CS690: Computational Genomics

Lecturer: Hamim Zafar — Lecture 22 (Differential Expression)
Scribe: Mubashshir Uddin [190516] — November 03, 2023

Disclaimer: *These notes aggregate content from several texts and have not been subjected to the usual scrutiny deserved by formal publications. If you find errors, please bring to the notice of the Instructor at hamim@cse.iitk.ac.in.*

1 DIFFERENTIAL GENE ANALYSIS

Differential gene analysis(DE or DGE in short) refers to the process of identifying how does the gene expression values effect the overall health and functioning of the cells. For example if we have a healthy cell, say cell A, and another tumor cell cell B, then there is a possibility that there is some observable changes in the gene expression values of these genes. The difference can simply lie within the changes in the values of particular genes, or even in the ratios of differences in the gene expression. Applying learning models to identify the bio-medically relevant parameters, is one of the apex task of computational genomics. Because such a technology if developed holds the key to micromanaging health. for example, we can have targeted vector RNA medicines that treat cancer by specifically identifying the cancer cells and then injecting them with an in place self-destruct signal.

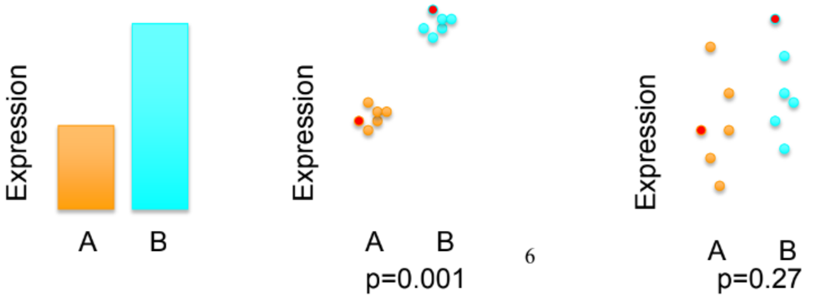

Fig. 1. p-0.001 is much less likely to be observed than p-0.27, but the first situation might be indicative of a disease. credit: lecture slides

The major problem in this task is that the gene expression values almost never directly correlate with the overall health of the cell. Further more there are significant **biases** expressed to different levels in the expression of genes themselves. For example,

(1) **Fragment length preference:** Bigger RNA segments(1000-2000 base pairs) are more likely to be effectively picked up by the genome sequencing machines, thus producing an artificially inflated counts of the genes that are represented by huge number of base pairs. Having lesser number of base pairs to a gene is not the best predictor of biological significance of the gene.
(2) **Positional bias:** There is a higher probability of getting good fragments from the end of a transcript. As the transcript is arranged from the 3' to 5' end, there is a higher transcript capture ratio near the 5' end, any of the genes that are located near this end will have an manipulated expression count. also being a positional effect this will have a continuously varying effect across the length of the transcript.
(3) **Sequence based bias:** Primers are short stretches of DNA that target unique sequences and help identify a unique part of genome. binding efficiency is also a continuous variable parameter, dependent even upon the test temperature for each different gene.
(4) **Fragment G-C Bias:** The G-C (nucleotide bases of the DNA) content of the organism proposes unique challenges to genome expression, higher G-C content generally meaning that the reads are less likely to be sequenced.

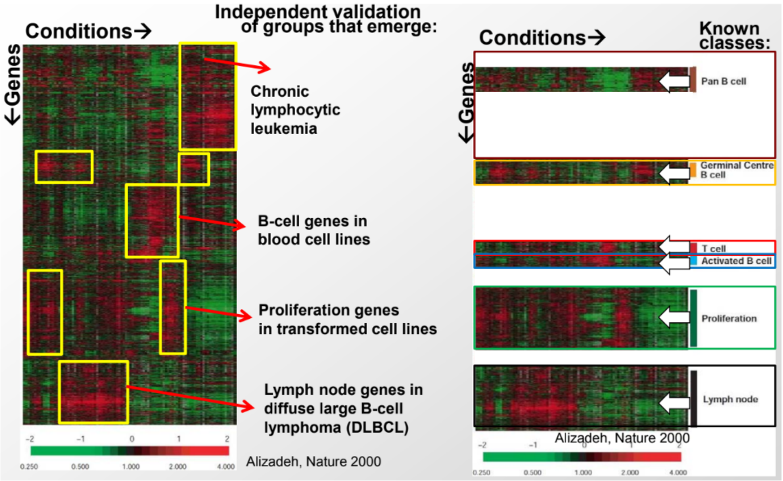

Fig. 2. gene expression and global changes to gene expression often relates to the occurrence of disease

identification of genes whose expression differs under distinct conditions is crucial for understanding the molecular basis of various biological processes, including diseases like cancer, responses to stress, and the regulation of essential cellular systems. These differential expressed genes (DEs) play a pivotal role in unraveling the complexities of cellular behavior, and they often serve as valuable starting points for developing models that depict the underlying systems being studied.

some examples that the medical community has been aware of from quite a while-

- Tumor Suppressor Genes such as TP53, BRCA1, and BRCA2, which are often associated with tumor suppression, can provide insights into the loss of regulatory control in cancer cells.
- Cyclins and Cyclin-Dependent Kinases (CDKs): Genes involved in the regulation of the cell cycle, such as CCND1 (cyclin D1) and CDK4, can be key players in understanding the control points governing cell division.

2 MODELLING THE VARIATION

To get to the bottom of the puzzle of finding how the genes are affected based on the cell state, we first need to figure out how to model the biological system itself and the different scales of noise that is being added to our system. Only when we have accurate models of both the noise and the system that we wish to analyse, then we can identify the small variations in these patterns and label them as outliers, i.e. the deferentially expressed genes that we have been looking for.

2.1 Mathematical modelling of the observed data (Probability distributions)

There are a few probability distributions that we often use to fit out experimental data and draw facts. Here is a quick refresher on how does the fitting works and what type of data they are best used to model:

- **Poisson Distribution:** Poisson distribution is a mathematical model that describes the probability of a given number of rare and random events occurring within a fixed interval of time or space. It assumes that these events happen independently of each other and at a constant average rate. The distribution is commonly used in various fields, such as telecommunications, traffic engineering, and epidemiology, to analyze and predict the occurrence of infrequent events.
- **Negative Binomial Distribution:** Models the number of failures in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of successes (denoted r) occurs. For example, if we flip a coin until we see three heads (r = 3), then the probability distribution of the number of tails will be negative binomial.

Since the replicates have a tendency to uptake very high variances, thus the poisson distribution becomes unusable for genes count. But it is still of utmost use to predict rate charts. Negative binomial gives a much better alternative to modelling such data

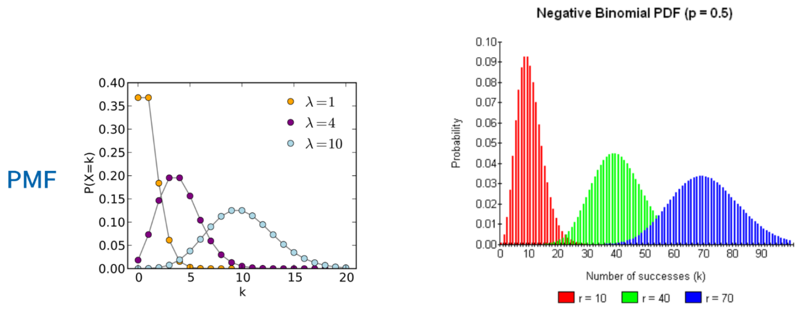

Fig. 3. Poisson and negative binomial distribution

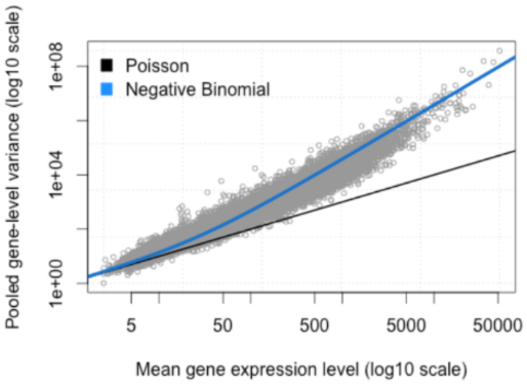

Fig. 4. NBD is better than poisson dist

2.2 RSEM

RSEM is a software package for estimating gene and isoform expression levels from RNA-Seq data. The RSEM package provides an user-friendly interface, supports threads for parallel computation of the EM algorithm, single-end and paired-end read data, quality scores, variable-length reads and RSPD estimation

RSEM already accounts for fragment length and positional bias. It also creates and holds specific variables to label the degree to which a particular transcript has not been sequenced.

3 HOW TO CHECK FOR DIFFERENTIAL EXPRESSED GENES

since we are having a fair bit large set of genes pool from which we want to find inconsistencies for expression we need to first narrow down our search:

3.1 Hypothesis testing

Hypothesis testing refers to creating strong hypothesis based on the current set of experimental conditions. Then setting up the experiment so that the hypothesis is deemed true.
Example: if we know that a particular gene is suspected to be related to cancers, then we identify the gene and setup a **NULL Hypothesis**, the null hypothesis in general makes a statement - that the observed differences are due to the many other genes present except our test genes. So for the NULL hypothesis to be false, our predictor gene should have a strong correlation with the observations.
basically differential expression is all about modelling the biological variability itself and provide a measure of significance. The way to find the default expected values of the gene expression is usually to train on a

large enough set of cells that are medically and physically considered to be healthy.

3.2 Single cell -omics data and spatial -omics data

The spatial transcriptomics data comes with an image and the cell position matrix associated with each reading, one of the major problems with ST data is that there is much higher noise and missing values, because of all the extra experimental steps required to capture the cell positions along with the gene expression.

3.3 Constant genes

Many of the present in viable cells are there necessarily to maintain the basic functionality of the biological system, i.e relating to cellular respiration, metabolism, waste management, etc. if there is a natural error in these genes then the cells are likely to be dead rather than alive. so any of the reads of these genes coming from a healthy sample will mostly contain the experimental noise. If we have enough examples working cells, we can figure out the default values of the expression of these genes, across a particular trajectory of the cells. this is known as the **rank invariance** of genes.

4 NORMALIZATION OF THE DATA

If we want to use our -omics data as a training example, into models that spot the biological significance in the data, it is very important to first normalised across all of the major sources of variations. Some normalisation's that we can carry out are:

- **Quantile normalization:** There should be no differences across multiple trials. Quantile normalization ensures that the distributions of expression values for each gene are the same across all samples. This is particularly important in gene expression studies where the goal is often to compare the relative expression levels of genes across different conditions or samples
- **Scale factor normalization:** Same distribution but different scale factors, eg.library size. We assume that the total mRNA quantity remains the same across samples and the variations come from the errors introduced during sequencing.
  - Mean normalization:
    $$\hat{y}' = y' - M^j + M$$
  - Variance normalization:
    $$\hat{y}^i_j = \frac{(y^j_i - M^j) \cdot \sqrt{v}}{\sqrt{v^j}}$$
- **Invariant set normalization:** all of the invariant sets that we are talking about need to be normalised across the observations. Different cells according to the different condition of sequencing will have different values of gene expression but the expression of the house keeping genes should be proportional. So using this we normalise all the genes also.

5 RPKM, FPKM AND TPM

Earlier RNA-seq measurements reported results in RPKM (Reads Per Kilobase Million) or FPKM (Fragments Per Kilobase Million).
However, TPM (Transcripts Per Kilobase Million) is the new trend.

These three metrics attempt to normalize for sequencing depth and gene length.

- **RPKM:** Count up the total reads in a sample and divide that number by 1,000,000 – this is our "per million" scaling factor. Divide the read counts by the "per million" scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM) Divide the RPM values by the length of the gene, in kilobases.
- **FPKM:** RPKM was made for single-end RNA-seq, where every read corresponded to a single fragment that was sequenced. FPKM was made for paired-end RNA-seq. With paired-end RNA-seq, two reads can correspond to a single fragment, or, if one read in the pair did not map, one read can correspond to a single fragment. The only difference between RPKM and FPKM is that FPKM takes into account that two reads can map to one fragment (and so it doesn't count this fragment twice).

- **TPM**is very similar to RPKM and FPKM. The only difference is the order of operations. Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK). Count up all the RPK values in a sample and divide this number by 1,000,000. This is your “per million” scaling factor. Divide the RPK values by the “per million” scaling factor. This gives you TPM. So you see, when calculating TPM, the only difference is that you normalize for gene length first, and then normalize for sequencing depth second. However, the effects of this difference are quite profound

When you use TPM, the sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of reads that mapped to a gene in each sample. In contrast, with RPKM and FPKM, the sum of the normalized reads in each sample may be different, and this makes it harder to compare samples directly [source]

6 TRIMMED MEAN

The Trimmed Mean of the M-values (TMM) is a normalization approach used in sequencing tasks, particularly in the analysis of RNA-seq data. The TMM method involves selecting a reference sample, calculating fold changes and absolute expression levels relative to that sample, and then trimming the genes based on these values to remove differentially expressed (DE) genes. The trimmed mean of the fold changes is then calculated for each sample. This trimmed mean is used to scale the read counts, adjusting for differences in library sizes.
In the TMM approach:

1. **Reference Sample:** A single sample is chosen as a reference, and the other samples are compared to it
2. **Fold Changes:** Fold changes and absolute expression levels are calculated relative to the reference sample.
3. **Trimming:** Genes are trimmed based on fold changes and absolute expression levels to eliminate differentially expressed genes.
4. **Trimmed Mean:** Genes are trimmed based on fold changes and absolute expression levels to eliminate differentially expressed genes.
5. **Scaling:** Read counts are scaled by this trimmed mean and the total count of their respective samples.

The TMM method is employed for normalization to mitigate the impact of variations in library sizes and other technical biases, allowing for more accurate comparisons of gene expression levels across different samples.
It’s worth noting that the edgeR package commonly uses TMM normalization. However, in the literature, the term TMM is more prevalent. Another normalization method mentioned in the context of RNA-seq is Median Ratio normalization (MRN), which is similar to TMM but aims to be more robust. In MRN, read counts are divided by the total count of their sample, and the median of the ratios of these values between conditions is used for normalization. The original counts are then adjusted by this median and their respective library size. [source]

7 P- VALUE OF THE MODEL

In the context of RNA-seq and differential gene expression analysis, the p-value plays a crucial role in hypothesis testing to determine whether a gene is differentially expressed under different experimental conditions. The fundamental idea is to assess the probability of observing the data under the assumption that the null hypothesis is true. The null hypothesis posits that there is no significant difference in gene expression between the compared conditions, implying that the mean expression of the gene is constant.
To conduct hypothesis testing, researchers formulate two hypotheses for each gene: the null hypothesis (H_0) and the alternative hypothesis (H_1) The null hypothesis suggests that the gene’s expression is consistent across conditions, while the alternative hypothesis asserts that there is a significant difference in expression between the conditions. In mathematical terms:

$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A \neq \mu_B$$

where μ_A and μ_B represent the mean expression levels under conditions A and B, respectively.

The p-value is a statistical metric that quantifies the probability of obtaining the observed data or more extreme results under the assumption that the null hypothesis is true. In the context of RNA-seq, it reflects how likely it is to observe the given gene expression profile if there were no true differences in expression between the conditions. A small p-value suggests that the observed data is unlikely to have occurred by random chance alone, providing evidence against the null hypothesis.

Researchers typically set a significance threshold, often denoted as α such as 0.05. If the calculated p-value is less than this threshold, the null hypothesis is rejected, indicating that there is sufficient evidence to suggest differential gene expression. On the other hand, a p-value greater than the threshold leads to the acceptance of the null hypothesis, suggesting that observed differences are likely due to random variability.

In summary, the p-value in RNA-seq analysis serves as a crucial measure of the strength of evidence against the null hypothesis of no differential expression. It provides a quantitative basis for decision-making, allowing researchers to assess the reliability of their findings and determine whether observed gene expression differences are statistically significant or could be due to random fluctuations.

8 REFERENCES:

(1) Lecture slides: https://drive.google.com/file/d/1j0hQIfqk5jOpejNA4DU0$_{f}ChWNmTxVZ8/view$
(2) My class notes: https://drive.google.com/file/d/1rYOqXbdt3FZCoqftDJRTalYHBxCEDlSh/view?usp=sharing
(3) https://academic.oup.com/bib/article/19/5/776/3056951
(4) https://www.investopedia.com/terms/t/trimmed$_{m}ean.asp$
(5) https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/
(6) https://deweylab.github.io/RSEM