Gene mutation analysis -Role of Open Source tools for personalized medication and designer drugs in a cost-effective and efficient manner

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering
in
Computer Science and Engineering

Submitted By
Tarun Preet Singh
(Roll No. 821132005)

Under the supervision of:
Dr. Maninder Singh
Associate Professor

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004
July 2014

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "Gene mutation analysis -Role of Open Source tools for personalized medication and designer drugs in a cost-effective and efficient manner.", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Computer Science and Engineering* submitted in Computer Science and Engineering Department of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Maninder Singh* and refers other researcher's work which are duly listed in the reference section.

The matter presented in the thesis has not been submitted for award of any other degree of this or any other University.

Tarun Preet Singh

(Tarun Preet Singh)

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

(Dr. Maninder Singh)
Associate Professor, CSED

Countersigned by

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

(Dr. S. K. Mohapatra)
Dean (Academic Affairs)
Thapar University
Patiala

ACKNOWLEDGEMENT

Words are scarce to express my gratitude to people who have stimulated and guided me while treading this journey. However I articulate my feelings from the bottom of my heart to…..

Dr. Maninder Singh, my mentor and guide, Associate Professor, CSED, Thapar University Patiala & Dr. Deepak Garg, Associate Professor and Head, CSED, Thapar University Patiala for not only providing the requisite research facilities but also for their support and valuable advice whenever I needed so.

Mr. Rajan Bir Singh, CIO, Max Healthcare for allowing me to carry on my research work at Max Healthcare Okhla.

My teachers, family and friends for giving me support and motivation.

Above all I pray and thank 'Almighty' with whose beatitude I could reach this far.

ABSTRACT

We have established Gene Mutation Analysis platform of software pipelines using Open Sources tools (BWA, GATK, Bioconductor, Samtools, VarScan2) on Linux based Operating System. We have implemented GATK best practices for mutation detection. Also, developed browser-based tools for visual analysis of the gene mutations using Open Source tools (JBrowse, HTML5, Javascript, Python). Browser-based visual tool enables clinicians to look atupstream/downstream of the mutation/SNP. The visual tool enables identification of mutation hot-spots, driver mutations and passenger mutation.

Quality Checks are made part of the software tools for minimizing possible false-positive detection of gene mutation/SNP. The quality checks based on minimum coverage, Phred scores, ratio of supporting variant and reference alleles. Mutations are further validated using COSMIC, HapMap, Geome Wide Association Studies and clinically flagged SNPs from dbSNP.

With help of the above platform and visual tools, the Oncologist is now able to do a triage – look at a patient, look at their lab results and look at their genetic profile visually; they are now able to do more personalized medicine.

Table of Contents

Certificate i

Acknowledgement ii

Abstract iii

List of Figures vi

List of Tables vii

Chapter I Introduction 1

1.1 What is Genomics? 1

1.2 Genetic Testing and Personalized Medicine 1

1.3 Need for developing genomics based system in India 2

1.4 Genome Wide Association Studies (GWAS) 2

1.4.1 Example of Her2/neu and Response to Breast Cancer Treatment 4

1.4.2 Example of BRAF and mutations 4

1.5 Genetic Information based Drug Target Identification 5

1.6 Role of Information Technology in gene mutation analysis 5

Chapter 2 State of the Art 7

2.1 Development of analysis pipelines using Bio-conductor and various standard next-generation tools 7

2.2 Theoretical Prediction of Effect of mutations 7

2.3 Integration of Publically available Gene Mutation Databases 8

2.3.1 COSMIC (Catalogue of Somatic Mutations in Cancer) Database 8

2.3.2 IARC TP53 Database 8

2.3.3 ARDB (androgen receptor gene mutations) 9

2.3.4 RCGDB 9

2.3.5 HapMap Database 9

Chapter 3 Problem Statement 10

3.1. Problem Statement 10

Chapter 4 Solution to the Problem 11
4.1 Strategy used for pre-processing of file before uploading to browser. Description of software strategy used and reason behind it 16
Chapter 5 Testing & Results (Experimental Results) 24
5.1 Drug Response 24
5.1.1 Search Resources 24
5.2. Quality Scores 26
5.3 Quality Checks
27
5.4. Drug Response in Non-public resources (PGMD from BioBase) 32
5.5. Experimental Results 35
Chapter 6 Conclusion and Future Scope 53
6.1. Web portal development for visualizing gene mutations and mapping to drug responses 53
6.2. Relational Database Structure 53
6.3. Python based Django based web-framework 53
6.4. Summary of Contribution 54
References 55

List of Figures

Figure 4.1	Parts of JBrowse	11
Figure 4.2	An example workflow showing Jbrowse as finalstep	13
Figure 4.3	Shows Jbrowse Architecture	14

List of Tables

Table 4.1	Features of Jbrowse compared to other web-based genome annotation browsers	15
Table 4.2	For pre-processing the files for the Jbrowse the tools used.	16
Table 5.1	Quality Check Of Samples(Germ line Cases)	28
Table 5.2	Germline Cases	29
Table 5.3	Somatic Cases	30
Table 5.4	Somatic cases	31
Table 5.5	Drug Response	32
Table 5.6	Drug Response	33
Table 5.7	Addl. Sample 69 info.	34
Table 5.8	Addl. SNPs found	50
Table 5.9	Addl. SNPs found	51
Table 5.10	Addl. SNPs found	52

Chapter 1
Introduction

The recent advances, in the human genome research, are opening the door to a new paradigm for medical sciences. Personalized medicine, the use of marker-assisted diagnosis and targeted therapies derived from an individual's genetic profile, will impact the way drugs are developed and medicine is practiced in coming years.

1.1 What is Genomics?

Genomics is the study of the complete genetic material (genome) of organisms. The field includes sequencing, mapping, and analyzing a wide range of RNA (ribonucleic acid) and DNA (deoxyribonucleic acid) codes of many species across the kingdoms of life. Post 2000, there are intensive efforts to determine the entire DNA sequence of many individual humans in order to map and analyze individual genes and alleles as well as their interactions. The primary goal that drives these efforts is to understand the genetic basis of heritable traits, and especially to understand how genes work in order to prevent or cure diseases. Recently, there are good quality resources publically available with reference genome sequences –like 1000 genomes [1], [2].

1.2 Genetic Testing and Personalized Medicine:

The fast paced progress in technology is revolutionizing the application of technology in the human life. There is an increasing demand on personalizing various aspects of human life with the help of technology. A similar trend is observed in medicine as well. The healthcare is increasingly being tailored according to one's strengths and weaknesses. In fact, the practice of personalized medicine approach dates back to ancient times, when Hippocrates (2500 years ago) prescribed sweet elixirs to some and astringent ones to others, depending on each patient's physical characteristics and seasons. The personalized medicine in the modern era, utilizes the person's genes (including his/her DNA sequence and its polymorphism), to make drugs better and safer. Moreover, the rapid technological advances are decreasing DNA sequencing costs and making it practical for large scale

applications. The sequencing price per genome is expected to reach $100 over the next few years. The advancement of sequencing technologies and decrease in cost has made it now possible for healthcare providers, to combining sequenced genomic data with other medical data to get a better picture of disease in an individual. The vision is that treatments will reflect an individual's illness, and not be a one treatment fits all, as is too often true today.

1.3 Need for developing genomics based system in India:

The year 2013 marked the 10th anniversary of the completion of Human Genome Project (HGP), the multibillion-dollar effort to read the human genetic code in its entirety. The HGP and related efforts have sparked a revolution that could alter the fields of health care. World over, there is a tremendous increase in the research and (business) investment in the field of genomics and genomics based therapy options. India is currently lagging behind the countries like - USA, China UK, Germany and Brazil – both in terms of investment and research. Such an approach (of increased research and investments) becomes more imperative for India for the fact that the number of new cases of breast cancer in India will increase from the current 115,000 to around 200,000 per year. Moreover, the age of onset of breast cancer is falling from 40s to 20s. Therefore, the burden of various cancer types would be, unfortunately, increasing in coming years in India. The genomics based screening for germline mutations provides a preventive option for the high risk individuals. Thus, genomics based approach promises to reduce the future burden of cancer in India.

1.4 Genome Wide Association Studies (GWAS)

The GWAS - in which multiple single-nucleotide polymorphisms (SNPs) are tested for association with one/more disease in large population - have revolutionized the search for genetic influences on complex traits [6]. A single nucleotide polymorphism is one in which there is a one nucleotide difference between two genes. For example, two sequenced DNA fragments, AAGCCTA versus AAGCTTA, have one differing nucleotide. In other words, these are different alleles of the same gene.

For example, SNP based personalization of dosage for managing simvastatin-induced myopathy risk in the context of *SLCO1B1* genotyping has recently been issued [7]. A non-synonymous coding single-nucleotide polymorphism (SNP), rs4149056, in *SLCO1B1* markedly increases systemic exposure to simvastatin and the risk of muscle toxicity. The above-mentioned guideline explores the relationship between rs4149056 (c.521T>C, p.V174A) and clinical outcome for all statins [7].

Since 2005, more than 1,600 publications have identified more than 2,000 replicated genetic associations with more than 300 common human diseases and traits [8]. There are three near future clinical applications for GWAS namely [8] :

- Predictive models to identify high risk patients, as in Type 1 Diabetes patients. For example in one of the cases of maturity-onset diabetes of the young or MODY, is due to mutations in the *HNF1A* gene. The *HNF1A* mutations could have important implications for patients and their relatives, as many patients with *HNF1A*-MODY are better managed with sulphonylureas than with metformin or insulin [8].
- Classifying disease subtypes of potential use for more precisely guided clinical trials, and targeted treatments (e.g. cancers). The cancer genomics based data generated by large consortia such as The Cancer Genome Atlas (TCGA) Research Network and the International Cancer Genome Consortium (ICGC) provides immense possibilities for researching for targeted therapy [9].
- Providing better information for screening drug candidates for toxicity and efficacy before clinical trials- GWAS is providing valuable information on gene-drug interactions with the potential to develop safer and more effective drugs as well as to reduce toxicities in the clinical use of existing medications. For example, treatment of hepatitis C, a common viral infection and a cause of liver cirrhosis. A GWAS of change in hemoglobin levels during ribavirin treatment identified inosine triphosphatase (ITPA) variants that can protect against ribavirin-induced anemia [8].

Although the potential for genomics to contribute to clinical care has long been anticipated, the pace of defining the risks and benefits of incorporating genomic findings into medical practice has been relatively slow [10].

1.4.1 Example of Her2/neu and Response to Breast Cancer Treatment:

One of success stories of genetic information based therapy, is the use of the *Her2/neu* gene as a predictor of breast cancer patients' responses to a drug called Herceptin. *Her2/neu* codes for a protein receptor which plays an important role in the signal transduction pathways involved with cell growth and differentiation. All breast tissue cells are coated with Her2 receptors, however overexpression (in about 15 to 20% of women with invasive breast cancer) causes abnormal cell growth. The tumors with overexpressed Her2/neu genes are said to be Her2-positive. Moreover, Herceptin is a much more effective treatment in women with Her2-positive tumors than in women with Her2-negative tumors. However, irrespective of Her2/neu status, Herceptin increases the risk of heart dysfunction. Considering the side effect of the Herceptin, U.S. Food and Drug Administration (FDA), has stated that physicians should prescribe Herceptin only to those women who are tested positive for Her2/neu. Otherwise, the risk of heart failure is just too great to justify use of this drug. Thus, women with advanced breast cancer are routinely tested for Her2/neu overexpression before any decisions are made about whether to prescribe Herceptin versus some other drug, because as good as Herceptin might be, to use Hippocrates' words, "the sweet ones do not benefit everyone."

1.4.2 Example of BRAF and mutations:

BRAF is the human gene responsible for the production of a protein called B-Raf, which is involved in sending signals inside cells to direct cell growth, and shown to be mutated in cancers. In 2011, a drug called vemurafenib, a B-Raf protein inhibitor, and the companion BRAF V600E Mutation Test were approved for the treatment of late stage melanoma. Vemurafenib only works in the treatment of patients whose cancer tests positive for the V600E BRAF mutation [11]. Around 60% of patients with melanoma have a BRAF mutation, and approximately 90% of those are the BRAF V600E mutation.

1.5 Genetic Information based Drug Target Identification:

The gene encoding leptin, a regulator of body fat discovered using genomic technologies [12], is not only proven to be a valuable drug target but blood leptin levels might be of use as a monitoring marker of drug-associated weight gain [13] or as a response to growth-hormone treatment in children [14].

1.6 Role of Information Technology in gene mutation analysis:

The amount of data being produced by sequencing, mapping, and analyzing genomes propels genomics into the realm of Big Data. Genomics produces huge volumes of data; each human genome has 20,000-25,000 genes comprised of 3 million base pairs. This amounts to 100 gigabytes of data. Sequencing multiple human genomes would quickly add up to hundreds of petabytes of data, and the data created by analysis of gene interactions multiplies those further. The medical discoveries of the future will largely depend on ability to process and analyse large genomic data sets, which continue to expand as the cost of sequencing decreases [15]. Moreover, the analysis of large sets of data, such as medication usage or hospital readmissions, can enable health care providers and policymakers to make smarter decisions and predict future trends. Among the various innovative applications of healthcare data, one interesting application could be to extract which prescriptions are cheapest or most cost-effective.

IT can be used for creating electronic tools and resources to collect and store patient health information, making it available to inform clinical decisions and improve safety while protecting patient privacy. There would be tremendous demand on integrating the electronic health record (EHR) systems and genomics data in the coming years [16]–[18]. There are number of quality IT system development currently underway across the globe for responding to the demands of the healthcare informatics. There are number of open source tools reviewed in the reference [19].

For example, medical histories of the individuals associated with their genomes, their drug treatments, their environments and microbiomes (the bacteria that live in your gut and your skin) can be used for prevention of disease and minimizing the re-admission.

Mount Sinai, in New York, USA, has used the data to help reduce the 30-day readmission rate for patients by 56% by determining which patients were most at risk, then taking precautions to prevent them returning to the hospital.

Chapter 2

State of the Art

High-throughput sequencing provides tools to conduct comprehensive analyses of all somatic alterations in the cancer genomes. The immense growth of information from the genome-scale investigations has promoted the development of new analytical frameworks and tools. An overview of the current state of cancer genomics and tools for accessing and analyzing cancer genomic data has been discussed in the reference [20].

2.1 Development of analysis pipelines using Bioconductor and various standard next-generation tools

Exome sequencing can be utilized for discovering cancer causing variants and mutations across hundreds of tumors. There are number of variation detection tools for the detection of somatic mutations in exome data from tumor–normal pairs. Among them, Bioconductor [21] (package: CancerMutationAnalysis), VarScan 2 [22], are most widely used tools. The bioconductor package, CancerMutationAnalysis, provides features for gene and gene-set level analysis methods for somatic mutation studies of cancer. The gene-level methods distinguish between **driver** genes (which play an active role in tumorigenesis) and **passenger** genes (which are mutated in tumor samples, but have no role in tumorigenesis).

Moreover, the genomics data analysis presents significant challenges. For example, sequencing coverage is non-uniform across targeted regions and varies from sample to sample [23], [24]. Repetitive and paralogous sequences can give rise to numerous false positives. The detection of somatic mutations in tumor genomes is even more challenging. The genomes of primary tumors are genetically heterogeneous [25], with frequent rearrangements [26] and copy number alterations (CNAs) [27]. Further, somatic mutations are relatively rare compared with germline variation, often representing <0.1% of variants in a tumor genome [28], [29].

2.2 Theoretical Prediction of Effect of mutations:

The known SNPs located in the dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) PolyPhen [30] (polymorphism phenotyping) and SIFT [31] (sorting intolerant from tolerant) database would be used for predicting the effect of mutation found. Moreover, the SIFT and PolyPhen tools can be used for loss-of-function or gain-of-function mutations [32]. The PROVEAN [33] database can also be used to predict whether a protein sequence variation affects protein function.

2.3 Integration of Publically available Gene Mutation Databases:

There are number of public resources which provide gene mutation information and drug related information. The exhaustive listing of such resources would be beyond the scope of this project. Below is brief description of few important databases:

2.3.1 COSMIC (Catalogue of Somatic Mutations in Cancer) Database [34]: All cancers arise as a result of DNA sequence abnormalities or mutations. Many of mutations may confer a growth advantage upon the cells in which they have occurred. These mutations accumulate in cells of the body over a person's lifespan. There are two different types of mutations – somatic and germline mutations. The **somatic** mutations happen on the genes that are usually located on autosomes (non-sex chromosomes). The **germline** mutations occur in germ cells (better known as the ovum or sperm). The germline mutations may occur *de novo* or be inherited from parents' germ cells. An example of germline mutations linked to cancer is the ones that occur in cancer susceptibility genes, increasing a person's risk for the disease. COSMIC is designed to store and display somatic mutation information and related details and contains information relating to human cancers.

2.3.2 IARC TP53 Database [35]: The tumor suppressor gene TP53 is frequently mutated in human cancers. More than 75% of all mutations are missense substitutions. The International Agency for Research on Cancer (IARC) TP53 database (www-p53.iarc.fr) compiles all genetic variations that have been reported in TP53. This database can be used for systematic analysis

to determine the functional properties that contribute to the occurrence of mutational "hotspots" in different cancer types and to the phenotype of tumors.

2.3.3 ARDB (androgen receptor gene mutations) [36]**:** The database now contains a number of mutations that are associated with prostate cancer treatment regimens (http://androgendb.mcgill.ca/).

2.3.4 RCGDB (The Roche Cancer Genome Database): This database provides single resources for genetic mutations from various sources updated till 2009. The somatic and germline mutations are numbered as per as the nomenclature of Human Genome Variation Society (http://www.hgvs.org/). The database can be accessed from the url http://rcgdb.bioinf.uni-sb.de/MutomeWeb/

2.3.5 HapMap Database [37]**:** The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to drugs.

Chapter 3

Problem Statement

The Genomic Testing was carried out by Max Healthcare's service provider Strand Life Sciences (SLS) - The next generation sequencing (NGS) detection was conducted by SLS for the purpose of mutation and analysis reports submitted in the form of hard copies. Genetic testing reports prepared, after testing saliva/tumor samples, were sent by SLS to Max Healthcare (MHC) clinicians. MHC has no platform (software pipelines) to validate the reports. No browser-based tools available for clinicians for visual analysis of results. For example, neighboring mutations/SNP and hot spots/regions. Robust Quality checks for the results were not part of few initial reports from SLS. The course of medication was decided based on symptoms, patient history, test reports (including genetic testing reports from SLS), morphology of disease.

3.1. Problem Statement- Gene mutation analysis-Role of Open Source tools for personalized medication and designer drugs in a cost-effective and efficient manner.

Although the year 2013 marked the 10^{th} anniversary of the completion of Human Genome Project, with USA having 841 High-throughput Gene Sequencing Machines, UK having 137, Germany having 142, China having 200 and Brazil having 83; India is found nowhere on the global gene sequencing map justifying the dire need to venture into this novel area.

Chapter 4

Solution to the Problem

Functional relationships in a genome are indicated by spatial relationships. Spatial relationships between different pieces of genomic data are conveyed by a genome browser and help users to form a hypothesis about their functional relationships. In the current scenario certain web-based genome browsers help users understand genomic data belonging to a specific region, but do not allow further understanding as navigation is required to other regions page-by-page. These page transitions do not show which genomic locus they are viewing and how the data points are related to one another. Other functions of a genome browser are helping solve biological cases, linking genomic databases and preparation of publication figures.

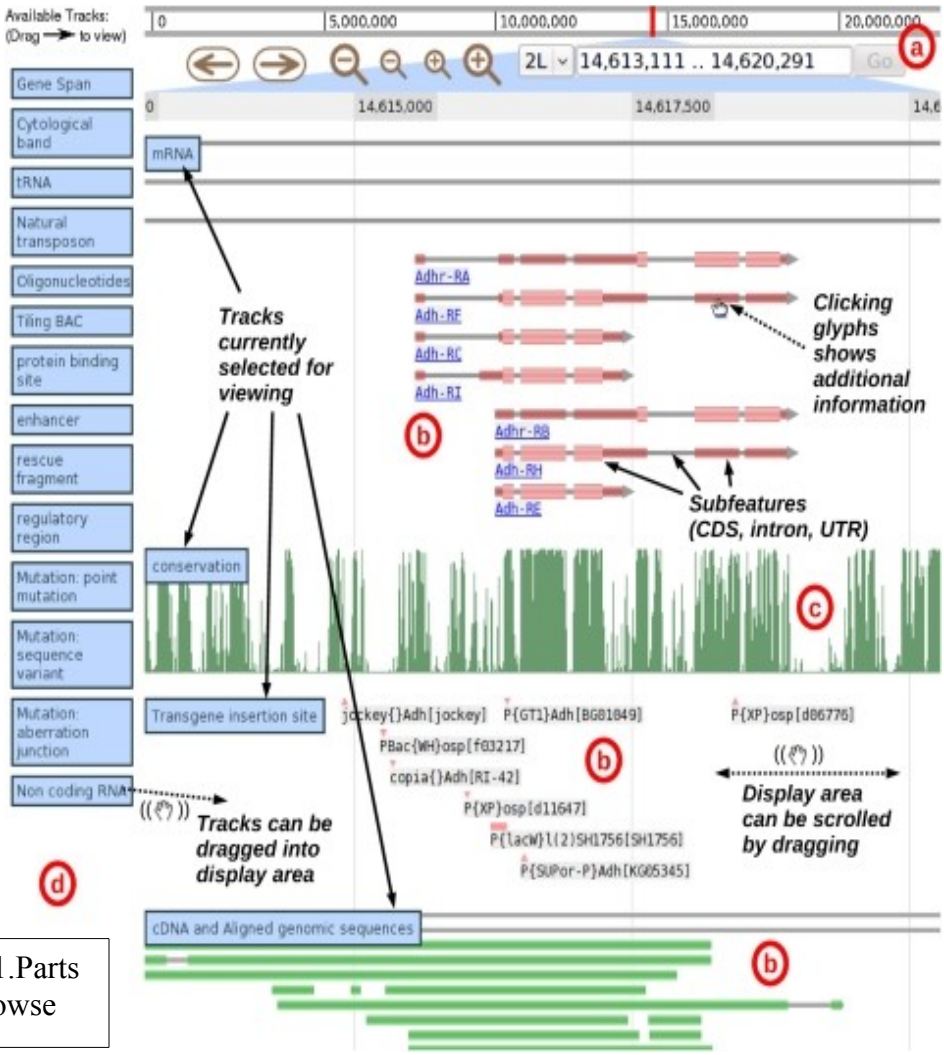


Fig 4.1.Parts of Jbrowse

Jbrowse is sufficiently compact and modular and can be used as a drop in component to web content management system. Browser functionality:-

1. Basic genome browser functionality-A user can view genome annotations against a reference ruler with an overhead bar giving a visual indication of chromosome position. Navigation by dragging the display left or right is done and using navigation buttons is also possible. Zooming in or out is done using analogous buttons. Direct navigation to a region of interest using region coordinates in the search-box is also possible. To add annotation tracks, drag them from a reservoir bar on the left and for removal drag them back off the main display. Tracks can be reordered by dragging. Live track manipulation requires no page reloads.

2. Multiple types of track-Basic features like a variety of simple glyphs and compound features like UTR/ exon / intron can be displayed. The quantitative tracks have a value

for every base. Feature tracks are displayed as histograms showing feature density at lower zoom levels.

3. Flexible data sources-

Data can be taken from GFF files for simple feature tracks or GFF3 files for compund feature tracks or from WIG files for quantitative tracks. Data can also be imported from a Bio::DB database using open-source applications that work on BioPerl libraries.

4. Makefile-driven workflow-

A workflow can be made using assembly, alignment , annotation and Jbrowse view of genome sequences. A workflow is a set of logical rules for transforming data types automatically like a Makefile. A Makefile prepares sequence and annotation files on a server for viewing by a Jbrowse client. Makefile also defines simple file format conversions like conversion of annotation files from BED format to GFF format. Downloading relevant FASTA, BED, GFF or WIG files into the top-level directory and executing “make jbrowse” option helps to apply Makefile rules.

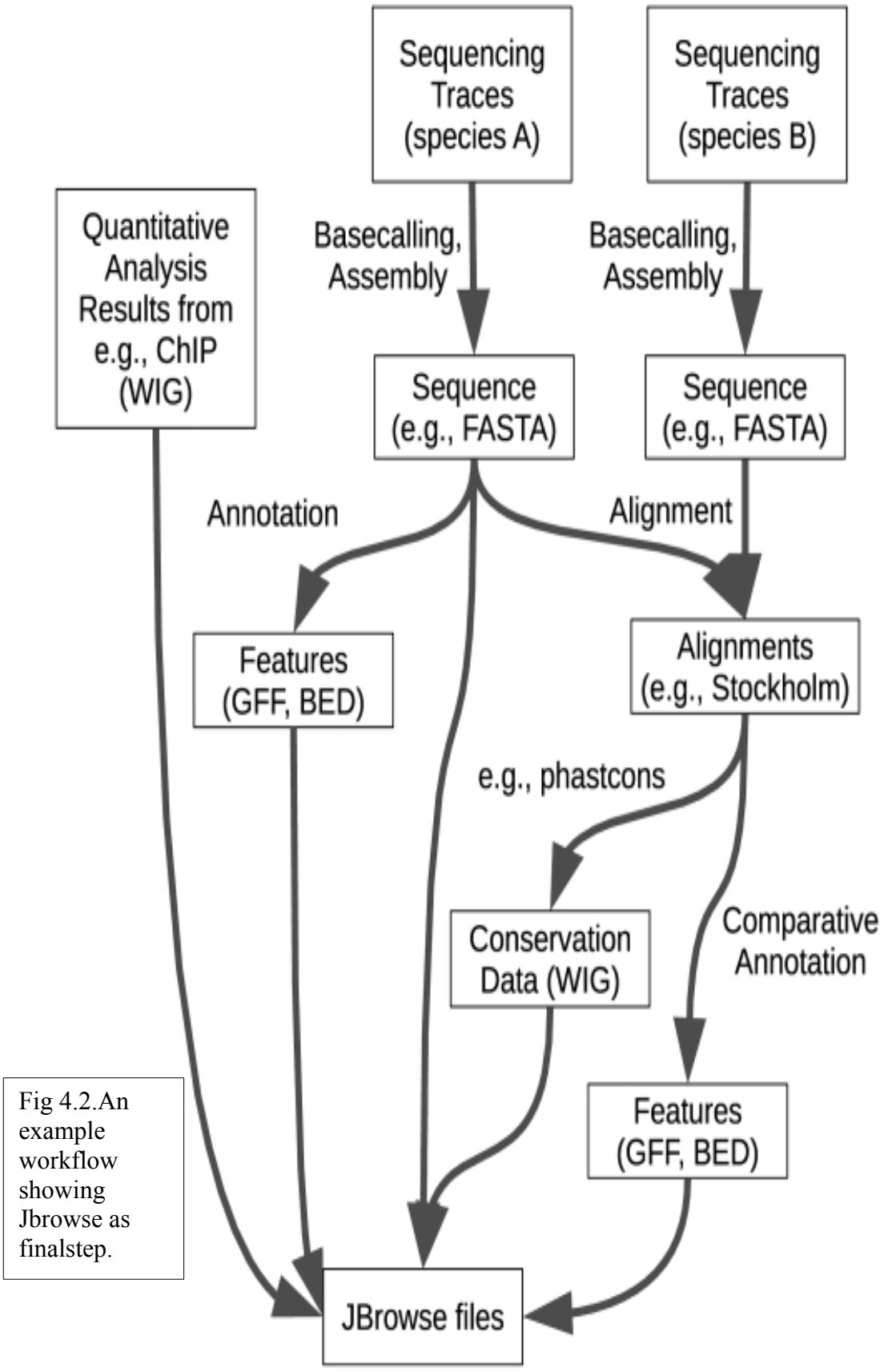


Fig 4.2.An example workflow showing Jbrowse as finalstep.

Server area shows data served by web server (rectangles), programs that generate data (arrow labels) and the data sources used by those programs (cylinders). The "Client" area shows the main pieces of code that run in the web browser, how they fit together, and what data they consume.

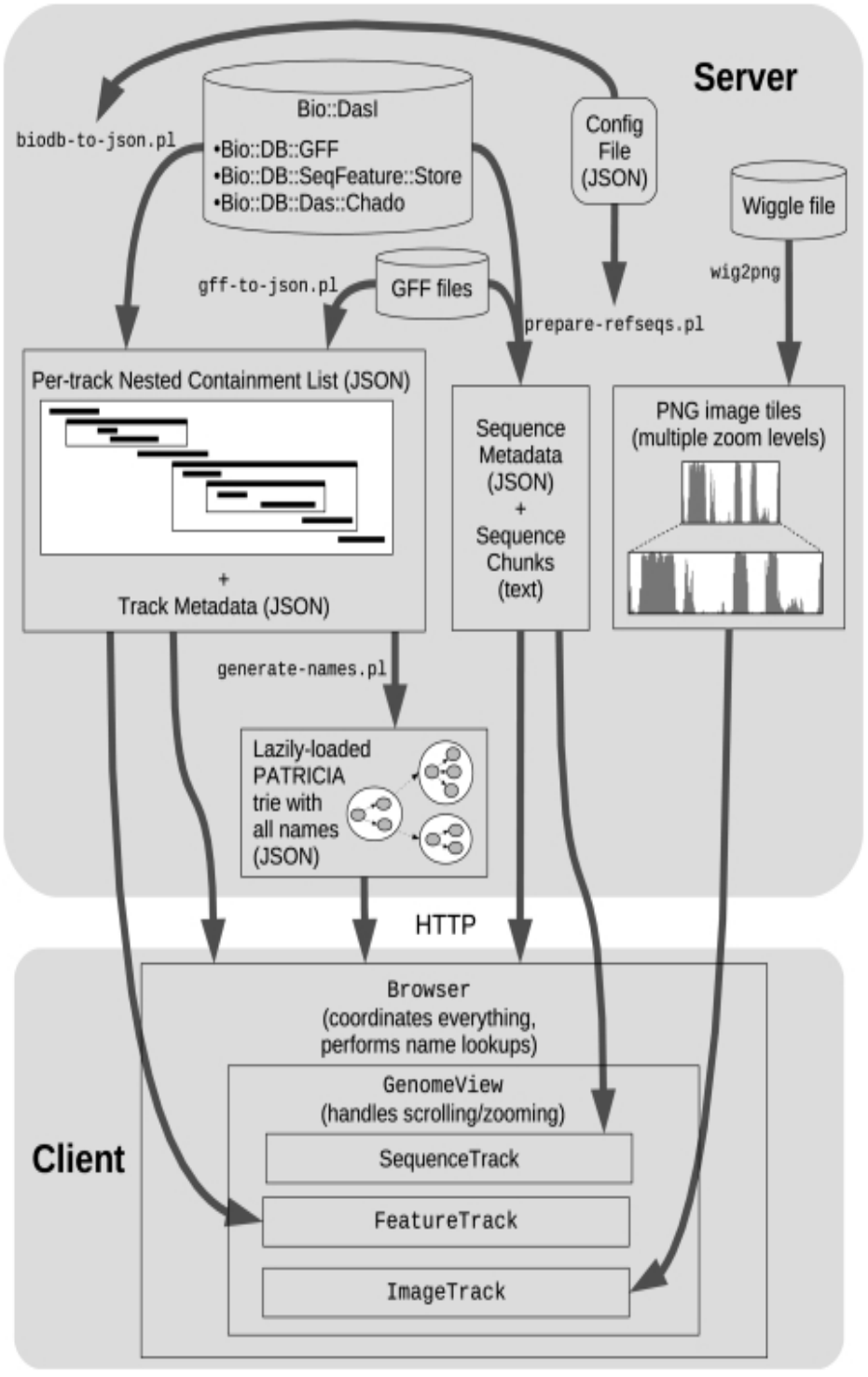


Fig4.3.Shows Jbrowse architecture,

5. Preservation of state-

To preserve the navigation state and track selection/ordering preferences of individual users; HTTP cookies are used. A user can close a browser window and open a new one, the browser will show the same genome annotations.

6. Configuration files-

Flexible configuration files allow the database administrator to customize the tracks and their behavior including glyphs and feature-click actions.

7. A "genome wiki"-

Jbrowse is compact, open source and has a simple Javascript API so can be embedded within other web applications. The core elements are Makefiles making Jbrowse a part of an extensible workflow system whereby users of a wiki can start analysis by uploading sequence, annotation or configuration files thus creating an open-source, portable, extensible wiki for sharing annotations.

Table 4.1. Features of Jbrowse compared to other web-based genome annotation

Table 1. Features of JBrowse compared to several other web-based genome annotation browsers

	Page reloads to pan	Page reloads to zoom	Page reloads to change track order	Page reloads to add/remove tracks	Feature tracks	Quantitative tracks	Click/mouse over shows feature-specific info?	Portable?	Open-source license for all users?	Firefox compatible?	Safari compatible?	Internet Explorer compatible?	Wiki integration
UCSC Genome Browser	1	1	2–3	1	√	√	√	√	x	√	√	√	x_P
UCSC Cancer Genomics Browser	1	0_D	n/a	0	x	√	√	√	x	√	√	√	x
Ensembl Genome Browser	1	1	n/a	2	√	√	√	√	√	√	√	√	x
NCBI Sequence Viewer 2.1	0	1	n/a	n/a	√	x	√	x	n/a	√	√	√	x
JGI Genome Browser	1	1	1	n/a	√	x	√	x	n/a	√	√	√	x
GBrowse 1.0	1	1	0_C	1	√	√	√	√	√	√	√	√	x
GBrowse 2.0	0_D	0_D	0_C	0	√	√	√	√	√	√	√	√	x
Anno-J	0_C	0_D	0	0	√	√		√	√	√	√	x	x
JBrowse	0_C	0_C	0_C	0	√	√	√	√	√	√	√	√	√

Code evolves; these comparisons are valid at the time of writing. √: Feature available; x: feature unavailable; n/a: not applicable; C: no page reloads are required; in addition, a smooth continuous transition is provided; D: no page reloads are required, but the transition is discontinuous; and P: planned for future release (Kuhn et al. 2009). References: UCSC Genome Browser (Kuhn et al. 2009); UCSC Cancer Genomics Browser (Zhu et al. 2009); Ensembl Genome Browser (Stalker et al. 2004); NCBI Sequence Viewer 2.1 (http://www.ncbi.nlm.nih.gov/projects/sviewer/); JGI Genome Browser (http://genome.jgi-psf.org/); GBrowse (Stein et al. 2002); Anno-J (Lister et al. 2008).

4.1 Strategy used for pre-processing of file before uploading to browser. Description of software strategy used and reason behind it:-

For a model organism genome database, each reference sequence would typically represent a chromosome. The large file containing all the chromosome can be divided into smaller parts as per as chromosome distribution of the organism.

Format	Tool Name	Link	Purpose
BED	BEDTools	https://code.google.com/p/bedtools/	A flexible suite of utilities for comparing genomic features.
BigWig	UCSC Tools	http://genome.ucsc.edu/goldenPath/help/bigWig.html	The bigWig format is for display of dense, continuous data that will be displayed in the Genome Browser as a graph
VCF	vcf-sort	http://vcftools.sourceforge.net/	VCF is conventions and extensions adopted by the 1000 Genomes Project for encoding structural variations. VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome.
VCF (Tabix)	tabix	http://vcftools.sourceforge.net/perl_module.html http://sourceforge.net/projects/samtools/files/tabix/	Perl scripts require that the VCF files are compressed by bgzip and indexed by tabix (both tools are part of the tabix package, available for download here). The VCF files can be compressed and indexed using the following commands
GFF3	Generic Feature Format Version 3	http://www.sequenceontology.org/gff3.shtml	GFF3 files are nine-column, tab-delimited, plain text files.
BAM	samtools	http://samtools.sourceforge.net	A BAM file (.bam) is the binary version of a SAM file. A SAM file (.sam) is a tab-delimited text file that contains sequence alignment data.
Genome Coverage	GenomeCoverage	http://bedtools.readthedocs.org/en/latest/content/tools/genomecov.html	To get coverage of genes.

Table 4.2. For pre-processing the files for the Jbrowse, above tools are used.

Formatting Reference Data:

The prepare-refseqs.pl script, available in bin directory, is used to format reference sequence data. In addition to formatting the sequence data, this script creates a track called "DNA" that displays the reference sequence. The multiple chromosomes are available from the "Data set by chromosome" link on the url hgdownload.cse.ucsc.edu for the particular genome.

In order to get gene names displayed in the reference sequences, we have used GFF3 file format. Using reference genome in GFF3 format, with gene co-ordinates marked. The single file of complete genome has been separated into smaller files in order to load into browser. Following unix commands are used for processing the reference file:

```
bin/prepare-refseqs.pl --fasta ../ref-chromosome/chr1.fa.gz --fasta ../ref-chromosome/chr2.fa.gz --fasta ../ref-chromosome/chr3.fa.gz --fasta ../ref-chromosome/chr4.fa.gz --fasta ../ref-chromosome/chr5.fa.gz --fasta ../ref-chromosome/chr6.fa.gz --fasta ../ref-chromosome/chr7.fa.gz --fasta ../ref-chromosome/chr8.fa.gz --fasta ../ref-chromosome/chr9.fa.gz --fasta ../ref-chromosome/chr10.fa.gz --fasta ../ref-chromosome/chr11.fa.gz --fasta ../ref-chromosome/chr12.fa.gz --fasta ../ref-chromosome/chr13.fa.gz --fasta ../ref-chromosome/chr14.fa.gz --fasta ../ref-chromosome/chr15.fa.gz --fasta ../ref-chromosome/chr16.fa.gz --fasta ../ref-chromosome/chr17.fa.gz --fasta ../ref-chromosome/chr18.fa.gz --fasta ../ref-chromosome/chr19.fa.gz --fasta ../ref-chromosome/chr20.fa.gz --fasta ../ref-chromosome/chr21.fa.gz --fasta ../ref-chromosome/chr22.fa.gz --fasta ../ref-chromosome/chrX.fa.gz --fasta ../ref-chromosome/chrY.fa.gz --out sample_data/json/hg19
```

For getting gene annotations (namely gene name and its start and end positions), we need to get the gene annotated file in gff3 format.

```
# Seperate chromosomes, repeat this command for each chromosomes
awk -F"\t" '{if ($1 == 1){print $0}}' Homo_sapiens.GRCh37.74.gff3 > chr1.gff
# Rename files into gff3
```

```
ls *.gff | cut -d'.' -f1 | xargs -I name mv name.gff name.gff3
# Add lines at top of the each file
ls chr*.gff3 | xargs -I name sed -i '1s/^/##gff-version 3\n/' name
```

The Homo_spaiens.GRCh37.74.gff3 can be downloaded from http://asia.ensembl.org/Homo_sapiens/Info/Index

Add following lines into tracks.conf for the gff3 file for each chromosomes. Change the path of chr1.gff3 as per as the your directory structure.

```
[ tracks . chr1_gff3 ]
storeClass = JBrowse/Store/SeqFeature/GFF3
urlTemplate = ../../raw/hg19/reference/chr1.gff3
type = CanvasFeatures
metadata.description = All the features in the Chr1.gff3
category = Miscellaneous
key = Genes on chr1
```

Formatting Feature Data:

The ucsc-to-json.pl script, available in bin directory, is used for converting range-based annotation data (genes, transcripts, etc) to range-indexed sets of static JSON files that are very fast for JBrowse to access. It adds a track configuration stanza to the trackList.json configuration file in its output directory.

- ucsc-to-json.pl - import UCSC database dumps (.sql and .txt.gz)

Above script uses data from a local dump of the UCSC genome annotation MySQL database. The MySQL dump can be downloaded from the url hgdownload.cse.ucsc.edu. In this database dump, a *.sql and *.txt.gz pair of files constitute a database table. Ucsc-

to-json.pl uses the *.sql file to get the column labels, and it uses the *.txt.gz file to get the data for each row of the table.

The reference-genome folder should contain all dump *.sql and *.txt.gz downloaded from the hgdownload.cse.ucsc.edu.

```
# To create a track for known genes.
bin/ucsc-to-json.pl --in hg19v2/ --track 'knownGene' --cssclass transcript --subfeatureClasses '{"CDS":"transcript-CDS", "UTR": "transcript-UTR"}' --arrowheadClass transcript-arrowhead --out sample_data/json/hg19

# To create a track for GWAS, HAPMAP, dbSNP, COSMIC database

bin/ucsc-to-json.pl --in ../ref-geneome --track 'gwasCatalog' --out sample_data/json/hg19
bin/ucsc-to-json.pl --in ../ref-geneome --track 'hapmapSnpsASW' --out sample_data/json/hg19
bin/ucsc-to-json.pl --in ../ref-geneome --track 'snp138Flagged' --out sample_data/json/hg19
bin/ucsc-to-json.pl --in ../ref-geneome --track 'cytoBandIdeo' --out sample_data/json/hg19
#bin/ucsc-to-json.pl --in ../ref-geneome --track 'cytoBand' --out sample_data/json/hg19
bin/ucsc-to-json.pl --in ../ref-geneome --track 'cosmic' --out sample_data/json/hg19
```

BAM file based features:

The BAM files are processed to add various tracks. The basic steps include (for each BAM files) are sorting and indexing of BAM files for faster loading into browser.

1. Make a separate jbrowseBAM folder in sample_data/raw/hg19 folder to store processed BAM files

```
mkdir –p sample_data/raw/hg19/jbrowseBAM
# copy all the BAM files into jbrowseBAM directory
cp <filename>.bam jbrowseBAM
```

2. Run following set of commands for each BAM files to make them Jbrowse compatible. Replace <filename> as per your case

```
samtools sort <filename>.bam <filename>.sorted.bam
samtools index <filename>.sorted.bam
```

3. Run following set of commands for each BAM files to make BED file and Genome Coverage files

```
bamToBed –i  <filename>.bam <filename>.bed

# Sort the BED file
sort –k 1,1 <filename>.bed <filename>.sorted.bed

# Get the gene boundary form UCSC using mysql command
mysql –user=genome –host=genome-mysql.cse.ucsc.edu –A –e \"select chrom,size from hg19.chromInfo\" > hg19.genome
# Use genomeCoverageBed to get gene coverage
genomeCoverage –bg –i <filename>.sorted.bed –g hg19.genome > <filename>.cov

# Get the chromosome sizes
wget "http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/fetchChromSizes.sh" .
chmod a+x fetchChromSizes.sh
./fetchChromSizes.sh hg19 > hg19.chrom.sizes
# Get the bigwig file
bedGraphToBigWig <filename>.cov hg19.chrom.sizes <filename>.bw
```

4. Add following lines for each BAM files to add BAM specific tracks information to the tracks.conf. For example, for P000227-S1314000129-Onc-GB-N_S5.bam file.

[tracks . P000227-S1314000129-Onc-GB-N_S5]
style.height = 7
histograms.storeClass = JBrowse/Store/SeqFeature/BigWig
histograms.urlTemplate = ../../raw/hg19/jbrowseBAM/P000227-S1314000129-Onc-GB-N_S5.bw
key = BAM - P000227-S1314000129-Onc-GB-N_S5
storeClass = JBrowse/Store/SeqFeature/BAM
urlTemplate = ../../raw/hg19/jbrowseBAM/P000227-S1314000129-Onc-GB-N_S5.sorted.bam
maxFeatureScreenDensity = 4
metadata.category = BAM
metadata.Description = BAM-format alignments of P000227-S1314000129-Onc-GB-N_S5
type = JBrowse/View/Track/Alignments2

5. Add following lines for each BAM files to add BAM specific coverage information to the tracks.conf file

[tracks . P000227-S1314000129-Onc-GB-N_S5-coverage]
storeClass = JBrowse/Store/SeqFeature/BAM
urlTemplate = ../../raw/hg19/jbrowseBAM/P000227-S1314000129-Onc-GB-N_S5.sorted.bam
metadata.category = BAM
metadata.Description = SNP/Coverage view of P000227-S1314000129-Onc-GB-N_S5
type = JBrowse/View/Track/SNPCoverage

key = sorted Coverage - P000227-S1314000129-Onc-GB-N_S5

6. Add following lines for each BAM files to add BAM specific XYPlot and Density plot

[tracks.bigwig-track-xyplot]
storeClass = JBrowse/Store/SeqFeature/BigWig
urlTemplate = ../../raw/hg19/jbrowseBAM/P000227-S1314000129-Onc-GB-N_S5.bw
category = Quantitative
type   = JBrowse/View/Track/Wiggle/XYPlot
key   = XYZ Coverage plot for P000227-S1314000129-Onc-GB-N_S5

[tracks.bigwig-track-density]
storeClass = JBrowse/Store/SeqFeature/BigWig
urlTemplate = ../../raw/hg19/jbrowseBAM/P000227-S1314000129-Onc-GB-N_S5.bw
category = Quantitative
type   = JBrowse/View/Track/Wiggle/Density
key   = Density plot for P000227-S1314000129-Onc-GB-N_S5

cat sample_data/raw/hg19/conf/bam.conf >> sample_data/json/hg19/tracks.conf

Adding mutation information based on VCF files

The mutations in the VCF files can be shown in the JBrowse after few processing like sorting and indexing of the VCF file.

1. Sort and index the VCF files

vcf-sort <filename>.vcf > <filename>.sorted.vcf
bgzip <filename>.sorted.vcf

tabix -p vcf <filename>.sorted.vcf.gz

2. Add following lines in the tracks.conf for each VCF files. Change the filenames as per as your case

[tracks . P000177-S131400087-GL-N]
storeClass = JBrowse/Store/SeqFeature/VCFTabix
urlTemplate = ../../raw/hg19/vcf/dbsnp-P000177-S131400087-GL-N_S1.sorted.vcf.gz
category = VCF
type = JBrowse/View/Track/CanvasVariants
key = VCF - P000177-S131400087-GL-N

Adding generated name and function information

Add following to tracks.conf

bin/add-json.pl '{ "dataset_id": "hg19", "include": ["../../raw/hg19/conf/functions.conf"] }' sample_data/json/hg19/trackList.json
bin/generate-names.pl --safeMode -v --out sample_data/json/hg19/

Chapter 5

Testing & Results (Experimental Results)

5.1 Drug Response:-

5.1.1Search Resources:

1.Public Resources:

- Clinical Trial: http://www.clinicaltrials.gov/
- PubMed: http://www.ncbi.nlm.nih.gov/pubmed
- PharmGKB: https://www.pharmgkb.org/
- DrugBank: http://www.drugbank.ca/
- Google Scholar: http://scholar.google.co.in/
- FDA Pharmacogenomic Biomarkers: http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogentics/ucm083378.htm

2.Non-public Resources:

- BioBase (PGMD) Trial Version: http://www. biobase-international.com
- ScienceDirect: http://www.sciencedirect.com/
- Nature, ACS, PNAS

Search Criterion:

1.Mutation/SNPs based

- Based on rsIDs
- Based on mutation co-ordinates (chr3:7577538)

2. Disease Condition and Drug Based:

- In Clinical Trial advance search based on conditions and interventions or full text of drug name, then looked for presence of mutation studies of gene of interest. Eg of FDA Approved drugs: Sunitinib, Cisplatin, Fluorouracil, Oxaliplatin:-

Example- rs1870377 (KDR, chr4: 55972974) drug response studies:-
1.PubMed [http://www.ncbi.nlm.nih.gov/pubmed/?term=rs1870377+cancer]

[1] The rs1870377 of KDR were associated with poor time to treatment failure (TTF) (p=0.029), while rs1870377 of KDR were associated with poor overall survival (OS) (p=0.001). In conclusion, rs1870377 may affect treatment outcome and toxicity in patients treated with sunitinib. [http://www.ncbi.nlm.nih.gov/pubmed/24123039]

[2] The VEGFR2 rs1870377 polymorphism might affect survival in patients with diffuse large B cell lymphoma (DLBCL), suggesting that angiogenesis might be related to poor survival in these patients. [http://www.ncbi.nlm.nih.gov/pubmed/22129133]

[3] The VEGFR2 rs1870377 is not found to be associated with sunitinib response or toxicity. The study performed on 101 patients with renal-cell carcinoma with decreased sunitinib response and tolerability. [http://www.ncbi.nlm.nih.gov/pubmed/22015057]

2.PharmGKB [https://www.pharmgkb.org/rsid/rs1870377]

[1] Genotypes AA + AT are not associated with response to cisplatin, fluorouracil and oxaliplatin in people with Stomach Neoplasms as compared to genotype TT. (P value=0.05 and n=94) [http://www.ncbi.nlm.nih.gov/pubmed/24090479]

Example using clinical trials-rs1870377 (KDR, chr4: 55972974) drug response studies:-

Clinical Trials [http://www.clinicaltrials.gov/ with text: KDR and conditions: Gliosarcoma

[1] Phase (Ph) II Bevacizumab + Erlotinib for Patients (Pts) With Recurrent Malignant Glioma (MG) [NCT00671970]

Bevacizumab + erlotinib was adequately tolerated in recurrent MG patients. However, this regimen was associated with similar PFS benefit and radiographic response when compared with other historical bevacizumab-containing regimens. [http://www.ncbi.nlm.nih.gov/pubmed/20716591]

[2] Phase 1 ongoing studies [Safety Study of XL184 (Cabozantinib) in Combination With Temozolomide and Radiation Therapy in the Initial Treatment of Adults With Glioblastoma (NCT00960492)]

[3] There are 441 clinical trials going on for Gliosarcoma and 4 having mention of KDR/VEGFR-2 genes [NCT00671970: Bevacizumab+Erlotinib; NCT01122888: cilengitide; NCT00960492: XL184+temozolomide] Drugs studied- FDA Approved drugs: Bevacizumab , Erlotinib , Temozolomide.Under FDA Approval (NDA): Cabozantinib (2012).Experimental Drugs: cilengitide.

5.2.Quality Scores:-

1. Phred [1] quality scores are quality scores assigned to each nucleotide base in automated sequencers. The most important use of Phred quality scores is the automatic determination of accurate, quality-based consensus sequences. Phred quality scores Q are defined as a property which is logarithmically related to the base-calling error probabilities P.

$Q = -10\log_{10}P$

Phred quality scores of 20 means: Probability of incorrect base call 1 in 100 and Base cal accuracy of 99%.

2. Base alignment quality [2] (BAQ) is a phred-like score representing the probability that read base is mis-aligned; it lowers the base quality score of mismatches that are near

indels (insertions or deletions). BAQ scores helps to rule out false positive SNP calls due to alignment artifacts near small indels.

3. Minimum Coverage [3]: This indicates number of nucleotides contributing to a portion of an assembly. In other words, on average each base has been sequenced a certain number of times.

This has been set to minimum of 500x, 400x or 100x atleast.

4. Strand-filter [3]: The strand-filter used to remove variant calls when >90% of the supporting reads come from one strand.

References for above-[1] Ewing B et.al. (1998), Genome Res. 8(3):186-194, [2] Li. H et.al. (2011) Bioinformatics. 15;27(8):1157-1158. [3] Heidi et.al (2013) ACMG clinical lab standard for NGS, Genetic In Medicine, 15, 733-747.

5.3Quality Checks:-

Germline cases: Quality Parameters: Total Reads of sequence, Number of reads supporting Reference Allele, Number of reads supporting Variant Allele, Q20: Bases having Phred score > 20

ID	Gene	Chromosome: Position	rsID	Coverage	Total Read with > Q20	Reference Allele Supporting Reads	Variant Allele Supporting Reads	Variant Allele Frequency (%)
069	KDR	chr4: 55972974	**rs1870377**	7998	7970	3979	3981	49.95
087	NSD1	chr5: 176637471	rs28932177	746	732	346	386	52.7
087	RECQL4	chr8: 145739069	NA	645	627	317	310	49.4
087	ERCC4	chr16: 14020502	NA	1208	1185	637	548	46.2
088	RECQL4	chr8: 145741748	rs199773279	305	207	99	108	52.17
099	RET	chr10: 43609989	rs75225191	739	730	371	359	49.18

Germline cases: Quality Parameters: Total Reads of sequence, Number of reads supporting Reference Allele, Number of reads supporting Variant Allele, Q20: Bas

Table 5.1 Quality Check Of Samples(Germline

ID	Gene	Chromosome: Position	rsID	Coverage	Total Read with > Q20	Reference Allele Supporting Reads	Variant Allele Supporting Reads	Variant Allele Frequency (%)
101	FANCA	chr16: 89833576	rs17233141	447	439	211	228	51.94
101	TSC2	chr16: 2134438	rs45517331	590	567	268	299	52.73
101	RECQL4	chr8: 145737373	rs36078464	412	403	250	151	37.47
128	RECQL4	chr8: 145737349	NA	333	243	123	120	49.3
128	MSH6	chr2: 48028273	NA	420	259	139	120	46.3
140	VHL	chr3: 10183876	rs61758376	301	138	78	60	43.48

Table 5.2 Germline Cases

Somatic cases: Quality Parameters: Total Reads of sequence, Number of reads supporting Reference Allele, Number of reads supporting Variant Allele, Q20: Bases having Phred score > 20

ID	Gene	Chromosome:Position	rsID	Type	Total Read with > Q20	Reference Allele Supporting Reads	Variant Allele Supporting Reads	Variant Allele Frequency (%)
069	KDR	Chr4: 55972974	rs1870377	Normal	7972	3990	3971	49.88
				Tumor	6499	4109	2371	36.59
075	KDR	Chr4: 55980456	rs2305949	Normal	7944	7943	5	0.06
				Tumor	7960	5748	2202	27.7
129	PIK3CA	Chr3: 178952088	NA	Normal	8002	7977	13	0.16
				Tumor	8000	4763	3125	40.3
134	TP53	Chr17: 7577538	rs11540652	Normal	7983	7884	20	0.25
				Tumor	7997	881	6995	88.77
140	TP53	Chr17: 7577106	NA	Normal	2529	2493	2	0.08
				Tumor	2358	1101	1238	52.93
140	RB1	Chr13: 48955565	NA	Normal	3120	3117	3	0.10
				Tumor	115	80	35	30.43
140	VHL	Chr3: 10183876	rs61758376	Normal	156	85	64	42.95
				Tumor	325	250	67	21.14
144	TP53	Chr17: 7578203	NA	Normal	7998	7959	4	0.05
				Tumor	5706	2496	3198	56.16

Table 5.3 Somatic Cases

Somatic cases: Quality Parameters: Total Reads of sequence, Number of reads supporting Reference Allele, Number of reads supporting Variant Allele, Q20: Bases having Phred score > 20

ID	Gene	Chromosome:Position	rsID	Type	Total Read with > Q20	Reference Allele Supporting Reads	Variant Allele Supporting Reads	Variant Allele Frequency (%)
166	VHL	Chr3: 10183817	NA	Normal	20	20	0	0.0
				Tumor	176	155	17	9.88
166	VHL	Chr3: 10183836	NA	Normal	20	20	0	0.12
				Tumor	176	156	17	9.83
166	PTEN	Chr10: 89717681	NA	Normal	4049	4040	8	0.19
				Tumor	459	406	53	11.5
166	RB1	Chr13: 49027241	NA	Normal	2173	2173	0	0
				Tumor	247	203	44	17.5
166	RB1	Chr13: 49037870	NA	Normal	6680	6665	15	0.22
				Tumor	312	257	55	17.6
166	RB1	Chr13: 49037937	NA	Normal	6680	6670	8	0.12
				Tumor	312	231	81	25.96
180	TP53	Chr17: 7577567	NA	Normal	2386	2374	6	0.25
				Tumor	7321	4233	3068	42.02
231	KRAS	Chr12: 25398281	NA	Normal	10852	10835	15	0.13
				Tumor	4184	2770	1414	33.7

Table 5.4 Somatic cases

5.4. Drug Response in Non-public resources (PGMD from BioBase):-

rsID	ID	Gene	Drug	Study Disease	Results	Reference
rs1870377	069	KDR	Bevacizumab (Avastin)	Carcinoma, NSCLC, Colorectal Neoplasms, Prostatic Neoplasms	HT and HFSR may be marker for clinical outcome. Individuals with prostate cancer experiencing HT had longer PFS following bevacizumab therapy than those without toxicity (n=60, P=0.0009) and bevacizumab+sorafenib in patients with solid tumors (n=27, P=0.52)	[PMID: 20630084]
rs1870377	069	KDR	Doctaxel	Stomach and Biliary Tract Neoplasms	With 63 patients, this SNP is associated with poor TTF (P=0.029) and poor OS (p=0.03)	[PMID: 24123039]
rs1870377	069	KDR	Sorafenib **	Carcinoma, NSCL, Colorectal Neoplasms, Prostatic Neoplasms,	HFSR was a **marker** for prolonged PFS during sorfenib therapy (n=113, P=0.0003). **HT was a risk factor for HFSR** in patients treated with bevacizumab+/sorafenib. Patients with this SNP experienced greater risk of developing HT (n=170, P=0.0154) and HFSR (n=170, P=0.0136)	[PMID: 20630084]

Table 5.5 Drug Response

rsID	ID	Gene	Drug	Study Disease	Results	Reference
rs187 0377	069	KDR	Sunitinib	Carcinoma, Biliary Tract Neoplasms, RCC, Stomach Neoplasms	In a retrospective study, the PFS was shown to significantly improved in 136 clear-cell metastatic RCC patients in genetic variant carrier patients treated with sunitinib as compared to non-carriers. Prospective validation advocated for further genetic association.	[PMID: 21097692]
					The SNP was not found to be associated with sunitinib response/toxicity in (95) patients with RCC	[PMID: 22015057]

Additional Information (sample 069):-

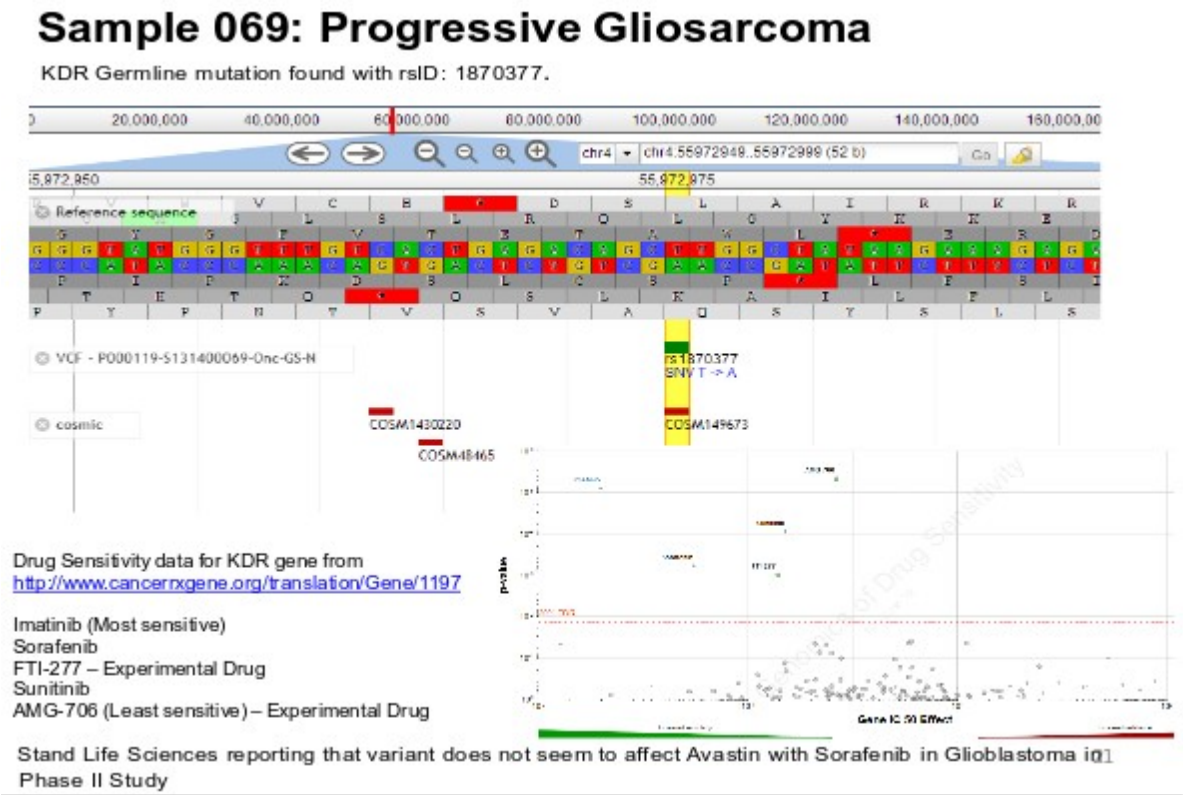
Few additional homozygous/heterozygous SNPs found in sample 069. One SNP has drugs response information. However, none of the drugs are appropriate for brain cancer as for high blood-brain barrier.

Table 5.6 Drug Response

ID	Gene	Chromosome:Position	rsID	Total Read with > Q20 or Q15	Reference Allele Supporting Reads	Variant Allele Supporting Reads	Variant Allele Frequency (%)	Drug Response Li
069	SMARCB1	chr22:24145675	rs5751738	7933	4048	3865	48.78	NA
069	GNA11	chr19:3119239	rs4900	1204	603	599	49.75	NA
069	TP53	chr17:7579472	rs1042522	3659	1699	1947	53.28	https://www.phar.org/rsid/rs10425 antineoplastic ag cisplatin, cyclophosphamid fluorouracil, pacli
069	FLT3	chr13:2861183	rs2491231	2168	1094	1067	49.33	NA

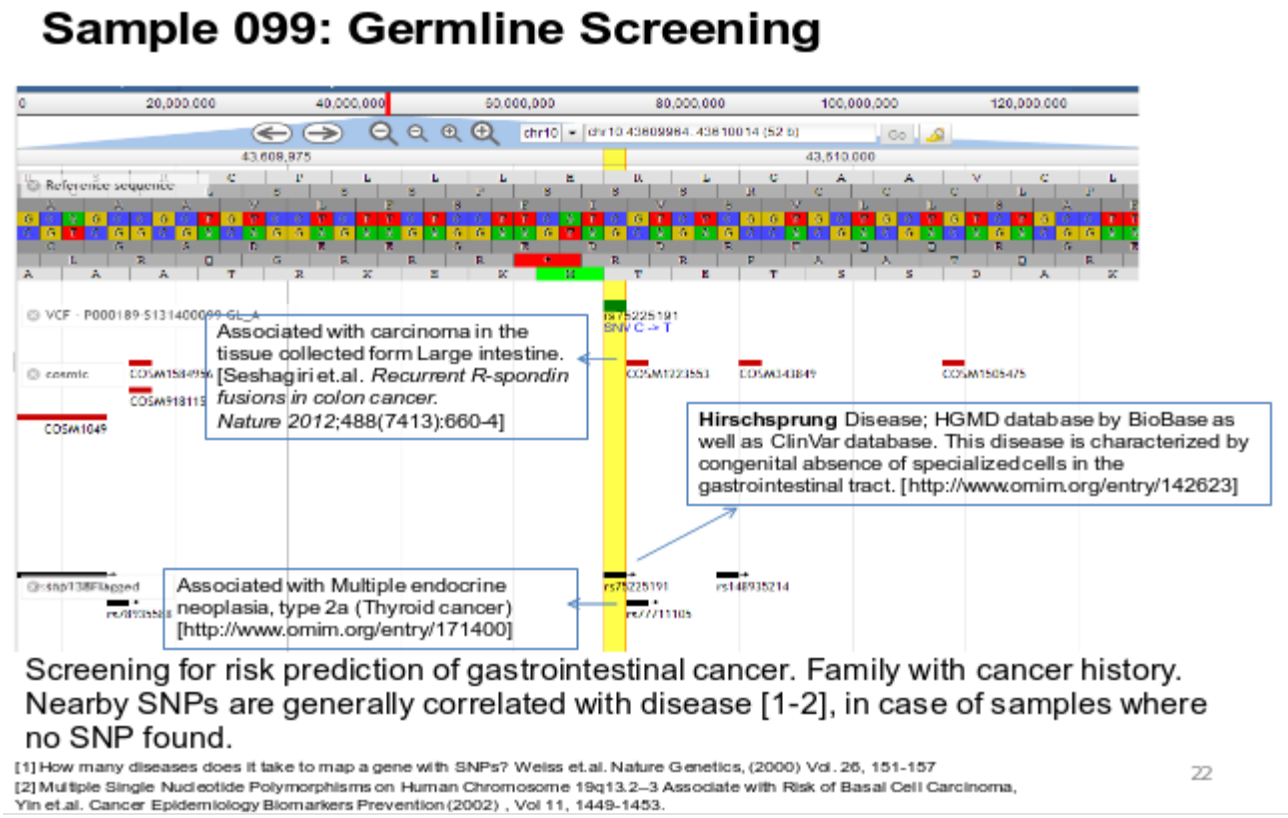
Table 5.7 Addl. Sample 69 info.

5.5. Experimental Results:-



Age-60 years
Gender-Male
Specimen type-FFPE(Tumor), Saliva(Normal)
Specimen type- Left Frontal Gliosarcoma
Lab id-069N, 069T
Clinical Indication- Progressive Gliosarcoma Grade 4, Ki-67 30%, GFAP positive with interspersed GFAP negative sarcomatous foci. Pre treatment.

Observation- Here germline mutation is shown on the KDR gene(rsid-1870377).
It is shown as somatic mutation also and we have taken support of COSMIC database in terms of sensitivity plot of different drugs tested from COSMIC database, for this mutation. Imatinib drug is not useful because of blood brain barrier. The clinician used Sorafenib which is a well known drug for brain cancer. We used Pubmed database for linking therapeutic options with the mutated gene. We also used Biobase. Pubmed is a public literature database having all abstracts. Imatinib is more suitable for blood cancer leukaemia. Another class is EGFR gene which has large number of drugs available and can help in sample 069. KDR has many synonyms like VEGFR-2 gene. If KDR does not have drug linkage associated with it then VEGFR-2 would have drug linkages. Going forward we can map genes with all their synonyms.



Age/DOB-11-3-1956

Gender-Male

Test- Cancer Predisposition Screening Test

Lab id- 099N

Specimen- Saliva

Clinical Indication- Risk prediction for gastrointestinal cancers. Family history-Grandfather and uncle had esophageal cancer.

Observation- ClinVar is the clinical variant of dbSNP. Patient has family history. This is saliva-sample only. No tumor. Patient came for checking predisposition. We are not reporting insertion-deletion and duplication at present because in most of the cases we get 1 nucleotide mutation. Insertion deletion and duplication have much higher false positive and false negative rates than those seen for SNP.

Sample 140: Somatic TP53 mutation

Indication: High grade urothelial carcinoma (chemotherapy: carboplatin & gemcitabine). 23
Father affected with lung cancer and sister affected with colon cancer.

Age/DOB- 16-12-1973

Gender-Female

Specimen type-FFPE Block(Tumor), Saliva(Normal)

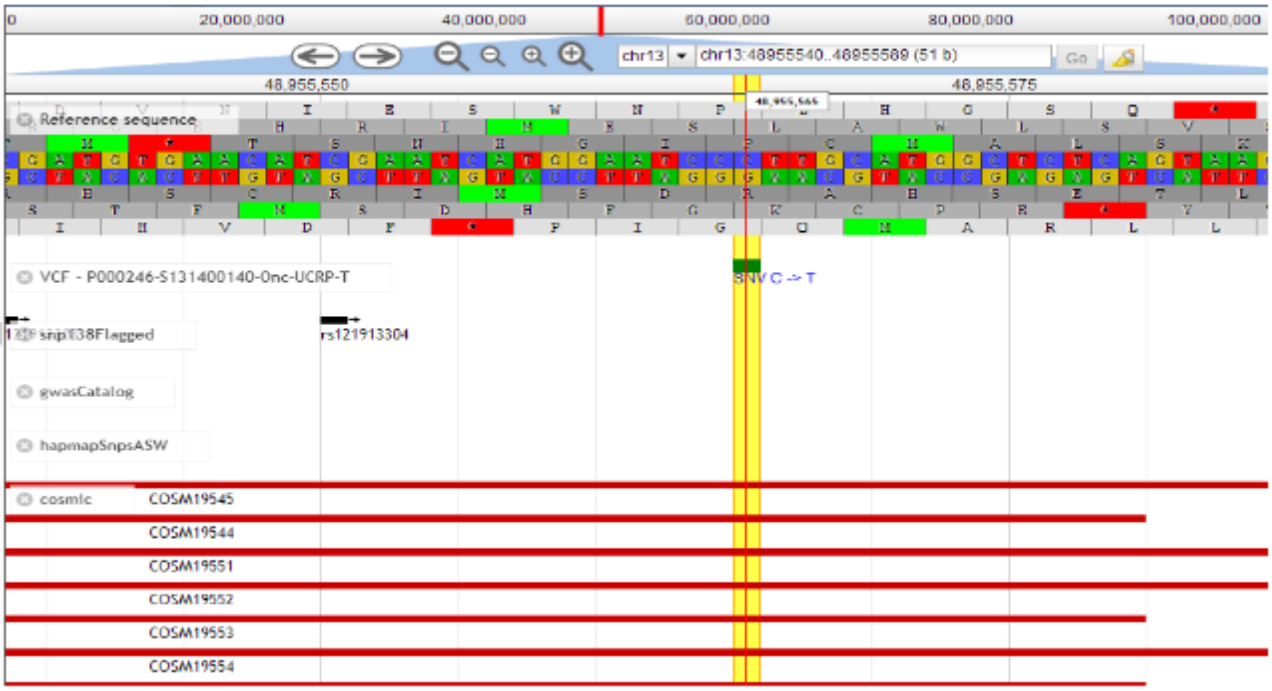
Specimen type- Left radical Nephrectomy & Retro caval lymph node

Lab id-140N, 140T

Clinical indication-High grade urothelial carcinoma, left renal pelvis-PT3N2, cancer stage-3, on chemotherapy (carboplatin and gemcitabine); Family history- father affected with lung cancer and sister affected with colon cancer.

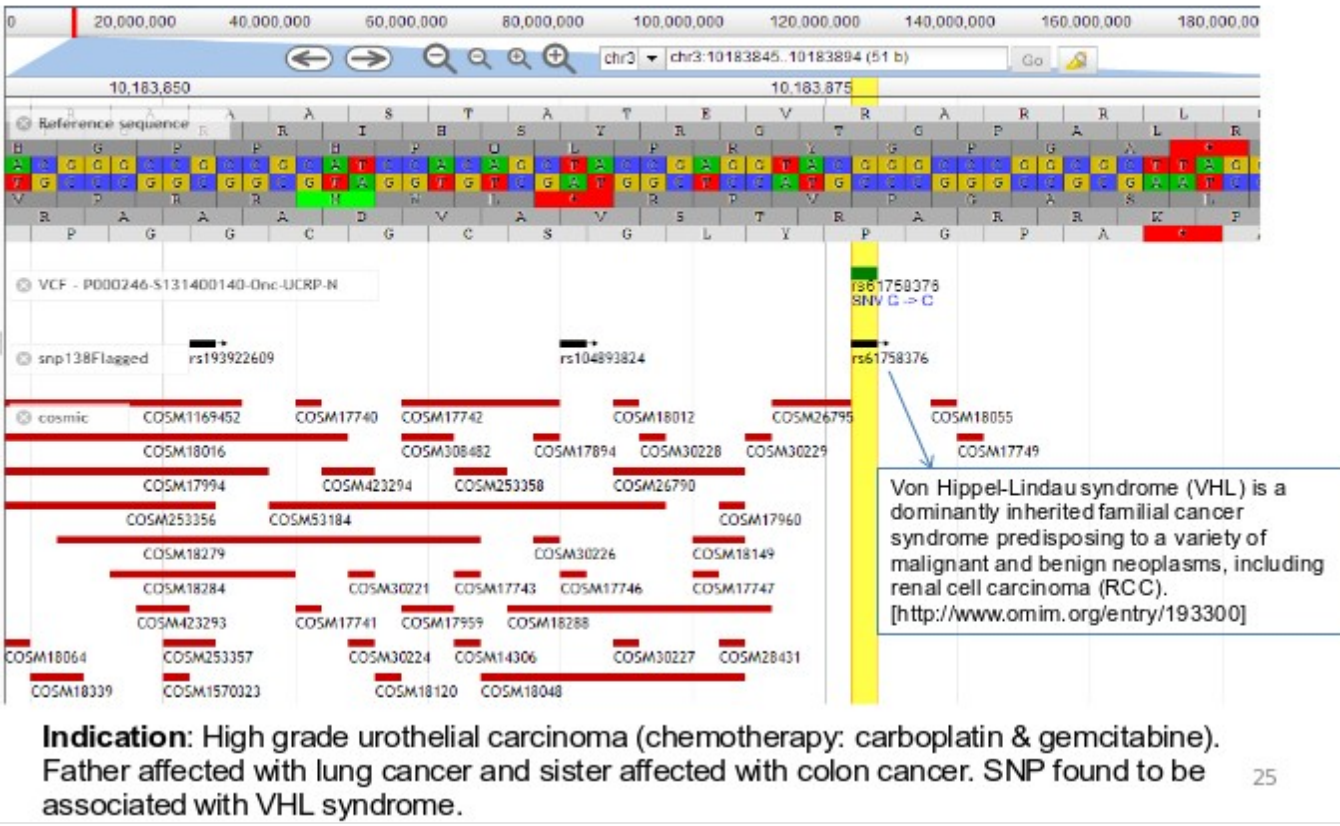
Observation- Somatic variations were detected in TP53 and RB1 genes and a germline variation was detected in VHL gene.

Sample 140: Somatic RB1 Mutation



Indication: High grade urothelial carcinoma (chemotherapy: carboplatin & gemcitabine). 24
Father affected with lung cancer and sister affected with colon cancer.

Sample 140: Germline Screening

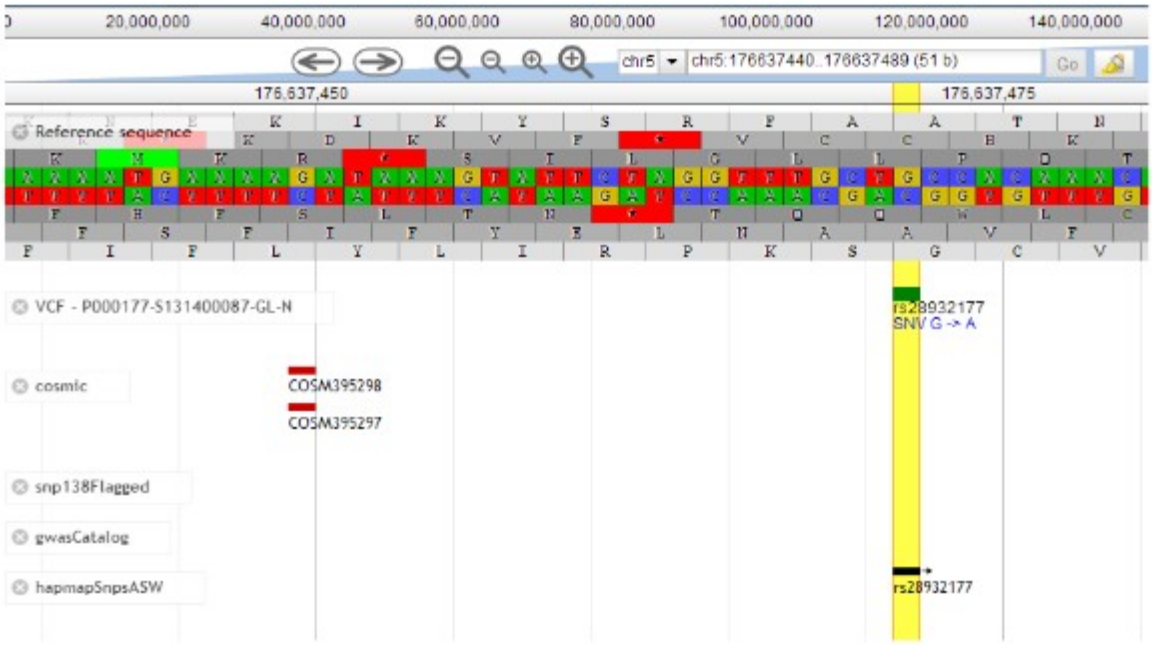


Indication: High grade urothelial carcinoma (chemotherapy: carboplatin & gemcitabine). Father affected with lung cancer and sister affected with colon cancer. SNP found to be associated with VHL syndrome. 25

Observation-TP53 is picked up as additional mutation which happens in case of cancer in general. Drug response is also linked with that.

This mutation is associated with VHL syndrome which is predisposed to many cancers, This patient is not yet diagnosed with cancer. VHL syndrome is shown on the upstream. This is makes him predisposed to variety of malignant and benign neoplasms, including renal cell carcinoma (RCC).VHL syndrome is shown by dbSNP.

Sample 087: Germline Screening



Indication: Unaffected, No family history of cancer
- No disease association of with this SNP

26

Age/DOB-15-01-1974

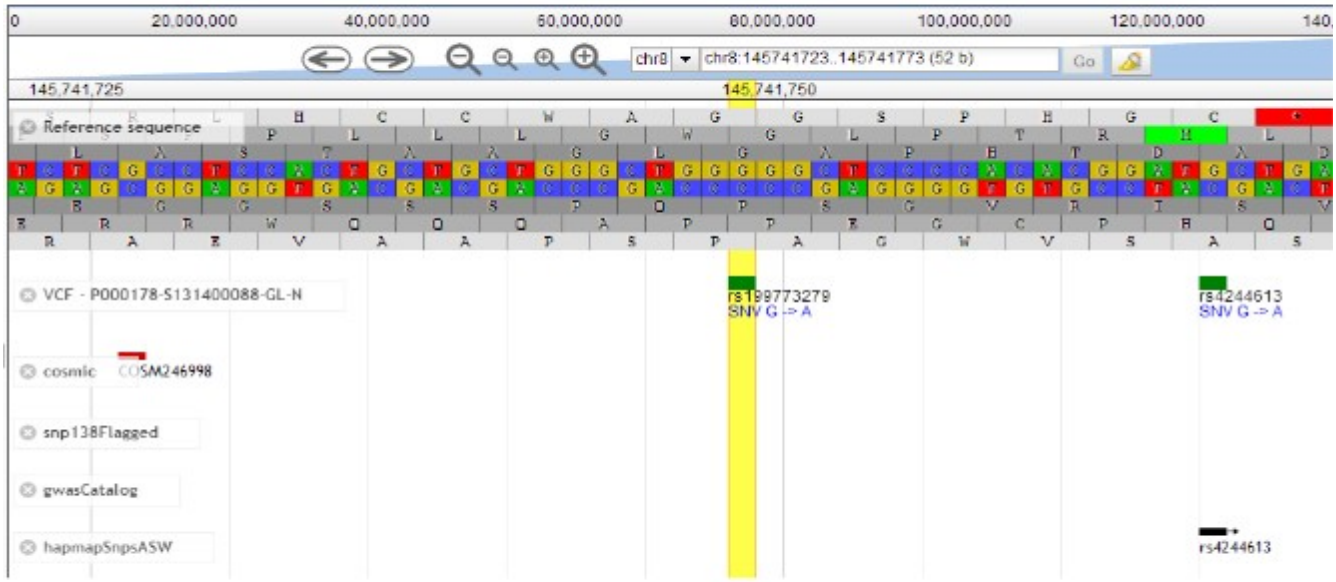
Gender- Female

Lab id- 087N

Specimen-Saliva

Clinical indications- Unaffected, No family history of cancer.

Sample 088: Germline Screening



Indication: Unaffected with no family history of cancer
- No disease association with this SNP

27

Age/DOB-45 years

Gender- Female

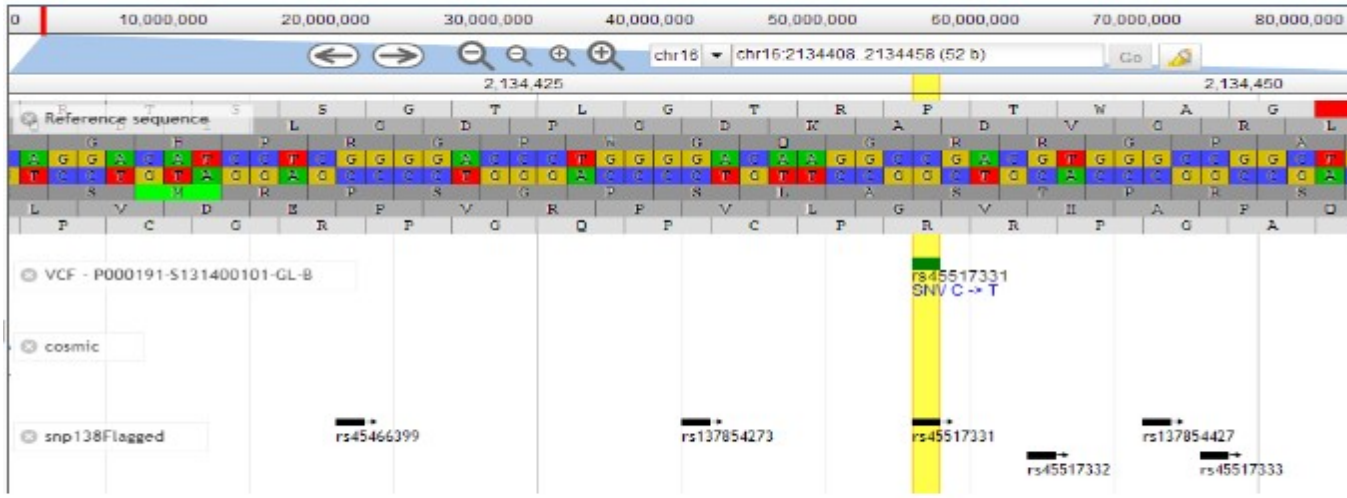
Lab id- 088N

Specimen-Saliva

Clinical indications- Unaffected, No family history of cancer.

Observations-Patient came for germline test only out of curiosity. No family history. No cancer found. No disease associated with this SNP. Cosmic support here is not effective because of being 20 nucleotides away from SNP.

Sample 101: Germline Screening



Indication: suspected predisposition as father suffered from stomach maligancy. The SNP has been marked clinically significant for *Tuberous sclerosis syndrome*.
[http://www.ncbi.nlm.nih.gov/medgen/C0041341].
Such patient can develop various cancers including renal-cell carcinoma

28

Age/DOB- 24 years

Gender- Male

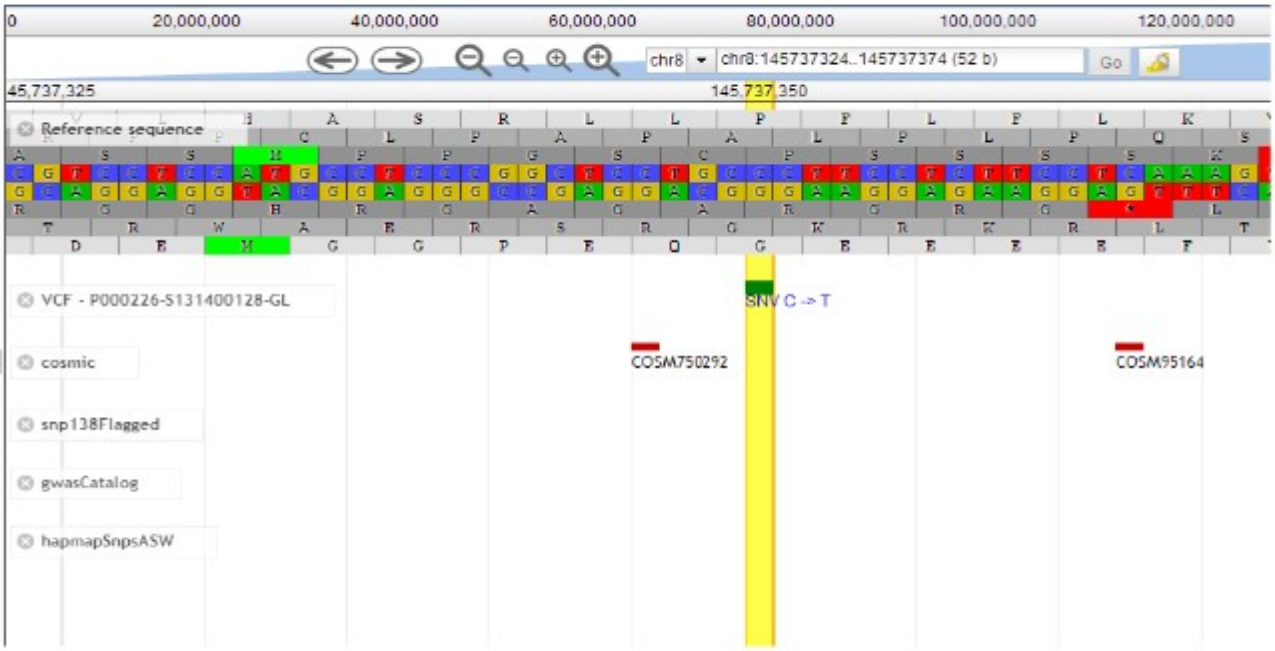
Lab id- 101N

Test- Cancer Predisposition Screening Test

Specimen- Saliva

Clinical Indications- Suspected predisposition to cancer.Oral Malignancy. Father suffered from abdominal malignancy.

Sample 128: Germline Screening



Indication: Screening for cancer. Family history of cancer
- No disease association for this mutation

29

Age/DOB- 30-12-1961

Gender- Male

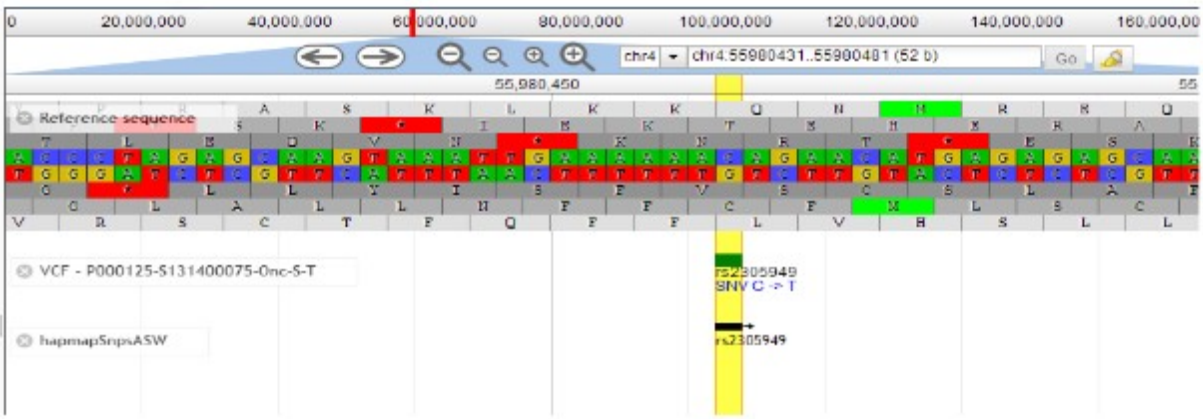
Test- Cancer Predisposition Screening Test

Specimen- Saliva

Clinical Indications- Patient came for risk prediction. Family history- Mother diagnosed with tumor in Uterus at the age of 65. Aunt diagnosed with tumor in uterus at the age of 55. Another aunt diagnosed at the age of 60 but cancer type unknown.

Observation-Only saliva sample. Family history of cancer. No cancer detected.

Sample 075: Somatic Mutations



Indication: Metastatic recurrent germ cell tumor. Chemotherapy with oxaliplatin and Gemcite.
- No disease association with this SNP

30

Age-69 years

Gender-Male

Specimen type-FFPE(Tumor), Saliva(Normal)

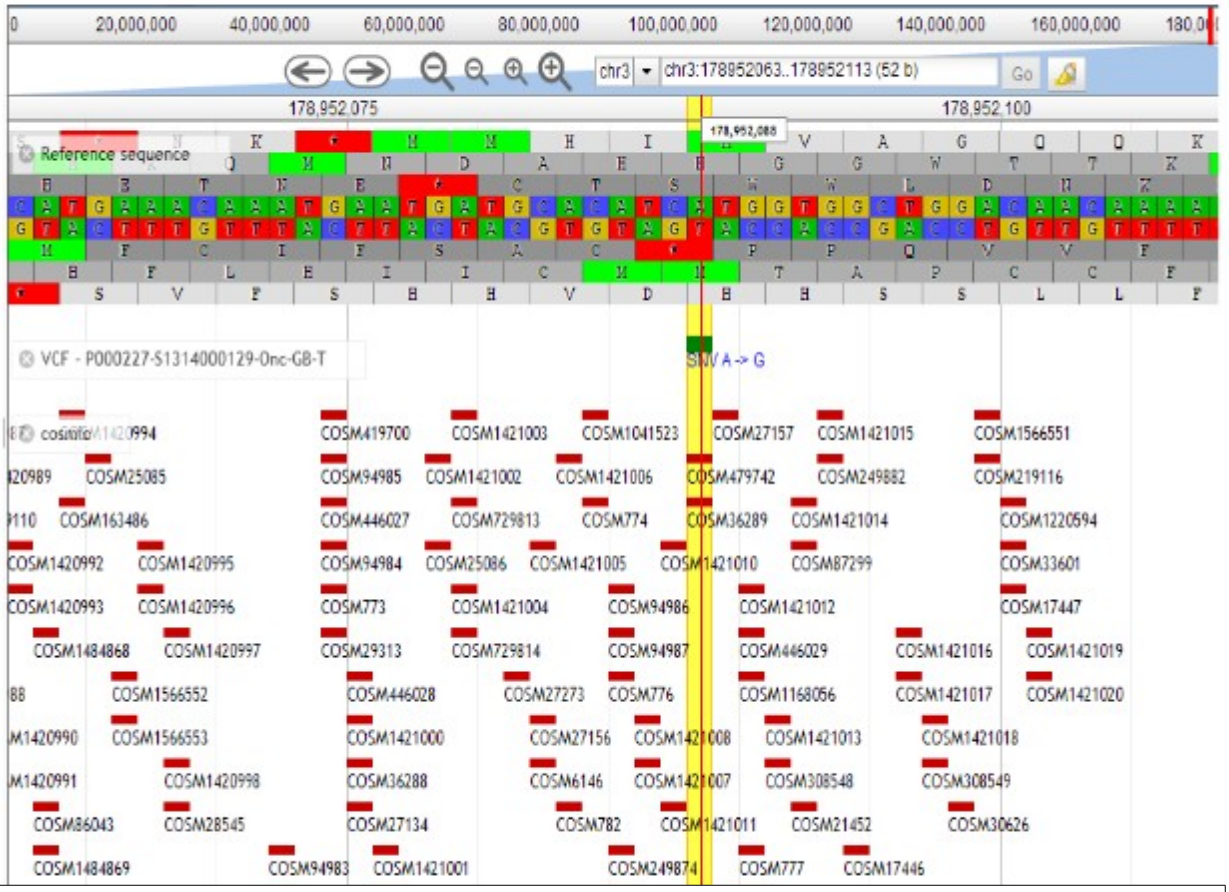
Specimen type- Retrocaecal deposit(FNAC)

Lab id-075N, 075T

Clinical Indication- Metastatic recurrent germ cell tumour, seminoma(tubular variant). Tumor cells marked with PLAP and c-kit.Treatment- Gemcite and oxaliplatin chemotherapy.

Observation-We are getting mutation but not associated with any disease as per COSMIC,dbSNP and Hapmap. Patient actually came with cancer. No disease association with this SNP.

Sample 129: Somatic PIK3CA mutations



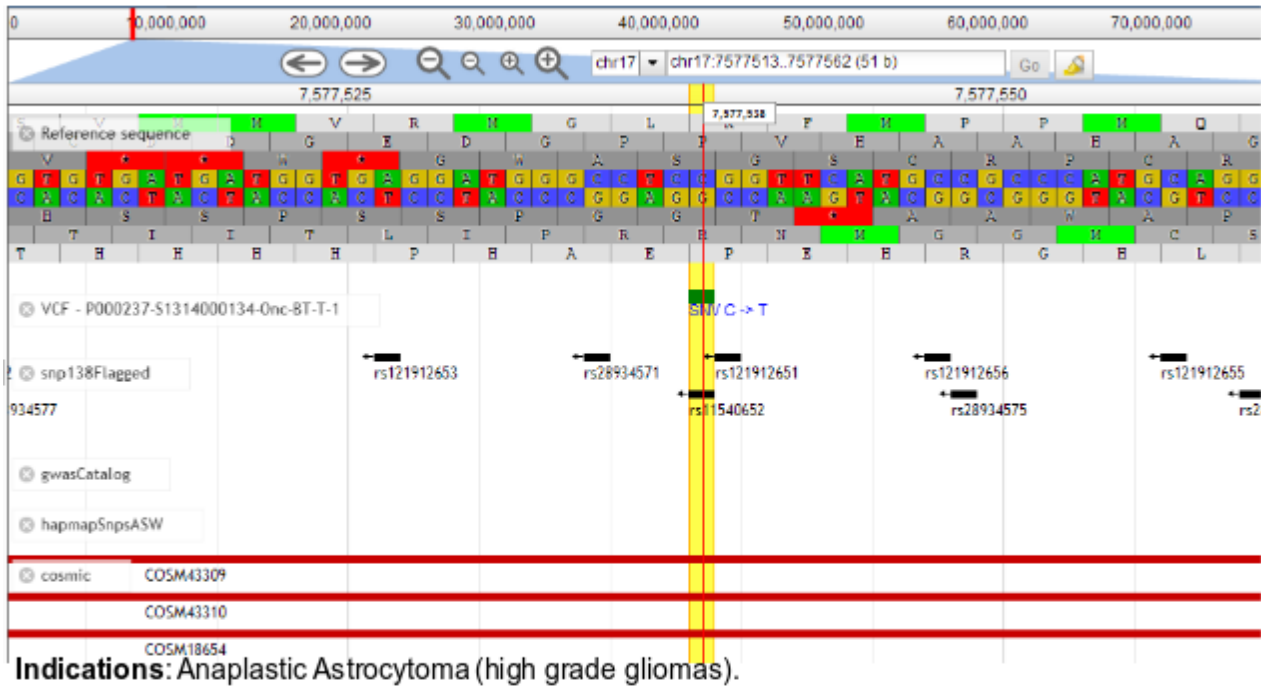
Age/DOB- 1-07-1945
Gender- Male
Specimen type-Saliva(Normal) & FFPE(Tumor)
Lab id-128N, 128T
Specimen site- Left temporal SOL
Clinical Indication- Glioblastoma- Grade 4, on concurrent temozolomide and radiation therapy.

Observation- Whatever therapeutic options and other information exists, COSMIC is showing at exact SNP. No direct evidence of therapeutic options based on the reported SNP mutations. We need to refine the substrategy to include greater degree of cancer biology information in terms of biochemical pathway and signal pathway to link any indirect evidence also.PIK3CA gene is very important in the cancer pathway where an upstream gene called MTOR for which there are a lot of known inhibitors.PIK3CA mutation has not given any direct drug response but one of the inhibitors of MTOR has relevance because the PIK3CA gene is susequently made dysfunctional. The drug is inhibiting MTOR in the same pathway where PIK3CA is playing an important role. The mutated gene like PIK3CA can be as long as 100 nucleotides.

Age/DOB- 22-10-2007
Gender- Male
Lab id-134N, 134T
Specimen type- Saliva(Normal), FFPE(Tumor)
Specimen site- Brain tumor(Left thalamic SOL Biopsy)
Clinical Indications- Anaplastic Astrocytoma, WHO Grade 3

Observations- LFS syndrome is associated with many cancers. dbSNP shows LFS. Patient has high grade brain cancer. The particular mutation is not associated with any disease but the nearby SNP shown by dbSNP on the downstream is associated with LFS.

Sample 134: Somatic TP53 mutation



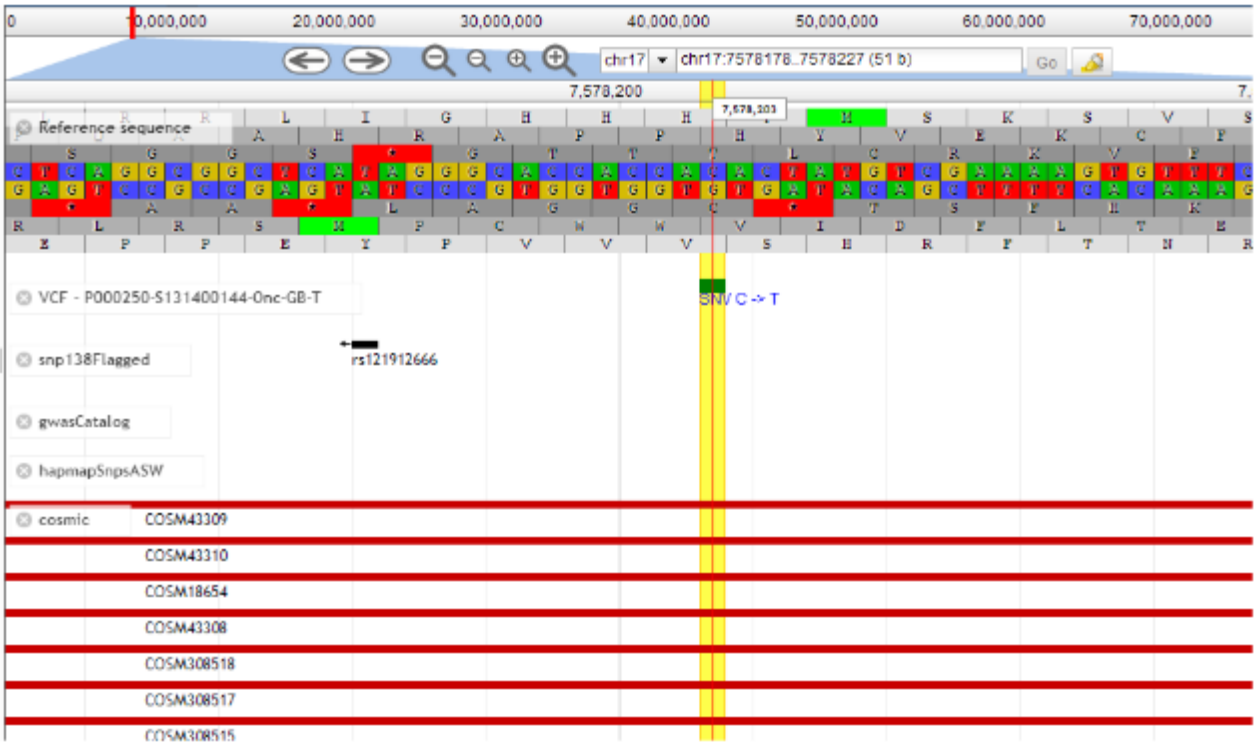
Indications: Anaplastic Astrocytoma (high grade gliomas).

Both the nearby SNPs are clinically marked significant for Li-Fraumeni syndrome. The Li-Fraumeni syndrome (LFS) is a clinically and genetically heterogeneous inherited cancer syndrome. LFS is characterized by autosomal dominant inheritance and early onset of tumors, multiple tumors within an individual, and multiple affected family members. The main hereditary predispositions for high grade gliomas are mainly neurofibromatosis type I, Li-Fraumeni syndrome. [http://www.omim.org/entry/151623] Chr 17: c.7577538C>T, TP53

Age/DOB- 12-02-1958
Gender- Female
Specimen type- Saliva(Normal) & FFPE Block(Tumor)
Lab id- 144N & 144T
Specimen type- Brain Tumor- Right frontal Glioma
Clinical Indications- Glioblastoma Multiforme, Grade 4, currently being treated with Bevacizumab.

Observations-COSMIC shows very large insertion or deletion. If we drill down it will give info for entire region and not for that pin pointed mutation.NGS is not technically suitable for large insertions or deletions.NGS is more suitable for 1 or 2 nucleotide mutations.

Sample 144: Somatic TP53 mutation



Indication: Glioblastoma Multiforme, currently being treated with Bevacizumab

Age/DOB- 15-04-1969
Gender-Female
Specimen type- Saliva(Normal) & FFPE (Tumor)
Lab id- 166N & 166T
Specimen type- Right frontal high grade glioma
Clinical Indications- Glioblastoma Multiforme

Observations-Somatic mutations in VHL, PTEN and RB1 genes were detected. This SNP is supported by COSMIC over 4 nucleotides.

Sample 166: Somatic Mutation

Indication: Glioblastoma Multiforme

Age/DOB- 1-3-1954

Gender- Female

Specimen type- Saliva(Normal), FFPE(Tumor)

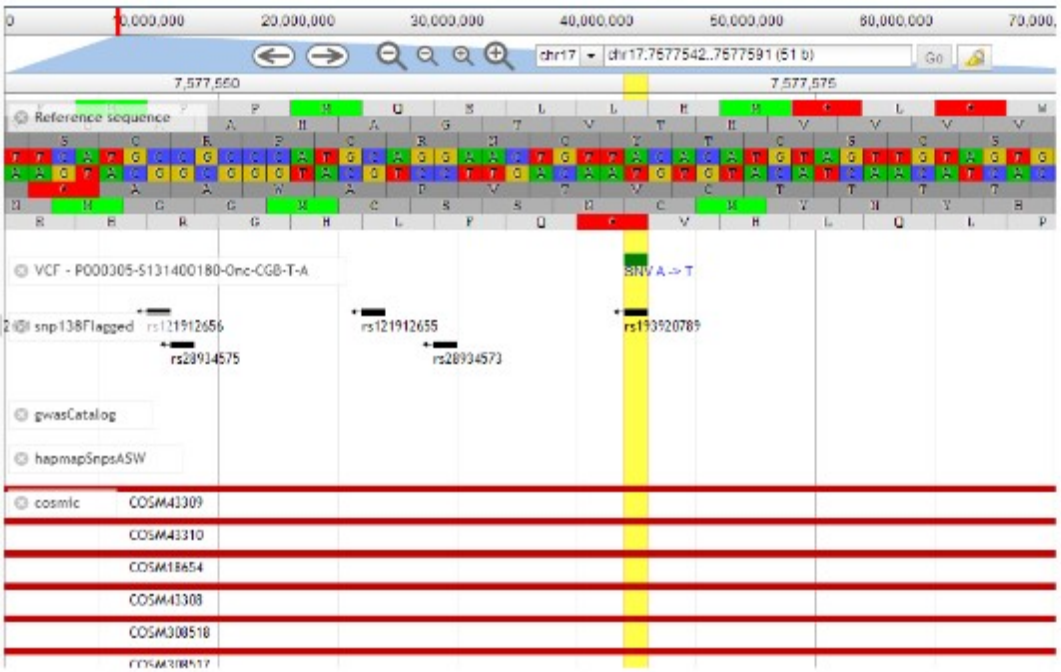
Lab id- 180N, 180T

Specimen site- Gall Bladder, Radical Cholecystectomy

Clinical Indication- Carcinoma gall bladder, G2 Stage,3B

Observations-Continuous redlines indicate either insertion or deletion so large number of nucleotides are either inserted or deleted.

Sample 180: Somatic Mutations



Indication: Carcinoma gall bladder

35

Additional SNPs report *

We have found additional SNPs across the samples. Following 4 rsids have been associated with drugs response(s) in PharmGKB (M. Whirl-Carrillo et.al (2012) Clinical Pharmacology & Therapeutics 92(4): 414-417 [https://www.pharmgkb.org/]).

rsID/Gene	Sample	TR (RD/AD) Variant Allele Freq (%)	Drug Response	Disease	Result	PubMed References ID
rs10821936 ARID5B	099 (GL)	155 (88/67) 43.23	Methotrexate	Precursor Cell Lymphoblastic Leukemia	Allele C is associated with **greater** methotrexate polyglutamate **accumulation** as compared to allele T (n=37, P=0.005) .	19684603
	088 (GL)	106 (59/47) 44.34				
rs13181 ERCC2 #	101 (GL)	181 (98/83) 45.86	Fluorouracil, leucovorin, oxaliplatin	Colorectal Neoplasms	Genotype GT is associated with **increased** PFS when treated with fluorouracil+oxaliplatin in patients with Colorectal Neoplasms as compared to genotype GG+TT (n=72, P=0.019).	21449681
	088 (GL)	104 (61/43) 41.35	Cisplatin, oxaliplatin, platinum compounds	Colorectal Neoplasms, Esophageal Neoplasms, Pancreatic Neoplasms	Genotype GT+TT may have **increased** survival when treated with platinum compounds as compared to genotype GG (n=121, P=0.003).	22026922

PFS: Progression-Free survival
TR: Total Read
RD: Reference Allele Supporting Read
AD: Variant Allele Supporting Read

Indications-
088: Unaffected, with no family history
099: Risk for gastrointestinal cancer. Family history of esophageal cancer.
101: Risk for oral cancer. Family history of abdominal cancer

* These SNPs are not reported by Strand Life Sciences, # Conflicting drug responses reported in literature[1]

Table 5.8 Addl. SNPs found

Additional SNPs report *

We have found additional SNPs across the samples. Following 4 rsids have been associated with drugs response(s) in PharmGKB (M. Whirl-Carrillo et.al (2012) Clinical Pharmacology & Therapeutics 92(4): 414-417 [https://www.pharmgkb.org/]).

rsID/Gene	Sample	TR (RD/AD) Variant Allele Freq (%)	Drug Response	Disease	Result	PubMed References ID
rs10821936 ARID5B	099 (GL)	155 (88/67) 43.23	Methotrexate	Precursor Cell Lymphoblastic Leukemia	Allele C is associated with **greater** methotrexate polyglutamate **accumulation** as compared to allele T (n=37, P=0.005) .	19684603
	088 (GL)	106 (59/47) 44.34				
rs13181 ERCC2 #	101 (GL)	181 (98/83) 45.86	Fluorouracil, leucovorin, oxaliplatin	Colorectal Neoplasms	Genotype GT is associated with **increased** PFS when treated with fluorouracil+oxaliplatin in patients with Colorectal Neoplasms as compared to genotype GG+TT (n=72, P=0.019).	21449681
	088 (GL)	104 (61/43) 41.35	Cisplatin, oxaliplatin, platinum compounds	Colorectal Neoplasms, Esophageal Neoplasms, Pancreatic Neoplasms	Genotype GT+TT may have **increased** survival when treated with platinum compounds as compared to genotype GG (n=121, P=0.003).	22026922

PFS: Progression-Free survival
TR: Total Read
RD: Reference Allele Supporting Read
AD: Variant Allele Supporting Read

Indications-
088: Unaffected, with no family history
099: Risk for gastrointestinal cancer. Family history of esophageal cancer.
101: Risk for oral cancer. Family history of abdominal cancer

* These SNPs are not reported by Strand Life Sciences, # Conflicting drug responses reported in literature [1]

Table 5.9 Addl. SNPs found

Somatic cases of additional SNPs *

rsID/Gene	Sample	TR (RD/AD) Variant Allele Freq (%)	Drug Response	Disease	Result	PubMed References
rs1042522 TP53	075 (SM)	Normal: 4438 (2087/2332) 52.77 Tumor: 3467 (1985/1453) 42.26	Carboplatin and Gemcitabine	Sarcoma, Breast and Ovarian Neoplasms, Grandular and Epithelial Neoplasms	Carboplatin: Homozygotes for the minor allele TP53-72Pro of the Arg72Pro SNP in the TP53 gene showed a **better response rate** (54.3%) than those for the major allele TP53-72Arg (29.1%; P = 4.4 × 10(-5)) irrespective of therapeutic regimens. Minor allele homozygotes had significantly longer progression-free and overall survivals than major allele homozygotes	18357466 and 23423487 (for gastric cancer adjuvant chemotherapy). 20940192 and 18618574 (response rate of carboplatin and germcitabine in NSCLC)

Mean Coverage for sample 075 is 6042 for Normal and 6226 for Tumor samples

PFS: Progression-Free survival
TR: Total Read
RD: Reference Allele Supporting Read
AD: Variant Allele Supporting Read

Indications-
075: Metastatic recurrent germ cell tumor, seminoma (tumor of testis). On Gemcitabine and Carboplatin

* These SNPs are not reported by Strand Life Sciences

1

Table 5.10 Addl. SNPs found

Max health care strategic output

Genetic Testing reports are now validated (for Oncology reports) in MHC for the mutations and drugs options using independent tools and sources. Established Gene Mutation Analysis platform of software pipelines using Open Sources tools (BWA, GATK, Bioconductor, Samtools, VarScan2) on Linux based Operating System. Moreover, we have implemented GATK best practices for mutation detection . Developed browser-based tools for visual analysis of the gene mutations using Open Source tools (JBrowse, HTML5, Javascript, Python). Browser-based visual tool enables clinicians to look at upstream/downstream of the mutation/SNP. The visual tool can be used to identify mutation hot-spots, driver mutations and passenger mutations. Quality Checks are made part of the software tools for minimizing possible false-positive detection of gene mutation/SNP. The quality checks based on minimum coverage, Phred scores, ratio of supporting variant and reference alleles. Mutations are further validated using COSMIC, HapMap, Geome Wide Association Studies and clinically flagged SNPs from dbSNP. Started working towards personalized medication based on suggestion for mutation-specific therapeutics from pharmacogenomics based database like PharmGKB (public), Clinical Trials (public) and BioBase (propriety)

Chapter 6

Conclusion and Future Scope

6.1.Web portal development for visualizing gene mutations and mapping to drug responses:

Need for development of database and web-portal:
The project envisages to use both document-oriented (NoSQL) and relational database (SQL) in order to pre-process as well as post-process the diverse data format of the output as well as the input data. For example, the genomic data would be stored both in ASCII (text) as well as binary format.

6.2.Relational Database Structure:

The dynamic MVC or MVT web-portal development:
The model-view-controller (MVC) pattern is an architectural pattern used primarily in creating Graphic User Interfaces (GUIs). The major premise of the pattern is based on modularity and it is to separate three different aspects of the GUI: the data (model), the visual representation of the data (view), and the interface between the view and the model (controller). The primary idea behind keeping these three components separate is so that each one is as independent of the others as possible, and changes made to one will not affect changes made to the others. In this way, for instance, the GUI can be updated with a new look or visual style without having to change the data model or the controller.

6.3.Python based Django based web-framework:

Django is a free and open source web application framework, written in Python, which follows the model–view–controller architectural pattern. Django's primary goal is to ease the creation of complex, database-driven websites. Django emphasizes reusability and "pluggability" of components, rapid development, and the principle of don't repeat

yourself. Python is used throughout, even for settings, files, and data models. Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models. Django can be used for building high-performing, elegant Web applications quickly.

The developed tools and application in this project would be used for visualization. The integrated data available on single platform can be used for providing near real-time access to gene mutation analysis and its (possible) mapping to the drug responses. The differential drug responses information associated with the gene mutations, would lead to practice of personalized medication in future.

6.4.Summary of Contribution-

We have established Gene Mutation Analysis platform of software pipelines using Open Sources tools (BWA, GATK, Bioconductor, Samtools, VarScan2) on Linux based Operating System. Moreover, we have implemented GATK best practices for mutation detection. Also, developed browser-based tools for visual analysis of the gene mutations using Open Source tools (JBrowse, HTML5, Javascript, Python). Browser-based visual tool enables clinicians to look at upstream/downstream of the mutation/SNP. The visual tool enables identification of mutation hot-spots, driver mutations and passenger mutation Quality Checks are made part of the software tools for minimizing possible false-positive detection of gene mutation/SNP. The quality checks based on minimum coverage, Phred scores, ratio of supporting variant and reference alleles. Mutations are further validated using COSMIC, HapMap, Geome Wide Association Studies and clinically flagged SNPs from dbSNP.

With help of the above platform and visual tools, the Oncologist is now able to do a triage – look at a patient, look at their lab results and look at their genetic profile visually; they are now able to do more personalized medicine.

References

[1] G. R. Abecasis, A. Auton, L. D. Brooks et al. “An integrated map of genetic variation from 1,092 human genomes.,” *Nature*, vol. 491, no. 7422, pp. 56–65, Nov. 2012.

[2] M. A. DePristo, E. Banks, R. Poplin et al. “A framework for variation discovery and genotyping using next-generation DNA sequencing data.,” *Nat. Genet.*, vol. 43, no. 5, pp. 491–8, May 2011.

[3] G. P. Sykiotis, G. D. Kalliolias, and A. G. Papavassiliou, “Pharmacogenetic principles in the Hippocratic writings.,” *J. Clin. Pharmacol.*, vol. 45, no. 11, pp. 1218–20, Nov. 2005.

[4] S. H. Katsanis, G. Javitt, and K. Hudson, “Public health. A case study of personalized medicine.,” *Science*, vol. 320, no. 5872, pp. 53–4, Apr. 2008.

[5] R. Drmanac, “The advent of personal genome sequencing.,” *Genet. Med.*, vol. 13, no. 3, pp. 188–90, Mar. 2011.

[6] T. A. Manolio, “Genomewide association studies and assessment of the risk of disease.,” *N. Engl. J. Med.*, vol. 363, no. 2, pp. 166–76, Jul. 2010.

[7] R. A. Wilke, L. B. Ramsey, S. G. Johnson et al. “The clinical pharmacogenomics implementation consortium: CPIC guideline for SLCO1B1 and simvastatin-induced myopathy.,” *Clin. Pharmacol. Ther.*, vol. 92, no. 1, pp. 112–7, Jul. 2012.

[8] T. A. Manolio, “Bringing genome-wide association findings into clinical use.,” *Nat. Rev. Genet.*, vol. 14, no. 8, pp. 549–58, Aug. 2013.

[9] V. Marx, “Drilling into big cancer-genome data,” *Nat. Methods*, vol. 10, no. 4, pp. 293–297, Mar. 2013.

[10] T. A. Manolio, R. L. Chisholm, B. Ozenberger et al. “Implementing genomic medicine in the clinic: the future is here.,” *Genet. Med.*, vol. 15, no. 4, pp. 258–67, Apr. 2013.

[11] P. B. Chapman, A. Hauschild, C. Robert et al. “Improved survival with vemurafenib in melanoma with BRAF V600E mutation.,” *N. Engl. J. Med.*, vol. 364, no. 26, pp. 2507–16, Jun. 2011.

[12] Y. Zhang, R. Proenca, M. Maffei, M. Barone et al. “Positional cloning of the mouse obese gene and its human homologue.,” *Nature*, vol. 372, no. 6505, pp. 425–32, Dec. 1994.

[13] K. I. Melkersson, A.-L. Hulting, and K. E. Brismar, “Elevated Levels of Insulin, Leptin, and Blood Lipids in Olanzapine-Treated Patients With Schizophrenia or Related Psychoses,” *J. Clin. Psychiatry*, vol. 61, no. 10, pp. 742–749, Oct. 2000.

[14] V. Tillmann, L. Patel, M. S. Gill et al. “Monitoring serum insulin-like growth factor-I (IGF-I), IGF binding protein-3 (IGFBP-3), IGF-I/IGFBP-3 molar ratio and leptin during growth hormone treatment for disordered growth.,” *Clin. Endocrinol. (Oxf).*, vol. 53, no. 3, pp. 329–36, Oct. 2000.

[15] A. O’Driscoll, J. Daugelaite, and R. D. Sleator, “‘Big data’, Hadoop and cloud computing in genomics.,” *J. Biomed. Inform.*, vol. 46, no. 5, pp. 774–81, Oct. 2013.

[16] I. S. Kohane, “Using electronic health records to drive discovery in disease genomics.,” *Nat. Rev. Genet.*, vol. 12, no. 6, pp. 417–28, Jun. 2011.

[17] J. Starren, M. S. Williams, and E. P. Bottinger, “Crossing the omic chasm: a time for omic ancillary systems.,” *JAMA*, vol. 309, no. 12, pp. 1237–8, Mar. 2013.

[18] E. R. Mardis, “Next-generation DNA sequencing methods.,” *Annu. Rev. Genomics Hum. Genet.*, vol. 9, pp. 387–402, Jan. 2008.

[19] J. Shendure and H. Ji, “Next-generation DNA sequencing.,” *Nat. Biotechnol.*, vol. 26, no. 10, pp. 1135–45, Oct. 2008.

[20] L. Chin, W. C. Hahn, G. Getz, and M. Meyerson, “Making sense of cancer genomic data.,” *Genes Dev.*, vol. 25, no. 6, pp. 534–55, Mar. 2011.

[21] R. C. Gentleman, V. J. Carey, D. M. Bates et al. “Bioconductor: open software development for computational biology and bioinformatics.,” *Genome Biol.*, vol. 5, no. 10, p. R80, Jan. 2004.

[22] D. C. Koboldt, Q. Zhang, D. E. Larson et al. “VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.,” *Genome Res.*, vol. 22, no. 3, pp. 568–76, Mar. 2012.

[23] S. B. Ng, E. H. Turner, P. D. Robertson et al. “Targeted capture and massively parallel sequencing of 12 human exomes.,” *Nature*, vol. 461, no. 7261, pp. 272–6, Sep. 2009.

[24] M. N. Bainbridge, M. Wang, D. L. Burgess et al. “Whole exome capture in solution with 3 Gbp of data.,” *Genome Biol.*, vol. 11, no. 6, p. R62, Jan. 2010.

[25] L. Ding, M. J. Ellis, S. Li et al. “Genome remodelling in a basal-like breast cancer metastasis and xenograft.,” *Nature*, vol. 464, no. 7291, pp. 999–1005, Apr. 2010.

[26] P. J. Campbell, P. J. Stephens, E. D. Pleasance et al. “Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.,” *Nat. Genet.*, vol. 40, no. 6, pp. 722–9, Jun. 2008.

[27] R. Beroukhim, C. H. Mermel, D. Porter et al. “The landscape of somatic copy-number alteration across human cancers.,” *Nature*, vol. 463, no. 7283, pp. 899–905, Feb. 2010.

[28] T. J. Ley, E. R. Mardis, L. Ding et al. “DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.,” *Nature*, vol. 456, no. 7218, pp. 66–72, Nov. 2008.

[29] E. R. Mardis, L. Ding, D. J. Dooling et al. “Recurring mutations found by sequencing an acute myeloid leukemia genome.,” *N. Engl. J. Med.*, vol. 361, no. 11, pp. 1058–66, Sep. 2009.

[30] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, “A method and server for predicting damaging missense mutations.,” *Nat. Methods*, vol. 7, no. 4, pp. 248–9, Apr. 2010.

[31] P. Kumar, S. Henikoff, and P. C. Ng, “Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.,” *Nat. Protoc.*, vol. 4, no. 7, pp. 1073–81, Jan. 2009.

[32] S. E. Flanagan, A.-M. Patch, and S. Ellard, “Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations.,” *Genet. Test. Mol. Biomarkers*, vol. 14, no. 4, pp. 533–7, Aug. 2010.

[33] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller et al. “Predicting the functional effect of amino acid substitutions and indels.,” *PLoS One*, vol. 7, no. 10, p. e46688, Jan. 2012.

[34] S. A. Forbes, N. Bindal, S. Bamford et al. “COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer.,” *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D945–50, Jan. 2011.

[35] A. Petitjean, E. Mathe, S. Kato et al. “Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database.,” *Hum. Mutat.*, vol. 28, no. 6, pp. 622–9, Jul. 2007.

[36] B. Gottlieb, L. K. Beitel, A. Nadarajah et al. “The androgen receptor gene mutations database: 2012 update.,” *Hum. Mutat.*, vol. 33, no. 5, pp. 887–94, May 2012.

[37] D. M. Altshuler, R. A. Gibbs, L. Peltonen et al. “Integrating common and rare genetic variation in diverse human populations.,” Nature, vol. 467, no. 7311, pp. 52–8, Sep. 2010.