Benchmarking of read mapping bias in allele specific expression analysis

Master Project in the Bioinformatics and Systems Biology Program at Chalmers University of Technology

Alva Rani James, 2013

Supervisor: Assist Professor Olof Emanuelsson, *KTH – Royal Institute of Technology, Science for Life Laboratory, Stockholm*

Co-supervisor: Dr Daniel Edsgärd, *KTH – Royal Institute of Technology, Science for Life Laboratory, Stockholm*

Examiner: Professor Olle Nerman, *Chalmers University of Technology, Gothenburg*

Department of Mathematical Sciences
CHALMERS UNIVERSITY OF TECHNOLOGY
Goteborg, Sweden SE- 412 96

Abstract

Most genes in diploid organisms have two “copies”; one copy inherited from each parent. If an individual has two different alleles (code variants) at a specific gene locus, then the individual is heterozygous at that locus. Allele specific expression (ASE) can be explained as the differential expression between the two different alleles of a gene in a single individual. There are several mechanisms that can cause ASE, e. g, it can be caused by a heterozygous variant in the promoter region, causing a difference in transcription factor binding affinity between the maternal and paternal allele. Accurate measurement and identification of ASE can be obtained by precise mapping of reads, generated from RNA next generation sequencing (RNA-seq), towards the reference genome of the organism. Mapping bias is a major technical hurdle in ASE studies which arises when we map short RNA-seq reads towards a reference genome. This arises mainly when the reads which carries non-reference alleles is not matching towards the reference genome gives out a lower mapping quality. In this thesis we investigated two proposed methods to reduce mapping bias: a read mapping program called GSNAP, and masking the reference genome with respect to single nucleotide variants. Masking the reference genome removed the mapping bias to a greater degree than GSNAP; however, the masking caused a considerable drop in read coverage. In conclusion, none of the two methods reduced the mapping bias satisfactorily, highlighting the importance to develop new or modified methods for mapping bias reduction.

Acknowledgement

This master thesis wouldn't have been possible without certain people in so many ways. First and foremost I would like to thank Assistant Professor Olof Emmanuelsson, KTH University, and my thesis supervisor, for accepting me into his group. I would like to thank his rewarding assistance and support during my tenure.

Each and every result in this thesis is accomplished by the immense help of my co-supervisor Dr Daniel Edgard. I would like to express my special thanks for his incredible patience, he patiently explained me all the details of my work. I am highly indebted for his contribution during each and every stage of my thesis, especially for all those great conversations on how to write a scientific paper and how to handle errors during analysis.

I owe my deepest gratitude to my Examiner Professor Olle Nerman, who helped me out for the thesis presentation and during report writing. I would like to extent my special thanks to Santhilal Subash my thesis Opponent, for that infectious level of energy and support all towards the end of my thesis.

Deep from my heart I vouch for the fact that the incredible love, guidance and prayers of my family is with me in whatever I pursue. Above all I would like to thank almighty God, without his grace nothing is possible.

Contents

1. INTRODUCTION

Next generation sequencing is the most revolutionized exploration of gene expression. Recent technical advances in the genome sequencing have enabled the sequence-based approaches for quantification of transcriptome. Next Generation sequencing is used for transcriptome sequencing inorder to estimate individual gene expression. The reads obtained from deep-sequencing technologies provides us with lot of information, such as from expression-level to the identification of loci with Allele Specific Expression (ASE) [26].

Understanding the genetic basis of variation in gene expression is an interesting topic in the genomic research field. Genetic variants such as Single Nucleotide Polymorphism (SNP), Copy number variation (CNV) and mutation results in phenotypic difference, which are due to the changes in gene expression. One of the most important types among such class of variations is Allele Specific Expression [2, 6]. Thus analyzing gene expression helps us to understand genotypic variation better.

Allele specific expression can be explained as the most preferred expression of one among two alleles in a gene, that is it can be either of two alleles in a gene within an individual (figure 1a and 1b). Accurate measurement of ASE can be obtained by precise mapping of sequence reads towards the reference genome. The major technical hurdle lies in ASE studies while mapping is mapping bias.

1.1.1 Aim

In this thesis work I applied and assessed two existing methods for reducing mapping bias. First, a short read alignment program called GSNAP. It can reduce the read mapping bias towards the reference genome by representing both alleles at genomic positions of known genetic variants. Second, I applied and assessed a method presented by Degner *et al* [5] where the known genetic variants are masked. Finally as a variant of this method I applied 'personalized masking method. In the following, ASE and reference genome mapping bias are explained in more detail.

1.1.2 Biological mechanism responsible for ASE

ASE is typically caused by sequence variation between the two chromosomes within an element that regulates gene expression. If the regulatory element is on the same DNA or RNA molecule as the element it regulates it is said to be *cis*-acting. Heterozygous genetic variation within *cis*-acting elements can cause ASE by three different mechanisms: (i) Differential transcription factor binding (ii) Differential binding of epigenetic factors, and (iii) differential mRNA processing including allele-specific alternative transcription initiation, allele-specific alternative splicing, allele-specific poly-adenylation, and mRNA degradation.

1.1.3 Epigenetic factors responsible for ASE

In an epigenetic context, the mechanisms such as X-inactivation (The process of inactivation of one of the X chromosome copies) and genomic imprinting leads to the silencing of one allele which in turn leads to the other allele being expressed higher than its heterozygous 'copy'. From recent studies it is clear that DNA methylation or nucleotide variation (modification of DNA by the addition of methyl group to the 5 position of cytosine) has a significant role in ASE. For

example, in a recent study on Pediatric Leukemia patients, the sample from bone marrow has validated about 16 % of ASE. From this observation ASE shows a direct quantitative correlation between CpG site methylation which clearly indicates that unequal epigenetic state of haplotype genome contributes towards allelic imbalance in gene expression [23, 28]. In addition to DNA methylation, histone modifications (for example acetylation or methylation of the histone tail or histone positioning) also contribute towards ASE [29].

1.1.4 Why is it interesting?

Genotype and phenotype mapping describes the basics of biological science. Many phenotypic differences results from genetic variants and these are mainly mediated through changes in gene expression. Thus gene expression studies allow better knowledge about genetic variation. ASE gathers research focus basically due to its ability to act as marker for the regulatory variants (such as *cis*, *trans* variants), contribution towards phenotypic variations, how its controls the gene expression level, and how it helps in disease susceptibility. In addition to that ASE can be used to identify expression Quantitative Trait Loci (eQTL), genomic loci that regulates gene expression.

1.1.5 ASE in disease susceptibility

Disease susceptibility can be identified from the heterozygous samples between two groups, e.g. from a sample of affected versus unaffected, providing with information on allele responsible for disease susceptibility from the ratio of allele expressions in the samples [9].

1.1.6 ASE as marker for regulatory variants

Often the gene is under allelic imbalance, that is the expression of allele is not equal then it states that the gene expression is under the *cis*-regulation. In an ASE study, the proportion of mRNA expression level of 2 allele of a heterozygous variant is measured under same cellular environment. In that condition one of the alternate allele acts as a within sample control of each other and hence there is no chance of an external or *Trans* acting elements to influence the gene expression, thus it explains that the gene expression is under control of *cis*-regulation. [24].

1.1.7 ASE to identify eQTL

eQTL can be explained as a genomic loci that regulate the expression of mRNA or protein, it's a type of inherited variance. That is when a genetic mutation regulates gene expressions such genetic variants are known as eQTL. There are two types of primary eQTL, *cis-eQTL* and *trans*-eQTL. If a *cis* allele of gene alters the expression of a allele in a heterozygote gene differently, then it is said to be *cis*-eQTL, otherwise it's a *trans*-eQTL. These eQTLs can only be accurately distinguished by ASE method, because from a total gene expression one cannot separate cis and *trans* eQTL since these two type of eQTL results in similar pattern across a group of individuals [27, 13].

Measurement of gene expression across multiple individuals is a long process, since the varying genomic and environmental conditions of different individuals reduce the statistical power to discover eQTL. In addition to that it is difficult to demonstrate the reliable correlation between SNP allele and gene expression levels when the differences of expression between haplotype are small. ASE provides an alternative method for addressing these limitations.

As an epitome to all the above mentioned reasons ASE is an elegant method of assessing expression within an individual rather than across the subjects hence it avoids major errors [13].

1.1.8 Haplotype

A haploid consists of half the number of chromosomes in somatic cells, diploid consists of two set of homologous chromosome, usually one set from father and another set from mother. Polyploidy consists of more than two pairs of homologous set of chromosomes. Triploid has three set of homologous chromosome and tetraploid consist of namely four set of homologous chromosome.

1.1.9 Phasing

Phasing is a method to distinguish the allele location, that is which of the allele exist together on the same chromosome. Phasing helps to find which genes from the parents which are inherited to the child.

A haplotype refers to DNA derived from a single chromosome.

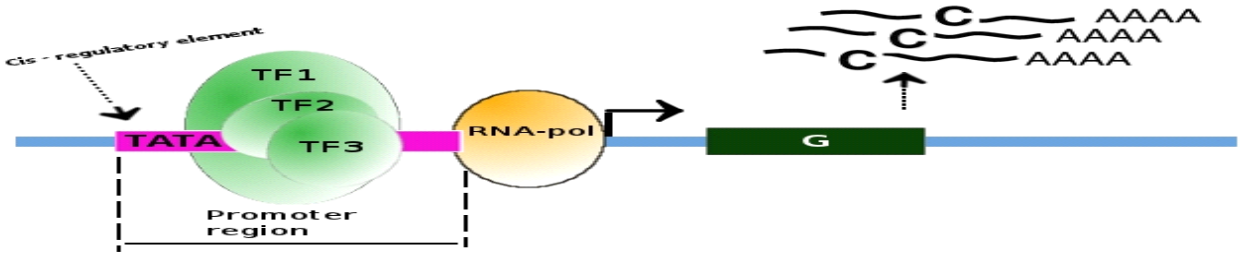


Figure 1a: Allele Specific expression with the paternal haplotype with high expression rate.

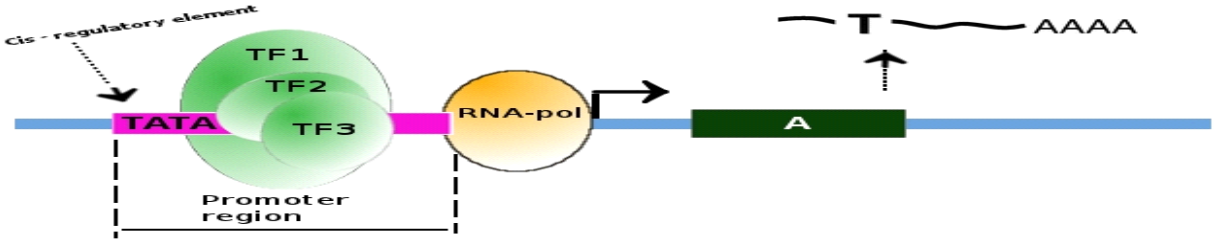


Figure1b: Allele Specific expression with the maternal haplotype with low expression rate

2 ASE MEASUREMENT TECHNIQUES

There are a few methods to measure ASE, in a simple way ASE can be measured by comparing the allelic transcripts within a sample by gene specific RT-PCR, or one can use the most latest next generation sequencing which provides more information such as intronic SNPs. Next generation sequencing is the most efficient tool for genome-wide quantification of DNA and RNA.

ASE measurement such as Allele specific expression PCR, pyro sequencing and Allele specific expression micro arrays has its own technical challenges such as the difficulty in designing the probe that is the probe will be similar for two alleles with a variation on only one location. And these are low-throughput technologies.

Next generation sequencing technology provides a very efficient way of measuring ASE in a very accurate and high throughput manner using the read counts [2].These reads are obtained from individual transcripts via high throughput sequencing. Next generation sequencing is a high throughput technology used for RNA sequencing (RNA-seq) to estimate the expression level of each gene. The first step is to isolate the messenger RNA from the sample and then these are converted to cDNA (complementary DNA) library of fragments with adapters attached to one or both ends. Sequencing these fragments using deep-sequencing technologies produce millions of short reads (figure 2).

These reads obtained can be paired end reads and Single end reads. A paired end read is obtained by sequencing both ends of the same DNA molecule. The two sequences you receive after sequencing both ends are termed as Paired End Sequencing. Single end is obtained by sequencing one end of the DNA molecule.

Reads are mapped towards a reference genome or transcriptome which will return the huge number of reads mapped to the transcriptome. These are read counts, which enables to quantify the transcriptome. Read counts are based on the number of reads mapped to the transcriptome or reference genome, which helps to analyze the expression level of transcripts [25, 26].

ASE studies generally depend on the accurate mapping of short reads towards the reference genome in the presence of sequence variation. RNA sequencing technique provides an efficient way for ASE studies.

2.1 RNA-Seq

RNA seq is the most elegant and powerful method for profiling, discovering and quantification of RNA transcripts. RNA-Seq helps the deep analysis of the transcriptome. Unlike the micro array experiments RNA-Seq provide more accurate measurement of known or unknown transcripts in a wider range. The sequences generated are mapped towards known libraries of exons of known transcripts, which is not present in micro array [7].

The complexity in gene expression level and regulation requires a more sensitive measurement of transcripts and hence to detect its structural abundance. In addition to gene expression RNA-seq discover information about alternative splicing events, ASE, and rare novel transcripts depending on the analysis tools used.

RNA-seq consists of following computation steps such as, read mapping, transcriptome reconstruction, transcript abundance estimation and differential expression analysis. Among the existing methods for ASE analysis RNA-seq is more computationally intense and provides both allelic and total expression data.

In RNA-seq experiment reads are mapped to transcriptome/genome as appropriate followed by counting the number of allele-specific reads that mapped back to heterozygous SNP (The SNP location where both the allele are different). By doing so one of the major analytical challenges here is mapping bias, that is mapped allelic reads will be biased by the allele present in the reference genome (Explained in detail in mapping bias section below) [7, 27].

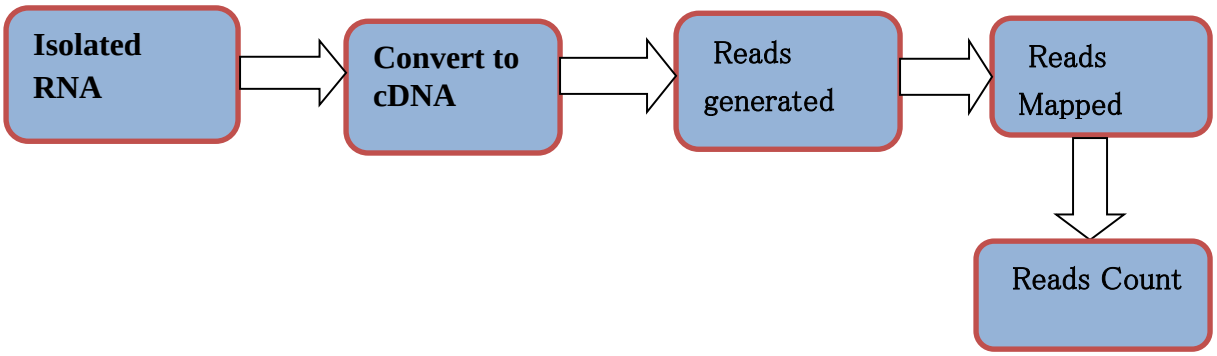


Figure 2: Workflow of RNA-seq, Starts with RNA isolation from a sample, followed by conversion to cDNA by fragmenting and reverse transcription. Reads are generated from a high-throughput sequencer. These reads are further mapped using alignment tools towards transcript set. Counts of the reads mapped to each gene obtained after mapping helps to estimate gene expression level [29].

Paired end reads are generated by sequencing the two ends of a fragment in the library, which will give out two reads. This is known as paired end sequencing. RNA-seq provides simplified data analysis workflow and it is a cost effective high through put technology. Compared to Genomic DNA, which is for genomic variation analysis; RNA-seq is for expression analysis.

2.1.1 Read Mapping

If a species has been sequenced, reads can be aligned, also known as 'mapped', to a reference genome of that species.

2.1.2 Mapping Bias

Achieving a perfect alignment of short reads to a genome is complicated. These difficulties are mainly caused by (i) a read is relatively short, (ii) genetic variation causes a personal sequence to differ from the reference genome and (iii) sequencing errors.

In order to estimate ASE accurately there should be a proper mapping of short reads towards the correct position in the reference genome. The major technical hurdle lies here. Especially when the short reads are mapped towards the reference genome there is a high chance for the reads which carry non-reference alleles to get discarded which in turn gives inaccurate results for ASE. Mapping bias is an inherent issue which occurs while assessing the ASE using sequencing data.

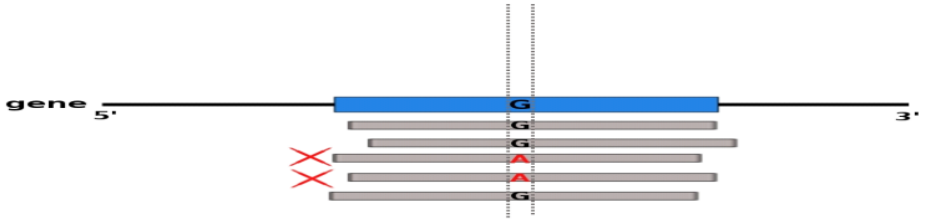


Figure 3: Illustrates the Mapping bias that is when a short read is mapped towards a reference sequence it can be mapped with the identical allele (G) with difference sequence whereas it discards the reads with non-reference allele (A) due to lower mapping quality. As a result, reads identical to the reference genome have a higher chance of being aligned as compared to reads carrying non reference alleles [28].

3. ASSESSED METHODS

In order to study the effects and methods for reducing mapping bias we identified few published methods which are used for reducing mapping bias. From these existing methods I assessed two methods for testing and benchmarking. The tests were done with synthetic simulated RNA-seq data comprising 16 samples.

3.1 Genetic variation masking of the reference genome

Recently there have been a few studies that addressed the ASE analysis with emphasis on reducing mapping bias issues. From the identified published methods *Degneret.al* in the paper "Effect of read mapping biases on detecting allele-specific expression from RNA-sequencing data" explains about the masking of SNP from each allele. They modify the reference genome by masking each SNP locus with a third base for example if A/T is identified in the reference sequence, then change A →G or C at that position.

The study managed to prove that masking of SNP do reduce the systematic bias towards the reference allele. In addition to that simulation studies helped to understand SNP are biased due to the read mapping. Simulations helped in removal of the false positive rate to a large extent [5].

3.1.1 Personalized masking

As a part of the masking we have also done a personalized masking analysis, this was done by masking the genomic variants in reference genome with respect to a single sample. The sample used for this is '1_LPS' and masking procedure is the same as this described above. The variants used correspond to the sample used for mapping.

3.1.2 Short read alignment program that makes use of known genetic variation

One way to reduce the mapping bias towards the reference genome is to use a short read alignment program which is capable of incorporating information about genetic variation. One such program is GSNAP, 'Genomic Short Read Nucleotide Alignment Program' [17]. Another program with a similar capability is Novo align, as was used in a study of ASE by Heap *et al [*13]. But GSNAP was chosen in our analysis since Novo align partly is proprietary software.

GSNAP creates a 'reference space' which is the combination of all major and minor alleles. By mapping reads to such generated 'reference spaces' the program avoids treating minor alleles as mismatches and helps to declare them as true genotypes in the alignment [15].

The authors of GSNAP claim that it is able to align reads against the entire genome even in the presence of substantial polymorphisms and sequence errors [15]. We were therefore interested in the extent to which this feature could alleviate the issue of read mapping bias and thereby improve the performance of allele specific expression analysis.

4. MATERIALS AND METHODS

In this thesis work we applied and assessed two existing methods for reducing read mapping bias towards the reference genome. First, I assessed a short read alignment program called GSNAP. It can reduce the read mapping bias towards the reference genome by representing both alleles at genomic positions of known genetic variants. Second, I applied and assessed a method presented by Degner *et al*. [5] where the reference genome sequence is masked with respect to known genetic variants. To be able to evaluate the performance of these two methods in terms of their ability to reduce the mapping bias, I used simulated synthetic RNA-seq data.

4.1.1 Allele specific synthetic RNA-seq data

Allele specific RNA-seq data had previously been generated by an in-house application and it was used as the benchmarking dataset in this study. The dataset contained synthetic RNA-seq data corresponding to 8 individuals, each subjected to two different conditions (white blood cells, untreated or treated with LPS, 'lipo-polysaccharide'), rendering a total of 16 samples.

Synthetic individual diploid genomes were created by sampling SNPs from the individuals of European ancestry in the 1000 Genome Project (Nov. 2010 release, 629 samples). Expression levels were sampled from phased real RNA-seq data of LPS-treated and untreated white blood cells. Finally, Illumina paired end reads were simulated with a program from the MAQ software

suite using the synthetic individual transcriptome and sampled expression levels. This program also simulates read base sequencing quality using base qualities from real data.

4.2 Application of a read alignment program that makes us of genetic variation

The synthetically generated RNA-seq reads were aligned using GSNAP (downloaded version: 12-06-2012). Apart from read-data three other types of input were supplied to GSNAP: (i) a reference genome (hg19) (ii) known single nucleotide variants and (iii) known splice sites.

Read-data is paired end (PE) fastq file, a total of 32 fastq files were used as input file, by giving each pair at a time. I used cloud computing system (uppmax cloud computer system) to run the command for the 32 files (Paired end 16 samples) with the aid of SHELL scripts jobs were submitted to the super computer.

4.2.1 GSNAP input: Known single nucleotide variants

As known SNVs I used the common variants from the database dbSNP, build 135. This includes SNVs with a MAF >1% in the global population where MAF (minor allele frequency) is the frequency at which the less common allele (the minor allele) occurs in a population.

4.2.2 GSNAP input: Known splice sites

Inputting known splice sites helps GSNAP to correctly align across intron-exon boundaries. GSNAP has a utility program 'gtf_splicesites' which creates a splice site index using a GTF file containing known exon positions. I used known exons from Ensemble version 59 as input 'Homo_sapiens.GRCh37.59.gtf'.

4.2.3 Read mapping

The synthetic reads were mapped with GSNAP to the human reference genome (hg19) using default parameters as well as piped with samtools for BAM output (detailed description of parameters is shown in Appendix section).

Initially I started with parameters such as -D reference directory, -d reference file, -A sam, -s splicesite, -v snpfile, -V snpdir, and quality parameter --quality-protocol=illumina which is equal to pred-scaled probability score 64. But majority of BAM output was with multi-mapped reads and was not able to perform further steps from PCR duplicate removal. In order to discard multi mapping of reads I set the quality parameter (--npaths=1 --quiet-if-excessive) and ran the mapping program. But this didn't help for all samples hence still there existed certain reads (reads from 19 BAM files) mapped to multiple loci of the reference genome.

4.2.4 Multi-mapped reads

Multi mapped reads are the reads which are mapped to multiple loci of the reference genome. Despite setting the unique-alignment flag of GSNAP, several multi-mapped reads persisted. A custom-Perl script was therefore implemented to remove remaining multi-mapped reads.

4.2.5 Processing of aligned reads to retrieve allele specific read counts

After alignment of reads to the reference genome and manual removal of multi-mapped reads

from the 19 BAM files, a number of computational steps were performed as to retrieve counts of each of the two alleles at variants called as heterozygous. The following steps describe the computational steps which we followed to retrieve the allele counts (Flow Chart 1a).

4.2.6 PCR-duplicate removal

To remove PCR duplicates I applied Picard Mark Duplicates to the aligned reads. This helped to filter out the duplicate reads that were identified in the paired end reads. Then the next step is to merge all the BAM files from each sample into a single BAM file this was accomplished by samtools merge. And then these BAM files were sorted, Samtools sort the BAM file based on the position in the reference genome. In order to allow a fast look up of the sorted BAM file we indexed them with Samtools index. This helps the program to read data and work with data in associated files more efficiently.

4.2.7 Variant calling

To find sequence variants we did variation calling with Samtools mpileup and bcftools was piped to get the output in VCF format [19]. Samtools mpileup helps the data storage in VCF format (Explained in detail in the file format section below). Firstly a list of all BAM files was created. Then it is needed to create a region file from the reference genome with the information of all genomic regions (that is the chromosome name its genomic region in the reference sequence) so we indexed the reference genome with Samtools faidx. Thereafter we extracted the variant information into VCF files for each of the BAM files (the command used here is explained in detail in appendix section command 2). Then we extracted data from all VCF files and a single list was created for further steps.

4.2.8 Variant filtering

Allele specific expression analysis can only be done using heterozygous variants as to be able to distinguish the two alleles. Furthermore, a certain read depth is required both for the variant calling to be reliable as well as to be able to reach statistical significance with respect to whether two alleles have different expression levels. We therefore extracted heterozygous variants with a minimum read depth of 10.

4.2.9 Allele counts

Samtools mpileup outputs number of reads mapped to each of the two alleles (allele counts). However, since the variant calling was performed using all samples as input as to get as high reliability as possible of the called variants, we reran samtools mpileup for each specific sample using the called variants as input. In this manner the allelic counts for the called variants were retrieved (the command used here is described in appendix section command 3).

4.2.10 Coverage calculation

To determine how many reads are mapped towards each base in reference genome we calculated coverage. In order to calculate the mapped read coverage on the exons we used bedtools. This provides the amount of the exome which is covered by at least *n* reads (read depth). For example, 80% of the exome may be covered with a read depth of at least 10. As exon annotation CCDS (Consensus Coding DNA Sequence) was used, which provides a complete set of annotation for the protein coding regions (exons) of the human genome [21]. The command used here is

described in appendix section (command 4). And therefore we calculated the coverage obtained from each sample, the formula used here is,

$$\textbf{\textit{Coverage}} = \frac{Readdepth * BasesCovered}{ExonLength}$$

Where depth read represents the reads covered at each position of the sample and it is an integer value. A base covered represents the length of bases covered in the exome. Exon length is the length of the exon.

4.2.11 Variant annotation

To annotate the called variants we used a custom PERL script. This program annotates variants with respect to various annotation databases, such as presence in dbSNP and within which gene a variant is located. And therefore we merged the annotated variants and converted it into RData structure for further statistical analysis.

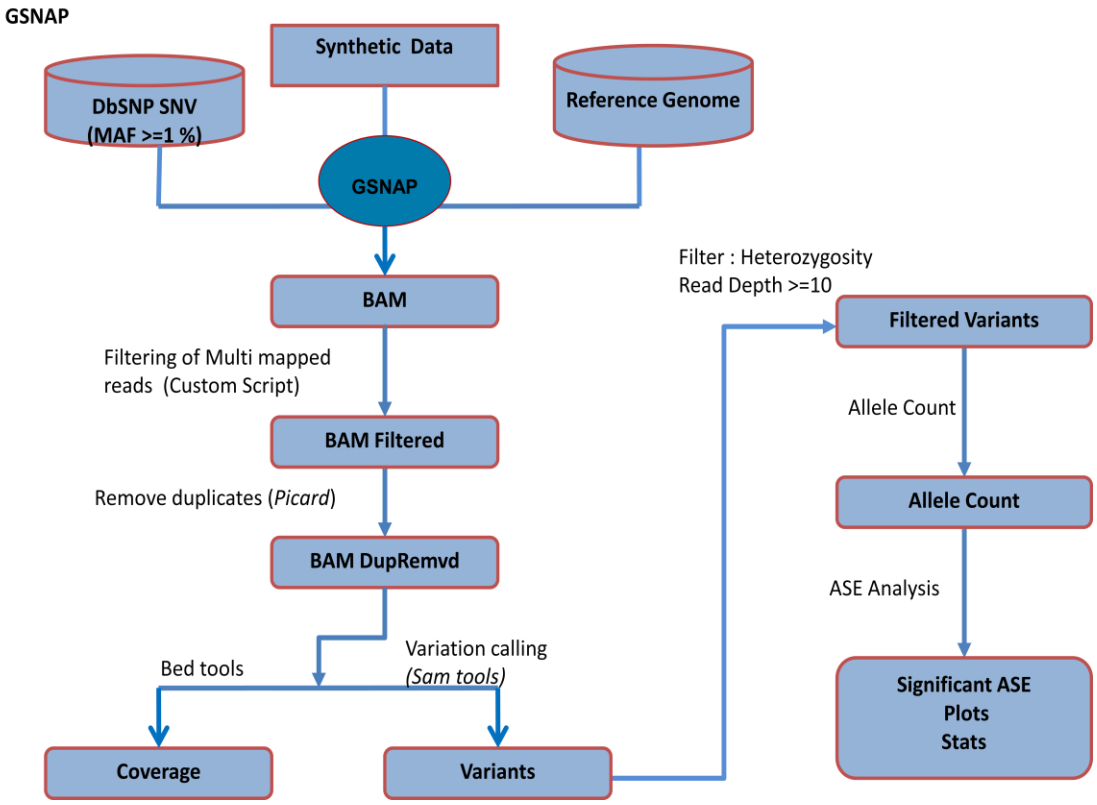
4.2.12 ASE analysis

The values obtained from the above methods are used for further ASE analysis. Statistical tests such as two-sided binomial test and multiple corrections of these tests were done on this data for plotting histogram for the mapping bias with minimal error rate. Read depth of alternative allele is taken from data obtained by allele counts and variant calling methods (described the materials and methods sections above). Alternative allele fraction represents the allele towards the alternative allele not in reference genome.

First started by plotting smoothed histograms for all alternative allele fractions, and then filtered the variants based on relative frequencies of alternative allele direction and significant P-value <=0.05.For both of them we plotted graphs (shown in results section).

Furthermore we checked the number of variants obtained before and after annotation.
In addition to that we analyzed the genes with multiple significant values for ASE within genes with minimum of 2 heterozygous variants. And then filtered all variants within the gene to make ASE significant using multiple corrected P-value.

In order to find out the occurrence of alternative allele within a gene we filtered on number of significant alternative alleles per gene when number of sample is >=1 as well as number of sample >=2 (Numerical values is shown in the results section).



Flow Chart 1a: Figure explains the pipeline I followed for the GSNAP method. The input used here is the Single Nucleotide Variant (SNV) database with MAF>=1% and the reference genome.

4.3 Genetic variation masking of the reference genome

To avoid mapping bias at positions where an individual genome may differ from the reference genome we masked known genetic variants from the reference genome (version hg19) [5]. The known genetic variants that were masked were the same European SNPs retrieved from the 1000 Genomes Project which is used to construct the synthetic RNA-seq data (see below). Masking was accomplished by changing the reference allele to a third allele that was neither the reference allele nor the alternative allele.

4.3.1 Read mapping to the masked reference genome

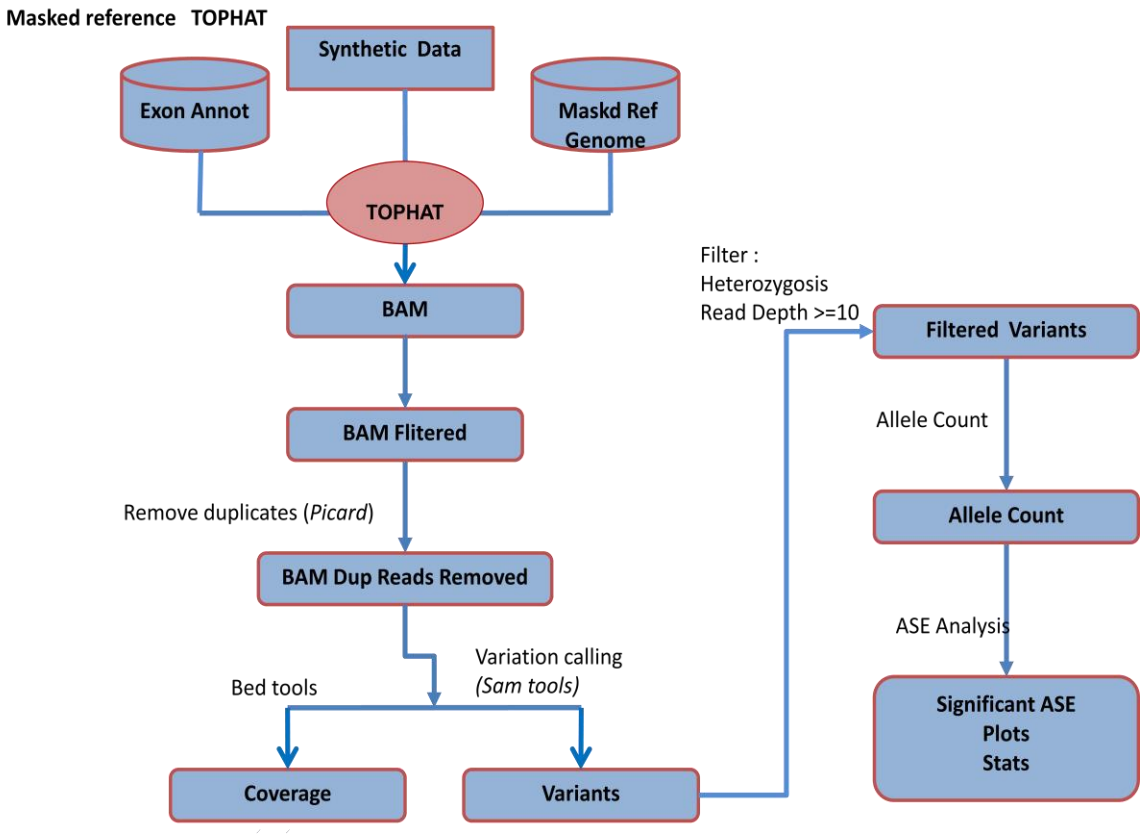
Alignment of sequencing reads to the masked reference genome was done with the RNA-seq alignment tool Tophat (version 1.4.0) [22]. Tophat uses the Bowtie aligner and we therefore Bowtie-indexed the masked reference genome with Bowtie, version 0.12.6.

The input given to Tophat was the synthetic RNA-seq reads the masked reference genome and an exon annotation file (ensemble version 59) (*flowchart 1b*).
Tophat has a required parameter which informs about the mean fragment length and this was set

to the mean fragment length used to generate the synthetic reads.

Tophat uses 'inner distance' which can be obtained by 'insert length -2*read length' where the insert length were different for each of the fastq files since the reads were fragmented by ultrasonic waves. The insert length is calculated by the bioanlayser plots and each fragment length varies from the other one. Read length is 100 for all samples used.

The Tophat command used in the analysis is described below in the appendix section. Tophat provides the output reads in BAM file format after mapping towards reference genome. Further steps are similar from the above method towards the results (section 2).



Flowchart 1 b: Explains the workflow I followed for the 'masked reference' method. Here in this method the inputs were the Exon annotation database and the masked reference genome, the mapping tool used is Tophat.

4.3.2 Personalized masking

As a part of the masking we have also done a personalized masking analysis. This was done by masking the genomic variants in the reference genome with respect to a single sample. The sample used for here is '1_LPS' and masking procedure is same as above.

Further steps are same as the above method.

5. RESULTS

Allele specific Expression analysis was performed on the Synthetic RNA –seq data from 8 individuals, a total of 16 samples. Each of these 16 samples was mapped towards the reference genome and transcriptome with TOPHAT (Masked reference and personalized masked reference) and GSNAP alignment program.

For all the methods the read coverage is calculated. Read coverage obtained from the GSNAP method shows a better coverage than the masked reference method and personalized masking method. The GSNAP method as described it assigns even minor mismatches as matches (due to reference space) and hence results in a better coverage. In the other two methods the genomic variants are masked, which results in relatively lower coverage. The table showing coverage obtained from all methods are shown below (table1a, table 1b, table 1c).

5.1 Read coverage for three methods is as following:

5.1.1 GSNAP: Short read alignment program and masked reference

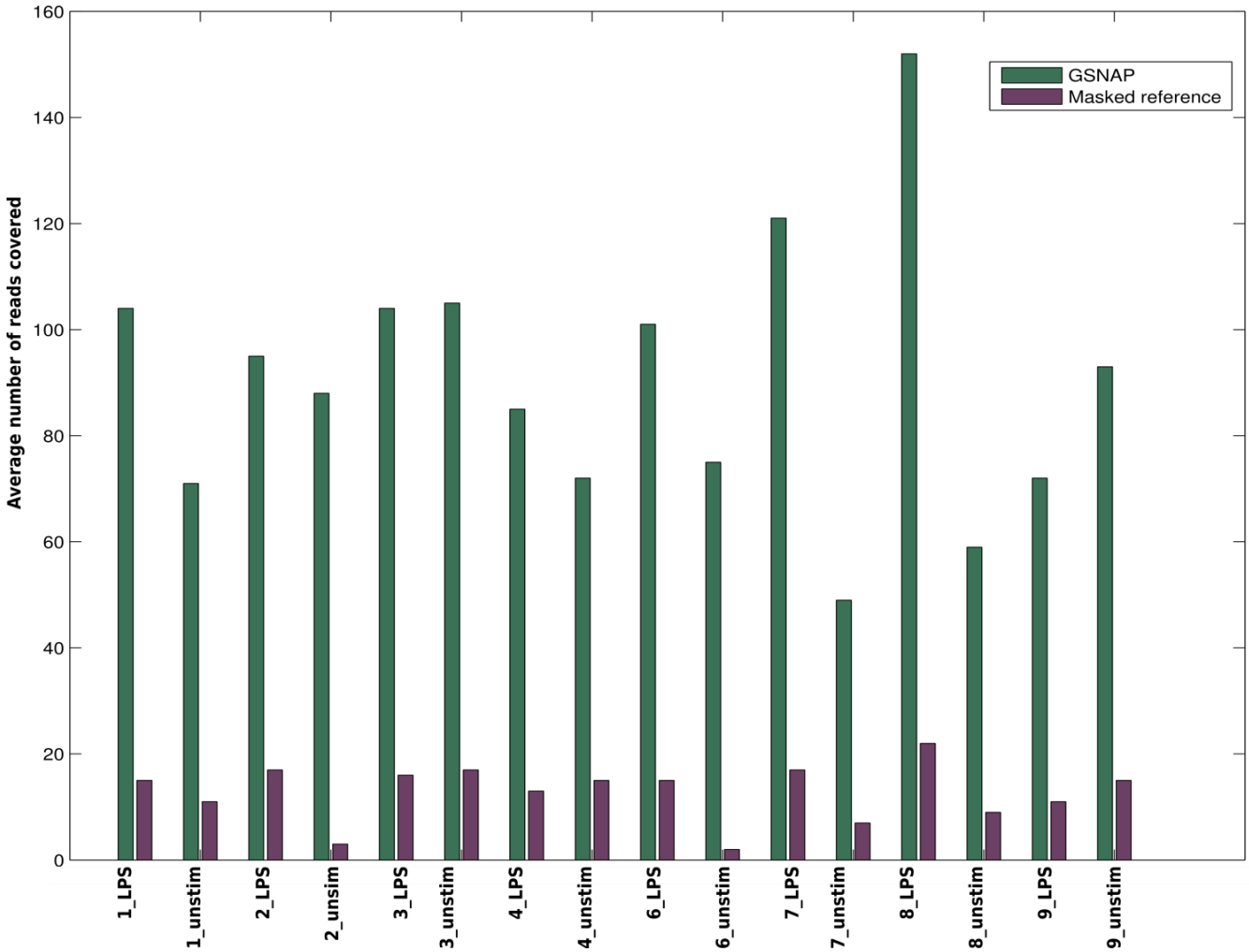


Figure 1a: Give the read coverage obtained from the 'masked reference' and GNSAP method.

5.1.2 Personalized masked reference

Sample	1_LPS
Coverage	11

Table 1c: Give the coverage obtained from personalized masking method for a single sample.

5.2 Mapping Bias for all Variants

5.2.1 GSNAP: Short read alignment program

Histograms (smoothed) of ASE were plotted for all methods. In the histogram we considered only alternate allele fraction from all heterozygous variants with a minimum read depth of 10. However both the methods (short read alignment and masked reference) still shows a certain amount of mapping bias. We have used two-sided binomial test here, and hence any heterozygous variants which have any deviation from 50 % are considered to be bias and in the histograms we can observe deviation from that, hence it shows that mapping bias still preexist. That is if there is no mapping bias existed the mean of these probability distribution should be on 0.5, but here we can see the mean is smaller so there is mapping bias existing toward the non-reference allele direction.

Whereas the histogram for mapping bias obtained from personalized masking method has shown a reduction in mapping bias compared to other two methods. This is mainly due to its genomic variants masked in reference genome. The read mapping bias at the SNP position in reference genome is due to the occurrence of a flanking sequence sharing sequence identity in other region of reference genome. With masking both the reference allele and alternate allele makes one base pair mismatch at correct location, but either allele will map to the corresponding position in the alternative location thereby biasing against correct mapping of allele that matches elsewhere [5].

In the histogram the x-axis represents allele fractions, which means the number of reads within samples has alternate allele. Y-axis represents the estimated probability density from the number of occurrence of reads.

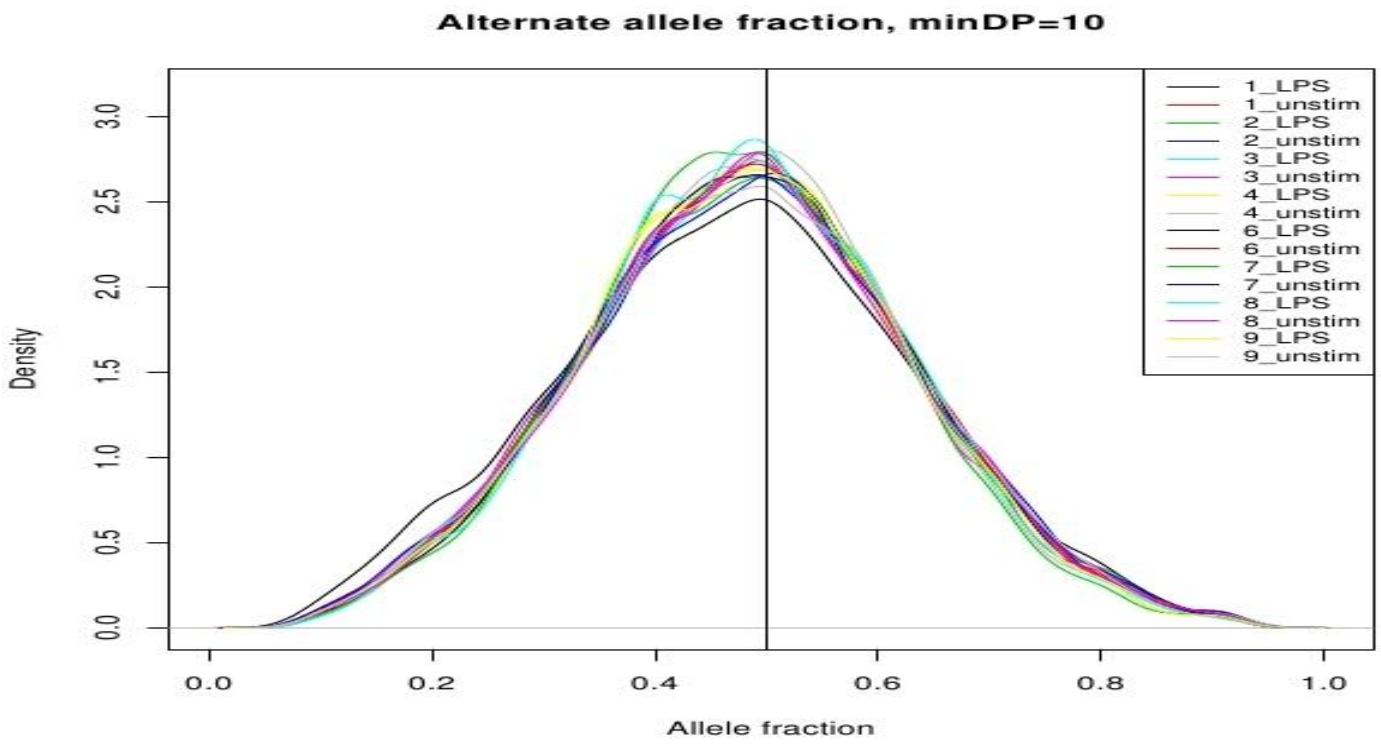


Figure 5 a: Smoothed histogram of alternative allele fraction for all variants from the GSNAP method. The histogram shows the mapping bias still exists.

5.2.2 Mean and Median

In order to get the numerical value of mapping bias we further calculated the mean and median obtained for each sample from the histogram. The following table defines the mean and median obtained from GSNAP method.

	GSNAP	
Sample	**Mean**	**Median**
1_LPS	0.467	0.467
1_unstim	0.479	0.479
2_LPS	0.478	0.478
2_unstim	0.477	0.477
3_LPS	0.476	0.476
3_unstim	0.476	0.476
4_LPS	0.476	0.476

4_unstim	0.477	0.477
6_LPS	0.477	0.477
6_unstim	0.471	0.477
7_LPS	0.477	0.471
7_unstim	0.476	0.476
8_LPS	0.476	0.476
8_unstim	0.474	0.476
9_LPS	0.470	0.473
9_unstim	0.478	0.478
Average	**0.477**	**0.477**

Table 2a: Shows the mean and median obtained from short reads alignment program. Obtained mean is 0.477 and median is 0.477.

5.2.3 Masked Reference

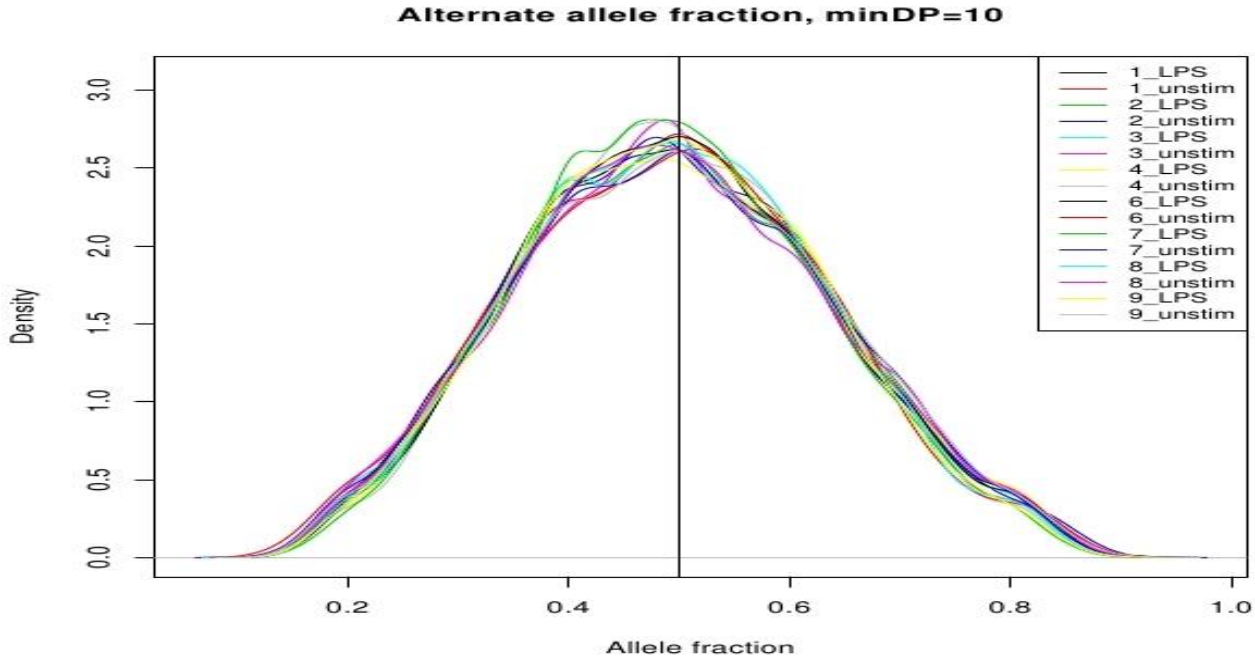


Figure 5 b: Smoothed histogram of alternative allele fraction for all variants from the masked

reference method.

5.2.4 Mean and Median

	Degner	
Sample	**Mean**	**Median**
LPS1	0.49238309	0.4893617
LPS2	0.49534388	0.5
LPS3	0.49346107	0.49324324
LPS4	0.49680817	0.49612403
LPS6	0.49702571	0.5
LPS7	0.49192292	0.4893617
LPS8	0.49329803	0.49555547
LPS9	0.49246006	0.5
UNSTIM1	0.49331384	0.5
UNSTIM2	0.49484297	0.49036044
UNSTIM3	0.49641828	0.49525971
UNSTIM4	0.49503177	0.49473684
UNSTIM6	0.49034261	0.49557522
UNSTIM7	0.4948537	0.5
UNSTIM8	0.49412747	0.49190633
UNSTIM9	0.49724266	0.5
Average	**0.49430476**	**0.49571779**

Table 2b: Table shows the mean median value obtained from masked reference method. The average mean value from all samples is 0.49, and median is 0.5

5.2.5 Personalized masked reference

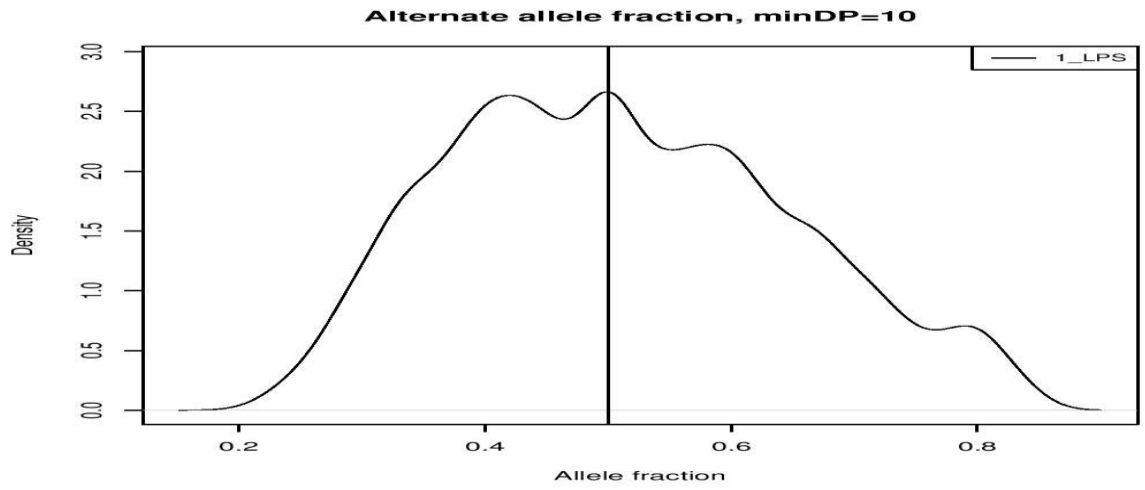


Figure 5 b: Smoothed histogram of alternative allele fraction for all variants from the personalized masked reference method

5.3 Mapping bias of variants Shown significant ASE

To get a more precise picture of mapping bias we filtered out the significant variants and plotted the histogram for those. From the obtained histogram, it shows that mapping bias persist in short mapping method and masked reference method. But it shows a reduction in personalized masking method. Histogram from all methods is plotted as follows. The peaks in left side indicate that majority of the ASE variants has expression shifted towards reference allele (figure 6a, 6b, 6c).

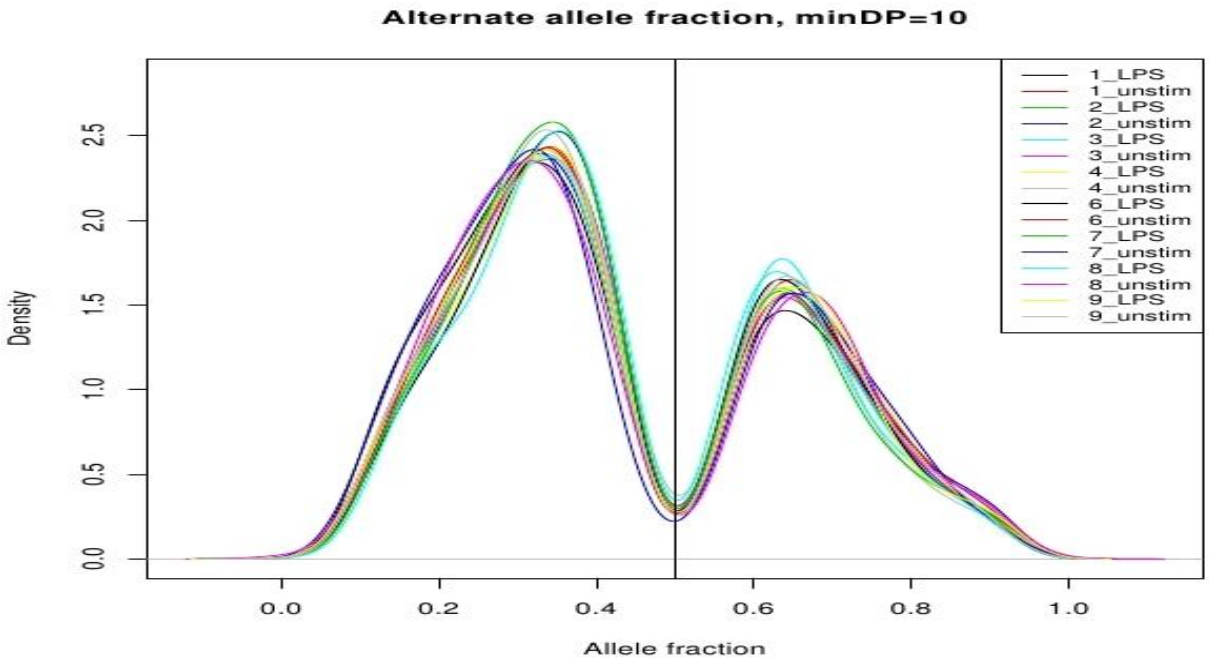


Figure 6 a: Smoothed histogram shows the alternative allele fraction for variants showing on 5% level of ASE. This Histogram Obtained from GSNAP method

5.2.3 Masked Reference

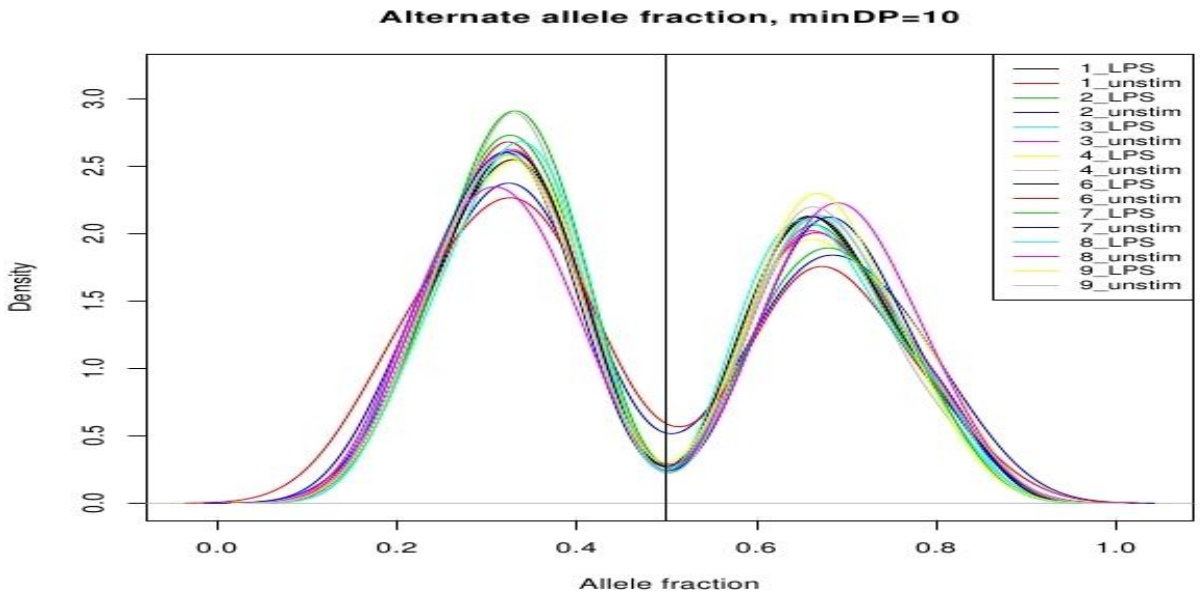


Figure 6 b: Smoothed Histogram shows the alternative allele fraction for variants showing 5% level of ASE.

5.2.5 Personalized masked reference

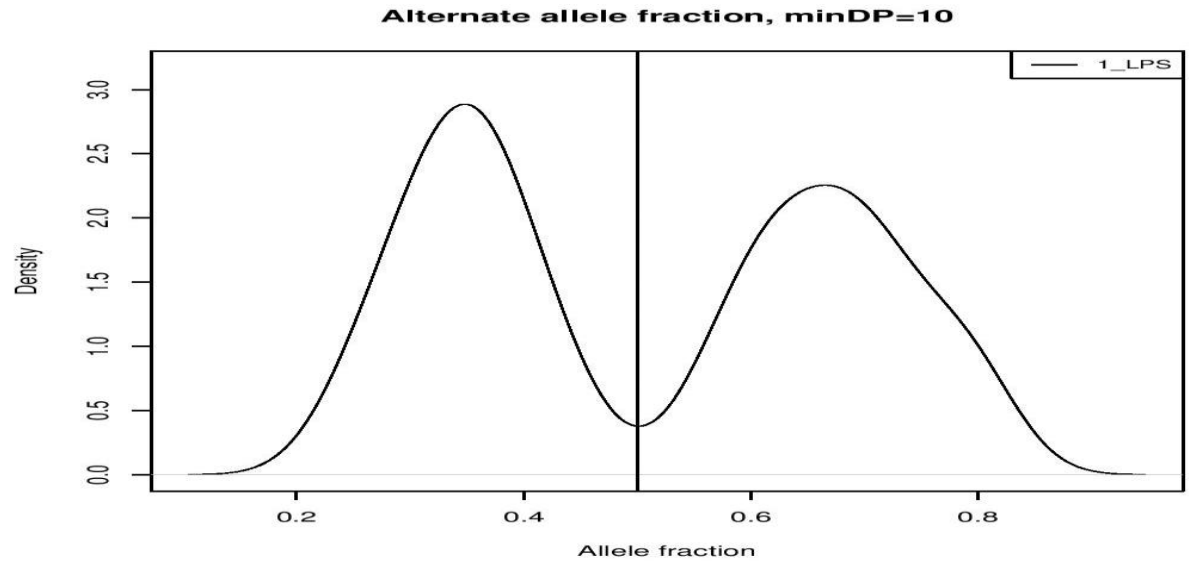


Figure 6c: Smoothed histogram shows the alternative allele fraction for variants showing 5 % level of ASE.

5.4 Allele Specific Analysis

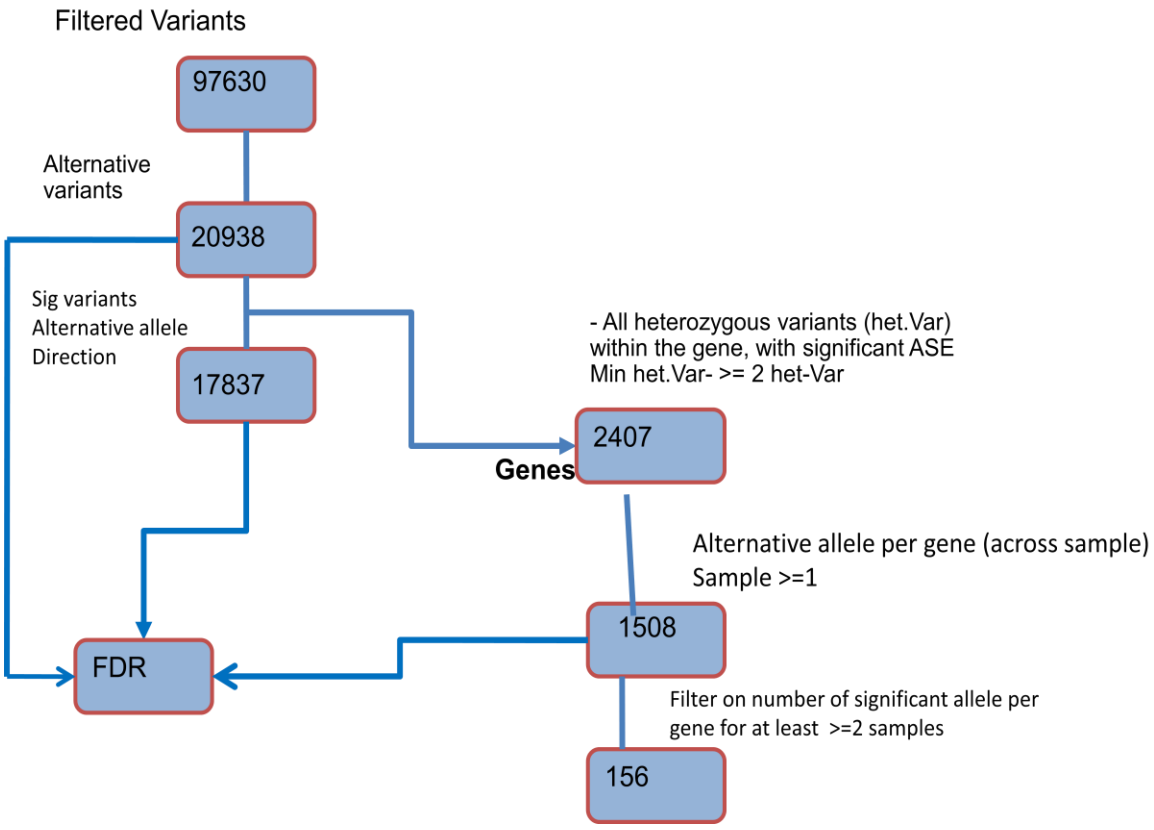
5.4.1 GSNAP: Short sequence alignment

In order to estimate the differently expressed allele within a gene we have followed some filters to all variants. Initially, we have taken the numerical value of all variants from all methods. Then we have filtered out the significant variants and calculated its value.

For GSNAP method we measured the subset of SNP from the both the reference genome and the dbSNP used for mapping, we have found around 80 % of similarity between both. In addition to that for both methods we have taken the value for alternate variants and calculated the value for significant variants towards alternative allele direction. In GSNAP the difference between significant variants and significant variants towards alternative direction is very little.

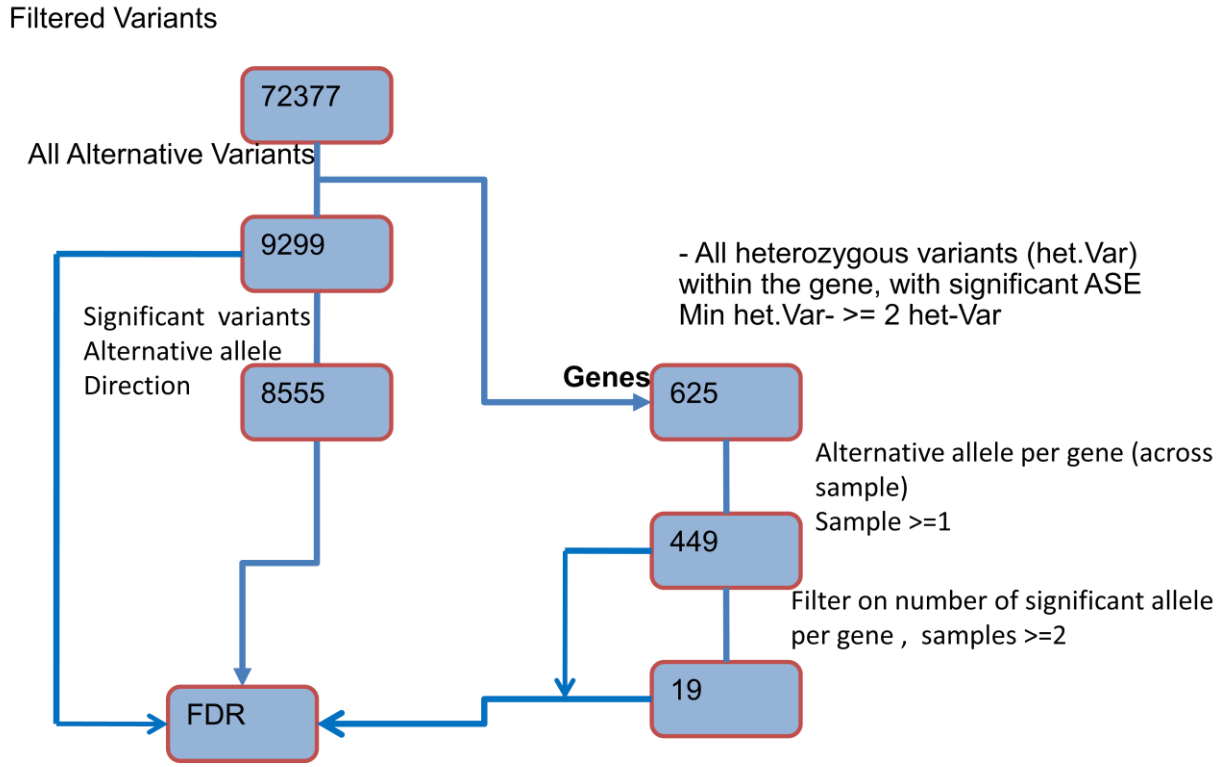
Whereas in masked reference method the difference is quite large, this could be mainly because of the masking of genomic variants in reference genome. And hence forth the variants towards alternative allele direction are comparatively less. Then FDR (False Discover Rate) has taken for both methods.

Further filtering was done for the fraction of significant ASE within the gene with a minimum of 2 heterozygous variants. The further filter is on the alternative allele expressed per gene, that is the common gene across a sample and the value is taken. When the sample is at least 2 the amount of allele expressed is quite low (flowchart 2a and 2b).



Flowchart 2a: Shows the numerical values obtained by GSNAP method for all variants and significant variants from the mapping bias estimation.

5.4.2 Masked Reference



Flowchart 2b: Shows the numerical values obtained by masked reference method for all variants and significant variants from the mapping bias estimation.

5.4.3 Personalized masking

We obtained a total of 5403variants after mapping with personalized masked reference genome. The amount for all alternative variants obtained is 706. Similarly we observed 1404 significant variants towards alternative allele direction. We found 62 significantly expressing ASE in a gene while filtering on fraction of significant ASE in gene. On other hand while filtering on the number of significant alternate allele per genes we got 43, while the number sample is 1 and 0 when more than one sample.

5.4.4 FDR Estimation

In order to increase the power of statistical tests in ASE analysis False Discovery Rate (FDR) for genes and variants with ASE are estimated.

For all methods both methods with respect to a true synthetic RNA-seq data set consisting of 16 samples which closely reflects the natural RNA –seq data and observed ASE expression. We had a set of expression from True RNA-seq data set so we used it as true positives and calculated the FDR and sensitivity. Both FDR and sensitivity values are calculated here for allele towards alternative allele

direction, before filtering and after filtering. From the three methods the personalized masking method gives better FDR value. In addition to that both FDR and sensitivity values are calculated when the number of sample is 1 and 2 (table 3a, table 3b, and table 4a and 4b).

$$Sensitivity = \frac{True\ Positive(TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)}$$

$$FDR = \frac{FalsePositive}{FalsePositive + TruePositive}$$

Both FDR and sensitivity is calculated in the following sections.

FDR

VARIANTS	GSNAP	Personalized masking	Masked Reference
Before Filtering	7.4 %	3.5%	1.5 %
After Filtering	6.3 %	0.09%	2.5 %

Table 3a: Defines the Sensitivity of variants before and after filtering.

GENES	GSNAP	Personalized masking	Masked Reference
When Number of Sample >=1	19 %	2.8%	23 %
When Number of Sample >=2	46 %	0	31 %

Table 3b: Defines the Sensitivity of genes when sample number is equal to or greater than 1 and 2 respectively.

Sensitivity

VARIANTS	GSNAP	Masked Reference	Personalized Masking	
Before Filtering	2.07%	9.9%	8.13%	
After Filtering	1.24 %	6.2 %	4.8%	

Table 4a: Shows the specificity for variants before and after filtering

GENES	GSNAP	Personalized masking	Masked Reference
When Number of Sample >=1	9.6%	0.16%	16%
When Number of Sample >=2	1%	0	2.7%

Table 4b: Shows the specificity for the Genes with sample number greater than and equal to 1 and 2 respectively.

6. DISCUSSION

In this thesis I have analyzed three different methods for mapping bias reduction for benchmarking the efficient one. From the results it is clear that mapping bias still preexists, none of the methods reduced mapping bias fairly which highlights the importance to develop new method for mapping bias reduction.

In the short sequence alignment program, initially there were a lot of technical issues such as the tool doesn't have any flags to handle multi-mapped reads and the tool wasn't able to handle the file size, so it was required to subdivide the input fastq file before mapping. However the coverage obtained is higher than the masked methods; this is mainly because the tool can handle minor allele as matches with help of 'reference spaces' (explained in methods section).

Furthermore by the simulated synthetic reads we were able to identify and remove a large number of false positives. The FDR value obtained is really good from all the methods. The study shows that the read coverage obtained from both masking methods (Degner and personalized masking) is very low, since all the genomic variants were masked.

RNA- seq is a very powerful tool for ASE studies, but we need to account for mapping bias. The thesis highlights the importance for developing tools for read mapping tools that can in cooperate multiple SNPs, particularly in the context of identification of ASE. Although there are a few tools like GSNAP has the ability to incorporate the variants, but these tools have limitations for regions with multiple SNPs in close proximity and thus it increases the search space increases rapidly. If you have a lot of variations there then it creates lot of difficulties in accurate mapping of these regions.

And also we need to be careful about what reference we are using while mapping if we are

studying about ASE. For example we can customize the reference genome according to the requirement such as masked reference, personalized masking or personalized enhanced masking, by adding alternate allele at SNP position.

Each and every human genome has its own Novel SNP/variants which are not found in the standard DbSNPs. Thus a 'Two way approach' can be implemented for addressing mapping bias in ASE study. Mostly bias is at SNP regions of the reference genome (i.e. there is no alternate allele), first approach could be creation of an enhanced reference genome, i.e. by incorporating all alternate alleles (from existing dbSNPs) to the exiting reference genome. Then map our short to that enhanced reference genome. Thus we are able to know the Novel SNPs and then incorporate these Novel SNP also to the enhanced reference genome, further use this reference genome for mapping this could also helps in reducing mapping bias (Ravi satya et.al).

In addition to that the study shows the need to design different statistical models for variability from multiple SNPs in gene, for example the selection of models such as Bayesian models which can in turn return good true positive as you are not biased by small and very consistent difference in SNP all over the gene.
Advance studies in ASE will definitely contribute towards clinical benefits especially, for example in the area of diagnosis, for instance by exploring ASE it helps in tumor type evaluation and classification [24]. Increasing the length of reads may help in mapping reads accurately to the reference genome.

7. REFERENCE

[1] R. Palacios, E. Gazave, J. Goni, G. Piedrafita, O. Fernando, A. Navarro, and P. Villoslada, “Allele-specific gene expression is widespread across the genome and biological processes,” *PLoS One,* vol. 4, no. 1, pp. e4150, 2009.

[2] B. J. Main, R. D. Bickel, L. M. McIntyre, R. M. Graze, P. P. Calabrese, and S. V. Nuzhdin, “Allele-specific expression assays using Solexa,” *BMC Genomics,* vol. 10, pp. 422, 2009.

[3] W. W. Wasserman, and A. Sandelin, “Applied bioinformatics for the identification of regulatory elements,” *Nat Rev Genet,* vol. 5, no. 4, pp. 276-87, Apr, 2004.

[4] A. P. Joseph, N. Srinivasan, and A. G. de Brevern, “Cis-trans peptide variations in structurally similar proteins,” *Amino Acids,* vol. 43, no. 3, pp. 1369-81, Sep, 2012.

[5] J. F. Degner, J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard, “Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data,” *Bioinformatics,* vol. 25, no. 24, pp. 3207-12, Dec 15, 2009.

[6] R. Liu, A. T. Maia, R. Russell, C. Caldas, B. A. Ponder, and M. E. Ritchie, “Allele-specific expression analysis methods for high-density SNP microarray data,” *Bioinformatics,* vol. 28, no. 8, pp. 1102-8, Apr 15, 2012.

[7] Zheng Jin Tu, "Getting Started with Illumina Genome Analyzer Data Analysis", *Computational Genetics Lab, University of Minnesota, ppt,* April 19 2011.

[8] Chris Park, Devan Correll, and Paul Oeth, "Measuring Allele-Specific Expression Using MassARRAY"**,** *John Hopkins Court, pp. 2-4,* September 8 2004.

[9] T. Pastinen, “Genome-wide allele-specific analysis: insights into regulatory variation,” *Nat Rev Genet,* vol. 11, no. 8, pp. 533-8, Aug, 2010.

[10] P. R. Buckland, “Allele-specific gene expression differences in humans,” *Hum Mol Genet,* vol. 13 Spec No 2, pp. R255-60, Oct 1, 2004.

[11] R. Tewhey, V. Bansal, A. Torkamani, E. J. Topol, and N. J. Schork, “The importance of phase information for human genomics,” *Nat Rev Genet,* vol. 12, no. 3, pp. 215-23, Mar, 2011.

[12] B. Tycko, “Mapping allele-specific DNA methylation: a new tool for maximizing information from GWAS,” *Am J Hum Genet,* vol. 86, no. 2, pp. 109-12, Feb 12, 2010.

[13] G. A. Heap, J. H. Yang, K. Downes, B. C. Healy, K. A. Hunt, N. Bockett, L. Franke, P. C. Dubois, C. A. Mein, R. J. Dobson, T. J. Albert, M. J. Rodesch, D. G. Clayton, J. A. Todd, D. A. van Heel, and V. Plagnol, “Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing,” *Hum Mol Genet,* vol. 19, no. 1, pp. 122-34, Jan 1, 2010.

[14] National Center for Biotechnology Information*, U.S National centre for Library,* http://www.ncbi.nlm.nih.gov/projects/SNP/docs/build135_announce.txt, 10.13.11

[15] T. D. Wu, and S. Nacu, “Fast and SNP-tolerant detection of complex variants and splicing in short reads,” *Bioinformatics,* vol. 26, no. 7, pp. 873-81, Apr 1, 2010.

[16] S. B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, and E. T. Dermitzakis, “Transcriptome genetics using second generation sequencing in a Caucasian population,” *Nature,* vol. 464, no. 7289, pp. 773-7, Apr 1, 2010.

[17] T. D. Wu, and C. K. Watanabe, “GMAP: a genomic mapping and alignment program for mRNA and EST sequences,” *Bioinformatics,* vol. 21, no. 9, pp. 1859-75, May 1, 2005.

[18] Genomes project publishes inventory of human genetic variation *1000 genome* http://www.1000genomes.org/node/10119, February 2011.

[19] Samtools, *sourceforce.net http*://samtools.sourceforge.net/mpileup.html *, December 2012.*

[20] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M. M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B. L. Maidak, J. Mudge, M. R. Murphy, T. Murphy, J. Rajan, B. Rajput, L. D. Riddick, C. Snow, C. Steward, D. Webb, J. A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, and D. Lipman, “The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes,” *Genome Res,* vol. 19, no. 7, pp. 1316-23, Jul, 2009.

[21] Health centre for genetic education, *NSW Government*, http://www.genetics.edu.au

[22] S. Campino, J. Forton, S. Raj, B. Mohr, S. Auburn, A. Fry, V. D. Mangano, C. Vandiedonck, A. Richardson, K. Rockett, T. G. Clark, and D. P. Kwiatkowski, “Validating discovered Cis-acting regulatory genetic variants: application of an allele specific expression approach to HapMap populations,” *PLoS One,* vol. 3, no. 12, pp. e4105, 2008.

[23] E. Turro, S. Y. Su, A. Goncalves, L. J. Coin, S. Richardson, and A. Lewin, “Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads,” *Genome Biol,* vol. 12, no. 2, pp. R13, 2011.

[24] R. V. Satya, N. Zavaljevski, and J. Reifman, “A new strategy to reduce allelic bias in RNA-Seq readmapping,” *Nucleic Acids Res,* vol. 40, no. 16, pp. e127, Sep, 2012.

[25] Wei Sun, Yijuan Hu, "eQTL mapping using RNA-seq data”, Department of Genetics, *Carolina Center of Genome Science*, pp.10-14, December 2012.

[26] Christopher G Bell,Stephan Beck, *Institute*, "Advances in the identification and analysis of allele-specific expression", *Medical Genomics*, *University College London Cancer*, *Genome Biomedicine*, pp.4-6, May 29 2009.

[27] C. G. Bell, and S. Beck, “Advances in the identification and analysis of allele-specific expression,” *Genome Med,* vol. 1, no. 5, pp. 56, 2009.

[28] B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey, “RNA-Seq gene expression estimation with read mapping uncertainty,” *Bioinformatics,* vol. 26, no. 4, pp. 493-500, Feb 15, 2010.

[29] Samtools, *sourceforce.net http*://samtools.sourceforge.net/mpileup.html *, December 2012.*

APPENDIX 1

1. Commands used

Command 1:

```
gsnap -D  refdir -d  ref  -A sam -s  splicefile -V  snpdir -v  snpfile –quality protocol=illumina  --npaths=1  -Q --quiet-if-excessive fastqfile .readp_1  fastqfile.readp_2 | samtools view -Sb ->outdir}/${sbatchfile}.bam
```

Command 2:

```
Samtools mpileup -q1 -d10000 -L10000 -DSugfref  -r  region -b  bamfilelist | bcftools  view  -vcg - | vcfutils.pl  varFilter>vcffile
```

Command 3:

```
Samtools mpileup -q1 -d10000 -L10000 –Dsugfreferencegenome -l 1_LPS.merged.hetvars 1_LPS.merged.bam | bcftools view - >*.mpileup.nocall.vcf
```

Command 4:

```
coverageBed –abam bamfile  -b  bed  -hist -split >ccds.bedtools.out2

coverageBed –abam bamfile  -b gtf  -hist –split >ensembl.bedtools.out2

genomeCoverageBed –ibam bamfile -g genome -split -max 1000 >genome.bedtools.out2
```

Command 5:

```
tophat --solexa1.3-quals -p   threads .  -o outdir –GTF gtffile  isize  --mate-std-dev  isizedev referencegenome  fastqfile_1   fastqfile_2
```

APPENDIX 2

2. FILE FORMATS

2.1. File formats in the analysis

Information about read sequences generated from a sequencing machine can be stored in a file format called ‘fastq’.

2.1.1 FASTQ

A fastq file has a minimum of four lines representing a single sequence and possibly more since the sequence can be wrapped. The first line starts with a ‘@‘character followed by a sequence identifier and a description. The second line is the raw sequence. The third line starts with a ‘+’ signifying the end of the sequence and it is optionally followed by a sequence identifier. The fourth line contains the base quality score for the bases on line 2, one base quality character for each base.

For example:
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAG
TTT
+
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65

2.1.2 SAM and BAM

Typically the positional information about where the reads were mapped along the genome as well as mapping quality information is stored in a file format called SAM (Sequence Alignment Map) format. Due to the size of these files they are typically compressed into binary versions for the SAM format, called BAM. This is a compact and index-able representation of nucleotide sequence alignments.

2.1.3 VCF file format

A VCF file is consisting of a header section with information about sample with the sample name at the 10^{th} column and a data section which is TAB delimited.VCF file provides information about genotype and read depth and with respect to sample name chromosome name, chromosome position, reference allele, alternative allele, ID which represents the variant identifier, QUAL which shows the phred-scaled probability of all samples being homozygous reference, INFO represents the variant information, and FORMAT colon delimited format of individual genotype [30].
Genotype represents the allele value, for a diploid the representation of genotype is 0/0, 1/0 or 1/1, where the 0 stands for the reference allele and 1 for alternative allele. Read depth represents the reads covered at that position of the sample and it is an integer value.