Study on the Differential Gene Expression of Normal vs Tumor Tissue in Locoregionally Advanced Laryngo-Hypopharyngeal Carcinoma

ABSTRACT:
This study was performed on mRNA profiles from human laryngo-hypopharyngeal carcinoma were generated by deep sequencing. Since the beginning of the twentieth century, two major options were available for the treatment of locally advanced laryngeal and hypopharyngeal squamous cell carcinomas (LA-LHCs): definitive radiation therapy (RT) with salvage surgery reserved in case of local failure or total laryngectomy with postoperative RT. Clinical investigations aimed at extending the indications of partial laryngectomy or exploring different protocols of RT using altered fractionation schedules or concurrent radiosensitizers. In this study, we aim to study differential gene expression of normal and tumoral tissue types in LA-LHCs).

METHOD:
The rapid adoption of high-throughput sequencing (HTS) technologies for genomic studies has resulted in a need for statistical methods to assess quantitative differences between experiments. We analyse 54 samples from the GSE184072 database. For our purpose, we have chosen DESeq2, interfaced through the R programming language via the Bioconductor repository.

DESeq2 uses a negative binomial distribution model to account for the over-dispersion commonly observed in RNA-seq data. This allows it to accurately capture the variance-mean relationship in the data, leading to more reliable estimates of differential expression. The estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions.

```
# Step 1: Load the data
# Step 2: construct a DESeqDataSet object ----------
dds <- DESeqDataSetFromMatrix(countData = data,
                              colData = colData,
                              design = ~ Tissue)
# set the factor level
dds$Tissue <- relevel(dds$Tissue, ref = "Normal")

# Step 3: Run DESeq ----------------------
dds <- DESeq(dds)
res <- results(dds)
```

MA plot shows the log average (A) on the x-axis and the log ratio (M) on the y-axis. Here, M stands for minus because log(A/B) = logA-logB.

Volcano plots display the statistical significance of differential experession of every single gene, usually through the negative base-10 log and base-2 log fold-change, respectively. Since the P-values have a negative transformation, the higher along the y-axis a data point falls, the smaller the P-value. It is generally used for volcano plots to include some threshold indicators for adjusted P-values to indicate which genes would be considered statistically differentially expressed based on the adjusted P-value of their difference between treatments. The log fold-change along the x-axis displays more considerable differences in the extreme values, with data points closer to 0 representing genes that have similar or identical mean expression levels. For volcano plots, a fair amount of dispersion is expected as the name suggests. A wider dispersion indicates two treatment groups that have a higher level of difference regarding gene expression. It is quite rare for a volcano plot to have most, or all data points clustered close to the origin.

RESULTS:

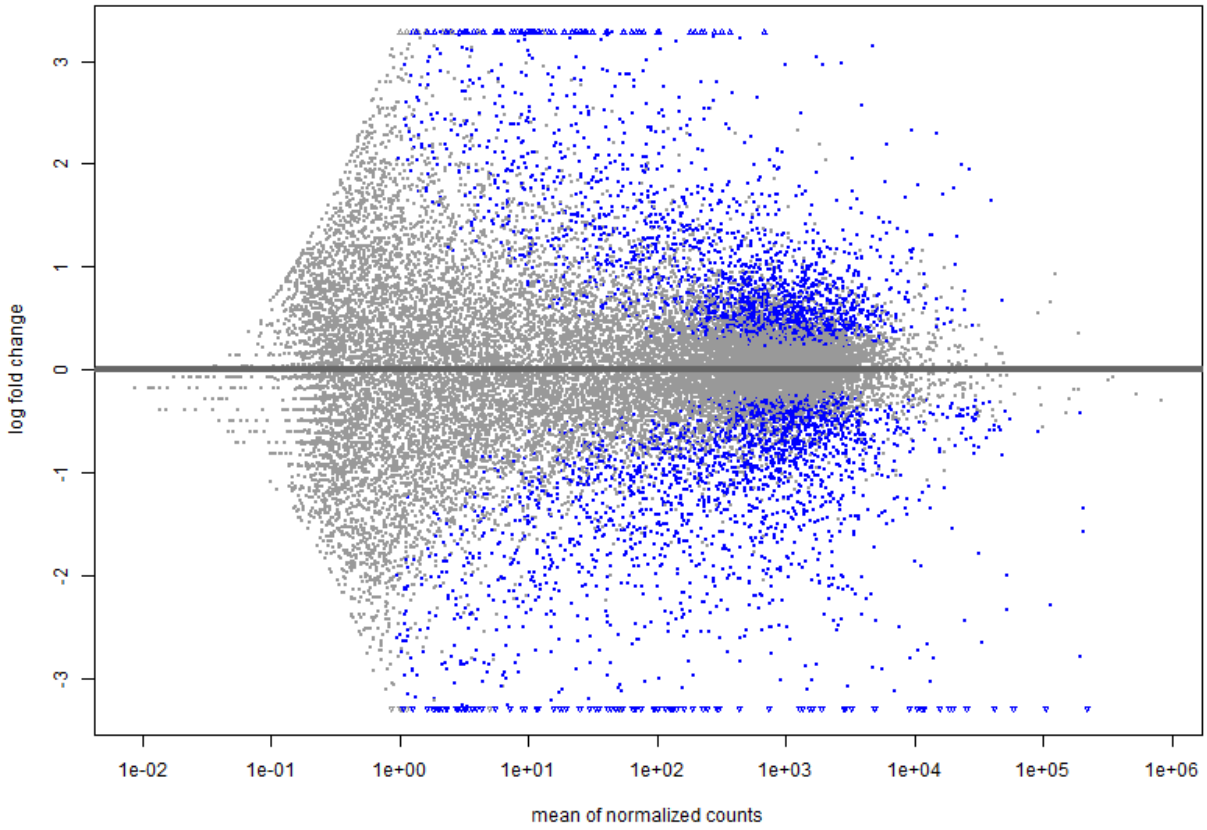

Figure 1: MA plot showing the differentially expressed genes. Genes in blue are significantly differentially expressed genes, they have adjusted p-values<0.05. The triangles at the edges of the plot indicate these genes have higher log-fold changes.

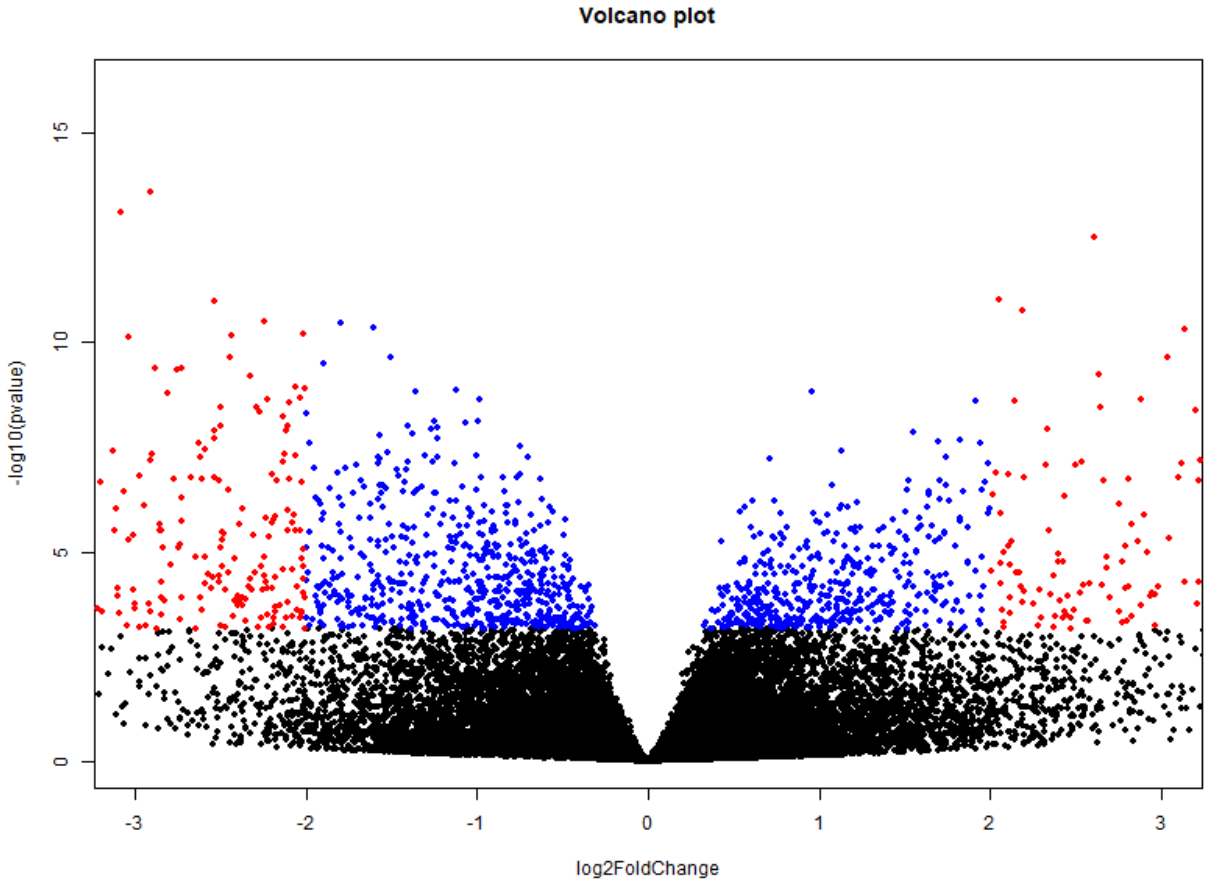

Figure 2: Volcano plot displaying both statistical significance and fold change information for each gene. Genes that are significantly differentially expressed are located toward the top of the plot (high significance) and toward the sides (large fold change). Significantly upregulated genes are colored red, significantly downregulated genes are colored blue.

DISCUSSION:

We have identified genes that are statistically significantly upregulated or downregulated in tumor cells as compared to the normal tissue. These genes can be further studied to understand their role in the gene regulatory network. In combination with a clinical dataset, we would be able to identify biomarkers for treatment of LA-LHCs.

DATA AND CODE AVAILABILITY:

The data has been downloaded from the National Center for Biotechnology Information Gene Expression Omnibus (www. ncbi.nlm.nih.gov/geo/) with the accession number GSE184072. All codes are available on my GitHub profile: https://github.com/Suhana101/Bioinformatics

REFERENCES:

1. Construction of a novel six-gene signature to predict tumour response to induction chemotherapy and overall survival in locoregionally advanced laryngeal and hypopharyngeal carcinoma
2. Laryngeal Preservation Strategies in Locally Advanced Laryngeal and Hypopharyngeal Cancers
3. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2
4. Interpretation of differential gene expression results of RNA-seq data: review and integration
5. Analyzing RNA-seq data with DESeq2
6. Global transcriptomic analysis of breast cancer and normal mammary epithelial cells infected with Borrelia burgdorferi