Differential Expression Analysis

Dulce Alejandra Carrillo Carlos

February 23, 2026

Introduction

Disseminated tumor cells (DTCs) are cancer cells that detach from the primary tumor and migrate to distant tissues, where they may remain dormant or initiate metastatic growth. Understanding the molecular differences between DTCs and normal mononuclear cells (MNCs) is crucial for identifying mechanisms underlying tumor dissemination and metastasis.

In this report, I analyzed RNA-seq data from the **SRP097735** project using the `recount3` and `DESeq2` R packages. The objective was to identify differentially expressed genes (DEGs) between DTC and MNC samples and to characterize their global transcriptomic differences.

Methodology

The analysis was conducted using:

- `recount3` for accessing RNA-seq datasets
- `DESeq2` for differential expression analysis
- `ggplot2` and `pheatmap` for visualization

The workflow included:

1. Data import
2. Quality filtering
3. Differential expression modeling
4. Exploratory visualization (PCA)
5. DEG visualization (Volcano, MA plot, Heatmap)

Data Import

```
library(recount3)
library(SummarizedExperiment)
library(DESeq2)
library(ggplot2)
library(pheatmap)
library(dplyr)
library(RColorBrewer)


options(recount3_url = "https://data.idies.jhu.edu/recount3/data/")

human_projects <- available_projects()

project_info <- subset(
```

```
    human_projects,
    project == "SRP097735" &
    project_type == "data_sources"
)

rse <- create_rse(project_info)
rse
```

The `SummarizedExperiment` object contains raw counts and sample metadata.

Data Filtering

Lowly expressed genes were removed to reduce statistical noise.

```
counts <- assays(rse)$raw_counts
dim(counts)
```

```
## [1] 63856    86
```

```
keep <- rowSums(counts >= 10) >= 5
counts <- counts[keep, ]
dim(counts)
```

```
## [1] 53808    86
```

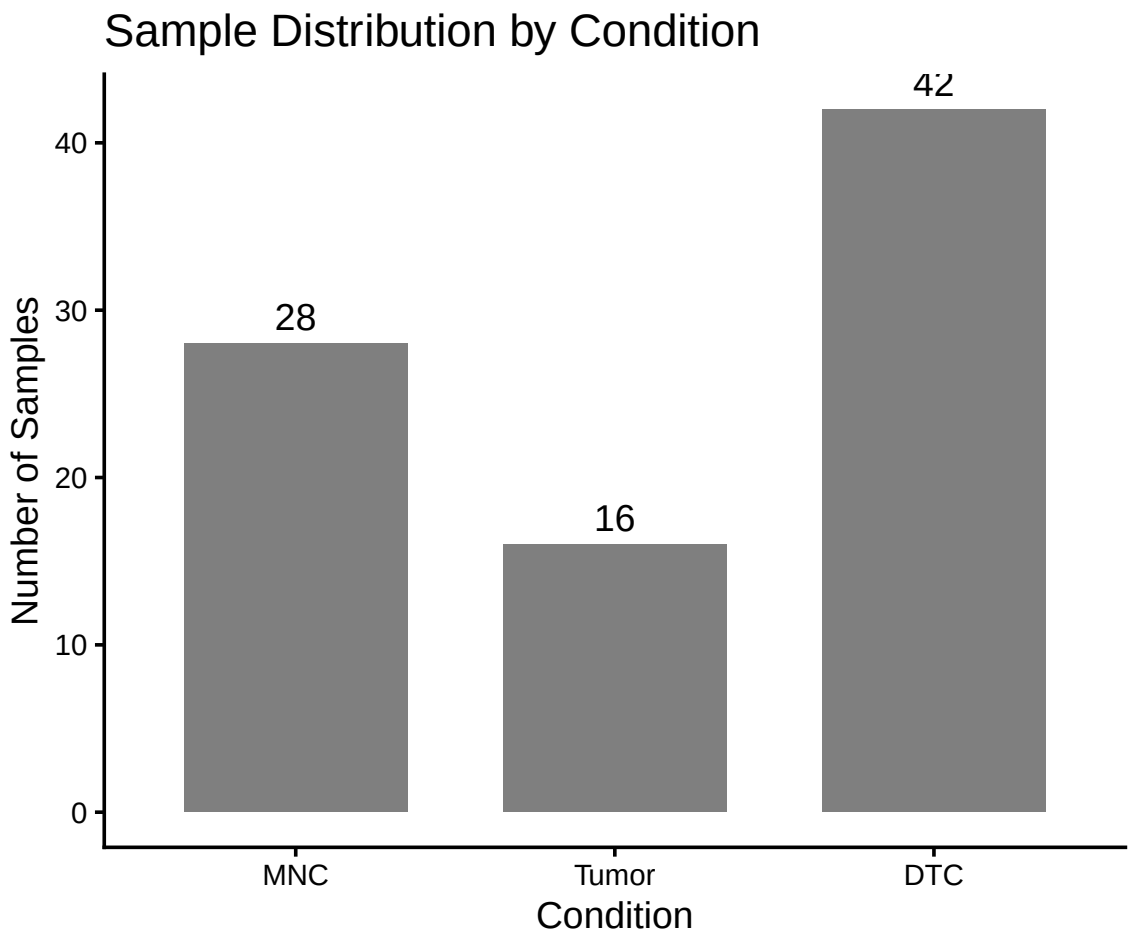
After filtering, only genes expressed in at least five samples with 10 counts were retained.

Metadata Preparation

```
# Extract sample attributes
attr <- colData(rse)$sra.sample_attributes

# Create condition variable
group <- ifelse(grepl("DTCs)", attr), "DTC",
         ifelse(grepl("MNCs)", attr), "MNC",
         ifelse(grepl("Tumor", attr), "Tumor", NA)))

# Show distribution table
table(group)
```

```
## group
##   DTC   MNC Tumor
##    42    28    16
```

```r
# Convert to factor
group <- factor(group, levels = c("MNC", "Tumor", "DTC"))

# Create metadata dataframe
coldata <- data.frame(condition = group)
rownames(coldata) <- colnames(rse)
coldata <- coldata[colnames(counts), , drop = FALSE]

# Prepare dataframe for plotting
plot_df <- as.data.frame(table(group))
colnames(plot_df) <- c("Condition", "Count")

ggplot(plot_df, aes(x = Condition, y = Count)) +
  geom_bar(stat = "identity", fill = "gray50", width = 0.7) +
  geom_text(aes(label = Count), vjust = -0.5, size = 5) +
  theme_classic(base_size = 14) +
  labs(
    title = "Sample Distribution by Condition",
    x = "Condition",
    y = "Number of Samples"
  ) +
  theme(legend.position = "none")
```



Differential Expression Analysis

```
dds <- DESeqDataSetFromMatrix(
    countData = counts,
    colData = coldata,
    design = ~ condition
)

dds <- DESeq(dds)
```

Main contrast:

```
res_DTC_vs_MNC <- results(
    dds,
    contrast = c("condition", "DTC", "MNC"),
    alpha = 0.05
)

res_DTC_vs_MNC <- res_DTC_vs_MNC[order(res_DTC_vs_MNC$padj), ]

summary(res_DTC_vs_MNC)
```

```
##
## out of 53747 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 10139, 19%
## LFC < 0 (down)     : 6963, 13%
## outliers [1]       : 0, 0%
## low counts [2]     : 2145, 4%
## (mean count < 2)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
head(res_DTC_vs_MNC)
```

```
## log2 fold change (MLE): condition DTC vs MNC
## Wald test p-value: condition DTC vs MNC
## DataFrame with 6 rows and 6 columns
##                     baseMean log2FoldChange     lfcSE      stat       pvalue
##                    <numeric>      <numeric> <numeric> <numeric>    <numeric>
## ENSG00000105613.9    71989.5        6.33088  0.168418   37.5902 3.11385e-309
## ENSG00000196361.9   144467.8        7.88729  0.215840   36.5423 2.36477e-292
## ENSG00000008735.13   91473.9        7.02627  0.200615   35.0236 9.83619e-269
## ENSG00000157851.16  149612.4        8.53728  0.249259   34.2507 4.25831e-257
## ENSG00000109132.6   244359.4        8.74736  0.255552   34.2293 8.86325e-257
## ENSG00000132563.15   55071.8        6.57917  0.192586   34.1622 8.80726e-256
##                           padj
##                      <numeric>
## ENSG00000105613.9  1.60871e-304
## ENSG00000196361.9  6.10856e-288
## ENSG00000008735.13 1.69389e-264
```

```
## ENSG00000157851.16 5.49993e-253
## ENSG00000109132.6  9.15805e-253
## ENSG00000132563.15 7.58350e-252
```

Genes with:

- Adjusted p-value < 0.05
- $|\log2\text{FoldChange}| > 1$

were considered significantly differentially expressed.

```
# Count upregulated and downregulated genes
upregulated <- sum(res_DTC_vs_MNC$padj < 0.05 & res_DTC_vs_MNC$log2FoldChange > 1, na.rm = TRUE)
downregulated <- sum(res_DTC_vs_MNC$padj < 0.05 & res_DTC_vs_MNC$log2FoldChange < -1, na.rm = TRUE)

upregulated
```

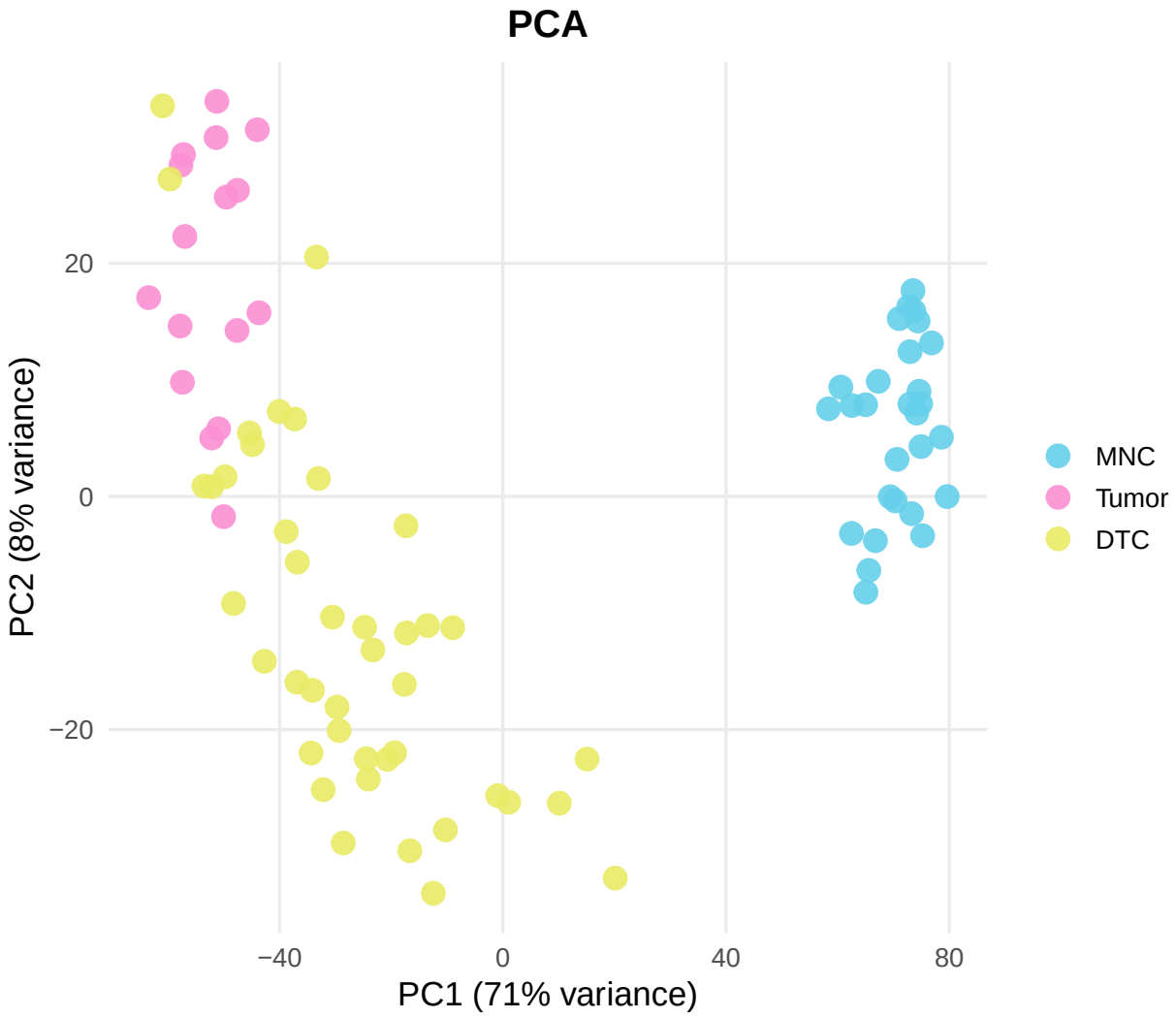
```
## [1] 7340
```

```
downregulated
```

```
## [1] 3172
```

Principal Component Analysis (PCA)

```
vsd <- vst(dds, blind = FALSE)

pcaData <- plotPCA(vsd, intgroup = "condition", returnData = TRUE)
percentVar <- round(100 * attr(pcaData, "percentVar"))

ggplot(pcaData, aes(PC1, PC2, color = condition)) +
  geom_point(size = 4, alpha = 0.9) +
  labs(
    title = "PCA",
    x = paste0("PC1 (", percentVar[1], "% variance)"),
    y = paste0("PC2 (", percentVar[2], "% variance)")
  ) +
  scale_color_manual(values = c(
    "MNC"   = "#65D0EB",
    "Tumor" = "#FC90D4",
    "DTC"   = "#E9EB65"
  )) +

  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    legend.title = element_blank(),
    panel.grid.minor = element_blank()
  )
```

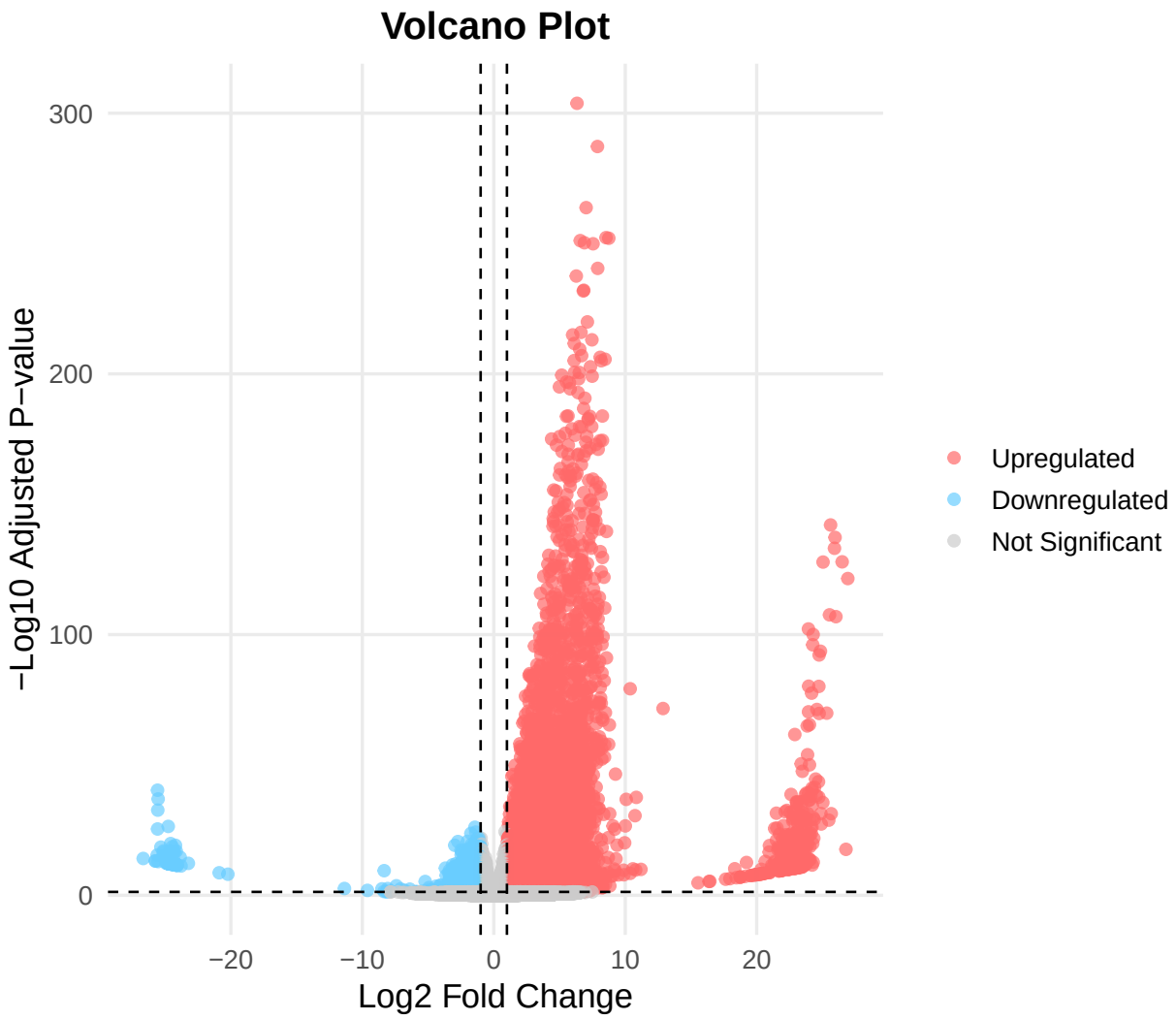


The PCA shows clear separation between DTC and MNC samples along the first two principal components, indicating that DTCs have a distinct global gene expression profile compared to normal mononuclear cells. The variance captured by PC1 and PC2 suggests that the major transcriptional differences are associated with the tumor dissemination state. Tumor samples, when present, cluster between MNCs and DTCs, implying that DTCs may share some transcriptional features with the primary tumor while also activating unique pathways for survival and migration.

Volcano Plot

```
res_df <- as.data.frame(res_DTC_vs_MNC)
res_df <- res_df[!is.na(res_df$padj), ]

# Define DEG categories
res_df$DEG <- "Not Significant"
res_df$DEG[res_df$padj < 0.05 & res_df$log2FoldChange > 1]  <- "Upregulated"
res_df$DEG[res_df$padj < 0.05 & res_df$log2FoldChange < -1] <- "Downregulated"
```

```r
# Set factor order for legend
res_df$DEG <- factor(res_df$DEG,
                     levels = c("Upregulated", "Downregulated", "Not Significant"))

ggplot(res_df, aes(log2FoldChange, -log10(padj), color = DEG)) +

  geom_point(size = 1.8, alpha = 0.7) +

  scale_color_manual(values = c(
    "Upregulated"   = "#FF6969",
    "Downregulated" = "#69CDFF",
    "Not Significant" = "grey80"
  )) +

  geom_vline(xintercept = c(-1, 1),
             linetype = "dashed",
             color = "black",
             linewidth = 0.5) +

  geom_hline(yintercept = -log10(0.05),
             linetype = "dashed",
             color = "black",
             linewidth = 0.5) +

  labs(
    title = "Volcano Plot",
    x = "Log2 Fold Change",
    y = "-Log10 Adjusted P-value"
  ) +

  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    legend.title = element_blank(),
    panel.grid.minor = element_blank()
  )
```

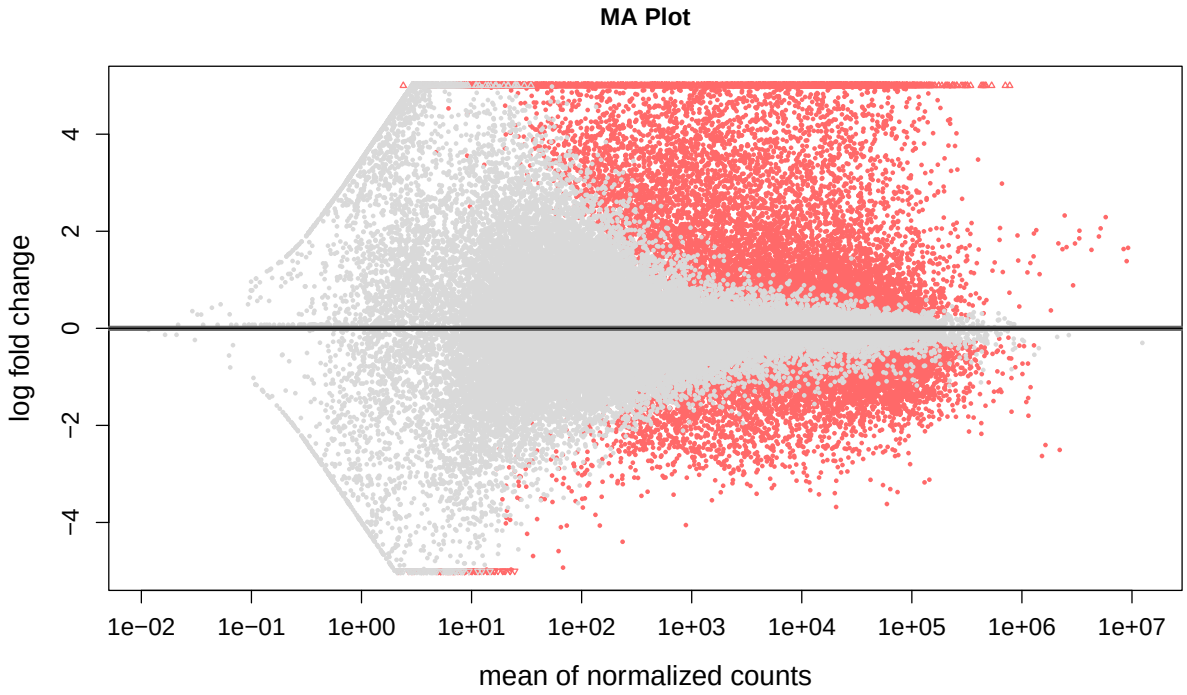


The volcano plot highlights the genes that are significantly up- or downregulated in DTCs compared to MNCs. Genes in red are strongly upregulated in DTCs, potentially reflecting activation of migration, adhesion, and survival pathways, whereas blue genes are downregulated, which may indicate suppression of immune or differentiation-related functions. This plot helps visualize not only the magnitude of expression changes but also their statistical significance, providing a prioritized list of candidate genes for functional studies.

MA Plot

```
par(
  mar = c(5, 5, 4, 2),
  cex.lab = 1.3,
  cex.axis = 1.2
)

plotMA(
  res_DTC_vs_MNC,
  ylim = c(-5, 5),
```

```
  colNonSig = "grey85",
  colSig = "#FF6969",
  cex = 0.5,
  main = "MA Plot"
)

abline(h = 0, col = "black", lwd = 1)
```

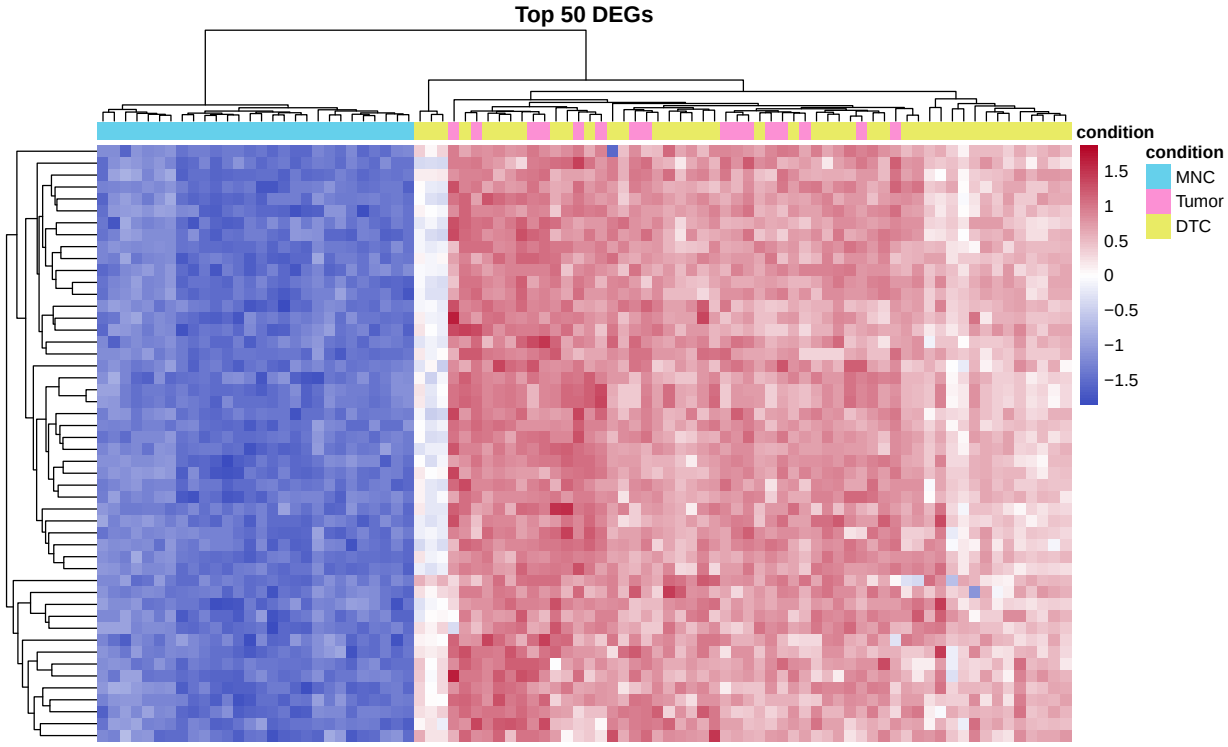


The MA plot displays the relationship between the average expression of genes and their log2 fold change. Highly expressed genes with large fold changes stand out, indicating that DTC-specific regulation affects both abundant and moderate-expression genes. This plot confirms that significant differential expression is not limited to lowly expressed genes, suggesting that DTCs reprogram central transcriptional programs rather than peripheral or rare transcripts.

Heatmap of Top 50 DEGs

```
# Select top 50 DEGs
top_genes <- head(rownames(res_DTC_vs_MNC), 50)

# Extract normalized expression matrix
mat <- assay(vsd)[top_genes, ]

# Ensure correct factor order
coldata$condition <- factor(coldata$condition,
                            levels = c("MNC", "Tumor", "DTC"))

ann_colors <- list(
```

```
  condition = c(
    "MNC"   = "#65D0EB",
    "Tumor" = "#FC90D4",
    "DTC"   = "#E9EB65"
  )
)

heat_colors <- colorRampPalette(c("#3B4CC0", "white", "#B40426"))(100)

pheatmap(
  mat,
  scale = "row",
  annotation_col = coldata,
  annotation_colors = ann_colors,
  show_rownames = FALSE,
  show_colnames = FALSE,
  cluster_rows = TRUE,
  cluster_cols = TRUE,
  color = heat_colors,
  border_color = NA,
  fontsize = 10,
  main = "Top 50 DEGs"
)
```



The heatmap shows hierarchical clustering of the top 50 differentially expressed genes, separating DTC and MNC samples into distinct clusters. Upregulated genes in DTCs cluster together, potentially representing coordinated transcriptional programs for survival and metastatic competence. Downregulated genes may correspond to pathways that are normally active in MNCs but are suppressed in DTCs. This visualization confirms the consistency and robustness of the identified DEGs across samples.

Results

Differential expression analysis of 53,747 genes with nonzero total read count identified 7,340 genes significantly upregulated and 3,172 genes significantly downregulated in DTCs compared to MNCs (adjusted p-value < 0.05, $|\log2\text{FoldChange}| > 1$). Out of the total, only 2,145 genes (4%) were filtered out due to low counts, and no outliers were replaced. The PCA shows clear separation between DTC and MNC samples, indicating distinct global transcriptional profiles. The volcano and MA plots highlight the magnitude and significance of differential expression across the transcriptome, while the heatmap of the top 50 DEGs demonstrates consistent clustering by sample type. Together, these analyses confirm a robust transcriptional distinction between DTCs and MNCs, capturing widespread gene expression differences even without functional pathway annotation.