Normal tissues CTA expression

Léa ROGUE

2025-03-24

List of Figures

Contents

This script generates plots to see the expression of CTAs in normal tissues from GTEx project. This allow us to visualize the most tumor specific CTAs. We added EGFR to compare with an housekeeping genes.

Load libraries

```r
library(ComplexHeatmap)
library(dplyr)
library(ggplot2)
library(colorRamp2)
library(plotly)
```

Functions

```r
# Compute scatter plot df
compute_scatter_data <- function(df_int_cta) {
    # Compute mean
    df_mean_int_all <- as.data.frame(rowMeans(df_int_cta))
    colnames(df_mean_int_all) <- "Intensity"
    df_mean_int_all$SYMBOL <- rownames(df_mean_int_all)

    # Merge data
    df_scatter <- merge(df_mean_int_all, df_mean_tissue, by = "SYMBOL")
    df_scatter <- merge(df_scatter, df_clust_cta_conv, by = "SYMBOL",
        all.x = TRUE)
    df_scatter <- merge(df_scatter, df_clust_cta_all, by = "SYMBOL",
        all.x = TRUE)

    # Clean and rename
    rownames(df_scatter) <- df_scatter$SYMBOL
    df_scatter <- df_scatter[, -which(colnames(df_scatter) ==
        "SYMBOL")]
    colnames(df_scatter) <- c("Intensity", "Mean_expression_tissues",
        "Cluster_conv", "Cluster_all")

    return(df_scatter)
}

# Plot function
create_plot <- function(df_scatter, cluster_col, nudge_x, nudge_y) {
    ggplot(df_scatter, aes(x = Mean_expression_tissues, y = Intensity,
        color = {
            {
                cluster_col
            }
        })) + geom_point(size = 1) + scale_color_manual(values = c(`1` = "#3498DB",
        `2` = "#E74C3C", `3` = "#2ECC71"), na.value = "#dadada") +
        geom_label(data = df_scatter[rownames(df_scatter) %in%
            housekp_genes, ], aes(label = rownames(df_scatter)[rownames(df_scatter) %in%
```

```
              housekp_genes]), nudge_x = nudge_x, nudge_y = nudge_y,
              size = 3, color = "black") + theme_minimal()
}
```

Load data

We use expression matrix generated by script 7, the list of CTA that impact survival probabilities in conventional chondrosarcomas and the intensities data from (E-MTAB-7462).

```
# Load mean expression of CTA in each tissue
df_expr <- read.table("../results/matrix_expr_cta_tissues_gtex.tsv",
    sep = "\t", header = TRUE, row.names = 1)

# Read list of CTA that impact survival probabilities
l_CTA_conv <- read.table("../data/CTA_signif_coxph_conv_indiv.txt",
    sep = "\t", header = FALSE)$V1

# Load intensities data
df_int <- read.table("../results/whole_gene_int_CTA_sign_imm_clean.tsv",
    sep = "\t", row.names = 1, header = TRUE)

# List of housekeeping genes
housekp_genes <- c("EGFR", "ERBB2", "FOLH1", "SSTR1", "SSTR2",
    "SSTR3", "SSTR4", "SSTR5")

# Metadata Read the metadata and select individuals that
# have survival data
df_metadata <- read.table("../results/metadata.tsv", sep = "\t",
    header = TRUE, check.names = FALSE, dec = ",")
df_metadata_surv <- df_metadata[, c("Patient", "OS.delay", "OS.event")]
df_metadata_surv <- na.omit(df_metadata_surv)

# Conventional chondrosarcoma
df_metadata_conv <- df_metadata[df_metadata$Histology != "N/A" &
    df_metadata$Histology != "benign" & df_metadata$Histology !=
    "dedifferentiated", ]
df_metadata_surv_conv <- df_metadata_conv[, c("Patient", "OS.delay",
    "OS.event")]
df_metadata_surv_conv <- na.omit(df_metadata_surv_conv)
```

I. Expression of all the CTAs in normal tissues

This part show that these genes are CTAs are exressed in the testis.

```
# Log10 to scale the data
df_log10 <- log10(t(df_expr) + 1)
df_log10_all_cta <- df_log10[, !colnames(df_log10) %in% housekp_genes]
df_log10_all_cta <- t(df_log10_all_cta)

# CTA annotation
```

```r
l_cta_validated <- read.table("../data/CTA_validated_list.txt",
    sep = "\t", header = FALSE)$V1

df_cta_validated <- df_log10_all_cta[, colnames(df_log10_all_cta) %in%
    l_cta_validated]
df_cta_putatives <- df_log10_all_cta[, !colnames(df_log10_all_cta) %in%
    l_cta_validated]

# Create heatmaps
colors = colorRamp2(c(min(df_log10_all_cta), max(df_log10_all_cta)),
    c("black", "red"))
ht_validated <- Heatmap(as.matrix(df_cta_validated), cluster_rows = FALSE,
    cluster_columns = FALSE, col = colors, border = NA, show_column_names = TRUE,
    show_row_names = TRUE, column_title = "Validated CTAs", column_names_gp = gpar(fontsize = 1),
    row_names_gp = gpar(fontsize = 10), heatmap_legend_param = list(title = "Expression"),
    rect_gp = gpar(col = NA))

ht_putatives <- Heatmap(as.matrix(df_cta_putatives), cluster_rows = FALSE,
    cluster_columns = FALSE, col = colors, show_column_names = TRUE,
    show_row_names = TRUE, column_title = "Putatives CTAs", column_names_gp = gpar(fontsize = 1),
    row_names_gp = gpar(fontsize = 10), rect_gp = gpar(col = NA))
# png('../results/figures/heatmaps/heatmap_cta_normal_tissues.png',
# width = 3000, height = 1800, res = 300)
draw(ht_validated + ht_putatives, merge_legend = TRUE)


# dev.off()
```
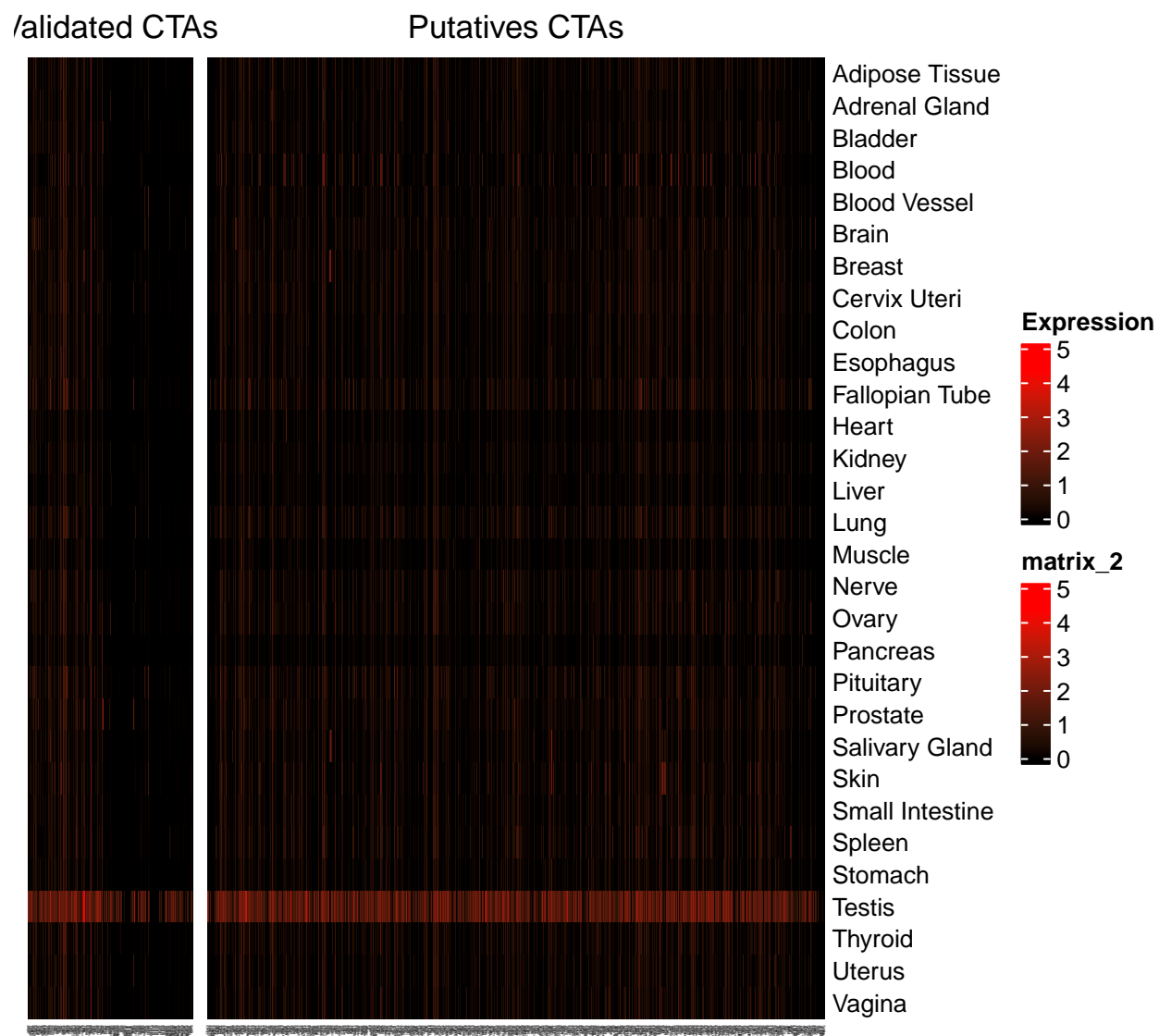
Figure 1: Heatmap of CTAs expression in normal tissues

So, this heatmap shows that genes are generally expressed in the testis and less expressed in other tissues.

II. Expression of CTAs that impact survival probabilities in conventional chondrosarcoma in normal tissues

This section shows the expression of CTAs that impact survival probabilities in conventional chondrosarcoma in normal tissues. This permit to see if these specific CTAs are expressed in normal tissues.

```r
# Select CTA
df_log10_conv_CTA <- df_log10[rownames(df_log10) %in% l_CTA_conv,
    ]

# Df to have genes and their correspondent color
df_clust_cta_conv <- read.table("../results/clusters_indiv/clusters_cta_signif_conv_indiv.tsv",
    sep = "\t", header = TRUE)
rownames(df_clust_cta_conv) <- df_clust_cta_conv$SYMBOL
df_clust_cta_conv <- df_clust_cta_conv[rownames(df_clust_cta_conv) %in%
    rownames(df_log10_conv_CTA), ]
df_clust_cta_conv$Cluster <- as.character(df_clust_cta_conv$Cluster)

# Colors of clusters
clust_colors <- c(`1` = "#3498DB", `2` = "#E74C3C", `3` = "#2ECC71")

# Heatmap
# pdf('../results/figures/heatmaps/heatmap_expr_cta_conv_gtex.pdf')
Heatmap(as.matrix(df_log10_conv_CTA), cluster_rows = TRUE, cluster_columns = TRUE,
    cluster_column_slices = TRUE, clustering_distance_columns = "euclidean",
    clustering_method_columns = "complete", show_column_dend = TRUE,
    col = colors, column_gap = unit(0, "mm"), row_gap = unit(0,
        "mm"), show_column_names = TRUE, show_row_names = TRUE,
    column_names_gp = gpar(fontsize = 7), row_names_gp = gpar(fontsize = 3),
    right_annotation = rowAnnotation(Cluster = df_clust_cta_conv$Cluster,
        col = list(Cluster = clust_colors)), heatmap_legend_param = list(title = "Expression Level"),
    rect_gp = gpar(col = "transparent", lwd = 0), )
```

```r
# dev.off()
```

We see that there is some groups that are not really expressed in normal tissues so they could be good biomarkers in chondrosarcomas.
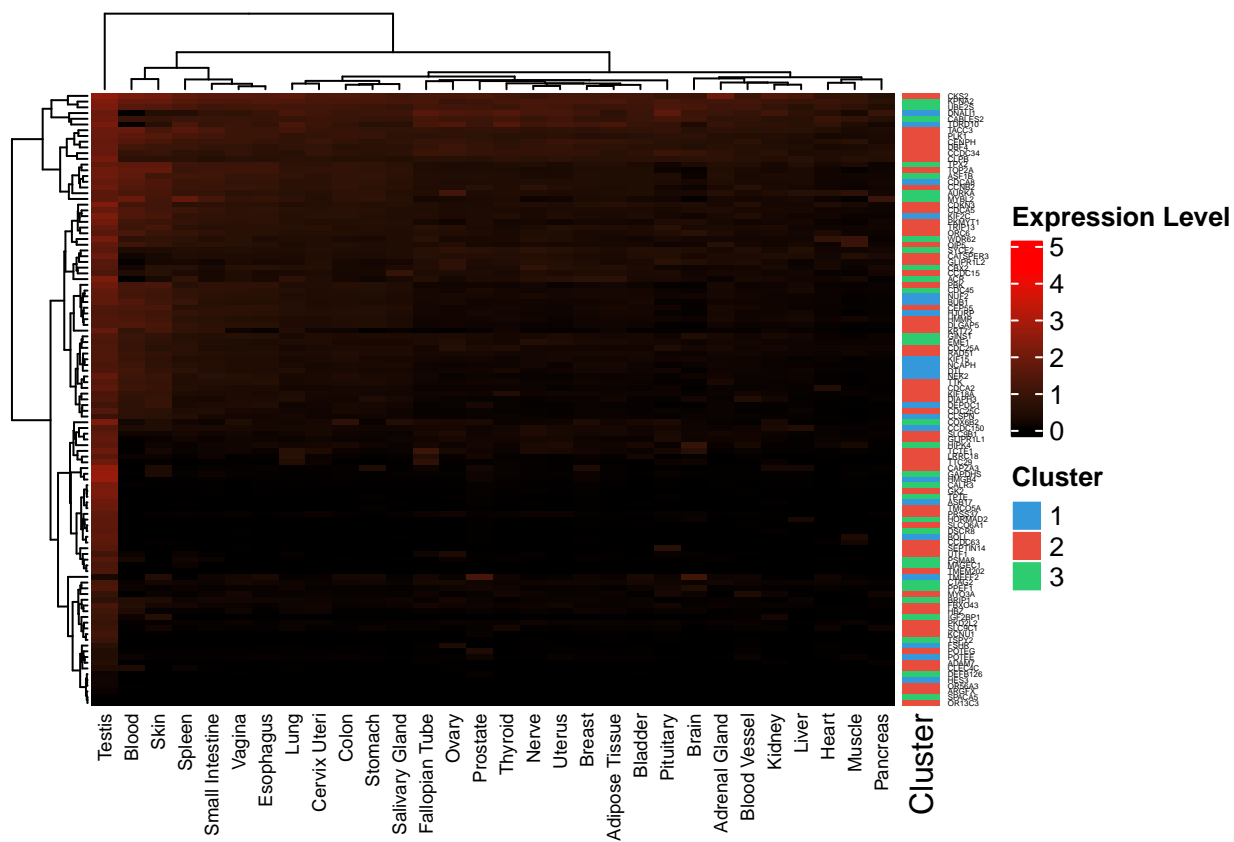
Figure 2: Heatmap of CTAs expression that impact survival in normal tissues

III. Scatter plot of CTAs intensities and their expression in normal tissues.

In this part, we want to see the less expressed CTAs in normal tissues and the more expressed CTAs in chondrosarcomas.

1) n = 82

a- Conventional survival CTAs list

```r
# add cluster info
df_clust_cta_all <- read.table("../results/clusters_indiv/clusters_cta_signif_coxph_all_indiv.tsv",
    sep = "\t", header = TRUE)
rownames(df_clust_cta_all) <- df_clust_cta_all$SYMBOL
df_clust_cta_all <- df_clust_cta_all[rownames(df_clust_cta_all) %in%
    rownames(df_log10), ]
df_clust_cta_all$Cluster <- as.character(df_clust_cta_all$Cluster)
colnames(df_clust_cta_all) <- c("Cluster_all", "SYMBOL")

# Select CTA intensities
df_int_cta <- df_int %>%
    filter(CTA != "NA") %>%
    select(-Signature, -CTA)
df_housekp_genes <- df_int[rownames(df_int) %in% housekp_genes,
    ]
df_int_cta <- rbind(df_int_cta, df_housekp_genes[, -c(1, 2)])

# Compute mean for each CTA in all tissues excluding Testis
df <- as.data.frame(t(df_expr[!rownames(df_expr) == "Testis",
    ]))
df_mean_tissue <- as.data.frame(rowMeans(df))
df_mean_tissue$SYMBOL <- rownames(df_mean_tissue)

# Compute mean for each CTA for all patients
df_scatter_82 <- compute_scatter_data(df_int_cta[, colnames(df_int_cta) %in%
    df_metadata_surv$Patient])

# Generate plot
p <- create_plot(df_scatter_82, Cluster_conv, 0.8, -0.3)
p + scale_x_log10() + labs(title = "Scatter plot of tumor specificity of CTAs (n = 82)",
    x = "Normal tissues expression (Log10 TPM mean)", y = "Chondrosarcoma CTA expression (intensities)"
    color = "Clustering gene\nimpacting survival\n(conventional CHS)")
```

Log10 scale

```r
# Generate plot
p <- create_plot(df_scatter_82, Cluster_conv, 4, -0.5)
```
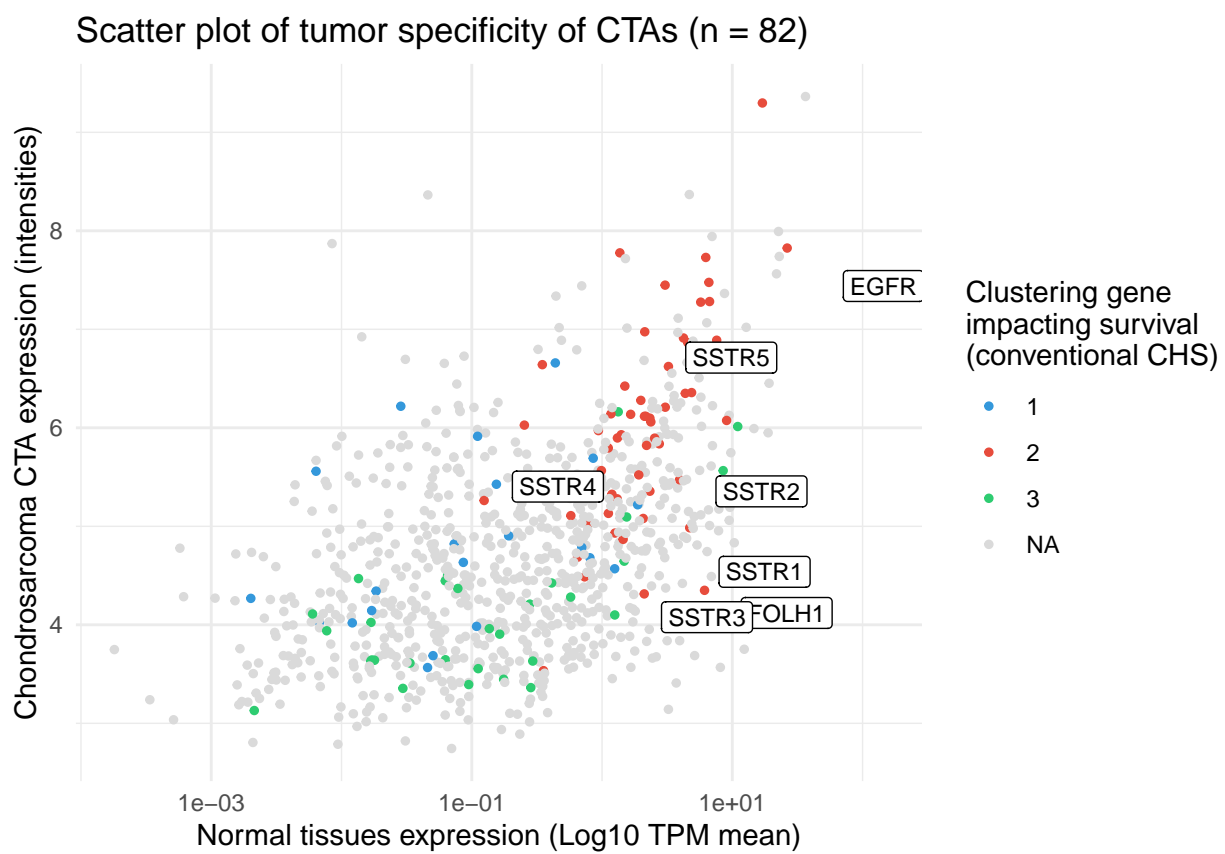
Figure 3: Scatter plot of CTAs intensities and their expression in normal tissues with clustering colors from n = 63 (log10 scale) with n = 82

```
p + labs(title = "Scatter plot of tumor specificity of CTAs (n = 82)",
    x = "Normal tissues expression (TPM mean)", y = "Chondrosarcoma CTA expression (intensities)",
    color = "Clustering gene\nimpacting survival\n(conventional CHS)")
```
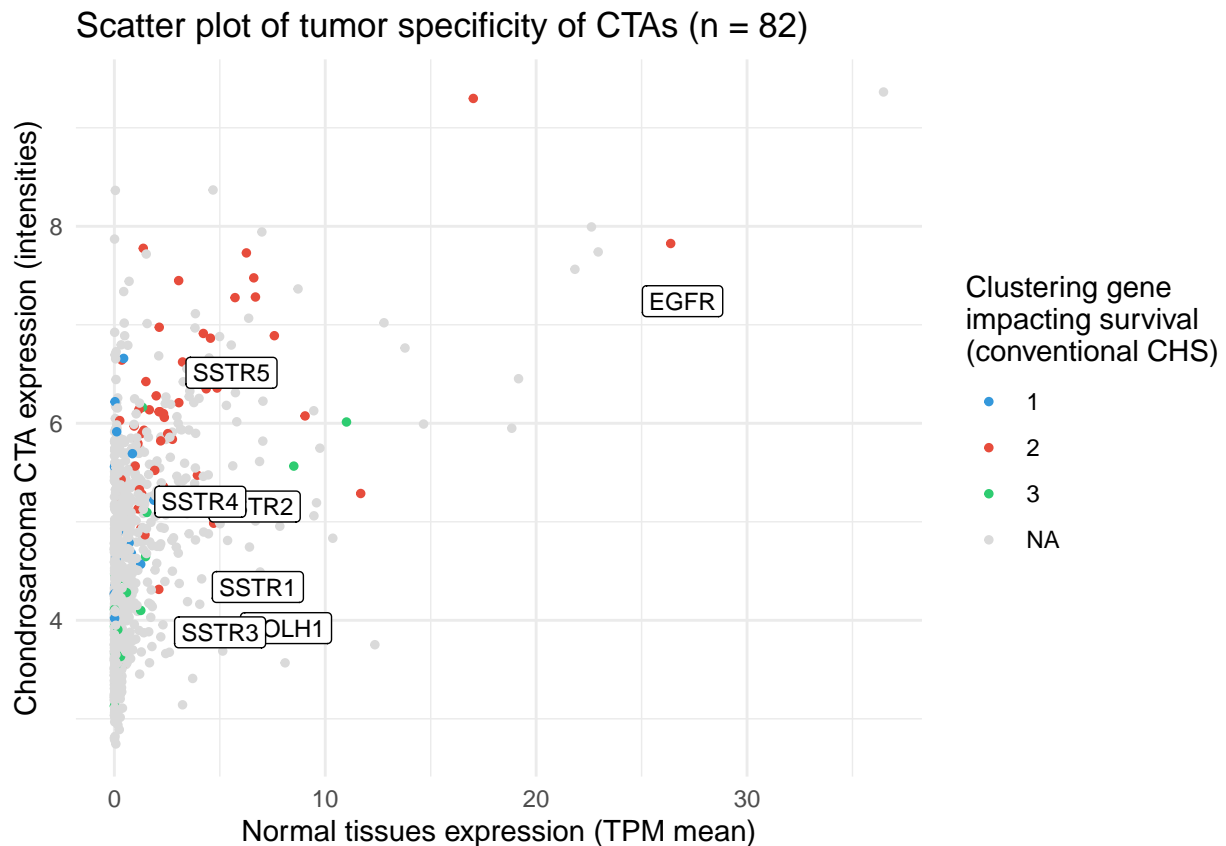

Figure 4: Scatter plot of CTAs intensities and their expression in normal tissues with clustering colors from n = 63 (linear scale) with n = 82

Linear scale

b- N = 82 survival CTAs list

```
# Generate plot
p <- create_plot(df_scatter_82, Cluster_all, 0.8, -0.3)
p + scale_x_log10() + labs(title = "Scatter plot of tumor specificity of CTAs (n = 82)",
    x = "Normal tissues expression (Log10 TPM mean)", y = "Chondrosarcoma CTA expression (intensities)"
    color = "Clustering gene\nimpacting survival\n(conventional CHS)")
```
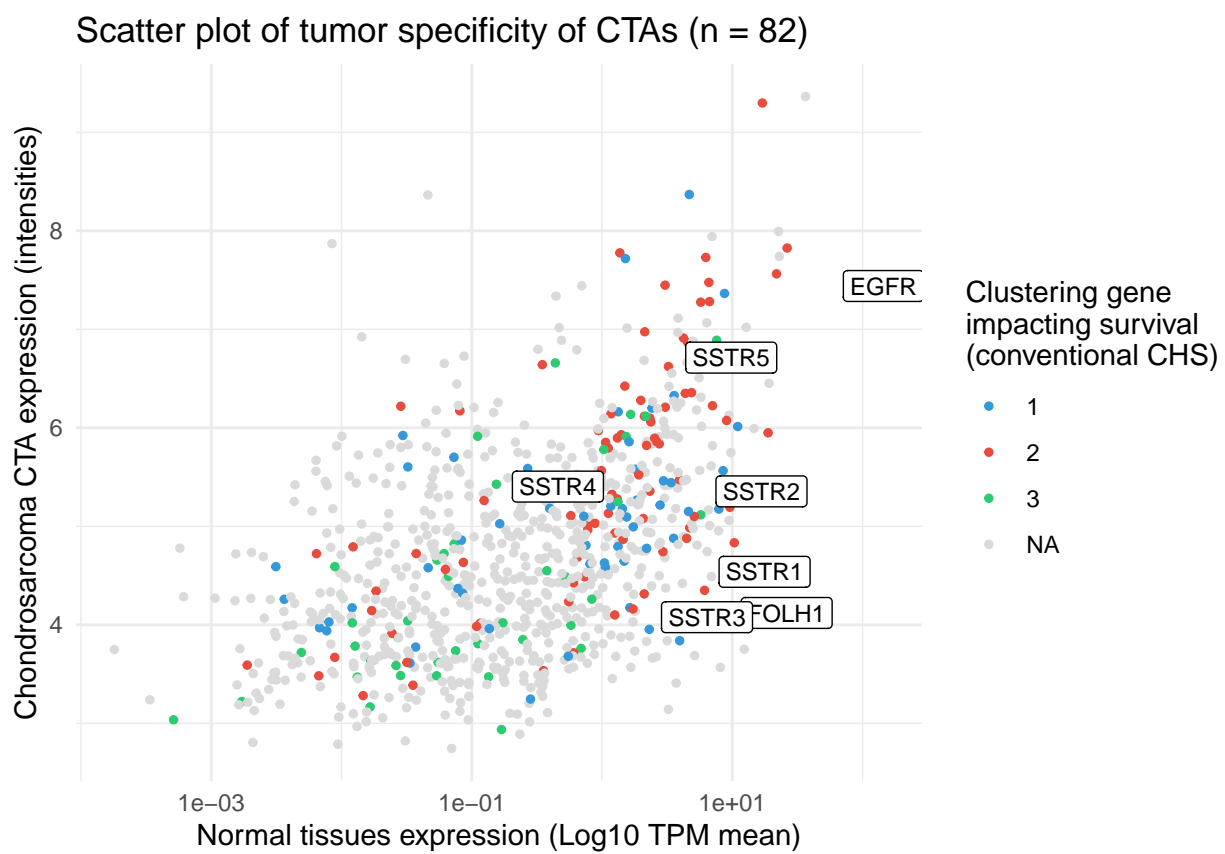
Log10 scale

Figure 5: Scatter plot of CTAs intensities and their expression in normal tissues with clustering colors from n = 82 (log10 scale) with n = 82

```
# Generate plot
p <- create_plot(df_scatter_82, Cluster_all, 4, -0.5)
p + labs(title = "Scatter plot of tumor specificity of CTAs (n = 82)",
    x = "Normal tissues expression (TPM mean)", y = "Chondrosarcoma CTA expression (intensities)",
    color = "Clustering gene\nimpacting survival\n(conventional CHS)")
```
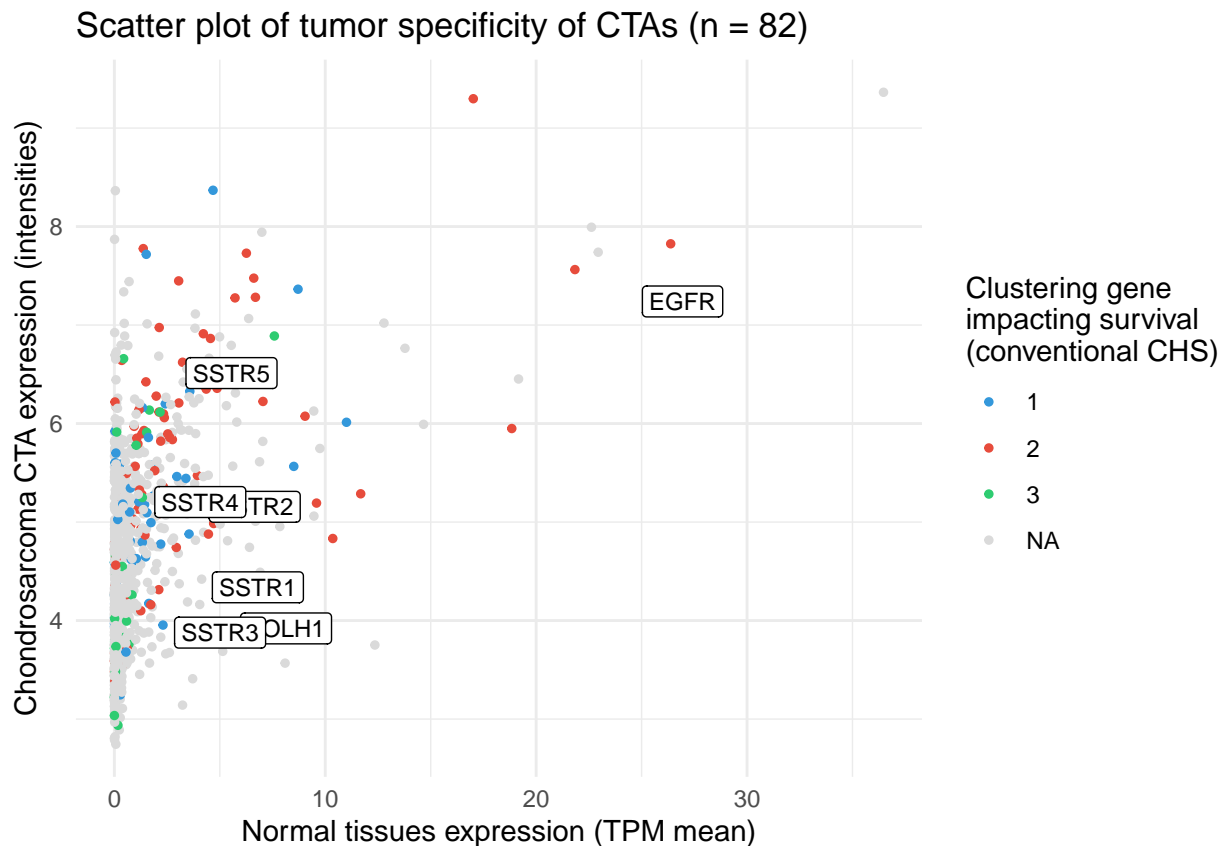

Figure 6: Scatter plot of CTAs intensities and their expression in normal tissues with clustering colors from n = 82 (linear scale) with n = 82

Linear scale

2) n = 63

a- Conventional survival CTAs list

```
# Compute mean for each CTA for all patients
df_scatter_63 <- compute_scatter_data(df_int_cta[, colnames(df_int_cta) %in%
    df_metadata_surv_conv$Patient])

# Generate plot
p <- create_plot(df_scatter_63, Cluster_conv, 0.8, -0.3)
```

```
p + scale_x_log10() + labs(title = "Scatter plot of tumor specificity of CTAs (n = 63)",
    x = "Normal tissues expression (Log10 TPM mean)", y = "Chondrosarcoma CTA expression (intensities)"
    color = "Clustering gene\nimpacting survival\n(conventional CHS)")
```
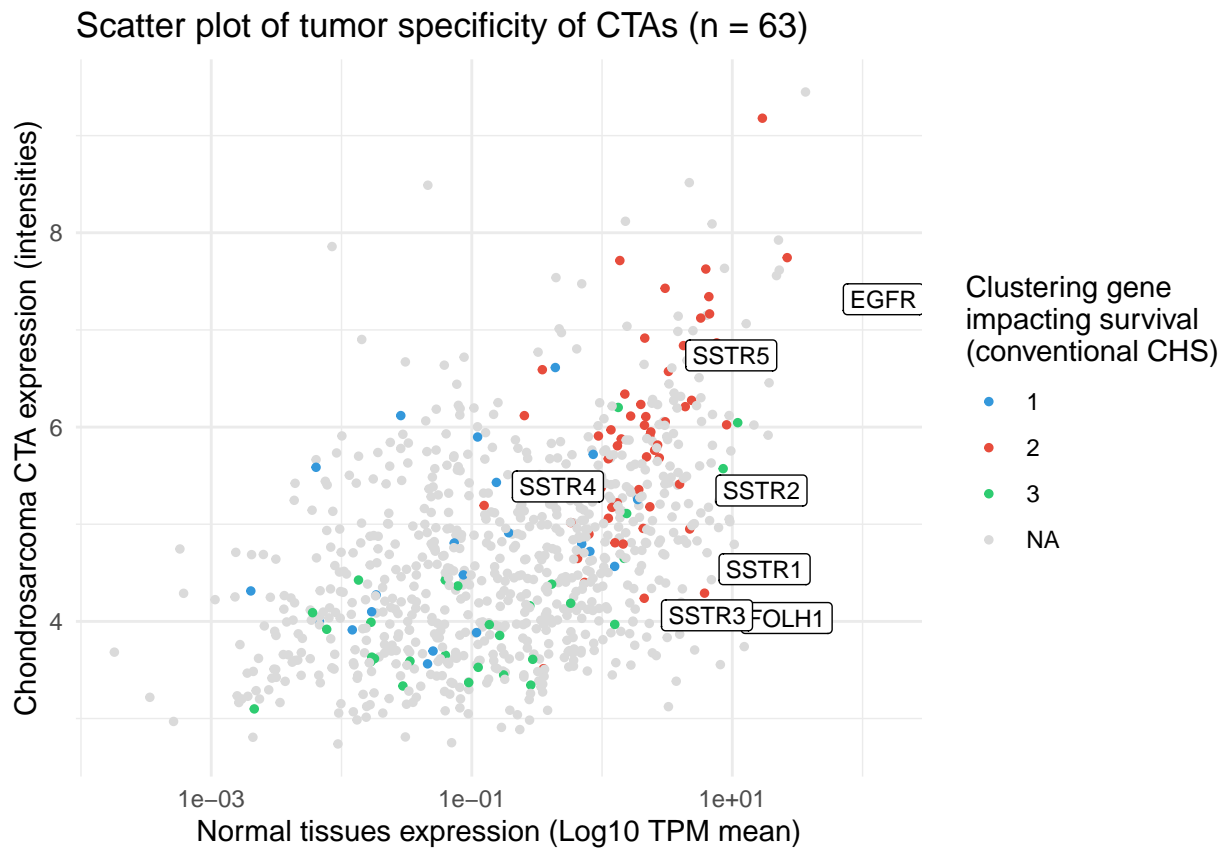

Figure 7: Scatter plot of CTAs intensities and their expression in normal tissues with clustering colors from n = 63 (log10 scale) with n = 63

Log10 scale

```
# Generate plot
p <- create_plot(df_scatter_63, Cluster_conv, 4, -0.5)
p + labs(title = "Scatter plot of tumor specificity of CTAs (n = 63)",
    x = "Normal tissues expression (TPM mean)", y = "Chondrosarcoma CTA expression (intensities)",
    color = "Clustering gene\nimpacting survival\n(conventional CHS)")
```

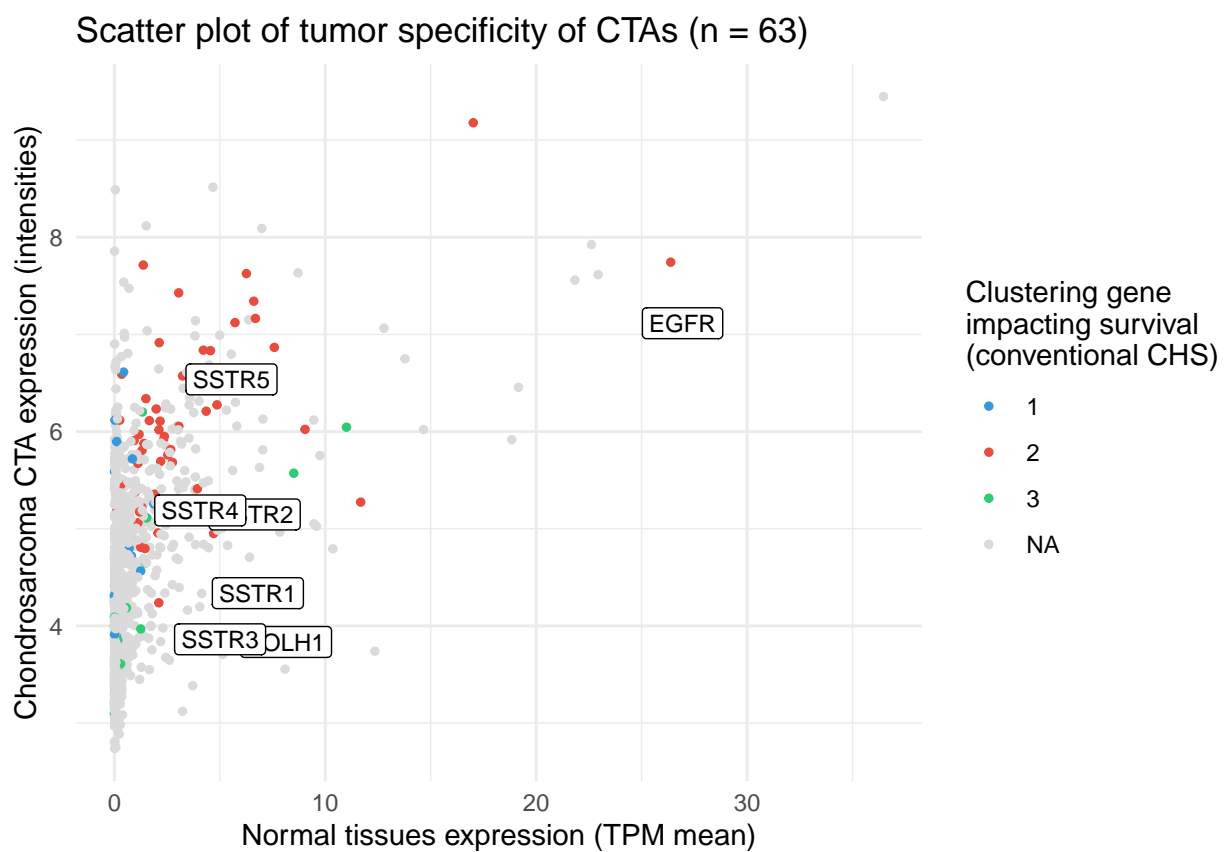**Linear scale**

b- N = 82 survival CTAs list

Figure 8: Scatter plot of CTAs intensities and their expression in normal tissues with clustering colors from n = 63 (linear scale) with n = 63

```
# Generate plot
p <- create_plot(df_scatter_63, Cluster_all, 0.8, -0.3)
p + scale_x_log10() + labs(title = "Scatter plot of tumor specificity of CTAs (n = 63)",
    x = "Normal tissues expression (Log10 TPM mean)", y = "Chondrosarcoma CTA expression (intensities)"
    color = "Clustering gene\nimpacting survival\n(conventional CHS)")
```
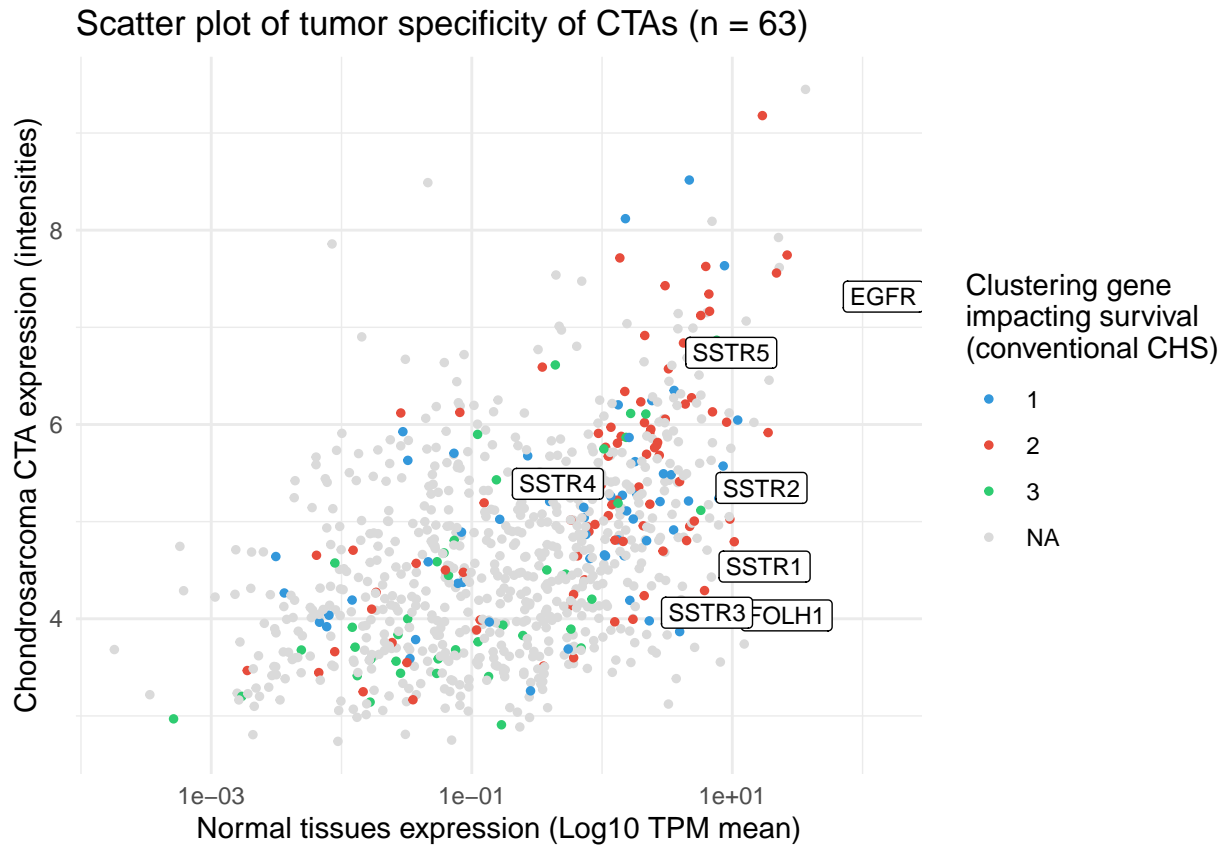

Figure 9: Scatter plot of CTAs intensities and their expression in normal tissues with clustering colors from n = 82 (log10 scale) with n = 63

Log10 scale

```
# Generate plot
p <- create_plot(df_scatter_63, Cluster_all, 4, -0.5)
p + labs(title = "Scatter plot of tumor specificity of CTAs (n = 63)",
    x = "Normal tissues expression (TPM mean)", y = "Chondrosarcoma CTA expression (intensities)",
    color = "Clustering gene\nimpacting survival\n(conventional CHS)")
```
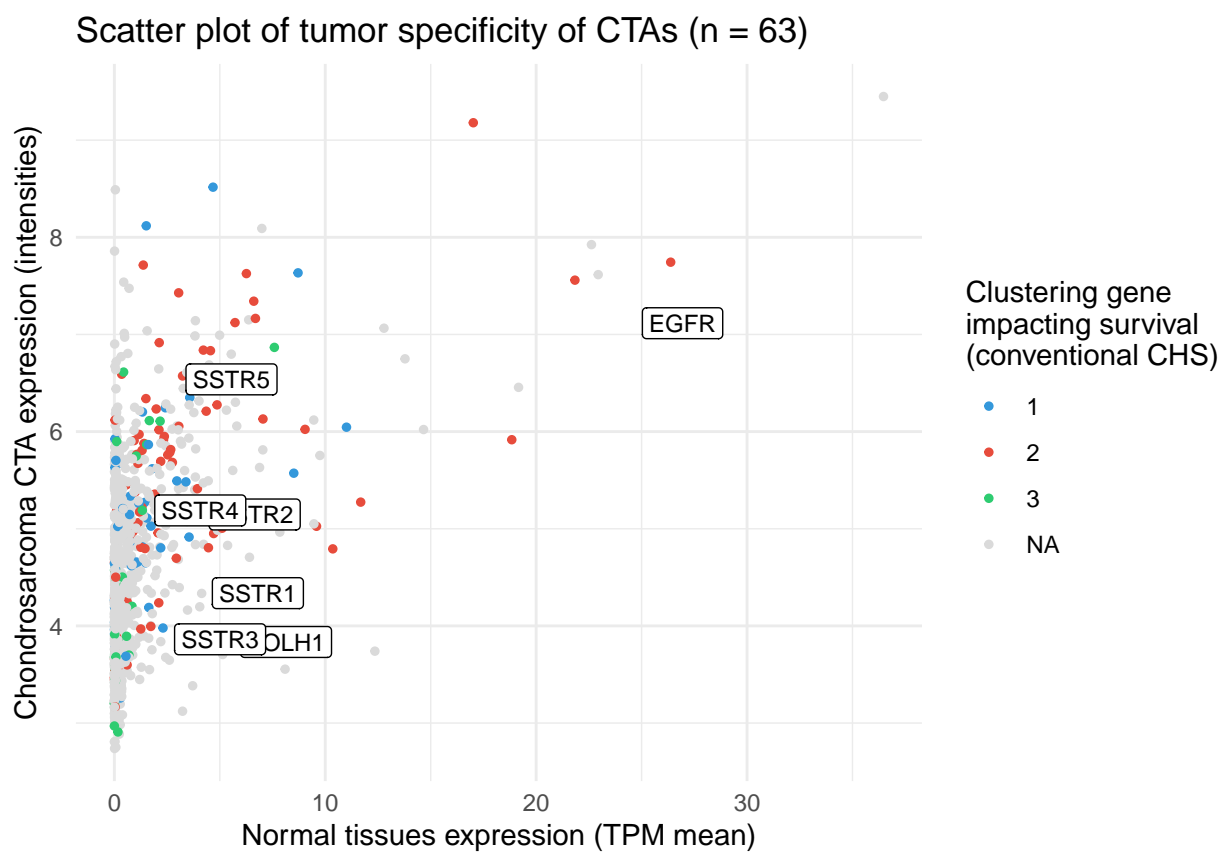
Linear scale

Figure 10: Scatter plot of CTAs intensities and their expression in normal tissues with clustering colors from n = 82 (linear scale) with n = 63

IV. Select interested CTAs thanks thresholds

In the literature, PSMA is a known marker and used to treat prostate cancer, so the expresion of PSMA gene (FOLH1) is used as a threshold with the median expression in chondrosarcoma

```r
# Select
selected_cta <- df_scatter_63[df_scatter_63$Intensity > median(df_scatter_63$Intensity) &
    df_scatter_63$Mean_expression_tissues < df_scatter_63["FOLH1",
        "Mean_expression_tissues"], ]

selected_cta <- selected_cta[!rownames(selected_cta) %in% housekp_genes,
    ]

# Save write.table(rownames(selected_cta), file =
# '../results/selected_cta_scatter.txt', quote = FALSE,
# row.names = FALSE, col.names = FALSE)
```

V. Generate plot for report

```r
# Replace
df_scatter_63$Cluster_conv <- ifelse(df_scatter_63$Cluster_conv ==
    "1", "a", ifelse(df_scatter_63$Cluster_conv == "2", "b",
    ifelse(df_scatter_63$Cluster_conv == "3", "c", df_scatter_63$Cluster_conv)))

df_scatter_63 <- df_scatter_63[!rownames(df_scatter_63) %in%
    housekp_genes, ]

save(df_scatter_63, file = "../results/df_scatter_63.RData")

# Generate plot
ggplot(df_scatter_63, aes(x = Mean_expression_tissues, y = Intensity,
    color = Cluster_conv)) + geom_point(size = 1) + geom_vline(xintercept = 4.139,
    linetype = "dashed", color = "black") + geom_hline(yintercept = median(df_scatter_63$Intensity),
    linetype = "dashed", color = "black") + scale_color_manual(values = c(a = "#dadada",
    b = "#dadada", c = "#dadada"), na.value = "#dadada") + theme_minimal() +
    scale_x_log10() + labs(title = "Scatter plot of tumor specificity of CTAs (n = 63)",
    x = "Normal tissues expression (Log10 TPM mean)", y = "Chondrosarcoma CTA expression (intensities)"
    color = "Clustering gene\nimpacting survival\n(conventional CHS)")
```

```r
# etiquette ?  pearson et HR
```

```r
# Adding data on pearson correlation
df_hot_cold <- read.table("../results/clusters_indiv/clusters_cta_pearson_63.tsv",
    header = T, sep = "\t")
colnames(df_hot_cold) <- c("Associated_immunophenotype", "SYMBOL")
df_scatter_63$SYMBOL <- rownames(df_scatter_63)
df_scatter_anno <- merge(df_scatter_63, df_hot_cold, by = "SYMBOL")

# Adding HR data
df_hr <- read.table("../results/results_coxph_cta_conv_indiv.tsv",
```
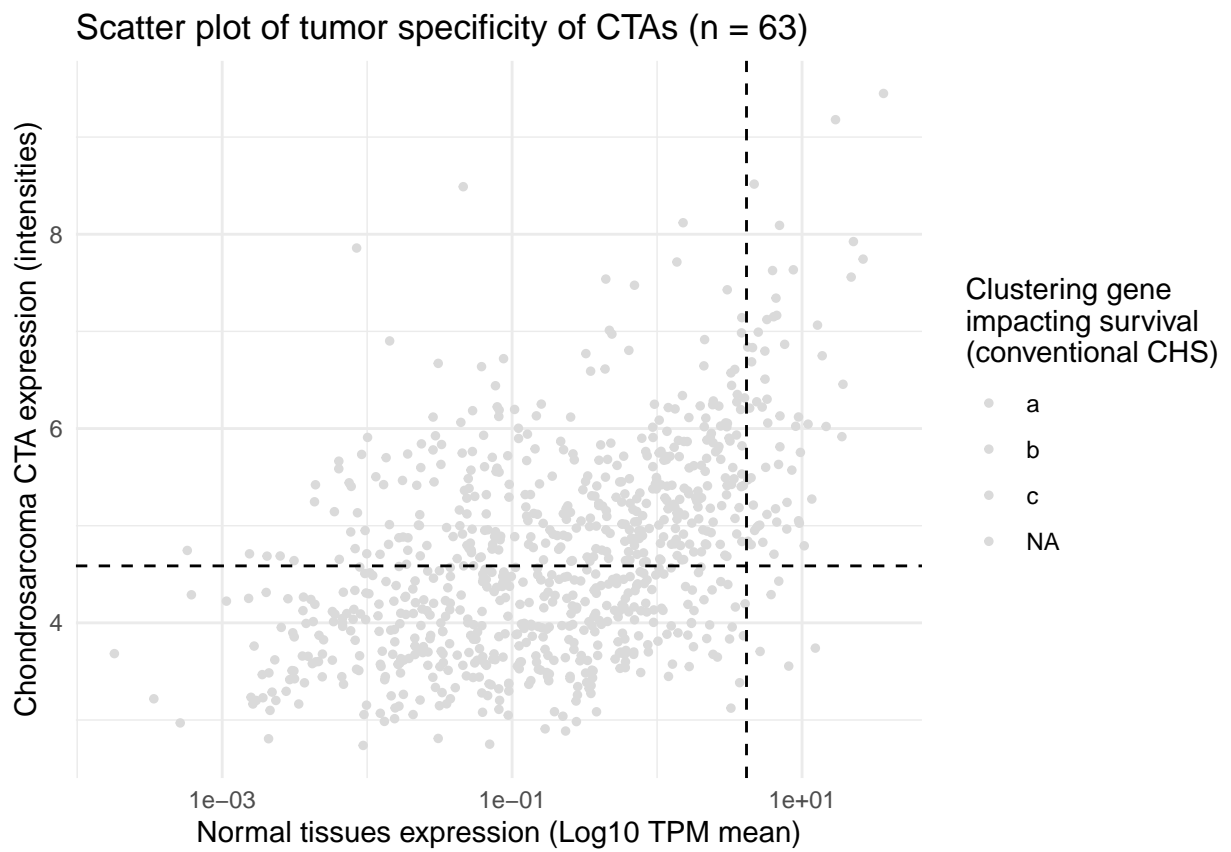
Figure 11: Scatter plot of CTAs intensities and their expression in normal tissues (n = 63)

```r
    header = T, sep = "\t")
df_hr <- df_hr %>%
    select(Variable, HR, Pvalue) %>%
    filter(Pvalue < 0.05) %>%
    rename(SYMBOL = Variable)
df_scatter_anno <- merge(df_scatter_anno, df_hr, by = "SYMBOL",
    all.x = T)

# Add HR value to add point size
df_scatter_anno$HR <- ifelse(is.na(df_scatter_anno$HR), 0.1,
    df_scatter_anno$HR)


# Generate plot
ggplot(df_scatter_anno, aes(x = Mean_expression_tissues, y = Intensity,
    color = Cluster_conv, size = HR, shape = Associated_immunophenotype)) +
    geom_point(alpha = 0.8) + geom_vline(xintercept = 4.139,
    linetype = "dashed", color = "black") + geom_hline(yintercept = median(df_scatter_63$Intensity),
    linetype = "dashed", color = "black") + scale_shape_manual(values = c(COLD = 15,
    HOT = 16)) + scale_size_continuous(range = c(1, 3)) + scale_color_manual(values = c(a = "#3498DB",
    b = "#E74C3C", c = "#2ECC71"), na.value = "#dadada") + theme_minimal() +
    scale_x_log10() + labs(title = "Scatter plot of tumor specificity of CTAs (n = 63)",
    x = "Normal tissues expression (Log10 TPM mean)", y = "Chondrosarcoma CTA expression (intensities)"
    color = "Clustering gene\nimpacting survival\n(conventional CHS)",
    shape = "Associated Immunophenotype", size = "Signficant HR")


p <- ggplot(df_scatter_anno, aes(x = Mean_expression_tissues,
    y = Intensity, color = Cluster_conv, size = HR, shape = Associated_immunophenotype)) +
    geom_point(alpha = 0.8) + geom_vline(xintercept = 4.139,
    linetype = "dashed", color = "black") + geom_hline(yintercept = median(df_scatter_63$Intensity),
    linetype = "dashed", color = "black") + scale_shape_manual(values = c(COLD = 15,
    HOT = 16)) + scale_size_continuous(range = c(1, 3)) + scale_color_manual(values = c(a = "#3498DB",
    b = "#E74C3C", c = "#2ECC71"), na.value = "#dadada") + theme_minimal() +
    scale_x_log10() + labs(title = "Scatter plot of tumor specificity of CTAs (n = 63)",
    x = "Normal tissues expression (Log10 TPM mean)", y = "Chondrosarcoma CTA expression (intensities)"
    color = "Clustering gene\nimpacting survival\n(conventional CHS)",
    shape = "Associated Immunophenotype", size = "Signficant HR")
```
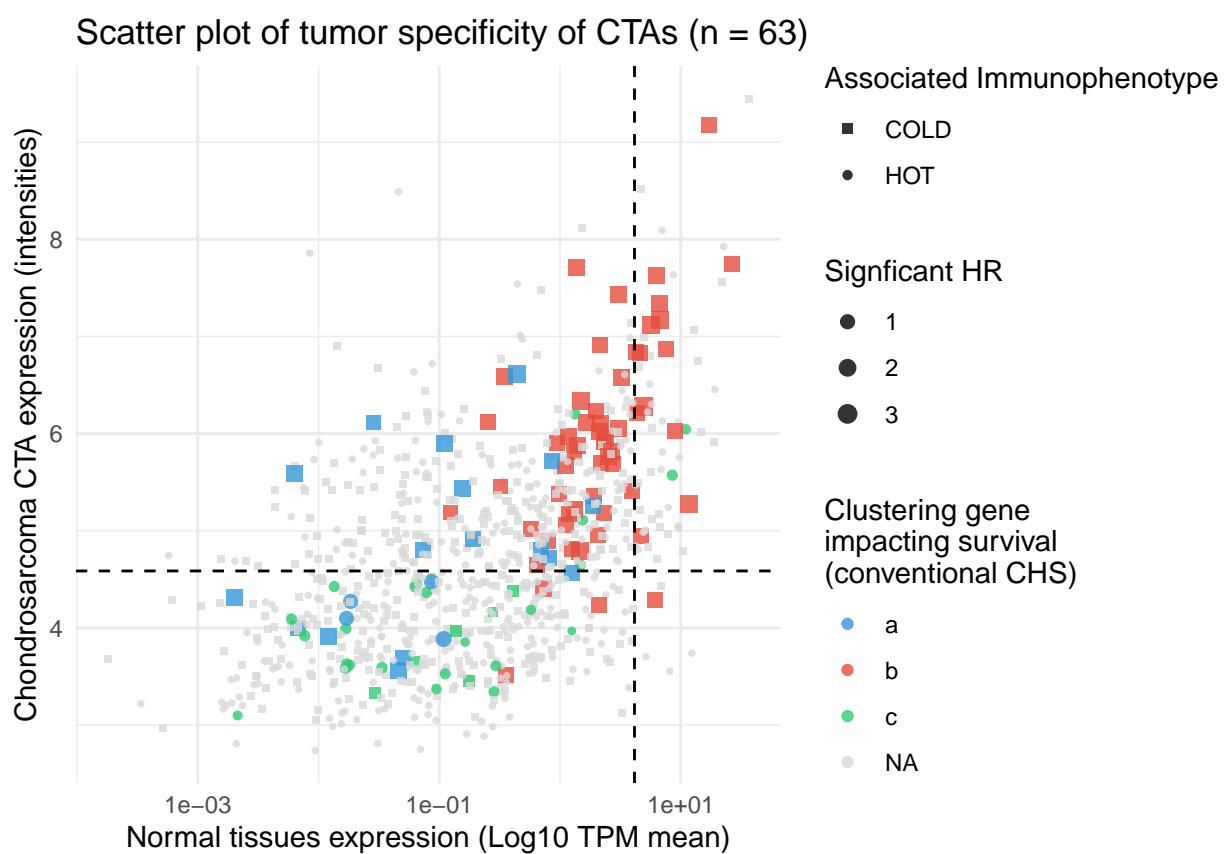
Figure 12: Scatter plot of CTAs intensities and their expression in normal tissues with clustering colors, size by HR (significant), and shape by associated immunophenotype (n = 63)