**Materials and Methods**

The Cancer Genome Atlas (TCGA) is a project discovering [genetic mutations](https://en.wikipedia.org/wiki/Genetic_mutation" \o "Genetic mutation) in different [cancer](https://en.wikipedia.org/wiki/Cancer" \o "Cancer)s by using [genome sequencing](https://en.wikipedia.org/wiki/Genome_sequencing" \o "Genome sequencing) and [bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics" \o "Bioinformatics) analysis. The goal of project was to generate statistically and biologically significant data and then meaningful conclusions can be drawn from the genomic data set generated. TCGA applies H[igh-throughput genome techniques](https://en.wikipedia.org/wiki/Multiplex_(assay)" \o "Multiplex (assay)) was used by TCGA to obtain tons of required genomic data, Large data analysis give us more accurate thus better understanding of the genetic basis of cancers. As a result, TCGA enhances people’s the ability to diagnose, treat, and prevent cancer.

RNA Sequencing

RNA sequencing, or called whole transcriptome shotgun sequencing, is a RNA sequencing technique uses [next-generation sequencing](https://en.wikipedia.org/wiki/Next-generation_sequencing" \o "Next-generation sequencing) to check the presence and quantity of [RNA](https://en.wikipedia.org/wiki/RNA" \o "RNA) in a biological sample at a given moment of time.RNA-Seq helps to search for differential gene expression,single nucleotide polymorphism, [exon](https://en.wikipedia.org/wiki/Exon" \o "Exon)/[intron](https://en.wikipedia.org/wiki/Intron" \o "Intron) boundaries,[alternative gene spliced transcripts](https://en.wikipedia.org/wiki/Alternative_splicing" \o "Alternative splicing),[post-transcriptional modifications](https://en.wikipedia.org/wiki/Post-transcriptional_modification" \o "Post-transcriptional modification), [gene fusion](https://en.wikipedia.org/wiki/Gene_fusion" \o "Gene fusion) and etc.

Central dogma of biology suggested that the gene expression level should be positively proportional to the concentration or read count of mRNA and its consequential protein product. Therefore, investigation of gene expression can be done by checking the quantity or concentration of peptide products and mRNA. High read counts or concentration of mRNA indicates high gene expression level and vice versa. Differential expression can be shown in difference in mRNA sequence and peptide products.The most commonly used tools for differential gene analysis are DESeq and edgeR packages from [Bioconductor](https://en.wikipedia.org/wiki/Bioconductor" \o "Bioconductor). Since our data type is count, usually Poisson distribution or negative binomial distribution is preferred.Both these tools use a model based on the negative binomial distribution because Poisson distribution assume the mean and variance of data are the same, variance is a function of mean, this will result overdispersion of data set than expected. In this case negative binomial distribution is more preferred since its mean and variance although related ,they are not the same. This can eliminated the dispersion level in data set.

GSE62944

GSE62944 is a data which reprocessed RNA-Seq data for 9264 tumor samples and 741 normal samples across 24 cancer types from The Cancer Genome Atlas(TCGA) with Rsubread package.Rsubread is a R package that has shown high concordance with other existing methods of alignment and summarization, but is simple to use and takes significantly less time to process data. GSE628944 used seed to vote alignment method. This strategy chooses the mapped genomic location for the read directly from the seeds. It uses a relatively large number of short seeds (called subreads) extracted from each read and allows all the seeds to vote on the optimal location.  It is sensitive because no individual subread is required to map exactly, nor are individual subreads constrained to map close by other subreads. It is accurate because the final location must be supported by several different subreads.

Data level is a method of data catagorization used within the TCGA network to facilitate researchers in communicating and locating their data of interest. There are four data levels: Level 1 is raw data, Level 2 is processed data, Level 3 is segmented and interpreted data and Level 4 is summarized data of region of interest. In this project, since we were interested in expression level of gene which is a kind of level 3 data. It was normalized by RPKM method. RPKM is a normalization method for comparing gene coverage values.If we use direct count of read to measure the gene expression level, gene with longer sequence length will have higher chance to be counted. This results a bias in data. RPKM normalization divide the mapped exon read count of gene by the total mapped gene read count times the length of exon read.

***RPKM=Total exon read/ (mapped reads(millions)\*exon length(KB))***