# 1 Introduction

## 1.1 Motivation

Phenotypes ultimately arise from genetic changes. This realization might be useful if for example, we want to characterize a patient's cancer or gain a better understanding of the underlying genes causing the disease. A simple and useful method to gain insights into how varying genetic expression leads to phenotypes is to directly compare genes across phenotypes. If there is significant differences in expression of a gene between two phenotypes then the intuition is that differentially expressed gene is somehow involved in the processes which lead to one of your phenotypes.

Unfortunately, this method is also prone to problems. Subtle changes in expression that are often missed by statistical tests can have huge effects on systems if the gene is upstream of many processes for example. Furthermore, the insights gained from analysis of a single gene rarely give any explanation of the underlying biology of why differential expression may lead to different phenotypes.

One popular method to get around these hurdles is to group and analyzes sets of genes togetehr. Gene sets are *a priori* groupings of genes made on the basis of similiar regulatory factors, involvement in the same pathways, etc. The benefit of doing this are numerous. First, the subtle changes of expression in a master regulator can be captured since genes involved downstream are included in the analysis. Furthermore, noise is mitigated by looking at the outputs of many genes. One variation of Gene Set Analysis (GSA) is single-sample Gene Set Enrichment Analysis (ssGSEA). This analysis takes a gene set and outputs an enrichment score, or a measure of how much the gene set as a whole is expressed.

*INSERT MORE/REVISE HERE*

## 1.2 My Project

My project was coming up with a single sample method for gene set enrichment analysis using a bayesian approach with Dr. David Gibbs. *More about that later*.

Broadly this problem can be broken down into the following sections:

- Researching current methods for ssGSEA

- Collecting gene expression data from patients with variable phenotypes
- Creating a bayesian approach to ssGSEA
- Creating ways to evaluate different methods

# 2 Project

The first two parts of this project, researching current approaches to ssGSEA and collecting expression data were done fairly simply from a literature review. There are a variety of ssGSEA methods online avaliable in open-source projects, and large caches of gene expression data along with gene set data is curated by institutions all around the globe.