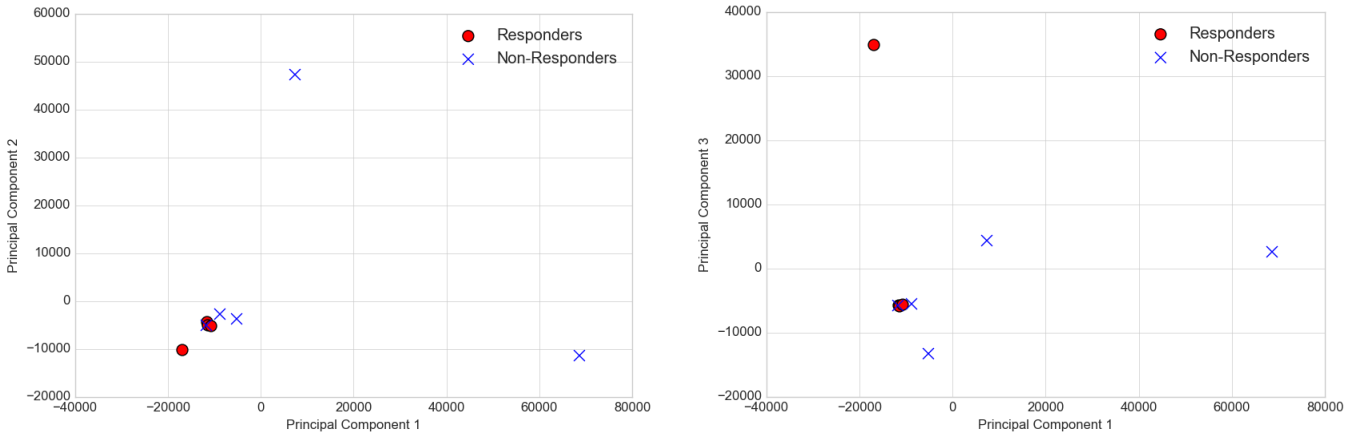


Figure 22 - PCA in ovarian samples

In principle component analysis, the first principle component is a linear combination of all the feature set (genes) that tries to capture the similarities between the data points. The second principle component should be a linear combination of all the feature set that tries to differentiate (separate) the data points above and below zero. Therefore, a successful PCA should show most data points in a small range over PC 1 and a clear separation between the data points over PC 2 or 3. However, the above figures show the exact opposite. Neither PC 1 could capture the similarities, nor PC 2 and 3 could separate between the responder and non-responder groups.

This tells us that every trial to find a linear combination of the genes that can distinguish between the two groups will fail. The reason for that might be the very small number of data points, compared to the relatively large number of features. In addition, it tells us that separation between the responders and non-responders based on the gene expression does not help identify the aberrant genes accountable for the drug response. We can conclude that every patient has her own story of aberrations that led to certain genes to be expressed at a certain level. Therefore, we need to find a higher level of abstraction that can separate the two groups. Working on the small scale of DNA, RNA and gene expression is not the best way to perform such analysis with a small number of patients.

Pathway Analysis

Let's zoom out from investigating at the genome level and look at a higher level of cell functions. We define *molecular pathways* to be a series of actions among molecules in a cell that leads to a certain cell function. Figure 24 shows an abstraction of a simple pathway scenario. Cell receptors bind to signal molecules triggering certain cell functions. For example, a gene regulatory network (GRN) is a collection of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression levels of mRNA and proteins based on the signals received from the receptors. One pathway typically engages more than one gene and one gene is a member of several pathways.

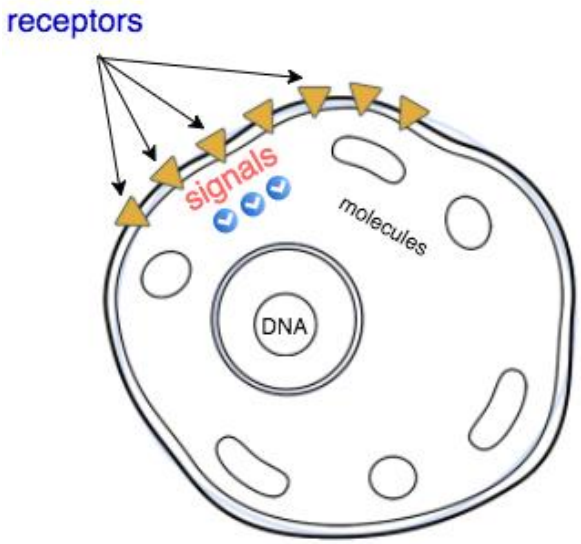


Figure 24 - Molecular pathways

The idea now is to investigate what pathways are involved in the aberrant genes for each patient separately, with the prospect that we can find common pathways among the non-responders that are not found in the responders. At that point, we can focus our analysis on one or two of the most significant pathways.

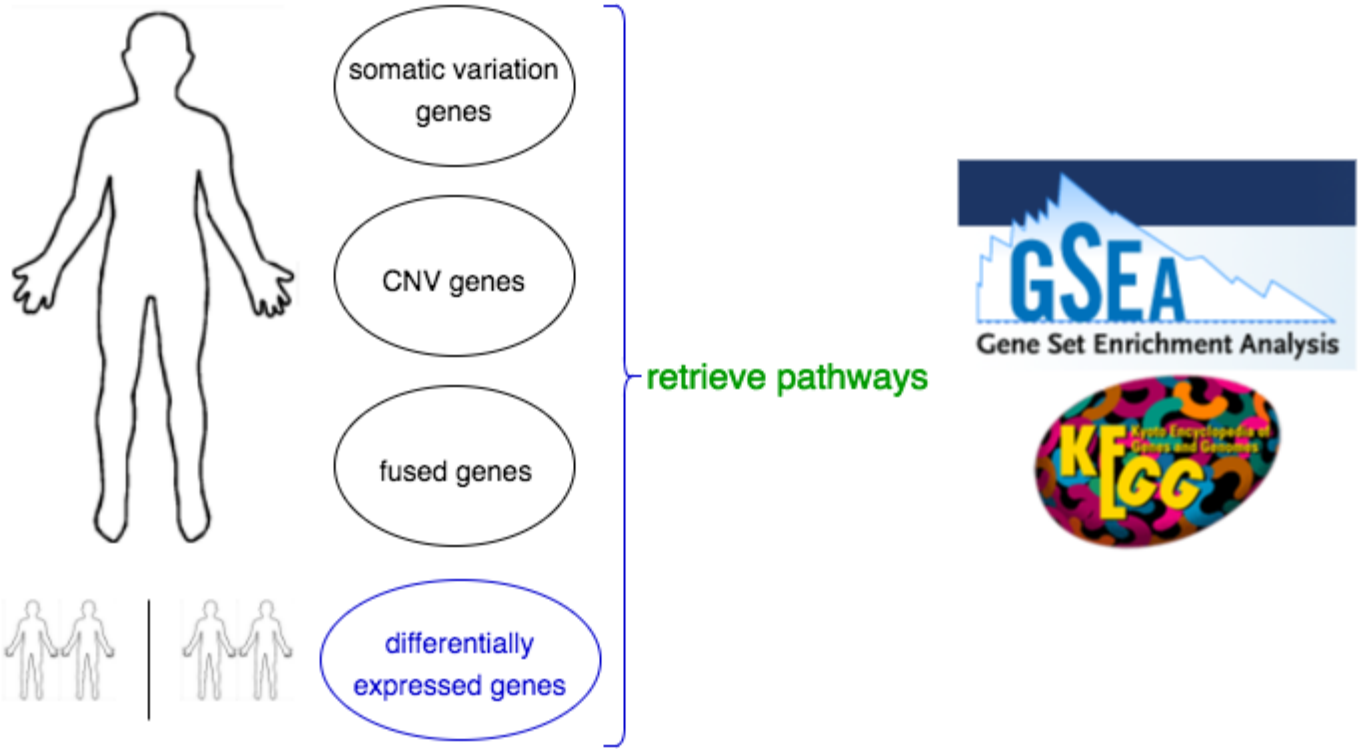


Figure 25 - Per patient pathway analysis

Figure 25 illustrates the pathway analysis that we performed. For each patient, we grouped together genes that have any kind of variations. We queried each matrix (M_2, M_3 and M_4) after performing the filtration steps discussed earlier (not all 19,963 genes) and produced a shortlist of aberrant genes for each patient. Then, we have fetched their corresponding pathways. We have used KEGG pathways and employed the Gene Set Enrichment Analysis tool for connecting to the KEGG pathways database [47,48]. After that, we constructed the pathway matrix P and drug response vector R as following:

$$P = \begin{bmatrix} & pthy_1 & pthy_2 & \cdots & pthy_{m-1} & pthy_m \\ p_1 & & & & & \\ p_2 & & & & & \\ \vdots & & & \ddots & & \vdots \\ p_n & & & & & \\ & & & \cdots & & \end{bmatrix}, \quad R = \begin{bmatrix} \\ : \\ \\ \end{bmatrix}$$

The value in each one of the P cells is equal to:

$$\frac{aberrant\ genes\ for\ p_i\ \cap\ gene\ set\ in\ pthy_j}{gene\ set\ in\ pthy_j} \times 100.0$$

for all $i \in [1, 2, \ldots, n]$ and $j \in [1, 2, \ldots, m]$, where m is the number of enriched pathways (m=133). We set the value equals to 0 if there are no aberrant genes in patient p_i engaged in pathway $pthy_j$. The union of all pathways we got from the above queries was: $\boldsymbol{m = 133}$. As an interpretation, the value in each cell of P tells us how much a given pathway is engaged in the aberrations for each patient. A smaller value means that there were few to no aberrations contributing to the drug response R $(0\ or\ 1)$. A larger value means that aberrations in that pathway contribute significantly to the drug response. The choice of these values will become justifiable in the next subsection.

Regression

Given the matrix P, can we give a weight to each pathway that determines its effect on the drug response? It turns out that regression is a convenient solution for this problem [49]. We represent the matrix P as the fixed input in the below illustration. Columns represent the pathways as features ($j \in [1, 133]$). Rows represent the patients' data points. Values are as depicted in the previous subsection. The objective of regression algorithms is to estimate the drug response $\hat{R}$ using a coefficients vector W as shown in Figure 26.

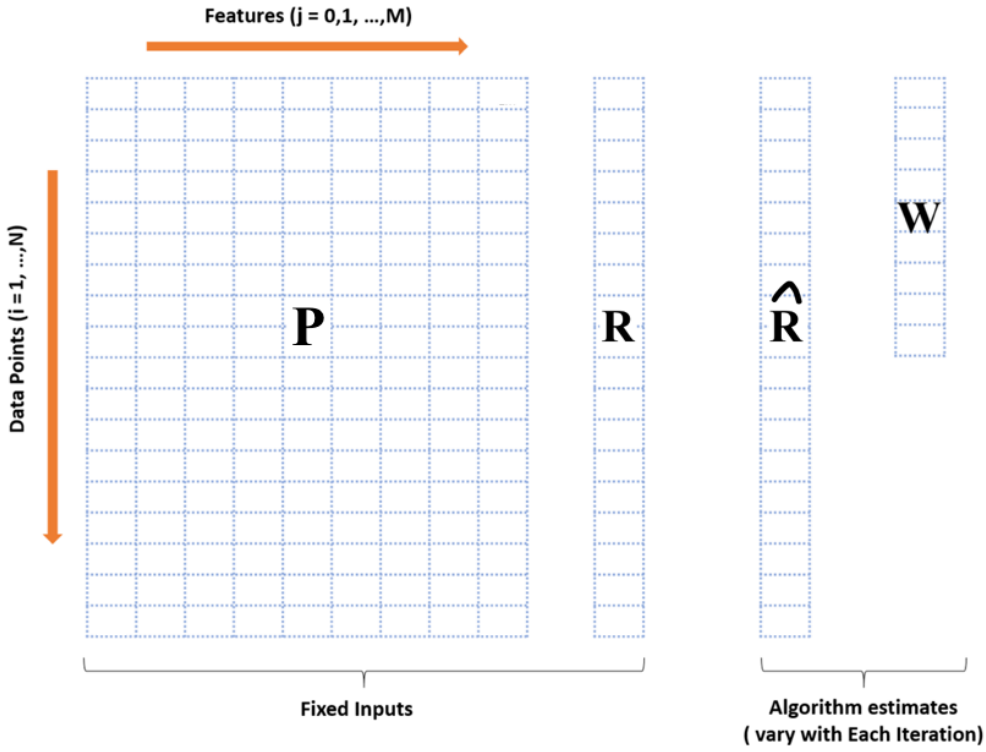


Figure 26 - Matrix setup for the regression analysis

Regression algorithms try to choose values for the W vector that minimize the following:

$$\min_{W} \|PW - R\|_2^2$$

Figure 27 shows an example of a curve (*blue*) resulted from running regression on a sample of data points that have only one feature p_0. However, we have 133 features in our matrix P and only 16 samples (data point). The choice of which regression algorithm to use is very critical in the drug response estimation. We want to obtain few pathways out of 133 pathways that discriminates the responding and non-responding groups. A common pitfall results from overfitting the curve to the data points – typically when the blue line tries to pass through every data point. Therefore, we decided to use Lasso regression.

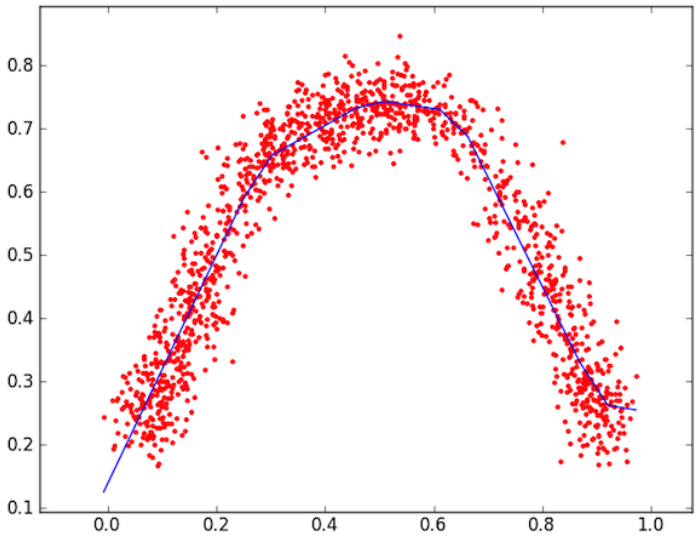


Figure 27 - Regression example in one dimension

Lasso

Lasso (Least Absolute Shrinkage and Selection Operator) penalizes the absolute size of the regression coefficient [50]. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models. Lasso avoids overfitting by adding a *regularization* parameter to the equation above as following:

$$\min_{W} \|PW - R\|_2^2 + \alpha \|W\|_1$$

The added term is called L1 regularization. Another advantage of the L1 regularization is that it leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero. This results to variable selection out of given n variables. So, running the Lasso regression with different values of α will lead to many pathways to be cancelled out

(given a coefficient of zero). Running Lasso on our matrix using different values of α results in the following coefficients (Table 6)

Table 6 – Coefficient values at different values of α

α	rss	intercept	coef_x_1	coef_x_2	coef_x_3	…	coef_x_133	Cancelled Pathways
10^{-15}	0.96	0.22	0.6	-0.37	0.0016	…	10^{-8}	**6**
10^{-10}	0.96	0.22	0.6	-0.37	0.0016	…	10^{-8}	**6**
10^{-8}	0.96	0.22	0.6	-0.37	0.0016	…	10^{-8}	**6**
10^{-5}	0.96	0.5	0.2	-0.13	0	…	10^{-6}	**103**
10^{-4}	1	0.9	0.1	-0.048	0	…	10^{-6}	**115**
10^{-3}	1.7	1.3	0.01	0	0	…	10^{-3}	**116**
10^{-2}	3.6	1.8	-0.55	0	0	…	10^{-3}	**120**
1	37	0.038	0	0	0	…	0	**133**
5	37	0.038	0	0	0	…	0	**133**
10	37	0.038	0	0	0	…	0	**133**

We have tried the Lasso regression with different weights of α to decide an appropriate value that can be used. As we can see in Table 6, the largest values of α made all the coefficients equal to 0. The smallest values of α only cancelled 6 pathways leaving us with 127 pathways to investigate, which is a large number. We have chosen the values of α in the middle rows (highlighted) as they cancel a large number of pathways, but leave us with the most important pathways that have an effect on the drug response.

We can deduce that positive coefficients mean that the corresponding pathways promote the drug response since they are trying to predict a value of $\hat{R}$ close to $R = 1$. Similarly, negative coefficients can be thought of promoting resistance as they are trying to predict a value of $\hat{R}$ close to $R = 0$. Table 7 shows the pathways that are not cancelled along with their corresponding coefficients for each value of the chosen alphas. We have highlighted the pathways that drive drug response with blue and pathways that drive resistance with red. Pathways that are not highlighted are opted out from our further analysis since they are either not related to gene regulations or their coefficients are mostly zero for α values.

Now, we can manually investigate the pathways that promote resistance and filter out those that are not related to cancer. For this step, we have looked at each pathway description on the KEGG database. Four of these pathways have biological meaningful fictions regarding tumor growth and are most likely related to the drug resistance in our study:

1. KEGG_CELL_CYCLE
   Brief description: Mitotic cell cycle progression is accomplished through a reproducible sequence of events, DNA replication (S phase) and mitosis (M phase) separated temporally by gaps known as G1 and G2 phases.
   Link: http://software.broadinstitute.org/gsea/msigdb/cards/KEGG_CELL_CYCLE.html
2. KEGG_HOMOLOGOUS_RECOMBINATION
   Brief description: Homologous recombination (HR) is essential for the accurate repair of DNA double-

strand breaks (DSBs), potentially lethal lesions.
Link: http://software.broadinstitute.org/gsea/msigdb/cards/KEGG_HOMOLOGOUS_RECOMBINATION.html

3. KEGG_INOSITOL_PHOSPHATE_METABOLISM
   Brief description: Inositol phosphate metabolism.
   Link: http://software.broadinstitute.org/gsea/msigdb/cards/KEGG_INOSITOL_PHOSPHATE_METABOLISM.html
4. KEGG_TGF_BETA_SIGNALING_PATHWAY
   Brief description: A wide spectrum of cellular functions such as proliferation, apoptosis, differentiation and migration are regulated by TGF-beta family members.
   Link: http://software.broadinstitute.org/gsea/msigdb/cards/KEGG_TGF_BETA_SIGNALING_PATHWAY.html

Table 7 - Pathways from Lasso regression

Pathway ID (KEGG)	coef at $\alpha = 10^{-5}$	coef at $\alpha = 10^{-4}$	coef at $\alpha = 10^{-3}$	coef at $\alpha = 10^{-2}$
KEGG_VASOPRESSIN_REGULATED_WATER_REABSORPTION	0.0425	0.039	0.014	0.0519
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION	-0.019	0	0	0
KEGG_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	-0.001	0	0	0
KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	-0.020	0	0	0
KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION	0.043	0	0	0
KEGG_N_GLYCAN_BIOSYNTHESIS	-0.027	-0.02	-0.01	-0.02
KEGG_PPAR_SIGNALING_PATHWAY	0.069	0.0863	0.090	0.0480
KEGG_FOCAL_ADHESION	0.066	0.105	0.153	0.036
KEGG_INSULIN_SIGNALING_PATHWAY	0.008	0	0	0
KEGG_OXIDATIVE_PHOSPHORYLATION	-0.051	-0.069	-0.088	0
KEGG_RNA_DEGRADATION	0.008	0	0.0005	0.005
KEGG_INOSITOL_PHOSPHATE_METABOLISM	-0.001	-0.008	-0.018	-0.017
KEGG_LONG_TERM_POTENTIATION	0.071	0.081	0.085	0.061
KEGG_PRIMARY_BILE_ACID_BIOSYNTHESIS	0	0.0003	0	0
KEGG_TGF_BETA_SIGNALING_PATHWAY	-0.078	-0.043	-0.015	0
KEGG_ADHERENS_JUNCTION	0.072	0.0625	0.0372	0
KEGG_CYSTEINE_AND_METHIONINE_METABOLISM	0.0015	0	0	0
KEGG_OOCYTE_MEIOSIS	0.0037	0	0	0
KEGG_TYROSINE_METABOLISM	0.0054	0	0	0
KEGG_NEUROTROPHIN_SIGNALING_PATHWAY	-0.040	-0.048	-0.019	0
KEGG_GLIOMA	0.0001	0	0	0
KEGG_CELL_CYCLE	-0.035	-0.038	-0.017	-0.022
KEGG_RENAL_CELL_CARCINOMA	0.0320	0	0	0
KEGG_VIBRIO_CHOLERAE_INFECTION	0	0.0185	0.0052	0.0051
KEGG_AMYOTROPHIC_LATERAL_SCLEROSIS_ALS	-0.022	0	0	0
KEGG_SPLICEOSOME	-0.054	-0.062	-0.065	-0.021
KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM	-0.002	0	0	0
KEGG_ARRHYTHMOGENIC_RIGHT_VENTRICULAR	-0.052	-0.058	-0.068	-0.031

KEGG_PURINE_METABOLISM	0.0427	0.0624	0	0.0688
KEGG_HOMOLOGOUS_RECOMBINATION	0	-0.004	-0.023	0
KEGG_UBIQUITIN_MEDIATED_PROTEOLYSIS	0.0128	0.0060	0.0033	0.0032
KEGG_OLFACTORY_TRANSDUCTION	0.0009	0	0	0
KEGG_VEGF_SIGNALING_PATHWAY	-0.004	0	0	0

Backtracking

Now that we have narrowed down our research to four pathways, the next step has been to look back at the aberrant genes for each non-responding patient, which engage in these four pathways. We compute the overlap matrix as following:

$$Overlap = \begin{bmatrix} & p_1 & \dots & p_n \\ \text{CELL CYC \_\_} & & \dots & \\ \text{HOMOLOG \_\_} & & \dots & \\ \text{INOSITOL \_\_} & & \dots & \\ \text{TGF BETA \_\_} & & \dots & \end{bmatrix}$$

Each cell of the overlap matrix represents a vector of aberrant genes for patient p_i that are a subset of the corresponding pathway. Then, we have looked at each row of this matrix and identify the most frequent genes that appear in all non-responding patients, but not appearing in the responding group. Then, we have ordered genes in a descending order from most frequent to least frequent. Results are presented in Table 8.

Table 8 - Genes driving drug resistance

Gene	Pathway	Significance to drug resistance	Biological Validation
PIK3CB	Inositol phosphate metabolism	High	PIK3CB (Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Beta) is a Protein Coding gene. PI 3-Kinases (phosphoinositide 3-kinases, PI 3-Ks) are a family of lipid kinases capable of phosphorylating the 3'OH of the inositol ring of phosphoinositides. They are responsible for coordinating a diverse range of cell functions including proliferation and survival.
PIK3C2B	Inositol phosphate metabolism	High	PIK3C2B (Phosphatidylinositol-4-Phosphate 3-Kinase Catalytic Subunit Type 2 Beta) is a Protein Coding gene. It belongs to the same family as PIK3CB.
RAD52	Homologous recombination	High	RAD52 (RAD52 Homolog, DNA Repair Protein) is a Protein Coding gene. Involved in double-stranded break repair. It plays a central role in genetic recombination and DNA repair by promoting the annealing of complementary single-stranded DNA and by stimulation of the RAD51 recombinase.
RPA4	Homologous recombination	High	RPA4 (Replication Protein A4) is a Protein Coding gene. As part of the alternative replication protein A complex, aRPA, binds single-stranded DNA and probably plays a role in DNA repair.
PITX2	TGF-beta signaling pathway	High	PITX2 (Paired Like Homeodomain 2) is a Protein Coding gene. It controls cell proliferation in a tissue-specific manner and is involved in morphogenesis.
MCM3	Cell cycle	High	MCM3 (Minichromosome Maintenance Complex Component 3) is a Protein Coding gene. It acts as component of the MCM2-7 complex (MCM complex) which is the putative replicative helicase essential for once per cell cycle DNA replication initiation and elongation in eukaryotic cells.
MRE11A	Homologous recombination	Medium	MRE11A (MRE11 Homolog A, Double Strand Break Repair Nuclease) is a Protein Coding gene. It is a component of the MRN complex, which plays a central role in double-strand break (DSB) repair, DNA recombination, maintenance of telomere integrity and meiosis.
POLD3	Homologous recombination	Medium	POLD3 (DNA Polymerase Delta 3, Accessory Subunit) is a Protein Coding gene. It is required for optimal DNA polymerase delta activity.
EP300	Cell cycle	Medium	EP300 (E1A Binding Protein P300) is a Protein Coding gene. Bromodomains (BRDs) are epigenetic reader domains that selectively recognize acetylated lysine residues on the tails of histone proteins, and are the only known protein modules that can target

			acetylated lysine residues.
INPPL1	Inositol phosphate metabolism	Medium	INPPL1 (Inositol Polyphosphate Phosphatase Like 1) is a Protein Coding gene. It plays a central role in regulation of PI3K-dependent insulin signaling, although the precise molecular mechanisms and signaling pathways remain unclear.
EP300	TGF-beta signaling pathway	Medium	EP300 (E1A Binding Protein P300) is a Protein Coding gene. Bromodomains (BRDs) are epigenetic reader domains that selectively recognize acetylated lysine residues on the tails of histone proteins, and are the only known protein modules that can target acetylated lysine residues.
RAD54L	Homologous recombination	Low	
MUS81		Low	
BLM		Low	
ORC6	Cell cycle	Low	
STAG1		Low	
ATR		Low	
CCND2		Low	
TFDP2		Low	
YWHAB		Low	
CHEK1		Low	
ORC4		Low	
HDAC1		Low	
WEE2		Low	
CDKN2C		Low	
CCNE1		Low	
PLCZ1	Inositol phosphate metabolism	Low	
PLCB1		Low	
INPP5E		Low	
INPP5B		Low	
INPP4A		Low	
ITPK1		Low	
PLCG1		Low	
ID1	TGF-beta signaling pathway	Low	
PPP2R1B		Low	
BMP2		Low	
BMP7		Low	
ACVR1		Low	

Chapter 5: Discussion

Cancer is a complicated disease and curing it requires enormous amount of efforts in many different directions. Some researchers have been focusing on understanding the disease development. Others have been studying possible treatments. In all directions, the ground truth is that the study of cancer genomics speeds up our progress toward curing the disease. Although every cancer patient has his/her own case of genomic aberrations that caused the disease, a comprehensive analysis of a large number of patients helps us find common aberrations that will serve as a key point-of-focus for further studies.

Given the genomic data of phase-1 clinical trial on 35 patients, our research question was to find candidate biomarkers for the drug resistance using limited number of samples (which is the case for all clinical trial studies). We run an unbiased, data driven and personalized approach to shed light on driver pathways and genes for drug resistance. Our story of following the data science approach can be described as looking for a needle in a haystack. Over 5 terabytes of data have been analyzed to extract a summary of the candidate genes that drive drug resistance when aberrant. Interestingly, the final results – the 4 pathways and the genes with higher score in Table 8 – potentially showed a biological meaning associated with drug response. We believe that aberrations in these genes can significantly contribute to the drug resistance shown in the given dataset. Of course, the wet lab validation of the candidate biomarker is required. Our goal is to generate a short list of candidate biomarker to facilitate the process of biomarker discovery.

During the lifetime of this study, we have faced a number of challenges that we had to deal with.

1. **Big data nature**: as we have seen in chapter 3, genomic data is very large in size. This nature implies a difficulty in designing and running automated pipelines. Computations takes a long time and if for some reason running computations failed, we do not want to re-run the whole pipeline from the beginning; that would be inefficient and time consuming. We have coded scripts that automate the pipeline execution.
2. **Small number of samples**: even though the size of the data is very large, they only represent a small number of samples. That's why conventional statistical analysis and machine learning techniques do not work. Using prior knowledge (pathways) and personalized approach (per patient analysis), we were able to tackle the problem from a different angle.
3. **Data heterogeneity**: the genomic features under study come from different resources and represent different meanings. In addition, they come in different file formats. Thus, we had to pre-process the data to come up with an abstract way of presenting them – using matrices.
4. **Integrative analysis**: due to the small number of samples and the fact that there is many missing information (represented as θ parameters), running our analysis on the micro-level of aberrant genes failed to classify the two groups. Consequently, we had to zoom out into a bigger view on the genomic features – pathways. Running regression on pathways was a successful trial in our case.

For the above challenges, conventional methods do not perform well. Linear and non-linear dimensionality reduction methods cannot find a smaller set of genes that do the classification. Applying regression on aberrant genes fails since the number of genes is very large relative to the number of samples. Perhaps the biggest limitation we have is the small number of samples that prevent us from

generalizing conclusions. The main advantage of the proposed approach is that analysis was done per patient. This means that we can gain a higher level of confidence with the results. Moreover, the data-driven and unbiased approach that does not make any assumption about the distributions of the data offer a strong aspect on solving such problems. On the flip side, this approach might not be the best one to follow if the number of samples is very large, since we will end up with a large number of pathways as features.

This study has generated tons of data that can be further investigated by clinicians and specialists. We can summarize our contribution in this study in the following points:

1. Developing the data analysis pipelines that anatomize raw sequencing data into meaningful information.
2. Researching and implementing different data analysis techniques – borrowed from the data science domain – on the processed data.
3. Developing a method by integrating prior knowledge (from pathways) with genomics data of limited number of sample to identify biologically meaningful candidate biomarkers. Finding clues to the main research question: “can we identify candidate genomic biomarkers that drive drug resistance?”, and suggesting answers.

Future Work

With advances in sequencing technologies and the availability of an increasing amount of high throughput genomics data, the use of advanced computational methods to analyze, integrate, and mine the huge amount of genomics data is an absolute necessity. We see this study as one building block toward a broader integrative analysis of genomics data that will also include microRNA, ChIP-Seq and Methylation data. Our future research steps are:

1. Cross-validation of the results using other software packages in the pipeline analysis.
2. Investigate the prevalence of the identified candidate biomarkers using TCGA data.
3. Integrate the pathway analysis part into the pipelines.
4. Develop a fully-automated pipeline that takes raw sequencing data and perform all the analysis in the steps followed in this study.
5. Include clinical data such as age, race and tumor stage into our feature set.

To conclude, our research focus would be on optimizing the computational pipelines and enhancing the unbiased and data-driven analysis approach.

List of Abbreviations

CNV – Copy Number Variation
DE – Differential Expression
DNA – Deoxyribonucleic Acid
ER – Estrogen Receptors
FC – Fold Change
FDR – False Discovery Rate
FPKM – Fragments Per Kilobase of transcript per Million mapped reads.
GRN – Gene Regulatory Network
Lasso – Least Absolute Shrinkage and Selection Operator
PCA – Principle Component Analysis
PR – Progesterone Receptor
RNA – Ribonucleic Acid
RNA-Seq – RNA Sequencing
RPK – Reads Per Kilobase
SAM – Sequence Alignment/Mapping
SGOL – Segment Gain Or Loss
SNP – Single Nucleotide Polymorphism
TCGA – The Cancer Genome Atlas
TNBC – Triple Negative Breast Cancer
TPM – Transcripts Per Million
VCF – Varian Calling Format
WES – Whole Exome Sequencing

References

1. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458:719–24.

2. Worldwide cancer statistics [Internet]. 2012. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer

3. Department of Defense Breast Cancer Research Program. The Breast Cancer Landscape [Internet]. Department of Defense Breast Cancer Research Program; 2016 Feb p. 11. Available from: http://cdmrp.army.mil/bcrp/

4. Azvolinsky A. Study Finds Ovarian and Basal-Like/Triple-Negative Breast Cancers Genetically Similar [Internet]. 2012. Available from: http://www.cancernetwork.com/triple-negative-breast-cancer/study-finds-ovarian-and-basal-liketriple-negative-breast-cancers-genetically-similar

5. JS R-F. Everything You've Always Wanted to Know About Triple-Negative Breast Cancers, Scientific Seminar [Internet]. London, UK; 2014. Available from: http://www.dako.com/us/index/knowledgecenter/kc_publications/white-papers-reviews-studies/breast-cancer-diagnostics-symposia/triple-negative-breast-cancers.htm?setCountry=true&purl=index/knowledgecenter/kc_publications/white-papers-reviews-

studies/breast-cancer-diagnostics-symposia/triple-negative-breast-cancers.htm?undefined&submit=Accept%20country#.WBDQIJMrKYU

6. Foulkes WD, Smith IE, Reis-Filho JS. Triple-Negative Breast Cancer. N. Engl. J. Med. 2010;363:1938–48.

7. Plaskocinska I, Shipman H, Drummond J, Thompson E, Buchanan V, Newcombe B, et al. New paradigms for BRCA1/BRCA2 testing in women with ovarian cancer: results of the Genetic Testing in Epithelial Ovarian Cancer (GTEOC) study. J. Med. Genet. 2016;53:655–61.

8. Mavaddat N, Peock S, Frost D, Ellis S, Platte R, Fineberg E, et al. Cancer Risks for BRCA1 and BRCA2 Mutation Carriers: Results From Prospective Analysis of EMBRACE. JNCI J. Natl. Cancer Inst. 2013;105:812–22.

9. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. The Lancet. 2011;378:1812–23.

10. Guha N, Kwan ML, Quesenberry CP, Weltzien EK, Castillo AL, Caan BJ. Soy isoflavones and risk of cancer recurrence in a cohort of breast cancer survivors: the Life After Cancer Epidemiology study. Breast Cancer Res. Treat. 2009;118:395–405.

11. Bast RC, Hennessy B, Mills GB. The biology of ovarian cancer: new opportunities for translation. Nat. Rev. Cancer. 2009;9:415–28.

12. UConn High-Performance Computing Cluster [Internet]. Available from: http://bioinformatics.uconn.edu/

13. Lahti L, Schafer M, Klein H-U, Bicciato S, Dugas M. Cancer gene prioritization by integrative analysis of mRNA expression and DNA copy number data: a comparative review. Brief. Bioinform. 2013;14:27–35.

14. Huang N, Shah PK, Li C. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. Brief. Bioinform. 2012;13:305–16.

15. Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale A-L. Principles and methods of integrative genomic analyses in cancer. Nat. Rev. Cancer. 2014;14:299–313.

16. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, et al. An Integrated Approach to Uncover Drivers of Cancer. Cell. 2010;143:1005–17.

17. Nabavi S. Identifying candidate drivers of drug response in heterogeneous cancer by mining high throughput genomics data. BMC Genomics [Internet]. 2016 [cited 2016 Oct 26];17. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2942-5

18. Jennings EM, Morris JS, Carroll RJ, Manyam GC, Baladandayuthapani V. Bayesian methods for expression-based integration of various types of genomics data. EURASIP J. Bioinforma. Syst. Biol. 2013;2013:13.

19. Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, Do K-A. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. Bioinformatics. 2013;29:149–59.

20. Görlich D, Kutay U. Transport Between the Cell Nucleus and the Cytoplasm. Annu. Rev. Cell Dev. Biol. 1999;15:607–60.

21. Alberts B. Molecular biology of the cell. Sixth edition. New York, NY: Garland Science, Taylor and Francis Group; 2015.

22. Sawicki MP, Samara G, Hurwitz M, Passaro E. Human Genome Project. Am. J. Surg. 1993;165:258–64.

23. Tucker T, Marra M, Friedman JM. Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. Am. J. Hum. Genet. 2009;85:142–54.

24. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38:1767–71.

25. Hosny A. GitHub repository: Nabavi Lab/TNBC [Internet]. 2016. Available from: https://github.com/NabaviLab/TNBC_Project

26. Andrews S. FastQC: A quality control tool for high throughput sequence data. 2010.

27. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

28. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002;12:996–1006.

29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

30. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

31. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 2011;43:491–8.

32. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline: The Genome Analysis Toolkit Best Practices Pipeline. In: Bateman A, Pearson WR, Stein LD, Stormo GD, Yates JR, editors. Curr. Protoc. Bioinforma. [Internet]. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2013 [cited 2016 Nov 9]. p. 11.10.1–11.10.33. Available from: http://doi.wiley.com/10.1002/0471250953.bi1110s43

33. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–76.

34. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

35. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5:557–72.

36. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC Known Genes. Bioinformatics. 2006;22:1036–46.

37. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164–e164.

38. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 2009;10:57–63.

39. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

40. Haas BJ, Zody MC. Advancing RNA-Seq analysis. Nat. Biotechnol. 2010;28:421–3.

41. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

42. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016;44:D710–6.

43. Nicorici D, Satalan M, Edgren H, Kangaspeska S, Murumagi A, Kallioniemi O, et al. FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data [Internet]. 2014 Nov. Report No.: 011650. Available from: http://biorxiv.org/lookup/doi/10.1101/011650

44. Zhang J, Feng B. cghMCR: Find chromosome regions showing common gains/losses. R package version 1.32.0 [Internet]. 2016. Available from: http://mirror.ufs.ac.za/bioconductor/packages/3.4/bioc/html/cghMCR.html

45. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013;29:1035–43.

46. Principal Component Analysis [Internet]. New York: Springer-Verlag; 2002 [cited 2016 Oct 26]. Available from: http://link.springer.com/10.1007/b98835

47. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. 2005;102:15545–50.

48. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27–30.

49. Chamberlain G. Multivariate regression models for panel data. J. Econom. 1982;18:5–46.

50. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B Methodol. [Internet]. 1996;58. Available from: http://www.jstor.org/stable/2346178?seq=1#page_scan_tab_contents