Master Thesis

Integrative gene based analysis of cancer copy number alteration, simple mutation and methylation data

Author(s):
Gupta, Saumya

Publication Date:
2012

Permanent Link:
https://doi.org/10.3929/ethz-a-007554600

Rights / License:
In Copyright - Non-Commercial Use Permitted
Integrative gene based analysis of cancer copy number alteration, simple mutation and methylation data

Master Thesis
Saumya Gupta
August 20, 2012

Advisors: Dr. Michael Baudis (Group leader)
Prof. Christian von Mering (Co-supervisor), Nitin Kumar (PhD)
Institute of Molecular Life Sciences, University of Zurich
I would like to express my gratitude to all those who gave me this opportunity to work, learn and add immensely to my personal experience. I would like to acknowledge my sincere gratitude to my supervisor and guide Dr. Michael Baudis, under whose guidance I had the opportunity to work. He has been helpful at every step of the study and his devotion and enthusiasm as a teacher has provided me inspiration that is not only limited to the present work, but also shall forever be my strength.

I am also extremely grateful to Nitin Kumar for his valuable suggestions, immense cooperation and encouragement at every step of my work. He has been very patient and always available to solve my queries and guided me thoroughly throughout my thesis. I would also like to express my heartiest gratitude to Prof. Christian von Mering and Prof. Dr. Mark D. Robinson for their valuable suggestions and encouragement during my thesis. I also thank Haoyang Cai and Ni Ai who were of great help during discussions for my work.

Especially, I would like to give my special thanks to my family and friends whose blessings and support enabled me to complete this work.

SAUMYA GUPTA
# Contents

## 1 Introduction

1.1 Motivation .................................................................................................................. 1

1.2 Overview .................................................................................................................... 2

1.2.1 Acquired capabilities of cancer .............................................................................. 2

1.2.2 Genomic alterations in cancer ................................................................................ 3

1.2.2.1 Point mutations .................................................................................................. 5

1.2.2.2 Copy number alterations .................................................................................. 5

1.2.2.3 DNA methylation .............................................................................................. 5

1.3 Objectives .................................................................................................................... 6

## 2 Materials and Methods

2.1 Data sources ................................................................................................................. 8

2.1.1 International Cancer Genome Consortium ......................................................... 8

2.1.2 arrayMap ............................................................................................................... 9

2.1.3 Ensemble Genome Browser .................................................................................... 9

2.2 Data pre-processing .................................................................................................... 10

2.2.1 ICGC data pre-processing .................................................................................. 10

2.2.2 arrayMap data pre-processing .............................................................................. 11

2.3 Statistical tests ........................................................................................................... 11

2.3.1 Binomial test ......................................................................................................... 12

2.3.2 Fischer's exact test ................................................................................................. 12

2.3.3 FDR adjustment .................................................................................................. 13

2.4 Methodology .............................................................................................................. 14

...
2.4.1 Frequency based gene enrichment ......................................................... 14
2.4.2 Copy number data coverage ................................................................. 14
2.4.3 Pathway enrichment analysis: Classical Approach ............................... 16
2.4.4 Pathway enrichment analysis: Permutations Approach ......................... 17

3 Results and Discussions ............................................................................. 19
3.1 ICGC data distribution .............................................................................. 19
3.1.1 Gene enrichment .................................................................................. 22
3.1.2 Comparison with Cosmic cancer database ........................................... 23
3.1.3 Pathway enrichment analysis ............................................................... 24
3.2 arrayMap data analysis ............................................................................ 27

4 Conclusion and Future Outlook ................................................................. 30

Bibliography .................................................................................................... 31
Chapter 1

Introduction

This chapter contains the introduction, motivation, overview of cancer and objectives of the work done here.

1.1 Motivation

Cancer is an evolutionary genomic disease, characterized by the accumulation of different types of DNA sequence and structural alterations including point mutations, insertions, deletions, chromosomal rearrangements, and loss or gain of whole chromosome arms, as well as by epigenetic changes (e.g. DNA methylation abnormalities). These alterations may target genes that are responsible for maintaining the normal cellular function. These mutations lead to uncontrolled proliferation of cells that can permeate through normal tissue and even metastasize to other organs [1] [17].

Different types of somatic genomic or epigenomic changes are a prerequisite for cancer development and progression. These changes vary from single base point mutations to copy number alterations (CNA) and from structural changes (e.g. translocations) to methylation. All or either of these genomic changes could be present in a sample analyzed. With the advancement of new technologies and integrative consortiums such as ICGC (International cancer genome consortium) and TCGA (The cancer genome atlas) data covering information about all kinds of genomic mutations is being made available to the research community. In order to understand the contribution of each genomic feature to the development of cancer, an integrative analysis is required.

Also, other challenges involved in cancer research are distinguishing "driver" mutations (present in cancer genes that confer growth advantage on the cells carrying them and are positively selected during the evolution of the cancer) from “passenger” mutations (present in an ancestor of the cancer cell that do not confer growth advantage), determining frequency of mutational occurrence at any genomic locus across various samples, comparing different stages of cancer development and identifying frequently perturbed biological pathway networks [15].

Therefore, there is an urgent need for the sequencing of cancer genomes so that this data can be thoroughly analyzed for the better understanding of cancer biology, including pathway information and types of mutations associated with different cancers. This means
of detecting cancer at early stages by providing information on genetic signatures of cancer will ultimately lead to improved treatments and therapies for cancer targeting a specific tissue.

1.2 Overview

Over the years, cancer has emerged as a disease involving dynamic changes in the genome [1]. Due to extensive research, it has been established that various types of mutation result in altering mainly two types of cancer causing genes i.e. oncogenes (increased activity in cancer) and tumor suppressor genes (decreased activity in cancer) that foster a cell with some special capabilities.

1.2.1 Acquired capabilities of cancer

Most of the cancer cells possess six essential changes in their physiology that lead to the malignancy. Each of these qualities acquired by the cells during cancer development pose a significant threat to the normal functioning of the mechanisms in-built in normal cells for maintaining their homeostasis. These qualities are acquired by entering through a wide range of mechanisms and pathways present in the cell [1]. The six essential capabilities are as follows:

1. Self-sufficiency in growth signals
   Normal cells need mitogenic growth signals (GS) to move from quiescent state (G₀) to active proliferative state whereas the tumor cells exhibit highly reduced dependence on exogenously derived growth signals, disrupting the homeostatic mechanism which is responsible for ensuring proper behavior of cells. This acquired GS autonomy is achieved by alteration of either extracellular growth signals, trans-cellular transducers of those signals or intracellular circuits that translate those signals.

2. Insensitivity to antigrowth signals
   In normal cells, many anti-proliferative signals are present to maintain the quiescent (G₀) state and homeostatic mechanism of cells. These cells may either re-emerge into proliferation when extracellular signals permit or can be induced to permanently relinquish their proliferative potential by making them enter the post-mitotic states. Tumor cells evade these anti-proliferative signals by disrupting the pathways that operate to block the transition through G1 phase of cell cycle. For example: Disruption of pRb pathways liberates E2Fs (transcription factors that control the progression from G1 to S phase) resulting in cell proliferation.

3. Evading apoptosis
   The tumor cell populations are determined both by rate of cell proliferation and the rate of cell attrition. Apoptosis (Programmed cell death) results in attrition. The
complete breakdown of cellular apoptotic machinery and resistance of a cell towards death is considered as one of the major causes of cancer development. For example: TP53 tumor suppressor gene (an important component of apoptosis pathway) inactivation leads to rapidly growing tumors with lesser number of cells with normal apoptotic function.

4. Limitless replicative potential
Normal cells have an intrinsic, cell-autonomous program limiting the cell multiplication which operates independently of cell-to-cell signaling pathways. Therefore, along with growth signal autonomy, insensitivity to anti-growth signals and evading apoptosis, this program should also be disrupted to impair cell multiplication potential in tumor cells.

5. Tissue invasion and metastasis
This results from the ability of tumor cells to invade adjacent tissues and travel to distant sites to find new colonies where they have plenty of nutrients and space to develop (metastasis). The proteins involved in the tethering of cells to their surroundings in a tissue are altered in cells possessing invasive or metastatic capabilities. For example: E-cadherin, a cell-to-cell interaction molecule responsible for suppressing invasion and metastasis in epithelial cancers, is functionally eliminated resulting in enhanced invasive and metastatic capability.

6. Sustained angiogenesis
In order to support unlimited growth, tumor cells need to have an ability of sustained growth of new blood vessels (angiogenesis) from pre-existing vessels for an adequate supply of oxygen and nutrients to the dividing cells. The induction and sustainability of angiogenesis seems to be an early to mid stage event in many human cancers that activate the angiogenic switch by changing the balance of angiogenesis inducers and countervailing inhibitors. For example: Increased expression of VEGF (Vascular endothelial growth factor) and FGFs (Fibroblast growth factors) due to altered gene transcription.

1.2.2 Genomic Alterations in Cancer
Every cancer cell carries a diploid genome. The set of differences in the DNA sequence of a cancer cell compared to the normal inherited genome composition is collectively termed as “somatic mutations”. This is in contrast to germline mutations, which are inherited from either of the parents or arise during gametogenesis or early embryonic development and can therefore be found in all cells of the body [2].

Only a small fraction of cancer related mutations may be fixed and inherited through the germline. Somatic mutations are acquired in “normal” tissue type specific ancestors of the cancer cells, sometimes due to the exposure to mutagens of both internal and
external origins. Although DNA replication itself has a low intrinsic error rate some processes such as DNA repair defects increase the mutational load. Mutation rates also increase in the presence of substantial exogenous mutagen exposures or various forms of radiation [2]. The accumulation of these mutations throughout the development of a human from a fertilized egg to an adult is illustrated in Fig 1.1.

Figure - 1.1 The lineage of mitotic cell divisions from the fertilized egg to a single cell within a cancer showing the timing of the somatic mutations acquired by the cancer cell and the processes that contribute to them. (Michael R. Stratton, Peter J. Campbell and P. Andrew Futreal: The cancer genome. 2009. Nature 458)

Somatic mutations in a cancer cell constitute different types of DNA sequence changes that are as follows:

- Point Mutation is addition, substitution or deletion of single base.
- Copy Number Alterations are insertions or deletions of small or large segments of DNA
- Chromosomal rearrangements, in which DNA has been broken and then rejoined to a DNA segment from elsewhere in the genome or on the same chromosome itself.

Apart from somatic mutations, the cancer may also have epigenetic changes that are modifications to chromosomes that alter gene-expression patterns occurring through DNA methylation, acetylation, or phosphorylation of histones and other proteins around which DNA is wound to form chromatin [3] or it may also have acquired new DNA sequences from external sources like viruses [2].
1.2.2.1 Point Mutations

Genetic alterations that alter only one or a few nucleotides along a DNA strand are called as point mutations.

The three-letter codon coding for one amino acid read by ribosomes may be changed by mutation in one of three ways [4]:

1. Nonsense mutations
   The mutated codon is a stop codon which leads to the premature termination of protein, producing a shorter protein rendering it in a non-functional state.

2. Missense mutations
   The mutated codon results in a codon that code for some other amino acid to be inserted into the protein. The function of the faulty protein depends on newly inserted aa. These types of mutations are also called as non-synonymous mutations that are crucial in cancer development.

3. Frame-shift mutations
   The loss or gain of 1 or 2 nucleotides causes the affected codon and all of the codons that follow to be misread leading to a non-functional protein with different sequence of amino acids.

Some nucleotides may not be 'read' by RNA polymerase due to some DNA damage which might be corrected by adding in nucleotides, even if it means putting in the wrong thing. This process is known as transcriptional mutagenesis and it may play a significant role in the development of cancer [5].

1.2.2.2 Copy Number Alterations

Copy number alterations (CNA) refers to the gains and losses of DNA segments. It differs from copy number variations that are generally found in germline cells [6] and mostly without a (known) relation to a specific disease phenotype.

Generally, copy number alterations are linked with gene expression. Mostly, tumor suppressor genes suffer loss of DNA and are under expressed whereas oncogenes suffering copy number gains are over expressed, both leading to tumor development [7].

Copy number alterations can be detected by using cytogenetic techniques such as fluorescent in situ hybridization, comparative genomic hybridization, array comparative genomic hybridization, and by virtual karyotyping with SNP arrays [23].
1.2.2.3 DNA Methylation

Methylation commonly occurs at DNA sequences called "CpG islands," which are cytosine and guanine rich regions, often found in gene promoters but which can also occur in other areas of a chromosome. DNA methylation can either lead to hypermethylated or hypomethylated states which can be both present in cancer or normal cells. In other words, a genomic region that is hypomethylated in normal tissue samples can be hypermethylated in tumor-cell samples. Although certain methylation changes are specific to cancer types there is no fixed pattern of to what extent these changes occur in cancer cells. However, a change in the methylation state of DNA can play a crucial role in tumor growth by oncogene activation, or in tumor suppressor gene silencing and chromosomal instability.

DNA hypermethylation can result in silencing of genes that are responsible for proliferation, apoptosis, DNA repair, and immortalization, giving rise to tumor growth. For example: Methylation of the cyclin-dependent kinase inhibitor CDKN2A, a tumor suppressor gene, leads to unlimited growth of breast and lung epithelial cells. DNA hypomethylation can also occur in many human cancers but still to be supported by more evidences to be established as one of the causes for tumorigenesis [3].

1.3 Objectives
This master thesis comprises of two projects. Each project is based on analyzing cancer genome mutation data from two different data sources.

The first project involves an integrative analysis of copy number alteration (CNA), somatic mutation and methylation data from various cancer types. Most of the work will be focused on ICGC (International cancer genome consortium) data. It involves a gene centric analysis of all kinds of genomic and epigenomic changes present in cancer [19]. The main focus of the project is to identify genes which are selectively targeted by different kinds of somatic genome changes and to evaluate a preferential bias of some genes being selectively targeted by only specific mutations types.

A part of the analysis will also focus on genes co-altered and/or are mutually exclusively affected by somatic changes. This will then be followed by a pathway centric analysis of different kinds of mutation data and their combination, to identify pathway modules preferentially targeted by specific mutations.

The second project involves the analysis of copy number alteration data from arrayMap database. The main focus of the project will be to identify the frequency of genes encompassed by copy number alterations analyzed for their various classes, determined based on their size range [20]. This approach will help in the identification of genes on
which a selective pressure for CNA accumulation exists, in contrast to genes which may be hit as “bystanders” in large scale CNA.

The analysis will be carried out separately for gains and losses and the data will be filtered for each window size i.e. the copy number alterations lesser or equal to a particular window size. This is followed by pathways centric analysis for the set of genes found significant in each window size.

A part of analysis will focus on the detection of copy number variations based on the genes found significant from the copy number alterations less than 1 mbp in size.
Chapter 2

Materials and methods

2.1 Data sources

There are multiple sources for cancer genome data. Many research groups collaborate and pour in both clinical and research data into huge databases which is then used by other groups to work further on. In this work, two of these databases have been used, the details of which are given below.

2.1.1 International Cancer Genome Consortium (ICGC)

All the three types of data i.e. Simple mutations, Copy Number Alterations and Methylation data for the integrative analysis has been obtained from International Cancer Genome Consortium (ICGC) dataset - Version 8 (March 15th, 2012) - [http://dcc.icgc.org/web/](http://dcc.icgc.org/web/) across 29 different cancer types and total 3561 donors. The number of samples for each tissue type is illustrated in Fig 2.1.

![Figure 2.1 - Overview of the ICGC Dataset](image-url)
ICGC serves as a repository for different cancer research groups worldwide that coordinate large-scale genome studies in tumours across 50 different cancer types and subtypes and submit different types of data e.g. Simple mutations, Copy Number Alterations, Methylation, Expression, Structural variations and miRNA sequencing data and exon junctions. ICGC uses the term simple mutations to refer to point mutations. Hence, the term simple mutation has been used throughout the thesis keeping the terminology of the data from ICGC.

It provides a defined catalogue for each tumour type and subtype to include all the genomic alterations present in cancer in different forms and make this data available to all the research communities as quickly as possible with minimal restrictions [8].

The Data Portal also hosts data from other large-scale cancer genome projects including The Cancer Genome Atlas (TCGA), Tumor Sequencing Project (TSP) and Johns Hopkins University apart from ICGC participating institutions across the world.

### 2.1.2 arrayMap

The CNA data was obtained from arrayMap database ([http://www.arraymap.org/cgi-bin/amHome.cgi](http://www.arraymap.org/cgi-bin/amHome.cgi)) across various cancer types.

arrayMap developed by the group "Theoretical Cytogenetics and Oncogenomics" at the Institute of Molecular Life Sciences of the University of Zurich, serves as an excellent source for copy number profiling data in different human cancers. This data can be used for meta-analysis as well as can be integrated with high resolution cancer copy number data [9]. The current data reflects:

- 42875 genomic copy number arrays
- 634 experimental series
- 256 array platforms
- 197 ICD-O cancer entities
- 480 publications (Pubmed entries)

### 2.1.3 Ensembl Genome Browser

Ensembl ([http://www.ensembl.org/info/about/intro.html](http://www.ensembl.org/info/about/intro.html)) is a collaborative project between European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI).

It serves as source of information about genes identified in different organisms. It annotates the genome, integrates this annotation with other available biological data and
make all this publicly available via web. This annotation can either be automatic (Genome-wide scale) or manual (Case-by-case scale). For most of the human genes, HGNC symbol from the HUGO Gene Nomenclature Committee which are approved symbols are given.

2.2 Data pre-processing
The data was available in different formats and thus could not be directly used for analysis. The primary reason the format of the data was non-uniform and inconsistent across different cancer types. Some of the data files were tab separated while others were either space separated or they had to be read line by line. Also, the data contained a lot of additional information which was not useful for gene-based analysis. Apart from that, the useful information from the data file was missing for various samples and sometimes they could not be filtered for the sample type. Thus, they had to be left out from the analysis. Hence, there was a need to pre-process data to overcome these challenges and the methodology for this is explained in detail below. Additionally, it is always useful to convert the data into a form where it can be manipulated and processed in an efficient manner (e.g. in R, matrix operations are faster than data frames).

2.2.1 ICGC Data Pre-processing
For integrative analysis, the data was filtered for only tumour samples including tumour xenografts and cell lines derived from tumour and excluding controls and pure cell lines.

Then, combined matrices of all cancer types were generated with unique Sample ID’s as rows and Gene Id’s as columns for each cancer type. These individual matrices were combined into a single matrix for all cancer types available in each type of data. The matrices generated for analysis are illustrated in Fig 2.2 and 2.3.

<table>
<thead>
<tr>
<th>ENSG00000001629</th>
<th>ENSG00000001631</th>
<th>ENSG00000017260</th>
<th>ENSG00000021645</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFS_MB1 &quot;0&quot; &quot;0&quot; &quot;0&quot; &quot;gain&quot; &quot;gain&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LFS_MB2 &quot;gain&quot; &quot;gain&quot; &quot;gain&quot; &quot;0&quot; &quot;0&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LFS_MB3 &quot;gain&quot; &quot;0&quot; &quot;0&quot; &quot;0&quot; &quot;0&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LFS_MB4 &quot;0&quot; &quot;0&quot; &quot;0&quot; &quot;0&quot; &quot;0&quot;</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.2 – Character Matrix of Copy Number Data
Figure 2.3 – Character Matrix of Methylation Data

All the three matrices were converted to binary matrices by setting cut-offs as follows:

a) For Methylation data:
   - Multiple values for same gene and sample were dealt with by taking the median of all the values.
   - Value < 0.1: Hypomethylated and Value >= 0.75: Hypermethylated was set to 1
   - 0.1 < Value < 0.75 was set to 0.

b) For Simple mutation data:
   - Values with non synonymous mutations were set to 1.
   - Values with any other type of mutation i.e. 3’ utr, 5’ utr, complex indel, complex substitution, frameshift coding, inframe indel, intergenic, synonymous coding etc. were set to 0.

c) For Copy number alterations data:
   - Values with gain or loss (genes altered) were set to 1 and all others including LOH (Loss of Heterozygosity) 0.

2.2.2 arrayMap Data pre-processing

Copy number data coverage was determined for various genomic intervals. Several window sizes of 100kbp, 200kbp, 500kbp, 1mbp, 2mbp, and 5mbp were used. Sex chromosomes were excluded from the analysis due to possible bias in some of the published series (e.g. use as normalization control in CGH experiments).

The analysis was carried out separately for gains and losses and the data was filtered for each window size i.e. the copy number alterations lesser than or equal to a particular window size.

2.3 Statistical Tests

There were multiple statistical tests that were used to analyze the data. The tests are summarized as below-
2.3.1 Binomial test

It is an exact test of the statistical significance of deviations from a theoretically expected distribution of binary data. The null hypothesis \( H_0 \) is that probability \( p \) is equal to some specified \( p \). It requires random sample of size \( n \) (or a sequence of \( n \) mutually independent trials) where each observation is categorized as “success”/“1” or “failure”/“0”, with \( p = \text{prob (success)} \) remaining constant for all trials. If we let \( s \) equal the number of successes, then \( s \) has a binomial \((n, p^*)\) distribution under \( H_0 \) [10]. Binomial formula is stated as follows:

\[
f(s; n, p) = \sum_{s=0}^{n} \binom{n}{s} p^s (1-p)^{n-s}
\]

In our analysis we used one-tailed binomial test to obtain a set of statistically significant genes from a background of large number of genes based on their frequency of occurrence in all the tumour samples. The null hypothesis was \( \alpha \leq 0.05 \) which is rejected only if the sample proportion is much greater than 0.05. The alternative hypothesis was \( \alpha \geq 0.05 \).

2.3.2 Fisher's exact test

Fisher's exact test is a statistical significance test used for categorical data given by classification in a form of contingency table. It is used to examine the significance of the non-random associations (contingency) between the two kinds of classification.

A matrix is formed in which the entries represent the number of observations for each category. Row sums, column sums and the total sum are calculated. The conditional probability \( (P_{\text{cut-off}}) \) of getting the actual matrix given the particular row and column sums is also calculated which act as cut-off and is a multivariate generalization of the hypergeometric probability function. Then, all possible matrices of nonnegative integers consistent with the row and column sums are calculated. For each one, the associated conditional probability is calculated, where the sum of these probabilities must be one [11].

In our analysis, the fisher matrix consisted of observed values of genes that are enriched or present in pathway database or both.

<table>
<thead>
<tr>
<th>Enriched Genes</th>
<th>Not Enriched Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Pathway</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>b</td>
</tr>
</tbody>
</table>
2.3.3 FDR Adjustment

A p-value of 0.05 determining the significance of an experiment indicates that there is a 5% chance that a result is false positive which is acceptable for one test. But doing a large number of tests can result in many false positives. This is known as multiple testing problem. The False Discovery rate approach determines adjusted p-values for each test by estimating the fraction of the false positives and also controls the number of false discoveries in those tests that is a significant result. For example as a p-value of 0.05 indicates that 5% of all tests will result in false positives, an FDR adjusted p-value of 0.05 would imply that 5% of the significant tests will result in false positives, which is much less than the first [12].

In our analysis, multiple testing was dealt with while using the binomial probability function on the data for different tumour samples and genes associated with them. This statistical procedure yielded a set of significantly altered genes where the number of false positives is to be kept as small as possible. This drove motivation to apply FDR correction on our set of genes to get the most accurate set of genes. Assuming that the true null hypotheses’ p-values are independent and uniform random variables i.e. the hypothesis tests are not correlated, the FDR was calculated using Benjamini and Hochberg method [13].
2.4 Methodology

The methodology used in the work is detailed as below.

2.4.1 Frequency based gene enrichment

It involves identifying significantly altered genes with the use of Binomial Test specifying a confidence interval of 95%. In the analysis, genes with total frequency of occurrence across all samples as 0 after making binary matrices were removed from the analysis to ensure more accuracy of results. The steps are shown below in Fig 2.4.

![Diagram showing steps for frequency based gene enrichment]

**Figure 2.4 – Steps for frequency based gene enrichment**

2.4.2 Copy number data coverage

The frequency of a gene encompassed by copy number alterations was determined for various genomic intervals where the intervals were defined for the size of copy number alteration. The window sizes considered were 100kbp, 200kbp, 500kbp, 1mbp, 2mbp, and 5mbp.

The analysis was carried out separately for gains and losses and the data was filtered for each window size i.e. the copy number alterations lesser than or equal to a particular window size. The structure of the character matrix with gene frequencies for each window size is illustrated in Fig 2.5 and 2.6.
A partial overlap was considered i.e. gene is considered to be a part of copy number alterations if either its end position or start position lies in the copy number alteration specific to a chromosome, then that particular gene is said to be hit by a copy number alteration.

In order to identify a set of significantly altered genes from the set of 20207 genes, binomial testing was used. Also, the significantly altered genes that are present in copy number alteration less than 1mbp in size are then used to detect the copy number
variations again considering the partial overlap. The steps to do this analysis are illustrated in Fig 2.7.

**Figure 2.7 – Steps for finding copy number coverage in Ensembl genes**

2.4.3 Pathway Enrichment Analysis: Classical Approach

After obtaining a list of genes involved in a given dataset, the next step is to map these genes to known pathways/Gene Ontology terms and identify pathways overrepresented in a given set of enriched genes. With our list of significantly altered genes, we mapped them to database for pathways from nature pathway interaction database (http://pid.nci.nih.gov/), Reactome, BioCarta and Kegg consisting of a total of 660 pathways. The pathway information was obtained using Graphite R Package [16]. The steps are illustrated below in Fig 2.8.
2.4.4 Pathway Enrichment Analysis: Permutations Approach

This approach for pathway enrichment is simply adopted to get the frequency independent list of significant pathways in each type of data [14]. In this approach, a new gene matrix is formed where any gene if mutated or altered is set as 1 and otherwise 0. Genes are represented in columns and rows as samples. The genes in the matrix are obtained from the background of ensemble genes common with the genes present in the data. Any gene if altered is set to 1.

Before applying permutations the pathway score for each pathway is calculated for the data which is termed as observed pathway score. The permutations are then applied to original matrix by changing the position of gene names in the columns and therefore changing the occurrence of mutation in a particular gene. The steps are illustrated below in Fig 2.9.
Figure 2.9 – Steps for pathway enrichment analysis - permutations approach
Results and Discussions

In this chapter, the results and discussion generated by this work are presented.

3.1 Integrative Analysis involving various kinds of mutational data obtained from ICGC data

Three types of mutation data i.e. methylation, point mutation and copy number alteration were used for carrying out frequency based gene and pathway enrichment analysis.

All the three data types were analyzed to get an idea about the amount of data present and how it is distributed across different cancer types. Dataset summary for the data from ICGC is illustrated below in Figure 3.1. The x-axis represents the different cancer types and the y-axis represents the type of data which is present for the same. e.g. for cancer type Liver_Cancer-NCC-JP has only simple mutation (SNP) and Liver_Cancer-INCA-FR has simple mutation(SNP) and CNA data.

Figure – 3.1 Summary of data available for each cancer type
The methylation, copy number alteration and simple mutation data were separately plotted to see the general distribution of data across all types of cancers. The plots of different data types are illustrated in Fig 3.2, 3.3 and 3.4

When focused on point mutation, there were very few non-synonymous mutations in comparison to other mutation types. Other types include: 3prime utr, 5prime utr, complex indel, complex substitution, frameshift coding, inframe indel, intergenic, splice site and synonymous coding.

![Figure 3.2 Distribution of simple mutation data across 17 cancer types](image)

In methylation data, many values were lying in the range of 0 - 0.2 and based on the cut-off of 0.1 many genes from the data came out to be hypomethylated where as fewer values were in the range of 0.7 – 1 as illustrated in Fig – 3.3. In the given figure, the x-axis represents methylation value and y-axis represents the density.
In copy number alteration data, there was a major percentage of the copy neutral LOH as shown in Fig 3.4. They were left from our analysis and only biallelic gains and losses were considered.
3.1.1 Gene enrichment

In this analysis, main focus was to look for genes that are frequently altered across all the three types of mutations independent of cancer type. These genes would be referred to as enriched genes in this work. The numbers of enriched genes (genes found significantly altered) in all the three types of data are summarized in Table 3.1.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Number of Samples</th>
<th>Total number of genes present</th>
<th>Number of enriched genes (at 95% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumour</td>
<td>1651</td>
<td>14408</td>
<td>Hypermethylated: 2363</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hypomethylated: 5777 (Total)</td>
</tr>
<tr>
<td>Controls</td>
<td>66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methylation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copy number alterations</td>
<td>207</td>
<td>51473</td>
<td>Gain: 1990</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Loss: 255 (Total)</td>
</tr>
<tr>
<td>Simple mutations</td>
<td>1070</td>
<td>47703</td>
<td>102</td>
</tr>
</tbody>
</table>

Table – 3.1 Number of enriched genes in each data type

It was more interesting was to see the intersection between these data types i.e. the genes found significant in all the three and also in pairs of two. Genes overlapping between all the three types are summarized below in Table 3.2.

<table>
<thead>
<tr>
<th>Ensembl Gene ID</th>
<th>Gene Information [21] [22]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000153707</td>
<td>Protein tyrosine phosphate</td>
</tr>
<tr>
<td>ENSG00000198626</td>
<td>Cardiac muscle ryanodine receptor-calcium release channel</td>
</tr>
<tr>
<td>ENSG00000127914</td>
<td>A-kinase anchor protein 9</td>
</tr>
<tr>
<td>ENSG00000141646</td>
<td>Mothers against decapentaplegic homolog 4 (Deletion target in pancreatic carcinoma 4)</td>
</tr>
<tr>
<td>ENSG00000168702</td>
<td>Low-density lipoprotein receptor-related protein 1B</td>
</tr>
<tr>
<td>ENSG00000178568</td>
<td>Receptor tyrosine-protein kinase erbB-4</td>
</tr>
</tbody>
</table>

Table - 3.2 Genes affected by all the three types of data

The intersection of all the three data types was very small but in comparison, the intersection between the pairs was much larger than the intersection of all three. The results are summarized in Table 3.3 and also in the form of a Venn diagram in Fig 3.5.
<table>
<thead>
<tr>
<th>Gene Overlaps</th>
<th>Number of enriched genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Mutation, Methylation and Copy number alterations</td>
<td>6</td>
</tr>
<tr>
<td>Methylation and Simple Mutations</td>
<td>39</td>
</tr>
<tr>
<td>Methylation and Copy number alterations</td>
<td>116</td>
</tr>
<tr>
<td>Copy number alterations and Simple mutations</td>
<td>17</td>
</tr>
</tbody>
</table>

Table – 3.3 Number of genes overlapping between different types of data

Figure – 3.5 Frequency based gene enrichment overlap in all the three types of data

Another alternative for the integrative analysis was to combine all the genes from 3 types of data and then carrying out a frequency based enrichment using the Binomial Theorem.

Using the Binomial Test approach, it was found that in a total of 52894 genes across 2928 samples, 3051 genes were enriched and for those set of genes 398 pathways came out to be significant.

### 3.1.2 Comparison with Cosmic cancer database

Cosmic stands for Catalogue of Somatic Mutations In Cancer (http://www.sanger.ac.uk/genetics/CGP/cosmic/add_info/). It is a freely available online source of somatic mutations and related details found in human cancers. Data is derived from papers in the scientific literature and large scale experimental screens from the Cancer Genome Project at the Sanger Institute. It includes the following:

- Samples which have been found to be negative for mutations during screening therefore enabling frequency data to be calculated for mutations in different genes in different cancer types.
- Samples which include benign neoplasms and other benign proliferations, in situ and invasive tumours, recurrences, metastases and cancer cell lines.
In our analysis, the set of significantly altered genes was compared with a total of 467 genes found in cosmic cancer database [18] that have been reported as targets of various types of mutations. The results are summarized in Table 3.4.

<table>
<thead>
<tr>
<th>Gene overlaps with COSMIC database</th>
<th>Number of enriched genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methylation</td>
<td>266</td>
</tr>
<tr>
<td>Copy number alterations</td>
<td>8</td>
</tr>
<tr>
<td>Simple mutations</td>
<td>20</td>
</tr>
<tr>
<td>Overlap of all the 4</td>
<td>2 (Details below)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene Information [21] [22]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000127914</td>
<td>A-kinase anchor protein 9</td>
</tr>
<tr>
<td>ENSG00000141646</td>
<td>Mothers against decapentaplegic homolog 4 (Deletion target in pancreatic carcinoma 4)</td>
</tr>
</tbody>
</table>

Table - 3.4 Number of genes found in Cosmic Database

The results are summarized in a Venn diagram in Fig 3.6 showing the overlap of cancer genes with the genes in cosmic database.

![Venn diagram](image)

Figure – 3.6 Comparison with Cosmic cancer database

### 3.1.3 Pathway Enrichment Analysis:

The pathway enrichment analysis was carried out on the significant set of genes separately for each data as well as for the combined set of genes from all the three types of data. Both
the classical approach using Fischer’s test and frequency independent permutations approach have been used.

For the combined set of genes **389 pathways** were found to be enriched at a cut-off of 0.05 for p-value. Various types of pathways were found enriched for each data type as well as the common pathways affected between all the data types have been found to be crucial in cancer development. The pathway information are summarized in table 3.5 and 3.6

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Total number of pathways</th>
<th>Number of pathways enriched</th>
<th>Classical Approach (Fischer’s test)</th>
<th>Permutations Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methylation</td>
<td>660</td>
<td></td>
<td>227</td>
<td>430</td>
</tr>
<tr>
<td>Copy number alterations</td>
<td>660</td>
<td></td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>Simple mutations</td>
<td>660</td>
<td></td>
<td>18</td>
<td>167</td>
</tr>
</tbody>
</table>

Table - 3.5 Number of pathways enriched for each data type

<table>
<thead>
<tr>
<th>Pathway overlap</th>
<th>Number of pathways</th>
<th>Pathway Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>All the three datatypes</td>
<td>1</td>
<td>Axon guidance</td>
</tr>
<tr>
<td>Methylation and Copy number alterations</td>
<td>3</td>
<td>Axon guidance, Hepatitis C and Jak-STAT signaling</td>
</tr>
<tr>
<td>Copy number alterations and Simple Mutations</td>
<td>1</td>
<td>Axon guidance</td>
</tr>
<tr>
<td>Methylation and Simple Mutations</td>
<td>15</td>
<td>(Listed below)</td>
</tr>
</tbody>
</table>


Table - 3.6 Commonly affected pathways by different types of mutations found using classical approach (Fischer’s test)

The numbers of pathways significantly altered found using the classical approach has been illustrated in Fig 3.7.
Therefore, after the frequency based gene and pathway enrichment analysis, we found out that there was no significant overlap between three data types and the genes targeted by different mutations differ a lot and only very few genes are affected by all the three types of mutations studied. As compared to the number of genes overlapping between all three data types, the number of genes intersecting between two types was significantly higher but even this was not very large as compared to the total number of genes affected by individual mutations.

Similarly, for the pathway enrichment analysis, the number of pathways commonly affected by three types of mutations was very small. On the other side, still some useful insights could be gained about the pathways that could be established as preferentially targeted by mutations to affect and result in cancer development. Also, the results generated by the new approach adopted for frequency independent pathway enrichment were not conclusive as the number of enriched pathways was very large as compared to the classical approach and it might be overestimating the significance of a pathway. This approach needs to be refined or run for very large number of permutations to get better results. This lies in the future scope of the project to achieve more accurate pathway enrichment.
3.2 arrayMap Data Analysis

In this analysis, the copy number coverage for different sizes of copy number alterations were calculated where the frequency of every gene being hit by a copy number alteration of particular size was calculated. The various sizes considered were - 100kbp, 200 kbp, 500kbp, 1Mbp, 2Mbp, 5Mbp and greater than 5 Mbp. Higher the frequency number for a particular gene implies that it was hit more frequently by copy number alterations of a particular size. The number of significantly hit genes was calculated by doing binomial testing and summarized in table 3.7.

The number of genes has been represented separately for copy number losses and gains. The total number of genes from Ensemble was 22207. For the set of significantly hit genes, pathway enrichment analysis was also carried out and many pathways were found enriched for several sizes of copy number alterations. The numbers of pathways enriched for each window size are summarized in table 3.7.

<table>
<thead>
<tr>
<th>Window Sizes for filtering copy number alterations data</th>
<th>Enriched Genes in each window size</th>
<th>Number of pathways enriched in each window size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gain</td>
<td>Loss</td>
</tr>
<tr>
<td>100 kbp</td>
<td>877 (12%)</td>
<td>1001 (12%)</td>
</tr>
<tr>
<td>200 kbp</td>
<td>1149 (15%)</td>
<td>1260 (15%)</td>
</tr>
<tr>
<td>500 kbp</td>
<td>1829 (25%)</td>
<td>1799 (21%)</td>
</tr>
<tr>
<td>1 mbp</td>
<td>2880 (39%)</td>
<td>2391 (28%)</td>
</tr>
<tr>
<td>2 mbp</td>
<td>4006 (54%)</td>
<td>3752 (43%)</td>
</tr>
<tr>
<td>5 mbp</td>
<td>5463 (73%)</td>
<td>5226 (61%)</td>
</tr>
<tr>
<td>&gt;5 mbp</td>
<td>6982 (94%)</td>
<td>8335 (97%)</td>
</tr>
<tr>
<td>0 (No window)</td>
<td>7437 (100%)</td>
<td>8631 (100%)</td>
</tr>
</tbody>
</table>

Table - 3.7 Copy Number Coverage in genes and number of pathways specifically hit by a copy number alteration of particular size

The data represented in table 3.7 is illustrated below in Fig 3.8 and Fig 3.9. The number of enriched genes increases as the copy number alteration window size is increased as the genes falling in the lower size window also fall in the category for higher size window. There were total of 176 genes that were significantly hit by copy number gains and 159 genes that were significantly hit by copy number losses, both less than 1mbp of size. Out of these genes 13 and 12 genes are found annotated in COSMIC database respectively for gains and losses. The top 10 genes hit by copy number gains less than 1mbp size based on their hit frequency were also computed and out of these genes 2 were found to be annotated in Cosmic Database. The list of genes is shown in the table 3.8.
<table>
<thead>
<tr>
<th>Ensemble Gene ID</th>
<th>Gene Information [21] [22]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000171862</td>
<td>Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN (Mutated in multiple advanced cancers)</td>
</tr>
<tr>
<td>ENSG00000138293</td>
<td>Nuclear receptor coactivator 4</td>
</tr>
<tr>
<td>ENSG00000170178</td>
<td>Homeobox protein Hox-D12</td>
</tr>
<tr>
<td>ENSG00000175879</td>
<td>Homeobox protein Hox-D8</td>
</tr>
<tr>
<td>ENSG00000128709</td>
<td>Homeobox protein Hox-D9</td>
</tr>
<tr>
<td>ENSG00000128710</td>
<td>Homeobox protein Hox-D10</td>
</tr>
<tr>
<td>ENSG00000196436</td>
<td>Nuclear pore complex-interacting protein-like 2</td>
</tr>
<tr>
<td>ENSG00000109339</td>
<td>Mitogen-activated protein kinase 10</td>
</tr>
<tr>
<td>ENSG00000140839</td>
<td>C-type lectin domain family 18 member B</td>
</tr>
<tr>
<td>ENSG00000081052</td>
<td>Collagen alpha-4(IV) chain</td>
</tr>
</tbody>
</table>

Table 3.8 List of genes frequently hit by copy number gains

To study the intersection between copy number alterations and variations, we considered the genes which were hit by copy number alterations less than or equal to 1mbp size. The total number of genes significantly hit by copy number alterations less than or equal to 1mbp size are: Gain - 2024 and Loss – 1992.

![Number of significantly altered genes by different sizes of copy number alterations](image)

Figure – 3.8 Number of significantly altered genes by different sizes of copy number alterations
Figure – 3.9 Number of significantly altered pathways by different sizes of copy number alterations

For these set of genes that are hit, copy number variation data obtained from Database of Genomic Variants from TCGA (http://projects.tcag.ca/variation) and ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) from EMBL-EBI, was mapped. In copy number gains, out of 2024 genes, 401 genes were found to be targeted by copy number variations. The number of copy number variations encompassed in each gene varied gene to gene from 1-43 copy number variations found in a gene. Out of the 1992 genes in copy number losses, 534 genes were found to be hit by copy number variations.

Thus, this analysis was carried out to see the size based preference of copy number alterations while targeting genes. Genes targeted more often by smaller copy number alterations and the genes that are targeted by larger copy number alterations could be easily deduced. Similarly, the pathways affected by these alterations were also deduced.

This information can be useful in distinguishing CNA’s (somatic mutations) with CNV’s (germline mutations) [6].
Chapter 4

Conclusion and Future Work

In this work, an integrative analysis on three different types of cancer genome data i.e. copy number alteration, DNA methylation and simple mutation data from ICGC was carried out. To perform an integrative analysis, the significantly altered genes were first identified for cancer related mutations. Computing the genes overlap helped identify genes which are affected by all the three types of mutations irrespective of cancer type. Since the gene overlap was not significant between all the three types of mutations, it can be concluded that there is no strong indication of different types of cancer related mutations targeting common genes. Similarly from pathway enrichment analysis, it was seen that different types of mutations affect different pathways which imply affecting different functional entities that might have some common regulation.

Another focus of this work was studying copy number alteration data from arrayMap database. Copy number coverage of genes for different sizes of copy number altered regions was studied and it was found that there are certain genes that are specifically altered by a particular size of copy number alteration. Out of these genes, the ones hit specifically by copy number altered regions of less than 1mbp size were considered and mapped for detecting those which are also hit by copy number variations. It was observed that some of the genes hit by copy number alterations both gains and losses, were also in the regions of germline copy number variations. Thus, it can be concluded that the genes also hit by copy number variations might not represent cancer associated genes.

The future scope of this work would be to analyse and identify pathways targeted by single or multiple gene events. More cancer types can also be included in the future as new data for different cancer types becomes available over time. Moreover, the expression data from cancer patients can be included in the analysis and studied for possible effect on the expression of genes affected by different types of mutations. Collaboration with hospitals to use the identified targets in finding more specific drugs for treating cancer would become a significant contribution to the field.
Bibliography


