Pattern Recognition in Clinical Data

Dual Degree Project

Saket Choudhary
Guide: Prof. Santosh Noronha

Department of Chemical Engineering

IIT Bombay

Certificate

Indian Institute of Technology, Bombay

This Dual Degree Project titled Pattern Recognition in Clinical Data by Saket Choudhary(09D02007) was prepared under my guidance and may be accepted for evaluation.

Prof. Santosh Noronha
Chemical Engineering Department
Date: June 26, 2014
Place: IIT Bombay, Mumbai

Declaration

I, Saket Choudhary, Roll No. 09D02007 understand that plagiarism is defined as any one or the combination of the following:

1. Uncredited verbatim copying of individual sentences, paragraphs or illustrations (such as graphs, diagrams, etc.) from any source, published or unpublished, including the internet
2. Uncredited improper paraphrasing of pages or paragraphs (changing a few words or phrases, or rearranging the original sentence order).
3. Credited verbatim copying of a major portion of a paper (or thesis chapter) with- out clear delineation of who did or wrote what.

I have made sure that all the ideas, expressions, graphs, diagrams, etc., that are not a result of my work, are properly credited.
I affirm that no portion of my work can be considered as plagiarism and I take full responsibility if such a complaint occurs. I understand fully well that the guide of the project report may not be in a position to check for the possibility of such incidences of plagiarism in this body of work.

Saket Choudhary
09D02007
Date: June 26, 2014
Place: IIT Bombay, Mumbai

Approval Sheet

Indian Institute of Technology, Bombay

This dissertation titled "Pattern Recognition in Clinical Data" by Saket Choudhary(09D02007) is approved for the degree of Master of Technology in Chemical Engineering.

Supervisor:

Prof. Santosh Noronha
Chemical Engineering Department
Date: June 30, 2014
Place: IIT Bombay, Mumbai

Chairperson:

Prof. P.V Balaji
Department of Bioscience and Bioengineering
Date: June 30, 2014
Place: IIT Bombay, Mumbai

Examiner:

Prof. Sharad Bhartiya
Chemical Engineering Department
Date: June 30, 2014
Place: IIT Bombay, Mumbai

Acknowledgements

To me a guide is one who keeps you motivated, gives you the freedom to explore, and ultimately tries to bring the best out of you. To me my guide Prof. Santosh Noronha, has been an incessant source of knowledge, without whose ideas, I would have probably given up the project long ago. Like most of the people, I have had mid-life breakdowns, he however has always kept me motivated.

I would also like to thank Dr.Rita Mulherkar, Advanced Centre for Treatment, Research and Education in Cancer (ACTREC) and Dr. Neelam Shirsat for valuable inputs.

I am greatly indebted to Dr. Parvez Syed and Dr. Sanjeeva Srivastava for introducing us to the microarray problem, which has contributed to enormously to the project in its later stage.

And immense thanks to the mini sever that sits back in the Protein Engineering Lab. I would not have survived sans a laptop otherwise.

Abstract

Biology's next big problem is, too much data to handle. With the advent of Next Generation Sequencing techniques, the rate of data production far exceeds our ability to make sense out of it. There has been a shift from the classical hypothesis-driven research to an *-omics* based research.

-omics focuses on the global analysis of a system as opposed to a individual component based analysis as in classical hypothesis driven research. Besides collecting the *-omics* data, there is an ever increasing collection of clinical data. The problem no longer lies in getting the DNA sequenced, but as to what inferences can be drawn from such a study. *-omics* and classical hypothesis driven research can complement each other, only if we overcome the bottleneck of *analyzing* this data. Too much data, not be too informative.

Here, we try to tackle a collection of problems from the *genomics* and *functional genomics* domains with an aim to tackle the bottlenecks by making use of mathematical tools. Mathematics as a *science* and as a *tool* can help mitigate some of these bottlenecks.

We address five different problems here. The first problem discusses how information from DNA sequencing can be used for characterizing the nature of mutations in Cancer. A unified workflow is presented which can aid biologists in prioritizing the damaging nature of these mutations.

In the second discussion, we demonstrate the earlier known presence of Human papillomavirus in Cervical cancer datasets.

The third problem presents a case study of benchmarking two alignment algorithms.

The fourth problem focuses on obtaining a set of bio-markers, from a proteomics microarray experiment, that can be used for prognosis of cancer.

With the fifth problem we present visualisation tools for biological data, that can be used in next generation sequencing and discovery domains for data analytics.

Contents

List of Figures

Chapter 1

Introduction

DNA or **D**eoxyribo **N**ucleic **A**cid is a molecule found in all eukaryotes including humans that carries hereditary information[5] from one generation to another. Nearly every cell in the human body has the same DNA sequence. DNA that is passed by the mother to its offspring in mitochondria is referred to as mtDNA(short for 'mitochondrial DNA').

DNA is essentially made up of four bases : Adenine(A), Guanine(G) (Purines) and Thymine(T), Cytosine(C) (Pyrimidines). Every base is attached to a sugar molecule(deoxyribose sugar) and a phosphate molecule resulting in a **nucleotide**. These nucleotides are bound to each other and the sequence of these bonds determines the traits of an individual.

Structurally DNA is a packed as a double stranded helix. The two strands are complementary. Adenine(A) on one strand always has a corresponding Thymine(T) on the other strand and similarly Guanine has a corresponding Cytosine on the other strand. Owing to the naming convention associated with the **C**(Carbon) atoms in the nucleotide rings, there is a 5' (pronounced as 'five prime') and a 3'(pronounced as 'three prime') end. Figure 1.1 shows how two complementary strands are arranged.

```
5'-ATGCCGTAATTGGCC-3'
3'-TACGGCATTAACCGG-5'
```

Figure 1.1: Complementary strands of DNA

The biological information in any individual is 'encoded' by the DNA which is divided into discrete units called *genes*. The entire hereditary information stored in the set of 23 chromosomes besides the mtDNA is called the 'Human Genome'.

The Genome is *packaged* in the form of *chromosomes*. Chromosomes are made up of DNA and proteins and is like a packet containing a chunk of the genome.

Humans have a set of 22 autosomes and one pair of sex chromosome. Autosomes are same in males and females while sex chromosomes of an individual determine the sex. Females have two copies of the X chromosome as XX, while males have one X and one Y chromosome.

Genes are the *instructing* machinery for protein manufacturing. Every person has two copies of gene one inherited from each parent. This can be easily seen from the fact that during zygote formation one set of chromosome comes from the father while the other set of the chromosomes comes from the mother.

The largest chromosome of any organism is called Chromosome 1, the next largest is called Chromosome 2 and so on. Different chromosomes contain different genes. Each chromosome contains many genes and genes are essentially made up of DNA and 'code' for proteins.

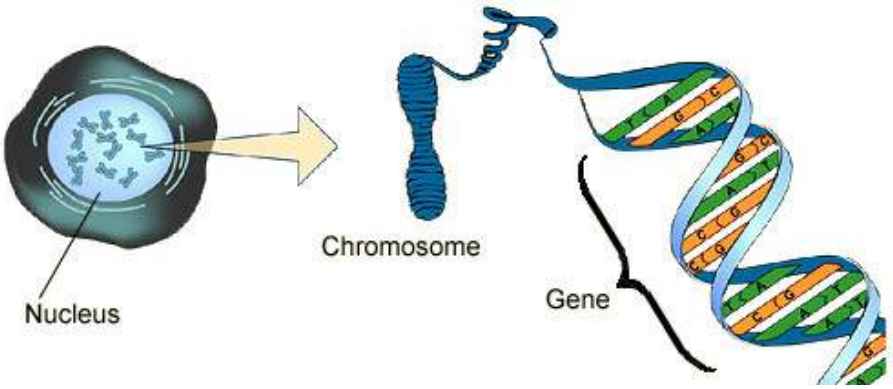


Figure 1.2: Chromosome,DNA and genes, http://www.bbc.co.uk/schools/gcsebitesize/science/add_aqa_pre_2011/celldivision/celldivision1.shtml

1.1 DNA Sequencing

Deciphering DNA sequences is essential for virtually all branches of biological research. Determining the sequence of bases gives us insight into the genetic variations associated. These genetic variations are in turn associated with level of protein expressions. These level of protein expressions in turn govern the susceptibility to diseases, the effectiveness of drugs and the probability of passing diseases down the generation. A better understanding at the molecular level will ensure specific therapeutic targets.

The Genome of an organism is the *blueprint* of how an organism functions. Genome sequencing is a pathway for finding genes more easily and quickly. De-

ciphering the gene map and the basic sequences that govern them correctly, will aid in early prognosis of diseases besides zeroing down upon a set of specific therapeutic targets for drug delivery.

1.1.1 Sanger Sequencing

Making use of dideoxy nucleoside triphosphates, Sanger designed an easy and reliable protocol to sequence the DNA [38]

Nucleoside: Base (A,T,G,C) bound to ribose or deoxyribose sugar. Chain of nucleosides forms a DNA sequence
Nucleoside Triphosphate: Nucleoside 1.3 bound with three phosphate molecules

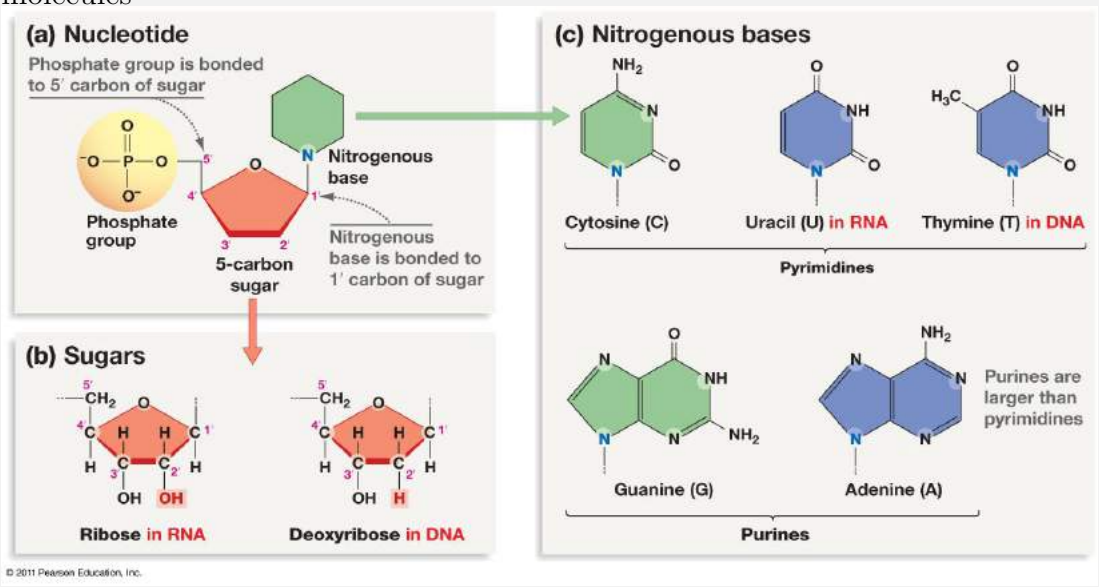


Figure 1.3: Nucleosides, http://www.uic.edu/classes/bios/bios100/lectures/dna.htm

DNA replication can be initiated in-vivo by providing it with the necessary dNTPs and a starting sequence called primers, and some enzymes. The enzymes(like DNA polymerase) catalyze DNA replication and add new nucleotides to the end of primer sequences such that the synthesizes strand is complementary to the original strand.

In Sanger Sequencing, DNA replication is initiated with the help of a DNA primer, however instead of a dNTP, a ddNTP is used. ddNTP is like regular DNA sans the 3 Hydroxyl(-OH) group , thus if added to the end of DNA, there is no way of further extending the chain. The key point lies in the fact that most of

Deoxyribose Nucleoside Triphosphates: Nucleoside triphosphates containing the deoxyribose sugars

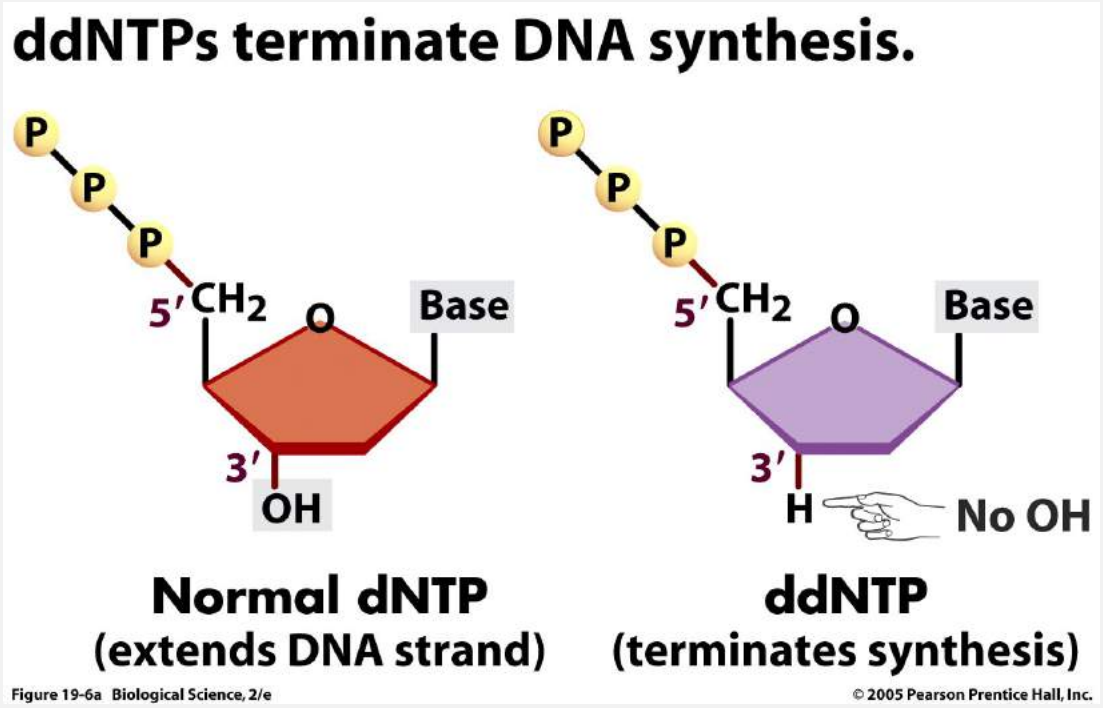


Figure 1.4: ddNTP and dNTP http://www.uic.edu/classes/bios/bios100/lectures/techniques.htm

the nucleotides are normal(dNTPs) and only some of them are ddNTPs. Either the primer is radio-labeled or the ddTPs are radio-labeled so that they can later be detected on a gel.

Using restriction enzymes the sequence is fragmented at random positions and then these sequences are pooled along with dNTPs, ddNTPs and enzymes required for replication. The first step generally involves PCR amplification of the DNA. As the DNA fragments are random and the chain may terminate at any random position, multiple copies of DNA are required. This ensures that each base site is covered more than once in the short and long fragments or reads, as we refer to them henceforth.

When we take multiple copies of the sequence and a ddNTP such as T, i.e. a ddTTP then most sequence will end in a dideoxy T But what is not known is where this T would be incorporated along the length. The sample will essentially be a mixture of long and short reads with 'T' occurring at the ends, at different lengths based on the site it gets incorporated. This is just an example for 'T', same applies for 'A', 'C' and 'G'.

As seen from 1.5 all the sequences started with one common primer, but all of them ended in a 'T'. The position at which each fragments terminates is random. Say we start with billion PCR fragments, possibly a million of them will end in a 'T'. After annealing the DNA, denaturing it into single strands, a primer is used with its 3' end lying next to DNA sequence of interest. Either the primer or the ddNTP is radio-labeled so that it can be detected later. This sample is then run on a gel with four different lanes labeled 'G','A','T' and 'C'.

The DNA sequence can be *read off* in the form of a ladder from the gel starting from the bottom right going to the top left as showing Figure 1.6. We start from the bottom right since the sequence at the bottom of the gel is the smallest(lighter) while the one at the top will be the largest(heavier). Since there are four lanes, each lane will have a sequence terminating in that particular nucleotide, and hence all fragments will be of different size.

1.1.2 Next Generation Sequencing

The traditional Sanger Sequencing was automated, giving rise to the 'First Generation Sequencing' [33] by using fluorescent labeled ddNTPs by labeling the four ddNTPs with fluorescent dyes of different wavelengths. This method makes sequencing easier and cheaper.

Next Generation Sequencing is a used to describe a collection set of technologies that emerged around 2005 which enabled high-throughput sequencing at a cheap cost [40]. Next Generation Sequencing(here after referred as NGS) are essentially built on top of the idea of Automated Sanger Sequencing, where thee

Primer Used : GAATGTCCTTTCTCTAAGTCCTAA

5'-GAATGTCCTTTCTCTAAGTCCTAA**T***
3'-CTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATC-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAG**T***
3'-CTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATC-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCC**T***
3'-CTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATC-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGA**T***
3'-CTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATC-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGG**T***
3'-CTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATC-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGGTAC**T***
3'-CTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATC-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGGTACT**T***
3'-CTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATC-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGGTACTTC**T***
3'-CTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATC-5'

Figure 1.5: Terminating chains with radio-labeled ddNTPs

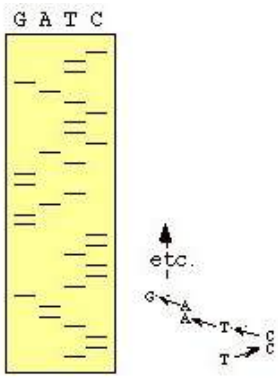


Figure 1.6: Autoradigram, Source: http://www.bio.davidson.edu/courses/molbio/molstudents/spring2003/obenrader/sanger_method_page.htm

DNA to be sequenced is first broken down into a library of small fragments by cutting the DNA strand at random positions using specific reagents(restriction enzymes). These small fragments can now be sequenced individually in parallel. These small fragments are referred to as 'reads'. These reads can then be re-assembled using a reference genome to get a contiguous original sequence. This approach is commonly referred to as **Shotgun Sequencing** [45]

Whole Genome Shotgun Sequencing: Whole Genome is broken down into small fragments. Since the cleavage takes place at random positions, the short fragments are resequenced multiple times in order to come up with a consensus sequence. It might however happen that certain regions remain un-sequenced(owing to the fragmentation being random) and will turn out as gaps, as indicated by Figure 1.7

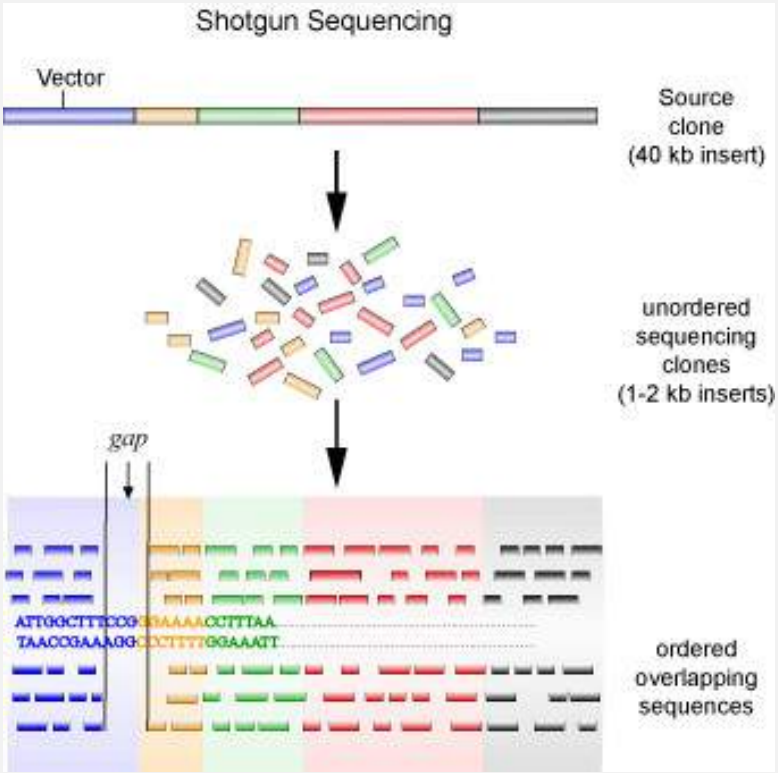


Figure 1.7: Shotgun Sequencing , Source: http://www.scq.ubc.ca/genome-projects-uncovering-the-blueprints-of-biology/

1.2 NGS Data Formats

NGS sequencers generate output as a FastQ file. FastQ file contains raw read sequences along with a quality score value referred to as the **Phred Quality Score**[14]

@HWI-ST1097:104:D13TNACXX:4:1101:1715:2142 1:N:0:CGATGT
GCGTTGGTGGCATAGTGGTGAGCATAGCTGCCTTCCAAGCAGTTAT
+
=<@BDDD=A;+2C9F<CB?;CGGA<<ACEE*1?C:D>DE=FC*0BA

Value String	**Description**
HWI	the unique instrument name
ST1097	the run id
D13TNACXX	flow cell id
4	flow cell lane
1101	tile number within flow cell
1715	'x'-coordinate of the cluster within the tile
2142	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
N	Y if the read fails filter (read is bad), N otherwise
0	when none of the control bits are on, otherwise it is an even number
CGATGT	index sequence
Line 1	Read identifier, end and bar-code for the read
Line 2	Read Sequence
Line 3	Marker(Same throughout the file)
Line 4	String of ASCII-encoded base quality scores, one character per base in the sequence

Figure 1.8: FastQ format, Partially adapted from http://en.wikipedia.org/wiki/FASTQ_format

As shown in 1.8 for every base there is a corresponding quality score associated with it. A quality value Q is defined as an integer mapping of the probability score p , i.e. the probability that the corresponding base is wrong.

$$Q_{sanger} = -10log_{10}p \tag{1.1}$$

where Q_{sanger} = Quality score for Sanger formatted files

$$Q_{solexa} = -10log_{10}\frac{p}{1-p} \tag{1.2}$$

where Q_{solexa} = Quality score for files as received from Solexa sequencer.

A notable point here is that different sequencers encode the Phred Scores differently Sanger format can encode a Phred quality score from 0 to 93 (Phred+33) using ASCII 33 to 126 Starting with Illumina 1.3 and before Illumina 1.8, the format encoded a Phred quality score from 0 to 62 (Phred+64) using ASCII 64 to 126

It is important to know which platform the reads came from, since the alignment algorithm essentially depends on these score values.

A NGS pipeline involves the following steps:

1. **Quality Control Checks** Before processing starts, it is necessary filter out reads that are low quality. Pre-processing may also involve trimming the reads to filter out poly-A tails.

2. **Alignment**: Pre-processed reads are aligned to a reference sequence. Alignment is important from the point of down stream analysis

3. **Variant Call**: Determine mutations after alignment

4. **Post Processing and Annotation** Generate re-calibrated mapping scores and annotate variants to determine the associated genes.

The workflow is summarized in 1.9

The output of a pipeline is a VCF(**Variant Call Format**)[16] which is a tab delimited file that lists down the reference and allelic nucleotides at various chromosome positions.

Most of the analysis in life sciences is dependent on variant calling, and VCF is a standard format for storing the variants.

1.3 Sequencing: Why?

The Genome of an organism is the *blueprint* of how an organism functions. If we are able to decipher the chromosome maps, the basic sequences that govern

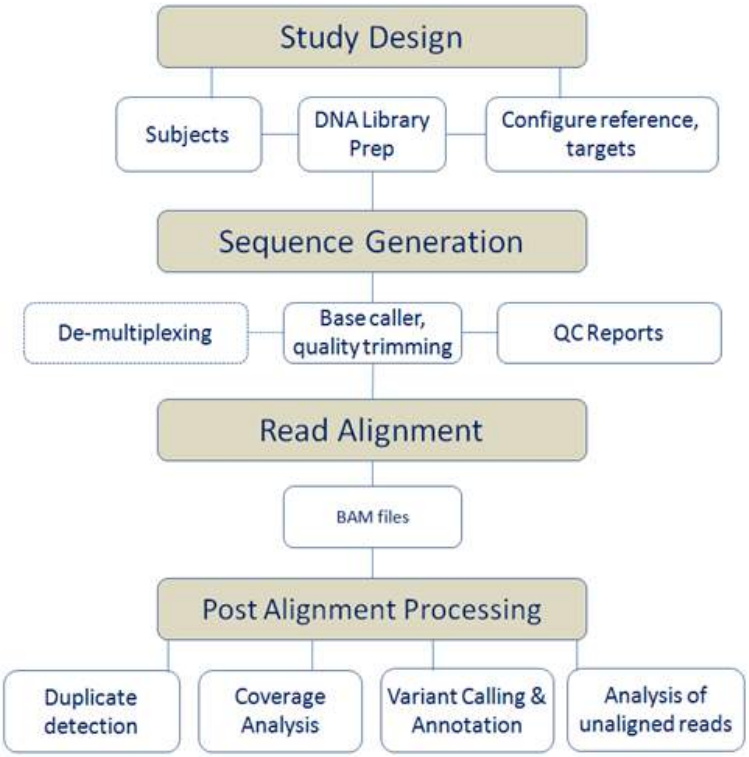


Figure 1.9: NGS Workflow,Adapted from http://cgf.nci.nih.gov/operations/bioinformatics.html

them diseases can not only be identified early but at the same time new drugs and treatments can be discovered.

At the very least Genome sequencing is a pathway for finding genes more easily and quickly. Locating genes gives insights into the protein expressions which in turn is related to tumors, drug delivery.

Chapter 2

Driver mutation identification

2.1 Driver Mutations

A **somatic mutation**, alternatively referred to as an **acquired mutation** is the set of mutations developed post zygote formation. The zygote might itself inherit mutated DNA from one or both of its parents, these are referred to as **germline mutation**.

Cancer is known to develop due to accumulation of somatic mutations [51]. However not all somatic mutations are equally important from the point of promoting tumor growth. Only a small subset of these mutations are directly involved in development and progression of cancer. Driver mutations are known to confer growth advantages to the cell besides being 'selected' positively in the tumor tissue. [46]

The problem of differentiating driver mutations from the somatic mutations has been studied using the following broad approaches[56]:

1. **Prediction of functional impact of the mutation**
2. **Machine learning approaches; classifier trained on known set of driver mutations**
3. **Difference in background mutation frequency of driver and passenger mutations**

Single Nucleotide Variation(SNV): A DNA sequence variation arising due to the assembled DNA not aligning with the reference DNA, perfectly
. Reference sequence: ATCGTAGGCTA

ATCGTAGGCTA
ATCGC*AGGCTA

At the marked position C occurs instead of an expected T, giving rise to a SNV. If this SNV inturn causes a different amino acid to be expressed this is referred to as a non synonymous mutation nsSNV.

Here we briefly discuss few of the approaches for driver mutation identification.

2.2 Polyphen2 [2]

Polyphen 2 [2] relies on functional impact prediction of the variants. Polyphen2 was primarily developed for studying the deleterious effects of non synonymous mutations. Based on eight sequence-based and three structure-based predictive feature, it applies a **naive-bayes** classifier to predict the posterior probability of a given mutation being a deleterious or damaging mutation. Though the approach is machine learning based, the set of features considered take into account the biological/functional impact of the mutations.

Out of initial 19 sequence-based and 13 structure-based features, a set of 11 features(8 sequenced based and 3 structure based) were determined using feature selection techniques. Feature selection is important as a noisy/irrelevant/redundant feature can affect the performance of the classifier.

Some of these features are:

1. **PSIC Score**: PSIC score [48] gives likelihood score of observing an amino acid at a particular position, given the substitution pattern of amino acids as in BLOSUM62 matrix.

2. **Sequence Identity to closest homologue**: Degree of closeness to the homologue carrying any amino acid different from the wild type allele

3. **Congruency to MSA**: Sequence identity for the amino acid at the given site with respect to its closest homologue in which this amino acid is observed.

4. **CpG context**: CpG context of transition matrix

5. **Change in hydrophobic propensity**

6. **Crystallographic B factor for conformational mobility**
7. **Alignment Depth**
8. **Change in amino acid volume**
9. **Existense of Pfam domain**

The steps involved in Polyphen2 are as in 2.2

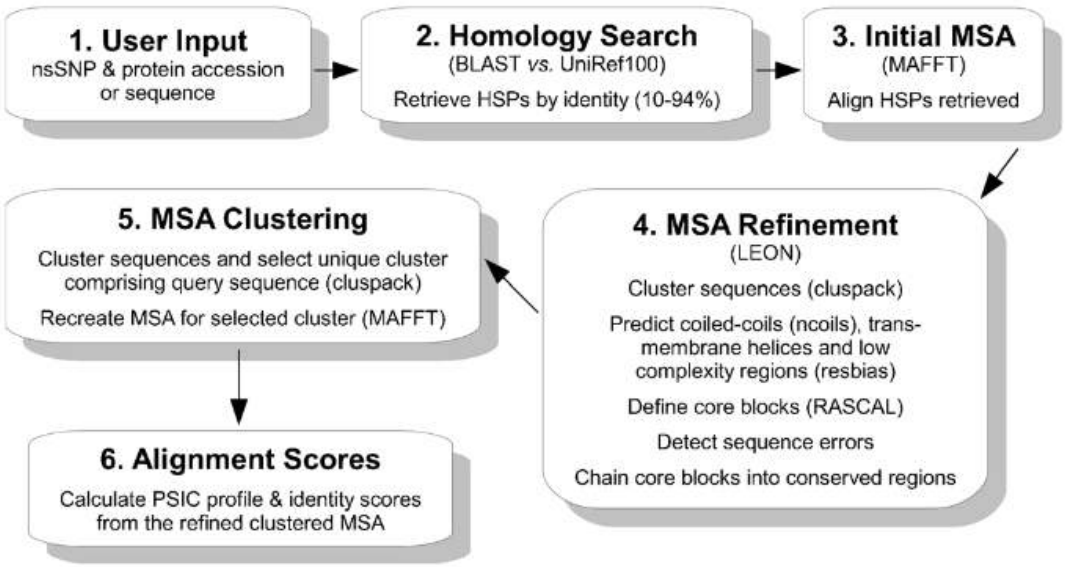


Figure 2.1: Polyphen2 Steps, Adapted from [2]

2.3 SIFT [26]

SIFT (**S**orting **I**ntolerant **F**rom **T**olerant) uses protein sequence homology to classify amino acid substitutions arising from non synonymous mutations in the DNA. Using the PSI-BLAST [3] algorithm SIFT searches for functionally related proteins in a protein database A multiple sequence alignment(MSA) is performed for all these proteins. With the MSA output, it is possible to generate a PSSM(Position Specific Scoring Matrix) where the rows list out all the available positions in the MSA and the columns list out all the 20 amino acids. Each position thus has 20 columns associated with each of the 20 amino acids

having a probability of occurrence associated with it. These rows are each divided by the corresponding maximum value in the row to obtain a SIFT scaled probability matrix. Given the mutation and hence the amino acid substitution SIFT can predict the nature of substitution(deleterious/normal) based on the corresponding probability value being lesser than a threshold. If an amino acid with higher probability(more conserved) is substituted with an amino acid with a lower probability value it is generally classified as "intolerant".

SIFT does not take into account the protein structure while assigning the probability and considers only the substitution model. In order to provide confidence values for an MSA with almost similar sequences a **conservation value** is assigned to each score such that a position where all the 20 amino acids are observed gets a zero score while the position where only one amino acid occurs is assigned a conservation score of $log_2 20$

2.4 Mutation Assessor [36]

'Causative' mutations(drivers) are simply **not** the *loss of function* mutations. In general the types of mutations can be classified into the following categories:

- **Loss of function**: Inactivate tumor suppressor proteins
- **Gain of function**: Activates normal genes transforming them to oncogenes
- **Drug Resistance Mutations**: Mutations that have evolved to overcome the inhibitory effect of drugs
- **Switch of function**: Intermediate between a loss and a gain of function mutation

The functional changes affecting a mutated protein sequence can be:

- **Change in stability** : Mutated protein might be unstable leading to lower steady state levels
- **Change in interaction with other proteins,ligands**: A mutated protein's interaction with other proteins/ligands is affected too

If a certain mutation confers an advantage to the cell in terms of replication rate, it is probably going to be selected while all those mutations that reduce its fitness have a higher chance of being eliminated from the population. This is one of the explanations behind a certain residue being conserved across MSA of homologous sequences. The hypothesis is then, of all possible tried combinations of residues among the population, the homologous

Mutation assessor algorithm builds upon SIFT algorithm by incorporating 3D structure of sequence homologs. The fact that a particular sequence has evolved by natural selection reflects the effects that it had over all levels: molecular, tissue and organ. With this hypothesis, some amino acid residues are classified as "specificity residues" based on the clustering homologous sequences and analyzing their functional specificity based on conservation of overall function. Thus, protein sub family MSA represents a model exhibiting likelihoods that a particular sequence belongs to the family. Some residues will be more frequent than others at specific positions and hence these probability values can be converted into a scoring function for predicting the functional impact of any mutation.

The entropy associated with column i is given by[36] S_i^c:

$$S_i^c = ln\frac{N!}{\Pi n_i(\alpha)} \tag{2.1}$$

$$\sum_{\alpha} = N \tag{2.2}$$

α=20 amino acids
$n_i(\alpha)$=Number of residues of type α occurring in i^{th} column
For α substituted by β amino acid, the change in entropy value:

$$\delta S_i^c(\alpha \longrightarrow \beta) = -ln\frac{n_i(\beta)+1}{n_i(\alpha)} \tag{2.3}$$

If α residue is conserved across the sequences and β is a point mutation $n_i(\alpha) >> n_i(\beta)$ and hence conservation score $\delta S(\alpha \longrightarrow \beta)$ would be high. The physical interpretation of this score is that the physical constraints (protein-protein interaction, protein-ligand interaction etc) govern the nature of residues at particular positions and hence it can be assumed that the functional impacts of these residues substitution is indirectly related to physico-chemical changes. To extend the assessment of conservation patterns, the sequences are clustered in order to create protein sub-families to obtain a conservation score at sub-family level.Clustering creates sub-family groups such that sequence diversity **within subfamilies** is minimized and the overall difference between subfamilies at specific positions is maximized.

Difference of entropy of substitution in column i of some sub-family m and the overall entropy of column i(taken as a reference):

$$\delta S_i^m = ln\frac{N^m!}{\Pi_\alpha n_i^m(\alpha)} - ln\frac{N^m!}{\Pi_\alpha \langle n_i^m(\alpha)\rangle} \tag{2.4}$$

where $n_i^m(\alpha)$: observed frequency of α amino acid at position i, and
$\langle n_i^m(\alpha)\rangle = \frac{n_i(\alpha)N^m}{N}$: Expected frequency of α at i
N^m : Number of subsequences in subfamily m
Objective function to be minimized :

$$\delta S = \sum_{i=1}^{L}\sum_{m=1}^{M} \delta S_i^m \tag{2.5}$$

Thus, the sub-families are determined using 2.5 and as with conservation score, a 'specificity score' can by calculated:

The specificity score associated with column i is given by S_i^m:

$$\delta S_i^m(\alpha \longrightarrow \beta) = -ln\frac{n_i^p(\beta)+1}{n_i^p(\alpha)} \tag{2.6}$$

where $n_i^p(\alpha), n_i^p(\beta)$ are number of α and β residues at position i in submfamily p

Using 2.3 and 2.6 a functional impact score is defined as:

$$FIS = \frac{\delta S_i^c + \delta S_i^m}{2} \tag{2.7}$$

Higher the FIS 2.7, more is the functional impact of the $\alpha \longrightarrow \beta$ substitution.

2.5 CHASM [8]

Carter et al. [8] describe how genes mutated with high frequency across a cohort of cancer samples and how analysis of **large** number of cancer samples can be helpful in driver mutation prediction. Methods based on mutation frequency can fail because genes that are mutated in a small fraction of tumors can still act as drivers.

A driver mutation arises because of *intolerable* mutation at specific residues, while passenger mutations are more like non synonymous single nucleotide polymorphisms. **Passenger mutations are neutral from the point of cancer cell fitness and hence an impact on protein can be present or absent**. nsSNP with higher minor allele frequency (MAF)have become part of the human genome and as such should contribute minimally to improving cancer cell

fitness. Hence the classifier should not be trained with passenger mutations being classified into a default category of high MAF nsSNPs.

Using Shannon entropy [39], mutual information was calculated for more than 89 predictive features and they were than ranked with feature with maximum mutual information ranked the first. Shannon entropy was chosen , rather than simple correlation coefficients as two features are not necessarily linearly associated. Not all features were selected as redundant features and noisy features can negatively impact classification.

2.5.1 Feature Selection

Starting with 80 candidate features for driver mutation identification, a list of 49 features were shortlisted to train the Random Forest classifier.

$p(X_i)$ represents the probability of occurrence of an event X_i Considering a series of events X_1, X_2, X_3...,X_n analogous 'series of packets' in communication theory , the information received at each step can be quantified on a log scale by:

$$\frac{1}{log_2(X_i)} = -log_2(p(X_i)) \tag{2.8}$$

The expected value of information from a series of events is called shannon entropy: $H(X)$:

$$H(X) = -\sum_i p(X_i)\log_2 p(X_i) \tag{2.9}$$

Mutual Information between two random variables X, Y is defined as the amount of information gained about random variable X due to additional information gained from the second, Y:

$$I(X,Y) = H(X) - H(X|Y) \tag{2.10}$$

Here:
X: Class Label[Driver/Passenger]
Y: Predictive Feature

and hence $I(X,Y)$ represents how much information was gained about the class label Y from knowledge of a feature X

Simplifying :

$$I(X,Y) = \sum p(x,y)log_2\frac{p(x,y)}{p(x)p(y)} \tag{2.11}$$

The above formulations will hold true only for predictive features which are categorical(can be classified into discrete classes e.g 1,2,3,4,5) in nature otherwise discrete variables are made categorical by classifying them based on thresholds to five equal bins. Each mutation can thus be associated with a class X=Driver/Passenger and a corresponding feature category(1,2,3...)

Using the mutation dataset p(X,Y) is determined for each combination of class and feature category and the features are ranked. So if a particular feature is always associated with a class across the mutations the mutual information is high as the feature can be alone used to predict the class. For e.g. feature Y always occurs as a low category(1,2,3) with label X="Driver" and high(4,5) with label X="Passenger" so depending on "positive" or "negative" correlation joint

Table 2.1: Empty Mutation Probability Matrix, C^*pG represents only C nucleotide was mutated in a CpG dinucleotide

	C^*pG	TpC^*	CpG^*	G^*pA	A	T	G	C
A								
T								
C								
G								

probability $p(X, Y) = 0$ or $p(X)$

2.5.2 In silico mutations

Synthetic passenger mutations were generated to train the classifier. Using curated datasets it is possible to come up with a mutation probability matrix as in 2.1 for different cancer types. This can then be used to generate passenger mutations where the rows show a multinomial probability distribution for the 8 contexts Passenger mutations are generated by mutating the wild type nucleotide corresponding to probabilities defined by the matrix

The

2.5.3 Training and Output

The Random Forest classifier was trained using COSMIC [24] database and synthetically simulated mutations.

Null Hypothesis: Mutations are passengers **Output of Random Forest Classifier** : Fraction of 'trees' that vote for the mutation to be under 'passenger' class, which gets translated into a CHASM score.

2.6 TransFIC [22]

TransFIC(TRANSformed Functional Impact for Cancer) is an ensemble of methods that combines the SIFT [26], Polyphen2 [2] and Mutation Assessor [36] scores. The motivation behind TransFIC lies in encompassing the effect of amino acid substitution ultimately on the functioning of the cell depending on the protein modification, which possibly confer a selective advantage to cancer cells for proliferation. Since all the nsSNVs that inhibit development has been eliminated by natural selection, the remaining nsSNVs in any gene define a 'baseline tolerance' level that survive without affecting the cell fitness and hence minor perturbations

to the otherwise conserved amino acid sequence needs to be accounted for. Simply put, he FIS generated by the three tools should be adjusted for the relevance of the gene/protein in cell operation

Using nsSNVs from 1000 Genomes Project [1], and annotation data Gene Ontology Biological Process (GOBP) and Molecular Function (GOMF) categories [4], canonical pathways (CP) [47] and Pfam domain [6], all the mutations were annotated. Annotation was performed in order to cluster genes into different groups with common functionality. For each annotation system, the nsSNVs of genes belonging to a particular cluster(each cluster has a set of functionally related genes. for e.g. all genes that regulate cell death) is pooled and all three scores from Polyphen2, SIFT and Mutation assessor are calculated. It was found that conserved genes belong to least tolerant group. Thus a mutation in an otherwise conserved gene is harmful from cancer point of view.

So a scaled FIS can be calculated such that two mutations affecting the same FIS affecting genes two entirely different germline tolerance should result in a higher FIS for mutation affecting gene with low tolerance. and the scaled score is given by:

$transfic = \frac{os-dm}{dstd}$ where
os = original SIFT/Polyphen/MA score
dm = mean score
dstd = Standard deviation of the score

Chapter 3

Galaxy Toolboxes for Driver Mutation Discovery

3.1 Galaxy

Galaxy [19] is a web based platform for data intensive biology. With the advent of Next Generation Sequencing and increased data generation by high throughput studies, there have been concerns over making analyses, accessible and reproducible. Galaxy provides an open source solution to track and manage data provenance.

Any software used for drawing conclusions from raw sequencing data might generate completely different results depending on the parameter values it is processed with. Simple threshold cut off for P values used to call a mutation a 'driver mutation' might affect the whole set of results. In order to ensure that results are reproducible all the steps followed, the command line parameters used for analysis should be documented properly.

With 'Galaxy workflows' it is possible to represent the entire data analysis pipeline in an intuitive graphical interface. These workflows can either be made public via a URL link or be provided as a supplementary material distributed as text file which can be 'imported' into any instance of Galaxy. These workflows act as 'log files' ensuring the same 'versions' and 'parameters' be used to run the analysis pipeline, ensuring reproducibility.

Zhang et al point out that there is less overlap between the softwares predicting driver mutations[56]. While trying to reproduce results from [56], it was non-trivial to convert the files to different data formats for every new tool used. Most of tools lacked a Galaxy plugin which would have otherwise taken care of data formats.

In order to tackle the issue of lack of any tool that can give a comprehensive

picture of the results of running all these softwares, in one go motivated us to come up with a set of 'Significant Mutation Toolbox' for Galaxy.

Here is a brief discussion of various tools used to predict driver mutation. Each section describes the input format required, and the steps required to convert VCF [16] file into appropriate input format. VCF file is standard output from a number of NGS pipelines and hence was chosen as the common input to all the tools.

3.2 Polyphen2

Polyphen2 is available as a webservice at http://genetics.bwh.harvard.edu/pph2/bgi.shtml. The server is sent an input file in appropriate format by the 'Polyphen2 Webservice' tool implemented in Galaxy.

3.2.1 Input Format

```
chr1:888659 T/C
chr1:1120431 G/A
chr1:1387764 G/A
chr1:1421991 G/A
chr1:1599812 C/T
chr1:1888193 C/A
chr1:1900186 T/C
```

Figure 3.1: Polyphen2 Input Format, http://genetics.bwh.harvard.edu/pph2/bgi.shtml

3.2.2 Galaxy Workflow

3.3 SIFT

SIFT toolbox is also implemented in Python and interacts with the webservice at http://provean.jcvi.org/genome_submit.php

From the VCF file only the relevant columns are 'cut' and 'trimmed' to send a request to the Polyphen2 webservice

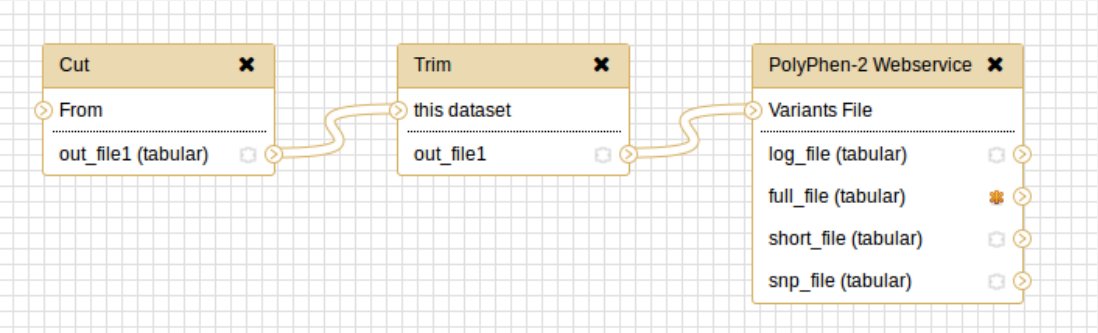


Figure 3.2: Polyphen2 Workflow as implemented in Galaxy

```
1,888659,T,C
1,1120431,G,A
1,1387764,G,A
1,1421991,G,A
1,1599812,C,T
1,1888193,C,A
1,1900186,T,C
```

Figure 3.3: SIFT/PROVEAN Input Format, http://provean.jcvi.org/genome_submit.php

3.3.1 Input Format

3.3.2 Galaxy Workflow

From the VCF file only the relevant columns are 'cut' and 'trimmed' to send a request to the SIFT webservice

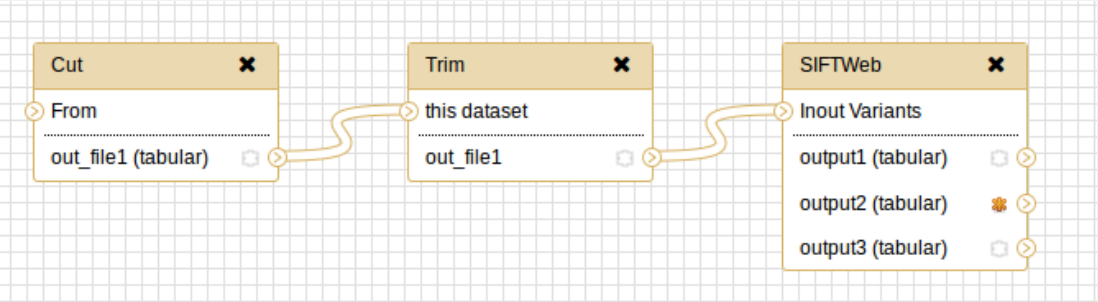


Figure 3.4: SIFT/PROVEAN Workflow as implemented in Galaxy

3.4 Mutation Assessor

Mutation Assessor supports an API http://mutationassessor.org/howitworks.php which is called from Galaxy. Mutation Assessor explicitly requires the human genome build[hg18/hg19] to be specified

3.4.1 Input Format

```
hg19,1,888659,T,C
hg19,1,1120431,G,A
hg19,1,1387764,G,A
hg19,1,1421991,G,A
hg19,1,1599812,C,T
hg19,1,1888193,C,A
hg19,1,1900186,T,C
```

Figure 3.5: Mutation Assessor Input Format, http://mutationassessor.org

3.4.2 Galaxy Workflow

From the VCF file some columns are trimmed while an extra column 'hg19' is appended

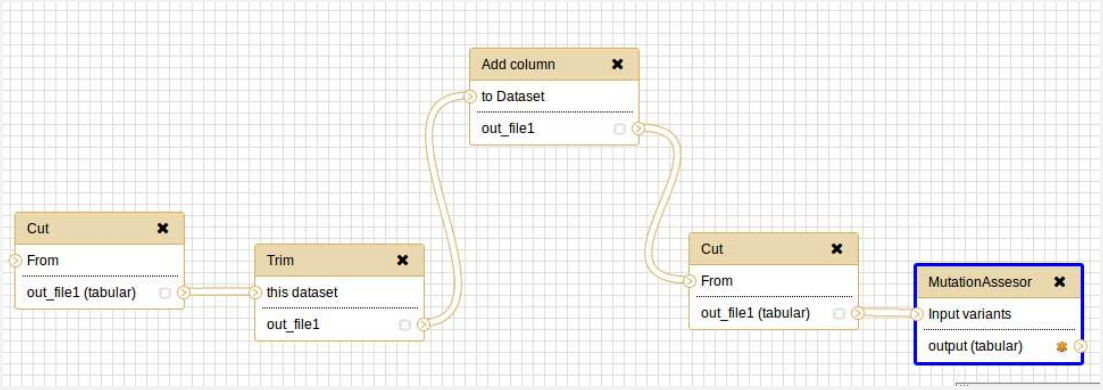


Figure 3.6: Mutation Assessor Workflow as implemented in Galaxy

3.5 TransFIC

TransFIC has an API that can be interacted from command line. PyCurl http://pycurl.sourceforge.net/ was used to interact with its API through Galaxy

3.5.1 Input Format

```
1 888659 888659 C
1 1120431 1120431 A
1 1387764 1387764 A
1 1421991 1421991 A
1 1599812 1599812 T
1 1888193 1888193 A
```

Figure 3.7: TransFIC Input Format, http://bg.upf.edu/transfic/home

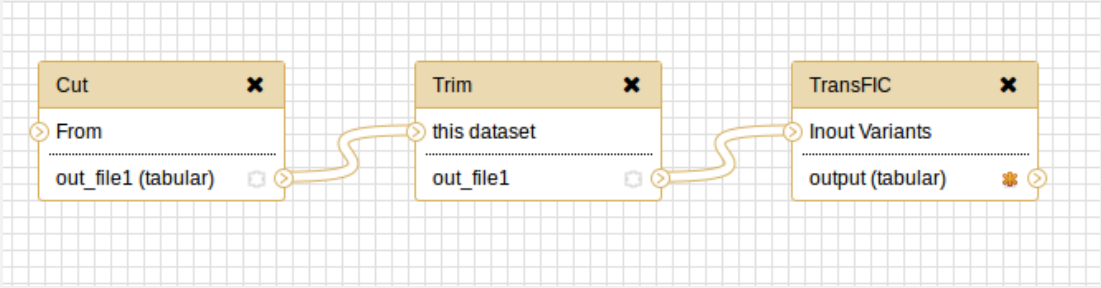


Figure 3.8: TransFIC Workflow as implemented in Galaxy

3.5.2 Galaxy Workflow

3.6 Condel

Condel and TransFIC have similar API

3.6.1 Input Format

```
1 888659 888659 C
1 1120431 1120431 A
1 1387764 1387764 A
1 1421991 1421991 A
1 1599812 1599812 T
1 1888193 1888193 A
```

Figure 3.9: Condel Input Format, http://bg.upf.edu/condel/home

3.6.2 Galaxy Workflow

3.7 Results and Discussion

The motivation behind coming with a Galaxy based tool was to come up with a comprehensive framework that will allow a user to compare the output of various tools predicting driver mutation. The set of tools can be run by importing each one of them to a workflow.

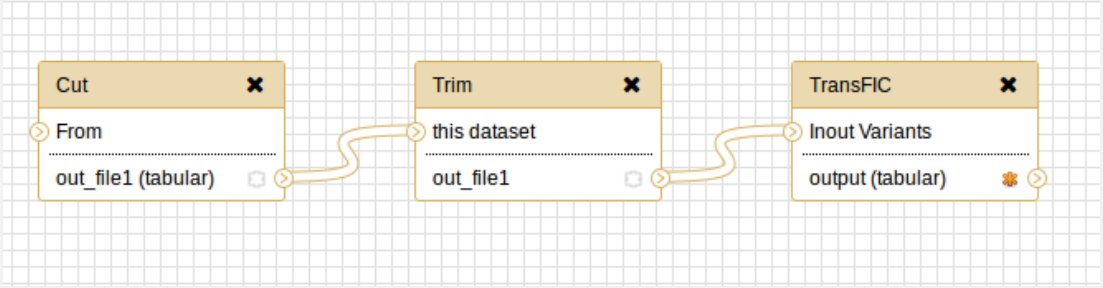


Figure 3.10: Condel Workflow as implemented in Galaxy

In order to have a visualize the user can choose to generate a 'heatmap'. This tool has utility in terms of reproducibility of the analysis and at the same time it is possible to create yet another ensemble of methods by plugging in the outputs of these tools to a completely new user defined tool. Hence these tools ported in Galaxy not only make analysis across all the methods available at 'one click' , but also provide a flexible framework to add new tools to refine the analysis.

Chapter 4

Galaxy Visualisation Toolbox: A Case study

Motivation

This chapter discusses a case study where we use the Galaxy based toolbox for assessing the deleteriousness of mutations.

We demonstrate how different scoring mechanisms score differently on the same set of mutations. The integration with Galaxy, aids the user by providing a reproducible and user-friendly way to interact with multiple tools at once.

4.1 Data and Method

In order to assess the prediction capability of various driver mutation prediction algorithms, we use a subset of mutations from the Catalog of Somatic Mutations in Cancer (COSMIC)[15] database. COSMIC database contains curated set of mutation data extracted from literatures studies on Cancer.

In order to create a proxy dataset to study the prediction capabilities of all these tools, we created set of workflows to parse the VCF file as obtained from COSMIC database. As already discussed, different tools use different input formats. A workflow based approach through Galaxy gives the end user the power to run multiple such analysis since the pre-processing is handled by these workflows itself.

An example workflow is shown in 4.1

Each tool generates it's score indicating the level of functional impact of each mutation. These scores carry different interpretations. For examples a Sift score of 0.05 indicates that the mutation is deleterious whereas a score of -1 on mutation assessor would indicate the mutation is likely benign. Most of the tool have

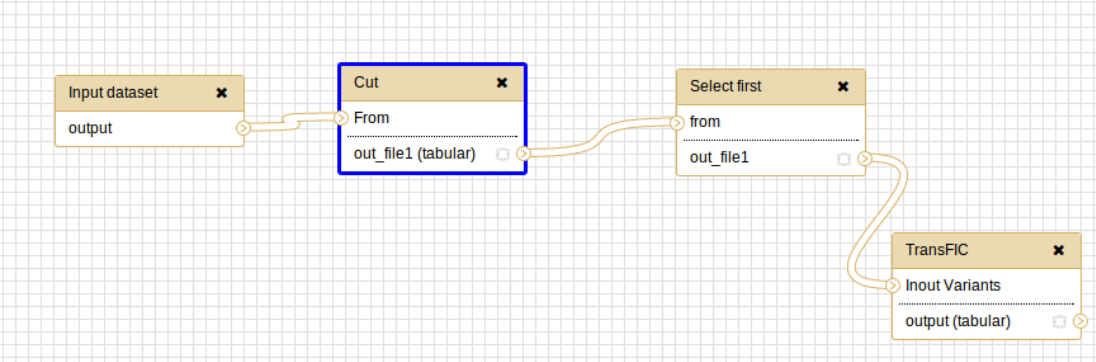


Figure 4.1: A Galaxy based workflow to process VCF files. The VCF files are converted to transFIC-friendly format

different ranges of scores, with different thresholds for calling a mutation driver or passenger.

Since the tools use different algorithms for prediction, the results are bound to be different. It has been reported [18] that these tools have varying accuracies and a pooled score can be a better predictor than individual scores.

For a particular mutation if all tools prediction is a 'benign', then the mutaion is likely benign too. Thus a *majority or pooled* consensus can help in narrowing down the disparities that exists. We try to tackle this by the use of a 'heatmap' integrated into Galaxy with Galaxy' visualization registry.

A heatmap such as in 4.2 gives an overview of the prediction scores for all the tools.

Methodology for transforming the scores

Since the scores generated have different ranges, in order to visualize the scores matrix(where the row represents a mutation and the columns represent scores from different tools). Thus, ach column represents the scores generated by a tool, say CONDEL for the input mutations(rows) . We transform these values by base-shifting the scores in each column by the minimum score in that column, folllowed by a scaling to bring the values in the range of 0 and 1.

$$n'_{ij} = \frac{n_{ij} - min(n_j)}{range_j} \tag{4.1}$$

where $min(n_j)$ represents minimum value of th column j, n_{ij} represents the original score for a mutation in i^{th} row given by tool j and $range_j$ represents the range of values(max-min) as given by the tool j over all possible mutations.

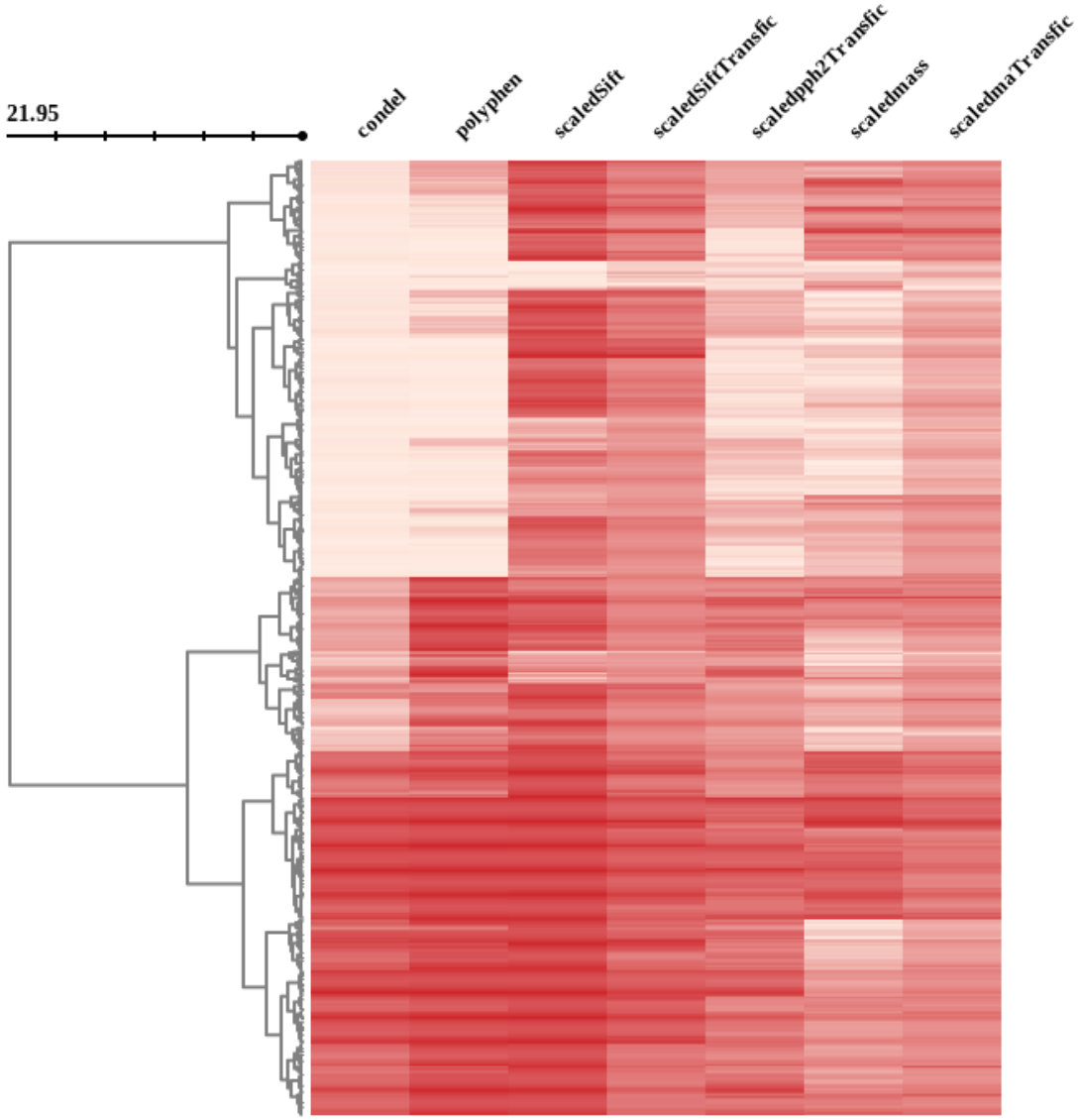


Figure 4.2: Heatmap representing the outputs of various tools. The framework is flexible enough to allow visualising output of more tools. The rows represent "chromosome:position" format. Darker shades of red represent damaging/ high functional impact mutations, lighter represents benign/low functional impact

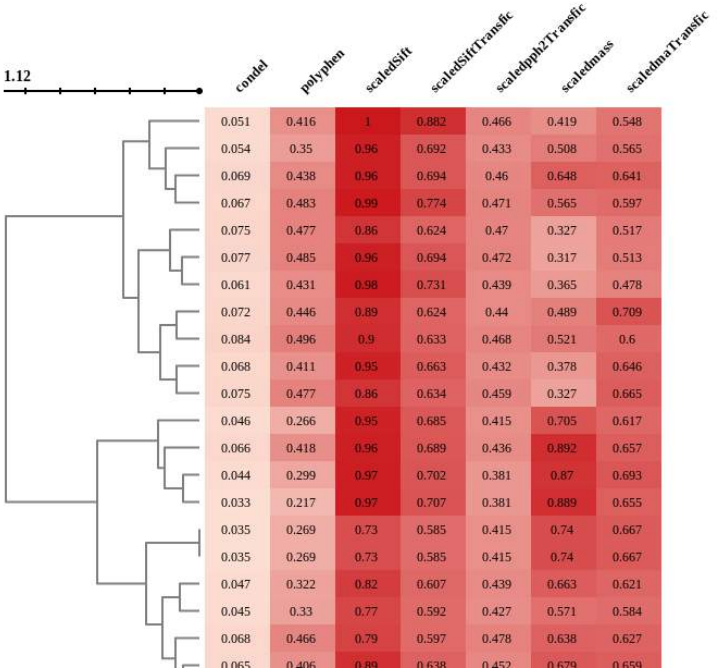


Figure 4.3: Heatmap Zoomed

Since the visual representation might not give insight into the actual score values, these can be visualized in the zoomed view.

InCHlib plugin of BioJS[21] was used to create these heatmaps. However, we have now successfully integrated this into Galaxy visualisation registry and hence visualisation and data processing are now possible from Galaxy itself

Conclusions

We presents a Galaxy based toolbox with visualization support that be a helpful tool to unify the predictions of different algorithms, thus leading to more true insights.

Chapter 5

Errors in Bioinformatics data analysis and Reproducible Research

With the availability of NGS at low costs, a lot of data is being generated everyday. It often is noticed that in pursuit of novel 'discovery', standardizing the data analysis pipeline is often ignored. This might not only lead to dubious conclusions, but will serve as a error prone guideline for further research. *Reproducibility is the hallmark of science* As Drummon points out in his paper [12], there are three aspects: Reproducibility, Statistical Replicability and Scientific Replicability. Reproducibility implies the experiment be replicable to at least an extent for other similar datasets. Statistical replicability addresses the problem of results-by-chance that may arise dues to limited data sets and above all Scientific replicability emphasizes the robustness and generalization of the result. While analyzing NGS data we noticed two common wrong practices:

- **Quality Score Encoding**: All the FastQ files were **assumed** to be encoded in Sanger format where the score is stored as (Phred score+33). Such an assumption will cause reads of lower quality encoded in (Phred+64) format appear as of high quality. So even though some bases are expected to not to be considered, blindly assuming the format to be be sanger causes erroneous analysis at the alignment step itself, which is going to propagate all the way down till post-processing, possible leading to dubious results
- **Ignorance on Quality Assurance of reads**: The pre-processing step is ignored, based on a blind assumption that the data is all fit to be processed in. The data might however have repeated fragments that might require grooming.

To solve the quality score encoding issue, we resort to using a guesser script as implemented at https://github.com/saketkc/NGS-Stuff/blob/master/guess_fastq_platform.py and to address the quality assurance and reproducibility at all stages, the following two system packages were made use of:

1. Galaxy: A web based tool for doing Bioinformatic analysis
2. bcbio-nextgen: A python based NGS pipeline

In order to perform bioinformatics analysis and, Galaxy can be made use of as to document the steps. Galaxy pages allow embedding of workflows illustratively that can shared via a simple url accessible publicly. The steps involved in Condel workflow are published as a Galaxy page in 5.1

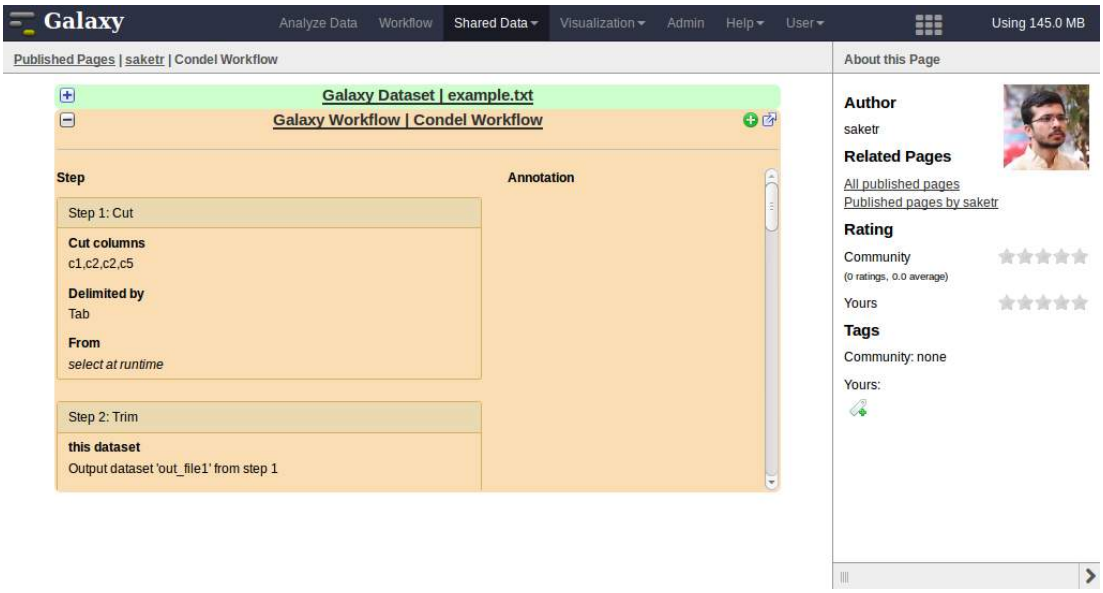


Figure 5.1: A published page of steps involved in Condel workflow accessible via a public URL. The workflow can be directly imported or downloaded too

Results and Discussions

NGS technology has numerous applications in life sciences. For example variant discovery is used in most of the applications. There are published protocols and best practice guidelines available[11], but they often are ignored in the haste of arriving at 'novel discoveries'. It is important that the standard set of protocols be followed to avoid down stream analysis turning out dubious. Any analysis pipeline

bcbio-nextgen pre-processes reads to detect any repeated segment among the

: **B1123 (1123_21092013_iplat-sort-prep)**

Reference organism	GRCh37	
Total	17,348,964	54bp paired
Aligned	16,157,535	(93.1%)
Pairs aligned	15,782,908	(91.0%)
Pair duplicates	6,891,418	(43.7%)
Insert size	221.7	+/- 34.4

Table 1: Summary of lane results

Sequence	Count	Percent	Match
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCG TATGCCGTC	149245	0.43	TruSeq Adapter, Index 6 (100% over 54bp)

Table 2: Overrepresented read sequences

reads

Figure 5.2: bcbio-nextgen report on repeated segments

should be benchmarked as in the case of bcbio-nextgen, besides requiring the least human intervention. At a broader level it is important that any new method being developed, or a software being used for analysis be properly documented, what parameters were used, what was the dataset and the results can should be reproducible. This is the main motivation to shift to Galaxy and bcbio-nextgen based pipelines.

Some of the guidelines that might help making an analysis 'reproducible':

- Any code being used for analysis, should be included and archived, either on the web or as supplementary material in case of a publication
- A log file that lists out all the commands that were run to generate the results should also be included
- If possible, the code should be available in *ready-to-run* format, web-based platforms like Galaxy would be possible options for hosting
- Not everyone is code literate, so emphasis should not be on learning a software but to understand how it works, what is the ideal set of parameters depending on the dataset

bcbio-nextgen reports of the quality profile of bases looking at which the user might want to trim the reads if quality falls below a threshold

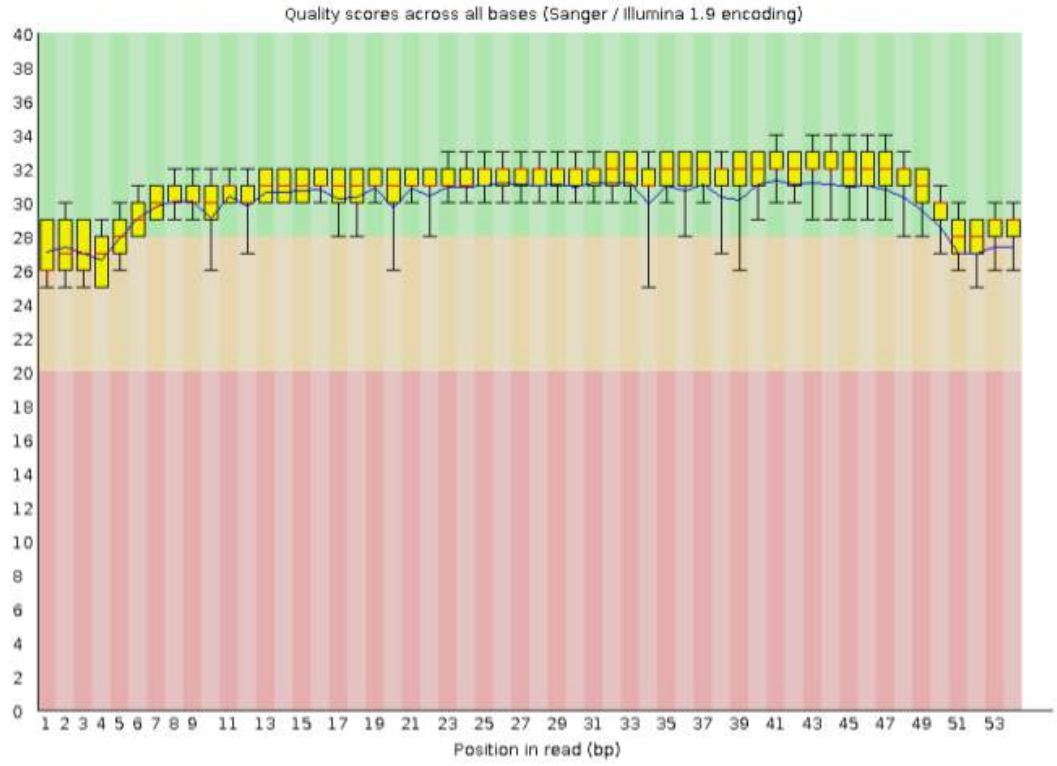


Figure 1:

Figure 5.3: bcbio-nextgen report on quality profile of reads

Chapter 6

Detecting Viral Genomes in Cancer tumors

Cervical cancers have been proven to be associated with Human Papillomavirus(HPV). Many of the cervical cancers have been found to contain HPV genomes integrated , though the sites of integration have been found to be relatively random [55]

Cervical Cancers and NGS

Cancer has been one of the widely studied disease using Next Generation Sequencing, the motivation being to tap into the changes at molecular level for a better understanding at genetic and genome level, and NGS is thus finding a routine role in diagnosis. [32]

Cervical cancer datasets from Indian women was put through an analysis to detect :

1. Any possible HPV integration
2. Sites of HPV integration

The motivation for such a study is from a prognosis angle. Instead of whole genome sequencing, it might be possible to predict onset of cervical cancer by doing a targeted sequencing at the sites where these virus have been detected in a cohort of samples, thus speeding up the whole process.

In order to detect viral genomes in the cervical tumors' dataset, the first step involves aligning the reads to the human genome. Since there are possible foreign(viral,bacterial etc) sequences, some of the reads will remain unaligned to the human genome reference. These unaligned reads can be extracted. Since the exact nature and origin of the reads is unknown, these were aligned with a

custom genome built by concatenating all viral genomes known to affect humans from NCBI[34]

Only a few samples were analyzed before further analysis on this project had to be abandoned over data privacy issues with the lab to which the data originally belonged. Alignment of unmapped reads extracted from one of the cervical cancer tumors is depicted in 6.2

Results and Conclusions

HPV sequences were detected in cervical cancer tissues. Though the project was discontinued before the exact integration sites of viral genome in human genome could be determined, such a study would enable targeted sequencing at these sites and can be used as an easy alternative to whole genome sequencing.

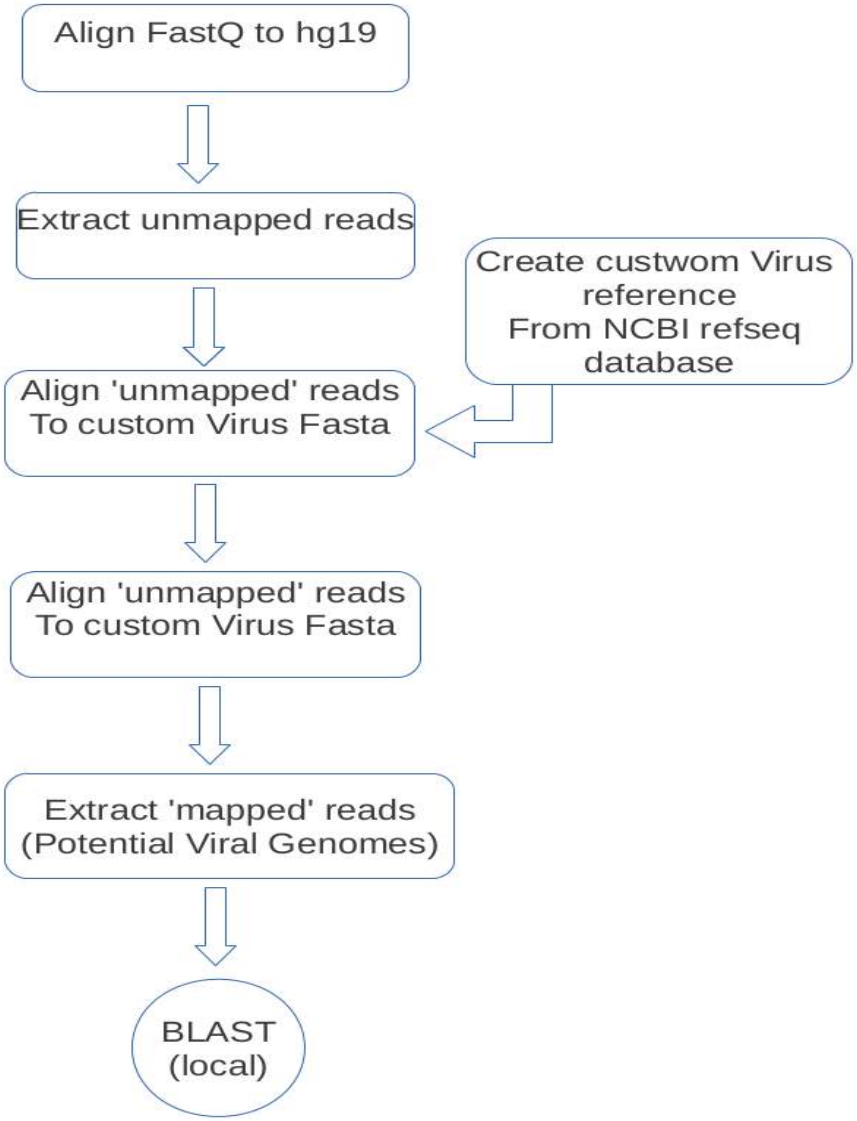


Figure 6.1: Steps to detect viral genomes in human NGS data

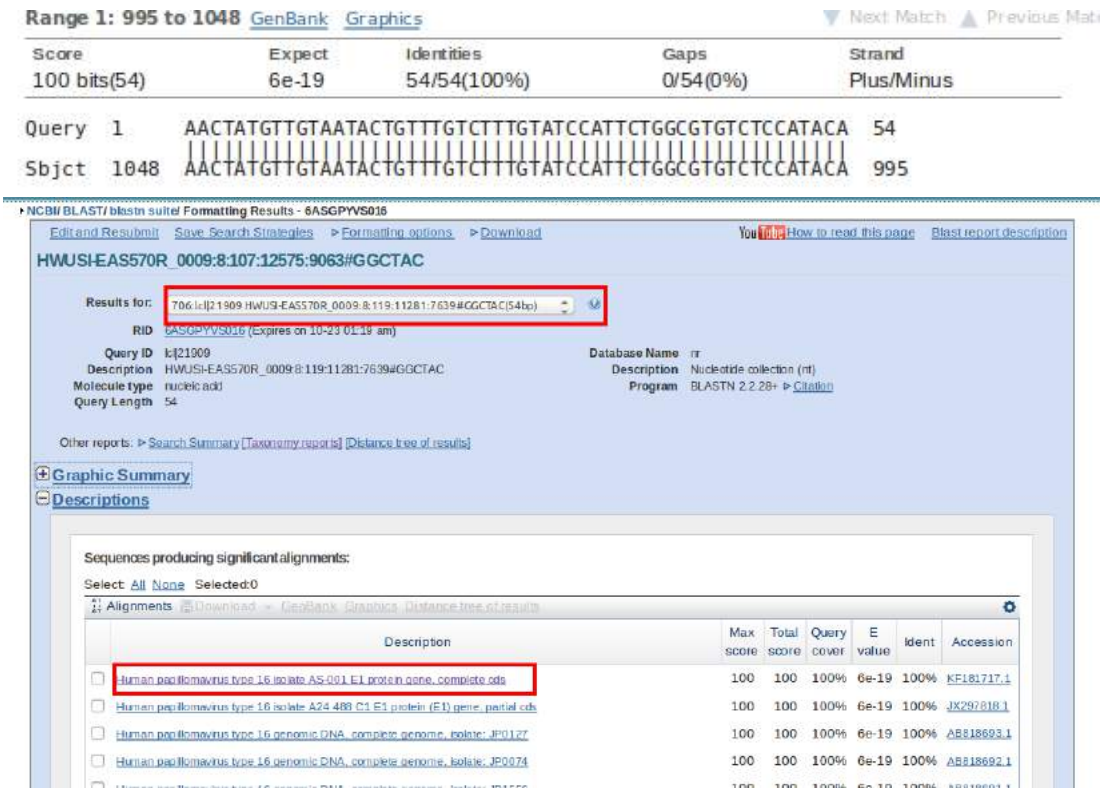


Figure 6.2: Reads unmapped to the human genome were aligned with custom built viral genome. All the reads mapping to this genome were then blasted. Some of the tissues showed an exact identity match between the read originally unaligned and the HPV16 genome sequence. Screenshots taken from NCBI BLAST [35]

Chapter 7

Benchmarking BWA with BWA-PSSM

BWA [27] is one of the widely used aligners for aligning short and long reads a given reference genome.

BWA internally uses a quality cut off. It thus ignores bases which have a quaity score below a certain threshold(-q option in bwa command line). BWA-PSSM[25] is a modification of the original BWA which takes quality scores into account while aligning in the form of a Position Specific Scoring Matrix(PSSM).

Consider the following sequence:
@read
ACT
+
III

Assuming Sanger encoded quality scores, all the base positions have a phred score of (73-33=40) . Given an error model of the sequencing platform, it is possible to come up with a matrix like:

	A	T	G	C
A				
T				
G				
C				

for all possible phred scores, which assigns to each possible score and a given nucleotide a score given by (i,j), emphasizing the probability that an observed nucleotide by the sequencer is indeed the same nucleotide

In order to benchmark BWA-PSSM against BWA, read pairs were simulated usinsg wgsim [28]. Using the hg19[NCBI build 37] reference genome, a custom

genome is simulated with SNPs and INDELS and then using this genome error free reads are generated and they are mapped to the synthetic genome by both BWA and BWA-PSSM algorithms.

A ROC 7.1 curve can be plotted since the number of reads that are expected to match is known apriori, This was done using the $wgsim_eval.pl$ script distributed with wgsim [28]

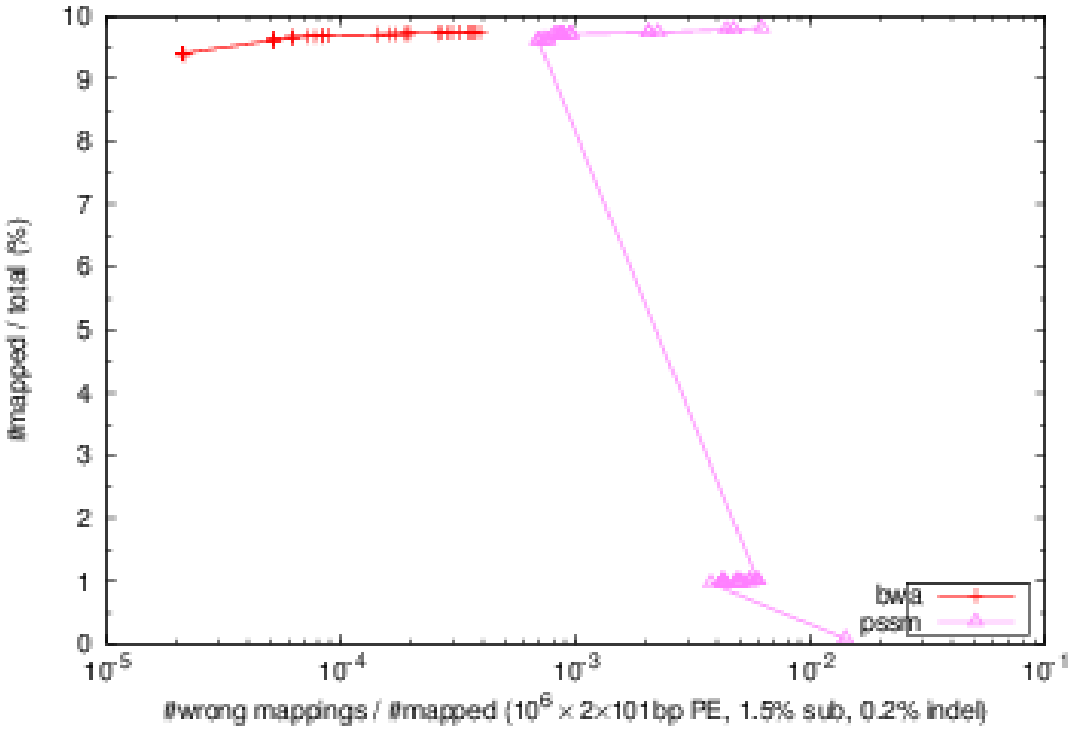


Figure 7.1: ROC curve for BWA v/s BWA-PSSM mappings

Results

As evident from the ROC, BWA-PSSM has a higher number of incorrectly mapped reads for the same true positives as BWA. BWA performs better.

BWA-PSSM internally uses the error model that might be optimized for a particular class of error profiles, which is a possible reason for its poor performance though in principle it is expected to perform better than BWA. The error model was not perturbed, which is one possible avenue for improvement.

Chapter 8

Analysis of Microarray Data

Organisation

This chapter is an introductory overview of the Microarray technology. The following sections discuss *-omics* research. The next follow up section describe the data analysis workflow for pre-processing the raw data. The penultimate section describes the outcome of this pre-processing work-flow when applied to a microarray study, followed by a discussion section on the overall recommendations for running such work-flows.

8.1 Introduction

One of the technologies that matches the scale of sequencing platforms in terms of the data involved is Microarray technology. Microarrays have been to study gene expression levels of thousands of genes at once. The other applications of microarray technology deals with gene variation analysis.

Microarray technology is one of the many techniques that emerged from the field of 'functional genomics'. 'Functional genomics' itself is a sub-field of molecular biology that primarily focuses on studying the dynamic aspects of 'genomics' such as the gene expression values. 'Genomics' on the other side, deals with the more *static* part of genome such as determining the whole sequence of genome.

8.2 -Omics Research

With the advent of the term 'Genomics', the other *-omics* terms were created, subsequently.

8.2.1 Genomics

The term 'Genomics' was coined by Dr. Thomas Roderick of the Jackson Laboratory at an international meeting on the feasibility of mapping the entire human genome, in 1986. [54] *Genomics* deals with analysis of *genome* sequences in order to analyse the associated function and structure. A sub-domain of this field includes *functional genomics* that focuses on the *functional* implications of genomics, thus studying genes and their products at the expression level. *Genomics* and *Genetics* are treated differently, since the latter is understood more from the point of view of study of genes from the point of view of *genetic* inheritance.

8.2.2 Proteomics

Proteomics is the functional and structural study of proteins. The scope of insight from Genomics is limited from a biological point of view, since the focus is on the *static* contents of the genome. Proteomics is a *functional* approach and hence can be more relevant from a biological point of view, given the focus is on protein levels. Protein levels determine cell physiology. Unlike genome content, the proteomic content has a dynamic nature and is constantly changing. One cell can have totally different proteome levels at different points of time, irrespective of the state of the cell(whether or not it has been *diseased/affected* by external or internal factors)

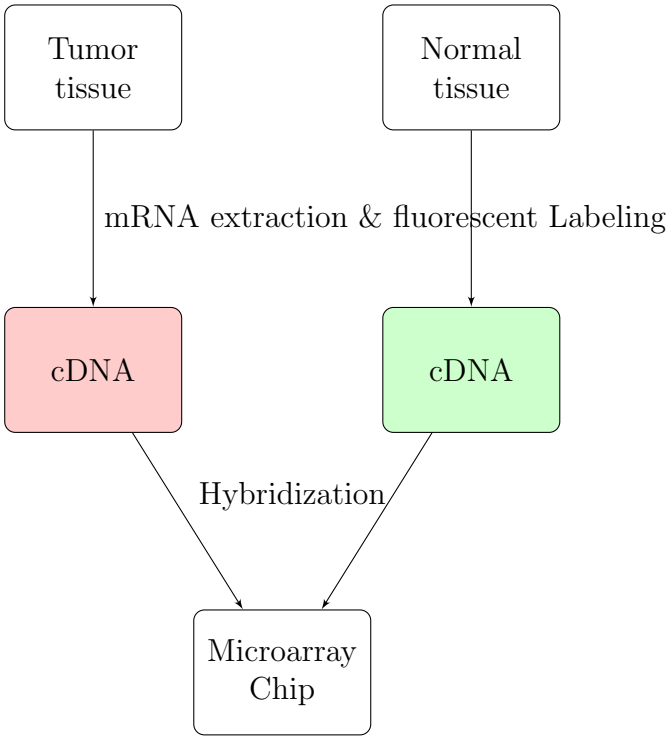
8.2.3 Transcriptomics

Transcriptome is the net total of all RNA content in a given cell type including mRNA, tRNA, rRNA, and other non-coding RNA. It is different from exons as exons consists of DNA transcribed to RNA in any cell type, whereas transcriptome is the transcribed RNA in a *specific* cell type. *Transcriptomics* or expression profiling is the study of these expression levels in a given cell type. High throughput transcriptomics is called *RNA-Seq*

8.3 Microarray Technology

8.3.1 Motivation

The motivation behind performing a microarray study is to identify those set of marker genes that could differentiate two or more conditions. In cancer studies, any particular insight into which set of genes are up or down regulated in the tumor tissue as compared to the benign tissues, can lead to potential therapeutic applications where the drug could be designed to control or repress the specific



gene's expression levels. Another potential output of a microarray study could be a *bio-marker*, a relatively small panel of *marker genes* that could be used in the place of whole microarray as a prognosis kit in order to determine the category of a particular sample(normal/disease), and hence if the whole expression profile of this unknown sample would be *similar* to those known to be suffering from the disease.

8.3.2 Experimental Design

A DNA microarray is essentially a lab-on-a-chip device with spots of fragments of DNA sequences attached. These spots contain small fragments of DNA that in turn would *hybridize* when exposed to a *target*. This hybridization is *quantifiable* by detection of intensity of fluorescent-dye signals. This signal intensity is proportional to the amount of sample-target bonding. A dual channel experiment would involve quantification of this intensity for two sets of samples, say *normal* and *disease*, under two different fluorescent signals. For example, the normal tissue could be labeled with a green dye and the tumor tissue with a red dye. These two samples are further hybridized on the same microarray chip. The relative intensities given by the red and green signals quantifies the level of difference between gene expression values across a normal-tumor pair.

8.3.3 Why Microarray?

Studying which genes are active and which genes are silenced in a *diseased* tissue versus the normal tissue can lead to a better understanding of the disease. Instead of studying each gene individually, microarray technology is a high-throughput way [44] to study thousand of genes at once. This is a shift from the classical hypothesis based testing, where the biologist's focus would have been to investigate one gene ata a time. This *omics* based high-throughput method involves study of thousands of genes at once, the hypothesis in most of the experiments remains the same: '*Majority of genes are not differentially expressed*' and as such the motivation is to identify those set of genes that are differentially expressed. A differentially expressed gene can either be over-expressed or under-expressed in a particular cell type, even though the genomic content does not vary between the cell types. The aim of the study is thus to identify the smaller set of differentially expressed gene which could either be further studied for a biological point of view or can in turn be treated as a set of *marker* to characterize the cell type, which can thus be used as diagnostic kit, thereby cutting down the need to profile from thousands of genes to a handful of them.

8.4 Microarray: A data science problem

There is an analogy between treating a microarray experiment to a classical clinical study. The traditional clinical study involved thousands to ten thousand cases with around 100 variables. A microarray study is just the transpose of this. Given the large number of variables involved , the system is largely *undetermined*. Given this sort of high dimensionality with the number of variables far exceeding the number of observations and hence there is a need to to get rid of this *curse of dimensionality*

8.5 Data Analysis

An output of a microarray experiment is a text file, (*.gpr* in case of Genepix platform). This file stores spot intensities as ready by the scanner. A foreground is determined by the image processing software after the spots has been aligned by a gridding software. Due to the intermediate steps involved the color intensity recorded by the software is prone to noise and probably bias. Bias may arise due to various other factors:

- Variation between chips arising due to manufacturing defects

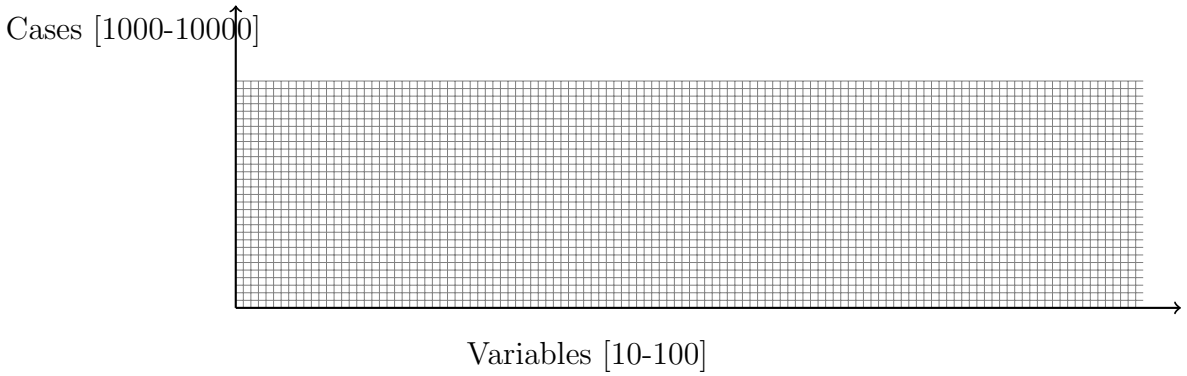


Figure 8.1: Traditional clinical studies

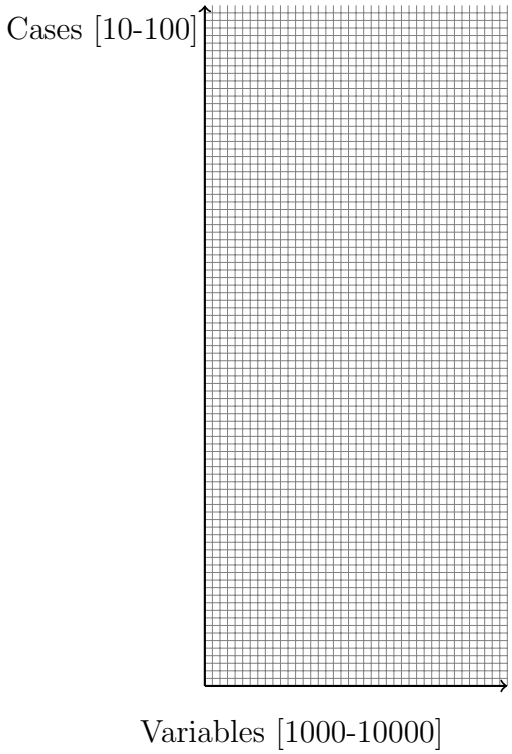


Figure 8.2: Curse of dimensionality with microarray and most other high throughput data

- Amount of dye used might differ between various samples, hence leading to certain chips showing an overall high/low intensity
- In case of dual channel experiments, there have been reports of a gene-specific dye bias [30]
- Measurement errors, arising from the scanner being either sensitive to certain regions of the chip or having different sensitivity for different intensity ranges

Most of these errors can be controlled via good a replication design.

Replication

Given that the experiments are always subject to random noise, mostly arising due to various environmental factors(for example room temperature might effect hybridization efficiency [29]), replication can be a good tool to adjust for the associated variations. Replication is possible at two levels:

- **Biological** : A biological replicate comes from a *new* biological sample. The aim of a microarray experiment to define the set of differentially expressed genes between two or more cohorts and as such the interest lies in the average behavior of the genes across the two cohorts. These two cohorts should be representative of their respective population and hence requires that more members of the population be made part of the study as biological replicates. Even though there are chances of higher variation arising due to more subjects, the results will tend to be less biased.
- **Technical**: Technical replication involves repeating the *same* subject. So multiple *samples* are created from the same subject. Though technical replicates can help reduce the variations involved due to random noise, the results however might be too specific for the cell line replicated and may not be true for the population.

Microarray Experiments: Inherent assumptions and myths

Myth: Gene expression values on a chip follow Gaussian distribution

Gene expression values on a single chip **do not** follow a Gaussian distribution. In fact given that microarray experiments can involve any subset of genes from the organism, it does not make sense to even assume that those genes will follow any kind of probability distribution.

A probability distributions assignment is to any process with repetitive measurements of the same instance of experiment. Hence though the gene expression values on a single chip experiment may define a *distribution* but **not** a probability distribution.

It can however be assumed that the gene expression values follow a normal distribution *across* arrays. This is an important assumption, which we encounter whilst performing parametric tests in order to identify the (statistically) differential expressed genes [17]

Microarray data analysis involves several steps, which can broadly be divided into the following steps:

- **Exploratory Data Analysis and Quality Assessment**: This is often an essential step to determine which methods should be employed to preprocess the raw data and gauge the overall quality of data.
- **Background Correction**: Involves separation of foreground signals from the background
- **Normalization**: Adjustment for *within-array* and *between-array* bias that might arise due to different experimental conditions, dye-bias etc.

8.6 Exploratory Data Analysis

Before performing any sort of normalization it is often helpful to plot the raw intensities so that they can be compared post-processing. The most often explored values are the foreground and background intensities. A lot of background signal in one of the samples might be an indicative of a faulty array.

8.7 Background Correction

The background intensities arise irrespective of true(foreground) intensities, often due to non specific binding of the dye to the spot, irrespective of the presence of the probe. Background correction involves *subtracting* the background intensities from the foreground intensities. The background intensity can in fact not be measured at the spot directly and is read from the spots nearby.

8.7.1 standard

A näive approach to obtain the true signal intensity is to subtract the foreground intensity from the background intensity. This however can result into final intensities being negative.

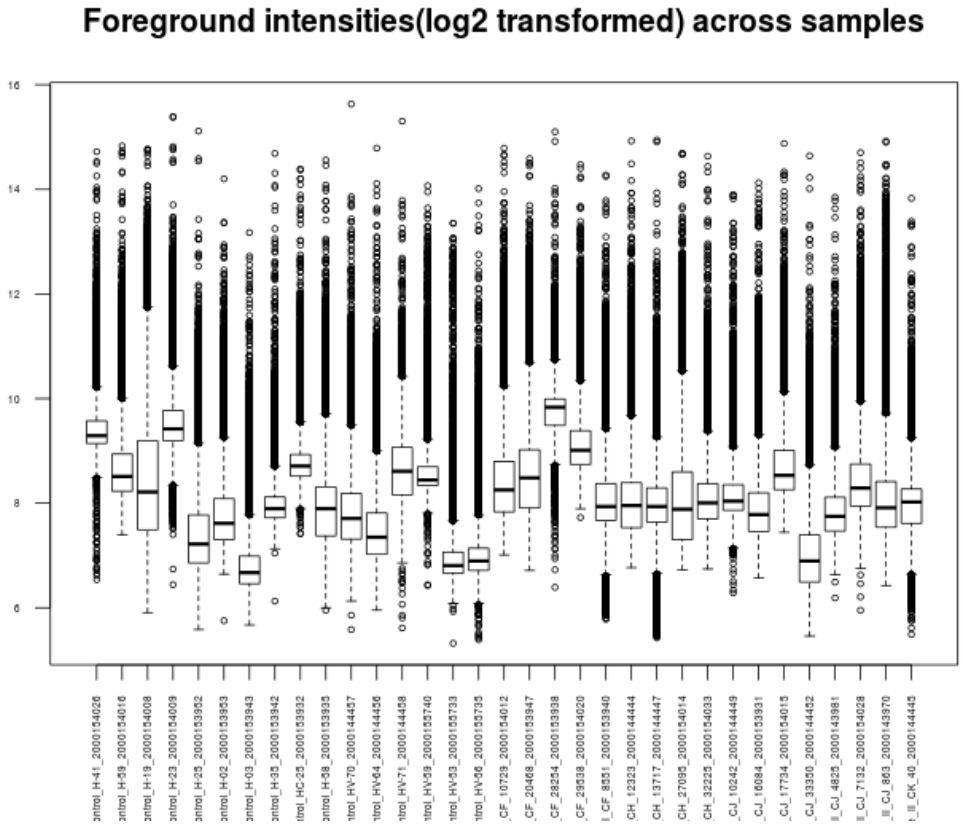


Figure 8.3: log_2 transformed foreground intensities of 17 control and 16 Grade II GBM patients

An observed intensity can be modeled as an additive of two intensities, the true intensity T and the residual background signal B, besides S_b, the additive background signal:

$$S = B + T + S_b \tag{8.1}$$

8.7.2 normexp method

The normexp method [41] models the background intensity to be normally distributed:

$$B \sim \mathcal{N}(m, \sigma^2) \tag{8.2}$$

and the True signal is modelled as an exponential:

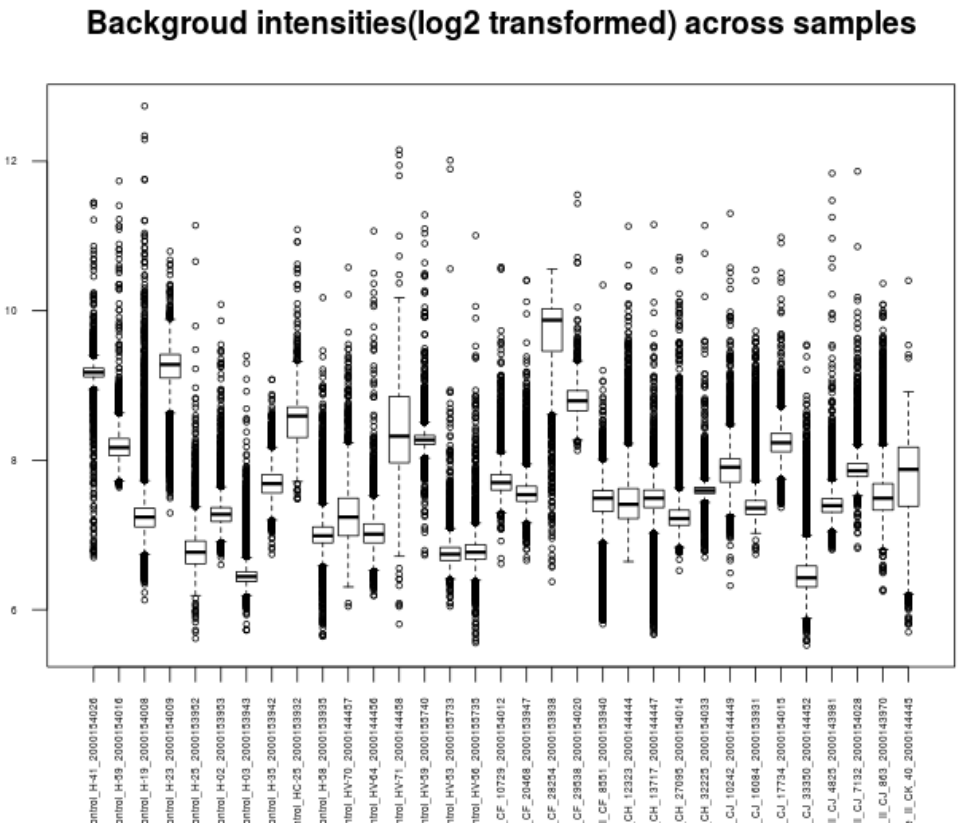


Figure 8.4: log_2 transformed background intensities of 17 control and 16 Grade II GBM patients

$$T \sim \frac{1}{\alpha} exp \frac{-t}{\alpha} \tag{8.3}$$

The noise B and true signal T can be assumed to be independent random variables and hence the joint distribution of B and T is given by:

$$f_{B,T}(b,t;\mu,\sigma,\alpha) = \frac{1}{\alpha} exp \frac{-t}{\alpha} \mathcal{N}(t;m,\sigma^2) \tag{8.4}$$

Consider the random variable $X = S - S_b$ which is essentially the background subtracted intensity[foreground-background]. The joint distribution of X,S is given by [42]:

$$f_{X,T}(x,t;\mu,\sigma,\alpha) = \frac{1}{\alpha} exp(\frac{\sigma^2}{2\alpha^2} - \frac{x-\mu}{\sigma}) \mathcal{N}(t;\mu_{X,T},\sigma^2) \tag{8.5}$$

$$\mu_{X,T} = x - \mu - \frac{\sigma^2}{\alpha} \tag{8.6}$$

On integrating and dividing the joint by marginal:

$$f_{S|X}(s|x;\mu,\sigma,\alpha) = \frac{\mathcal{N}(t;\mu_{X,T},\sigma^2)}{1-\mathcal{N}(0;\mu_{X,T},\sigma^2)} \tag{8.7}$$

The estimated signal given the observed intensity x is given by:

$$E(T|X=x) = \mu_{X,T} + \frac{\sigma^2\mathcal{N}(0;\mu_{X,T},\sigma^2)}{1-\mathcal{N}(0;\mu_{X,T},\sigma^2)} \tag{8.8}$$

normexp produces positive intensities.

8.7.3 normexp+offset

normexp method can be stabilised for variance effects by shifting the baseline by adding a small offset to move the corrected intensities from zero.

8.7.4 edwards

Edwards method [13] by subtracting the background from the foreground only when the difference is larger than a threshold. Otherwise, it is replaced by a smooth monotonic function

8.7.5 rma

The RMA(Robust Multi-Array Average) method works by partitioning the distribution of smoothed intensities around its mode. The rma-75 and rma-mean estimators [31] aim at correcting for the bias created by the rma algorithm.

8.8 Between Array Normalization

Between array normalization is important if the box plots of the arrays are not consistent throughout width-wise. We obtain the following boxplot after performing '*normexp+offset*' background correction. The choice of this particular background correction comes from the discussion given in Ritchie et al. [37].

MA plots

For dual channel arrays, visualising the intensity difference between the Red and Green channel, would generally involve a scatter plot of the R and G channel intensities. The expected behaviour of this scatter plot is a diagonal. A slight modification of this strategy is to plot, M and A:

$$M_i = log_2(\frac{R_i}{G_i}) \tag{8.9}$$

$$A_i = \frac{1}{2} log_2(R_i G_i) \tag{8.10}$$

Thus plotting M_i versus A_i should in principle return a 45 degree rotated plot. Since the underlying hypothesis of a Microarray experiment is that *majority* of genes are not differentially expressed, hence a MA-plot would involve just a single line passing through x-axis, since the expected intensity of red and green channels are expected to bye similar. There are however points scattered around this line, depicting under and over expressed genes, potentially differentially expressed.

For a single channel the R is analogous to the single channel intensity

8.8.0.1 cyclicloess

Loess is a *non-parametric* methods used for smoothing scatter plots based on a locally weighted regression [10]. It involves fitting a polynomial to a subset of data at each point in the data set. The fitting in turn is based on weighted least squares. This involves a user specified input with n the degree of polynomial to fit and a smoothing factor α. Values of α determines the proportion of dataset to be used for fitting the polynomial of degree n.

The polynomial fitting procedure gives most weight to the points lying closest to the point at which the polynomial estimation is taking place.

A polynomial $f(A_i)$ can be fit , using loess. The residuals are given by:

$$\bar{M_i} = M_i - f(A_i) \tag{8.11}$$

$$\bar{A_i} = A_i \tag{8.12}$$

These residuals will constitute a normalized MA plot. In terms of M_i and A_i:

$$log_2(R_i) = \bar{A_i} + \bar{M_i}/2 \tag{8.13}$$

$$log_2(G_i) = \bar{A_i} - \bar{M_i}/2 \tag{8.14}$$

For single channel arrays the MA plots involve plotting the different M, (equivalent to the background corrected intensity) versus a *redefined* A, where A can be taken to be average of all arrays, or be obtained from a pairwise comparison everytime. The 'cyclic' in cyclicloess arises because this loess procedure is run over every possible pair of array chip comparisons.

8.8.1 Quantile Normalization

Quantile Normalization normalizes the value in two or more datasets by essentially making the distribution of the probe intensities identical statistically. This in turn is motivated by the concept of **Q-Q plot**. A **Q-Q plot** is a graphical method to compare the values of two probability distributions. Given a case of plotting values of the reference dataset as the abscissa and those of the test dataset as the ordinate, the graph depict a diagonal line, given that the test and reference dataset have identical probability distributions. This is simple to imagine since the highest value in the test dataset would correspond to the highest value in the reference dataset, and so on .

Quantile normalization thus makes the distribution of the test dataset identical to that of the reference dataset by associating the highest value in the test dataset to the highest value in the reference dataset, the next highest value in the test dataset to the second highest value in the reference dataset and so on. However there is no such *reference* dataset in any microarray dataset.

The problem of performing quantile normalization where there is no reference dataset can be tackled by going back to the motivation of **Q-Q plot**. Extending the concept of Q-Q plots to n-dimensions, if all the n data vectors have identical distribution, a Q-Q plot would generate a straight line with its direction vector as:

$$d = (\tfrac{1}{\sqrt{n}}, \tfrac{1}{\sqrt{n}}, ..., \tfrac{1}{\sqrt{n}})$$

Thus it is possible to make a set of data have identical distribution if it is projected along this diagonal d.

Consider k^{th} quantile data vector $q_k = (q_k1, q_k2, ..., q_kn)$ for $k = 1, 2, 3, ...m$ for m spots on n arrays. A linear transformation of this vector given by:

$$Q = (q_k.d).d$$

$$Q = (\tfrac{1}{n}\sum_{j=1}^{n} q_m j, ..., \tfrac{1}{n}\sum_{j=1}^{n} q_m j)$$

Thus the reference dataset is created by extracting the mean quantile values from the given data vector. Typical steps to perform quantile normalization for

1	5	6	7
2	1	4	2
3	3	5	5
4	2	3	8

Table 8.1: Original Data

1	1	3	2
2	2	4	5
3	3	5	7
4	5	6	8

Table 8.2: Data with sorted columns

a given m spots on n arrays:

1. Create a mxn matrix say X
2. Sort the columns in X so that the 1^{st} row has the highest values across all columns, 2^{nd} has the second highest values across all rows and so on.Call this array X_sort
3. Calculate the row wise means of the sorted array X_sort
4. Rearrange X_sort replacing the rearranged values by the corresponding means calculated above.

A short example demonstrating the approach:
Sorted columns:

8.9 Differential Expression

After the pre-processing, the next step is to identify the set of genes that are differentially expressed. This step essentially involves performing statistical tests

1	1	3	2	2.75
2	2	4	5	3.25
3	3	5	7	4.5
4	5	6	8	5.75

Table 8.3: Row means

1	2.75
2	3.25
3	4.5
4	5.75

Table 8.4: Rank Mappings

1	4	4	3
2	1	2	1
3	3	3	2
4	2	1	4

Table 8.5: Rank Matrix:

on the pre-precessed data and assigns rank to the genes based on them. Differentially expressed genes are the genes whose expression levels are outliers to standard state that the other genes exhibit. The underlying hypothesis, to test of gene i is differentially expressed or not is:

$$H_0 = \text{Gene } i \text{ is not differentially expressed}$$
$$H_1 = \text{Gene } i \text{ is differentially expressed}$$

Test statistics are used to summarize the evidence in the data underlying H_0

8.9.1 Fold Change

Let x_i^C and x_i^D represent the $log2$ expression levels of the gene i in Control and Disease sample respectively. One definition of log Fold Change is given by [52]:

$$FC_{ratio} = \frac{\bar{x_i}^C}{\bar{x_i}^D}$$

where $\bar{x_i}^C$ and $\bar{x_i}^D$ represents the mean expression level of gene i among Controls and Disease samples respectively. However fold changes have also been calculated as : [23]

2.75	5.75	5.75	4.5
3.25	2.75	3.25	2.75
4.5	4.5	4.5	3.25
5.75	3.25	2.75	5.75

Table 8.6: Normalised Matrix

$$FC_{difference} = \bar{x_i}^C - \bar{x_i}^D$$

Fold Change cut off is a naive approach used to determine *differentially expressed* genes. The genes following a fold change above a certain threshold, generally between 1.8 and 3.0 are regarded as differentially expressed. Biologists tend to prefer this method of short-listing, even though there are serious problems with this approach. A fold change based cut off ignores the inherent variance that might exists in the short-listed genes, and considers only mean values of expression. Thus, genes with larger values of variance will often tend to be short-listed, just because of the noise involved. Other way round, the highly expressed genes will not be shortlisted because of the lower variability involved and hence lower chances of showing [9]. It is however been suggested that the decision to use a fold change based cut off is biological [53]. The choice of a fold change can be justified if large absolute changes are indeed relevant to that particular experiment, ignoring the underlying noise.

8.9.2 t test

8.9.2.1 Welch's t test

The Welch's t-test uses the following statistic:

$$z_i = \frac{\bar{x_i}^C - \bar{x_i}^D}{s_i} \tag{8.15}$$

where s_i is the **non-pooled** variance:

$$s_i = \sqrt{\frac{sc_i^2}{N_C} + \frac{sd_i^2}{N_D}} \tag{8.16}$$

where sc_i and sd_2 are the standard deviations with sample sizes N_C and N_D for the control and disease respectively. z_i has degrees of freedom df given by:

$$df = \frac{(a_c + a_d)^2}{\frac{a_c^2}{N_c - 1} + \frac{a_d^2}{N_d - 1}}$$

where :

$$a_c = \frac{sc_i^2}{N_c}$$

This z_i statistic follows a t-distribution:

$$z_i \sim t_i$$

the associated p-value is given by:

$$p - value = 2 * P(t_i \geq |z_i|)$$

8.9.2.2 Pooled variance t-test

For pooled variance t-test, the assumption is that the control and disease samples have the equal population variance and hence s_i is given by:

$$s_i = \sqrt{\frac{(N_C-1)sc_i^2+(N_D-1)sd_i^2}{N_C+N_D-2}}$$

and z_i has $N_C + N_D - 2$ degrees of freedom.

8.9.3 Linear Models for Microarray

Smyth et al. [43] suggested linear models for modeling microarray experiments. Consider N set of samples in total with the gene g's expression values in the n samples given by:

$$y_g^T = (y_{g1}, y_{g2}, ..., y_{gn})$$

y_g^T contains the normalized $log2$ pre-processed intensities. Then,let the expectation of $y_g$be given by:

$$E(y_g) = X\alpha_g$$

Where X is the design matrix and α_g is an unknown coefficient vector. The variance of y_g is given by:

$$var(y_g) = W_g\sigma_g^2$$

where W_g is a weight matrix, and σ_g^2 represents unknown gene wise variance. Consider β_g as the log-fold change for gene g. Instead of classical hypothesis testing where the test is $H_0 : \beta_g = 0$ versus $H_1 = \beta_g \neq 0$, the test is conducted on thresholded values $H_0 : |\beta_g| \leq \tau$ versus $H_1 : |\beta_g| > \tau$, where τ is pre-specified log-fold change.

Assume the contrast to be tested is $\beta_g = c^T\alpha_g$ where c^T is a contrast matrix like X. Since α_g is unknown, given the response vectors and X it is possible to

fit a linear model to obtain an estimate of coefficient vector as $\hat{\alpha_g}$ such that the covariance is given by:

$$var(\hat{\alpha_g}) = V_g \sigma_g^2$$

where V_g is independent from σ_g^2 and is positive definite.

Thus the estimate of β_g is given by $\hat{\beta}_g = c^T \alpha_g$ Assuming $\hat{\beta}_g$ to be normally distributed without forcing the normal distribution on y_g. $\hat{\beta}_g$ is assumed to be normally distributed with mean β_g and can be approximated as :

$$\hat{\beta}_g | \beta_g, \sigma_g^2 \sim \mathcal{N}(\beta_g, v_g \sigma^2)$$

where

$$v_g = c^T V_g c$$

the variance s_g^2 is assumed to follow a scaled χ^2 distribution.

$$s_g^2 | \sigma_g^2 \sim \frac{\sigma_g^2}{d_g} \chi_{d_g}^2$$

where d_g represents the residual degrees of freedom for gene g.

Under the above assumptions, the statistic t_g follows a t-distribution with d_g degrees of freedom:

$$t_g = \frac{\hat{\beta}_g}{s_g \sqrt{v_g}}$$

8.9.4 Correcting for multiple comparison

Determining the set of differentially expressed genes involves multiple hypothesis testing, where the null hypothesis H_0 states that the gene is not differentially expressed. Thus for each gene, the test statistic tests H_0,since these tests are performed on 1000 or more genes. This potentially leads to an increased chance of false positives.

Considering a sample of 1000 genes with 30 of them differentially expressed. A $p-value$ is the probability of obtaining a result that is equal or more than the actually observed given that the null hypothesis H_0 is true. The significance of a smaller $p-value$ lies in the fact that it implies that the observed values might be very rare given H_0 is true or H_0 is not true at all. Let the pre-defined threshold for the $p-value$ be α where we reject H_0 falsely with a probability of α.

With a p value cut off of 0.05, this would also imply of the remaining $1000 - 30 = 970$ non-differentially expressed genes,$0.05x970 = 49$ are false positives Thus the number off positives are greater than the truly expressed genes.

In order to avoid this bias, arising due to over-fitting, the levels of significance can be adjusted for multiple testing. The *Bonferroni* correction corrects this error by reducing the level of significance by a factor of n, the total samples.

8.10 Materials and Methods

Datasets

Though DNA based microarrays are more common, here we tackle analysis of microarray from a proteomics study.

The data for the study was obtained from Glioblastoma multiforme(or GBM) patients. GBM is known to arise from the glial cells responsible for homeostasis. Homeostasis(biological homeostasis) is a property of the human body via which the body regulates the levels of certain variables in order to ensure that the internal metabolic reactions can be carried out in the right set of conditions. Regulation of blood's pH at a value of 7.365 is thus a result of homeostasis.

Clinicians, classify GBM into two broad categories:

- Low Grade[Grade II]: Cells are non-anaplastic and the tumor is benign. The current methods of diagnosis reply MRI scans
- High Grade[Grade III and Grade IV] Cells are anaplastic and hence are dividing rapidly forming malignant tissues.

The data was collected from the following samples:

1. **Controls**: A set of 17 control samples, collected from *normal* patients. It is worth noting that the definition of **Controls** itself is a bit subjective. While comparing the gene expression levels with the *disease* samples, we assume these set of controls to be **'healthy'**, and a representative sample of the *healthy population*
2. **Grade II**: A set of 16 samples with Grade II glioma.
3. **Grade III**: A set of 16 samples with Grade III glioma.
4. **Grade IV**: A set of 16 samples with Grade IV glioma.

The assignment of grades to the diseased samples was an outcome of the Clinician's analysis of the MRI and related tests. GenePix platform and software were used to

Defining the Question

Before performing any data analysis, it is very important to clearly define the question to be answered. The motivation behind performing this data analysis was:

1. Determine the set of differentially expressed genes in Grade II, Grade III and Grade IV samples with respect to the *healthy* controls
2. Compare the set of differentially expressed genes via a pairwise comparison between (Grade II, Grade III), (Grade III, Grade IV) and (Grade II, Grade IV)
3. Determine a smaller panel of *marker genes* that can potentially be used to differentiate the various Grades of GBM from control and possibly amongst themselves

Exploratory Data Analysis

One of the key ideas, that needs to be taken care of is that the pre-processing steps should be run on all the datasets at once, rather than running them on any two cohorts taken together. This is important for any inferences that can be drawn out after the downstream analysis.

Consider the box plots for foreground and background intensities. Though a background intensity is always expected, *subtraction* of this background intensity from the foreground should yield a similar looking boxplot across all arrays. However as evident from 8.4, this does not seem to be the case and this is the motivation to perform a normalization across all arrays

Normalization

The underlying hypothesis in a microarray study is that the expression levels of most of the genes are expected to be constant. Thus, a box plot with these expression values, should thus exhibit a behavior such that the mean expression levels across the samples remain the same. We employ 'quantile' based normalization here.

With normalization, the pre-processing work-flow is complete for the next downstream analysis for finding differentially expressed genes.

Differentially expressed genes

One of the good ways to visualize the set of differentially expressed genes is a Volcano plot. The plot plots log odd scores versus the log fold change. Thus

for a gen to be differentially expressed, just being above a certain fold change threshold is not sufficient s it must satisfy must have higher log odd score too.

8.11 Discussion

This chapter discusses the individual components of a pre-processing workflow, besides discussing the Mathematics, behind most of the methods. As a small case study, we present the list of differentially expressed genes in Grade-2 of GBM samples as compared to the Controls, after required pre-processing.

Though the list of differentially expressed genes, the purpose of the problem defined in the beginning of this Chapter is not met yet. Had the question being asked just focussed on identifying those set of genes that show a marked behavior between Control and Disease samples, the list of differentially expressed gene would be a close answer. However we are also interested in finding if these expression profiles can be used to differentiate these cohorts. These set of differentially expressed genes could themselves be large in number. From the point of biomarker design, a long list of these genes would not solve the process. The näive solution of selecting the first few differentially expressed genes, doe not work either because these top ranking genes might be part of single pathway and hence a change in an upstream gene will trigger the expression of the following downstream genes in the pathway. Hence even though these genes are the top ranked in differential expression, they might collectively be unable to differentiate the control and disease samples, since the *information* gain from such a list might be minimal.

The next chapters discuss the approach we take to potentially come up with such a bio-marker.

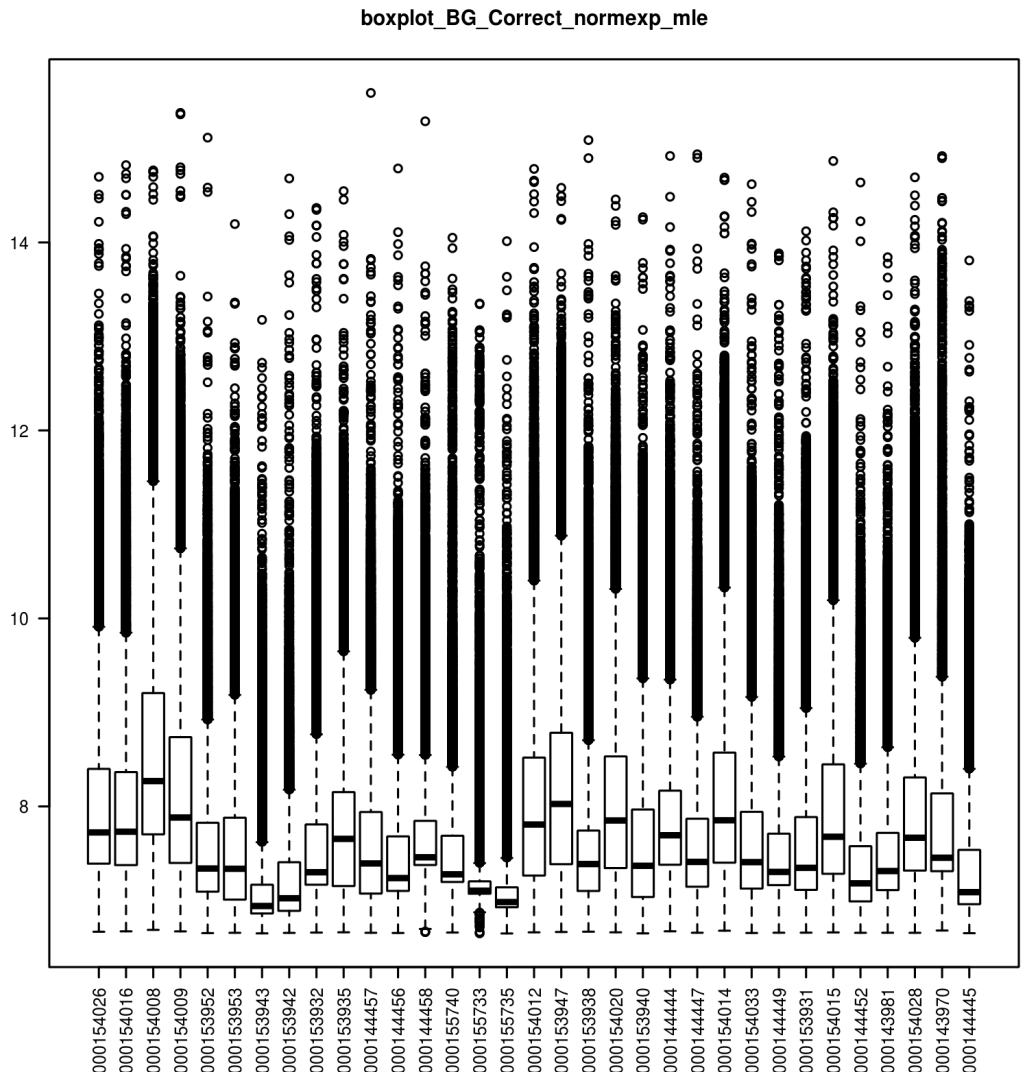


Figure 8.5: Boxplots after background correction using 'normexp+offset'. Offset=100

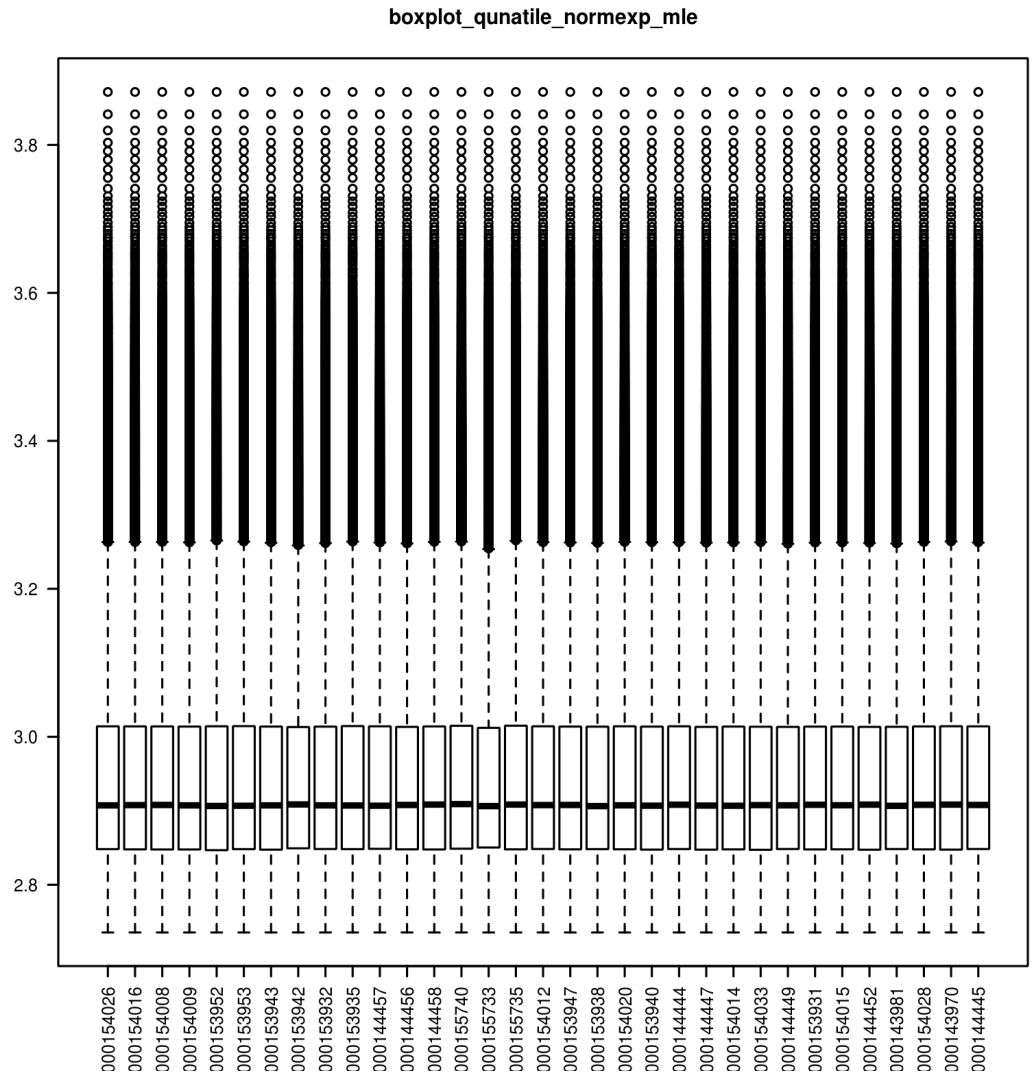


Figure 8.6: Boxplots after 'quantile' normalisation of background corrected raw values using 'normexp+offset'

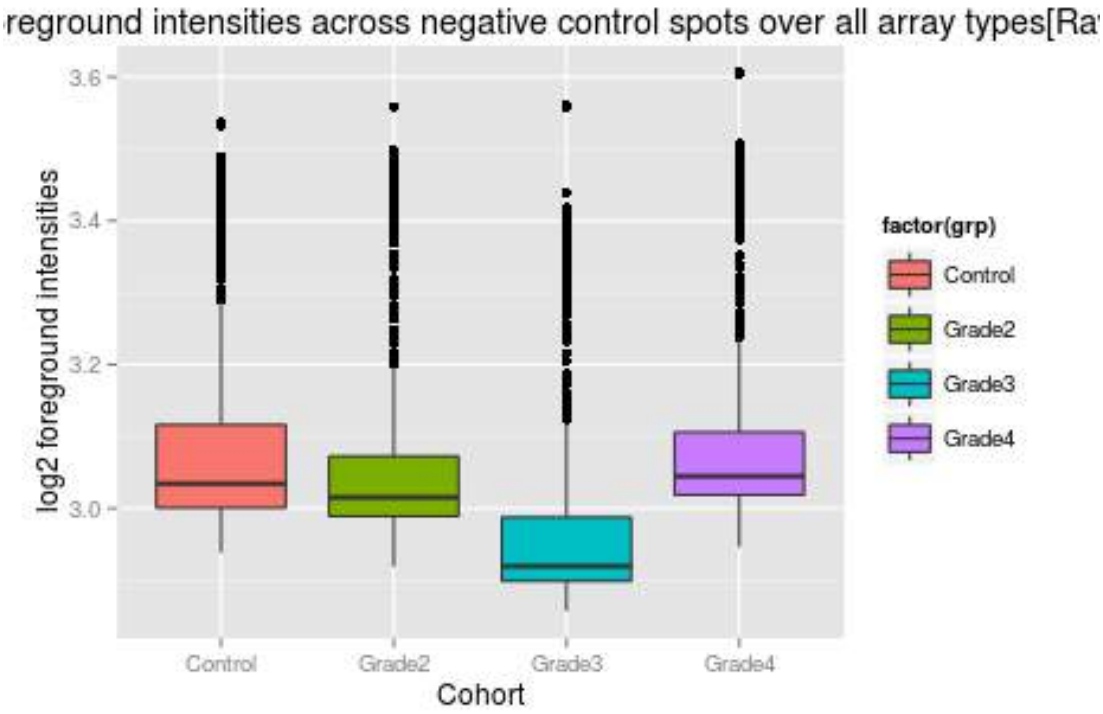


Figure 8.7: Raw foreground log2 transformed intensities across negative control spots

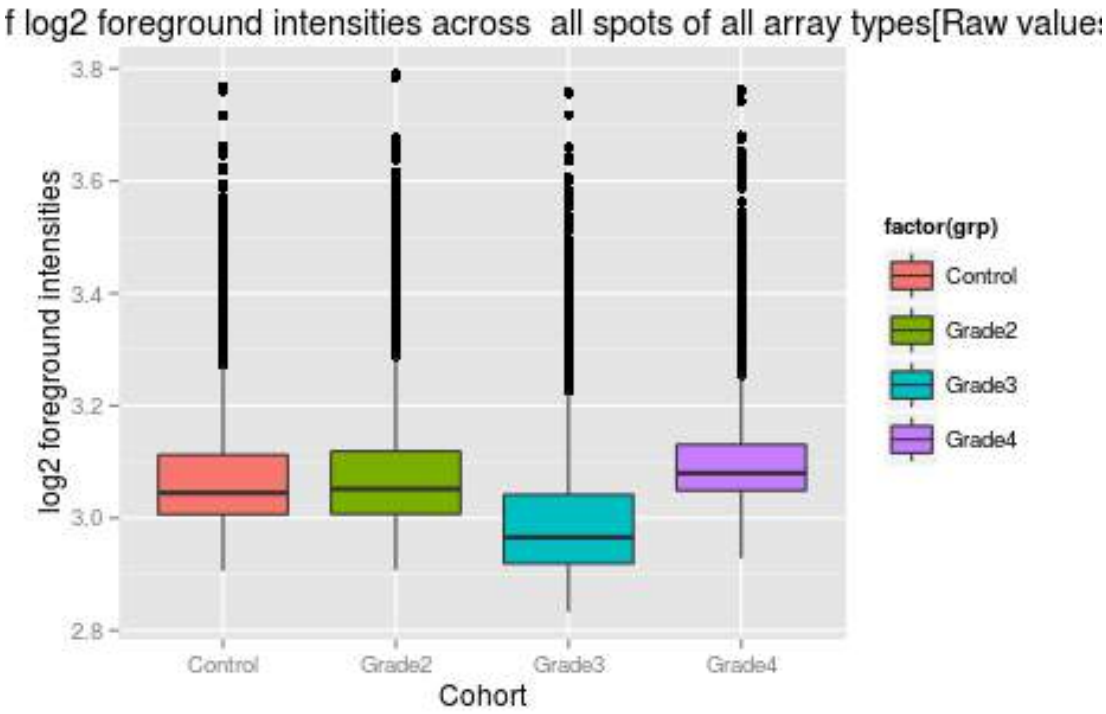


Figure 8.8: Raw foreground log2 transformed intensities across all spots

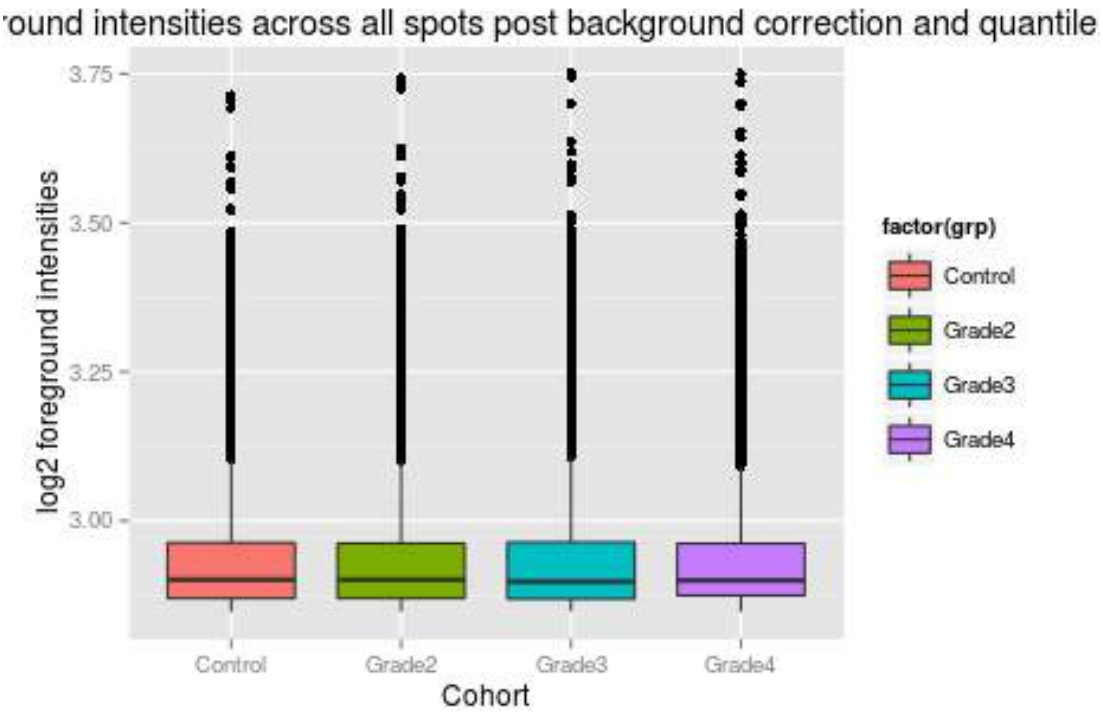


Figure 8.9: Foreground intensities post quantile normalization and background correction

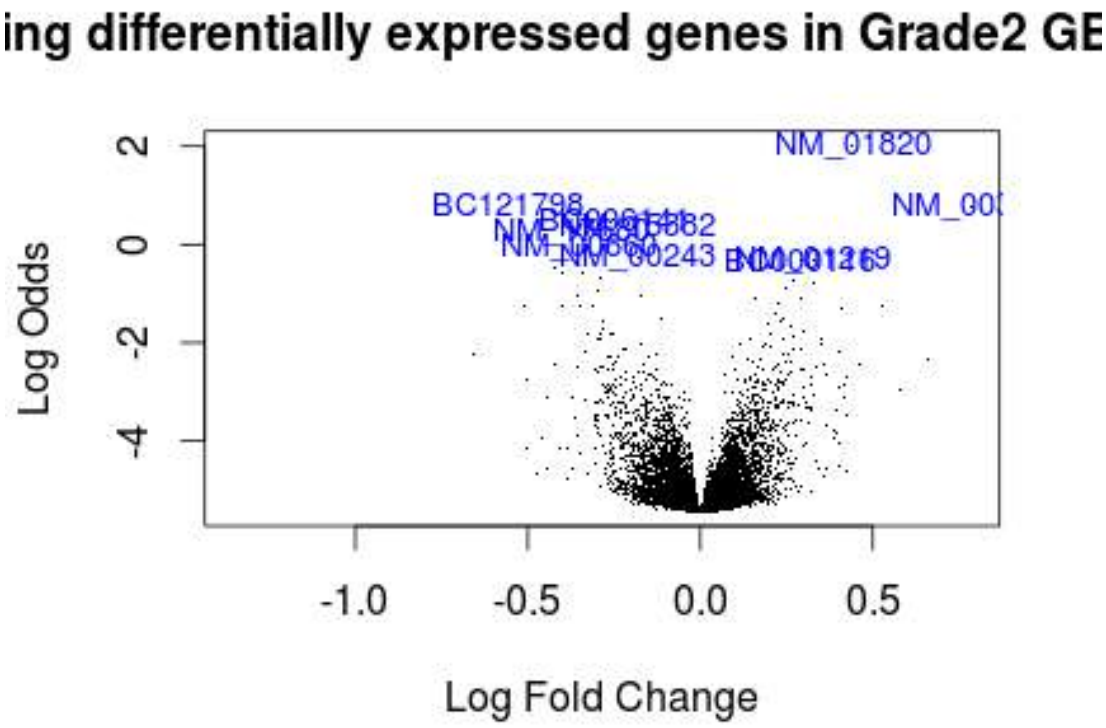


Figure 8.10: Volcano plot highlighting the statistically significant genes

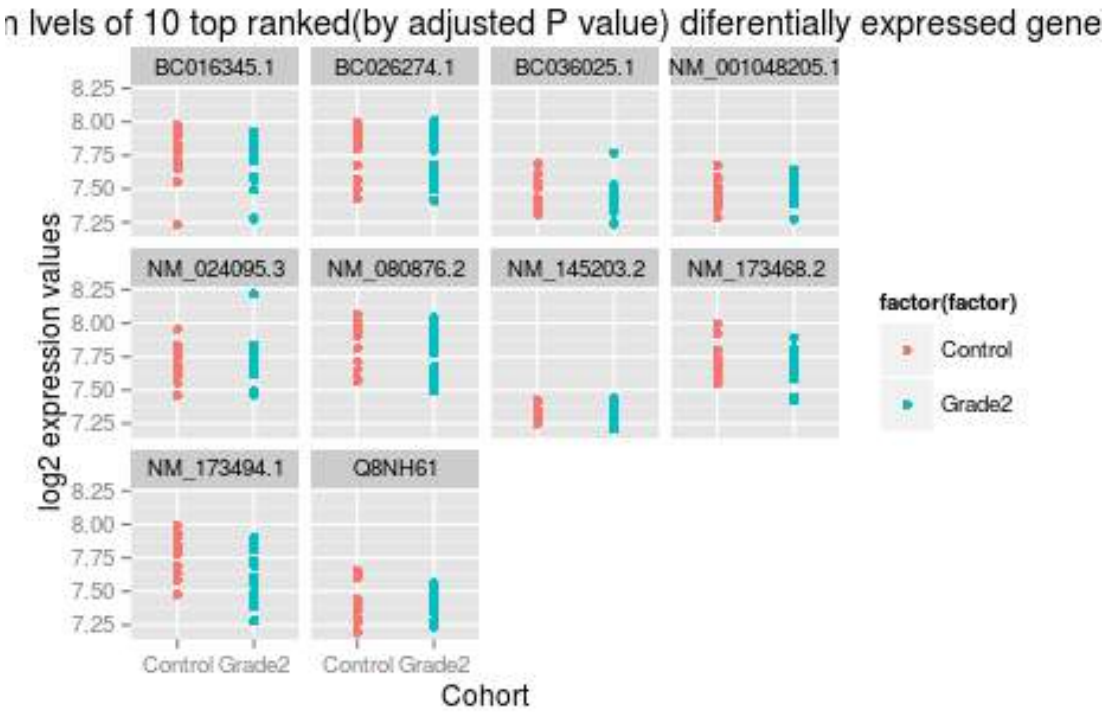


Figure 8.11: Expression levels of the top 10 differentially expressed genes

Chapter 9

Correspondence analysis

Motivation

The set of differentially expressed gene might be large on it's own. A large list of genes though may hemp in differentiating the samples[class wise], but may not be an economical design. As we discussed earlier, a list of top ranked such genes might be too uninformative and might not give a solution our problem.

The shortlist of these differentially expressed is not *short* itself. In order to reduces this dimensionality, we employ a dimensionality reduction technique. The aim is to project the data in higher dimensions to lower dimensions, without affecting the overall relationships.

Correspondence analysis is one such dimensionality reduction method. The aim in such methods is not to come with a classifier that might organize data into two separate classes, but to come up with a projection in a lower-dimension, where more relevant features can be further shortlisted.

The shortlist coming from a correspondence analysis can be treated as a set of markers whose values are associated with the classes in a statistically significant way rather than by mere chance. Hence there is strong reason to believe that this further shortlisting will help make the set of markers smaller, while at the least retaining the accuracy, if not improving.

9.1 Introduction

Correspondence Analysis is a multivariate statistical technique applied to nominal variables [7].

Let $N = IxJ$ denote the data matrix. Converting the N matrix to P such that:

$$P = \frac{N}{\sum_i \sum_j n_i j} \tag{9.1}$$

The *row masses* are represented by:

$$r_i = \sum_{j=1}^{J} p_i j \tag{9.2}$$

The *column masses* are represented by:

$$c_j = \sum_{i=1}^{I} p_i j \tag{9.3}$$

For row and column masses, the diagonals are given by:

$$D_r = diag(r) \tag{9.4}$$

$$D_c = diag(c) \tag{9.5}$$

Algorithm

Distance between two rows i and i' is given by:

$$d^2(i, i') = \sum_{j=1}^{J} \frac{1}{c_j} (\frac{p_{ij}}{r_i} - \frac{n_{i'j}}{r'_i})^2 \tag{9.6}$$

Column wise distance can be defined similarly.

These distances are not euclidean distances, but euclidean distances weighted by the inverse of the corresponding frequency. This kind of weighting ensures that the distances are *standardized* variance-wise. Thus, for larger proportions or smaller proportions, the differences are standardized.

Even if the rows i and i' are replaced by their sum of rows, then distances between columns would not change. The same holds for the columns being replaced by their column sum. The data table can now be though of as a cloud of points.Cloud of points $N(I)$ is the set of elements $i \in I$ with mass $= r_i$ and similarly cloud of points $N(J)$ is the set of elements $j \in J$ with mass $=c_j$.These both clouds of points have masses adding up to one each. Distances can also be defined for both the set of cloud of points as shown above. The inertia for i^{th} row profile is thus defined as:

$$Row inertia = Row mass * Square of distance from the centroid of the rows \tag{9.7}$$

9.1.0.1 The significance of Chi-squared distance

The underlying hypothesis for CA is that the rows and columns are independent. In a contingency table the theoretical value of a cell at (i, j) is given by, assuming the abov hypothesis is true :

$$E_{i,j} = r_i * c_j \tag{9.8}$$

However the *observed* value at (i, j) is p_{ij}. Thus the Chi-square distance is alculated as :

$$\chi^2 = n \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \tag{9.9}$$

Consider the centroi z of the row vector points:

$$z = [c_1, c_2, ..., c_J] \tag{9.10}$$

The distance between any i^{th} row and it's centroid is given by, using the distance relation between rows from above:

$$d_{iz}^2 = \sum_{j=i}^{J} \frac{(\frac{p_{ij}}{r_i} - c_j)^2}{c_j} \tag{9.11}$$

which can be rewritten in terms of the centroid $\mu_{ij} = r_i c_j$ as:

$$d_{iz}^2 = \frac{1}{r_i} \sum_{j=i}^{J} \frac{(p_{ij} - \mu_{ij})^2}{\mu_{ij}} \tag{9.12}$$

Thus row inertia:

$$r_i d_{iz}^2 = \sum_{j=i}^{J} \frac{(p_{ij} - \mu_{ij})^2}{\mu_{ij}} \tag{9.13}$$

The column inertia can be defined similarly.
Consider the residual matrix S:

$$S_{ij} = |\frac{p_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}}}| \tag{9.14}$$

In order to decompose S to lower dimensions consider SVD decomposition of S:

$$S = U D_\alpha V^T \tag{9.15}$$

where U,V are orthonormal $VV^T = 1$ and $UU^T = 1$ and D_α is a diagonal matrix with entries in descending order as λ_1, λ_2,....

The scores of the rows is then given by:

$$F = D_r^{\frac{-1}{2}} U D_\alpha \tag{9.16}$$

and the column scores are given by:

$$G = D_c^{\frac{-1}{2}} V D_\alpha \tag{9.17}$$

The dimension of these score matrices is $min(I-1, J-1)$ and essentially represent the *coordinates* of these row vectors in the higher-dimensional subspace. The points in this space are so arranged that the euclidean distances between two points corresponds to the Chi-square distance in the original matrix. Even though these scores are representative of the original rows, in terms of the Chi-square distance, visualizing graphically is possible only in 2 or 3 dimensions. Hence we consider only the first two components of each row to visualize them on a 2D surface. The distance between the points on the graph is a close estimate of the original Chi-square distance.

In order to quantify the amount of inertia represented by this plot, we consider the following score:

$$\phi^2 = \sum_{i=1}^{I} r_i d_{iz}^2 \tag{9.18}$$

and the amount of inertia captured by he first two principal axes is given by:

$$\frac{\lambda_1^2 + \lambda_2^2}{\phi^2} \tag{9.19}$$

The row and column scores can be plotted in one 2D graph after proper scaling and is called as a *CA biplot*. On such a biplot the relationships between the row points and column points can be inferred the following way:

- Given a row point X and a column point Y, if the angle between line joining the centroidorigin of the 2D plot and line joining Y and centroid is acute, it essentially points that association between X and Y is high. A right angle denotes zero association and an obtuse angle denotes negative association
- The distance on biplot are proportional to χ^2 distances in the original higher dimension
- The farther away a point is from the centroid, the higher is that row's contribution to the value of statistic

Discussions

We make use of correspondence analysis to find genes that are associated strongly with either the Controls or the disease(GBM) samples, treating one grade at a time.

Figure 9.1: Correspondence Analysis of Grade4 samples as compared to Controls. The genes located along the diagonals have association with the Grade4/Control samples. Association can be negative or positive. Control and Grade4 samples are separated along the second axis. However the separation is not distinct.

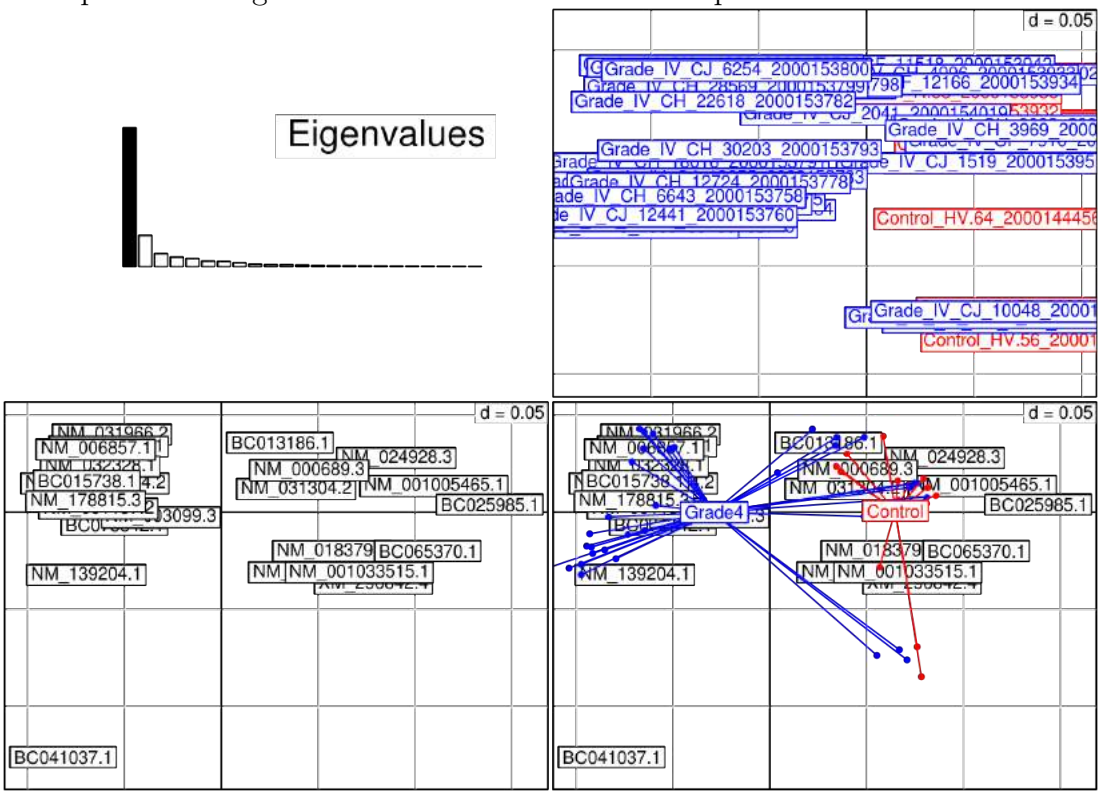
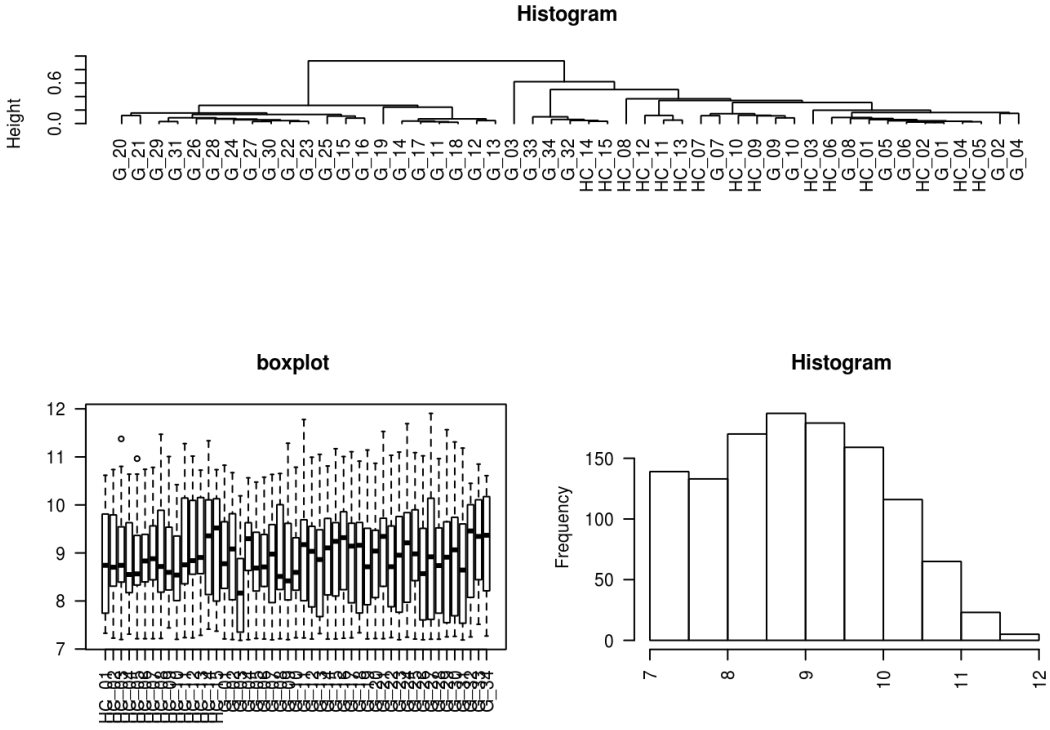


Figure 9.2: Hierarchical clustering with Average linkage for Grade4 and Control samples. Though there are two distinct clusters, there is an intermixing of groups too



Chapter 10

Classification of Microarray Data

Problem

Given the datasets have already been classified into two classes: *Control* and *Grade II*. The focus of the underlying problem is to come up with a marker set of genes, which can be used to differentiate new samples into Control and Disease. This belongs to the class of *supervised learning* problems, where the classes to which each sample should belongs to is available. Another class of problems deals with *unsupervised learning*, where the focus is to discover the classes to which the data can be classified into. An example of this would be classifying the disease samples into further sub-categories(Grade II, Grade III, Grade IV). We tackle the former here.

We follow the approach as shown in to solve our problem:

10.1 Curse of Dimensionality : Feature Selection

Given the large number of *attributes* associated with a microarray study with often a small number of observations, it becomes difficult to classify, since large number of observations are required for several combinations of attribute values.

In order to prevent the *curse of dimensionality*, the set of genes can be restricted. The rationale behind using correspondence analysis was to reduce the set of attributes to those genes which have been found to have strong association with the 'control' or 'disease' classes. Thus rather than feeding the set of 500 different attributes, we consider only 100 genes which are shown to have strong association with the control and disease samples. These set of 100 attributes are thus fed to a SVM classifier to build a binary classifier model.

However it is till necessary to determine which set(s) of genes are the most relevant for building a classifier model. The *best* features can be selected either by employing *filter* methods such as signal to noise ratio [20] or *wrapper* methods such as *recursive feature elimination(RFE)*. We discuss *RFE* in detail here.

Recursive Feature Selection

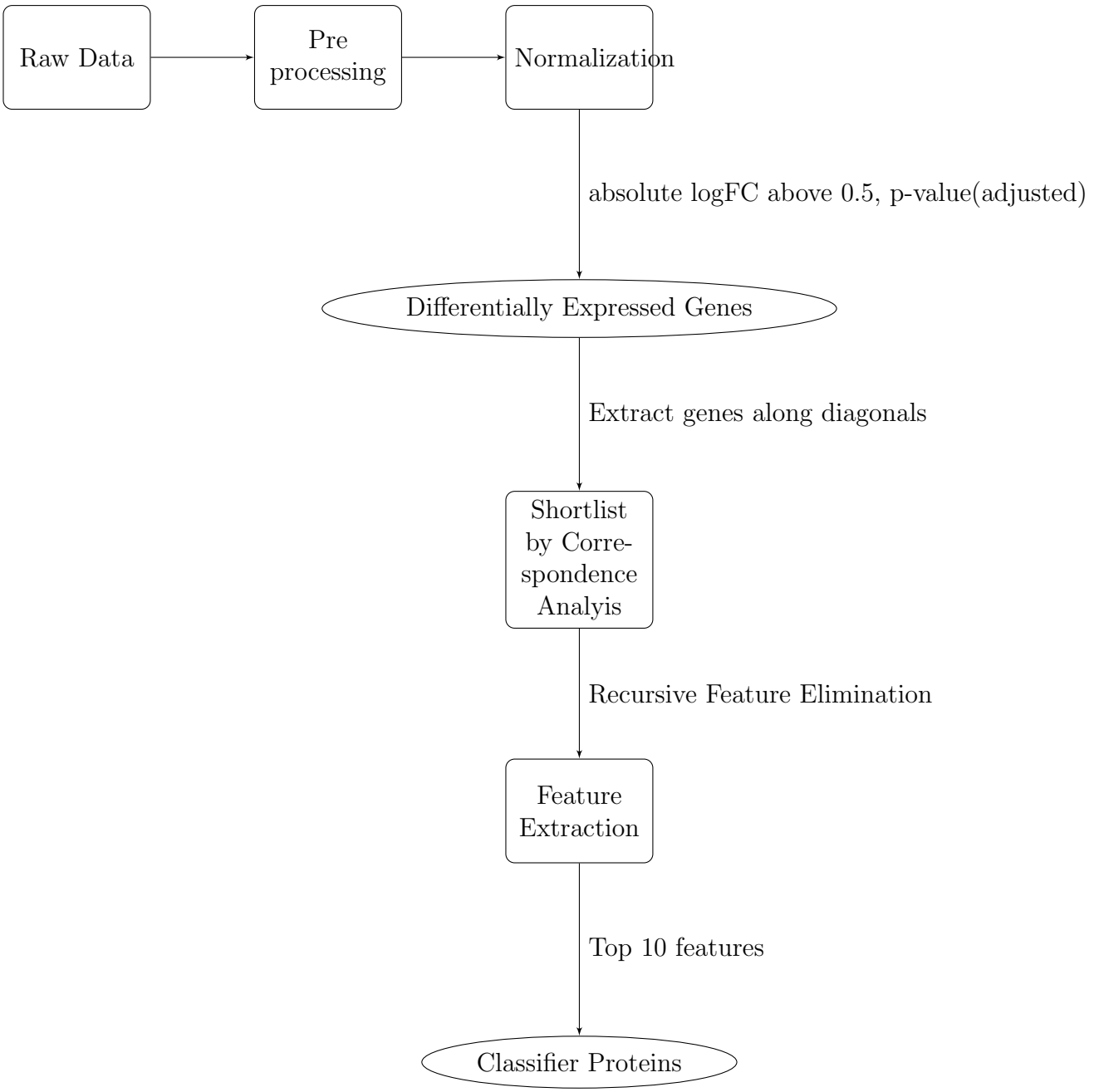
RFE is based on a simple principle of eliminating those w_i that have low magnitude, since the contribution of such elements contribute less to the classification function. For a linearly separable dataset, this procedure is executed by first calculating the weights of all attributes :

$$S_j = |w_j| \tag{10.1}$$

These weights are then sorted in descending order and one or more features are removed from the bottom of the list. Using the shorter set of attributes, a SVM is re-trained. And this procedure is repeated till we settle down to a pre-decided length of features.

10.2 SVM Classification

Support Vector Machines are binary classifiers. Given a training set of (points,labels) (x_i, y_i) where $x_i \in \mathbf{R}$ and $y \in -1, 1]$. The idea is to search for a hyperplane that would separate the points with $y_i = 1$ from $y_i = -1$. There could be multiple hyperplanes like that, the focus is however only on the hyperplane that with


Raw Data
Pre processing
Normalization
absolute logFC above 0.5, p-value(adjusted)
Differentially Expressed Genes
Extract genes along diagonals
Shortlist by Correspondence Analyis
Recursive Feature Elimination
Feature Extraction
Top 10 features
Classifier Proteins

maximum-margins(on both sides). Any such hyperplane satisfies:

$$w.x - b = 0 \tag{10.2}$$

If the data is linearly separable, two hyperplanes can be found :

$$w.x - b = 1 \tag{10.3}$$

$$w.x - b = -1 \tag{10.4}$$

The distance between the two hyperplanes is $\frac{2}{||w||}$. Thus minimizing $||w||$ would yield the required the hyperplane.

In order to prevent misclassification, the following constraints are required:

$$(w.x_i - b) \geq 1 for x_i belonging to class 1 \tag{10.5}$$

and

$$(w.x_i - b) \leq -1 for x_i belonging to class -1 \tag{10.6}$$

which can be combined as:

$$y_i(w.x_i - b) \geq 1 \tag{10.7}$$

and the objective function to be minimised under this constraint is : $||w||$

10.3 Cross Validation and SVM

Given the small set of observations, coming up with a predictor and testing it on the same dataset will present a rosy but a wrong picture. This is the classical problem of *over-fitting* where the prediction function work with maximum accuracy on the training dataset, however will often perform not so well for an entirely new dataset (validation dataset). In problem like the microarray experiments where the training data is limited, it is often difficult to define a separate validation dataset. Thus in order to avoid over-fitting, part of the training data set is with held as test dataset. Consider k-fold validation. the model is trained using k-1 observables of the dataset and the remaining 1 dataset serves as the test dataset. This performance is averaged over all possible values of the k-1 substitutes of the data.

The SVM employed uses a linear kernel. The choice of a linear kernel over other non-linear kernels is justified, since the data matrix is till in higher dimensions[49]

10.4 Results and Discussions

We used RFE to select a list of 30 genes that could potentially be used as a set of biomarker for prognosis.

These set of proteins are:

A detailed table is present in Appendix 1.

Here we focus on the pattern of classification, rather than the classifiers themselves. The 'Brier' score measure the accuracy of probabilistic predictions. It measures the mean squared error between the predicted probability and the actual observation. Thus, a smaller score indicates a more accurate classifier.

A pattern worth noticing is how, a larger number of features does not always guarantee a lower brier score and hence a higher accuracy. This may look paradoxical, but the paradox is resolvable.

The argument goes like this.: A new gene included in the shortlisted set might not be equally informative as the rest. In many cases in fact introduction of a new gene as an attribute leads to poor classification accuracy as the gene is *noisy*. This discussion is part of Golub et. al [20] where the coefficient used for ranking the genes is given by:

$$w_i = \frac{\mu_i(+) - \mu_i(-)}{\sigma_i(+) - \sigma_i(-)} \tag{10.8}$$

where μ_i represents the mean and σ_i represents the standard deviation of gene i expression values for two class of samples, $+$ and $-$. Larger w_i represents stronger association of gene i with class $+$. As a first step towards selecting the best features, one could shortlist the gene with larger w_i or lower w_i. The intermediate values of w_i do not led to informative features.

However in the above discussion an implicit assumption has been that the features are *orthogonal* (independent) to each other. Each gene i is ranked individually , assuming complete independence of the genes. Thus neglecting any kind if co-relation that might exist between two or more genes, possibly due to them being part of the same pathway.

The *RFE* method of feature selection tackles this issue by assigning weights to each attribute, tackling all the attributes at once. The genes are then ranked with the genes with maximum weights being ranked at the top. There is no implicit or explicit orthogonality assumption here.

A note on the results

The shortlisted gene list after pre-processing are all summarized in Appendix 1, append 2 and Appendix 3.

"Number of features"	Brier	Specificity	Sensitivity
1	0.427	0.067	0.971
2	0.449	0.000	1.000
3	0.355	0.200	0.882
4	0.269	0.667	0.853
5	0.176	0.867	0.941
6	0.140	0.867	0.941
7	0.180	0.867	0.941
8	0.204	0.667	0.941
9	0.138	0.800	0.941
10	0.134	0.800	0.941
11	0.123	0.867	0.971
12	0.125	0.800	0.971
13	0.121	0.800	0.971
14	0.114	0.800	0.971
15	0.149	0.867	0.882
16	0.171	0.733	0.941
17	0.175	0.800	0.912
18	0.173	0.800	0.882
19	0.151	0.800	0.941
20	0.124	0.867	0.971
21	0.118	0.800	0.941
22	0.109	0.933	0.941
23	0.119	0.933	0.941
24	0.133	0.867	0.941
25	0.155	0.800	0.941
26	0.167	0.800	0.912
27	0.141	0.933	0.941
28	0.146	0.800	0.941
29	0.148	0.867	0.941
30	0.152	0.867	0.941

Figure 10.1: Features and their Brier scores for Control v/s Grade4

However, there has been no step of validation involved either experimentally or a through a thorogh literature search to verify if the shortlisted sets of genes are known to be associated in any way. A naive approach would have been to lookup which categories of MeSH [http://www.nlm.nih.gov/mesh/meshhome.html] do the articles citing this fall under. This would essentially point out if there has ben an earlier study on these sets of genes related to cancer. However, we decided against such an approach.

Chapter 11

Visualisation tools for Bioinformatics

Need/Rationale

With the advent of Next Generation Sequencing, there has been a boom in the amount of data being generated. This vast amount of biological data presents a great challenge from the point of understanding and interpretation. Availability of visualization tools can act as a possible solution by utilizing the bandwidth of human vision for interpretation[50].

11.1 Phred Score Viewer

Phred scores are the quality scores assigned to the nucleotides as sequenced by automated sequencers. Phred scores are an indicator of the quality of the fastq files. See Chapter 2 for a discussion on fastq files and phred scores.

11.1.1 Implementation details

The phred score viewer is a pure javascript based implementation. Thus, it can be used to render box plots of the quality scores via web browsers, independent of the platform of the user.The user can draw conclusions about the distribution of the scores by looking at the box plots.

11.2 Human Genetic Variation Viewer

No two humans are genetically identical,even though there sequences are 99.9% similar. The genetic difference arise at various sites and are common both within

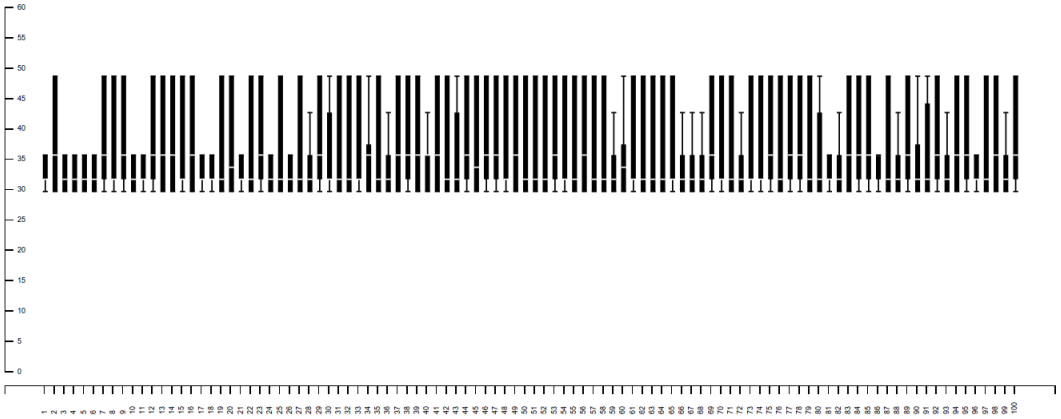


Figure 11.1: Box plot plotted using a javascript based phredscore plotter, for a 100-read based sequence

and among populations. A study of these genetic variations has both evolutionary and medical significance and can essentially help in deciphering the impacts of variations across populations.

Most of these variations have now been cataloged into databases such as the dbSNP, which is a catalog of the human single nucleotide polymorphisms. However, the amount of data is too large and too scattered to draw any conclusion from it.

Human genetic variation viewer, is an effort that tackles this very problem. With a visual map of genetic variations, it is not only easy to visualize all the genetic variations at once, but at the same time draw inferences. the user can configure the viewer to load a certain protein and can visualize variants which have been scored for their impacts, for example via SIFT and Polyphen algorithms. These values indicate the deleteriousness of a certain variant position.

11.2.1 Implementation details

Human genetic variation viewer is also a javascript based implementation that fetches the list of variants from a web service, given the Entrez ID. The map is a stacked bar chart highlighting the damaging, benign and intermediate variants at the protein level.

11.2.2 Conclusion

Given the platform independence and scalability of the tools, they can add utility to biological data visualization, making it simple to infer quality scores and study

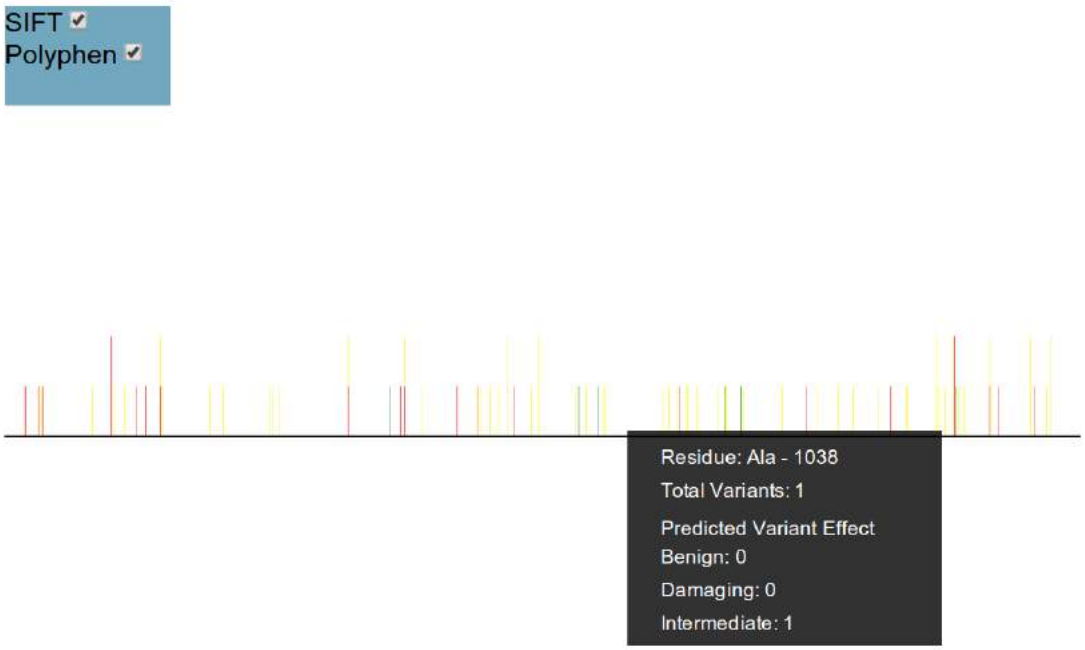


Figure 11.2: Stacked bar charts showing the frequency of damaging(red), benign(green) and intermediate(yellow) mutations in a protein

the impact of various mutations. The tools are designed to be configurable by the end user, and hence are user-friendly.

Chapter 12

Conclusions

Over the course of this project we demonstrated the presence of HPV in Cervical cancer patients, by computational techniques. The knowledge about sites of integration could be potentially be used as a therapeutic target.

We also benchmarked two algorithms for aligning next generation sequencing whole(exome) sequencing data. BWA shows better performance over BWA-PSSM, possibly because BWA-PSSM models the rror qualities that are not applicable in general.

We built a toolbox using a set of pre-existing tools for predicting driver and passenger mutations. These set of tools are thus available to use now, through Galaxy. We also built a set of workflows that allow the conversion of VCF files to each tool's required input format. The user can thus run multiple tools on the same dataset without the need of pre-processing every time. This whole framework also provides a heatmap visualization of the output, thus making the whole analysis more easily interpretable.

The microarray problem required a small set of markers that could potentially be used to differentiate the various stages of glioma. For each stage, we generate such a list using recursive feature elimination and k-fold cross validation. We however, do not re-validate our findings by checking the biological significance of these genes.

We also develop two visualization tools for visualizing phred scores in a fastq file and human genetic variations.

All the code for the project has been Open Sourced:

- **NGS Scripts**: https://github.com/saketkc/NGS-Stuff
- **Microarray Scripts**: https://bitbucket.org/saketkc/proteomics_analysis
- **Galaxy Toolbox** : https://github.com/saketkc/galaxy_tools

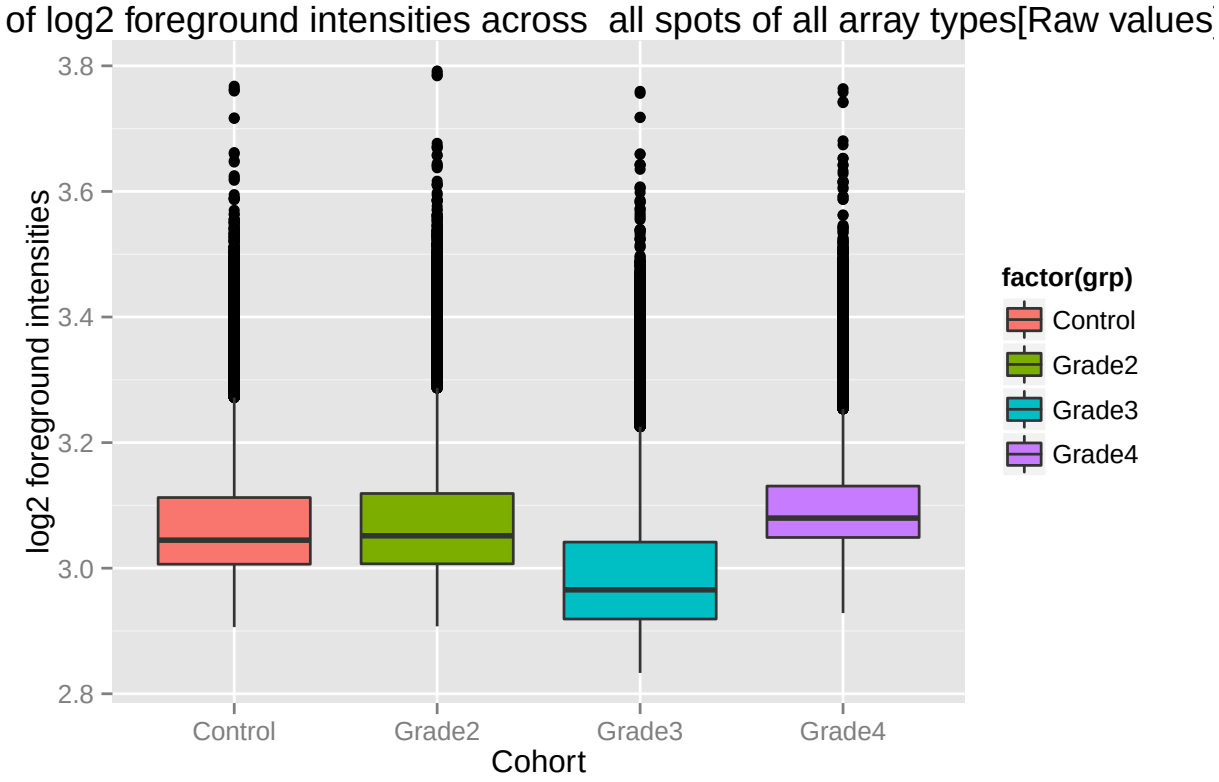
Appendix 1: Analysis of GBM Grade4 samples vs Control

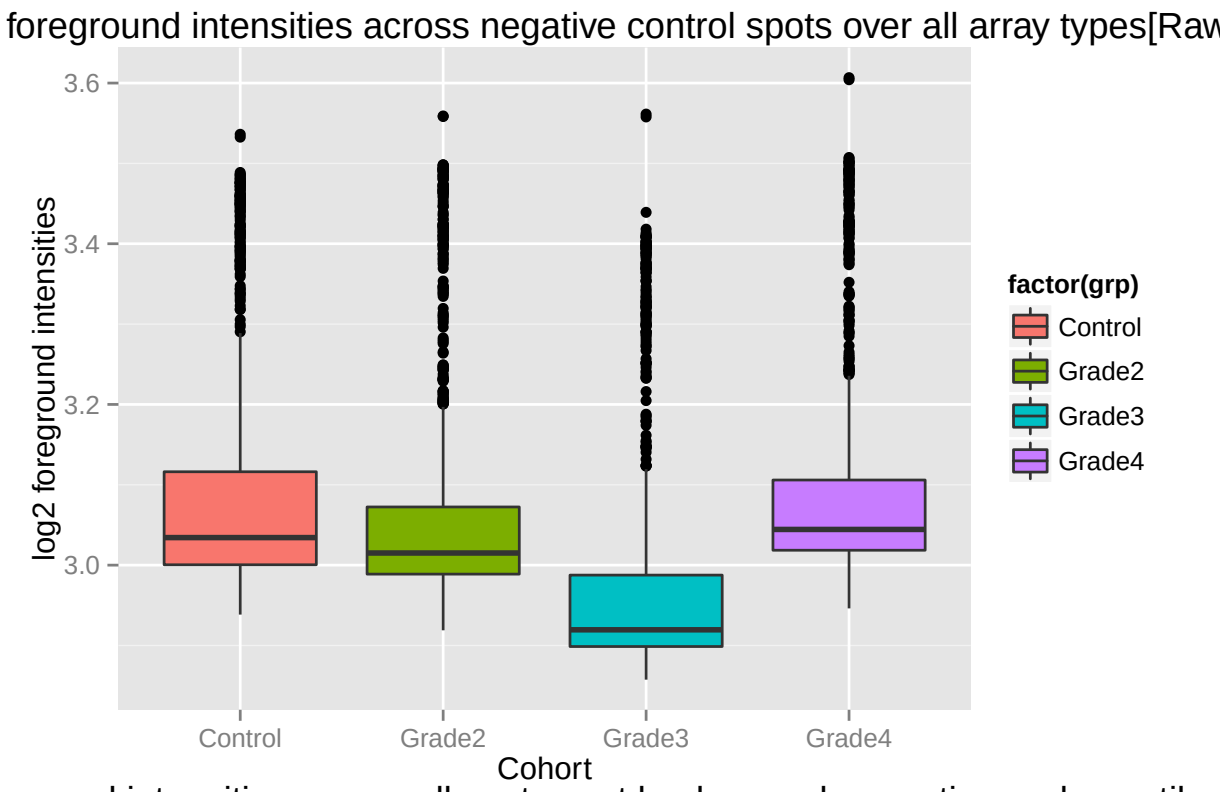
Grade4vsControl

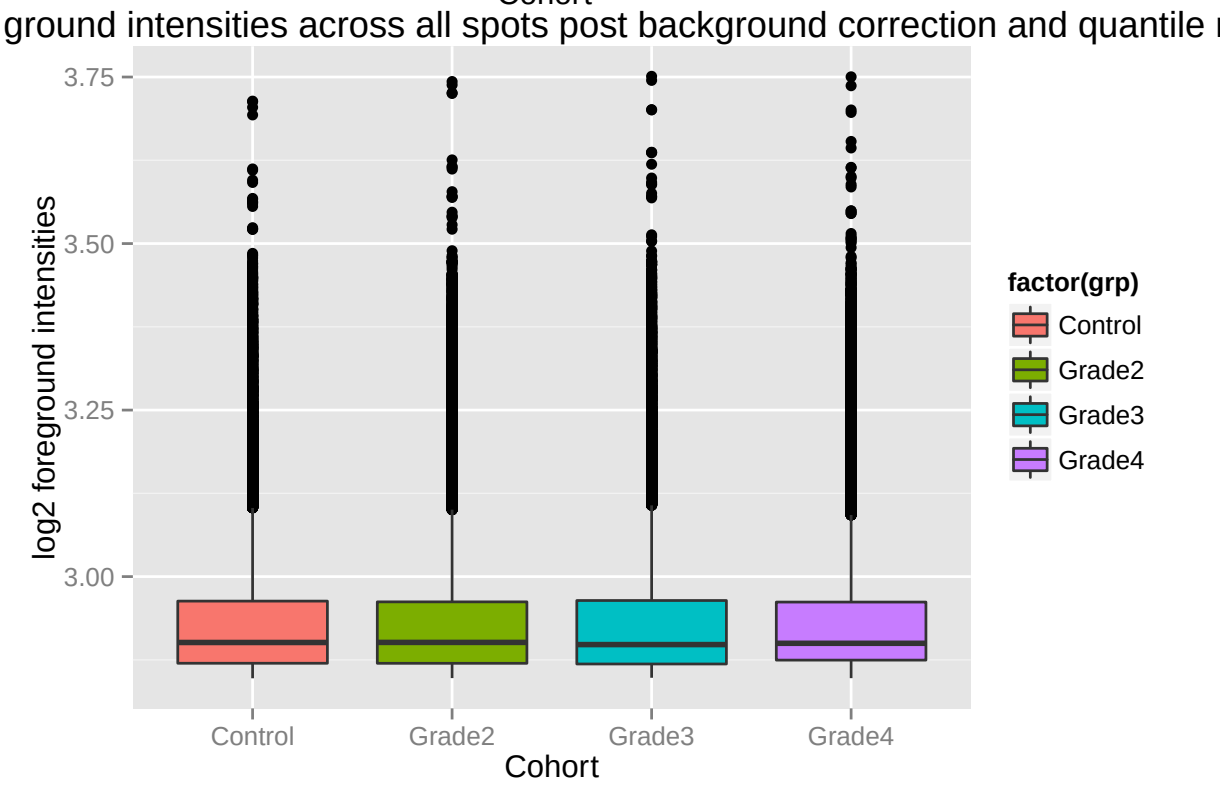
Saket Choudhary

Wednesday 25 June 2014

```
## Loading required package: statmod
```

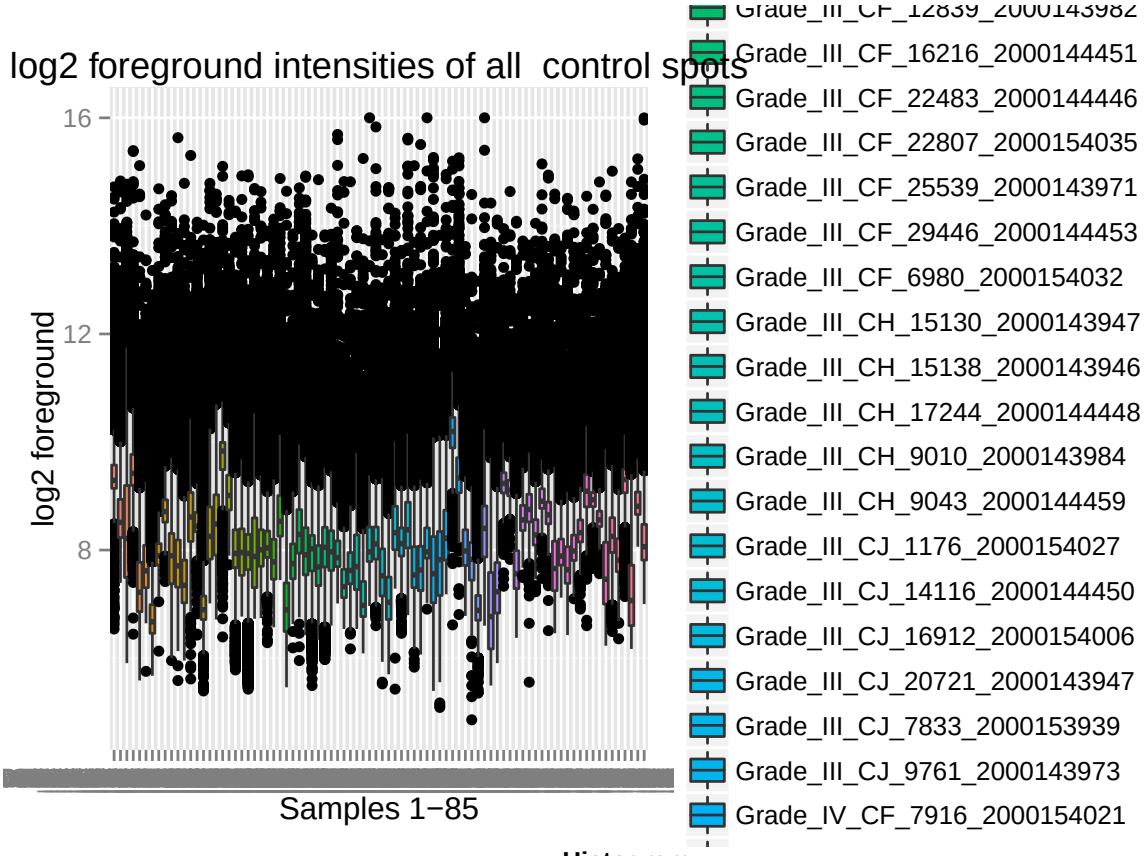


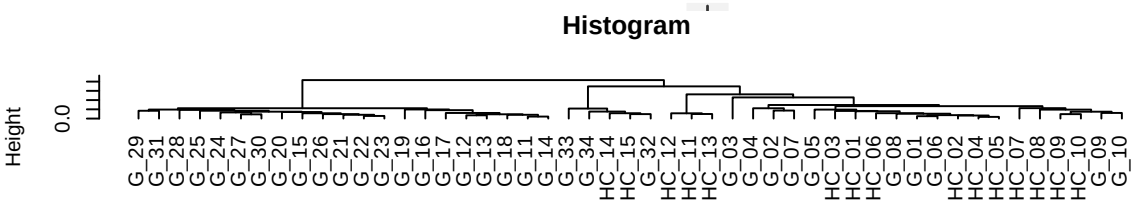


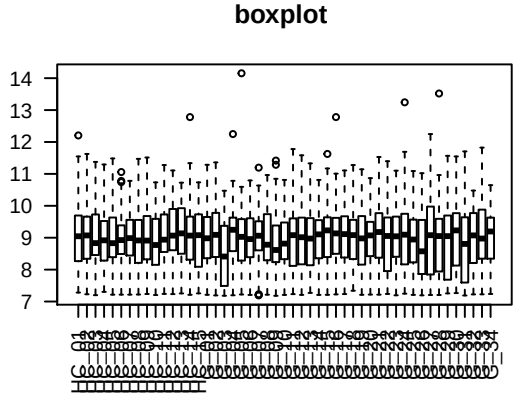


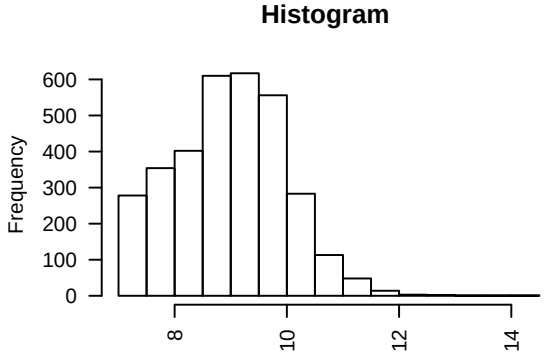

```
## No id variables; using all as measure variables
```

```
## Warning: Removed 128 rows containing non-finite values (stat_boxplot).
```


log2 foreground intensities of all control spots
log2 foreground
16
12
8
Samples 1−85
Grade_III_CF_12839_2000143982
Grade_III_CF_16216_2000144451
Grade_III_CF_22483_2000144446
Grade_III_CF_22807_2000154035
Grade_III_CF_25539_2000143971
Grade_III_CF_29446_2000144453
Grade_III_CF_6980_2000154032
Grade_III_CH_15130_2000143947
Grade_III_CH_15138_2000143946
Grade_III_CH_17244_2000144448
Grade_III_CH_9010_2000143984
Grade_III_CH_9043_2000144459
Grade_III_CJ_1176_2000154027
Grade_III_CJ_14116_2000144450
Grade_III_CJ_16912_2000154006
Grade_III_CJ_20721_2000143947
Grade_III_CJ_7833_2000153939
Grade_III_CJ_9761_2000143973
Grade_IV_CF_7916_2000154021


Histogram
Height
0.0
G_29
G_31
G_28
G_25
G_24
G_27
G_30
G_20
G_15
G_26
G_21
G_22
G_23
G_19
G_16
G_17
G_12
G_13
G_18
G_11
G_14
G_33
G_34
HC_14
HC_15
G_32
HC_12
HC_11
HC_13
G_03
G_04
G_02
G_07
G_05
HC_03
HC_01
HC_06
G_08
G_01
G_06
HC_02
HC_04
HC_05
HC_07
HC_08
HC_09
HC_10
G_09
G_10


boxplot
14
13
12
11
10
9
8
7

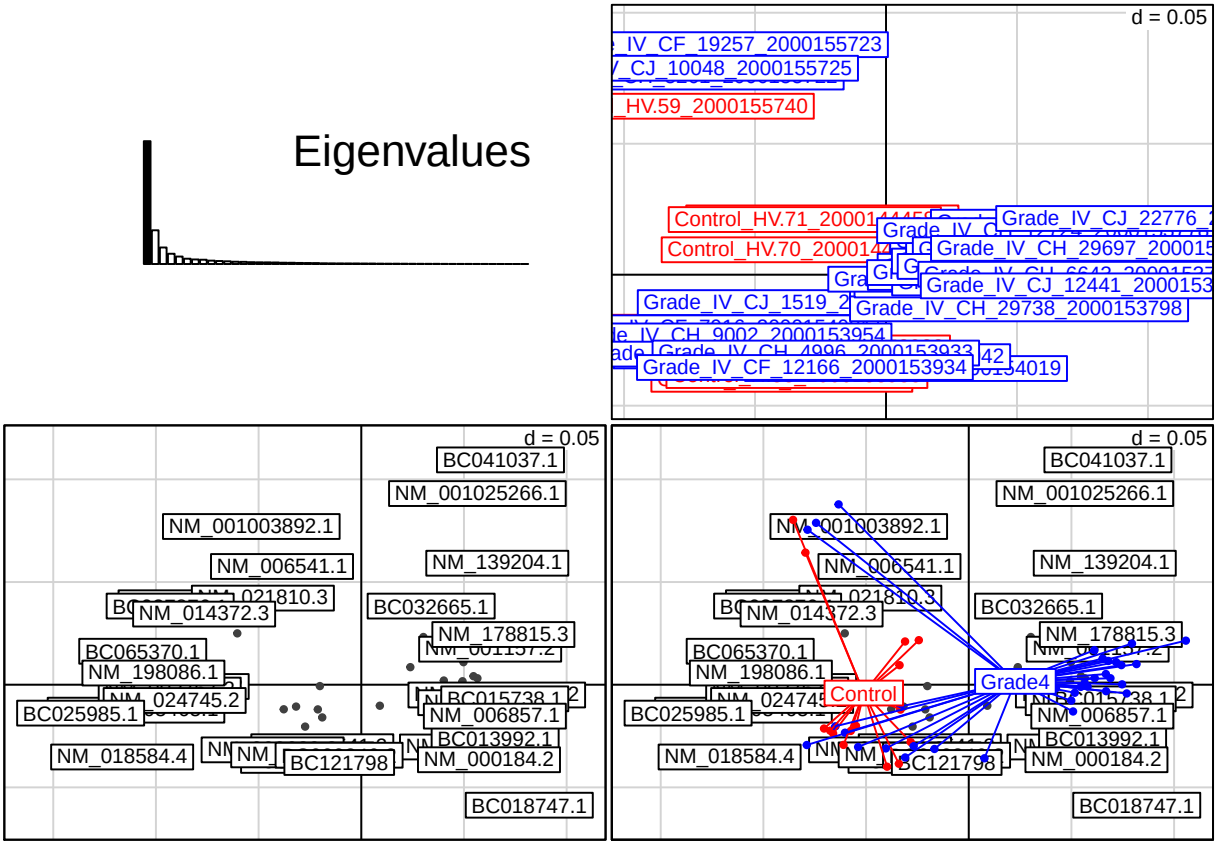

Histogram
Frequency
600
500
400
300
200
100
0
8
10
12
14

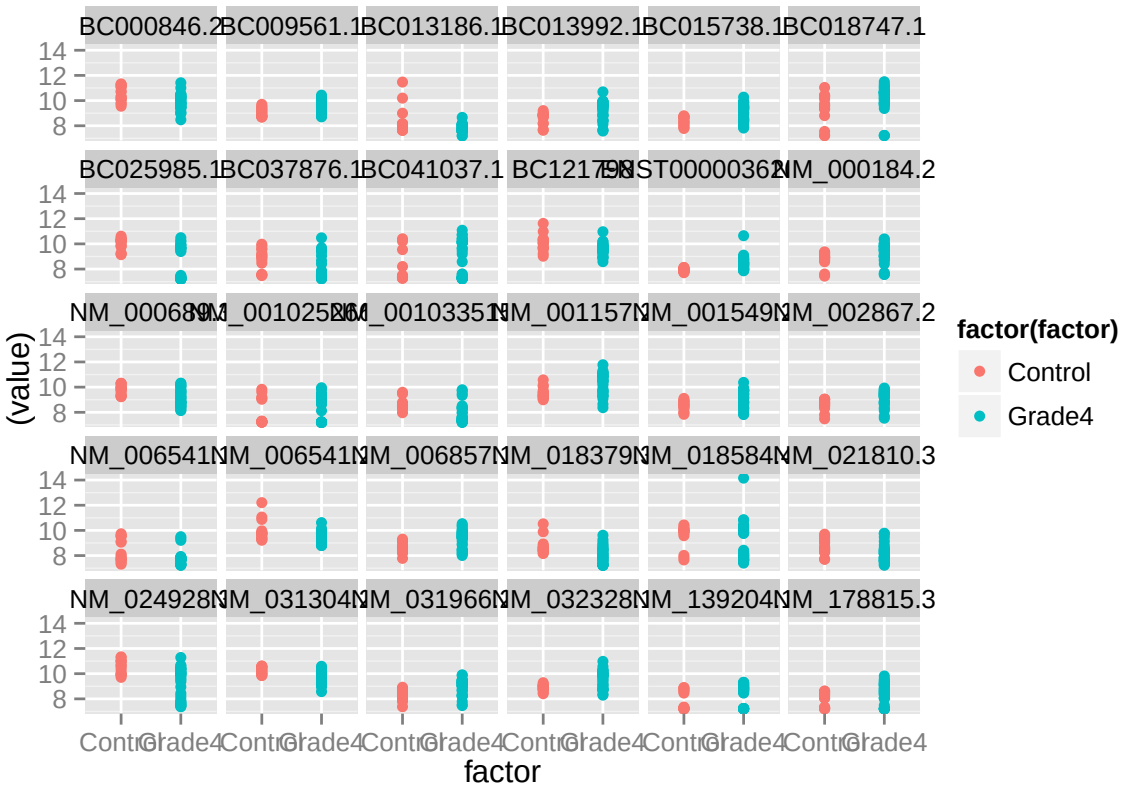
```
## NULL
```

```
## NULL
##  [1] "NM_031966.2"    "NM_002867.2"    "NM_001827.1"    "BC075842.1"
##  [5] "BC041037.1"     "NM_018379.3"    "NM_001005465.1" "BC065370.1"
##  [9] "NM_024928.3"    "NM_148910.2"    "NM_002767.2"    "NM_018584.4"
## [13] "NM_198086.1"    "XM_290842.4"    "NM_006541.1"    "NM_021810.3"
## [17] "BC037876.1"     "NM_001003892.1" "NM_014372.3"    "NM_001033515.1"
## [21] "NM_024745.2"    "NM_017924.2"    "NM_032347.1"    "NM_001001394.2"
## [25] "BC018747.1"     "NM_032328.1"    "NM_001025266.1" "BC013992.1"
## [29] "NM_031304.2"    "NM_021979.2"    "NM_000689.3"    "BC013009.2"
## [33] "BC015738.1"     "BC025985.1"     "NM_000184.2"    "NM_006857.1"
## [37] "NM_001157.2"    "NM_178815.3"    "NM_001549.2"    "NM_139204.1"
##  [1] "NM_031966.2"     "NM_000884.2"     "NM_002867.2"
##  [4] "NM_001827.1"     "NM_006541.2"     "ENST00000362035"
##  [7] "BC075842.1"      "BC041037.1"      "NM_018379.3"
## [10] "BC065370.1"      "NM_018584.4"     "XM_290842.4"
## [13] "NM_006541.1"     "NM_021810.3"     "BC037876.1"
## [16] "NM_001003892.1"  "NM_014372.3"     "NM_001033515.1"
## [19] "NM_017924.2"     "BC009561.1"      "BC018747.1"
## [22] "NM_001155.3"     "NM_001025266.1"  "BC013992.1"
## [25] "BC010450.1"      "BC013186.1"      "NM_000689.3"
## [28] "NM_001801.2"     "BC000846.2"      "BC121798"
## [31] "BC067254.1"      "NM_024815.3"     "NM_000184.2"
## [34] "NM_001157.2"     "BC032665.1"      "NM_001008491.1"
## [37] "BC000446"        "NM_178815.3"     "NM_001549.2"
## [40] "NM_139204.1"
##
##
## | NF|  Brier|FeatureList
## |--:|------:|:-----------------------------------------------------------------------------
## |  1| 0.4272|NM_021810.3
## |  2| 0.4494|NM_021810.3 BC037876.1
## |  3| 0.3546|NM_021810.3 BC037876.1 BC009561.1
## |  4| 0.2694|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1
## |  5| 0.1758|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2
## |  6| 0.1398|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2
## |  7| 0.1798|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4
## |  8| 0.2037|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## |  9| 0.1380|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 10| 0.1335|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 11| 0.1232|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 12| 0.1254|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 13| 0.1209|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 14| 0.1141|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 15| 0.1493|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 16| 0.1707|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 17| 0.1751|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 18| 0.1734|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 19| 0.1515|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 20| 0.1244|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 21| 0.1176|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 22| 0.1085|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 23| 0.1192|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 24| 0.1325|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 25| 0.1555|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
```

```
## | 26| 0.1671|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 27| 0.1409|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 28| 0.1457|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 29| 0.1477|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM
## | 30| 0.1522|NM_021810.3 BC037876.1 BC009561.1 NM_006541.1 BC000846.2 NM_002867.2 NM_018584.4 NM

## Using ID as id variables
```




BC000846.2
BC009561.1
BC013186.1
BC013992.1
BC015738.1
BC018747.1
BC025985.1
BC037876.1
BC041037.1
NM_000184.2
NM_006541.1
NM_006541.1
NM_006857.1
NM_018379.1
NM_018584.4
NM_021810.3
NM_024928.1
NM_031304.1
NM_031966.1
NM_032328.1
NM_139204.1
NM_178815.3
(value)
14
12
10
8
Control
Grade4
factor
factor(factor)
Control
Grade4

References

[1] GR Abecasis, David Altshuler, A Auton, LD Brooks, RM Durbin, Richard A Gibbs, Matt E Hurles, Gil A McVean, DR Bentley, A Chakravarti, et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010. 21

[2] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010. vii, x, 13, 14, 20

[3] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997. 14

[4] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000. 21

[5] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *The Journal of experimental medicine*, 79 (2):137–158, 1944. 1

[6] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik LL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl 1):D138–D141, 2004. 21

[7] JP Benzecri. coll., 1973. *L'analyse des données. Tome I: La taxinomie. Tome II: L'analyse des correspondances*, 1973. 70

[8] Hannah Carter, Sining Chen, Leyla Isik, Svitlana Tyekucheva, Victor E Velculescu, Kenneth W Kinzler, Bert Vogelstein, and Rachel Karchin. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer research*, 69(16):6660–6667, 2009. vii, 17

[9] J M Claverie. Computational methods for the identification of differential and coordinated gene expression. *Human molecular genetics*, 8(10):1821–32, January 1999. ISSN 0964-6906. URL http://www.ncbi.nlm.nih.gov/pubmed/10469833. 57

[10] William S Cleveland and Clive Loader. Smoothing by local regression: Principles and methods. In *Statistical theory and computational aspects of smoothing*, pages 10–49. Springer, 1996. 53

[11] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, 43(5):491–498, 2011. 34

[12] Chris Drummond. Reproducible research: a dissenting opinion. 2012. 33

[13] David Edwards. Non-linear normalization and background correction in one-channel cdna microarray studies. *Bioinformatics*, 19(7):825–833, 2003. 52

[14] Brent Ewing, LaDeana Hillier, Michael C Wendl, and Phil Green. Base-calling of automated sequencer traces usingphred. i. accuracy assessment. *Genome research*, 8(3):175–185, 1998. 8

[15] Simon A Forbes, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, David Beare, Mingming Jia, Rebecca Shepherd, Kenric Leung, Andrew Menzies, et al. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic acids research*, page gkq929, 2010. 29

[16] 1000 Genomes. VCF format. http://vcftools.sourceforge.net/specs.html, 2013. 9, 23

[17] P. J. Giles and D. Kipling. Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*, 19(17): 2254–2262, November 2003. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg311. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btg311. 49

[18] Florian Gnad, Albion Baucom, Kiran Mukhyala, Gerard Manning, and Zemin Zhang. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC genomics*, 14(Suppl 3): S7, 2013. 30

[19] Jeremy Goecks, Anton Nekrutenko, James Taylor, T Galaxy Team, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010. 22

[20] T. R. Golub. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, October 1999. ISSN 00368075. doi: 10.1126/science.286.5439.531. URL http://www.sciencemag.org/cgi/doi/10.1126/science.286.5439.531. 77, 80

[21] John Gómez, Leyla J García, Gustavo A Salazar, Jose Villaveces, Swanand Gore, Alexander García, Maria J Martín, Guillaume Launay, Rafael Alcántara, Noemi Del Toro Ayllón, et al. Biojs: an open source javascript framework for biological data visualization. *Bioinformatics*, page btt100, 2013. 32

[22] Abel Gonzalez-Perez, Jordi Deu-Pons, Nuria Lopez-Bigas, et al. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med*, 4(11):89–89, 2012. vii, 20

[23] Lei Guo, Edward K Lobenhofer, Charles Wang, Richard Shippy, Stephen C Harris, Lu Zhang, Nan Mei, Tao Chen, Damir Herman, Federico M Goodsaid, Patrick Hurban, Kenneth L Phillips, Jun Xu, Xutao Deng, Yongming Andrew Sun, Weida Tong, Yvonne P Dragan, and Leming Shi. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature Biotechnology*, 24(9):1162–1169, September 2006. ISSN 1087-0156. doi: 10.1038/nbt1238. URL http://www.nature.com/doifinder/10.1038/nbt1238. 56

[24] Sanger Institute. VCF format. http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/, 2013. 20

[25] Peter K. BWA PSSM. https://github.com/pkerpedjiev/bwa-pssm, 2013. 41

[26] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols*, 4(7):1073–1081, 2009. vii, 14, 20

[27] Heng Li. Fast and accurate short read alignment with burrowswheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. 41

[28] Heng Li. Wgsim. https://github.com/lh3/wgsim, 2009. 41, 42

[29] Ruixiao Lu, Geun-Cheol Lee, Michael Shultz, Chris Dardick, Kihong Jung, Jirapa Phetsom, Yi Jia, Robert H Rice, Zelanna Goldberg, Patrick S Schnable, et al. Assessing probe-specific dye and slide biases in two-color microarray data. *BMC bioinformatics*, 9(1):314, 2008. 48

[30] Marie-Laure Martin-Magniette, Julie Aubert, Eric Cabannes, and Jean-Jacques Daudin. Evaluation of the gene-specific dye bias in cdna microarray experiments. *Bioinformatics*, 21(9):1995–2000, 2005. 48

[31] Monnie McGee and Zhongxue Chen. Parameter estimation for the exponential-normal convolution model for background correction of affymetrix genechip data. *Statistical applications in genetics and molecular biology*, 5(1), 2006. 52

[32] Cliff Meldrum, Maria A Doyle, and Richard W Tothill. Next-generation sequencing for cancer diagnostics: a practical perspective. *The Clinical Biochemist Reviews*, 32(4):177, 2011. 37

[33] Michael L Metzker. Sequencing technologies-the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2009. 5

[34] NCBI. Ncbi ftp. ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/viruses.txt, 2013. 38

[35] NCBI. Ncbi blast. http://blast.ncbi.nlm.nih.gov/Blast.cgi, 2013. xi, 40

[36] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118–e118, 2011. vii, 15, 16, 20

[37] Matthew E Ritchie, Jeremy Silver, Alicia Oshlack, Melissa Holmes, Dileepa Diyagama, Andrew Holloway, and Gordon K Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23 (20):2700–2707, 2007. 52

[38] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977. 3

[39] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001. 18

[40] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008. 5

[41] Wei Shi, Alicia Oshlack, and Gordon K Smyth. Optimizing the noise versus bias trade-off for illumina whole genome expression beadchips. *Nucleic acids research*, 38(22):e204–e204, 2010. 50

[42] Jeremy D Silver, Matthew E Ritchie, and Gordon K Smyth. Microarray background correction: maximum likelihood estimation for the normal-exponential convolution. *Biostatistics (Oxford, England)*, 10(2): 352–63, April 2009. ISSN 1468-4357. doi: 10.1093/biostatistics/kxn042. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2648902&tool=pmcentrez&rendertype=abstract. 51

[43] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):Article3, 2004. ISSN 1544-6115. doi: 10.2202/1544-6115.1027. 58

[44] Jens Sobek, Kerstin Bartscherer, Anette Jacob, Jvrg D Hoheisel, and Philipp Angenendt. Microarray technology as a universal tool for high-throughput analysis of biological systems. *Combinatorial chemistry & high throughput screening*, 9(5):365–380, 2006. 46

[45] R Staden. A strategy of dna sequencing employing computer programs. *Nucleic acids research*, 6(7):2601–2610, 1979. 7

[46] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009. 12

[47] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005. 21

[48] Shamil R Sunyaev, Frank Eisenhaber, Igor V Rodchenkov, Birgit Eisenhaber, Vladimir G Tumanyan, and Eugene N Kuznetsov. Psic: profile extraction

from sequence alignments with position-specific counts of independent observations. *Protein engineering*, 12(5):387–394, 1999. 13

[49] Yuchun Tang, Yan-Qing Zhang, Zhen Huang, Xiaohua Hu, and Yichuan Zhao. Recursive fuzzy granulation for gene subsets extraction and cancer classification. *Information Technology in Biomedicine, IEEE Transactions on*, 12(6):723–730, 2008. 79

[50] Ying Tao, Yang Liu, Carol Friedman, and Yves A Lussier. Information visualization techniques in bioinformatics during the postgenomic era. *Drug Discovery Today: BIOSILICO*, 2(6):237–245, 2004. 83

[51] Ali Torkamani and Nicholas J Schork. Prediction of cancer driver mutations in protein kinases. *Cancer research*, 68(6):1675–1682, 2008. 12

[52] V G Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9): 5116–21, April 2001. ISSN 0027-8424. doi: 10.1073/pnas.091062498. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=33173&tool=pmcentrez&rendertype=abstract. 56

[53] Daniela Witten and Robert Tibshirani. A comparison of fold-change and the t-statistic for microarray data analysis. *Department of Statistics, Stanford University technical report*, 2007. 57

[54] Satya P Yadav. The wholeness in suffix-omics,-omes, and the word om. *Journal of biomolecular techniques: JBT*, 18(5):277, 2007. 44

[55] Tingxi Yu, Matthew J Ferber, Tak Hong Cheung, Tong Kwok Hung Chung, Yick Fu Wong, and David I Smith. The role of viral integration in the development of cervical cancer. *Cancer genetics and cytogenetics*, 158(1): 27–34, 2005. 37

[56] Jing Zhang, Jie Liu, Jianbo Sun, Chen Chen, Gregory Foltz, and Biaoyang Lin. Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing. *Briefings in bioinformatics*, page bbt042, 2013. 12, 22