Summary Report

1 Introduction

This document provides summary for processing and filtering one raw VCF file (/mnt/data/SeqData/gse81089-lung-cancer/outLungTopPlusSAMHg38/L400T_raw.vcf) as well as annotating the filtered VCF file through the Somatic Mutation Annotator through ANNOVAR in BRB-SeqTools. We generate the following files in the variant annotation process:

- A gene list (/mnt/data/ting_test/6_test072116/gse81089/annovarhg38topsam/L400T_raw_genelist.txt) containing nonsynonymous and splicing variants which are not known polymorphisms unless in COSMIC.
- An annotation table (/mnt/data/ting_test/6_test072116/gse81089/annovarhg38topsam/L400T_raw_annoTable.txt) for the detected variants.
- An annotated VCF file (/mnt/data/ting_test/6_test072116/gse81089/annovarhg38topsam/L400T_raw_annotated.vcf) associated with the annotation table.

2 Variant Annotation Process

The raw VCF file is processed and filtered in the following steps:

1. We keep those variants that pass the criterion that the variant call quality QUAL ≥ 50, the read depth DP ≥ 10 and the mapping quality MQ ≥ 50.
2. We decompose and left normalize the remaining variants.
3. We remove those variants reported in dbSNP database but keep those variants reported in COSMIC database.
4. Nonsynonymous and splicing variants are identified from the remaining variants for further analyses.
5. The remaining variants are annotated through ANNOVAR.
6. A gene list is retrieved for the variants through ANNOVAR, which may be a potential list related with the data of interest.

3 Summary Statistics

Table 1 summarizes the stastics related with the variant annotation process via ANNOVAR.

Table 1: Statistics summary associated witht the variant annotation via ANNOVAR.

Statistics	Count
Total number of variants in the raw VCF file	137592
Number of variants left after the filter QUAL >= 50, DP >= 10, MQ >= 50	18401
Number of variants remaining after removing variants reported in dbSNP while keeping variants in COSMIC	3410
Number of variants (out of 3410 variants) that are nonsynonymous or splicing ones	579

Statistics	Count
Number of variants (out of 579 variants) that are reported in COSMIC	11
Number of genes associated with 579 variants	503

We also provide a statistics table for the nonsynonymous and splicing variants kept for annotation. Table 2 summarizes the effects the nonsynonymous variants have.

Table 2: Nonsynonymous and splicing variants after filtering.

Region	Effect	Count
Exonic	Frameshift deletion	4
Exonic	Frameshift insertion	3
Exonic	Stoploss	0
Exonic	Stopgain	3
Exonic	Mis-sense	569
Splicing	/	0
Total	/	579

4 Charts

We summarize here statistics of gene annotations for 3410 variants that pass the quality, read depth and mapping quality filtering criteria. These variants are annotated by RefSeq and UCSC Known Gene annotation sources. We draw figures for the proportion of variants that hit different regions such as exonic and intronic regions as shown in Figure 1, and for the proportion of exonic with differents functional effects (e.g., synonymous, nonsynonymous) as shown in Figure 2.

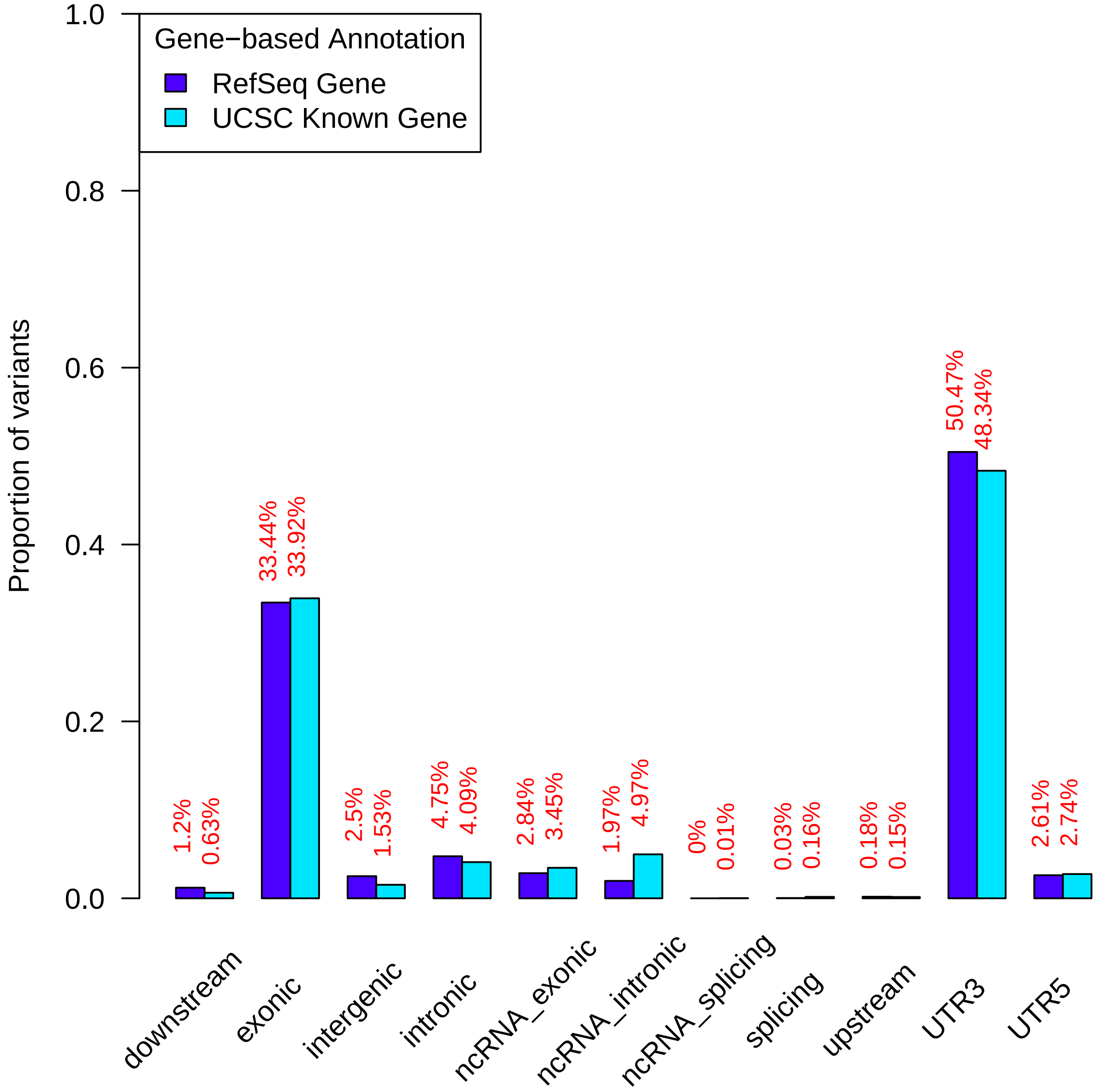

Figure 1: Proportion of variants that hit different regions based on RefSeq and UCSC Known Gene annotation sources.

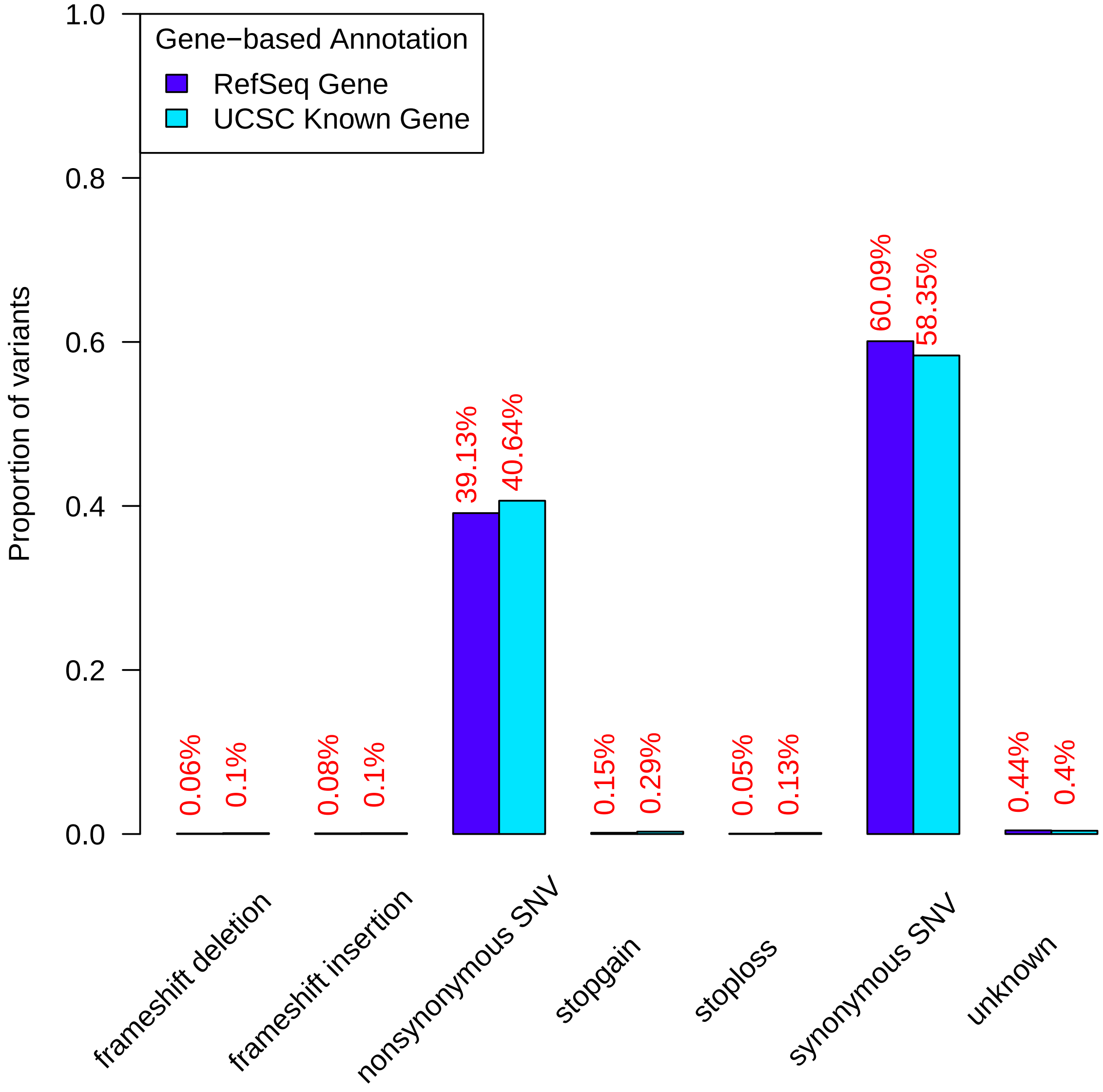

Figure 2: Proportion of exonic variants with their functional effects based on RefSeq and UCSC Known Gene annotation sources.