Aalto University
School of Science
Degree Programme in Computational and Systems Biology (euSYSBIO)

Sailendra Pradhananga

Association studies of exome sequencing data of lung cancer patients undergoing gemcitabine/carboplatin chemotherapy with myelosuppression toxicity

Master's Thesis
Espoo, July 31, 2015

Supervisors: Prof. Joakim Lundeberg
KTH Royal Institute Of Technology
Prof. Samuel Kaski
Aalto University
Advisor: Benjamin Sigurgeirsson
Science for life Laboratory

Aalto University
School of Science
Degree Programme in Computational and Systems Biology (euSYSBIO)

ABSTRACT OF MASTER'S THESIS

Author: Sailendra Pradhananga

Title:
Association studies of exome sequencing data of lung cancer patients undergoing gemcitabine/carboplatin chemotherapy with myelosuppression toxicity

Date: July 31, 2015 **Pages:** viii + 74

Major: Computational Systems Biology **Code:** BB201X

Supervisors: Prof. Joakim Lundeberg
Prof. Samuel Kaski

Advisor: Benjamin Sigurgeirsson

Chemotherapeutic drugs such as carboplatin/gemcitabine administerd to non small cell lung cancer (NSCLC) patients frequently induce myelosuppression toxicity potentially leading to reduction or removal of drugs. We set out to identify the genetic variants associated with toxicity induced myelosuppression by whole exome sequencing (WES) 216 NSCLC patients and associating biallelic variants with different quantitative and qualitative measurements of myelosuppression phenotypes.

WES identified on average 29834 variants in each patient. Biallelic variants from combined patients genotype were associated with each myelosupression phenotype - Thrombocytopenia (TPK), Leukopenia (LPK) and Neutropenia (NPK) using quantitative Log-transformed (LN) and Empirical normal quantile transformation(ENQT) phenotypes and qualitative high/low toxicity study design in linear and logistic regression methods. Additionally, gene-based SKATO tests were performed for transformed quantitative phenotypes to investigate enrichment of rare and common variants.

Due to sample size limitation, none of the variants reached multiple corrected Bonferroni significant or FDR-BH p - values. However, variants with p-value $< 1.00 \times 10^{-3}$ in each study design were evaluated for high toxicity. We found five, one and two variants for TPK, LPK and NPK respectively associated in all quantitative and qualitative single variant association study. Furthermore, single variant rs4808 in *CAPZA2* and rs8018462 in *SLC7A7* genes were identified by Gene-based SKATO test for TPK and LPK phenotypes. This results could implicate association of *CAPZA2* and *SLC7A7* to TPK and LPK myleosupression. However, validation and replication of the variants and genes needs to be further studied in an independent studies.

Keywords: Exome sequencing, Lung Cancer, Myelosupression, Thrombocytopenia, Leukopenia, Neutropenia

Language: English

Acknowledgements

Foremost, I would like to thank the European commission and it's Education, Audiovisual and Culture Executive Agency for providing me with opportunity and scholarship to pursue my master's degree in Computational and Systems Biology (euSYSBIO). I would like to express my sincere gratitude to Prof. Erik Aurell, Prof. Juho Rousu and Prof. Isabel Sá Correia; the coordinators of the program, for organizing an excellent program and wish all the best for its continuity. I would like to thank Karin Knutsson, Paivi Koivunen and Ana Barbosa for taking care of all the practical matters. I am also thankful to all the professors, teachers, staffs from all the three universities and everyone who supported and helped in each and every step of my learning.

I am grateful to Prof. Joakim Lundeberg (KTH Royal Institute Of Technology) and Prof. Samuel Kaski (Aalto University) for being my thesis supervisors and providing valuable insights about the project.

I would also like to express my sincere gratitude to Benjamin Sigurgeirsson for being my advisor and providing all the support and guidance throughout the thesis project. I have learned a lot from him and without him it would have been impossible to complete project. He has provided with all the guidance and knowledge needed to complete the project.

I would also like to thank Science for Life Laboratory (ScilifeLab) and UPPMAX for providing all the necessary infrastructure to carry out the project. Also, I would like to thank all the members in ScilifeLab genomics platform for providing a friendly atmosphere at work.

Last but not least, I would like to remember my family; without their love, care, support and blessings I would not have been where I am today.

Espoo, July 31, 2015

Sailendra Pradhananga

Abbreviations and Acronyms

GWAS	Genome wide association study
AC	Adenocarcinoma
LCC	Large cell lung cancer
SCC	Squamous cell lung cancer
NSCLC	Non small cell lung cancer
CTC	Common toxicity criteria
VCF	Variant calling files
NGI	National genomics infrastructure
SNP	Single nucleotide polymorphism
MAF	Minor allele frequency
QC	Quality control
SKAT	Sequence kernel association test
SKATO	Sequence kernel association test optimal
RBC	Red blood cells
WBC	White blood cells
TPK	Thrombocytopenia
LPK	Leukopenia
NPK	Neutropenia
LN	Logarithm normalized
ENQT	Empirical normal quantile transformation
bwa	Burrows-Wheeler alignment
sam	Sequence alignment/Map
indel	Insertion and Deletion
IBD	Identity by descent
IBS	Identity by state
ADS	Adverse drug reaction
FDR	False discovery rate
SVA	Single variant association
HSC	Haematopoeitic stem cells
LPI	Lysinuric protein intolerance

CAST	Cohort Alleleic Sum test
CMC	Combined multivariate and collapsing

Contents

Chapter 1

Introduction

1.1 Cancer Pharmacogenomics

Pharmacogenomics is the application of modern genomic medicine in drug therapy. It deals with the interaction between the human genetic components and effect of the drug uptake mechanisms - pharmacokinetics and pharmocodymanics. One aspect of pharmacokinetics is time duration a drug remains in a body fluids after administration of a certain dose. The primary objective of pharmacokinetics is to increase efficacy and decrease toxicity of a drug. Pharmacodynamics studies effect of drugs on body indicating desired results from certain administered doses [1].

Pharmacogenomics research aims at identifying genes or gene variants involved in the interaction between the drugs and body. Genetic variants can have profound influence on effect and dose requirement to produce the desired effect. Pharmacogenomics have potential to elucidate adverse and positive influence of drug based on these genetic make-up of individuals. Modern-day advancement in genotyping technologies from microarray to massively parallel DNA sequencing provides unprecedented potential to interrogate the nucleotide to single base-pair level. Germline variations within patients help in understanding individualized response to a drug [2]. This understanding results in correct dosing and effective treatment strategies for various human diseases. Specifically, these pharmacogenomics approaches are taken towards cancer and neurological disorders [3]. These cancer chemotherapeutic drugs target cellular machineries involved in cancer growth. However, these drugs can induce adverse reaction in normal cells leading to undesired complications.

Genome-wide association studies are used to interrogate relationship between phenotype of interest and genotype of an individuals. In regards with

the pharmacogenomics, genome wide association studies consider the traits as the drug dose dependent responses or the adverse event profiles. Association methods are used to discover novel associations between the drugs and genes cases and control i.e cases being the patients that show adverse drug reaction and controls reacting 'normally' from drugs [2–4]. A study in genetic variation in *TPMT* gene was associated with myelosupression after 6-mercaptopurine (6-MP) and 6-thioguanine (6-TG) therapy [5]. Another study of pharmacogenomics influencing the drug therapy is based on the effect of the genetic variation of *UGT1A1* gene in irinotecan- induced neutropenia [6, 7]. Similar study associated the variations in *CYP2D6* gene with Tamoxifen (an oestrogen inhibitor) induced toxicity [8]. Another successful study associated the genetic variants with *DPYD* with the 5- Fluorouracil toxiciteis. Specific mutation in *DPYP*2A* gene are associated with the 5-FU associated leucopenia and severe mucositis [9].

However, chemotherapy toxicity traits are multi-genic with smaller influence and follow complicated underlying biological mechanisms. Most of these phenotypes are probably complex traits dependent upon multiple SNPs in modifiers genes that have the small effect [3] resulting into association efforts to be underpowered and difficult to replicate [2]. Hence, often the association signals do not reach the genome-wide significance although they may be contributing to the drug adverse reaction to some extent.

1.2 Lung cancer and chemotherapy

Lung cancer is the most lethal of all the cancer types. According to World Heath Organization (WHO) [10] fact sheet of 2015, lung cancer caused 1.56 million death worldwide in 2012. With overall survival rate of 18%, it was estimated for 26% of the all cancer deaths in 2014 and thus, the leading cause of cancer death in the USA [11].

Chemotherapy with standard platinum agents are frequently administered to patients with advanced lung cancer [12]. Platinum based drugs such as cisplatin, carboplatin and oxaliplatin are widely used. Platinum based agents thwart cellular process forming DNA adducts and lead to apoptosis [13]. The standard chemotherapeutic treatments for lung cancer are based on using platinum based agents with other agents, known as third generation drugs [14]. Microtubule - targeting agents such as paclitaxel, docetaxel, or vinorelbine and DNA-damaging agents such as gemcitabine or irinotecan are paired with the platinium-based agents in chemotherapeutic treatment [12].

Chemotherapeutic drugs are administered to various cancer patients in different regimens and doses based on the somatic mutation profiles and are

aimed at inhibiting cancer cell growth and genomic integrity [15]. However, these chemotherapeutic drugs can induce various adverse reaction mechanism. Adverse reaction such as toxicity not only impacts the quality of life but sometime leads to reduction in dose or even to circumvention of the treatment in extreme of conditions [16].

Drug-toxicity induced myelosuppression is one of side effects caused due to these platinium based chemotherapies [17]. Myelosuppression is the debilitating condition that leads to decreased immunity, oxygen carrying capacity and normal blood clotting activity in individuals. The condition is characterized by suppression in bone marrow activity which leads to decrease in production of platelets, white blood cells and red blood cells [18]. Myelosuppressive effect characterized by decreased production of white blood cells (WBCs) causes leukopenia in cancer patients. Specifically, chemotherapeutic drugs induces neutropenia, condition characterized by decreased count of a specific type of WBCs - neutrophils in blood . Additionally, these drugs can lead to decreased platelets in blood results causing thromobocytopenia resulting in poor blood clotting [17, 18].

1.3 Exome sequencing and association studies

Genome wide association studies are modern powerful tools to understand human genetics. It includes screening of genomewide variants for association to complex traits. These techniques identify common, low penetrant variants at greater statistical power and resolution than conventional linkage studies or candidate gene studies [19]. Last decade witnessed exponential rise in GWAS for many complex traits [20]. However, as these methods are based on SNP tags it can identify risk alleles that are usually in linkage with causal variants [21].

Whole exome sequencing involves the sequencing of all protein coding region of human genome and have been extensively used as discovery tool in identification of genes in Mendelian disorders [22, 23]. With the rapid development and steep decrease in cost of next generation sequencing technologies, whole exome sequencing association studies have been an emerging tool in the study of complex traits genetic architecture [24, 25]. The major advantage of whole exome sequencing from microarray based genotyping method is it provides unbiased variant discovery and direct association with phenotype [21]. An early example of exome sequencing identified variants in *DCTN4* as a modifier in chronic *Pseudomonas aeruginoas* infection in cys-

tic fibrosis [25]. Similarly a schizophrenia study identified a polygenic rare mutations using exome sequencing association study [26]. A pharmacogenomic study associated risk of multiple rare variants in *KCNE1* and *ACN9* for drug-induced long QT interval syndrome [27].

1.4 Statistical association methods for genotype-phenotype correlation

1.4.1 Single variant tests for association studies

A commonly used genetic variation for the association tests are single nucleotide polymorphism (SNPs). These tests involve testing each SNP independently for association to the phenotype [20] . Various statistical methods are developed based on this study design [21]. Single variants statistical tests such as chisquared (χ^2) test, Fischer exact test, Cocharan-Amritage tests and logistic regression are used in the association of variants with the diseases/healthy design in Diabetes, Melanoma, and Alzheimer's disease [28–30]. These methods test enrichment of allele in case and control groups. While, quantitative phenotypes such as blood cholesterol level, body weight and measurements are tested for association to genetic variants using linear regression methods [30–32].

1.4.1.1 Linear regression for quantiative association

Regression methods are based on the dependence between response variables (Y) and the several or single predictor variables (X). In the association study, the outcome variables are either quantitative or binary based on the study design. Linear regression is based on the linear relationship between the quantitative traits and genotypes. These linear regression analysis assume: 1) quantitative traits being normal distributed; 2)genotype groups have same variance and independent from each other [20]. A simple linear regression model with single independent variables such as genotype G and quantitative phenotype Y is given by

$$E[Y] = \beta_0 + \beta_G G \tag{1.1}$$

where β_G is the parameter for the genotype and $E[Y]$ is mean of phenotype. Furthermore, using the variable X, we can add covariates to the above equation, such as age and gender to give the fuller model as

$$E[Y] = \beta_0 + \beta_G G + \beta_X X \tag{1.2}$$

where additional parameters β_X for the covariates accounts for adjustment of the model to the new additional variables in the regression analysis [33].

The null hypothesis for single SNP regression analysis assumes no difference between the quantitative trait means and genotype classes implicating no association between the phenotype and genotype classes while the alternate hypothesis assumes there is no association between genotype and phenotype.

1.4.1.2 Logistic regression for case-control association

The categorical response variables in case-control study design code response variables with binary outcomes. For example: diseased individuals are coded as 1 while healthy ones are coded 0 . For these binary outcomes, the linear regression is modeled using transformation of outcome with logistic function (logit). Logit function predicts probability of diseased group in a given genotype classes. The logistic transformation (logit) is given as $\log \frac{p}{1-p}$ where p is the conditional probability of the discrete variables given the whole data, $Pr(Y = 1 \mid X = x)$. This value of the logit is equated with the genotype groups of the individuals [34]. The simple logistic regression for the association studies is given as:

$$\log \frac{p}{1-p} = \beta_0 + \beta_G G \tag{1.3}$$

The simple logistic models assess the relationship between the dichotomous dependent variable (Y) and predictor genotype variables (G). Logisitic regression are extensively used in association studies as the model is flexible enough to incorporate other interesting clinical variables [34].

1.4.2 Gene based association tests

Sequencing based association methods provide an unprecedented opportunity to interrogate and understand rare and common variants implication in a complex traits. Traditional single-variant studies are underpowered to detect rare variants as few of the individuals in study group have rare variants. Hence, genes/regions are defined where single variants are aggregated together. The rationale is applying these approaches would enrich rare variants in a region and decrease the number of tests for multiple corrections [35, 36]. Broadly classifying, gene/region based tests are classified as burden and non-burden methods.

Burden tests are based on collapsing or aggregating the variants in a single genomic regions and associating these regions with the phenotype of interest [21]. In a recent study, burden tests identified polygenic rare mutations

in schizophrenia [26]. Gene-based association test such as Cohort Alleleic Sum test (CAST) [37], Combined multivariate and collapsing (CMC) [38], Weighted sum test (WST) [39] are developed based on the burden test principle. These methods are based on evaluating enrichment of rare mutations between cases and controls and assume rare variants influence phenotype in same direction and with equal magnitude. However, most variants sequenced in a gene/region could have either no effect on phenotype with only few influencing the phenotype. Thus, collapsing all variants into genes will have spurious association and loss in statistical power [36, 40].

Another category of tests are non-burden tests which are independent of magnitude and directionality of the variant effect. Statistical tests such as C-alpha [40] and Sequence Kernel Association test (SKAT) [36] is a non-burden tests developed for gene based association studies. SKAT are kernel machine regression method that aggregates variants information through the kernel function and uses variance component test for variant association in a gene/region. For each regions/genes SKAT calculates p value for association while adjusting for covariates. However, the tests suffer when large number of variants are causal in same direction. Lee *et al.* 2012 [41] developed the generalized form of SKAT- SKAT Optimal, a data adoptive methods which includes linear combination of burden and SKAT, and based on the parameter provided to identify the optimal test which maximizes the power of the study .

1.5 Structure of thesis

The thesis consists of the analysis of the exome sequence of 216 cancer patient cohort treated with combination of chemotherapy drugs: carboplatin and gemcitabine. In this retrospective study the lung cancer patients with myelosuppression toxicity measured as individual nadir values of thrombocytopenia (TPK), neutropenia (NPK), and leukopenia (LPK) for individual patient, are associated with the exome variants of the patient. The thesis addresses two major goals: firstly, we describe the bioinformatic analysis and quality control pipeline of the exome sequencing and secondly, the study of the association test in quantitative and qualitative case-control study designs.

The work-flow of the overall thesis project is described below in the Figure 1.1. In the following chapter, I describe methodologies and rationales from the study. In chapter 4 and 5, I discuss the results from whole study and conclude with the future directions in chapter 6.

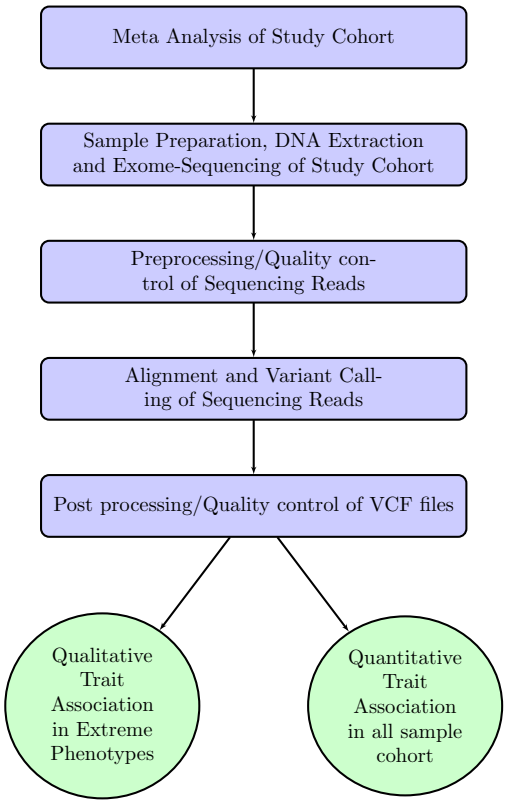

Figure 1.1: Overall flow-chart of project

Chapter 2

Materials and Methods

2.1 Study cohort

A total of 216 patient diagnosed with Non-small cell lung cancer (NSCLC) were included in the study. All patients were scheduled to be treated with carboplatin and gemcitabine for four cycles and received at least one cycle of carboplatin and/or gemcitabine chemotherapy. After the chemotherapy cycle, the nadire values for leukocytes, neutrophils, platelets and haemoglobin are monitored. Based on the observed nadir values, the patient cohort was graded as 0, 1, 2, 3 or 4 based on the Common Toxicity Criteria (CTC) [42] grade set up by National Cancer Institute (NCI).

2.2 Whole exome sequencing of the patient cohort

In the current project, DNA was extracted from the patient blood samples and libraries were prepared using Nextera® Rapid capture exomes kit. The sequencing of the individual samples was performed on Illumina®Hiseq2500 platform to generate read lengths of 2×150 base pairs. The exome sequencing was done at the National Genomics Infrastructure (NGI) platform at Science For Life Laboratory, Solna, Stockholm.

2.3 Preprocessing of raw sequencing reads

The exome sequencing FASTQ files are provided from the NGI platform. The format includes sequence reads and the quality score associated with the each nucleotide in the sequence [43]. Some quality control measures were applied

before mapping reads with reference genome. Quality measure and adapter removal was performed with a utility program, Trim Galore[44]. TrimGalore is programming script to trim adapter sequences and low quality ends using Cutadapt [45]. The program keeps reads with quality threshold of 25 on Phred scale and discards read pairs from analysis if either read in a pair is shorter than 25 bp.

2.4 Mapping and variant calling of sequencing reads

After trimming, the reads are aligned to the reference genome and then the mapped reads are processed through some further quality control before GATK pipeline tools are applied to call variants. Details of these steps are described below and summarized in the flowchart in Figure 2.1 below.

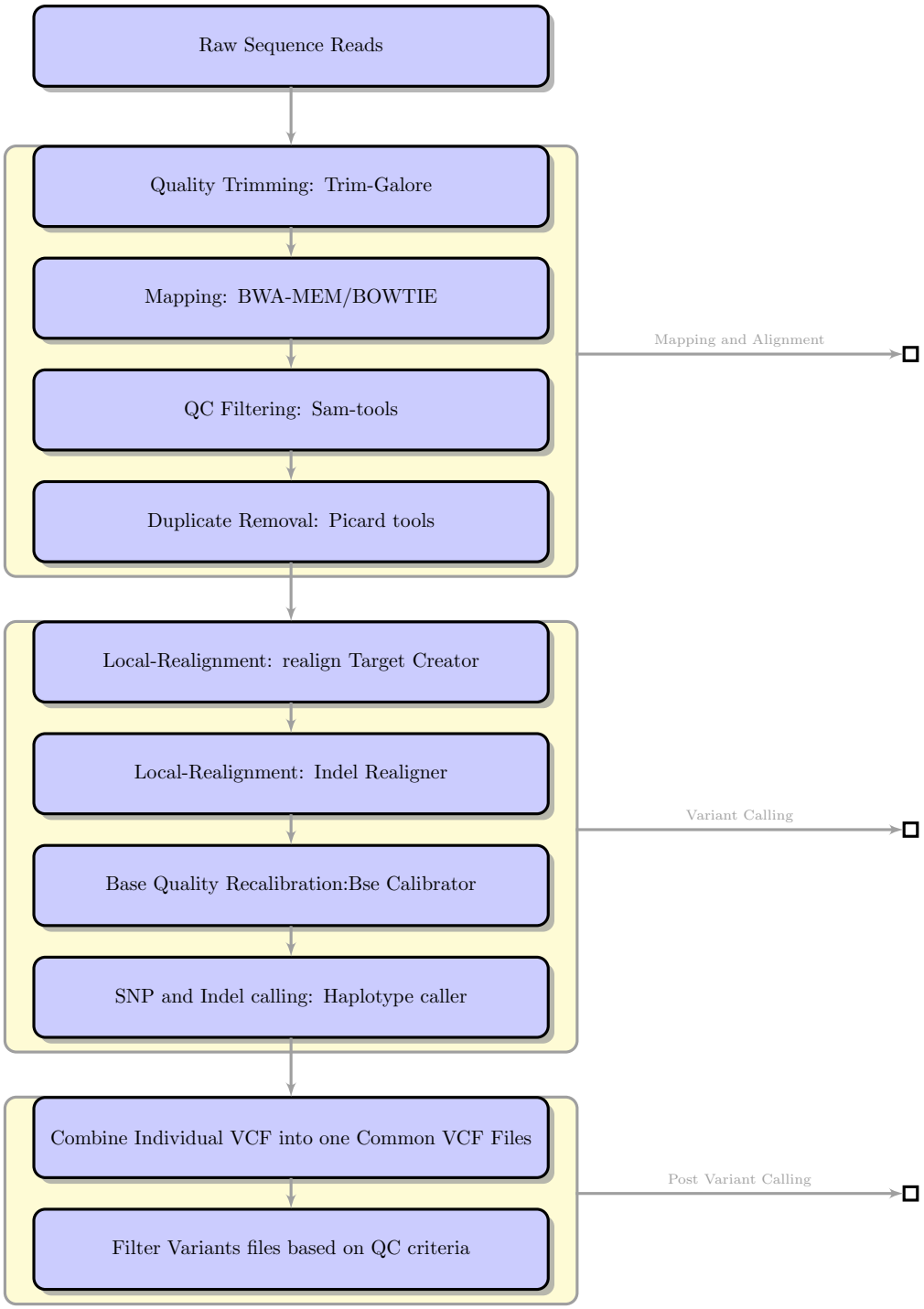

Figure 2.1: Overall Flow-chart of GATK pipeline and toolkits used

The trimmed reads are aligned to GRCh37 /hg19 human reference genome (UCSC Genome Browser) with Burrows - Wheeler Aligner (bwa/0.5.9) [46] software package. The *bwa mem* command is used to align the sequence with the read length greater than 100 base pair. The resulting sequence alignment/map (*sam*) files were converted to *bam* files using SAMtools [46]. The SAMtool consists of the *C* implementations for the manipulation of *sam* and *bam* files [47]. Obtained *bam* files were processed with $MarkDuplicates$ command in picard `(http://broadinstitute.github.io/picard/)` software tool to mark the duplicate reads from the mapped reads. Picard consists of the java toolsets for next-generation sequencing data manipulation.

Post processing of these aligned reads was done using GATK (v3.3-0) where the reads were processed to command-line tools $IndelRealigner$ and Base Quality Score Recalibration$BQSR$. The $IndelRaligner$ tools perform local alignment of aligned reads around indels. The main objective is reduce the mismatches bases by locally realigning the aligned reads at indel positions. Base quality recaliberator recalibrates the base quality scores of aligned reads such that a quality score generated are closer to the actual probability of mismatching in the reference genome.

Variant calling in aligned reads is done using $HaplotypeCaller$ package from GATK and performed on the targeted regions in an individual sample. Variant calls were initially made on the individual samples and written in raw gvcf file. $HaplotypeCaller$ estimate the probability that a given site is variant or non variant given the likelihood of the haplotype generated from the read data. The individual gvcfs are collectively collected and formatted to generate the multi-sample Variant call Format (VCF) file. The raw variants in the VCF files were flagged if the quality scores < 50,FisherSB filter > 60, quality by depth < 5. Thus, obtained VCF file is termed as $rawVCF$ (called henceforth) file which contain variant informations of all sample cohort. Finally, all the variants in VCF file was annotated using SNPEff [48] which annotates and predicts the effects of variants on genes (such as amino acid changes).

2.5 Quality control of $rawVCF$

The $rawVCF$ file consists of 211691 variants across the 216 samples. The variants were filtered using vcftools [49] *–remove- filtered all* and *–min-meanDP* commands based on quality measures of the variants outputted by GATK pipeline. The criteria for filtration was that the variants should passed the GATK filter and have a mean sequencing depth ≥ 10. A total of 156049 variants passed the filter and were stored in a file called *Filter VCF*

(henceforth). The quality control metrics such as reads counts mapped to hg19, reads mapped to the target, transition/transversion (Ti/Tv) ratio was evaluated using plinkseq [50] tool *i-stats* command. The summary statistics of variants are reported in Appendix A.1. An average of 29834 variants were identified in whole sample at the genotype rate of 0.99. The Ti/Tv ratio was on average 2.180 per individual. The genotyping rate of 0.95 was considered as the threshold for the accurate genotype of the individuals as shown in the Figure 2.2. The genotyping rate provides information regarding quality and quantity of variants called in the sample.

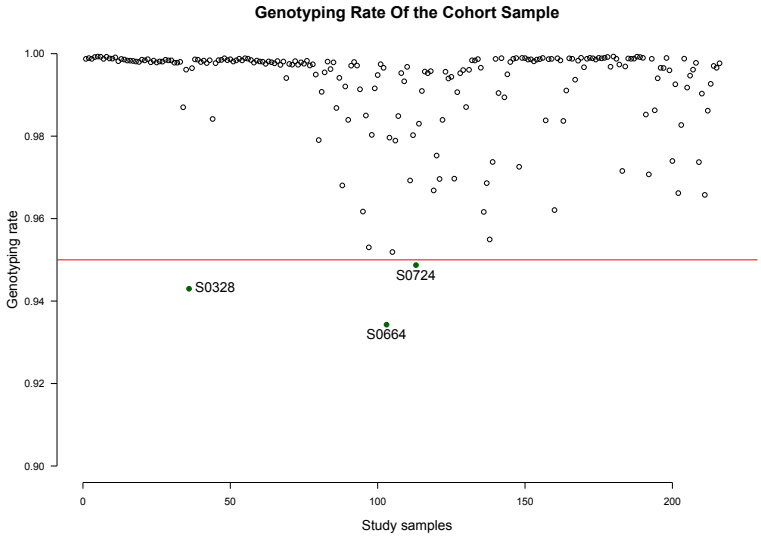

Figure 2.2: Genotyping rate for all sample. Three of the samples: S0724, S0328, S0664 have a genotyping rate lower than 0.95, threshold genotyping rate in study.

The mean alternate allele counts in the study cohort was found to be 29832 and further to find outlier sample number of read length was plotted against the alternate count as shown in Figure 2.3.

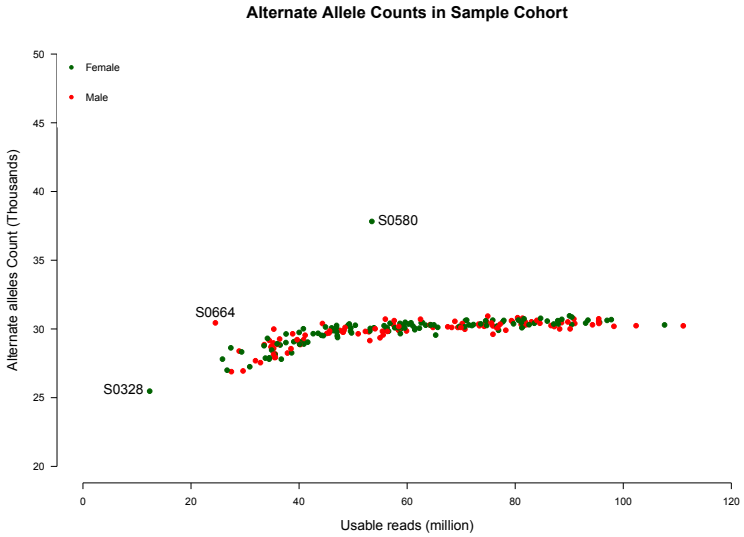

Figure 2.3: Alternate allele count in the cohort in both genders. Three samples S0580, S0664, S0328 are considered as an outliers based on the alternate allele counts.

2.6 IBS clustering in Cohorts

Identity by state is a method to measure the similarity between unrelated patients. Based on the genotype called on the filtered variant files, we carried out identity based on state (IBS) clustering of the samples. IBS clustering was performed using *–cluster* command in plink [30]. The clustering of the whole sample is seen in the Figure 2.4.

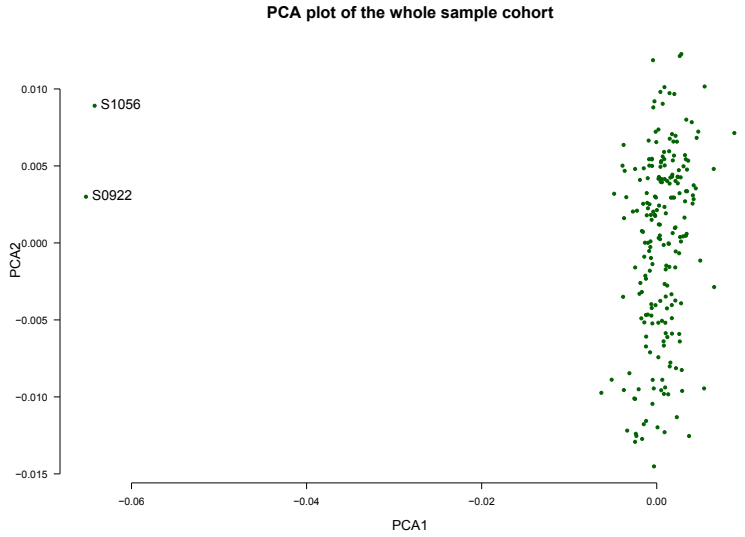

Figure 2.4: Identity by similarity clustering in whole sample in $Filter1VCF$. Two samples S0922 and S0156 are out-clustered from the rest of the samples.

All the samples apart from two samples S1056 and S0922 are clustered together which defines the samples are homogeneous with the same ancestral descent. We confirmed the similar metadata information for the two of the out-clustered samples and considered it could possibly be sequenced twice. On further inquiry we found that the two samples were from same patient and decided to drop sample S0922 from further analysis. Furthermore, we performed pairwise IBD analysis in the sample cohort. Pairwise Clustering measures the relatedness between the individuals by calculating the estimates of getting too similar samples by random chance. The command *–genome* was run in plink files of the $Filter1VCF$.

The pair-wise IBS clustering results samples S0580 and S0664 are related to all the other samples in the cohort implying these samples could be contaminated during sample preparation. Furthermore, these samples were flagged as outliers in alternate count allele analysis. S0328 was also flagged as being an outliers in the alternate allele count analysis. Thus, we removed S0580, S0664, S0328, and S0922 from down-streaming analysis based on these result. Hence a total of 212 exome samples were used for the rest of the analysis.

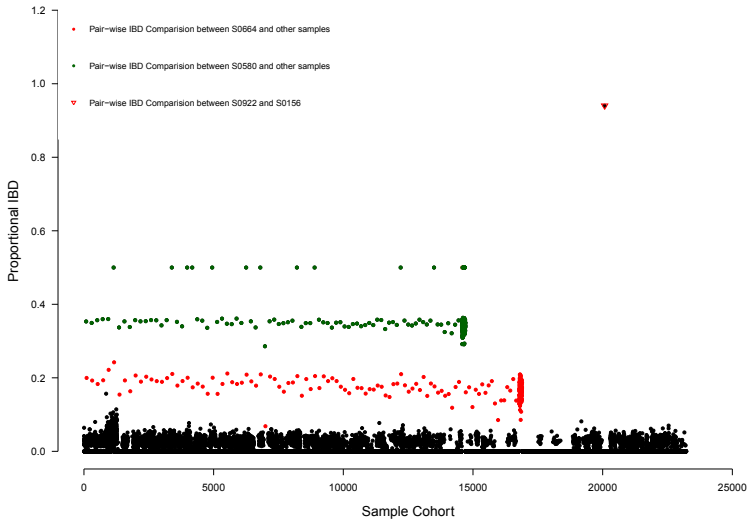

Figure 2.5: Pair-wise IBS clustering in whole sample in $Filter1VCF$ file. The red dots indicate the pair-wise comparison between the S0664 and rest of samples while the green dots indicate the pair-wise comparision between S0580. Similarly, the outliers show the pairwise comparision between S0922 and S0156.

2.7 Filtration of $Filter1VCF$

The vcftool command *-exclude* was used to remove outliers S0580, S0664, S0328, and S0922 the from $Filter1VCF$ file in the study cohort. Furthermore, we only kept the variants with the genotyping rate greater than 95% in 212 sample cohort. The total variant count after filtration in the whole sample was 152042. We called the file as *Common_rare* VCF. Based on the minor allele frequency (MAF) of 0.01 in the sample, the *Common_rare* VCF file was further separated in *Common* and *Rare* VCF files. The $CommonVariant$ files consist of variants with the MAF > 0.01 in study cohort while rare variants MAF < 0.01. The final variant count in the Common and Rare variants were 74281 and 77761 respectively.

2.8 Quantitative Association Tests

In the present study, the measures of nadir values of myleosupression phenotype-leukopenia, neutropenia, and thrombocytopenia were defined as quantitative

trait of interest. The initial blood concentration was measured before and after administration of chemotherapy treatment reported as baseline and nadir count respectively. Specifically, rank (ENQT) and logarithm (LN) normalization of the individual nadir values for the each phenotypes were considered as the quantitative traits. These quantitative traits represent the effect of the chemotherapy treatment on the patient's adverse drug reaction (ADRs). The missing values in the phenotypes were coded as -9 for the association test to make the file compatible with plink phenotype file format. We analyzed the QC filtered variants with the phenotype in single marker (SNV) and gene based association tests for both transformed data.

2.8.1 Single Variant Association test for Quantitative Traits

A total of 72855 biallelic variants in *Common* VCF were used in the association studies. We used the default additive genotype model in linear regression for all common variants. The linear regression was performed using $--linear$ command in plink [30].

Two types of multiple correction: strict Bonferroni adjusted p-value $< 6.75 \times 10^{-7}$ and less conservative FDR-BH of p-value < 0.05 were considered for the variant to be statistically significant in the association test. The bonferroni correction adjusts threshold p-value from 0.05 to new corrected threshold p-value of $0.05/k$ (k =number of tests, here 72855) while FDR estimates the proportion of the significant results that are false positive. [20, 51]

Since our sample cohort is small, the statistical significant p-values for the variants were unable to be obtained. Thus, we applied an alternate strategy: we divided the variants into high and low toxicity based on the β values obtained from the regression association. The positive β values were regarded as the variants that were associated with low toxicity while the negative β values variants as the high-toxicity associated variants. This approach was applied to only variants with p $< 10^{-3}$ in the linear association studies.

SNPs associated with high toxicity were mapped and annotated using the bioinformatic tool SNPnexus [52] and associated pathway are analyzed using LifeMap [53].

2.8.2 Gene/Region based Association test

For the gene/region based association, we investigated the effect of both rare and common variants of the genes to the individual toxicity phenotype. Thus,

we used the *CommonRare* VCF file - the bi-allelic variants for gene based association tests. The rationale for using both common and rare variants in association test is we consider an adverse drug reaction as the complex trait with equal contribution of both rare and common variants in the predisposition of the toxicity phenotype. Additionally, aggregating variants into gene/region provides us with an opportunity to study the effect of rare and commmon variants in toxicity. We set the parameters for equal contribution of rare and common variants in SKATO analysis in SKAT [54] package in R[55].

2.8.3 Gene/Region definition for association studies

For the Gene based association study, we initially mapped the variants in the *CommonRare* VCF to the corresponding standard Refseq [56] genes using the $--assoc$ command in plink with the baseline TPK values. The SNV generated in *.assoc* files was then mapped to the corresponding genes, exon and exons ± six base-pairs regions of the genome using the gene reporting tool in plink. The $--gene-report$ command in plink mapped the SNV to corresponding region and generated a gene region file.

We investigated three definition of the gene/regions - gene only, exon only and exons ± six in the association with the quantitative phenotypes. First we considered the standard gene segment as defined in Refseq and USSC Genome Browser [57] databases. We downloaded the corresponding genomic coordinates for genes and assigned the longest transcript as standard genes. We found 155239 variants in *CommonRare* VCF mapped to the standard 19142 genes in Refseq. In order to reduce the redundancy of tests in variants associated with Single variant test, we only considered using the genes with the number of variants greater than one. So that we can identify the combined effect of rare and common variant in the phenotype. That lead to the 153095 variants in the study cohort.

However, we found that there were 11275 variants that were present within same genomic coordinates. This could be due to orientation of the gene; as different genes could have same genomic position but could differently oriented (sense and antisense) or in some condition fusion genes were also seen. In order to reduce these discrepancies, we further segmented individual genes into different regions such as exons ± six base-pairs and exon only.

This definition of gene/region into exons minimized the number of repetitive variants found within same genomic coordinates but in different genes. However considering only exome as region definition lead to exclusion of intronic variants. Hence to incorporated the splice site variants in further

down-streaming gene/region based association tests we further defined region into exons ± six base-pairs. As seen in the Table 2.1 exons ± six base-pairs definition of gene region leads to 109588 variants identified in the 15334 genes in Refseq.

Region Defination	**Variant Count**				
	Total Variants Count	Genes with Multiple Variants	Unique Variants	Total Genes Before processing	Total Genes After processing
Genes from Refseq	155239	153095	141820	19142	16261
Exon ±6 basepair	114047	111300	109588	18220	15334
Exon only	111144	108373	106704	18117	15178

Table 2.1: Variants count in defination of exome. Total count refers to the number of variants before processing of the genes. Genes with multiple variants refers to the genes where variant count > 1 . Unique variants refers to the variants non redundant variants in the genes.

These exon± six basepair regions were then used in SKATO test in SKAT R package [54] using both LN transformed and ENQT for each of the toxicity phenotypes. Similar to the single variant association test we considered two types of multiple corrected p-values for the a gene to be statistically significant. We set the threshold of strict Bonferroni correction of p-value $< 3.2 \times 10^{-6}$ and more flexible FDR-BH threshold at < 0.05.

2.9 Case/Control Based Association Studies in Extreme Phenotypes

Qualitative study design includes taking into consideration the binary phenotypes such as diseased/non-diseased or high/low toxicity group. In the current study, we were provided with the Common Toxicity Criteria (CTC) [42] score of individual patient phenotype. We used CTC score to classify the patients to either high or low toxicity group. The rationale for the extreme phenotype study is finding the variants enriched within these individual groups.

2.9.1 Definition of extreme High Toxicity cases and Low Toxicity control group from the study Cohort

For each of the individual myelosuppression toxicity phenotypes we classified patients as high toxicity (cases) with the CTC score of either 3 or 4 and as low toxicity (controls) groups of CTC score 0 or 1. Number of patients in each group for each phenotype is shown in the Table 2.2. The number of patients in different CTC group for TPK, LPK and NPK phenotypes as shown in Figure 2.6

Phenotype	**High Toxicity(Cases)**	**Low Toxicity (Control)**
Thrombopenia (TPK)	75	93
Leucopenia (LPK)	49	91
Neutropenia (NPK)	97	76

Table 2.2: Number of patients in each phenotype. The different toxicity phenotypes of the individual patients are given in the first column and the corresponding cases of High toxicity group with CTC score of either 0 or 1 and control group of Low toxicity with CTC score of either 3 or 4 are tabulated in each successive columns.

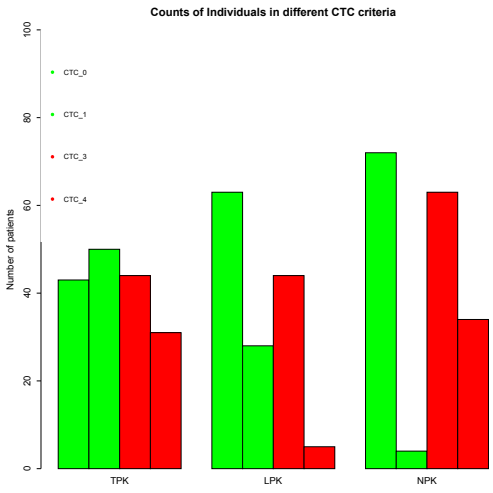

Figure 2.6: The color red indicates the patient in high toxicity cases while green depicts patients in low toxicity control in each phenotype. As seen in the figure there are 5 patients with CTC score of 4 for LPK while there are 4 patients with CTC score of 1 in NPK phenotypes.

2.9.2 Single Variant Association test for Qualitative phenotypes

The Single variant association study was performed on bi-allelic variants with MAF > 0.01 in *Common* VCF file for each phenotype case and control group. We performed logistic association for the all common variants. The phenotype information was provided for the individual phenotype in fam files where the cases were coded with 2 and the control group as 1. The logistic regression was performed using $--logistic$ command in plink [30]. The bialleic single nucleotide polymoprhisms (SNP) with multiple corrected Bonferroni p-value $< 6.8 \times 10^{-7}$ or less lenient BH-FDR p-value < 0.05 after multiple correction was considered as statistically significant.

As in the quantitative analysis, the statistical significant p-values for the variant were unable to obtained due to small sample size. Thus, we applied alternate strategy: we took the variants with p $< 1.0 \times 10^{-3}$. The variants were then annotated to respective genes using SNPnexus tools. Those genes were analyzed in LifeMap GeneAnalytics tools. The rationale for the alternate analysis is the potential variants for the phenotype are enriched in each sample cohort as we have divided it into cases and control.

Chapter 3

Results

3.1 Summary Statistics of the study cohort clinical data

The study cohort consists of non small cell lung carcinoma (NSCLC) patients treated with carboplatin and gemcitabine chemothereupaetic drugs. The phenotypic characters of the lung cancer patients are shown in Table 3.1 below.

The study cohort consists of 115 female and 101 male NSCLC patients. 61% of patients have been diagnosed with adenocarcinoma lung cancer histological subtype. And 90% of patients have a smoking history of either being a current smoker (44%) or former smoker (46%). Most of the patients (70%) in the study cohort are diagnosed with lung cancer in advanced stages - IIIa/IIb/IV.

142 patients in advanced stages of lung cancer in study cohort were provided with combination of carboplatin and gemcitabine while 74 of them are treated with adjuvant treatment. Adjuvant treatment mode are additional chemotherapy drugs, carboplatin and gemcitabine given after the surgery to reduce the cancer risk.

Clinical Features		Patients
Gender of the Samples		
	Female	115
	Male	101
Age of treatment		
	Overall	64.5 (60-71)*
	Female	64 (59.50-70.00)*
	Male	67 (61-72)*
Histological Subtype		
	Adenocarcinoma (AC)	133
	Squamous Cell Carcinoma (SCC)	41
	Large Cell Carcinoma (LCC)	10
	Uncharacterized Non-Small Cell Lung Cancer	31
Smoking History		
	Current	95
	Former	100
	Never	21
Pathological Stages		
	Stage Ia/Ib	40
	Stage IIa/IIb	29
	Stage IIIa/IIIb	64
	Stage IV	81
Treatment Type		
	Advanced disease	142
	Adjuvant treatment	74

Table 3.1: Clinical Features of Study Cohort. The figures in the right indicate the number of patients with the clinical features. The median age of the patients with the inter-quantile range in the brackets

Additionally, the cancer histology and pathological condition were evaluated in male and female cohorts independently. We found that in both the group of cohort, adenocarcninoma is the most abundant histological condition with patients depicting the pathological stage-IV lung cancer. In summary, we found that our study cohort to be homogeneous in both gender groups based on pathological and cancer histology phenotype as shown in the Figure 3.1.

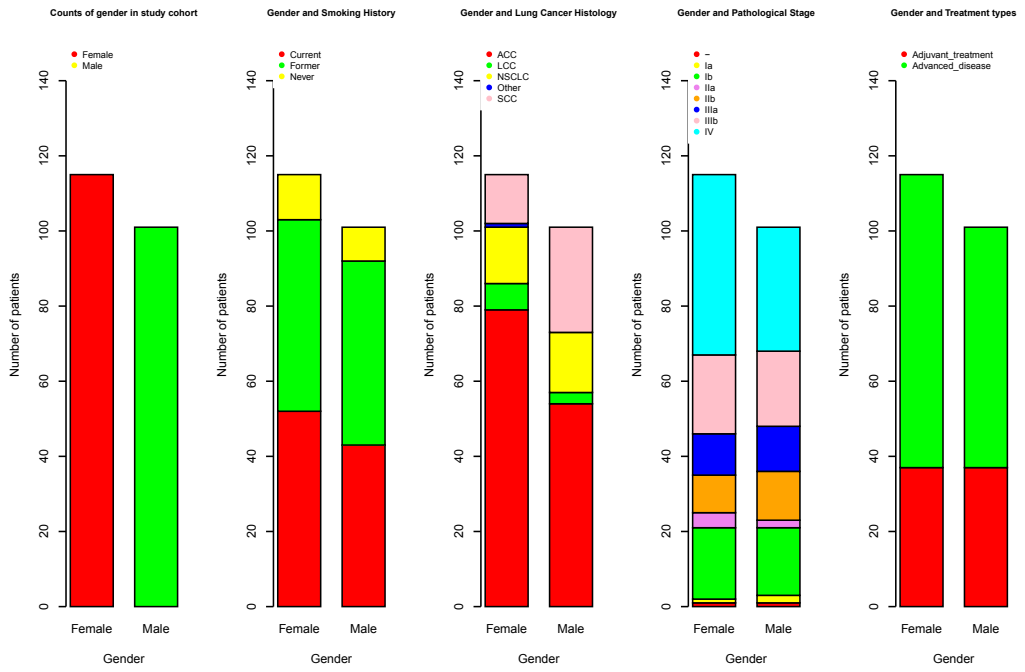

Figure 3.1: The gender and lung cancer phenotype of the study cohorts. The barplot depicts the number of the female patients are higher than the male patients with the most of the patients are treated for advanced treatment. However the proportion of the patents with the different stages of the lung cancer are similar in both the sexes in the study cohort.

Similarly, we investigated the lung cancer histological subtypes and pathological stages in the three groups of smoking history. We found that more than half the patients had adenocarcinoma histology in all the smokers including the current and former smoker and even in never smoking groups. Also, we found that half of the smokers and non-smokers had the advanced pathological stages of lung cancer as shown in Figure 3.2.

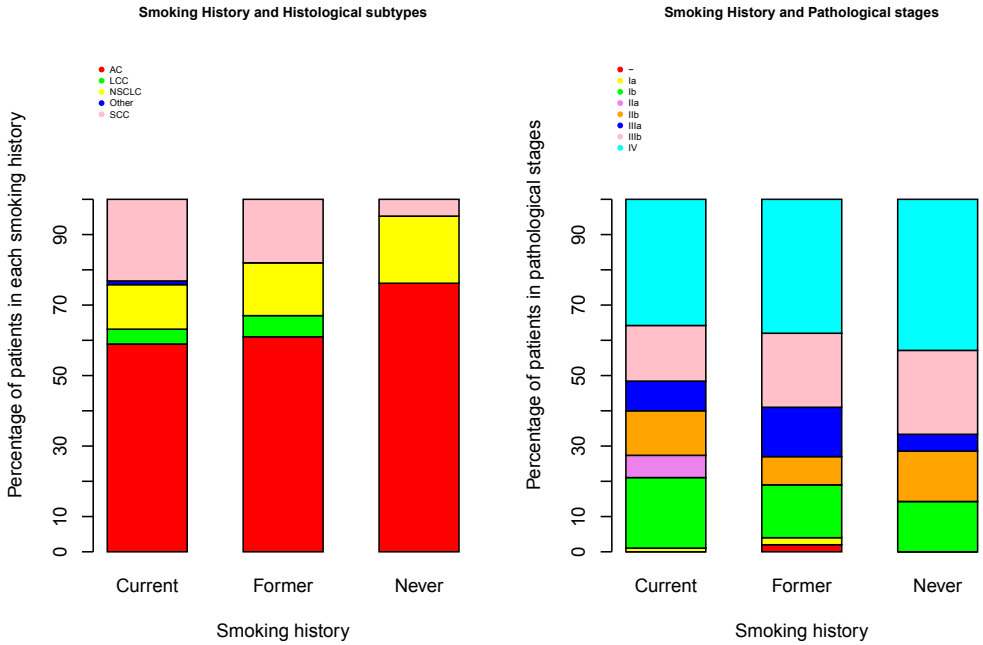

Figure 3.2: Smoking history and the lung cancer phenotype of the study cohorts in each of the categories. More than 50% of the current and former smokers were diagnosed with adenocarcinoma lung cancer which is half of the total study cohort. The adenocarcinoma lung cancer are present in 76.19% of the non-smokers

3.2 Transformation of the Nadir TPK, LPK and NPK

The change in the initial baseline and nadir values of the individual patient platelets, leukocytes and neutrophils counts are measured and provided as baseline information. The myelosuppression toxicity phenotype are provided in the nadir values of the individual measured after the adminstration of the drugs. The Figure 3.3 below shows the change in the individual myleosuppression toxicity in all the CTC groups in each phenotype.

The nadir values represent the lowest blood count after chemotherapy [58]. From Figure 3.3, we can say that nadir values of each phenotype is skewed with the presence of extreme of outliers. These outliers represent patients who are unaffected by the administration of the drugs. Therefore, we used normalization techniques such as logarithm (LN) and Empricial Normal

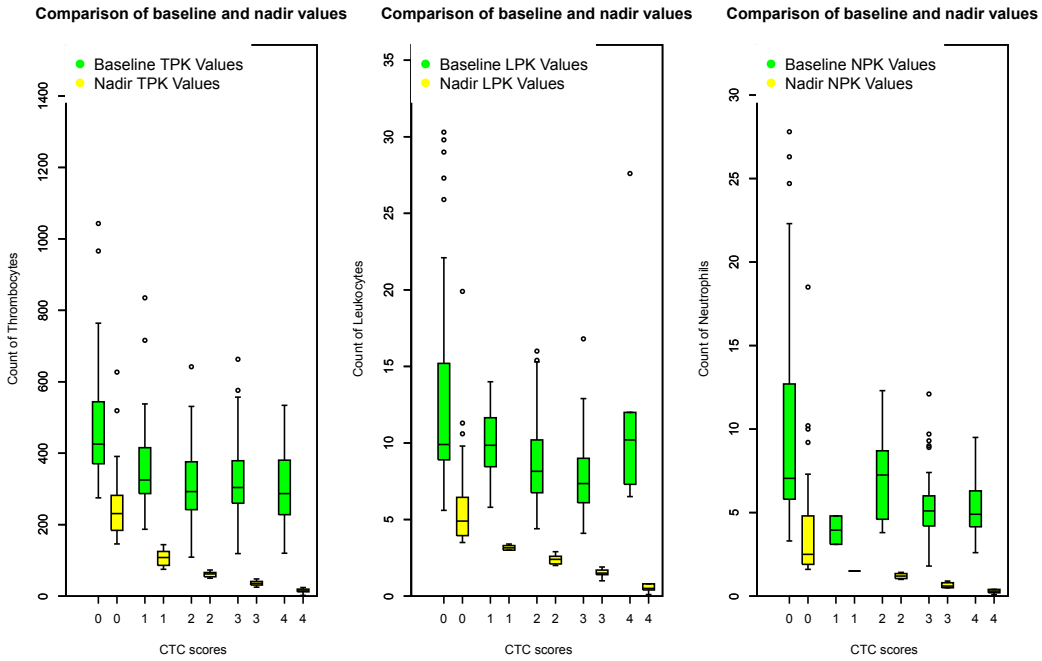

Figure 3.3: The figure illustrates the decrease in the count of the thrombocytes, leukocytes and neutrophil after the patients administered with combination of gemcitabine and carboplatin. Additionally, boxplot illustrates the distribution of the baseline and nadir values of each phenotype.

Quantile transformation (ENQT), a rank-based transformation to transform the data. The normalcy of the data is tested with Shapiro test in R[59]. The Table 3.2 and Figure 3.4 below show the result of the normality test for untransformed, logarithm transformed and ENQT transformed data.

Phenotype	**Shapiro-Wilk Score**	**p-value**
Nadir TPK	0.7944	4.103×10^{-16}
Log-TPK	0.9909	0.1937 *
ENQT-TPK	0.9987	0.9999 *
Nadir LPK	0.7582	2.2×10^{-16}
Log-LPK	0.9498	$7.736\ \times10^{-07}$
ENQT-LPK	0.9978	0.9926 *
Nadir NPK	0.6447	2.2×10^{-16}
Log-NPK	0.9886	0.1141 *
ENQT-NPK	0.9948	0.7268 *

Table 3.2: Shapiro-Wilk Test for different phenotype data. The null hypothesis of the test assumes the data to be normality distributed at p-value > 0.01. The values * indicates the data is normally distributed.

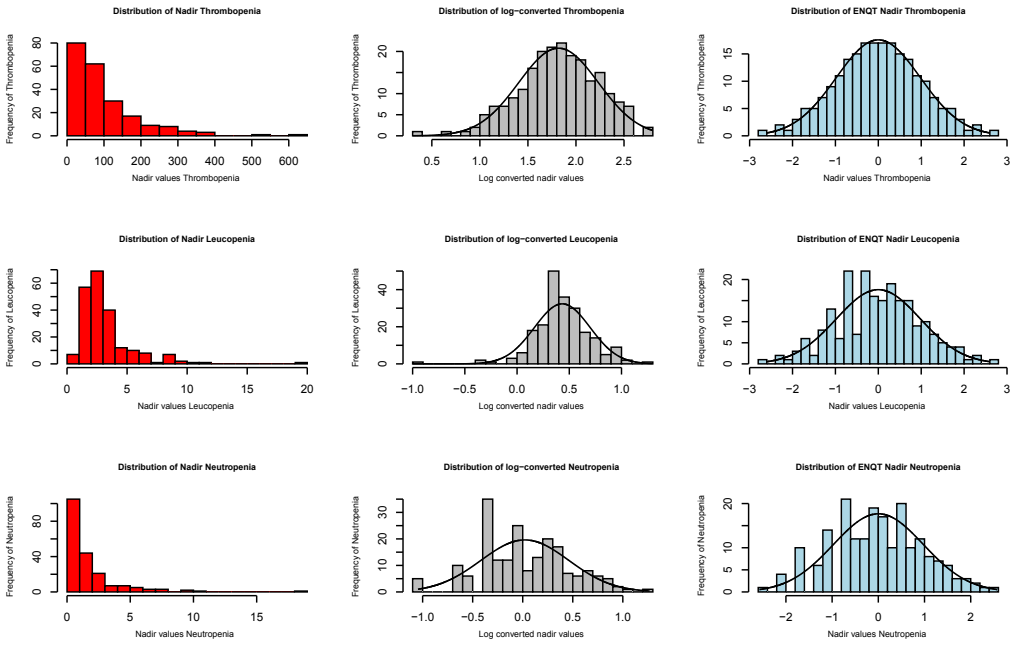

Figure 3.4: The figure illustrates histogram plot of different transformed thrombocytes, leukocytes and neutrophils Nadir values.

3.3 Read Counts in Mapping and Alignment of the sequenced reads

We calculated the read counts in each step of the mapping steps of the sequencing reads to the reference genome. Figure 3.5 shows the counts of reads in each step of exome sequencing.

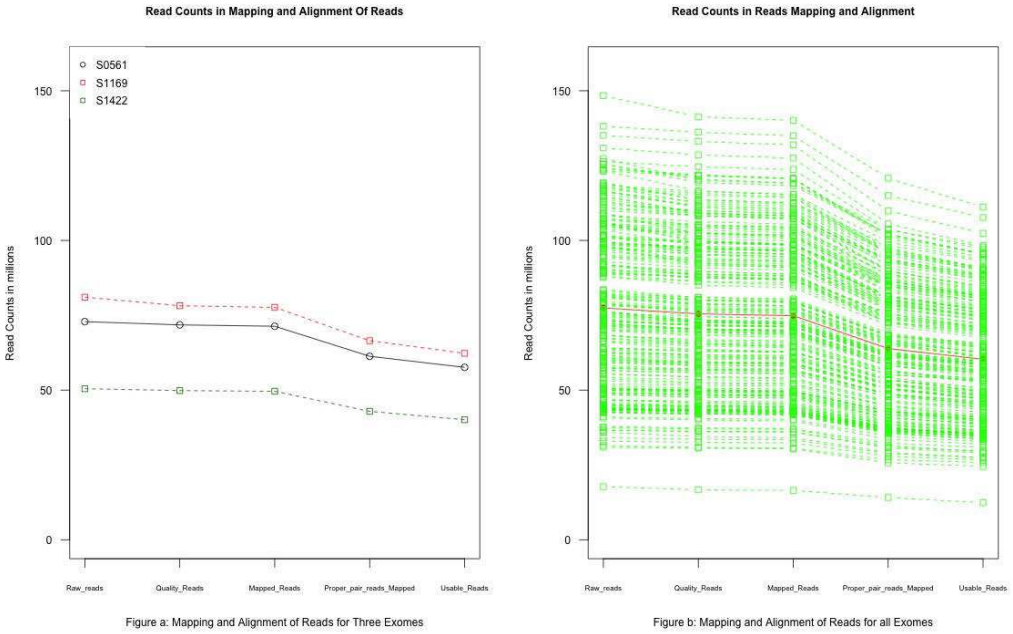

Figure 3.5: Read Counts in each quality-control step of the exome alignment and sequencing. Figure in right(a) shows the read counts in the three samples while figure (b) shows the read counts in the overall cohort. The mean read counts of the whole sample cohort is denoted as the red line in the figure (b)

From above Figure 3.5, it can be seen that during each processes of mapping and alignment pipeline of the sample, there is substantial decrease in the read count. In average we see 22.12% decrease in the read counts from raw read counts to the usable reads for variant calling. Similarly another index for measuring the efficiency of the sequencing reactions are the coverage at each base of the reads. The coverage at each for the final mapped reads were extracted from using *HSMetric* toolkit. The Table 3.3 below shows the target coverage of the three sample exomes.

We compared the variant called in three random exomes that were mapped

Sample	Target Coverage at 2X	Target Coverage at 10X	Target Coverage at 20X	Target Coverage at 30X	Target Coverage at 40X	Target Coverage at 50X	Target Coverage at 100X
S0561	95.88	91.31	85.94	79.97	73.42	66.49	34.45
S1169	95.32	89.47	83.03	76.41	69.52	62.50	31.66
S1422	93.23	82.63	71.93	62.48	54.01	46.41	19.61

Table 3.3: Base Target Coverage for three samples.

with two different alignment methods: Bowtie and BWA MEM. Variants called on three exomes S0561, S1169, and S1442 from both alignment methods were PCA plotted as shoen in the Figure 3.6. There is a complete match between the variants called by the two alignment methods.

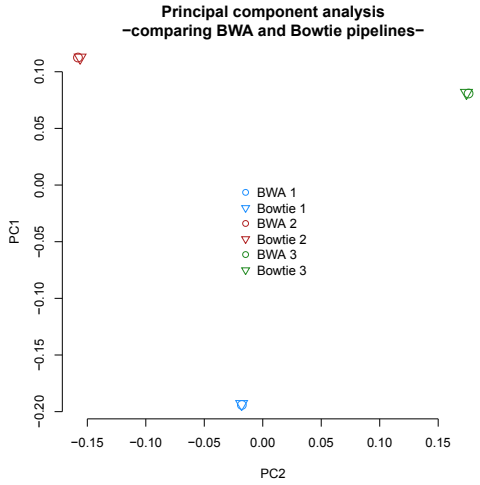

Figure 3.6: PCA plot of the exomes of S0561, S1169, S1442 mapped with *Bowtie* and *BWA-MEM* methods. The figures depicts both alignment methods identified near identical variants in these three samples. Thus, it illustrates that the we can further continue the analysis using the Bowtie/BWA-MEM methods.

3.4 Quantitative trait single variant association test

3.4.1 Thrombocytopenia (TPK)

The Q-Q plot 3.7 shows probability distribution of p-values from empirically distributed p-values of single marker association test for common variants in log transformed and ENQT nadir thrombocytopenia values. Both transformation methods result in variants with the p-value in concordance with expected p-values. However, none of biallelic single variant reached the statistical significance Bonferroni corrected p-value of 6.75×10^{-6} or FDR corrected p-value of < 0.05.

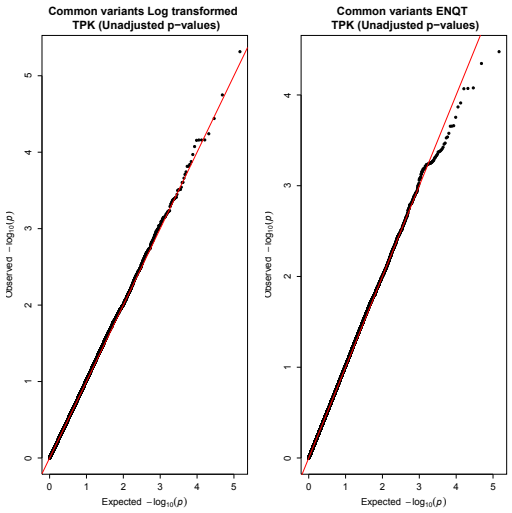

Figure 3.7: Q-Q plots of the LN and ENQT TPK

Manhatton plots, Figure 3.8 and Figure 3.9 visualizes chromosome position of biallelic variants. Lowest p-valued variant rs149407483 with 4.84×10^{-6} mapped to chromosome 19 for LN TPK while rs145707160 and rs4440539 with p-value 3.336×10^{-5} and 4.450×10^{-5} mapped in chromosome X and 7 for ENQT TPK respectively.

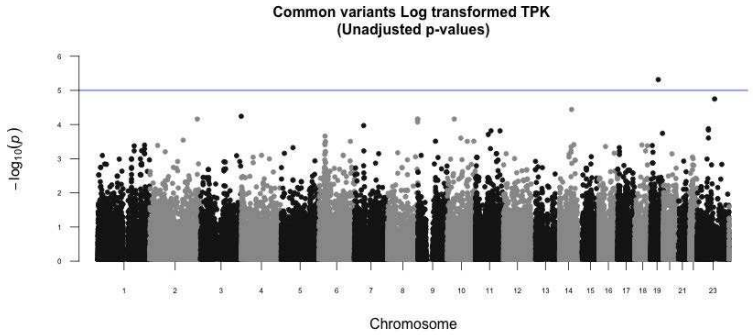

Figure 3.8: Manhatton Plot: LN-TPK

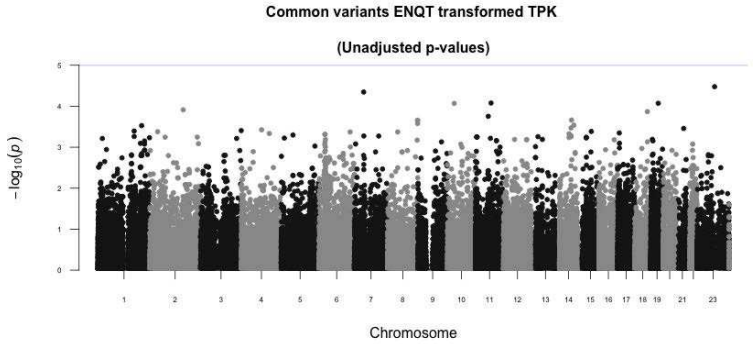

Figure 3.9: Manhatton Plot: ENQT-TPK

Both the transformation methods identified 82 and 79 variants associated with TPK at p-value $< 10^{-3}$. As both methods were introduced to normalize nadir TPK values, we concluded that using intersection of the transformation methods we were able to identify at least 58 variants associated with the TPK phenotypes. This is depicted in Venn diagram 3.10. In the alternate analysis of variants with p $< 10^{-3}$ and negative β values, we identified 46 variants associated high toxicity in LN TPK phenotype and 31 variants in ENQT TPK. 25 of the variants were common in the both the method as shown in the Venn diagram 3.11.

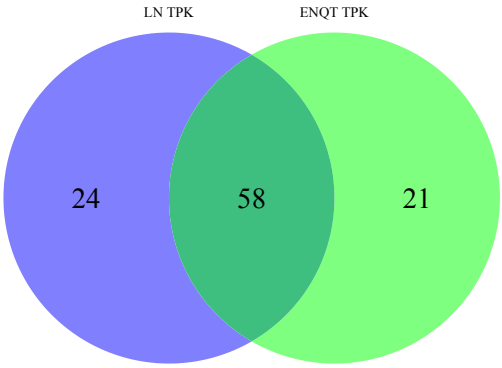

Figure 3.10: SNP comparision between the two transformed phenotypes at $p < 10^{-3}$. A total of 103 SNPs were identified to be associated in both the transformation methods at $p < 10^{-3}$ of which 58 variants are identified in both the transformations.

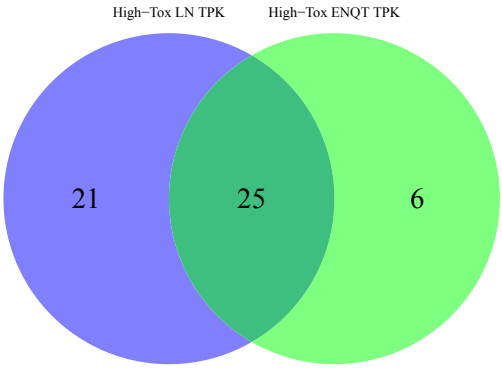

Figure 3.11: Venn diagram depicting the common variants associated with the high toxicity group (negative β values) in the regression analysis. 25 common variants were associated in both the phenotypes for high TPK in cohort

We annotated 25 common variants from both the transformed phenotypes into the genes using the SNPnexus [52] (a SNP annotation tool) and analyzed gene associated pathway in LifeMap GeneAnalytics tools [53]. Pathways such

as *Ion Transport by P-type ATPase* and *Factors involved in megakaryocyte development and platelet production* were found to be over-represented with high toxicity associated genes. Genes *FXYD1*, *FXYD7* were over-represented in Ion Transport by P-type ATPase pathway while *CAPZA2*, *JMJD1C* in factors involved in megakaryocyte development and platelet production pathway. These pathway were curated and derived from reactome database [60].

3.4.2 Leukopenia (LPK)

Similarly, results from linear regression analysis of single variant associated with LN and ENQT LPK are shown in Q-Q plot 3.12. None of the variants result in statistically significant Bonferrroni corrected p-values of 6.75×10^{-6}. However, we found two SNPs rs79823754 and rs111710000 significant at p-value < 0.05 for FDR-BH for LN transformed Nadir values. These SNPs mapped to genes *HDAC7* and *OLFM3* respectively.

In contrary to LN LPK, none of the variant results into statistically significant p-value for ENQT leukopenia phenotypes. SNP rs8018462 gave the lowest p-value at 2×10^{-5}. The SNP was mapped to *SLC7A7* gene. A Manhatton plot of variants and chromosome position of transformed LPK phenotypes are shown in the following Figure 3.13, 3.14

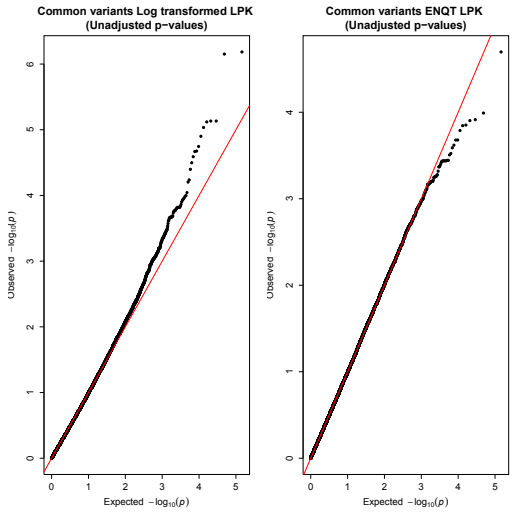

Figure 3.12: Q-Q plots for the transformed phenotype for LPK. The Q-Q plot for the LN LPK phenotype shows deviation from the theoretical empirical red line indicating distribution of the p-values are deviating from normal distribution in observed p-values

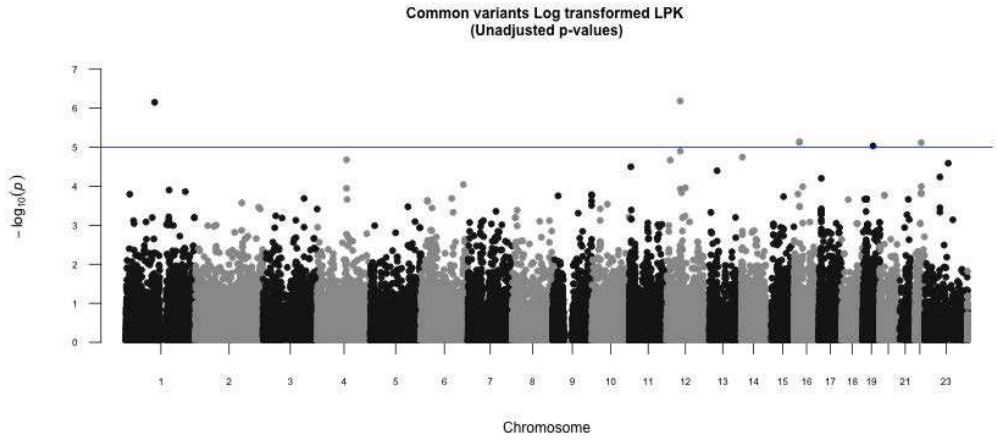

Figure 3.13: Manhatton Plot: LN-LPK

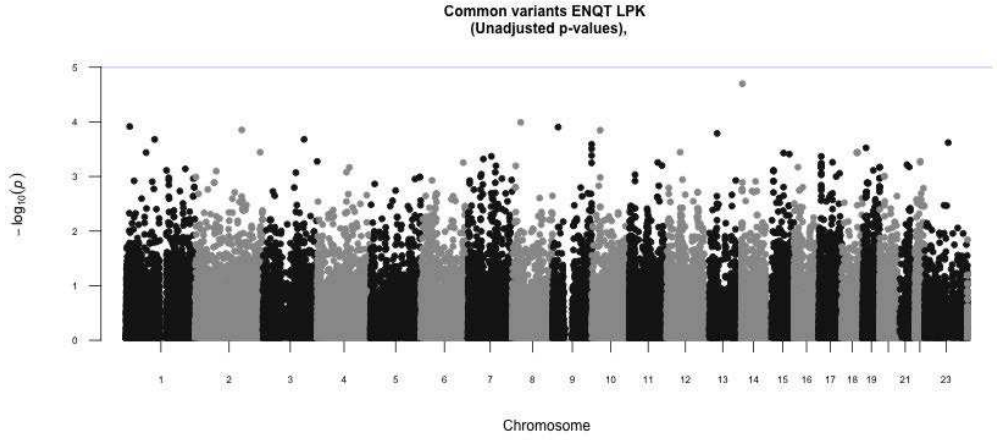

Figure 3.14: Manhatton Plot: ENQT-LPK

A comparative study of single variants at p-value $< 10^{-3}$ from both transformation methods identified total of 136 and 66 variants. There were 55 variants identified from both transformation methods as seen in the Figure 3.15. An alternative analysis of high toxicity variant at p-value $< 1.00 \times 10^{-3}$ and negative β values depicted 15 common variants for both transformed phenotypes as seen in the Figure 3.16.

We annotated 15 common variants from both the transformed phenotypes into the genes using the SNPnexus [52] and analyzed pathway associated in LifeMap GeneAnalytics tools [53]. Two genes *DNMT1* and *HDAC7* were over represented in *Macrophage Differentation and Growth Inhibition*

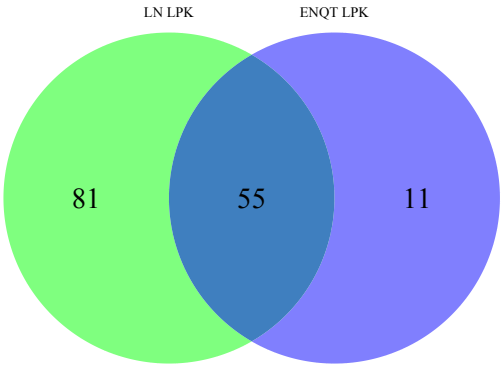

Figure 3.15: SNP comparision between the two transformed phenotypes at $p < 10^{-3}$. A total of 147 SNPs were identified to be associated in both the transformation methods at $p < 10^{-3}$ of which 55 variants are identified in both the transformations.

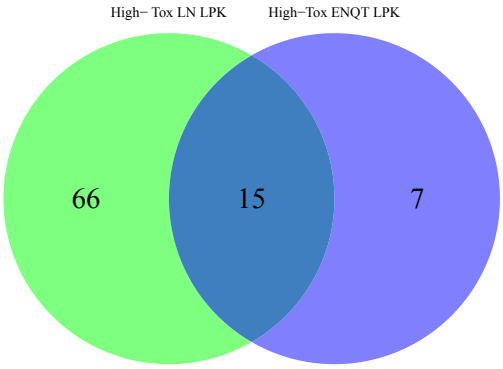

Figure 3.16: Venn diagram depicting the common variants associated with the high toxicity group (negative β values) in the regression analysis.15 common variants were associated in both the phenotypes for high LPK in cohort

by MEts and DNA Methylation and *Transcriptional Repression* pathways. These pathways are curated from Ingenuity pathway knowledge databases [61].

3.4.3 Neutropenia (NPK)

The SNV regression analysis, as shown in Figure 3.17, identified variant rs143522213 with p-value of 2.979×10^{-5} and 3.010×10^{-5} for the LN and ENQT NPK. The variant mapped to chromosome 3 as shown in Figure 3.18, and 3.19. There were 16 patients with the missing nadir values. These values were coded as -9 in phenotype file and association studies were carried out.

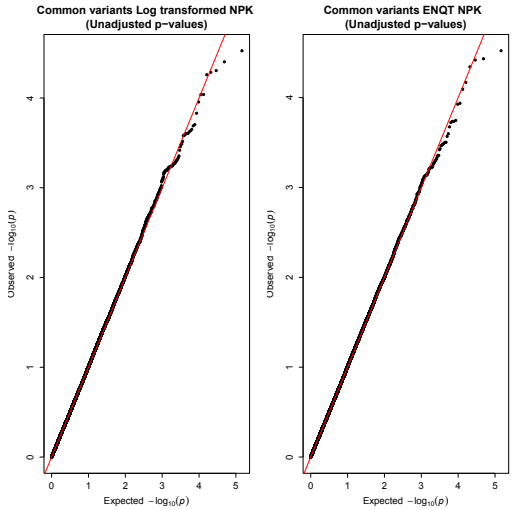

Figure 3.17: Q-Q plot for the transformed phenotype for NPK

As seen in Figure 3.20 72 of variants are associated with either of the transformed phenotype at p-value $< 1 \times 10^{-3}$. 31 high toxicity variants are identified for both transformed phenotypes as seen in the Figure 3.21. Upon annotation of common 31 variants in SNPnexus [52] and pathway analysis in LifeMap GeneAnalytics tools, genes *CYFIP2* and *ITGAE* are over represented in *E-cadherin signaling in the nascent adherens junction* and are curated in NCBI Biosystem pathway[62].

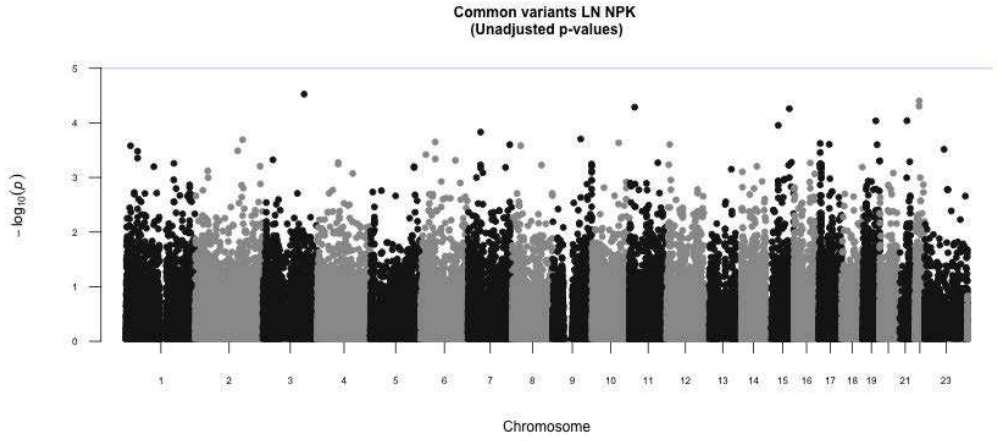

Figure 3.18: Manhatton Plot: LN-NPK

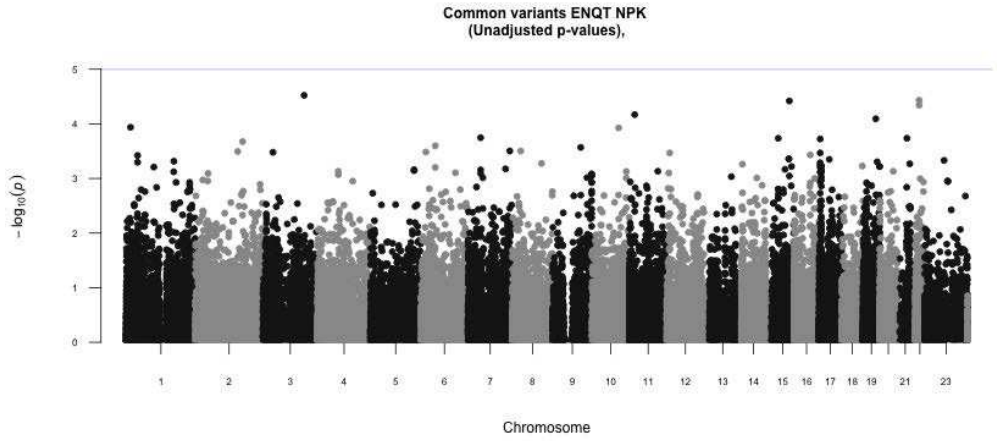

Figure 3.19: Manhatton Plot: ENQT-NPK

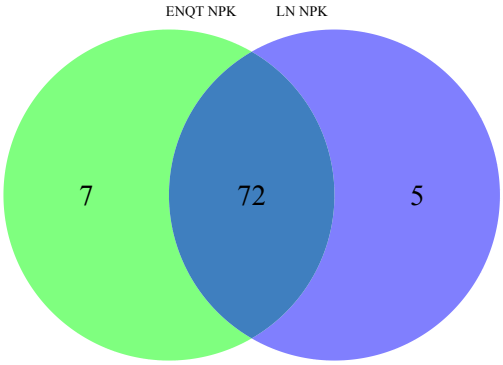

Figure 3.20: SNP comparision between the two transformed phenotypes at $p < 10^{-3}$.103 SNPs were identified to be associated in both the transformation methods at $p < 10^{-3}$ of which 72 variants are identified in both the transformations.

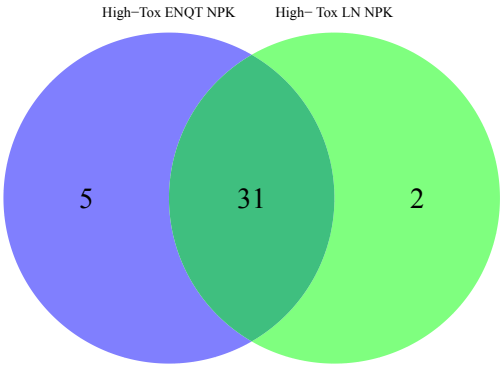

Figure 3.21: Venn diagram depicting the common variants associated with the high toxicity group (negative β values) in the regression analysis. 31 common variants were associated in both the phenotypes for high NPK in cohort

3.5 Qualitative trait single variant association study

3.5.1 Thrombcytopenia (TPK)

73429 bialleic variants from the 168 Case-Control Thrombocytopenia group were association with the logistic regression methods. Q-Q plot 3.22 and Manhatton plots 3.27 plot depicts the distribution and position of the variants high/low toxicity TPK phenotype.

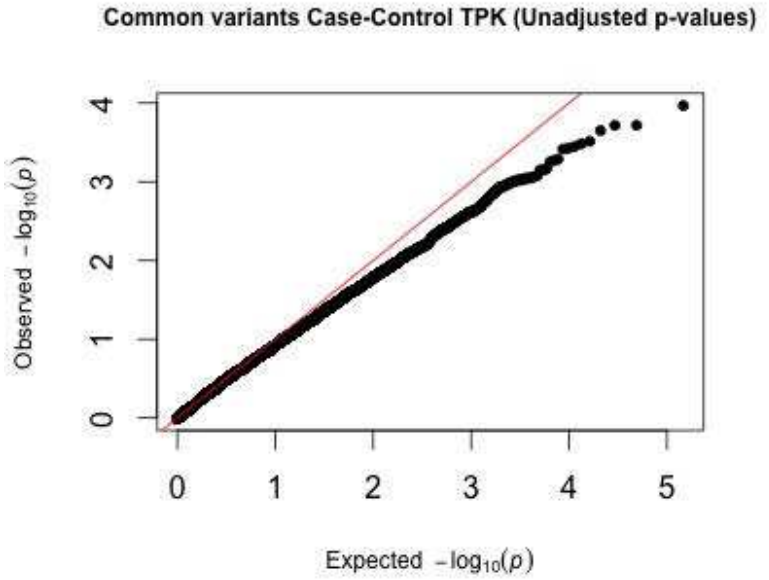

Figure 3.22: Q-Q plot of high/low toxicity group of TPK. The reference red line depicts theoretical line between the expected and observed distribution. The distribution of the observed p-values is deviating from the expected distribution. The possible explanation for the deviation could be we undertook only the cases of high and low toxicity which could cause the skewness in the distribution.

As in quantitative analysis, p-values for individual bi-allelic SNPs are underpowered to reach genome wide significance of Bonferroni corrected threshold of $< 6.8 \times 10^{-7}$ or adjusted Benjamini-Hochberg FDR threshold of < 0.05. The highest ranked SNP from the SNV association was intronic variant,

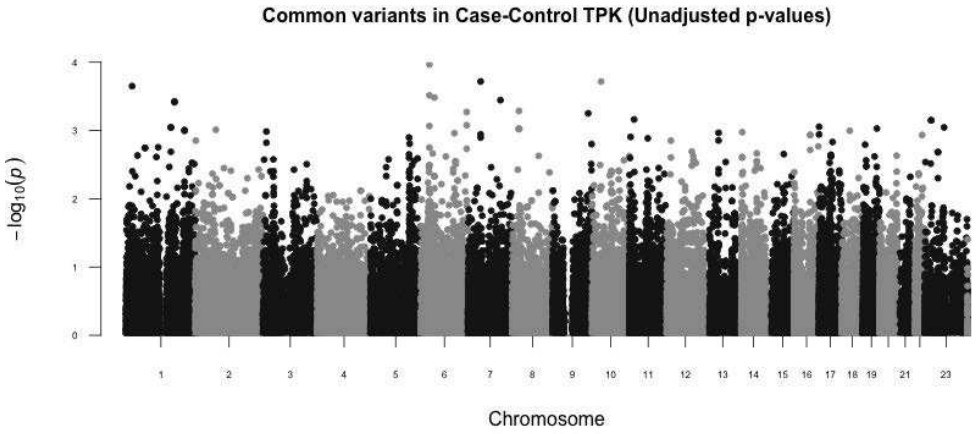

Figure 3.23: Manhatton plot of High/Low Toxicity group of TPK.The unadjusted p-value were taken on the y-axis with the SNV position on the x-axis. None of the SNPs reached the genome-wide significance of $< 6.8 \times 10^{-7}$

$rs66772001$ with p-value of 1.07×10^{-4}. The SNP mapped to *HLA-C* gene.

In the alternative analysis approach, 27 biallelic SNPs with p-value $< 1.00 \times 10^{-3}$ were identified. Upon SNPs annotation in SNP nexus [52] and pathway analysis in GeneAnalytics tools, Genes *ITGB1*, *LAMB2* were over represented in cell adhesion_ECM remodeling pathway curated in GeneGo Metago database [60].

3.5.2 Leukopenia (LPK)

78333 variants in 140 high/low toxicity leukopenia cohort were analyzed in logistic regression analysis. The distribution of p-values and position of association SNPs are shown in Q-Q plot 3.24 and Manhatton plot 3.27.

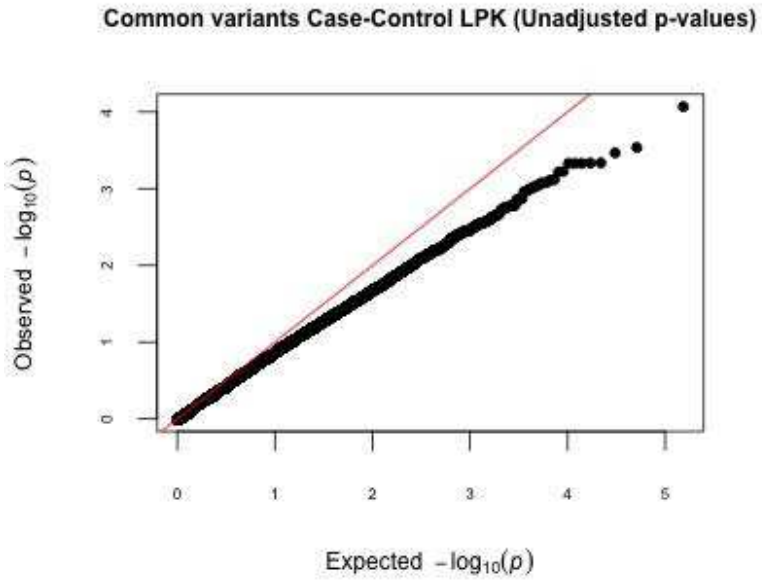

Figure 3.24: Q-Q plot of high/Low Toxicity group of LPK.The reference red line depicts theoretical line of a perfect match between the expected and observed distribution.

None of the variants reached statistically significant multiple corrected Bonferroni and FDR p-values. SNP rs61735550 was the top hit with the p-value of 8.54×10^{-5} and mapped to ZFHX3 gene on chromosome 16:72958322 position.

19 SNPs with p-value $< 1 \times 10^{-3}$ were mapped to the corresponding genes using the SNP nexus [52]. These SNPs mapped to 14 unique genes in RefSeq databases. These genes were upon analysis in LifeMap Gene toolkit found Protein digestion and absorption pathway, curated in KEGG pathway database [63] over represented. Two genes COL24A1 and SLC7A are associated in protein digestion and absorption pathway.

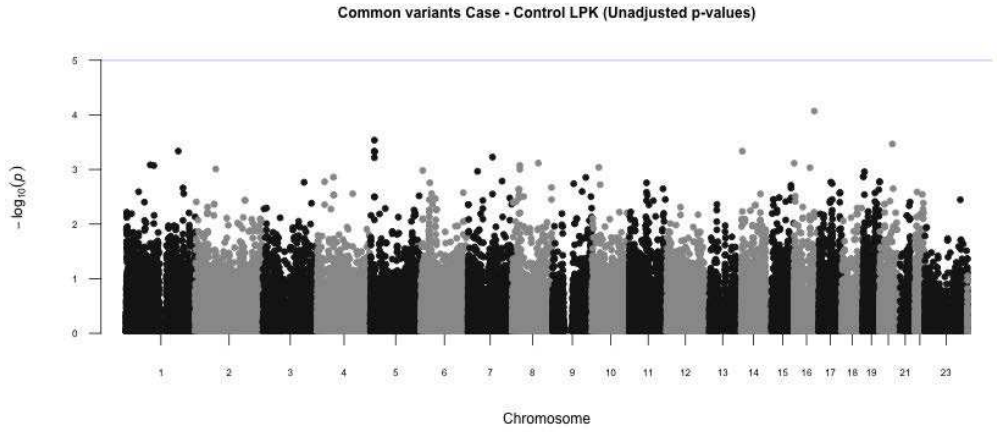

Figure 3.25: Manhatton plot of high/Low Toxicity group of LPK.The unadjusted p-value were taken on the y-axis with the SNV position on the x-axis. None of the SNPs reached the genome-wide significance of $< 6.8 \times 10^{-7}$

3.5.3 Neutropenia (NPK)

97 high toxicity and 76 low toxicity NPK sample cohort with 75354 bialleic variants are associated using logisitic regression. The visualization of the association test are shown in following Q-Q plot 3.26 and Manhatton plot 3.27.

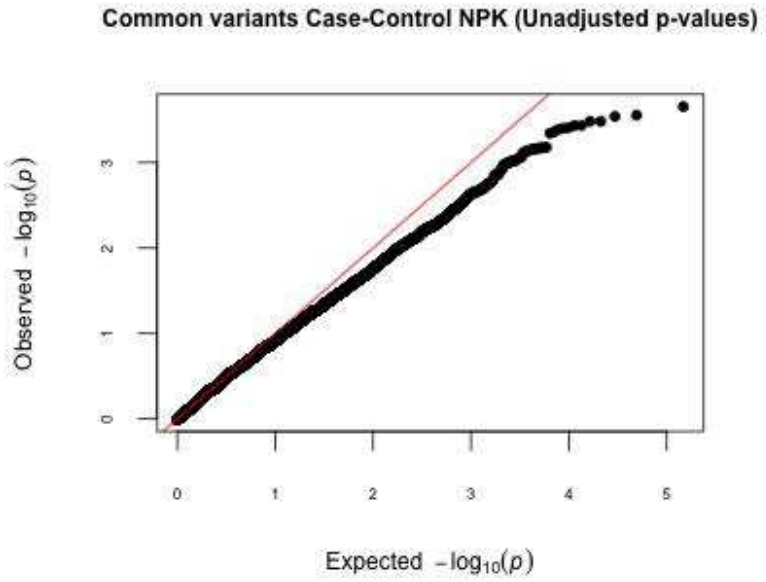

Figure 3.26: Q-Q plot of High/Low Toxicity group of NPK.The reference red line depicts theoretical line of a perfect match between the expected and observed distribution. The observed p-values is deviating from the expected distribution. The possible explanation for the deviation could be we undertook only the cases of high and low toxicity which could effect the distribution.

None of the variants were able to reach the Bonferronin multiple test correction p-value of $< 6.8 \times 10^{-7}$ or adjusted Benjamini-Hochberg FDR threshold of < 0.05. A SNP rs2301664 was the top hit with the p-value of 2.2×10^{-4}. The SNP mapped to SV2B gene on the chromosome 15:91827264 position.

30 SNVs with p-value with $< 1 \times 10^{-3}$ mapped to corresponding gene using SNP-Nexus tool. These mapped genes were run through the LifeMap GeneAnalytics tools to observe the genes represented in the pathways. Two genes *KLRK1* and *KLRC4-KLRK1* are genes represented in Malaria pathway. However, same two genes are also represented in the pathway relating to Immune response Role of *DAP2* receptors in NK cells implicating role in our Neutropenia phenotype.

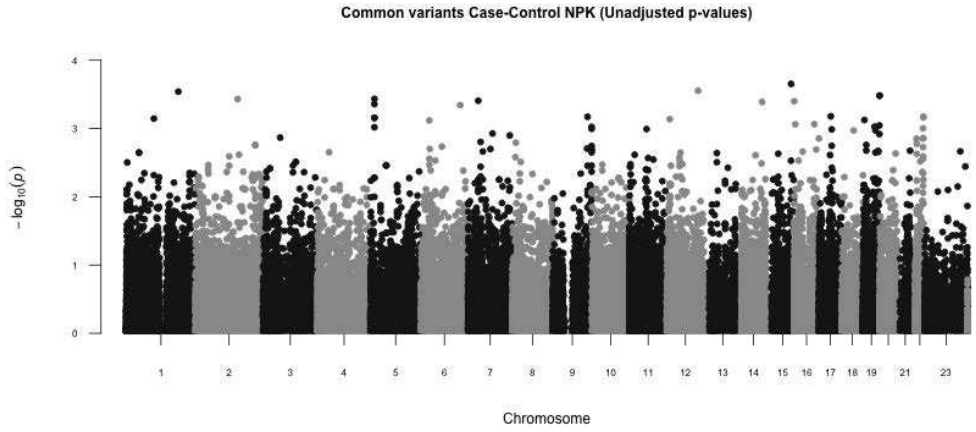

Figure 3.27: Manhatton plot of High/Low NPK Toxicity group.The unadjusted p-value were taken on the y-axis with the SNV position on the x-axis. None of the SNPs reached the genome-wide significance of $< 6.8 \times 10^{-7}$

3.6 Quantitative trait gene based association test

3.6.1 Thrombcytopenia (TPK)

The Q-Q plot distribution of the genes identified by SKATO test associated with both phenotypes are plotted and shown in Figure 3.28. The highest ranked genes for LN and ENQT TPK were *UBXN7* and *MYL7* at p-value 1.13×10^{-4} and 5.87×10^{-5} respectively. As in the SNV analysis, we analyzed genes using alternative strategy with p-value $< 1.00 \times 10^{-3}$. 15 and 13 genes are found to be associated for LN and ENQT modified phenotypes. 11 of the genes were found common in both the transformed phenotype. The Venn diagram 3.31 below representation of the genes found in both the transformed methods.

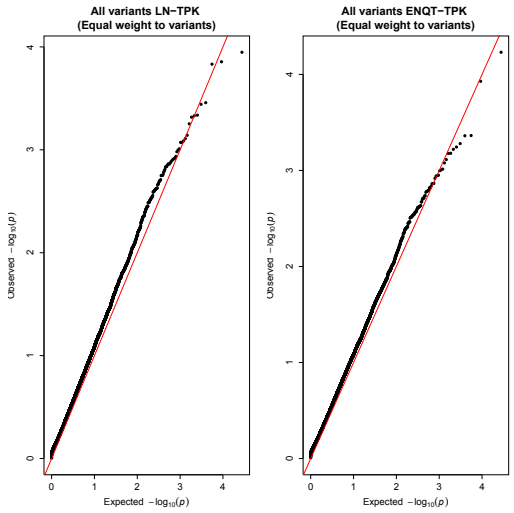

Figure 3.28: Q-Q plots: Gene based SKATO for the LN-transformed TPK. Both the transformed phenotypes have the qq-plot that deviates from the empirical p-value distribution

Four genes *ZSCAN26*, *CAPZA2*, *TRIM27* and *UBXN7* were identified in both SKATO gene tests and high toxicity single nucleotide variant regression tests at p-value $< 1.00 \times 10^{-3}$. The genotype frequency of the variants in 212 sample cohort is shown in Appendix A Table A.2. Both common variants rs4808 and rare variant rs374052696 are found to enriched at p-value $< 1.00 \times 10^{-3}$ in *CAPZA2*. Another interesting gene *ZSCAN26* are enriched with five rare variants in study cohort with single common variant allele.

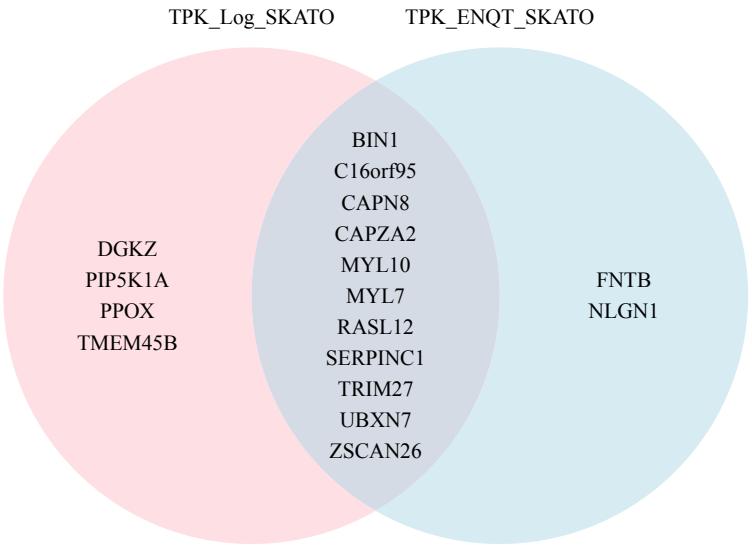

Figure 3.29: Comparison of the Genes identified by SKATO association methods in LN and ENQT TPK.

3.6.2 Leukopenia (LPK)

The Q-Q plot 3.30 distribution of p-values obtained for each transformed LPK phenotypes results *CFAP126*, *FTMT*, *RPL19*, *GSTK1*, *LONP2*, *GTF2E2*, *TSPO2* to be statisitcally significant values of adjusted FDR BH p-value < 0.05 and Bonferroni corrected p-value of $< 3.2 \times 10^{-6}$. However, for ENQT LPK no such significant genes are found. The top ranked gene is *TSPO2* at unadjusted p -value $< 5.85 \times 10^{-5}$. This discrepancy in the genes associated with LN and ENQT is caused due deviation of the LN-LPK phenotype from the normal distribution of the phenotype as seen in the earlier with Shapiro-Wilk test Table 3.2.

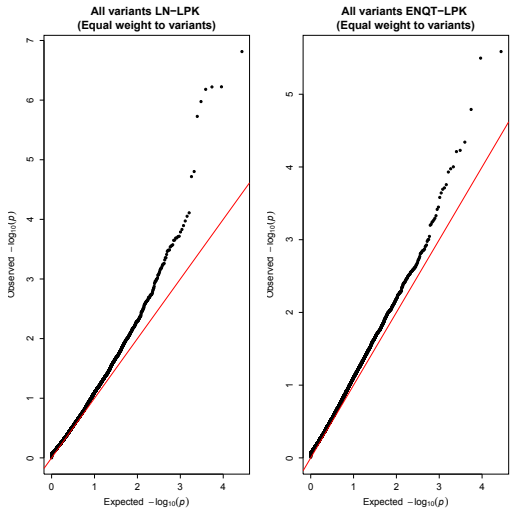

Figure 3.30: Q-Q plots-Gene based SKATO for the LN and ENQT

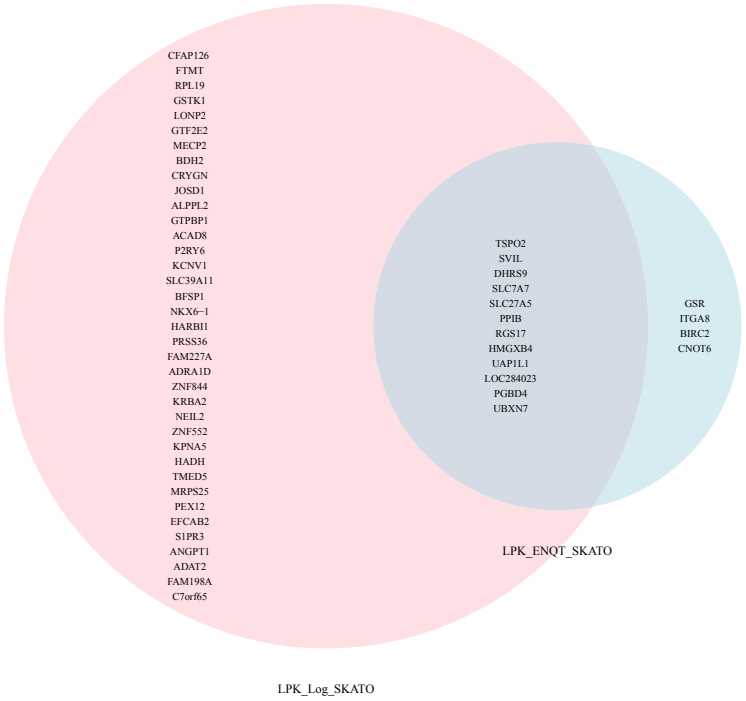

Figure 3.31: Comparision of the Genes identified by the both transformation

Thus, as in earlier studies, all genes with p-values $< 1.00 \times 10^{-3}$ from both phenotypes were compared. 49 LN-LPK and 16 ENQT-LPK associated genes were identified at $< 1.00 \times 10^{-3}$. Upon comparative analysis of genes 12 genes were identified with both methods as represented in the Venn diagram 3.31.

In order to identify genes associated with high toxicity, genes identified from SKATO gene association test was compared with the high toxicity genes identified in the Single variant association for both the phenotypes. *SVIL*,*SLC7A7*, *RGS17*, *HMGXB4* ,*UBXN7* genes were identified in both of the methods. The Appendix Table A.3 shows counts of the variants in the cohort samples in the identified genes.

3.6.3 Neutropenia (NPK)

Identical analysis pipeline was carried out with the LN and ENQT modified NPK phenotypes. No statistical significant p-values threshold was achieved for each of the transformed phenotype. Apart from small sample size, 16

patients were missing data. These samples were subsequently non-used in the association tests which further reduced sample cohort. Near identical Q-Q plot 3.32 was seen for both phenotypes. The highest ranked gene was *HOMER2* for both LN and ENQT transformed phenotypes with p-value of 2.98×10^{-5} and 2.87×10^{-5} respectively. As in previous comparative analysis, 20 and 15 genes were identified by the both LN and ENQT transformed methods at p-value $< 1 \times 10^{-3}$. All of the genes in ENQT were identified by LN NPK phenotype as seen in Figure 3.33.

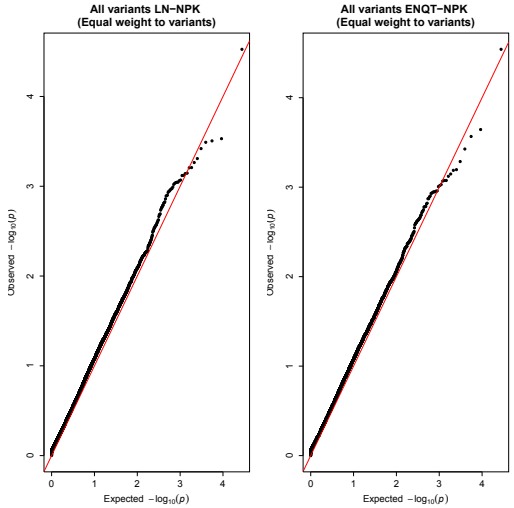

Figure 3.32: Q-Q plots-Gene based SKATO for the LN-transformed NPK

We compared the genes identified by both the phenotypes at p-value $< 1 \times 10^{-3}$ for both the single variant and gene variant association tests. Two genes HOMER2 and ZZEF1 were identified in both test. Upon further analysis, both rare and common variants were seen in NPK sample cohort. Appendix table A.4 shows the variants counts in the sample cohort for two genes.

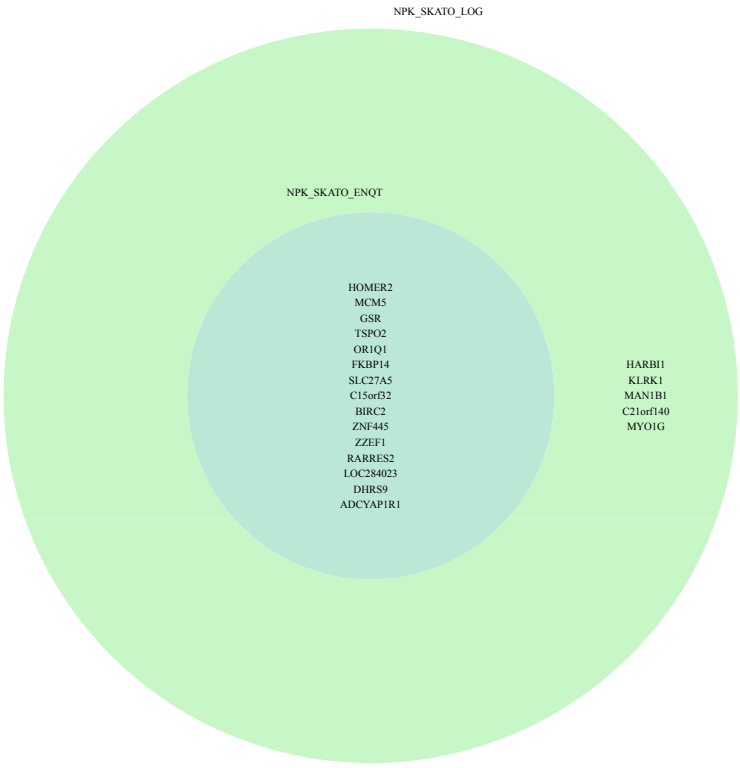

Figure 3.33: Comparison of the Genes identified by the both transformation in NPK

Chapter 4

Discussion

In the current study, we exome sequenced 216 NSCLC patients treated with the combination of carbolpatin and gemcitabine. Many patients showed different grades of adverse bone marrow myleosuppression upon the administration of drug. We investigated genetic variants associated with the individual phenotypes by analyzing the quantitative and qualitative traits based on individual phenotype.

4.1 Quantitative and Qualitative Single Variant Association tests

In the both quantitative and qualitative study, sample size of 216 study cohort made it infeasible to hunt down multiple corrected statistically significant p-values for genetic variants. Thus, we resorted to the alternative analysis strategy where the variants with p-values $< 1 \times 10^{-3}$ were taken into consideration. For the quantitative phenotypes we categorized the variants into high and low toxicity based on the β values for the linear regression analysis. Similarly, logistic regression was carried out in all the myelosupression phenotypes of case/control group with high toxicity as the patients with the CTC score of either 3 and 4. This strategy was adopted to find true positives variants that are associated in all of phenotype definition. 5 biallelic variants were identified in SNV analysis for both quantitative and qualitative TPK phenotype. The Figure 4.1 summarizes the variants identified in all of two of transformed quantitative phenotype and qualitative case-control TPK cohorts.

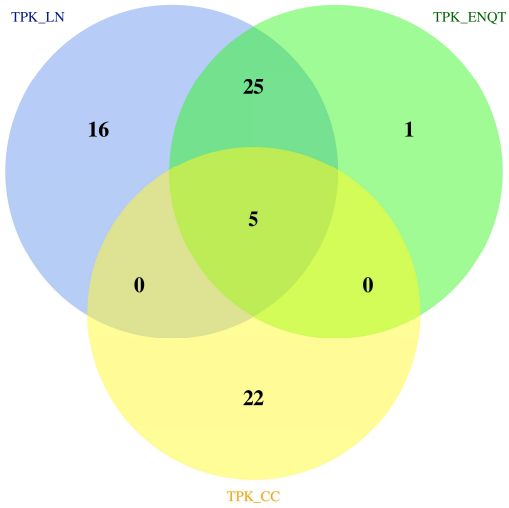

Figure 4.1: Biallelic SNP in all methods with p-value $< 1.0 \times 10^{-3}$. TPK_LN and TPK_ENQT refers to the log transformed nadir values for nadir thrombopenia values and CC represents the $case - control$ cohort of high/low toxicity TPK phenotypes.

SNPs rs56070322, rs10496192, rs61739531, rs4808, rs2298141 were identified by the above three methods. The SNPs mapped to genes *KIF17*, *ALMS1*, *MYO1G*, *CAPZA2*, *ITGB1* respectively. Interestingly, *CAPZA2* was also identified by the quantitative $gene-based$ SKATO association methods in the both transformed phenotypes. This supports that the idea that both rare and common variants could be enriched in the gene. Furthermore, SNPs rs374052696, rs4808 and 7:116502628 in the gene were enriched in the TPK case-control cohort as shown in Table 4.1. rs374052696 was present in a high toxicity patient while 7:116502628 and rs4808 was enriched in the high toxicity patients with the odd ratio of 2.51 and 2.86 respectively. Variant rs374052696 present in high toxicity patients is a 5-prime-UTR variant present at -41 position from the transcription site and reported at MAF of 0.0002 in 1000 genome project while 7:116502628 is an in-frame deletion that leads to GCC deletion. This provides inclination of *CAPZA2* and associated variants in the thrombocytopenia phenotype in the study cohort.

And potentially validates the advantage of using whole exome sequencing in genotyping patients rather than microarray based methods as we can profile all the common and rare variants in a gene.

Gene	SNP	Alt Alle	Freq Alt Alle	Freq Ref Alle	Ref Allele	ChisQ	p-value	Odd Ratio
CAPZA2	rs374052696	T	0.006667	0	C	1.23	0.2673	NA
CAPZA2	7:116502628	T	0.04	0.0163	TGCC	1.77	0.1834	2.514
CAPZA2	rs4808	T	0.32	0.1398	C	15.7	7.409×10^{-3}	2.896

Table 4.1: CAPZA2 variants in TPK Case-Control Cohort

Similar approach applied to LPK phenotypes identified a single variant rs8018462 in all of the study design as shown in Figure 4.2

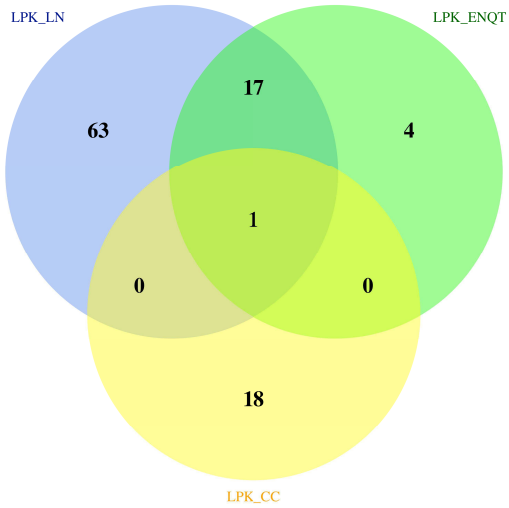

Figure 4.2: Biallelic SNP in all methods with p-value $< 1.0 \times 10^{-3}$ for LPK phenotype.Biallelic SNP in all method with p-value $< 1.0 \times 10^{-3}$. *LPK_LN* and *LPK_ENQT* refers to the log transformed nadir values for nadir leukopenia values and CC represents the $case - control$ cohort of high/low toxicity TPK phenotypes.

The rs8018462 SNP maps to *SLC7A7* gene at position 14:23282110 a common variant in our sample cohort. Additionally, *SLC7A7* gene was also identified in the gene based association SKATO test. Furthermore, in case control 12 variants were identified in the case-control LPK cohort were observed as shown in Table. Couple of SNPs such as rs1805062, rs8018462, rs1805059,rs2281677 in the SLC7A7 genes are present in the odd ratio greater than 2.0 which might indicate the variants int the genes are associated with the leukopenia phenotype in our study cohort. However, variants rs373156106 and rs199522527 were absent in LPK cases which might suggest these vari-

ants might provide protective advantage in our study cohort.

Gene	SNP	Alt Alle	Freq Alt Allele	Freq Ref Allele	Ref Allele	ChisQ	p-value	Odd Ratio
SLC7A7	rs143575981	A	0.0102	0.01099	G	0.003703	0.9515	0.9278
SLC7A7	rs1061040	C	0.8367	0.8846	T	1.273	0.2591	0.6685
SLC7A7	rs373156106	A	0	0.005495	G	0.5404	0.4623	0
SLC7A7	rs199522527	A	0	0.005495	T	0.5404	0.4623	0
SLC7A7	rs1805062	T	0.02041	0.005495	C	1.337	0.2476	3.771
SLC7A7	rs1805061	G	0.1429	0.1319	A	0.06558	0.7979	1.097
SLC7A7	rs8018462	G	0.6939	0.4451	A	15.84	6.89e-05	2.826
SLC7A7	rs11568438	A	0.02041	0.01099	G	0.4013	0.5264	1.875
SLC7A7	rs1805059	T	0.7174	0.5	C	11.76	0.0006041	2.538
SLC7A7	rs45479698	T	0.0102	0.02198	C	0.5035	0.478	0.4588
SLC7A7	rs2281677	G	0.6735	0.4725	A	10.36	0.001285	2.302
SLC7A7	rs28364570	G	0.06122	0.07143	A	0.1049	0.7461	0.8478

Table 4.2: SLC7A7 Variants in Case-Control LPK phenotype

Similar analysis pipeline for the transformed NPK phenotypes and the case-control cohort sample of high toxicity NPK led to the identification of the two SNPs rs55842403, and rs1049172 in all SNV association test at p-value $< 1.0 \times 10^{-3}$. The Venn diagram 4.3 summarizes number of SNPs identified in the association test in all methods used. rs55842403 and rs1049172 mapped to genes *LPPR5* and *KLRK1* respectively. However, the gene based association methods didn't identify these genes as these SNP have a single nucleotide polymorphism called by GATK variant calling pipelines. Additionally, our definition of region/gene included only genes that had the more than one variants.

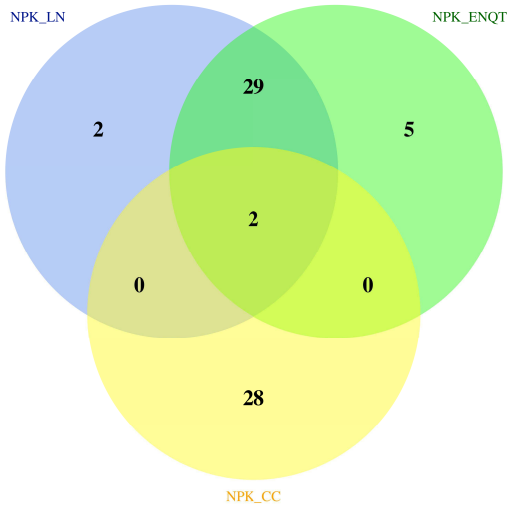

Figure 4.3: Biallelic SNP in all method with p-value $< 1.0 \times 10^{-3}$ for NPK phenotype.Biallelic SNP in all method with p-value $< 1.0 \times 10^{-3}$. *NPK_LN* and *NPK_ENQT* refers to the log transformed nadir values for nadir leukopenia values and CC represents the $case - control$ cohort of high/low toxicity TPK phenotypes.

4.2 Biological Interpretation of Associated Genes

In the study, we found that *CAPZA2* gene was associated with the high toxicity thrombocytopenia phenotypes in the patient cohort. *CAPZA2* gene codes protein, F-actin-capping protien subunit 2, a member of F-actin capping protein and regulates the growth of the actin filament. *CAPZA2* gene is related to megakaryocyte development and platelet production pathway in curated reactome database [60]. Precursor cells megakaryocytes are derived from haemotopoietic stem cells (HSC), primarily in bone marrow. These cells differentiates into circulating platelets with development of cytoplasmic the structural and functional characteristics [64, 65]. Furthermore, pegylated recombinant human megakaryocyte growth and development factor (PEG-rHuMGDF) has been shown to reduce severe thromobocytopenia in chemotherapy treatment in cancer patients [66].

SLC7A7 variants were found to be associated with leukopenia phenotype with p-value $< 1.0 \times 10^{-3}$ for both case/control and quantitative phenotypes. *SLC7A7* are cationic and neutral amino acid transporter gene necessary for normal and abnormal cell growth and proliferation [67]. Studies have suggested *SLC7A7* gene mutations are responsible for rare recessive disorder,

Lysinuric protein intolerance (LPI) [68, 69].A Japanese and Korean studies have shown Lysinuric protein intolerance study patients presented with pulmonary disease, haematological abnormalities of which leukopenia is one of the clinical features [70, 70]. Moreover, study has shown the mutations in SLC7A7 leading to a dysfunctional LPI macrophages [71] which might implicate its role in immune response.

Chapter 5

Future perspectives

In the current study, we investigated to identify variants associated with the myleosupression phenotypes at $p < 1 \times 10^{-3}$ using both quantitative and qualitative study design. However, these variants and genes are unable to achieve statistically significance due to limited sample size in the study. Since toxicity phenotypes are hypothesized to be multigenic traits, it might need large number of sample size to reach statistical significance. At the meantime, it is essential to understand that it is substantially hard to assemble a homogeneous study patient cohort treated with same drugs in same cancer patients. This is to our knowledge the largest effort to dissect myelosuppression toxicity in non-squamous lung cancer patient treated with carboplatin/gemcitabine with sequencing technologies.

Furthermore, the variants identified in the study needs to be validated. One method of validation of the results would be functional assay of identified genes and variants in-vitro conditions. The functional studies could be performed either using cell-line or knock-down mutations in model animals that is usually done in candidate based genetic studies. Currently, we are undertaking functional studies in CAPZA and variants in TPK phenotype to validate the findings in our study. Another approach for validation could be using replication of the association study in independent, identical NSCLC patients. However, the replication study should be performed in same population as the original association study.

Currently, in the project we performed whole exome sequencing for the extracting variants from the patient samples. However, exome consists of 2 % of the whole genome and apart from exomic variants, intronic variants are reported to play important role in toxicity and disease. Hence a whole genome sequencing of current patient sample and association of variants would be able to identify and elucidate biological mechanisms in finer details. In order to improve our understanding, we have scaled up to whole genome sequencing

of 98 sample cohort and starting data analysis of study cohorts.

Finally, in order to understand the effect of the variants we could integrate different data analysis from multitude of omics technologies such as transcriptomics, proteomics to further investigate the association of myelosuppression toxicity in NSCLC patients.

Bibliography

[1] B Meibohm and H Derendorf. Basic concepts of pharmacokinetic/pharmacodynamic (pk/pd) modelling. *International journal of clinical pharmacology and therapeutics*, (35):401–13, 1997.

[2] Heather E Wheeler, Michael L Maitland, M Eileen Dolan, Nancy J Cox, and Mark J Ratain. Cancer pharmacogenomics: strategies and challenges. *Nature Reviews Genetics*, 14(1):23–34, 2013.

[3] Erika L Moen, Lucy A Godley, Wei Zhang, and M Eileen Dolan. Pharmacogenomics of chemotherapeutic susceptibility and toxicity. *genetic testing*, 15:33, 2012.

[4] Konrad J Karczewski, Roxana Daneshjou, Russ B Altman, Fran Lewitter, and Maricel Kann. Chapter 7: pharmacogenomics. *PLoS Comput Biol*, 8(12):e1002817, 2012.

[5] Lynne Lennard, Jon A Van Loon, John S Lilleyman, and Richard M Weinshilboum. Thiopurine pharmacogenetics in leukemia: Correlation of erythrocyte thiopurine methyltransferase activity and 6-thioguanine nucleotide concentrations. *Clinical Pharmacology & Therapeutics*, 41(1):18–25, 1987.

[6] LSLMJTGFEERLMJ Iyer, S Das, L Janisch, M Wen, J Ramirez, T Karrison, GF Fleming, EE Vokes, RL Schilsky, and MJ Ratain. Ugt1a1* 28 polymorphism as a determinant of irinotecan disposition and toxicity. *The pharmacogenomics journal*, 2(1):43–47, 2002.

[7] Federico Innocenti, Deanna L Kroetz, Erin Schuetz, M Eileen Dolan, Jacqueline Ramírez, Mary Relling, Peixian Chen, Soma Das, Gary L Rosner, and Mark J Ratain. Comprehensive pharmacogenetic analysis of irinotecan neutropenia and pharmacokinetics. *Journal of Clinical Oncology*, 27(16):2604–2614, 2009.

[8] Kazuma Kiyotani, Taisei Mushiroda, Naoya Hosono, Tatsuhiko Tsunoda, Michiaki Kubo, Fuminori Aki, Yutaka Okazaki, Koichi Hirata, Yuichi Takatsuka, Minoru Okazaki, et al. Lessons for pharmacogenomics studies: association study between cyp2d6 genotype and tamoxifen response. *Pharmacogenetics and genomics*, 20(9):565–568, 2010.

[9] André BP van Kuilenburg. Dihydropyrimidine dehydrogenase and the efficacy and toxicity of 5-fluorouracil. *European journal of cancer*, 40(7):939–950, 2004.

[10] World Health Organization. WHO fact sheet, April 2015. `http://www.who.int/mediacentre/factsheets/fs297/en/`, [Online; accessed 28-February-2013].

[11] Rebecca Siegel, Jiemin Ma, Zhaohui Zou, and Ahmedin Jemal. Cancer statistics, 2014. *CA: a cancer journal for clinicians*, 64(1):9–29, 2014.

[12] Kouya Shiraishi, Takashi Kohno, Chiharu Tanai, Yasushi Goto, Aya Kuchiba, Seiichiro Yamamoto, Koji Tsuta, Hiroshi Nokihara, Noboru Yamamoto, Ikuo Sekine, et al. Association of dna repair gene polymorphisms with response to platinum-based doublet chemotherapy in patients with non–small-cell lung cancer. *Journal of Clinical Oncology*, pages JCO–2010, 2010.

[13] X Chen, Y Wu, H Dong, C Y Zhang, and Y Zhang. Platinum-based agents for individualized cancer treatmen. *Current molecular medicine*, 13(10):1603–1612, 2013.

[14] John Goffin, Christina Lacchetti, Peter M Ellis, Yee C Ung, William K Evans, Lung Cancer Disease Site Group of Cancer Care Ontario's Program in Evidence-Based Care, et al. First-line systemic chemotherapy in the treatment of advanced non-small cell lung cancer: a systematic review. *Journal of Thoracic Oncology*, 5(2):260–274, 2010.

[15] Roy S Herbst, Diane Prager, Robert Hermann, Lou Fehrenbacher, Bruce E Johnson, Alan Sandler, Mark G Kris, Hai T Tran, Pam Klein, Xin Li, et al. Tribute: A phase iii trial of erlotinib hydrochloride (osi-774) combined with carboplatin and paclitaxel chemotherapy in advanced non–small-cell lung cancer. *Journal of Clinical Oncology*, 23(25):5892–5899, 2005.

[16] S Cao, S Wang, H Ma, S Tang, C Sun, J Dai, C Wang, Y Shu, L Xu, R Yin, et al. Genome-wide association study of myelosuppression in

non-small-cell lung cancer patients with platinum-based chemotherapy. *The pharmacogenomics journal*, 2015.

[17] William P McGuire, William J Hoskins, Mark F Brady, Paul R Kucera, Edward E Partridge, Katherine Y Look, Daniel L Clarke-Pearson, and Martin Davidson. Cyclophosphamide and cisplatin compared with paclitaxel and cisplatin in patients with stage iii and stage iv ovarian cancer. *New England Journal of Medicine*, 334(1):1–6, 1996.

[18] bone marrow suppression. (n.d.). Mosby's medical dictionary, 8th edition, April 2009.
`http://medical-dictionary.thefreedictionary.com/bonemarrowsuppression`, [Online; accessed 12-April-2015].

[19] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

[20] William S Bush and Jason H Moore. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, 2012.

[21] Nathan O Stitziel, Adam Kiezun, Shamil Sunyaev, et al. Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol*, 12(9):227, 2011.

[22] Bahareh Rabbani, Mustafa Tekin, and Nejat Mahdieh. The promise of whole-exome sequencing in medical genetics. *Journal of human genetics*, 59(1):5–15, 2014.

[23] Michael J Bamshad, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure. Exome sequencing as a tool for mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755, 2011.

[24] Ron Do, Sekar Kathiresan, and Gonçalo R Abecasis. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Human molecular genetics*, 21(R1):R1–R9, 2012.

[25] Mary J Emond, Tin Louie, Julia Emerson, Wei Zhao, Rasika A Mathias, Michael R Knowles, Fred A Wright, Mark J Rieder, Holly K Tabor, Deborah A Nickerson, et al. Exome sequencing of extreme phenotypes identifies dctn4 as a modifier of chronic pseudomonas aeruginosa infection in cystic fibrosis. *Nature genetics*, 44(8):886–889, 2012.

[26] Shaun M Purcell, Jennifer L Moran, Menachem Fromer, Douglas Ruderfer, Nadia Solovieff, Panos Roussos, Colm O Dushlaine, Kimberly Chambert, Sarah E Bergen, Anna Kähler, et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487):185–190, 2014.

[27] Peter Weeke, Jonathan D Mosley, David Hanna, Jessica T Delaney, Christian Shaffer, Quinn S Wells, Sara Van Driest, Jason H Karnes, Christie Ingram, Yan Guo, et al. Exome sequencing implicates an increased burden of rare potassium channel variants in the risk of drug-induced long qt interval syndrome. *Journal of the American College of Cardiology*, 63(14):1430–1437, 2014.

[28] Fengju Song, Christopher I Amos, Jeffrey E Lee, Christine G Lian, Shenying Fang, Hongliang Liu, Stuart MacGregor, Mark M Iles, Neal I Lindeman, Grant W Montgomery, et al. Identification of a melanoma susceptibility locus and somatic mutation in tet2. *Carcinogenesis*, page bgu140, 2014.

[29] Gyungah Jun, Hirohide Asai, Ella Zeldich, Elodie Drapeau, CiDi Chen, Jaeyoon Chung, Jong-Ho Park, Sehwa Kim, Vahram Haroutunian, Tatiana Foroud, et al. Plxna4 is associated with alzheimer disease and modulates tau phosphorylation. *Annals of neurology*, 76(3):379–392, 2014.

[30] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[31] Li Ma, Jing Yang, H Birali Runesha, Toshiko Tanaka, Luigi Ferrucci, Stefania Bandinelli, and Yang Da. Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the framingham heart study data. *BMC medical genetics*, 11(1):55, 2010.

[32] Camilla Helene Sandholt, Kristine Højgaard Allin, U Toft, A Borglykke, Rasmus Ribel-Madsen, T Sparso, Johanne Marie Justesen, Marie Neergaard Harder, Torben Jørgensen, Torben Hansen, et al. The effect of gwas identified bmi loci on changes in body weight among middle-aged danes during a five-year period. *Obesity*, 22(3):901–908, 2014.

[33] Petra Buvzkova. Linear regression in genetic association studies. *PloS one*, 8(2):e56976, 2013.

[34] David J Balding. A tutorial on statistical methods for population association studies. *Nature Reviews GeneStics*, 7(10):781–791, 2006.

[35] Dan-Yu Lin and Zheng-Zheng Tang. A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics*, 89(3):354–367, 2011.

[36] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[37] Stephan Morgenthaler and William G Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1):28–56, 2007.

[38] Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

[39] Bo Eskerod Madsen and Sharon R Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384, 2009.

[40] Benjamin M Neale, Manuel A Rivas, Benjamin F Voight, David Altshuler, Bernie Devlin, Marju Orho-Melander, Sekar Kathiresan, Shaun M Purcell, Kathryn Roeder, Mark J Daly, et al. Testing for an unusual distribution of rare variants. *PLoS Genet*, 7(3):e1001322, 2011.

[41] Seunggeun Lee, Michael C Wu, and Xihong Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775, 2012.

[42] Andy Trotti, Roger Byhardt, Joanne Stetz, Clement Gwede, Benjamin Corn, Karen Fu, Leonard Gunderson, Beryl McCormick, Mitchell Morris, Tyvin Rich, et al. Common toxicity criteria: version 2.0. an improved reference for grading the acute effects of cancer treatment: impact on radiotherapy. *International Journal of Radiation Oncology* Biology* Physics*, 47(1):13–47, 2000.

[43] Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2010.

[44] Krueger Felix. Trim galore!, 2013. `http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`, [Online; accessed 28-February-2013].

[45] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

[46] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[47] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[48] P. Cingolani, A. Platts, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, and D.M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

[49] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[50] Purcell Shaun. Plink/seq, 2014. `https://atgu.mgh.harvard.edu/plinkseq/`,[Online; accessed 28-February-2013].

[51] William S Noble. How does multiple testing correction work? *Nature biotechnology*, 27(12):1135–1137, 2009.

[52] Abu Z Dayem Ullah, Nicholas R Lemoine, and Claude Chelala. A practical guide for the functional annotation of genetic variations using snpnexus. *Briefings in bioinformatics*, 14(4):437–447, 2013.

[53] Ron Edgar, Yaron Mazor, Ariel Rinon, Jacob Blumenthal, Yaron Golan, Ella Buzhor, Idit Livnat, Shani Ben-Ari, Iris Lieder, Alina Shitrit, et al.

LifeMap Discovery: The embryonic development, stem cells, and regenerative medicine research portal. 2013.

[54] Seunggeun Lee, with contributions from Larisa Miropolsky, and Michael Wu. *SKAT: SNP-Set (Sequence) Kernel Association Test*, 2015. R package version 1.0.7.

[55] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[56] Kim D Pruitt, Garth R Brown, Susan M Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M Farrell, Jennifer Hart, Melissa J Landrum, Kelly M McGarvey, et al. Refseq: an update on mammalian reference sequences. *Nucleic acids research*, 42(D1):D756–D763, 2014.

[57] Donna Karolchik, Robert Baertsch, Mark Diekhans, Terrence S Furey, A Hinrichs, YT Lu, Krishna M Roskin, M Schwartz, Charles W Sugnet, Daryl J Thomas, et al. The ucsc genome browser database. *Nucleic acids research*, 31(1):51–54, 2003.

[58] nadir (n.d.). Mosby's medical dictionary, 8th edition, April 2009. `http://medical-dictionary.thefreedictionary.com/bonemarrowsuppression`, [Online; accessed 4-July-2015].

[59] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[60] David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 42(D1):D472–D477, 2014.

[61] Ingenutiy Systems. `https://www.qiagen.com/se/products/genes%20and%20pathways/pathway%20details?pwid=136`, [Online; accessed 4-July-2015].

[62] Lewis Y Geer, Aron Marchler-Bauer, Renata C Geer, Lianyi Han, Jane He, Siqian He, Chunlei Liu, Wenyao Shi, and Stephen H Bryant. The ncbi biosystems database. *Nucleic acids research*, page gkp858, 2009.

[63] Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic acids research*, 42(D1):D199–D205, 2014.

[64] Sunita R Patel, John H Hartwig, and Joseph E Italiano Jr. The biogenesis of platelets from megakaryocyte proplatelets. *Journal of Clinical Investigation*, 115(12):3348, 2005.

[65] Varda R Deutsch and Aaron Tomer. Megakaryocyte development and platelet production. *British journal of haematology*, 134(5):453–466, 2006.

[66] Russell L Basser, Craig Underhill, Ian Davis, Michael D Green, Jonathan Cebon, John Zalcberg, Jamie MacMillan, Brian Cohen, Jennifer Marty, Richard M Fox, et al. Enhancement of platelet recovery after myelosuppressive chemotherapy by recombinant human megakaryocyte growth and development factor in patients with advanced cancer. *Journal of Clinical Oncology*, 18(15):2852–2861, 2000.

[67] Songhua Fan, Delong Meng, Tao Xu, Yuanyuan Chen, Jingkun Wang, Xiaoying Li, Hongyan Chen, Daru Lu, Juxiang Chen, and Qing Lan. Overexpression of slc7a7 predicts poor progression-free and overall survival in patients with glioblastoma. *Medical Oncology*, 30(1):1–7, 2013.

[68] Giuseppe Borsani, Maria Teresa Bassi, Maria Pia Sperandeo, Alessandro De Grandi, Anna Buoninconti, Mirko Riboni, Marta Manzoni, Barbara Incerti, Antonio Pepe, Generoso Andria, et al. Slc7a7, encoding a putative permease-related protein, is mutated in patients with lysinuric protein intolerance. *Nature genetics*, 21(3):297–301, 1999.

[69] David Torrents, Juha Mykkänen, Marta Pineda, Lidia Feliubadaló, Raúl Estévez, Rafael de Cid, Pablo Sanjurjo, Antonio Zorzano, Virginia Nunes, Kirsi Huoponen, et al. Identification of slc7a7, encoding y+ lat-1, as the lysinuric protein intolerance gene. *Nature genetics*, 21(3):293–296, 1999.

[70] Akio Koizumi, Norio Matsuura, Sumiko Inoue, Maki Utsunomiya, Jun-Ichi Nozaki, Kayoko Inoue, and Yuhei Takasago. Evaluation of a mass screening program for lysinuric protein intolerance in the northern part of japan. *Genetic testing*, 7(1):29–35, 2003.

[71] Amelia Barilli, Bianca Maria Rotoli, Rossana Visigalli, Ovidio Bussolati, Gian C Gazzola, Rita Gatti, Carlo Dionisi-Vici, Diego Martinelli, Bianca M Goffredo, Mariona Font-Llitjós, et al. Impaired phagocytosis in macrophages from patients affected by lysinuric protein intolerance. *Molecular genetics and metabolism*, 105(4):585–589, 2012.

Appendix A

First appendix

ID	No. ALT	No. MIN	No. HET	No. VAR	RATE	SING	TITV	PASS	PASS Sin	QUAL	DP
S0143	30245	22331	18699	158553	0.999	463	2.023	22331	188	143678	NA
S0156	30172	22330	18581	158583	0.999	509	2.035	22330	234	145227	NA
S0160	30512	22481	18856	158554	0.999	494	2.048	22481	219	144403	NA
S0162	30641	22537	19109	158623	0.999	474	2.041	22537	198	142552	NA
S0164	30739	23021	19689	158643	0.999	488	2.012	23021	213	143158	NA
S0170	30729	22832	19810	158634	0.999	475	2.048	22832	200	142167	NA
S0172	30933	23512	19470	158554	0.999	1011	2.044	23512	749	142381	NA
S0174	30228	22467	18749	158633	0.999	494	2.088	22467	218	145628	NA
S0177	30459	22191	18189	158570	0.999	503	2.032	22191	227	139713	NA
S0178	30954	22953	19407	158562	0.999	500	2.067	22953	224	143826	NA
S0225	30196	22256	18609	158610	0.999	530	2.051	22256	254	142733	NA
S0229	30699	22645	19197	158464	0.998	540	2.057	22645	267	141723	NA
S0237	30734	22804	19316	158550	0.999	494	2.09	22804	218	143465	NA
S0240	30409	22505	18877	158528	0.999	486	2.054	22505	216	142811	NA
S0243	30400	22516	18666	158490	0.998	484	2.087	22516	209	143958	NA
S0253	30428	22616	18899	158485	0.998	527	2.066	22616	255	143996	NA
S0258	30321	22423	19120	158473	0.998	483	2.086	22423	208	146239	NA
S0259	30604	22400	18728	158452	0.998	462	2.056	22400	186	142318	NA
S0264	30679	22957	19330	158436	0.998	496	2.071	22957	221	144332	NA
S0267	30820	22660	19278	158522	0.999	523	2.079	22660	257	142090	NA
S0269	30783	22872	19022	158495	0.998	725	2.064	22872	452	142113	NA
S0274	30572	22714	19270	158540	0.999	512	2.039	22714	236	144230	NA
S0279	29998	22128	18574	158422	0.998	612	2.115	22128	337	143632	NA
S0280	29793	22231	18255	158487	0.998	501	2.154	22231	225	146290	NA
S0282	30077	22100	18699	158415	0.998	519	2.159	22100	244	144884	NA
S0286	30112	22170	18273	158454	0.998	466	2.147	22170	191	145216	NA
S0287	29648	21822	18218	158447	0.998	493	2.148	21822	218	147216	NA
S0290	30274	22441	18782	158517	0.999	504	2.131	22441	229	146387	NA
S0295	30405	22609	19035	158496	0.998	588	2.149	22609	314	145738	NA
S0306	30450	22551	18924	158495	0.998	510	2.15	22551	236	143701	NA
S0307	30272	22227	18654	158406	0.998	487	2.117	22227	212	144173	NA
S0309	30157	22469	18791	158407	0.998	509	2.14	22469	233	147981	NA
S0313	30485	22584	19013	158438	0.998	512	2.152	22584	237	145460	NA
S0322	28625	20989	17243	156689	0.987	448	2.151	20989	180	151345	NA
S0325	30075	22194	18730	158139	0.996	521	2.09	22194	245	145001	NA
S0328	25474	18541	13921	149703	0.943	371	2.237	18541	128	162352	NA
S0330	29854	21879	18435	158197	0.996	501	2.169	21879	225	146422	NA
S0333	30478	22688	19145	158534	0.999	518	2.138	22688	243	144647	NA
S0338	30003	22047	18217	158521	0.999	505	2.157	22047	229	145463	NA
S0340	29854	21883	18073	158432	0.998	482	2.156	21883	209	145030	NA
S0341	29902	22090	18472	158488	0.998	494	2.102	22090	223	146774	NA
S0347	30224	22202	18518	158392	0.998	488	2.131	22202	216	144303	NA
S0349	30194	22219	18759	158501	0.998	487	2.16	22219	212	143988	NA
S0351	28595	20946	16888	156239	0.984	445	2.213	20946	185	151474	NA
S0353	30109	22238	18499	158388	0.998	491	2.183	22238	216	145089	NA
S0365	29907	22166	18331	158504	0.998	504	2.133	22166	228	146494	NA
S0376	30614	22530	19071	158513	0.998	500	2.144	22530	224	142620	NA
S0380	30365	22498	18845	158576	0.999	508	2.15	22498	232	144815	NA
S0397	30100	22252	18612	158504	0.998	565	2.15	22252	290	146659	NA
S0400	30349	22422	18747	158538	0.999	491	2.136	22422	216	145353	NA
S0404	30214	22279	18588	158451	0.998	477	2.129	22279	202	146260	NA
S0406	30330	22291	18714	158495	0.998	510	2.098	22291	235	143085	NA
S0407	30604	22696	18813	158559	0.999	592	2.086	22696	319	142947	NA
S0409	30321	22431	18726	158495	0.998	534	2.094	22431	258	144838	NA
S0411	30520	22506	18680	158577	0.999	544	2.111	22506	268	143188	NA
S0415	30308	22420	18648	158561	0.999	500	2.094	22420	225	144204	NA
S0416	30117	22416	18744	158509	0.998	469	2.162	22416	193	145258	NA
S0428	30283	22423	18795	158410	0.998	477	2.133	22423	203	143276	NA
S0437	30464	22563	18906	158487	0.998	496	2.138	22563	221	143123	NA
S0439	30303	22499	18526	158453	0.998	733	2.108	22499	463	142432	NA
S0445	29900	22387	18450	158443	0.998	483	2.138	22387	209	146241	NA
S0447	30038	22278	18663	158363	0.998	485	2.124	22278	209	145442	NA
S0451	30415	22708	19038	158454	0.998	546	2.134	22708	270	144935	NA
S0459	30314	22335	18345	158424	0.998	516	2.119	22335	241	142761	NA

ID	No. ALT	No. MIN	No. HET	No. VAR	RATE	SING	TITV	PASS	PASS Sin	QUAL	DP
S0466	29899	22123	18261	158382	0.998	513	2.119	22123	237	146383	NA
S0469	29555	21839	18267	158469	0.998	523	2.114	21839	248	148653	NA
S0472	30086	22184	18481	158321	0.997	503	2.091	22184	230	144637	NA
S0477	30391	22418	18891	158450	0.998	618	2.098	22418	346	144717	NA
S0488	30262	22482	18730	157815	0.994	473	2.098	22482	203	145179	NA
S0490	29999	22000	18271	158357	0.998	492	2.071	22000	221	143948	NA
S0492	30295	22394	18801	158330	0.997	490	2.102	22394	216	144177	NA
S0494	29979	22244	18442	158465	0.998	481	2.095	22244	207	146177	NA
S0496	30077	22305	18570	158327	0.997	494	2.098	22305	223	144629	NA
S0497	30304	22407	19028	158420	0.998	516	2.122	22407	246	145855	NA
S0498	30183	21975	18374	158368	0.998	519	2.093	21975	245	143762	NA
S0500	30233	22292	18485	158478	0.998	501	2.1	22292	225	142835	NA
S0512	29943	22351	18630	158298	0.997	504	2.05	22351	232	147166	NA
S0525	30356	22353	18659	158344	0.997	482	2.11	22353	211	144356	NA
S0527	29655	21806	18157	157950	0.995	463	2.117	21806	191	146374	NA
S0532	28832	21213	17326	155427	0.979	464	2.213	21213	201	150015	NA
S0552	28551	21047	17118	157283	0.991	452	2.159	21047	182	151408	NA
S0559	29990	22613	18156	158033	0.995	816	2.121	22613	547	145472	NA
S0561	30606	22980	19071	158449	0.998	859	2.109	22980	590	142329	NA
S0579	30093	22089	18485	158160	0.996	501	2.092	22089	227	144411	NA
S0580	37823	31505	30699	158419	0.998	498	2.194	31505	229	147326	NA
S0584	29009	21469	17657	156664	0.987	460	2.178	21469	198	149486	NA
S0591	29647	22004	18078	157822	0.994	491	2.125	22004	217	148559	NA
S0594	27805	20479	16296	153681	0.968	443	2.248	20479	188	153559	NA
S0600	29747	21997	18424	157491	0.992	483	2.154	21997	212	147573	NA
S0601	28693	21109	16968	156205	0.984	456	2.142	21109	190	151407	NA
S0604	30395	22319	18671	158292	0.997	509	2.099	22319	236	144380	NA
S0607	29893	21940	17993	158433	0.998	485	2.084	21940	211	144366	NA
S0611	30071	22279	18543	158296	0.997	479	2.112	22279	206	145107	NA
S0620	30014	22174	18437	157382	0.991	432	2.073	22174	164	146308	NA
S0626	27689	20221	16074	152675	0.962	415	2.155	20221	160	155139	NA
S0629	29163	22354	17310	156373	0.985	999	2.12	22354	739	147158	NA
S0631	26893	19982	15801	151297	0.953	405	2.139	19982	158	158389	NA
S0648	28880	21336	17263	155626	0.98	480	2.118	21336	220	148924	NA
S0650	29636	21823	18196	157418	0.992	478	2.1	21823	208	148029	NA
S0651	29677	21725	17877	157931	0.995	475	2.062	21725	204	145105	NA
S0653	30217	22459	18756	158348	0.997	495	2.129	22459	223	145872	NA
S0659	30110	22292	18087	158209	0.997	487	2.083	22292	217	143850	NA
S0664	30443	24418	22296	148320	0.934	351	2.193	24418	106	163973	NA
S0671	28859	21315	17111	155520	0.98	454	2.082	21315	191	151211	NA
S0677	26946	19723	15334	151116	0.952	415	2.158	19723	166	155236	NA
S0680	28777	21259	17014	155407	0.979	456	2.136	21259	196	150183	NA
S0681	28980	21339	17379	156352	0.985	462	2.139	21339	197	149778	NA
S0682	29708	21910	17990	158006	0.995	500	2.117	21910	230	146265	NA
S0683	29878	22184	18379	157689	0.993	465	2.132	22184	194	146824	NA
S0687	29822	22080	18271	158246	0.997	476	2.075	22080	207	147979	NA
S0693	28179	20857	16869	153871	0.969	443	2.111	20857	186	155092	NA
S0718	28779	21151	16950	155616	0.98	478	2.143	21151	215	147532	NA
S0724	26999	19905	15681	150612	0.949	393	2.195	19905	142	157665	NA
S0728	29072	21644	17852	156057	0.983	492	2.168	21644	225	148998	NA
S0732	29217	21443	17573	157317	0.991	481	2.096	21443	208	147950	NA
S0761	29785	22095	18180	158064	0.996	542	2.12	22095	268	147406	NA
S0762	29841	22043	18216	158004	0.995	483	2.074	22043	210	147089	NA
S0773	29784	21929	18172	158080	0.996	487	2.119	21929	219	147010	NA
S0774	27877	20699	16570	153488	0.967	474	2.14	20699	211	153359	NA
S0790	28390	20933	16815	154831	0.975	433	2.115	20933	172	151708	NA
S0791	27925	20542	16397	153929	0.97	413	2.136	20542	158	154204	NA
S0801	29054	21452	17404	156204	0.984	446	2.104	21452	180	146652	NA
S0804	29765	22047	18431	158058	0.996	470	2.102	22047	199	148598	NA
S0807	30254	22216	18698	157798	0.994	491	2.093	22216	222	145093	NA
S0812	29905	22103	18302	157860	0.994	505	2.109	22103	235	143912	NA
S0828	27802	20674	16871	153941	0.97	430	2.145	20674	180	154309	NA
S0832	29376	21593	17863	157272	0.991	495	2.11	21593	227	147639	NA

ID	No. ALT	No. MIN	No. HET	No. VAR	RATE	SING	TITV	PASS	PASS Sin	QUAL	DP
S0837	29851	21965	18345	158003	0.995	469	2.093	21965	198	146107	NA
S0862	29658	22269	18581	158121	0.996	488	2.067	22269	214	148610	NA
S0864	29268	21668	17768	156698	0.987	460	2.078	21668	193	146790	NA
S0867	29153	21466	17537	158132	0.996	426	2.066	21466	154	148251	NA
S0873	30709	22564	18972	158498	0.998	544	2.044	22564	274	142600	NA
S0883	30237	22255	18533	158491	0.998	464	2.08	22255	190	145310	NA
S0886	30555	22768	18919	158544	0.999	524	2.066	22768	248	144863	NA
S0895	30014	22148	18234	158212	0.997	504	2.128	22148	232	145430	NA
S0900	27802	20432	16326	152660	0.962	416	2.149	20432	156	153266	NA
S0922	28330	20752	16789	153769	0.969	261	2.136	20752	1	150760	NA
S0933	27250	19879	15628	151600	0.955	442	2.145	19879	191	153795	NA
S0934	28465	20888	16988	154584	0.974	446	2.12	20888	187	150547	NA
S0935	30518	22573	19151	158552	0.999	519	2.098	22573	244	144195	NA
S0940	28871	21239	16982	157238	0.99	435	2.112	21239	168	147887	NA
S0944	30641	22609	18998	158580	0.999	539	2.091	22609	264	142108	NA
S0947	29155	21681	17817	157073	0.989	505	2.08	21681	237	147521	NA
S0956	29565	21826	18031	157959	0.995	465	2.119	21826	196	147910	NA
S0958	29830	22152	18350	158432	0.998	513	2.093	22152	245	147473	NA
S0979	30553	22666	18917	158554	0.999	527	2.153	22666	253	142518	NA
S0984	30191	22195	18549	158584	0.999	461	2.077	22195	186	142878	NA
S0986	28247	20437	16515	154398	0.973	453	2.114	20437	196	149567	NA
S1003	30644	22826	19249	158589	0.999	485	2.101	22826	210	144259	NA
S1025	30511	22461	18710	158581	0.999	501	2.061	22461	228	142697	NA
S1032	30253	22309	18770	158527	0.999	494	2.108	22309	219	143706	NA
S1041	30414	22459	18873	158535	0.999	509	2.058	22459	233	144184	NA
S1054	30643	22640	19223	158459	0.998	537	2.091	22640	261	142796	NA
S1056	30622	22473	18973	158526	0.999	299	2.113	22473	24	142614	NA
S1066	30865	22902	19535	158544	0.999	530	2.047	22902	254	143797	NA
S1070	30307	22453	18888	158588	0.999	519	2.08	22453	245	143695	NA
S1071	28920	21260	17379	156185	0.984	507	2.112	21260	240	148524	NA
S1079	30414	22469	18765	158543	0.999	517	2.059	22469	241	144086	NA
S1089	30548	22829	18716	158554	0.999	813	2.057	22829	537	143909	NA
S1109	27548	20236	16166	152732	0.962	400	2.14	20236	152	154671	NA
S1111	30155	22292	18843	158576	0.999	472	2.106	22292	196	145239	NA
S1116	30592	22537	18969	158494	0.998	456	2.086	22537	180	142367	NA
S1123	28903	21372	16896	156163	0.984	775	2.137	21372	521	148415	NA
S1126	29536	21626	17869	157335	0.991	518	2.092	21626	248	146876	NA
S1132	30294	22543	19090	158574	0.999	608	2.074	22543	335	143302	NA
S1135	30425	22564	18829	158557	0.999	513	2.106	22564	237	145265	NA
S1153	30143	22319	18720	157752	0.994	469	2.078	22319	199	145558	NA
S1156	30111	22345	18563	158485	0.998	499	2.126	22345	223	145347	NA
S1168	30625	22554	18914	158594	0.999	549	2.083	22554	274	142758	NA
S1169	30059	22347	18687	158230	0.997	486	2.119	22347	213	145335	NA
S1180	30247	22338	18952	158557	0.999	518	2.063	22338	243	144908	NA
S1193	30470	22719	19055	158592	0.999	513	2.077	22719	237	144343	NA
S1197	30113	22283	18607	158575	0.999	537	2.116	22283	261	144899	NA
S1205	30737	22738	19269	158528	0.999	493	2.062	22738	218	143145	NA
S1208	30373	22385	18899	158592	0.999	513	2.128	22385	237	144681	NA
S1211	30006	21980	18246	158574	0.999	500	2.133	21980	224	144682	NA
S1214	30088	22471	18554	158592	0.999	474	2.132	22471	198	144394	NA
S1217	30302	22286	18679	158626	0.999	504	2.049	22286	229	141923	NA
S1218	29356	21897	18322	158250	0.997	478	2.097	21897	205	149530	NA
S1219	30236	22397	18478	158643	0.999	506	2.068	22397	230	144932	NA
S1220	29614	22057	17080	158556	0.999	554	2.097	22057	279	144510	NA
S1228	29820	22095	18198	158336	0.997	585	2.097	22095	312	146504	NA
S1231	28140	20755	16476	154238	0.972	439	2.124	20755	183	151032	NA
S1232	30363	22857	19233	158260	0.997	494	2.048	22857	220	146012	NA
S1240	30358	22529	18997	158568	0.999	499	2.15	22529	223	144058	NA
S1244	30776	23032	19355	158557	0.999	638	2.078	23032	364	142952	NA
S1254	30542	22632	19014	158562	0.999	460	2.068	22632	185	143954	NA
S1259	30401	22469	18825	158639	0.999	472	2.051	22469	197	143593	NA
S1272	30293	22686	19000	158623	0.999	515	2.089	22686	239	145840	NA
S1273	30624	22715	19189	158592	0.999	502	2.026	22715	227	143752	NA
S1283	29316	21788	17889	156410	0.985	474	2.085	21788	211	148110	NA

ID	No. ALT	No. MIN	No. HET	No. VAR	RATE	SING	TITV	PASS	PASS Sin	QUAL	DP
S1292	28151	20724	16735	154106	0.971	425	2.144	20724	173	153921	NA
S1319	30095	22302	18444	158555	0.999	528	2.144	22302	254	146414	NA
S1325	28880	21136	17047	156573	0.986	482	2.095	21136	215	148267	NA
S1327	29542	21859	17852	157806	0.994	510	2.122	21859	243	146345	NA
S1361	29845	21957	18407	158208	0.997	460	2.1	21957	189	146585	NA
S1364	30088	22338	18647	158202	0.997	567	2.125	22338	296	146971	NA
S1366	30369	22505	18876	158588	0.999	485	2.088	22505	210	145381	NA
S1368	29712	21982	18077	158112	0.996	465	2.144	21982	193	146604	NA
S1374	28263	20855	16915	154621	0.974	502	2.114	20855	248	152764	NA

Table A.1: Summary Statistics of indiviudal genotype

Gene	**Variant**	**Genotype in Cohort**		
CAPZA2	rs374052696	T/T=0	T/C=1	C/C=210
CAPZA2	7:116502628	T/T=0	T/TGCC=9	TGCC/TGCC= 202
CAPZA2	rs4808	T/T=12	T/C=64	C/C=136
TRIM27	rs41270608	A/A=0	A/G=2	G/G=210
TRIM27	6:28887823	C/C=0	C/T=3	T/T=209
TRIM27	rs143463783	A/A=0	A/G=1	G/G=211
TRIM27	rs2230683	C/C=3	C/T=30	T/T=176
UBXN7	rs61742253	C/C=0	C/T=2	T/T=210
UBXN7	rs73213957	G/G =0	G/A=5	A/A=207
ZSCAN26	rs76463649	G/G =0	G/A=2	A/A=210
ZSCAN26	rs16893892	G/G =0	G/A=2	A/A=210
ZSCAN26	rs11965538	A/A =6	A/G=37	G/G=162
ZSCAN26	rs11965542	A/A=0	A/G=2	G/G=210
ZSCAN26	rs187327081	T/T=0	T/C=1	G/G=211
ZSCAN26	6:28244225	C/C=0	C/G=1	G/G=211

Table A.2: Common genes and variants identified by high Toxicity single variant Association and Gene based association studies in whole cohort

Gene	Variant	Genotype in Cohort		
HMGXB4	rs148351517	T/T= 0	T/C=5	C/C=207
HMGXB4	rs1053593	T/T=88	T7G=98	G/G=26
HMGXB4	22:35661305	A/A=0	A/G=1	G/G=211
HMGXB4	22:35661371	T/T=0	T/C=1	C/C=211
HMGXB4	rs2272789	C/C=101	C/T=88	T/T= 23
RGS17	rs2295230	C/C=17	C/A=90	A/A=105
RGS17	rs41292882	A/A=0	A/G=27	G/G=185
SLC7A7	rs143575981	A/A=0	A/G=4	G/G=208
SLC7A7	rs1061040	C/C=160	C/T=51	T/T= 1
SLC7A7	rs373156106	A/A=0	A/G=1	G/G=211
SLC7A7	rs199522527	A/A=0	A/T=1	T/T=1
SLC7A7	rs1805062	T/T=0	T/C=6	C/C=206
SLC7A7	rs1805061	G/G=1	G/A=58	A/A=153
SLC7A7	rs8018462	G/G=62	G/A=103	A/A=47
SLC7A7	14:23282335	T/T=0	T/C =1	C/C=211
SLC7A7	rs11568438	A/A=0	A/G=8	G/G=204
SLC7A7	rs1805059	T/T=74	T/C=85	C/C=45
SLC7A7	rs45479698	T/T=0	T/C=8	C/C=204
SLC7A7	rs2281677	G/G=63	G/A=106	A/A=43
SLC7A7	14:23284892	A/A=0	A/G=1	G/G=211
SLC7A7	rs28364570	G/G=1	G/A=1	A/A=180
SVIL	rs56022643	A/A=0	A/G=4	G/G=208
SVIL	rs1057952	C/C=27	C/T=43	T/T=137
SVIL	rs7921306	C/C=9	C/T=30	T/T=173
SVIL	rs11007607	G/G=12	G/A=85	A/A=115
SVIL	10:29762907	C/C=0	C/G= 1	G/G= 207
SVIL	rs61737920	T/T=1	T/C=25	C/C=186
SVIL	rs10763720	A/A=4	A/G=55	G/G=153
SVIL	rs11007612	G/G=4	G/A=51	A/A=157
SVIL	rs1056782	G/G=41	G/A= 129	A/A=42
SVIL	rs56817459	A/A=0	A/G=42	G/G=170
SVIL	rs146267453	A/A=0	A/C=1	C/C=211
SVIL	rs7070135	A/A=1	A/C=41	C/C=170
SVIL	rs145392867	T/T=0	T/G=2	G/G=210
SVIL	rs7070678	T/T=37	T/G=112	G/G=63
SVIL	10:29813439	G/G=0	G/A=1	A/A=211

Gene	**Variant**	**Genotype in Cohort**		
SVIL	10:29820187	A/A=0	A/G=1	G/G=211
SVIL	rs17756919	T/T=40	T/C=98	C/C=74
SVIL	rs41284748	A/A=1	A/G=31	G/G=180
SVIL	rs1328323	C/C=45	C/T=103	T/T=64
SVIL	rs147010426	T/T=0	T/C=5	C/C=207
SVIL	rs150826046	A/A=0	A/G=1	G/G=211
SVIL	rs1247696	T/T=204	T/C=7	C/C=1
SVIL	rs7076239	C/C=45	C/T=103	T/T=64
SVIL	rs142262993	T/T=0	T/C=1	C/C=211
SVIL	rs138539716	C/C=0	C/T=1	T/T=211
SVIL	rs143011277	A/A=0	A/G=1	G/G=211
SVIL	rs141506698	T/T=0	T/C=2	C/C=210
SVIL	10:29839785	T/T=0	T/C=1	C/C=211
SVIL	rs10160013	G/G=15	G/A=75	A/A=122
SVIL	rs17834991	G/G=7	G/A=69	A/A=136
SVIL	rs1270874	C/C=120	C/A=77	A/A=15
SVIL	10:29839886	G/G=0	G/T=1	T/T=211
SVIL	rs3740003	G/G=15	G/A=76	A/A=121
SVIL	rs3740002	G/G=18	G/A=66	A/A=126
SVIL	rs1547169	T/T=15	T/C=74	C/C=123
SVIL	rs375845375	C/C=0	C/G=1	G/G=211
SVILP1	10:30993387	C/C=0	C/G=1	G/G=210
SVILP1	rs112090325	A/A=0	A/G=2	G/G=209
SVILP1	rs11008192	A/A=79	A/G=93	G/G=39
SVILP1	rs79612491	C/C=0	C/T=1	T/T=211
SVILP1	rs10826848	A/A=178	A/G=32	G/G=2
SVILP1	rs141761009	A/A=0	A/G=2	G/G=210
SVILP1	rs1826619	C/C=144	C/A=60	A/A=8
UBXN7	rs61742253	C/C=0	C/T =2	T/T=210
UBXN7	rs73213957	G/G =0	G/A=5	A/A=207

Table A.3: Table: Gene and SNP identified by High Toxicity Single Nucleotide and Gene based test

Gene	Variant	Genotype in Cohort		
HOMER2	rs34287296	A/A=0	A/G=6	G/G=206
HOMER2	rs74416301	A/A=0	A/G=10	G/G=202
HOMER2	rs79448007	C/C=0	C/T=3	T/T=209
HOMER2	rs76145073	C/C=0	C/T=1	T/T=211
HOMER2	rs200280749	A/A=0	A/G=2	G/G=210
ZZEF1	rs62072392	G/G=5	G/A=67	A/A=140
ZZEF1	rs116870033	G/G=0	G/A=1	A/A=211
ZZEF1	17:3912217	A/A=0	A/G=1	G/G=211
ZZEF1	rs76915727	A/A=0	A/G=6	G/G=206
ZZEF1	rs112497098	A/A=0	A/G=1	G/G=211
ZZEF1	rs8075562	A/A=11	A/G=70	G/G=131
ZZEF1	rs201134194	A/A=0	A/G=1	G/G=211
ZZEF1	rs72827323	C/C=0	C/G=3	G/G=209
ZZEF1	rs711177	G/G=2	G/C=23	C/C=187
ZZEF1	rs35511240	A/A=0	A/G=3	G/G=209
ZZEF1	rs1006954	A/A=5	A/G=56	G/G=151
ZZEF1	rs35284780	T/T=0	T/C=16	C/C=196
ZZEF1	rs781831	C/C=31	C/T=117	T/T=64
ZZEF1	rs34719232	A/A=0	A/G=1	G/G=211
ZZEF1	rs781853	C/C=26	C/T=115	T/T=71
ZZEF1	rs781852	G/G=28	G/A=113	A/A=71
ZZEF1	rs146287047	A/A=0	A/G=4	G/G=208
ZZEF1	rs117408376	A/A=0	A/G=1	G/G=211
ZZEF1	rs781825	G/G=28	G/A=113	A/A=71
ZZEF1	rs4790555	C/C=0	C/A=1	A/A=211
ZZEF1	17:3970468	C/C=0	C/G=1	G/G=211
ZZEF1	rs9891850	T/T=0	T/G=9	G/G=203
ZZEF1	rs139226355	A/A=0	A/G=1	G/G=211

Gene	**Variant**	**Genotype in Cohort**		
ZZEF1	rs34760976	A/A=1	A/G=53	G/G=158
ZZEF1	17:3978633	T/T=0	T/C=1	C/C=211
ZZEF1	rs143736611	T/T=0	T/C=2	C/C=210
ZZEF1	rs7207986	A/A=3	A/G=62	G/G=147
ZZEF1	rs8065185	A/A=5	A/G=59	G/G=148
ZZEF1	rs78806449	G/G=1	G/A=30	A/A=181
ZZEF1	17:3994109	G/G=0	G/A=1	A/A=211
ZZEF1	rs12947597	T/T=3	T/C=60	C/C=149
ZZEF1	rs7222392	C/C=6	C/T=58	T/T=148
ZZEF1	rs143093880	A/A=0	A/G=1	G/G=211
ZZEF1	17:4015912	C/C=0	C/T=1	T/T=211
ZZEF1	rs150456516	A/A=0	A/G=1	G/G=211
ZZEF1	rs58625333	G/G=8	G/A=81	A/A=123
ZZEF1	rs138134000	C/C=0	C/T=11	T/T=201
ZZEF1	rs117738178	T/T=0	T/C=5	C/C=207
ZZEF1	rs188631556	T/T=0	T/A=1	A/A=211
ZZEF1	rs111724159	A/A=0	A/G=6	G/G=205

Table A.4: Table: Gene and SNP identified by High Toxicity Single Nucleotide and Gene based test NPK