In this Case study II, we used TCGA head and neck cancer data to perform RNA-seq Data analysis. Following is a brief summary of steps of RNA-seq data analysis:

1) Make sure OOMPA related packages like oompaBase, oompaData, PreProcess are downloaded and loaded to the R session. We require gplots for heatmap.2 function and package edgeR for normalization.

3) Firstly, load the data set (In our case we used TCGA head and neck cancer data).

4) Calculate RNA-seq raw counts. These raw counts will be transformed using one of the following methods:-

(i) RPKM(Read per Kilobase per Million mapped reads): Divide RPM by the length of genes in kilobases.

(ii) FPKM (Fragments per kilobase per million): same like RPKM but accounts for paired-end reads.

(iii) TPM (Transcripts per kilobase million): like FPKM but accounts for gene length first and then library size.

5) RNA-seq analysis would be very similar to how we began with microarray data. We will use boxplot of log2 data since it's illuminative and there won't be large outliers.

6) We perform data reduction by calculating PCA (principal component analysis) for our log2 transformed data using samplePCA function in OOMPA package. We plot the result after reduction.

7) We use principle coordinate analysis (a classical form of MDS) by using cmdscale function in R.

8) To measure the distance between two objects, we can use measurements like:

(i) Maximum: abs(max(x-y))

(ii) Manhattan: sum(abs(x-y))

(iii) Correlation: (1-cor(x,y))/2

(iv) Euclidean distance

Distances based on Correlation will be more informative than Euclidean distance.

9) Before statistical analysis, filtering needs to be performed where we filter based on the mean expression. The threshold for filtering is subjective process, but generally mean count <10 across samples can be eliminated.

10) We perform normalization using the calcNormFactors function in package limma. We make use of DGEListclass which contains digital gene expression data.

11) Mean-variance modeling can be done using limma-trend or limma-voom approach. In our study we used limma-voom approach.

12) We create design matrix, then transform and calculate weights using voom function from limma package.

13) We use boxplots to see data is normalized.

14) Next is linear transformation using lmFit, eBayes and topTable function.

15) We use sapply() to find count of genes that pass p-value thresholds.

16) Volcano plot can be used to display results from differential expression testing.