**Title Page**

Umarani Perasani

Master of Science in Bioinformatics

Mr. Naveen Kamireddi

NxGen MDx

April 2021

**Table of contents**

1. Learning Objectives
2. Introduction
3. Description of Work
4. Internship Discussion
5. Employer Details
6. **Learning Objectives**

The objective of the internship was to provide the student with an introductory experience in bioinformatic analysis, including learning to setup and use bioinformatic command line tools, as practice implement certain parts of the bioinformatic analysis pipeline in R code and interpret the significance of obtained results. The specific topic of the analysis was annotating the effect of SNP data and implementing code for SNP normalization. The learning objectives of this student's internship were:

1. Study the theoretical background regarding the effects of SNPs on human health
2. Study the theoretical background on the analysis of SNP data, including variant detection and annotating the effects of SNPs.
3. Setup proper bioinformatic pipeline
4. Implement SNP normalization algorithm in R code
5. Analyze a test dataset of SNPs and interpret their results
6. Keep a work log of internship
7. Prepare a presentation on the results of internship
8. Creative thinking and problem-solving capability
9. **Introduction**

This report explains about the learning process while working as a Bioinformatics Intern at NxGen MDx. This report submitted as a requirement for course PSM 690, Internship winter, 2021 Grand Valley State University. Development of affordable sequencing technology offered commercially available services for the screening of genetic diseases, enabling unprecedented advances in personalized medicine. In brief, genetic screening allows the discovery if parents carry a genetic disease that you might pass on to their child. Screening can also reveal chromosomal disorders that may affect the health of their child. Genetic screening is a way to look at genes and detect if they are a carrier of an inherited genetic disease, such as cystic fibrosis, that might pass on. Other screens allow detection of non-inheritable genetic conditions that can lead to conditions like Down syndrome. The ideal time to undergo genetic screening is prior to pregnancy. Without screening, it is impossible to know whether parents are carriers of an inheritable genetic condition that can be passed on to their children. The human body is made of trillions of cells which contain genetic information, which is inherited from parents in the form of DNA (deoxyribonucleic acid). A gene is a segment of DNA that is responsible for providing instructions, for how to make a specific protein that the human body needs for normal growth and development. If there are changes (or mutations) in the recipe of a gene, that protein will be affected in some way and may not be able to complete its job in the human body. Genes are located on larger structures called chromosomes. Humans typically have 23 pairs of chromosomes, which are inherited from parents giving two different copies of each gene. Generally, genes can be either "recessive" or "dominant."

A recessive genetic disease occurs when both copies of a particular gene have pathogenic variants. If an organism has one copy of a gene with a variant and one normal copy of that same gene, it is considered a "carrier" for that recessive genetic disease even though not affected by that disease. For carriers to pass on a recessive gene, both partners would have to be carriers of the same genetic disease

A dominant genetic disease occurs when one of the two copies of a particular gene has a pathogenic variant. Only one of the copies of the gene needs to have a pathogenic variant for an individual to be affected by a dominant genetic disease (recessive diseases require both copies to have variants).

The X and Y chromosomes (often called the sex chromosomes) determine whether a newborn is male or female. Males have one X and one Y, while females have two X chromosomes. Some disorders are caused by variants or alterations of the X chromosome. Generally, if a woman has a pathogenic variant on one of the genes located on an X chromosome, she is considered a carrier — and has a 50% chance of passing that variant onto her children. If she passes on the variant on to a male child, that child can be affected with the condition. Affected males will not pass the variant onto their sons, but daughters of affected males will be carriers of the variant. Female carriers rarely develop symptoms of the condition, but their daughters have a 50% chance of being carriers.

Variant annotation is a crucial step in linking sequence variants with changes in phenotype. Annotation results can have a strong influence on the ultimate conclusions of disease studies. Incorrect or incomplete annotations can cause researchers both to overlook potentially disease-relevant DNA variants and to dilute interesting variants in a pool of false positives. There are many different types of information that can be associated with variants, and a first commonly used resource is using databases which contain variants that have previously been described. One popular example is dbSNP, a free, public archive for genetic variation within and across different species. It is hosted by NCBI in collaboration with NHGRI and although the name implies SNPs; it actually includes a range of molecular variation. One fundamental level of variant annotation involves categorising each variant based on its relationship to coding sequences in the genome and how it may change the coding sequence and affect the gene product. To do this we will be using a tool called SnpEff, a variant effect predictor program. Our understanding of the protein-coding sequences in the genome is summarized in the set of transcripts we believe to exist. Thus, variant annotation depends on the set of transcripts used as the basis for annotation.

Variant information on SNPs is collected and stored in variant call files, or VCF for short. Pre-processing these variants are key for successfully using SnpEff. There are two steps in such pre-processing:

1) Decomposing: this step takes multiallelic variants and expands them into distinct variant records; one record for each REF/ALT combination.

2) Normalize: This is a 3-step process. First the alternative and reference alleles are right aligned. Then any suffix is removed that is shared between the reference and alternative alleles. Finally, any prefix shared between the reference and alternative alleles is removed and position is incremented by the number of characters removed from each.

1. **Description of the Work**

As part of the aims of the internship, the student has successfully studied the theoretical background regarding the effects of SNPs on human health and analysis of SNP data, including variant detection and annotating the effects of SNPs. The algorithm for normalizing SNPs was implemented in R and can be seen below:

library (BSgenome.Hsapiens. UCSC.hg19)

library(dplyr)

uma<-read.table(file = "uma\_example.txt",header = T,sep="\t")

uma$chr<-uma$Chrom

uma$start<-uma$Position -1

uma$end<-uma$Position-1

del<-uma%>%select(Chrom,Position)

del$Position<-del$Position-1

variant<-getSeq(x=BSgenome.Hsapiens.UCSC.hg19,

names=del$Chrom,

start=del$Position,

end=del$Position)

variant<-data.frame(variant)

uma$Variant<-variant$variant

uma$Variant

Ref<-unite(data = uma,col = "Ref",Variant:Ref, sep = "")

uma$Ref<-Ref$Ref

uma$Ref

write.table(uma,"uma\_variant\_annotated.tsv",sep = "\t")

After proper normalization of the test SNP dataset was used as input for SnpEff. The following results were obtained.

|  |
| --- |
|  |
| Figure 1. Summary table of the results from SnpEff, test SNP dataset used as input |

|  |
| --- |
|  |
| Figure 2. Variant rate details of the results from SnpEff, test SNP dataset used as input |

|  |
| --- |
|  |
| Figure 3. Number of effects by impact of the results from SnpEff, test SNP dataset used as input |

|  |
| --- |
|  |
| Figure 4. Number of effects by type and region of the results from SnpEff, test SNP dataset used as input |

As can be seen from SnpEff reports, most variants appear in intron regions and have limited influence on phenotype. Most attention should be paid to variants inside exons as they can affect protein function. Also, variants inside the organizational site of the gene such as regulators can influence gene expression.

1. **Internship Discussion**

All of the objectives and aims of the internships have been achieved in due time and properly reported. During the internship, I have gained skills in R coding, bioinformatic tools use and results analysis. Apart from scientific skills, also gained skills in proper time management, efficiency, and other soft skills.

* Did the PSM coursework properly prepare the student for the scientific content of the internship?
* Did the PSM coursework properly prepare the student for the professional content of the internship?
* What challenges did you experience during the internship? What you could have or did to overcome them?
* What is your overall evaluation of the internship experience?

1. **Employer Details**