

GSE211398_race_analysis

Temitope

2025-11-24

R Markdown

This dataset is gotten from the public dataset titled “Comparative gene expression profiling analysis of RNA-seq data for pancreatic adenocarcinoma specimens obtained from Black and White patients”. <https://pubmed.ncbi.nlm.nih.gov/36812168/>

This analysis seek to answer the question: does the tumour effect differ by race?

```
# 1. Load required libraries
library(DESeq2)

## Loading required package: S4Vectors

## Loading required package: stats4

## Loading required package: BiocGenerics

## Loading required package: generics

##
## Attaching package: 'generics'

## The following objects are masked from 'package:base':
## 
##     as.difftime, as.factor, as.ordered, intersect, is.element, setdiff,
##     setequal, union

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':
## 
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
## 
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, saveRDS, table, tapply, unique,
##     unsplit, which.max, which.min
```

```

## 
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:utils':
## 
##     findMatches

## The following objects are masked from 'package:base':
## 
##     expand.grid, I, unname

## Loading required package: IRanges

## 
## Attaching package: 'IRanges'

## The following object is masked from 'package:grDevices':
## 
##     windows

## Loading required package: GenomicRanges

## Loading required package: GenomeInfoDb

## Loading required package: SummarizedExperiment

## Loading required package: MatrixGenerics

## Loading required package: matrixStats

## 
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
## 
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

```

```

## Loading required package: Biobase

## Welcome to Bioconductor
##
##      Vignettes contain introductory material; view with
##      'browseVignettes()'. To cite Bioconductor, see
##      'citation("Biobase")', and for packages 'citation("pkgname")'.

##
## Attaching package: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##       rowMedians

## The following objects are masked from 'package:matrixStats':
##       anyMissing, rowMedians

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.5.2

library(pheatmap)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:Biobase':
##       combine

## The following object is masked from 'package:matrixStats':
##       count

## The following objects are masked from 'package:GenomicRanges':
##       intersect, setdiff, union

## The following object is masked from 'package:GenomeInfoDb':
##       intersect

## The following objects are masked from 'package:IRanges':
##       collapse, desc, intersect, setdiff, slice, union

```

```

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, setequal, union

## The following object is masked from 'package:generics':
##
##     explain

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggrepel)
library(AnnotationDbi)

##
## Attaching package: 'AnnotationDbi'

## The following object is masked from 'package:dplyr':
##
##     select

library(org.Hs.eg.db)

##
# 2. Read count data and convert gene IDs to symbols
count_data <- read.delim("GSE211398_raw_counts_GRCh38.p13_NCBItab.txt", row.names = 1)
count_data <- count_data[rowSums(count_data) > 0, ]

# Convert gene IDs to symbols
convert_gene_ids <- function(gene_ids) {
  # Remove version numbers if present
  clean_ids <- gsub("\\..*", "", gene_ids)

  # Map to gene symbols
  symbols <- mapIds(org.Hs.eg.db,
    keys = clean_ids,
    column = "SYMBOL",
    keytype = "ENTREZID",
    multiVals = "first")
  return(symbols)
}

```

```

# Apply conversion
gene_symbols <- convert_gene_ids(rownames(count_data))

## 'select()' returned 1:1 mapping between keys and columns

# Create a data frame to handle duplicates properly
gene_mapping <- data.frame(
  original_id = rownames(count_data),
  symbol = gene_symbols,
  stringsAsFactors = FALSE
)

# Remove rows where symbol conversion failed
gene_mapping <- gene_mapping[!is.na(gene_mapping$symbol), ]

# Calculate mean expression for each gene
mean_expression <- rowMeans(count_data)

# Add mean expression to mapping
gene_mapping$mean_expr <- mean_expression[gene_mapping$original_id]

# For duplicate symbols, keep the one with highest mean expression
gene_mapping <- gene_mapping %>%
  group_by(symbol) %>%
  arrange(desc(mean_expr)) %>%
  slice(1) %>% # Keep first row (highest expression) for each symbol
  ungroup()

# Create new count matrix with unique gene symbols
count_data_symbols <- count_data[gene_mapping$original_id, ]
rownames(count_data_symbols) <- gene_mapping$symbol

print(paste("Original genes:", nrow(count_data)))

## [1] "Original genes: 30384"

print(paste("Genes with unique symbols:", nrow(count_data_symbols)))

## [1] "Genes with unique symbols: 29239"

print("Sample of gene symbols:")

## [1] "Sample of gene symbols:"

print(head(rownames(count_data_symbols), 20))

## [1] "A1BG"      "A1BG-AS1"   "A1CF"      "A2M"       "A2M-AS1"   "A2ML1"
## [7] "A2MP1"     "A3GALT2"    "A4GALT"    "A4GNT"     "AAAS"      "AACS"
## [13] "AACSP1"    "AADAC"     "AADACL2"   "AADACL3"   "AADACL4"   "AADACP1"
## [19] "AADAT"     "AAGAB"

```

```

# Check if we have common pancreatic cancer genes
pancreatic_genes <- c("KRAS", "TP53", "CDKN2A", "SMAD4", "TSPAN8", "AGR2", "POSTN", "TFF1", "CP")

found_genes <- pancreatic_genes[pancreatic_genes %in% rownames(count_data_symbols)]

print(paste("Found pancreatic cancer genes:", paste(found_genes, collapse = ", ")))

## [1] "Found pancreatic cancer genes: KRAS, TP53, CDKN2A, SMAD4, TSPAN8, AGR2, POSTN, TFF1, CP"

# 3. EXTRACT METADATA FROM SERIES_MATRIX FILE
series_matrix <- readLines("GSE211398_series_matrix.txt")

# Extract sample titles
sample_title_line <- grep("!Sample_title", series_matrix, value = TRUE)
sample_titles <- unlist(strsplit(gsub('!Sample_title\t"', '', sample_title_line), '\t'))
sample_titles <- gsub("^", "", sample_titles)

# Extract sample GEO accessions
sample_geo_line <- grep("!Sample_geo_accession", series_matrix, value = TRUE)
sample_geo <- unlist(strsplit(gsub('!Sample_geo_accession\t"', '', sample_geo_line), '\t'))
sample_geo <- gsub("^", "", sample_geo)

# Extract characteristics
characteristics_lines <- grep("!Sample_characteristics_ch1", series_matrix, value = TRUE)

# Parse characteristics
extract_characteristics <- function(line) {
  content <- gsub('!Sample_characteristics_ch1\t"', '', line)
  content <- gsub("^", "", content)
  unlist(strsplit(content, '\t'))
}

tissue_info <- extract_characteristics(characteristics_lines[1])
race_info <- extract_characteristics(characteristics_lines[2])
gender_info <- extract_characteristics(characteristics_lines[3])
age_info <- extract_characteristics(characteristics_lines[4])

# Clean and create metadata
metadata <- data.frame(
  sample_id = sample_geo,
  tissue = ifelse(grepl("adenocarcinoma", tissue_info), "Tumor", "NonTumor"),
  race = gsub("race: ", "", race_info),
  gender = gsub("gender: ", "", gender_info),
  age = as.numeric(gsub("age: ", "", age_info)),
  stringsAsFactors = TRUE
)
rownames(metadata) <- metadata$sample_id

# Reorder metadata to match count data
metadata <- metadata[colnames(count_data_symbols), ]

# Verify metadata
print("Sample distribution:")

```

```

## [1] "Sample distribution:"
```

```

print(table(metadata$tissue, metadata$race))
```

```

##
##          Black White
## NonTumor      3     9
## Tumor         5    11
```

```

head(metadata)
```

```

##           sample_id tissue race gender age
## GSM6469488 GSM6469488 Tumor White Male 72.8
## GSM6469489 GSM6469489 Tumor White Male 78.5
## GSM6469490 GSM6469490 Tumor White Male 76.8
## GSM6469491 GSM6469491 Tumor White Female 56.2
## GSM6469492 GSM6469492 Tumor Black Female 73.4
## GSM6469493 GSM6469493 Tumor White Male 67.2
```

```

# Check sample distribution
sample_summary <- metadata %>%
  group_by(tissue, race) %>%
  summarise(
    n = n(),
    mean_age = mean(age, na.rm = TRUE),
    sd_age = sd(age, na.rm = TRUE),
    .groups = 'drop'
  )
print(sample_summary)
```

```

## # A tibble: 4 x 5
##   tissue   race     n mean_age sd_age
##   <fct>   <fct> <int>    <dbl>   <dbl>
## 1 NonTumor Black     3     39.3   18.1
## 2 NonTumor White    9     47.7   11.7
## 3 Tumor    Black     5     66.8   11.2
## 4 Tumor    White    11     67.2   7.99
```

```

# 4. Create DESeq2 object
dds <- DESeqDataSetFromMatrix(
  countData = count_data_symbols,
  colData = metadata,
  design = ~ tissue + race + tissue:race
)
```

```

# 5. Run DESeq2
dds <- DESeq(dds)
```

```

## estimating size factors
```

```

## estimating dispersions
```

```

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## -- replacing outliers and refitting for 541 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing

```

6. ALL KEY COMPARISONS

```

## 6a. Tumor vs Non-Tumor (across all races)
res_tumor_vs_normal <- results(dds, name = "tissue_Tumor_vs_NonTumor")
sig_tumor <- res_tumor_vs_normal[res_tumor_vs_normal$padj < 0.05 & !is.na(res_tumor_vs_normal$padj), ]
sig_tumor <- sig_tumor[order(sig_tumor$padj), ]

print(paste("Number of DEGs between tumor and non-tumor:", nrow(sig_tumor)))

## [1] "Number of DEGs between tumor and non-tumor: 2582"

print("Top 10 tumor vs non-tumor genes:")

## [1] "Top 10 tumor vs non-tumor genes:"


top_tumor_genes <- as.data.frame(head(sig_tumor, 10))
top_tumor_genes$gene_symbol <- rownames(top_tumor_genes)
print(top_tumor_genes[, c("gene_symbol", "log2FoldChange", "padj")])

##          gene_symbol log2FoldChange      padj
## 1 IGHV3-15       IGHV3-15 21.85430 4.267674e-24
## 2 IGHV3-21       IGHV3-21 20.79325 1.773988e-22
## 3 MUC17          MUC17   27.11863 1.773988e-22
## 4 IGKV1D-33     IGKV1D-33 22.62090 7.880731e-21
## 5 KRT16          KRT16   23.58299 8.593489e-20
## 6 UTY             UTY    25.21620 8.593489e-20
## 7 IGLV2-18       IGLV2-18 20.48105 3.177769e-19
## 8 P2RY10         P2RY10   19.06878 3.210142e-19
## 9 FCRL5          FCRL5   19.63746 5.536206e-19
## 10 MUC2           MUC2    22.33241 1.252835e-18

```

```

## 6b. Race effect (across all tissues)
res_race <- results(dds, name = "race_White_vs_Black")
sig_race <- res_race[res_race$padj < 0.05 & !is.na(res_race$padj), ]
sig_race <- sig_race[order(sig_race$padj), ]

print(paste("Number of DEGs between races:", nrow(sig_race)))

## [1] "Number of DEGs between races: 334"

print("Top 10 race-associated genes:")

## [1] "Top 10 race-associated genes:"

top_race_genes <- as.data.frame(head(sig_race, 20))
top_race_genes$gene_symbol <- rownames(top_race_genes)
print(top_race_genes[, c("gene_symbol", "log2FoldChange", "padj")])

##          gene_symbol log2FoldChange      padj
## 1      IGHV3-15     24.05654 1.191766e-33
## 2      IGHV3-21     21.66123 1.167015e-27
## 3       FCRL5      21.05130 1.521716e-24
## 4    IGKV1D-33     22.83885 2.002550e-24
## 5    IGKV1-16     22.88549 8.773786e-23
## 6        UTY       25.08795 8.773786e-23
## 7    ADGRG7      21.57027 5.409480e-22
## 8    P2RY10      19.20953 5.409480e-22
## 9   IGLV2-18     20.33955 1.126361e-21
## 10   IGHV4-4     21.61821 2.155190e-21
## 11    IGKJ5      21.71431 3.504131e-21
## 12      MUC2      21.31686 1.711034e-19
## 13  IGHV10R15-1   IGHV10R15-1    20.33627 3.115665e-19
## 14 LOC102724760 LOC102724760    20.22679 6.187598e-18
## 15      PKP1      PKP1      20.17533 3.554007e-16
## 16      BTLA      BTLA      19.58205 6.866188e-16
## 17      KRT16      KRT16      19.76065 6.866188e-16
## 18      MUC17      MUC17      20.94462 1.744367e-15
## 19      TIMD4      TIMD4      20.07639 1.861585e-12
## 20     ELK2AP     ELK2AP      19.54860 6.416125e-12

## 6c. INTERACTION: Does tumor effect differ by race?
res_interaction <- results(dds, name = "tissueTumor.raceWhite")
sig_interaction <- res_interaction[res_interaction$padj < 0.05 & !is.na(res_interaction$padj), ]
sig_interaction <- sig_interaction[order(sig_interaction$padj), ]

print(paste("Number of genes where tumor effect differs by race:", nrow(sig_interaction)))

## [1] "Number of genes where tumor effect differs by race: 123"

```

```

print("Top 10 interaction genes:")

## [1] "Top 10 interaction genes:"


top_interaction_genes <- as.data.frame(head(sig_interaction, 10))
top_interaction_genes$gene_symbol <- rownames(top_interaction_genes)
print(top_interaction_genes[, c("gene_symbol", "log2FoldChange", "padj")])

##          gene_symbol log2FoldChange      padj
## IGHV3-15     IGHV3-15    -22.06786 2.837554e-19
## UTY          UTY        -26.09945 1.493746e-15
## FCRL5        FCRL5     -19.91086 6.122298e-15
## IGHV3-21     IGHV3-21    -19.21337 6.122298e-15
## IGKV1D-33    IGKV1D-33   -21.26518 6.174927e-14
## P2RY10       P2RY10     -18.39733 8.516423e-14
## ADGRG7        ADGRG7    -20.32747 9.860211e-13
## MUC2          MUC2      -21.02201 2.983747e-12
## IGLV2-18      IGLV2-18   -18.65143 4.010054e-12
## IGKV1-16      IGKV1-16   -20.52867 6.380199e-12

```

VISUALIZING ALL COMPARISONS

```

## 7a. Volcano plot for Tumor vs Non-Tumor
volcano_tumor <- as.data.frame(res_tumor_vs_normal)
volcano_tumor$gene <- rownames(volcano_tumor)
volcano_tumor$significant <- volcano_tumor$padj < 0.05 & abs(volcano_tumor$log2FoldChange) > 1
volcano_tumor$significant[is.na(volcano_tumor$significant)] <- FALSE

# Label top genes
volcano_tumor$label <- ifelse(volcano_tumor$padj < 0.001 & abs(volcano_tumor$log2FoldChange) > 2,
                                 volcano_tumor$gene, "")

ggplot(volcano_tumor, aes(x = log2FoldChange, y = -log10(pvalue))) +
  geom_point(aes(color = significant), alpha = 0.6) +
  geom_text_repel(aes(label = label), size = 3, max.overlaps = 10) +
  scale_color_manual(values = c("grey", "red")) +
  labs(title = "Tumor vs Non-Tumor DEGs",
       x = "Log2 Fold Change (Tumor/NonTumor)",
       y = "-Log10 P-value") +
  theme_minimal() +
  theme(legend.position = "none")

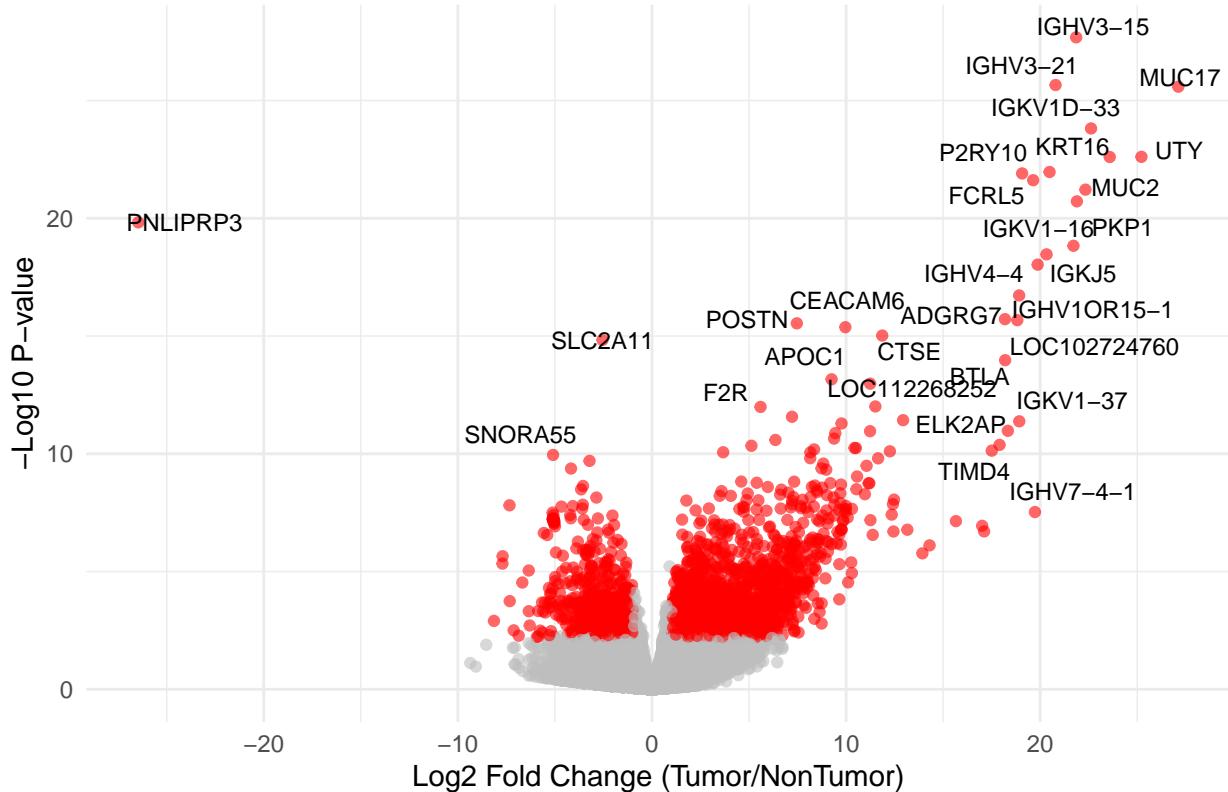
## Warning: Removed 168 rows containing missing values or values outside the scale range
## ('geom_point()').

## Warning: Removed 3891 rows containing missing values or values outside the scale range
## ('geom_text_repel()').

## Warning: ggrepel: 451 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps

```

Tumor vs Non-Tumor DEGs



```

## 7b. Volcano plot for Race effect
volcano_race <- as.data.frame(res_race)
volcano_race$gene <- rownames(volcano_race)
volcano_race$significant <- volcano_race$padj < 0.05 & abs(volcano_race$log2FoldChange) > 1
volcano_race$significant[is.na(volcano_race$significant)] <- FALSE

volcano_race$label <- ifelse(volcano_race$padj < 0.001 & abs(volcano_race$log2FoldChange) > 2,
                                volcano_race$gene, "")

ggplot(volcano_race, aes(x = log2FoldChange, y = -log10(pvalue))) +
  geom_point(aes(color = significant), alpha = 0.6) +
  geom_text_repel(aes(label = label), size = 3, max.overlaps = 10) +
  scale_color_manual(values = c("grey", "blue")) +
  labs(title = "White vs Black DEGs",
       x = "Log2 Fold Change (White/Black)",
       y = "-Log10 P-value") +
  theme_minimal() +
  theme(legend.position = "none")

```

```

## Warning: Removed 168 rows containing missing values or values outside the scale range
## ('geom_point()').

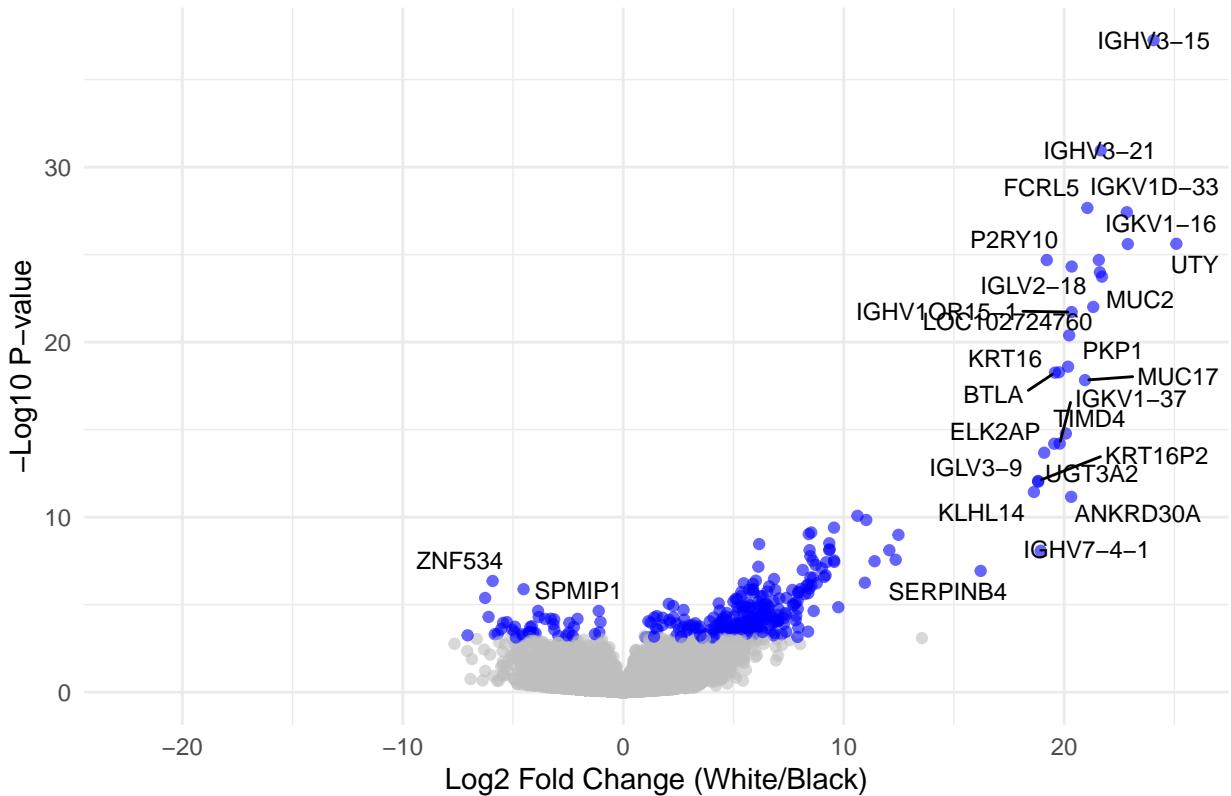
## Warning: Removed 1126 rows containing missing values or values outside the scale range
## ('geom_text_repel()').

## Warning: ggrepel: 57 unlabeled data points (too many overlaps). Consider

```

```
## increasing max.overlaps
```

White vs Black DEGs



```
## 7c. Volcano plot for Interaction
volcano_interaction <- as.data.frame(res_interaction)
volcano_interaction$gene <- rownames(volcano_interaction)
volcano_interaction$significant <- volcano_interaction$padj < 0.05 & abs(volcano_interaction$log2FoldCh
volcano_interaction$significant[is.na(volcano_interaction$significant)] <- FALSE

volcano_interaction$label <- ifelse(volcano_interaction$padj < 0.001 & abs(volcano_interaction$log2FoldCh
                                         volcano_interaction$gene, "")

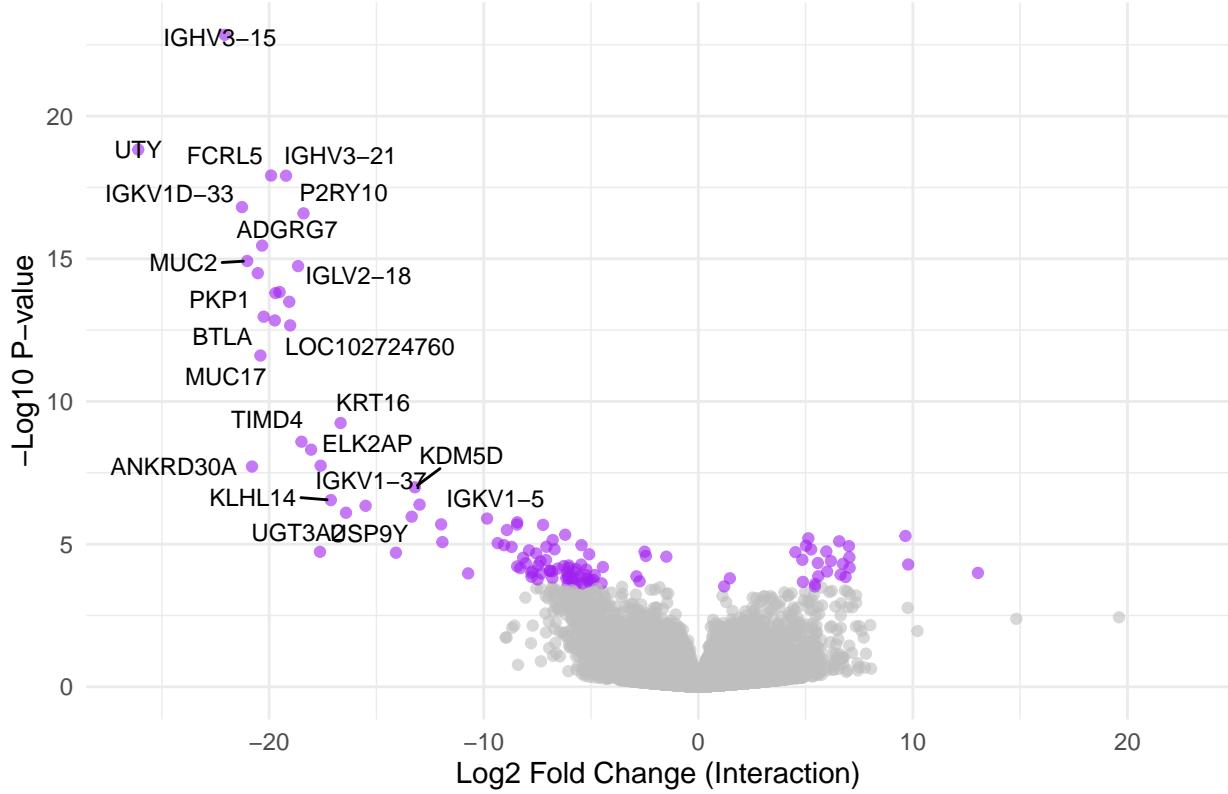
ggplot(volcano_interaction, aes(x = log2FoldChange, y = -log10(pvalue))) +
  geom_point(aes(color = significant), alpha = 0.6) +
  geom_text_repel(aes(label = label), size = 3, max.overlaps = 10) +
  scale_color_manual(values = c("grey", "purple")) +
  labs(title = "Interaction: Genes where tumor effect differs by race",
       x = "Log2 Fold Change (Interaction)",
       y = "-Log10 P-value") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## Warning: Removed 168 rows containing missing values or values outside the scale range
## ('geom_point()').
```

```
## Warning: Removed 2242 rows containing missing values or values outside the scale range
## ('geom_text_repel()').
```

```
## Warning: ggrepel: 6 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

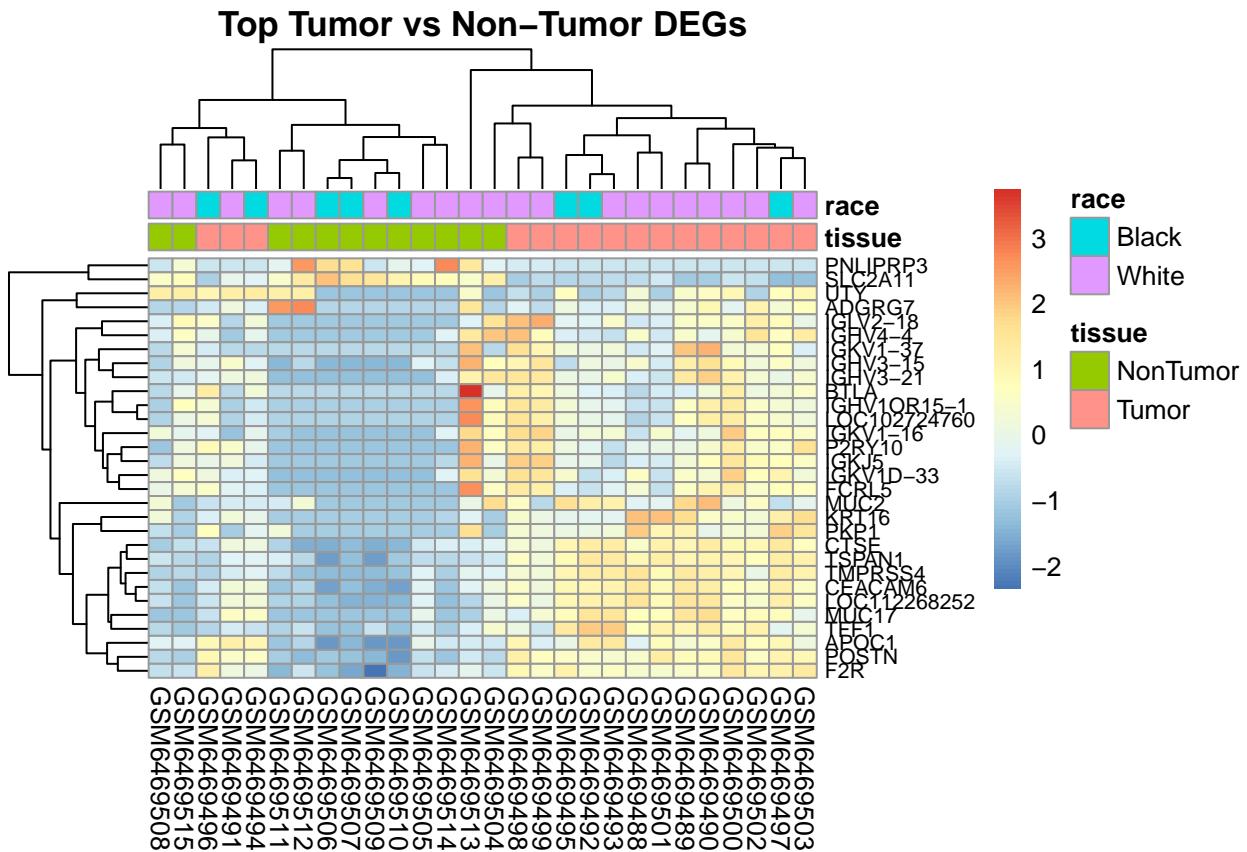
Interaction: Genes where tumor effect differs by race



8. HEATMAPS FOR EACH COMPARISON

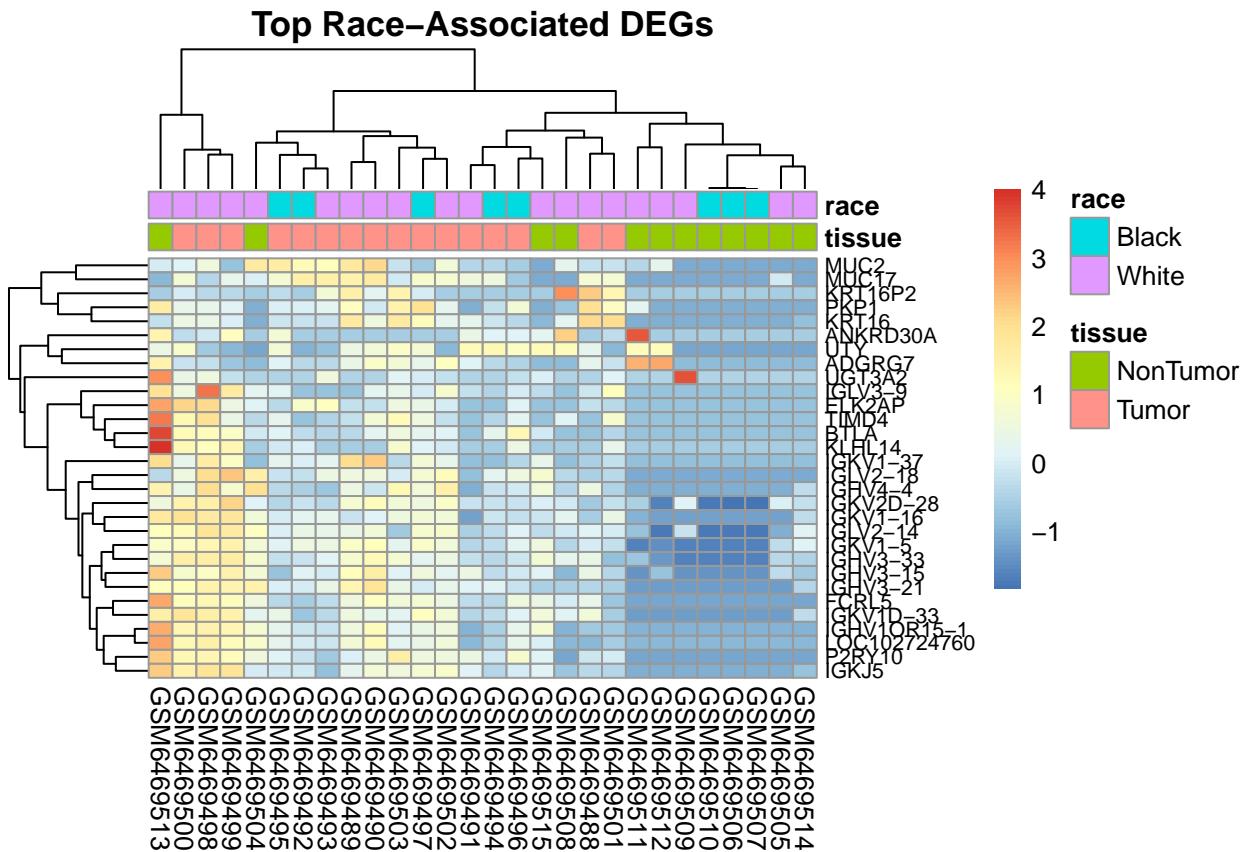
```
## 8a. Heatmap for Tumor vs Non-Tumor DEGs
if (nrow(sig_tumor) > 0) {
  vsd <- vst(dds, blind = FALSE)
  top_tumor_genes <- rownames(sig_tumor)[1:min(30, nrow(sig_tumor))]
  mat_tumor <- assay(vsd)[top_tumor_genes, ]
  mat_tumor <- t(scale(t(mat_tumor)))

  pheatmap(mat_tumor,
    annotation_col = metadata[, c("tissue", "race")],
    main = "Top Tumor vs Non-Tumor DEGs",
    show_rownames = TRUE,
    fontsize_row = 8)
}
```



```
## 8b. Heatmap for Race DEGs
if (nrow(sig_race) > 0) {
  top_race_genes <- rownames(sig_race)[1:min(30, nrow(sig_race))]
  mat_race <- assay(vsd)[top_race_genes, ]
  mat_race <- t(scale(t(mat_race)))

  pheatmap(mat_race,
    annotation_col = metadata[, c("tissue", "race")],
    main = "Top Race-Associated DEGs",
    show_rownames = TRUE,
    fontsize_row = 8)
}
```



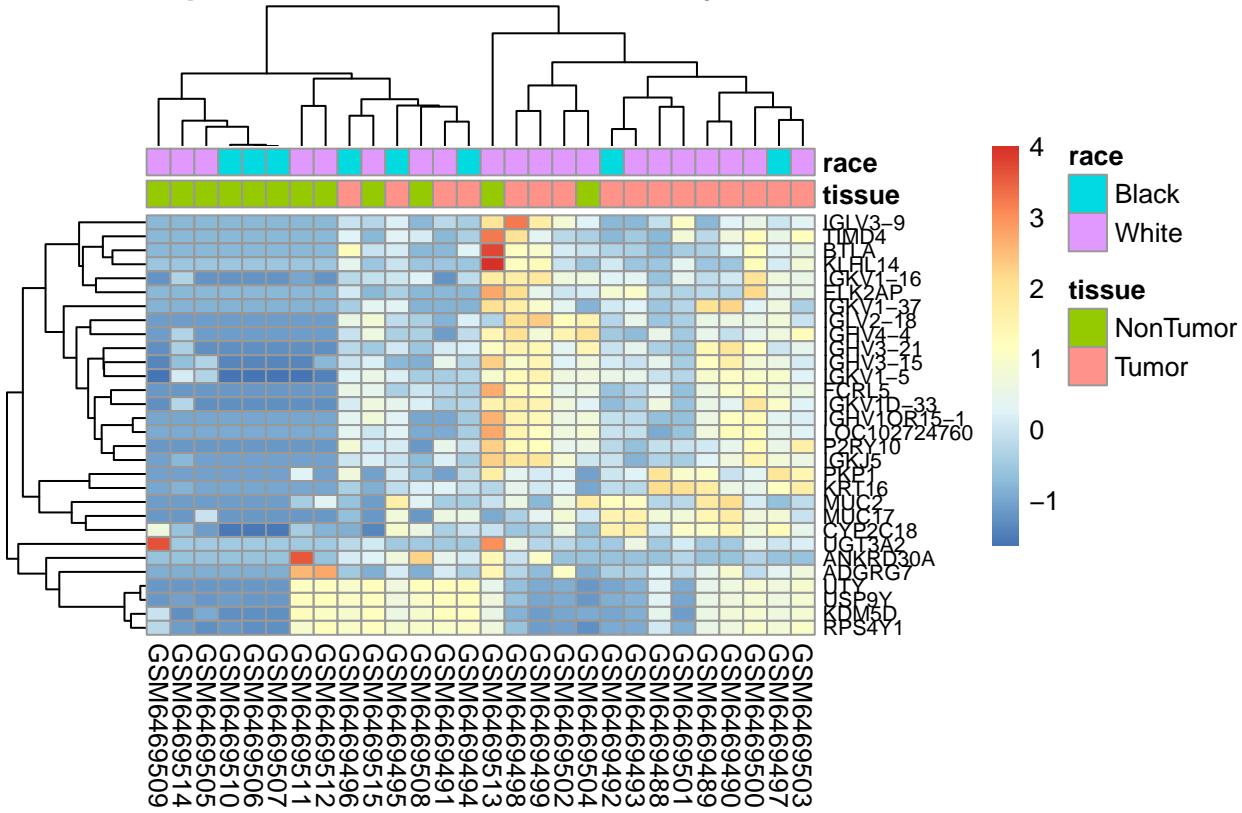
```

## 8c. Heatmap for Interaction genes
if (nrow(sig_interaction) > 0) {
  top_interaction_genes <- rownames(sig_interaction)[1:min(30, nrow(sig_interaction))]
  mat_interaction <- assay(vsd)[top_interaction_genes, ]
  mat_interaction <- t(scale(t(mat_interaction)))

  pheatmap(mat_interaction,
            annotation_col = metadata[, c("tissue", "race")],
            main = "Top Genes: Tumor Effect Differs by Race",
            show_rownames = TRUE,
            fontsize_row = 8)
}

```

Top Genes: Tumor Effect Differs by Race



```
# 9. Checking if SPECIFIC pancreatic GENES are expressed
genes_to_check <- c("TSPAN8", "AGR2", "POSTN", "TFF1", "CP", "KRAS", "TP53", "CDKN2A", "SMAD4")

print("== SPECIFIC GENE ANALYSIS ==")

## [1] "== SPECIFIC GENE ANALYSIS =="

for (gene in genes_to_check) {
  if (gene %in% rownames(count_data_symbols)) {
    # Plot expression
    plot_data <- plotCounts(dds, gene = gene, intgroup = c("tissue", "race"), returnData = TRUE)

    p <- ggplot(plot_data, aes(x = tissue, y = count, color = race, shape = race)) +
      geom_point(position = position_jitter(width = 0.2), size = 3) +
      labs(title = paste("Expression of", gene),
           x = "Tissue", y = "Normalized Counts") +
      theme_minimal()
    print(p) # Print results for this gene
    cat("\n---", gene, "---\n")
    if (gene %in% rownames(res_tumor_vs_normal)) {
      tumor_result <- res_tumor_vs_normal[gene, ]
      direction <- ifelse(tumor_result$log2FoldChange > 0, "UP in Tumor", "DOWN in Tumor")
      cat("Tumor vs Normal: FC =", round(2^tumor_result$log2FoldChange, 2),
          "| padj =", format(tumor_result$padj, scientific = TRUE), "|", direction, "\n")
    }
  }
}
```

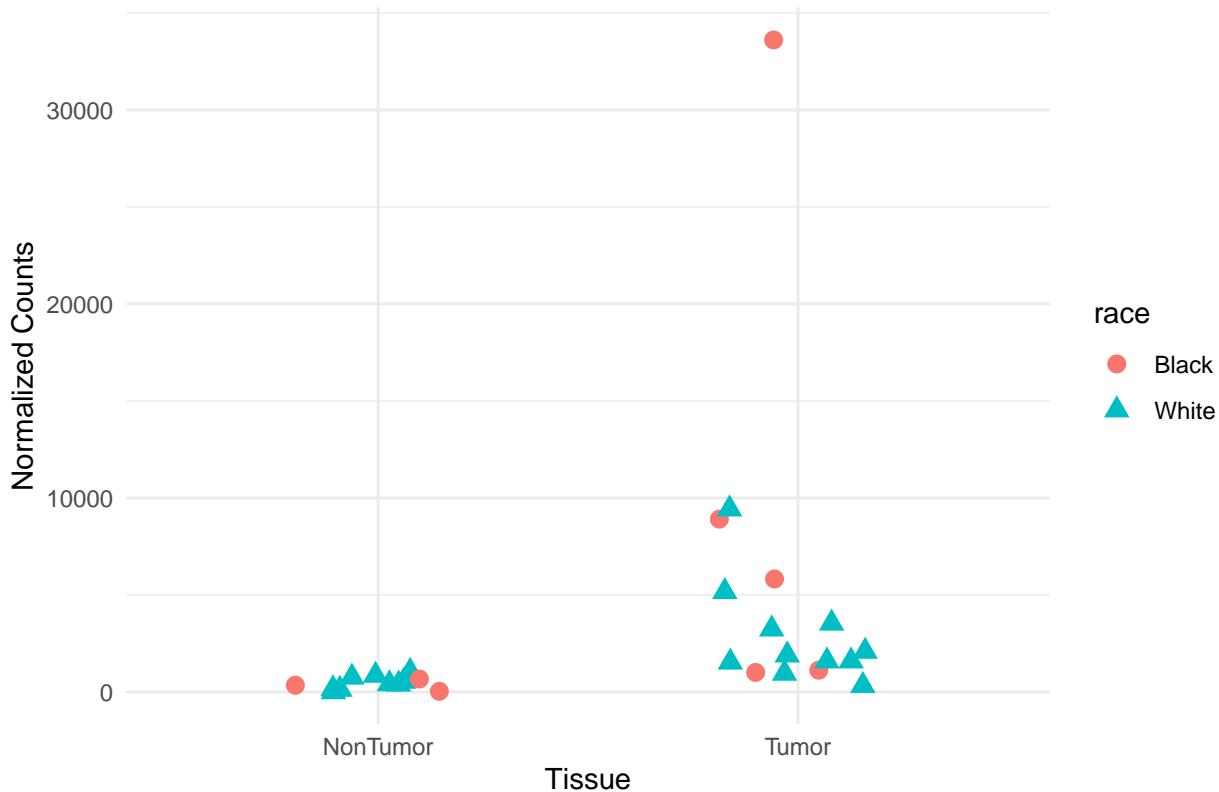
```

if (gene %in% rownames(res_race)) {
  race_result <- res_race[gene, ]
  direction <- ifelse(race_result$log2FoldChange > 0, "UP in White", "UP in Black")
  cat("Race effect: FC =", round(2^race_result$log2FoldChange, 2),
      "| padj =", format(race_result$padj, scientific = TRUE), "|", direction, "\n")
}

if (gene %in% rownames(res_interaction)) {
  interaction_result <- res_interaction[gene, ]
  direction <- ifelse(interaction_result$log2FoldChange > 0, "Stronger in White", "Stronger in Black")
  cat("Interaction: FC =", round(interaction_result$log2FoldChange, 2),
      "| padj =", format(interaction_result$padj, scientific = TRUE), "| Tumor effect", direction,
      "\n")
} else {
  cat("\n", gene, "not found in dataset\n")
}
}

```

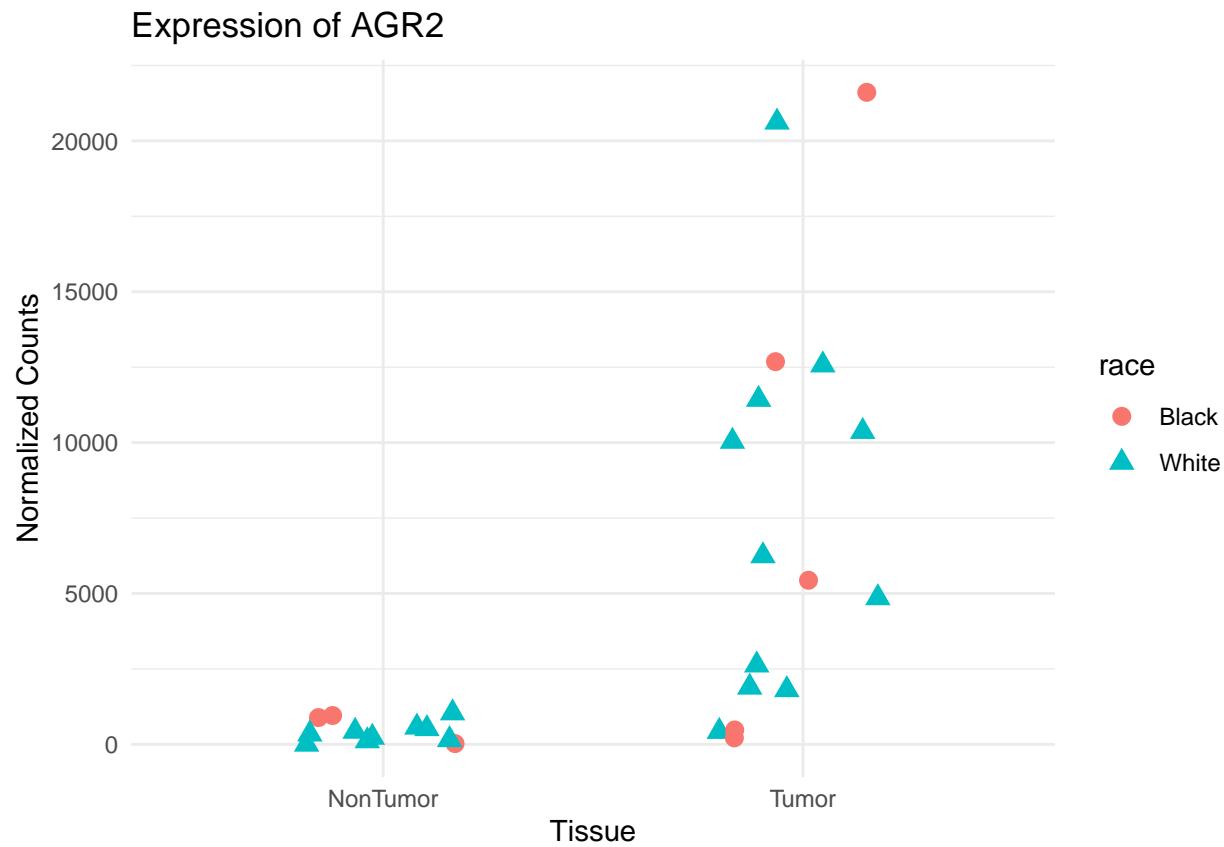
Expression of TSPAN8



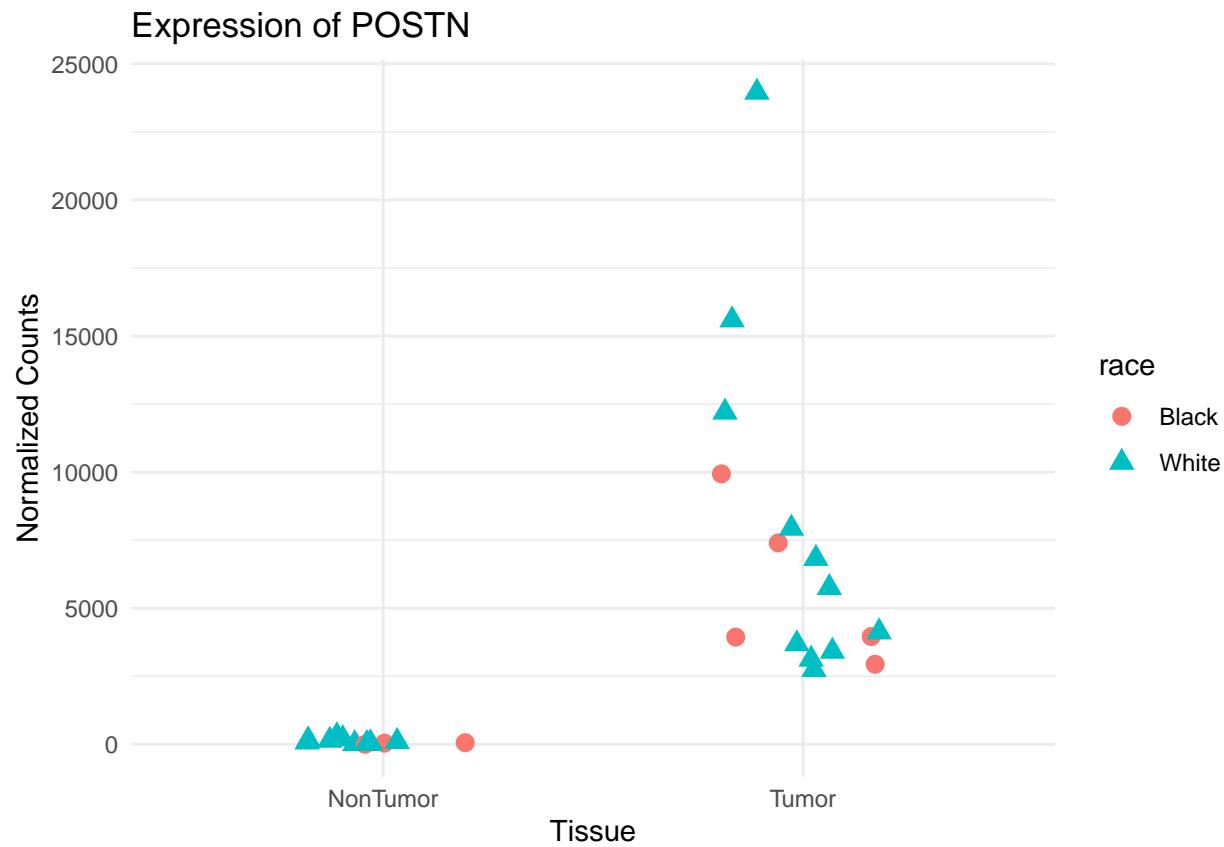
```

## 
## --- TSPAN8 ---
## Tumor vs Normal: FC = 29.17 | padj = 3.386687e-04 | UP in Tumor
## Race effect: FC = 1.38 | padj = 9.622917e-01 | UP in White
## Interaction: FC = -2.29 | padj = 7.432812e-01 | Tumor effect Stronger in Black

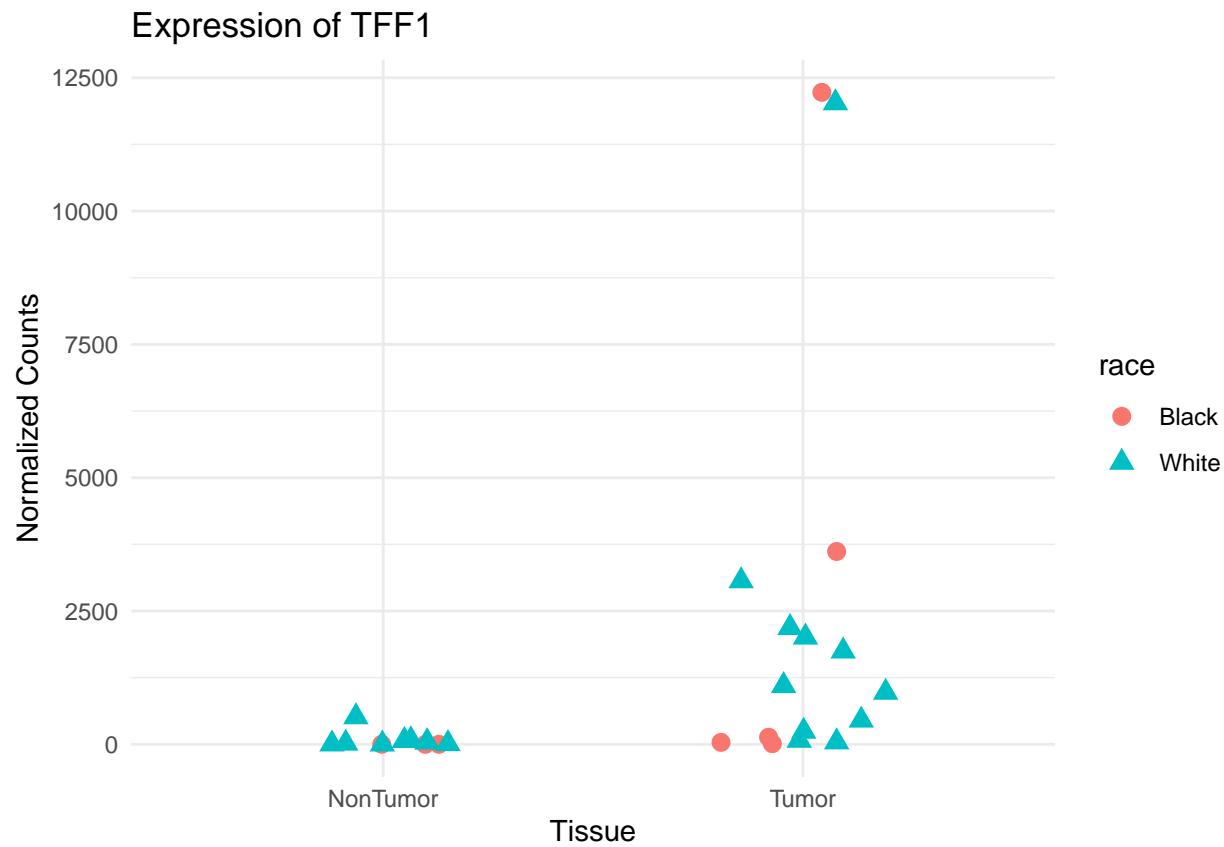
```



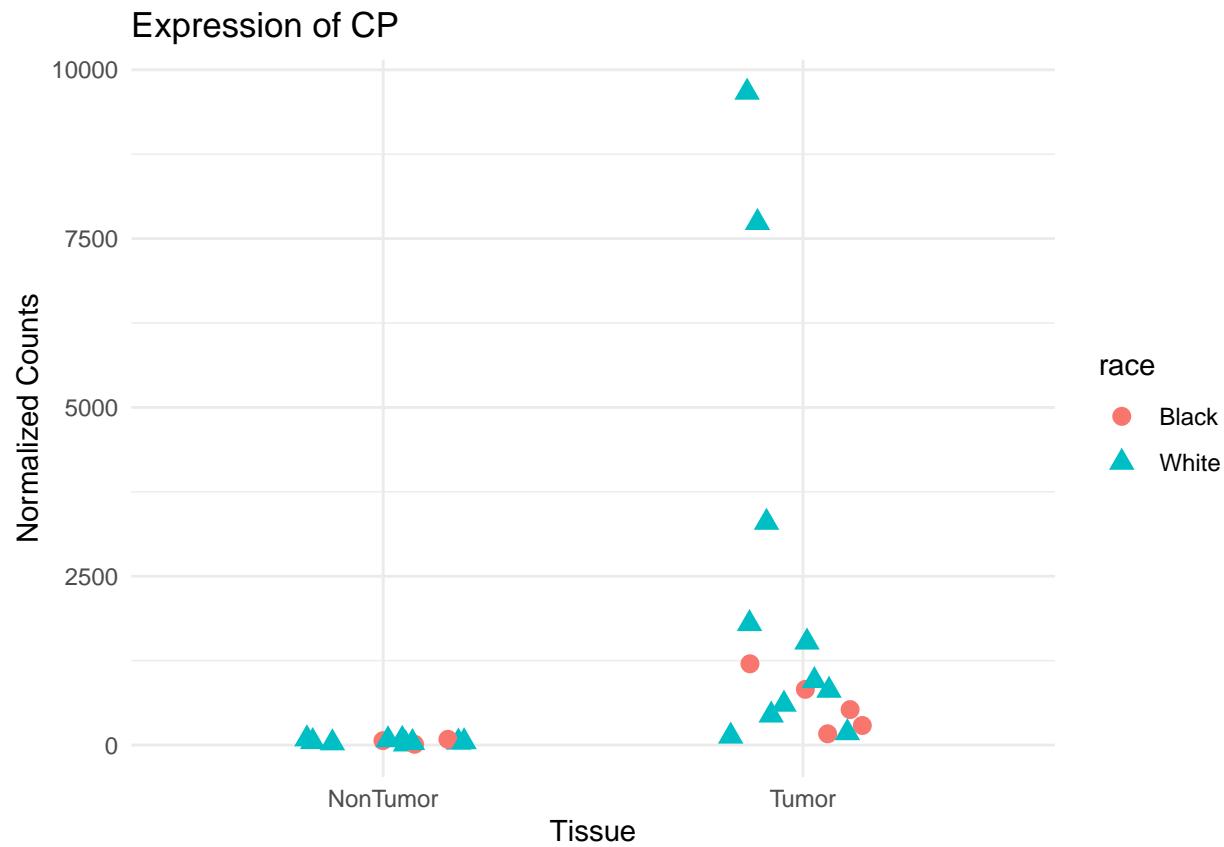
```
##  
## --- AGR2 ---  
## Tumor vs Normal: FC = 13 | padj = 1.834723e-02 | UP in Tumor  
## Race effect: FC = 0.6 | padj = 9.208789e-01 | UP in Black  
## Interaction: FC = 0.63 | padj = 9.969893e-01 | Tumor effect Stronger in White
```



```
##  
## --- POSTN ---  
## Tumor vs Normal: FC = 176.42 | padj = 3.126186e-13 | UP in Tumor  
## Race effect: FC = 3.74 | padj = 3.47969e-01 | UP in White  
## Interaction: FC = -1.38 | padj = 9.079409e-01 | Tumor effect Stronger in Black
```

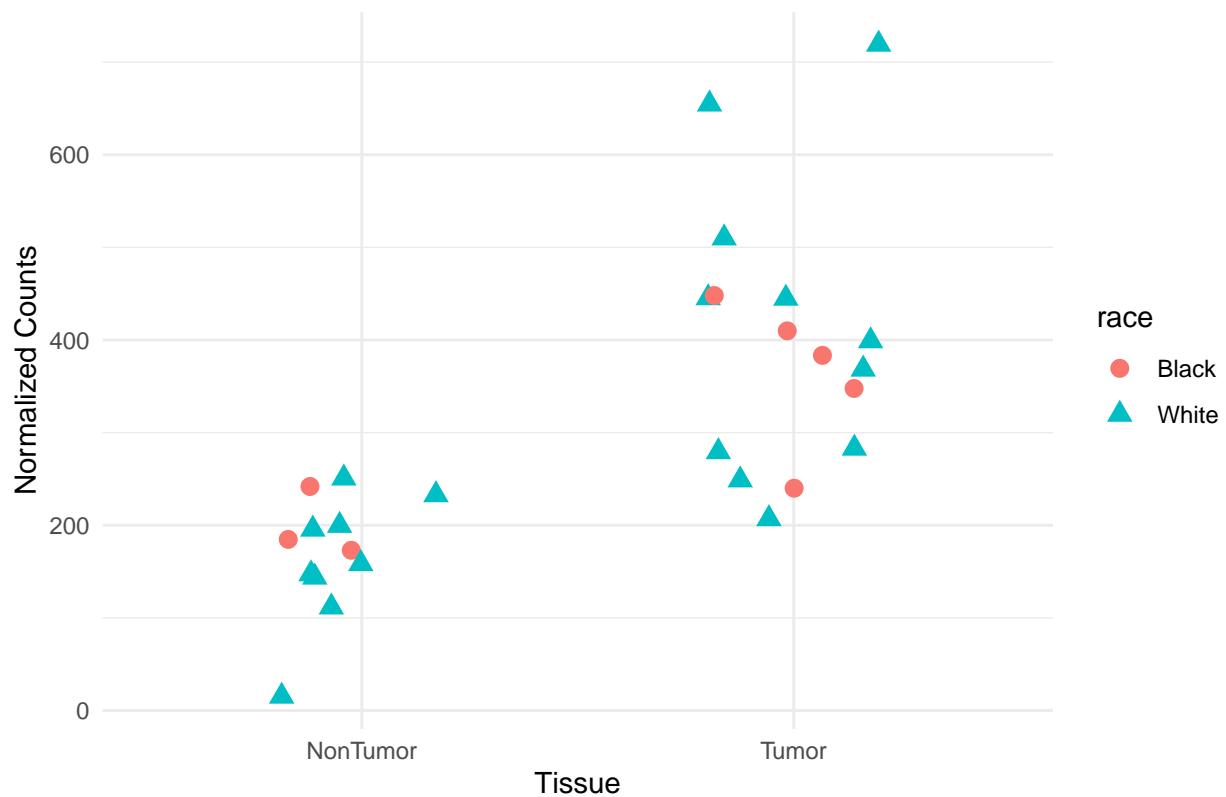


```
##
## --- TFF1 ---
## Tumor vs Normal: FC = 7838.04 | padj = 2.657379e-09 | UP in Tumor
## Race effect: FC = 213.15 | padj = 2.249158e-03 | UP in White
## Interaction: FC = -8.3 | padj = 1.779011e-02 | Tumor effect Stronger in Black
```



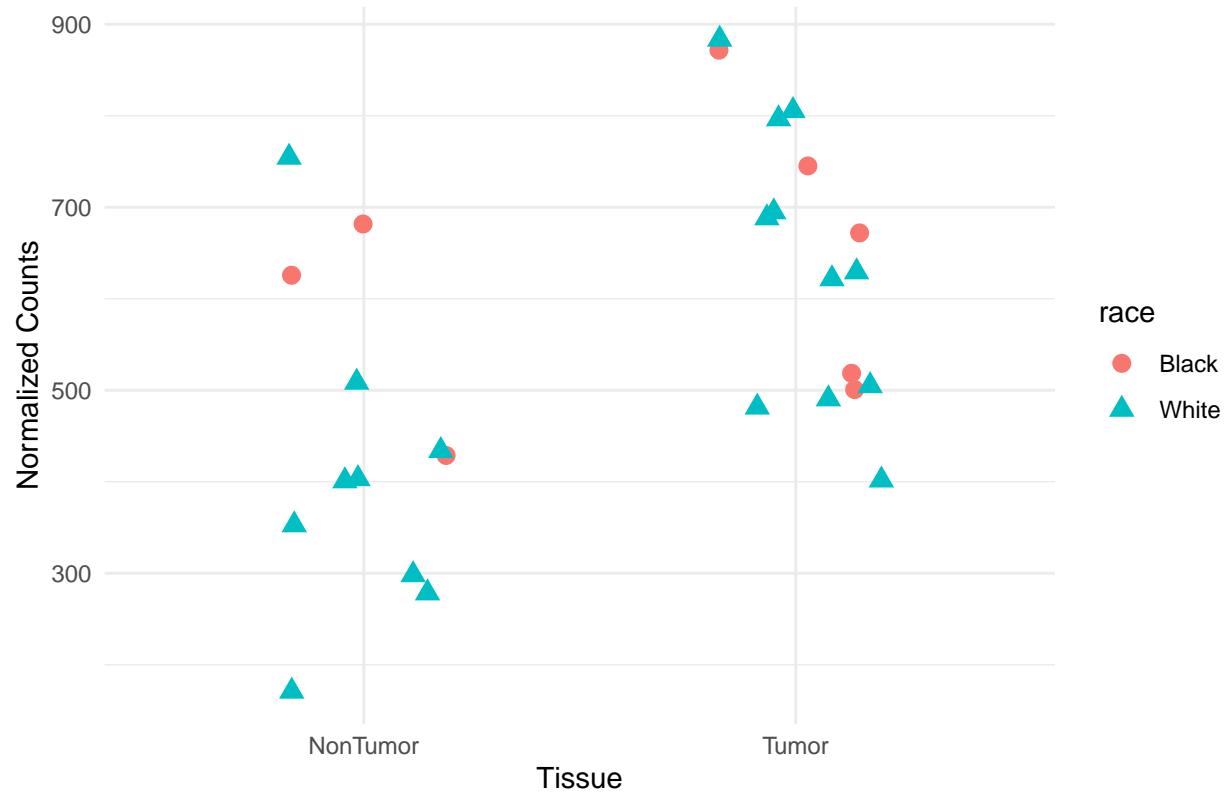
```
##  
## --- CP ---  
## Tumor vs Normal: FC = 11.49 | padj = 6.314581e-03 | UP in Tumor  
## Race effect: FC = 0.98 | padj = 9.97813e-01 | UP in Black  
## Interaction: FC = 2.07 | padj = 7.461765e-01 | Tumor effect Stronger in White
```

Expression of KRAS



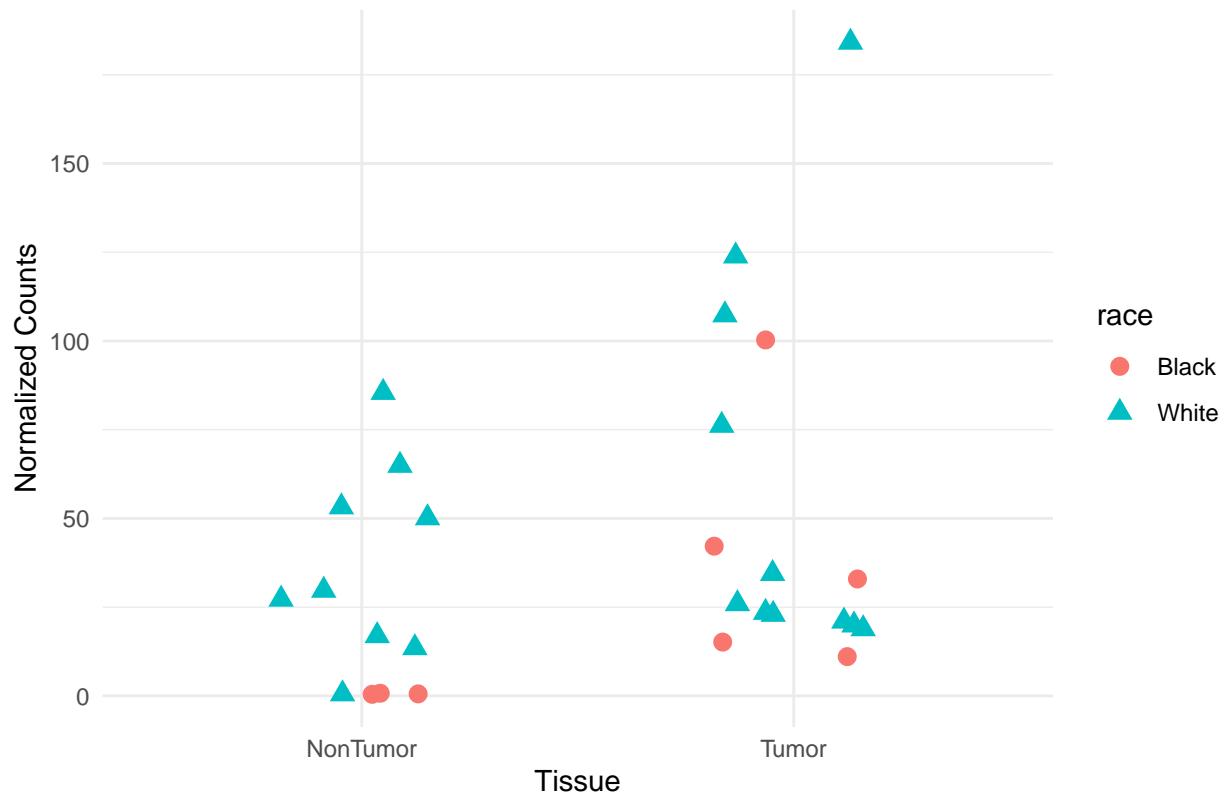
```
##  
## --- KRAS ---  
## Tumor vs Normal: FC = 1.84 | padj = 2.110918e-01 | UP in Tumor  
## Race effect: FC = 0.81 | padj = 9.197895e-01 | UP in Black  
## Interaction: FC = 0.48 | padj = 9.867653e-01 | Tumor effect Stronger in White
```

Expression of TP53

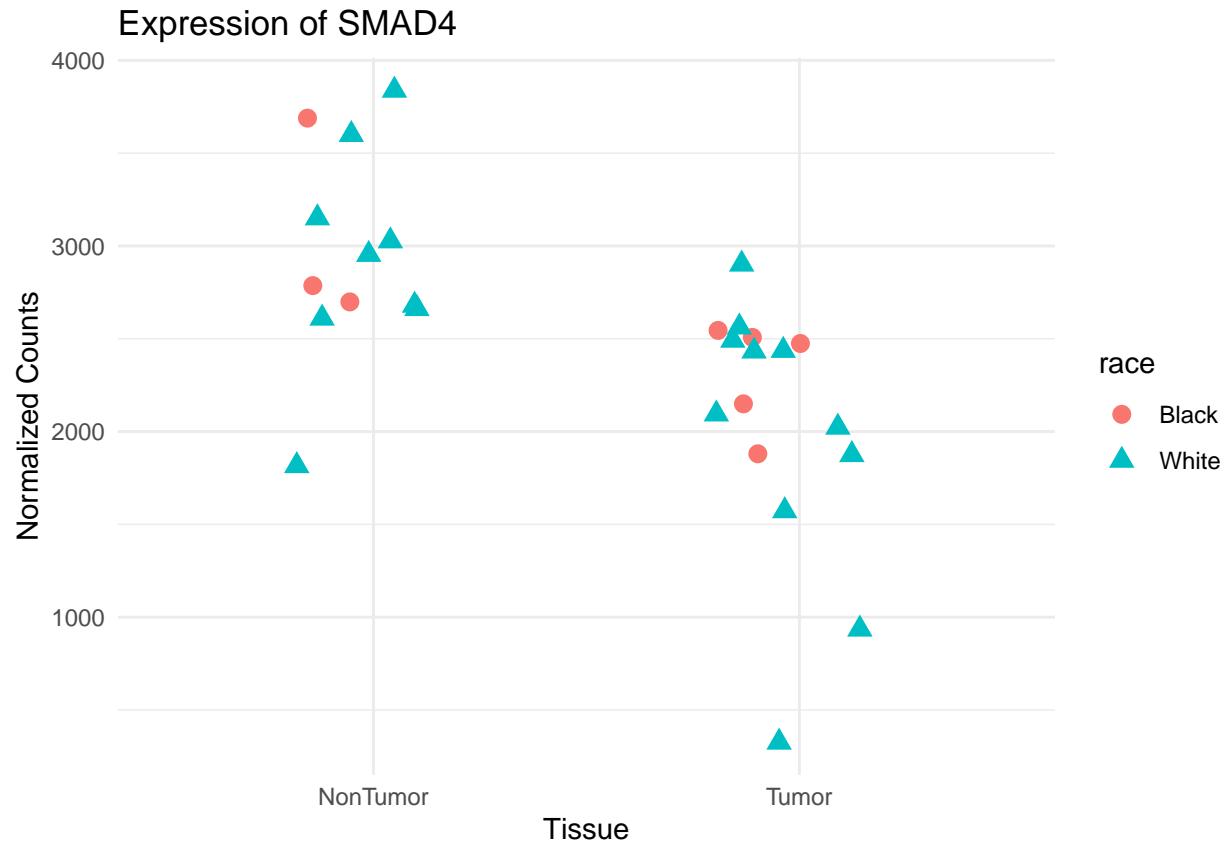


```
##  
## --- TP53 ---  
## Tumor vs Normal: FC = 1.14 | padj = 7.582079e-01 | UP in Tumor  
## Race effect: FC = 0.69 | padj = 5.938159e-01 | UP in Black  
## Interaction: FC = 0.47 | padj = 9.334379e-01 | Tumor effect Stronger in White
```

Expression of CDKN2A



```
##  
## --- CDKN2A ---  
## Tumor vs Normal: FC = 104.07 | padj = 1.095659e-04 | UP in Tumor  
## Race effect: FC = 98.37 | padj = 2.89949e-04 | UP in White  
## Interaction: FC = -6.03 | padj = 1.554318e-02 | Tumor effect Stronger in Black
```



```
##
## --- SMAD4 ---
## Tumor vs Normal: FC = 0.76 | padj = 5.45444e-01 | DOWN in Tumor
## Race effect: FC = 0.96 | padj = 9.912771e-01 | UP in Black
## Interaction: FC = -0.17 | padj = 9.969893e-01 | Tumor effect Stronger in Black
```

```
# 10. SUMMARY STATISTICS
print(paste("Tumor vs Non-Tumor DEGs:", nrow(sig_tumor)))
```

```
## [1] "Tumor vs Non-Tumor DEGs: 2582"
```

```
print(paste("Race-associated DEGs:", nrow(sig_race)))
```

```
## [1] "Race-associated DEGs: 334"
```

```
print(paste("Genes with different tumor effects by race:", nrow(sig_interaction)))
```

```
## [1] "Genes with different tumor effects by race: 123"
```

```
# Save results with gene symbols
write.csv(as.data.frame(sig_tumor), "tumor_vs_normal_deg.csv")
write.csv(as.data.frame(sig_race), "race_deg.csv")
write.csv(as.data.frame(sig_interaction), "interaction_genes.csv")
```