.CEL files are produced by the array scanner software and contain the measured probe intensities.

Genomic data consists of different components - stored as standardised data structures through Bioconductor - ‘ExpressionSet’.

ExpressionSet = expression data ; description of samples ; metadata/technology used for the experiment/annotations ; description of the experiment.

RMA :
- Background adjustment is important to remove non-specific hybridisation and the noise of optical detection system.
- Normalisation ensures expression values are comparable across different microarray chips.
- Summarisation I the last step where the background adjusted and normalised intensities of all probes summarised into one quantity.

Summarisation Plot using raw CEL data:
- Turn off background correction and normalisation.
- Summarisation step combines probe level intensities to one summarised expression value per transcript for each sample.
- Histogram shows the distribution of these expression values across all samples.
- To ensure that there are no extreme outliers.
- The plot needs to look roughly gaussian.

PCA Plot using RMA raw CEL data :
- Performed on RAW cel data after ram normalisation.
- Every point represents one sample, with the colour indicating its source.
- PC1 captures largest source of variation in the data.
- PC2 captures the technical / biological factor in the data.
- In this PCA the samples cluster by their source.

Mean Expression Analysis :
- Baseline expression of each gene in the dataset.
- Between groups gives information on how each gene behaves in each cell type - spot genes enriched in different cell types.
- log2FC tells how strong the difference is, if positive higher in abT, if negative higher in macrophage from mixed background.
- Scatter Plot :
    - Points along y=x indicate genes expressed similarly in both groups.
    - Points above diagonal indicate gene expressed more in macrophages.
    - Points below diagonal indicate genes expressed more in abT.
    - Red Line = potential DE genes.
    - Density colouring shows where points cluster most, near the diagonal = shared / housekeeping genes.

Variance and Standard Deviation :
- Variance measures how spread out the data is.
- Standard Deviation gives the average distance from the mean in the same units as the data.

High variance/SD → gene expression changes a lot between samples (potentially biologically meaningful).

Low variance/SD → gene expression is stable across samples (likely housekeeping or uninformative).

- Variance Plot :

    - Points along diagonal = shared.
    - Points above diagonal = macrophage variable.
    - Points below diagonal = abT variable.
    - Dense cluster near origin = low variance in both groups.
    - Outlier = unusual high variance genes.

- SD Plot :
    - Points along diagonal = shared.
    - Points above diagonal = macrophage variable.
    - Points below diagonal = abT variable.
    - Dense cluster near origin = low SD non informative.
    - Outlier = unusual high SD.

Mean vs Variance :
    - Points near bottom left = low mean, low SD = background/noise.
    - Points with high mean, low SD = housekeeping genes.
    - Points with low mean, high SD = Low count genes / artefacts.
    - Points with high means, high SD = Biologically interesting.
- Overall trend SD increases with mean.

Effect Size / MA Plot :
- To visualise differences in gene expression between two groups.
- It is the difference between group means.
    - Points near y= 0 indicate not differentially expressed genes.
    - Points above y = 0 indicate genes higher in abT.
    - Points below y=0 indicate genes higher in macrophages with mixed background.
    - Extreme points at high A indicate potential DE genes.
    - Extreme points at low A indicate noise.

Pooled Standard Deviation :
- Used to compute standard error between two groups when assuming equal variance.
- Used when sample sizes differ to give weighted means, in this case both conditions have the same number of samples.

Student t-test :
- t-statistics measures the strength of evidence for effect size, taking into account variability and size.
- Plot for t-statistics vs effect size:
    - Points near t = 0 indicate little variance.
    - Points far from t = 0 indicate high variance.
- Ideal DE genes have large effect size and large t-statistics.
- Plot for pooled variance vs p value:
    - Points on the left size = low SD, low p-value = reliable DE genes.
    - Points on the left side = low SD, high p-value = no real differences.
    - Points on right side = high SD, high p-value = unreliable for DE.
    - Points on right side = high SD, low p value = rare, may indicate noisy genes needs validation.

Calculating expected number of false positives :
- Indicates the amount of biologically significant genes providing a strong evidence.
- FDR :
    - Controls multiple testing.
    - Gene above threshold likely differential.

* Inspected the statistical behaviour of the data.
* Validated logic behind ‘limma’.
* Test showed 48% of genes significant.

* Strong separation between groups in mean-variance and MA plots confirming real signal not noise.
* Can be used for cross validation.

Linear Models :
- Find genes that are differentially expressed between experimental conditions. Here I compare the samples of different cell types.
- To identify which groups to compare run a command to identify the number of samples of each cell type.
- Focus on groups with at least 2-3 samples per condition.

Stromal Cell, C57BL/6J	11
abT Cell	6
Macrophage, mixed background	6
Neutrophil, C57Bl/6J	4
abT Cell, C57BL/6J	3
Eosinophil, C57BL/6J	3
Macrophage, C57BL/6J	3

- Running ‘limma’ without contrasts:
    - Comparison to the baseline = category that is alphabetically first = abT cells.
    - Down = 204
    - Not Sig = 5049
    - Up = 445
    - Top 10 Genes : Rag1 ; Endou ; Rag2 ; Arpp21 ; Lig4 ; Aqp11 ; Mir181a-1 ; H2-D1

- Running ‘limma’ with contrast:
    - Comparing abT and macrophages from mixed background.
    - Down = 1060 (mf)
    - Up = 596 (abT)
    - Top 10 Genes : Gata6 ; Ltbp1 ; Rp1 ; Rag1 ; Gdpd3 ; Garnl3 ; Selp ; Rag2 ; Arpp21 ; Lrg1

Rag1, Rag2, Arpp21 : canonical T cell markers consistent with abT lineage. Rag1 and Rag2 - antigen receptor rearrangement.
The rest of the genes are classic macrophage markers - adhesion, lipid metabolism, TGF-B signalling.

Lrg1 and H2-D1 found down regulated - linked to immune activation and antigen presentation.

Multi Density Plot :
- Shows distribution of all expression values across samples or groups.
- Helps assess normalisation.

Resources :
https://gtk-teaching.github.io/Microarrays-R/04-MetaData/index.html

https://dputhier.github.io/ASG/practicals/microarrays_student_test/DenBoer_Student_test.html#student_test_and_p-value

https://bioconductor.org/packages/release/workflows/vignettes/maEndToEnd/inst/doc/MA-Workflow.html