Identification of Differentially Expressed Genes in RNA-seq Data of *Arabidopsis thaliana*: A Compound Distribution Approach
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4827276/#:~:text=A%20gene%20is%20declared%20differentially,experimental%20conditions%20is%20statistically%20significant.

- “A gene is declared differentially expressed if an observed difference or change in read counts or expression levels between two experimental conditions is statistically significant.”
    - Statistical distributions are used to approximate the pattern of differential gene expression.
    - Such genes are selected based on a combination of expression change threshold and score cutoff, which are usually generated by statistical modeling.
- “This approach was applied in RNA-seq count data of *Arabidopsis thaliana* and it has been found that compound Poisson distribution is more appropriate to capture the variability as compared with Poisson distribution.”
    - Compound distribution model
    - “The Poisson assumption, however, does not account for biological variability in the data”
        - “However, it constrains the variance of the modeled variable to be equal to the mean”
    - “The negative binomial distribution has two parameters, encoding the mean and the dispersion, which allows modeling of more general mean–variance relationships. The negative binomial distribution, which requires an additional dispersion parameter to be estimated, is often used to deal with the biological variability in the data.”
    - “Negative binomial as compound Poisson is more capable of capturing the variability as compared with Poisson distribution and hence identified more differentially expressed genes in case of RNA-seq data.”
    - “Therefore, the mixture of Poisson distribution with gamma mixing weights (Poisson–gamma mixture) that results in negative binomial distribution was fitted to the up-regulated gene expression RNA-seq data. The fitted plot is shown in Figure 4. It is seen that it covers more variability and takes care of overdispersion.”
- “To measure gene expression (or transcript abundance), the sequencing reads obtained are aligned to a known reference genome sequence and the proportion of reads matching a given transcript is used as quantification of its expression level followed by statistical testing of difference in quantification values between samples”
- “TopHat and Cufflinks (Trapnell et al., 2012) are free, open-source software tools for gene discovery and comprehensive expression analysis of high-throughput mRNA sequencing (RNA-seq) data.”
    - “Later on, Cuffdiff 2 (Trapnell et al., 2013) was developed, which is an algorithm that estimates expression at transcript-level resolution and controls for variability evident across replicate libraries.”
- **“Several R packages are available for expression analysis, like DEGseq (Wang et al., 2010). The Bioconductor software package edgeR (Anders and Huber, 2010;**

**Robinson et al., 2010) has been developed to examine replicated gene count data using an overdispersed Poisson model.”**

- **Negative binomial distributions: DESeq, edgeR, and baySeq**
- **EBSeq**
- **Limma-voom, maSigPro**

Differential Gene Expression

https://www.ncbi.nlm.nih.gov/books/NBK10061/

“The three postulates of differential gene expression are as follows:

1. Every cell nucleus contains the complete genome established in the fertilized egg. In molecular terms, the DNAs of all differentiated cells are identical.
2. The unused genes in differentiated cells are not destroyed or mutated, and they retain the potential for being expressed.
3. Only a small percentage of the genome is expressed in each cell, and a portion of the RNA synthesized in the cell is specific for that cell type.”

Differential Gene Expression Analysis

https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/rna-sequencing/performing-a-rna-seq-experiment/data-analysis/differential-gene-expression-analysis/

- “Differential expression analysis means taking the normalised read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups.”

**https://hbctraining.github.io/Training-modules/planning_successful_rnaseq/lessons/sample_level_QC.html**

- “The goal of differential expression testing is to determine which genes are expressed at different levels between conditions.”

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8121839/

- “As a consequence, many newly identified (sub)populations are missing and relationships between cell populations might be inaccurate. A striking example of this inadequacy is neuronal cell populations. Recent single-cell studies have identified hundreds of populations[4,13,14], including seven subtypes and 92 cell populations in one study only[5]. In contrast, the Cell Ontology currently includes only one glutamatergic neuronal cell population without any subtypes.”
- “For instance, if a dataset is annotated at a low resolution, it might contain T cells, while a dataset at a higher resolution can include subpopulations of T cells, such as CD4+ and CD8+ T cells. We need to consider this hierarchy of cell populations in our representation, which can be done with a hierarchical classifier. This has the advantage that cell population definitions of multiple datasets can be combined, ensuring consistency.”
-