Aim 2 Analysis: Characterizing The Effect of CD7 in Restricting Multipotent Progenitor Potential

In this notebook, we will be completing our analysis of Days 7 - Day 13 of the CD34+ Cells in LEM.

The Research Questions of this Aim are:

1. Is There a Difference in CD7 expression between Days 7, 10 and 13, across all HSCs or within each Population?
2. Is There a Statistical Difference in CD7 Expression Between Ra-C- cells (Primitive), and Ra-C+/C+ Populations (Morer Mature)?
3. Does the Increase in CD7 Expression statisticallly correlate with an Increase in Known T-Cell Gene Expression?
4. What are the Top Transcriptional Drivers/Regulators in CD7+ Cells, Compared to CD7- Cells?
5. Do $CD7^{+}$ cells show pathway enrichment for general immune or alternative lineage programs (e.g., APC, myeloid, NK) instead of T-lineage-specific programs?

Pre-Processing and Data Preparation

Let's begin by loading any of our necessary libraries, and loading our CITE-Seq data into this notebook:

In [1]:
```
# Loading Libraries

library(BiocSingular) # We need this to use the BioConductor libraries that work on
library(SingleCellExperiment) # We need this to use the SingleCellExperiment data s
library(ggplot2) # we need this to make ggplot visualizations #nolint
library(tidyr) # we need this to manipulate data #nolint
library(dplyr) # we need this to manipulate data #nolint
library(patchwork) # to display plots side by side. #nolint
library(ggforce) # Allows me to display circles on ggplots. #nolint
library(limma) # helps with differential expression analysis #nolint
library(IRdisplay) # lets me display JPEGs in the notebook #nolint
library(org.Hs.eg.db) # lets me do gene annotation #nolint
library(clusterProfiler) # lets me do gene set enrichment analysis #nolint
library(broom) # lets me manipulate data #nolint
library(enrichplot) # lets me visualize gene set enrichment analysis #nolint
library(tidyverse) # lets me manipulate data #nolint
```

```
Loading required package: SummarizedExperiment

Loading required package: MatrixGenerics

Loading required package: matrixStats


Attaching package: 'MatrixGenerics'


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: GenomicRanges

Loading required package: stats4

Loading required package: BiocGenerics


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, intersect, is.unsorted, lapply, Map, mapply,
    match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
    Position, rank, rbind, Reduce, rownames, sapply, saveRDS, setdiff,
    table, tapply, union, unique, unsplit, which.max, which.min


Loading required package: S4Vectors


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: GenomeInfoDb

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
```

```
Attaching package: 'tidyr'

The following object is masked from 'package:S4Vectors':

    expand


Attaching package: 'dplyr'

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

    combine

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, union

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

    count

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':
```

```
    intersect, setdiff, setequal, union


Attaching package: 'limma'


The following object is masked from 'package:BiocGenerics':

    plotMA


Loading required package: AnnotationDbi


Attaching package: 'AnnotationDbi'


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

    select




clusterProfiler v4.14.4 Learn more at https://yulab-smu.top/contribution-knowledge-m
ining/

Please cite:

Guangchuang Yu, Li-Gen Wang, Yanyan Han and Qing-Yu He.
clusterProfiler: an R package for comparing biological themes among
gene clusters. OMICS: A Journal of Integrative Biology. 2012,
16(5):284-287


Attaching package: 'clusterProfiler'


The following object is masked from 'package:AnnotationDbi':

    select


The following object is masked from 'package:IRanges':

    slice


The following object is masked from 'package:S4Vectors':

    rename
```

```
The following object is masked from 'package:stats':

    filter


enrichplot v1.26.6 Learn more at https://yulab-smu.top/contribution-knowledge-minin
g/

Please cite:

Guangchuang Yu. Using meshes for MeSH term enrichment and semantic
analyses. Bioinformatics. 2018, 34(21):3766-3767,
doi:10.1093/bioinformatics/bty410

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ forcats   1.0.0     ✓ readr     2.1.5
✓ lubridate 1.9.4     ✓ stringr   1.5.1
✓ purrr     1.0.2     ✓ tibble    3.2.1
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✕ lubridate::%within%()     masks IRanges::%within%()
✕ dplyr::collapse()         masks IRanges::collapse()
✕ dplyr::combine()          masks Biobase::combine(), BiocGenerics::combine()
✕ dplyr::count()            masks matrixStats::count()
✕ dplyr::desc()             masks IRanges::desc()
✕ tidyr::expand()           masks S4Vectors::expand()
✕ clusterProfiler::filter() masks dplyr::filter(), stats::filter()
✕ dplyr::first()            masks S4Vectors::first()
✕ dplyr::lag()              masks stats::lag()
✕ ggplot2::Position()       masks BiocGenerics::Position(), base::Position()
✕ purrr::reduce()           masks GenomicRanges::reduce(), IRanges::reduce()
✕ clusterProfiler::rename() masks dplyr::rename(), S4Vectors::rename()
✕ lubridate::second()       masks S4Vectors::second()
✕ lubridate::second<-()     masks S4Vectors::second<-()
✕ clusterProfiler::select() masks AnnotationDbi::select(), dplyr::select()
✕ purrr::simplify()         masks clusterProfiler::simplify()
✕ clusterProfiler::slice()  masks dplyr::slice(), IRanges::slice()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

Once again, refer to `d0_analysis.ipynb` to look at the process by which I have created my sce object. For now, I am going to load it in, instead of repeating the process to save time. Then, I will be adding phenotype data and population markers to this sce object as columns (as down in `Aim1_Analysis.ipynb`).

In [2]:
```r
load("data/phenotype_with_ID.RData")
load("data/merge2.RData")

# Adding Phenotype data to the SCE object
pheno.d7 <- rep("CD34+CD45RA-CLEC12A-", 3039)
names(pheno.d7) <- colnames(merge2)[1:3039]

pheno.merge2 <- c(pheno.d7, pheno.d10, pheno.d13)
```

```
colData(merge2)$Phenotype <- pheno.merge2


# Cleaning Up the Phenotype Data so it belongs to the 3 populations

# Define phenotype groups
phenotype_groups <- list(
    Raneg_Cneg = c("CD34+CD45RA-CLEC12A-", "CD34-CD45RA-CLEC12A-"), # Ra-C-
    Rapos_Cneg = c("CD34+CD45RA+CLEC12A-", "CD34-CD45RA+CLEC12A-"), # Ra+C-
    Cpos = c("CD34-CD45RA-CLEC12A+", "CD34+CD45RA-CLEC12A+", "CD34+CD45RA+CLEC12A+"
    Other = c("CD10+", "CD14CD15+") # Pro -B #Pro-NM #FW Gating from a flow cytomet
)

# Assign group labels to phenotypes
group_labels <- sapply(pheno.merge2, function(phenotype) {
    group <- names(phenotype_groups)[sapply(phenotype_groups, function(g) phenotype
    if (length(group) > 0) group else "Other"
})

# Add group labels to colData of the SCE object
colData(merge2)$Group <- group_labels
```

Next, lets add a column indicating which day each observation belongs too, as we will be doing time sensitive analysis in the coming sections as well:

In [3]:
```
# Extract the day information from cell names
colData(merge2)$Day <- gsub(".*Day_([0-9]+).*", "\\1", rownames(colData(merge2)))
# Convert to a factor (optional, for better categorical handling)
colData(merge2)$Day <- factor(colData(merge2)$Day, levels = sort(unique(colData(mer


# Define columns to keep
cols_to_keep <- c("Group", "Day", "Phenotype")

# Create a lighter version of the SCE object
merge2_light <- merge2
colData(merge2_light) <- colData(merge2_light)[, cols_to_keep]

# Checking characteristics of the new light version
merge2_light
colnames(colData(merge2_light))
```

```
class: SingleCellExperiment
dim: 36601 12181
metadata(12): Samples scDblFinder.stats ... scDblFinder.stats
  scDblFinder.threshold
assays(2): counts logcounts
rownames(36601): MIR1302-2HG FAM138A ... AC007325.4 AC007325.2
rowData names(3): ID Symbol Type
colnames(12181): cell1Day_7 cell2Day_7 ... cell5137Day_13
  cell5138Day_13
colData names(3): Group Day Phenotype
reducedDimNames(9): PCA.cc UMAP.cc ... PCA TSNE
mainExpName: Gene Expression
altExpNames(1): Antibody Capture
```

'Group' · 'Day' · 'Phenotype'

Nice! As a Sanity Check, let's explore the makeup of the 3 populations, and their timepoints, using a simple tibble:

In [4]:
```
table(colData(merge2_light)$Group, colData(merge2)$Day)
```

```
             10   13    7
Cpos       1407 2219    0
Other       264 1023    0
Raneg_Cneg 1815  608 3039
Rapos_Cneg  766 1040    0
```

- Day 7 data exists only for Raneg_Cneg (i.e., primitive/early cells). Makes sense as only Ra-C- were sorted on Day 7, making it the input population.
- Day 10 and Day 13 have more mature populations like Cpos and Rapos_Cneg.

This makes sense developmentally and supports the idea that CD7 expression might increase over time as cells mature.

Finally, since CD7 **does** exist inthe ADT Assay of the Cite-Seq data (that is, Fangwu tagged the antibody with a fluorochrome during experimental setup), we will be able to use the Flow Cytometry results to to gate CD7+ and CD7- populations.

We will do this in conjunction with CD7 gene RNA Distribution and, annotate CD7hi population and CD7lo populations.

In [5]:
```
# Check available assays in the Antibody Capture altExp
assayNames(altExp(merge2, "Antibody Capture"))

# Check if "CD7" is among the rownames
"CD7" %in% rownames(altExp(merge2, "Antibody Capture"))

# Or view all available markers:
rownames(altExp(merge2, "Antibody Capture"))
```

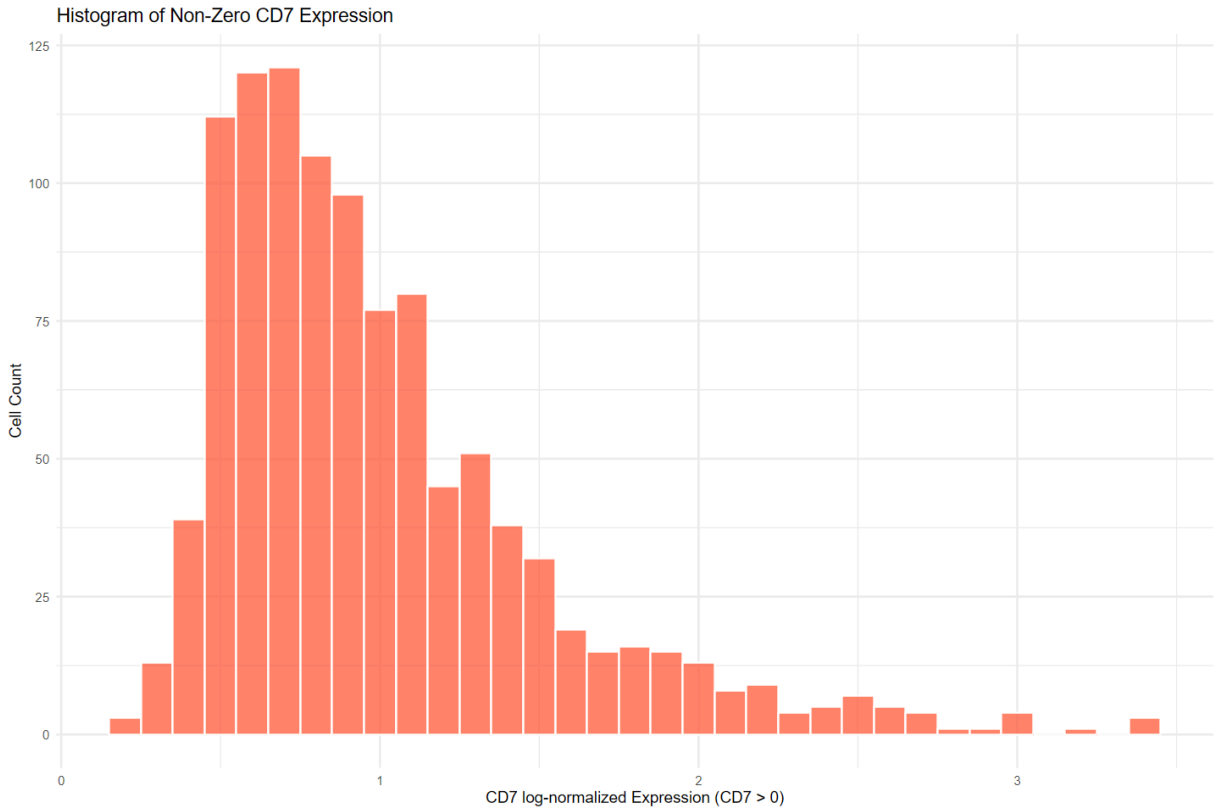
'counts' · 'logcounts'

FALSE

'TOTALSEQB_CLEC12A' · 'TOTALSEQB_CD45RA' · 'TOTALSEQB_CD10' · 'TOTALSEQB_CD7' · 'TOTALSEQB_CD34' · 'TOTALSEQB_CD14' · 'TOTALSEQB_CD15'

In [6]:
```
# Creating a Violin Plot to Generate CD7 Distribution Across All Cells Overall

options(repr.plot.width = 12, repr.plot.height = 8)

# 1. Extract CD7 expression from the RNA assay
cd7_expr <- logcounts(merge2_light)["CD7", ]
```

```r
# 2. Add CD7 expression to metadata
colData(merge2_light)$CD7_gene <- cd7_expr

# Filter to cells with CD7_expr > 0
# becuase otherwise the plot becomes right skewed and useless
nonzero_df <- as.data.frame(colData(merge2_light)) %>%
  filter(CD7_gene > 0)

# Plot histogram of non-zero CD7 expression
ggplot(nonzero_df, aes(x = CD7_gene)) +
  geom_histogram(binwidth = 0.1, fill = "tomato", color = "white", alpha = 0.8) +
  labs(
    title = "Histogram of Non-Zero CD7 Expression",
    x = "CD7 log-normalized Expression (CD7 > 0)",
    y = "Cell Count"
  ) +
  theme_minimal()
```



Nice! We can see the clean distribution using this histogram! Here's my interpretation of it:

- There's a dense peak between ~0.5 to ~1.5, with a long right-skewed tail extending up to ~3.2.
- This suggests a natural break point between low-but-detectable CD7 expression and robust $CD7^+$ expression.

So this is the gating structure I will be implementing, based on this:

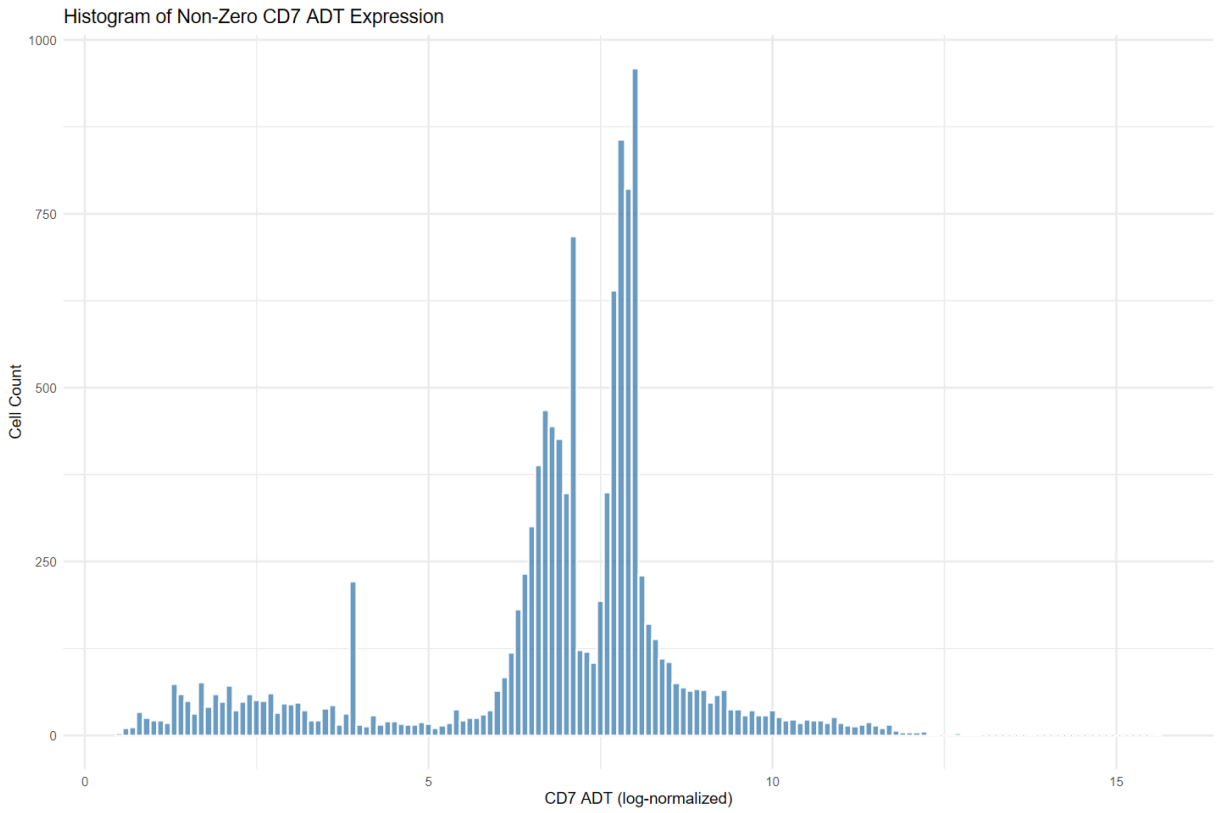
CD7 Status	Expression Range (CD7_expr)	Rationale
CD7_0	≤ 0	Truly negative
CD7-lo	> 0 & < 1	Weak expression; transitional or noise
CD7-hi	≥ 1	Stronger, confident expression

Let implement it in the code chunk below:

In [7]:
```
colData(merge2_light)$CD7_geneStat <- case_when(
  merge2_light$CD7_gene == 0 ~ "CD7_0",
  merge2_light$CD7_gene >= 1 ~ "CD7_hi",
  TRUE ~ "CD7_lo" # Optional middle category
)
```

Great - now lets do the same with the ADT flow cytometry data as well:

In [8]:
```
# Extract CD7 ADT signal from Antibody Capture assay
cd7_adt <- logcounts(altExp(merge2_light, "Antibody Capture"))["TOTALSEQB_CD7", ]
colData(merge2_light)$CD7_adt <- cd7_adt

# Filter to non-zero ADT values to avoid heavy skew
adt_nonzero <- as.data.frame(colData(merge2_light)) %>%
  filter(CD7_adt > 0)

# Plot histogram
ggplot(adt_nonzero, aes(x = CD7_adt)) +
  geom_histogram(binwidth = 0.1, fill = "steelblue", color = "white", alpha = 0.8) +
  labs(
    title = "Histogram of Non-Zero CD7 ADT Expression",
    x = "CD7 ADT (log-normalized)",
    y = "Cell Count"
  ) +
  theme_minimal()
```



Interesting, looking at this plot, here's what I see:

- Left peak: Around ~2.5–4.5 → likely $CD7^-$ or low-expressing population.
- Right peak: Sharp and dominant around ~7.5–9 → clearly $CD7^+$ cells.
- There's a clear valley between ~5.5 and 6.5, which is ideal for a gating threshold.

The **bimodal distribution** is super clear here. Based on what I see, here's the gating strategy I propose:

Gating Strategy for CD7 ADT:

Gate	Range (CD7_adt)	Rationale
CD7-	`≤ 1`	Undetectable expression (log-space 0)
CD7~	`> 1 & < 6`	Ambiguous or transitional zone
CD7+	`≥ 6`	Robust protein-level CD7 expression

In [9]:
```
colData(merge2_light)$CD7_adtStat <- case_when(
  merge2_light$CD7_adt <= 1 ~ "CD7-",
  merge2_light$CD7_adt >= 6 ~ "CD7+",
  TRUE ~ "CD7~"
)

colnames(colData(merge2_light))
```

'Group' · 'Day' · 'Phenotype' · 'CD7_gene' · 'CD7_geneStat' · 'CD7_adt' · 'CD7_adtStat'

Finally, as a sanity check, I want to compare the levels of Protein expression and RNA expression across the entire population, to see if it correlates generally:

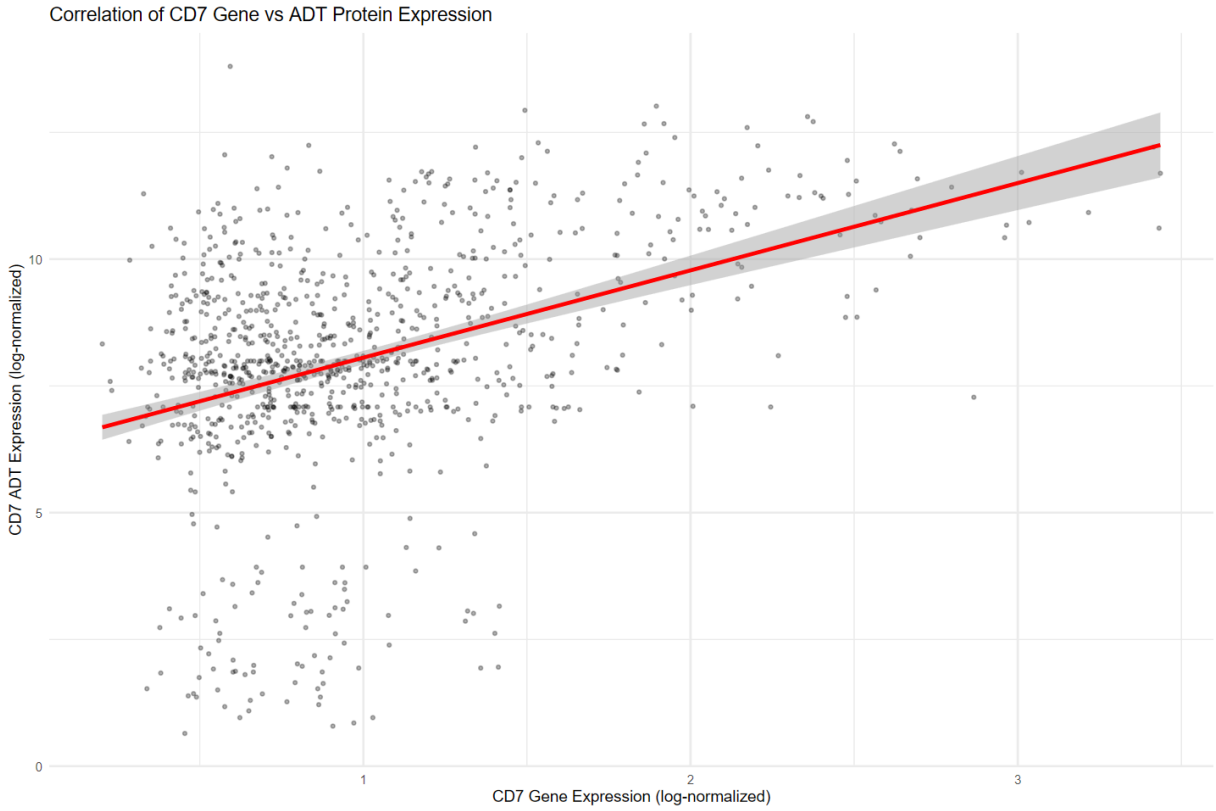
In [10]:
```
# Extract both columns from colData
corr_df <- as.data.frame(colData(merge2_light)) %>%
  filter(CD7_gene > 0, CD7_adt > 0, Group != "Other") # optional: restrict to nonze


library(ggplot2)

ggplot(corr_df, aes(x = CD7_gene, y = CD7_adt)) +
  geom_point(alpha = 0.3, size = 0.7) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(
    title = "Correlation of CD7 Gene vs ADT Protein Expression",
    x = "CD7 Gene Expression (log-normalized)",
    y = "CD7 ADT Expression (log-normalized)"
  ) +
  theme_minimal()

# Pearson correlation (linear relationship)
cor.test(corr_df$CD7_gene, corr_df$CD7_adt, method = "pearson")

# Optionally: Spearman correlation (rank-based)
cor.test(corr_df$CD7_gene, corr_df$CD7_adt, method = "spearman")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
        Pearson's product-moment correlation

data:  corr_df$CD7_gene and corr_df$CD7_adt
t = 13.127, df = 963, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3347217 0.4418440
sample estimates:
      cor
0.3895997
```

```
Warning message in cor.test.default(corr_df$CD7_gene, corr_df$CD7_adt, method = "spe
arman"):
"Cannot compute exact p-value with ties"
```

```
        Spearman's rank correlation rho

data:  corr_df$CD7_gene and corr_df$CD7_adt
S = 98623959, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3415054
```



In [11]:
```
table(CD7_gene = merge2_light$CD7_gene > 0, CD7_adt = merge2_light$CD7_adt > 0)
```

```
        CD7_adt
CD7_gene FALSE  TRUE
   FALSE   432 10685
   TRUE     24  1040
```

Oh Dear! This is quiet scary. It appears that there are 10,685 cells that are producing the CD7 ADT protein, but have CD7 gene expression. Look's like this is a case of non-specific binding (or as Ross called, it, CITE-Seq being "Sticky").

We will fix this by re-labelling cells with Zero CD7 expression as being "CD7-", regardless of protein level.

In [12]:
```
# Identify which cells have CD7_gene == 0
zero_gene_idx <- which(colData(merge2_light)$CD7_gene == 0)

# Set their ADT expression to 0
colData(merge2_light)$CD7_adt[zero_gene_idx] <- 0

# Set their ADT status to "CD7-"
colData(merge2_light)$CD7_adtStat[zero_gene_idx] <- "CD7-"

# Check the updated data frame
table(CD7_gene = merge2_light$CD7_gene > 0, CD7_adt = merge2_light$CD7_adt > 0)
```

```
        CD7_adt
CD7_gene FALSE  TRUE
   FALSE 11117     0
   TRUE     24  1040
```

Great! We are now ready to begin working on the first Research Question of this analysis.

Analysis Workflow

RQ1: Is There a Difference in CD7 expression between Days 7, 10 and 13, across all HSCs or within each Population?

CD7 may be a marker that increases as cells progress towards a T-lineage. Understanding its temporal dynamics across differentiation days should be the first sanity check we conduct to see if it is actually doing anything. The overall population of cells are maturing over time in LEM culture (which is designed to support Lymphoid specificiation). CD7 is a surface marker that is belived to be enriched in early T-lineage committed cells.

Given that the progression of populations from Day 7 to Day 13 involves a lot of maturation and lineage specification, one would expect to see higher levels of CD7 as time progresses across the overall CD34+ HSC population as a whole, and within each of the individual population subsets.

This is what we will explore and interpret in this section. We will do so in 2 phases:

- **Phase 1 - Qualitative CD7 RNA and ADT Trends over Time**: We will plot CD7 RNA Expression across Days 7, 10 and 13, (across all HSCs, and within each population group), and then visualize the proportion of CD7+ positivity as time progresses, helping see qualitatively if CD7 increases with time..
- **Phase 2- Statistical Testing of CD7 Trends Over Time**: We will conduct statistical testing to discern differences between CD7 expression level (ADT or GeneExpression) and their populations.

Lets begin with implementing phase 1

Phase 1: Qualitative Visualization and Interpretation of General Trends

Let's begin by generating our first faceted plot,containing CD7 expression Vs. Day (All HSCs, filtering CD7+/Low only), and the same information faceted by population (3 Pops, filtering CD7+/Low only):

In [13]:
```
# Implementing the BoxPlot Visualizations for CD7 Expression Across All Popuulation

options(repr.plot.width = 18, repr.plot.height = 9)

# Filter out CD7- cells
cd7_filtered_df <- as.data.frame(colData(merge2_light)) %>%
  filter(CD7_geneStat != "CD7_0") %>%
  filter(Group != "Other")
```

```r
# Reorder Day as a factor with levels in desired order
cd7_filtered_df$Day <- factor(cd7_filtered_df$Day, levels = c("7", "10", "13"))

# Plot 1: Overall Boxplot (All groups combined)
CD7_gene_expression_plot_HSC <- ggplot(cd7_filtered_df, aes(x = Day, y = CD7_gene))
  geom_boxplot(fill = "steelblue", alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.2, alpha = 0.2, size = 0.5) +
  labs(
    subtitle = "For All HSCs",
    x = "Day",
    y = "CD7 Log-Normalized Gene Expression"
  ) +
  theme_minimal()

# Plot 2: Faceted by Population Group
CD7_gene_expression_plot_GROUP <- ggplot(cd7_filtered_df, aes(x = Day, y = CD7_gene
  geom_boxplot(fill = "darkorange", alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.2, alpha = 0.2, size = 0.5) +
  facet_wrap(~Group) +
  labs(
    subtitle = "By Population",
    x = "Day",
    y = ""
  ) +
  theme_minimal()

CombPlot_CD7_gene_expression <- CD7_gene_expression_plot_HSC + CD7_gene_expression_
  plot_annotation(title = "CD7 Gene Expression Across Time by HSC Population & Grou
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    plot.subtitle = element_text(size = 15, face = "italic"),
    strip.text = element_text(size = 12)
  )
```

Great! Now, let's do the same with our ADT data:

In [14]:

```r
options(repr.plot.width = 18, repr.plot.height = 9)

# Filter out CD7- (ADT = 0) cells for meaningful distribution
adt_filtered_df <- as.data.frame(colData(merge2_light)) %>%
  filter(CD7_adt > 0 & Group != "Other")

# Ensure Day is treated as an ordered factor
adt_filtered_df$Day <- factor(adt_filtered_df$Day, levels = c("7", "10", "13"))

# Plot 1: CD7 ADT expression over time (All HSCs)
CD7_adt_count_plot_HSC <- ggplot(adt_filtered_df, aes(x = Day, y = CD7_adt)) +
  geom_boxplot(fill = "steelblue", alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.2, alpha = 0.1, size = 0.5) +
  labs(
    subtitle = "All HSCs",
    x = "Day",
    y = "CD7 ADT log-normalized Expression"
```

```
  ) +
  theme_minimal()

# Plot 2: Faceted by Group
CD7_adt_count_plot_GROUP <- ggplot(adt_filtered_df, aes(x = Day, y = CD7_adt)) +
  geom_boxplot(fill = "darkorange", alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.2, alpha = 0.1, size = 0.5) +
  facet_wrap(~Group) +
  labs(
    subtitle = "By Population",
    x = "Day",
    y = ""
  ) +
  theme_minimal()


CombPlot_CD7_adt_expression <- CD7_adt_count_plot_HSC + CD7_adt_count_plot_GROUP +
  plot_annotation(title = "Proportion of CD7 ADT Log Fluorescence Across Time by HS
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    plot.subtitle = element_text(size = 15, face = "italic"),
    strip.text = element_text(size = 12)
  )
```

Amazing! Now, let's generate the Proportion plots based on the GeneExpression data:

In [15]:
```
options(repr.plot.width = 18, repr.plot.height = 9)

# Ensure Day is a factor in the right order
cd7_filtered_df$Day <- factor(cd7_filtered_df$Day, levels = c("7", "10", "13"))

# Create the normalized stacked bar plot
CD7_gene_prop_plot_HSC <- ggplot(cd7_filtered_df, aes(x = Day, fill = CD7_geneStat)
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(
    subtitle = "All HSCs",
    x = "Day",
    y = "Proportion of Cells",
    fill = "CD7 Status"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"), # nolint
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    plot.subtitle = element_text(size = 15, face = "italic"),
    strip.text = element_text(size = 12)
  ) +
  theme_minimal()

CD7_gene_prop_plot_GROUP <- ggplot(cd7_filtered_df, aes(x = Day, fill = CD7_geneSta
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
```

```
  labs(
    subtitle = "By Population",
    x = "Day",
    y = "",
    fill = "CD7 Gene Status"
  ) +
  facet_wrap(~Group) +
  theme_minimal()

CombPlot_CD7_gene_prop <- CD7_gene_prop_plot_HSC + CD7_gene_prop_plot_GROUP + plot_
  plot_annotation(title = "Proportion of CD7 Gene Expression Across Time by HSC Pop
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    plot.subtitle = element_text(size = 15, face = "italic"),
    strip.text = element_text(size = 12)
  )
```

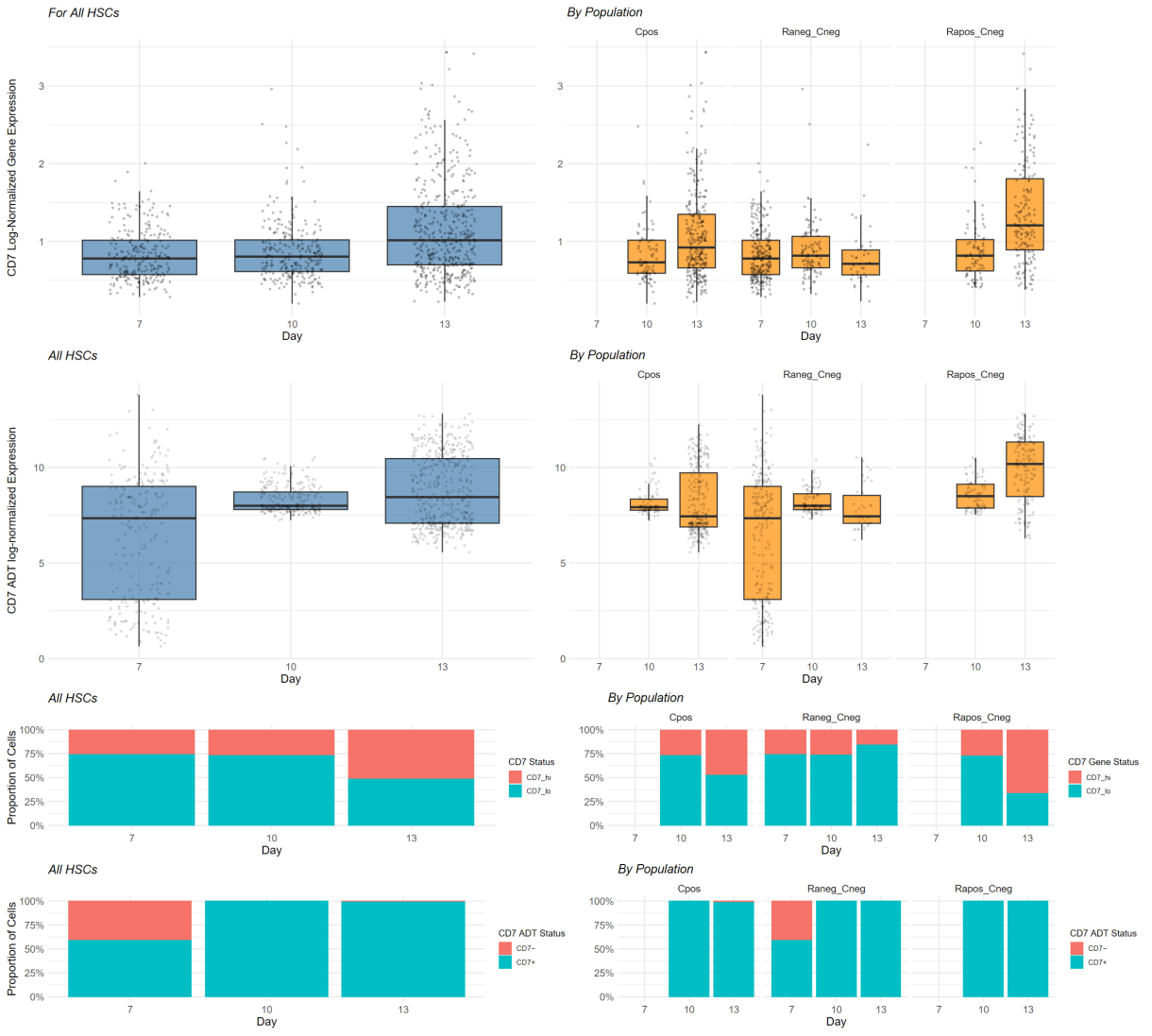
Great! Now, we repeat with the ADT data:

In [16]:
```
# Ensure Day is a factor in the right order
adt_filtered_df <- as.data.frame(colData(merge2_light)) %>%
  filter(CD7_adtStat != "CD7-" & Group != "Other")

# Ensure Day is treated as an ordered factor
adt_filtered_df$Day <- factor(adt_filtered_df$Day, levels = c("7", "10", "13"))


# ---- Plot 1: Overall Proportion Across All HSCs ----
CD7_adt_prop_plot_HSC <- ggplot(adt_filtered_df, aes(x = Day, fill = CD7_adtStat))
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(
    subtitle = "All HSCs",
    x = "Day",
    y = "Proportion of Cells",
    fill = "CD7 ADT Status"
  ) +
  theme_minimal()

# ---- Plot 2: Faceted by Population Group ----
CD7_adt_prop_plot_GROUP <- ggplot(adt_filtered_df, aes(x = Day, fill = CD7_adtStat)
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(
    subtitle = "By Population",
    x = "Day",
    y = "",
    fill = "CD7 ADT Status"
  ) +
  facet_wrap(~Group) +
  theme_minimal()

# ---- Combine the two plots ----
```

```
CombPlot_CD7_adt_prop <- CD7_adt_prop_plot_HSC + CD7_adt_prop_plot_GROUP +
  plot_layout(ncol = 2) +
  plot_annotation(
    title = "Proportion of CD7 ADT Expression Across Time by HSC Population & Group
    theme = theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"))
  ) & # nolint
  theme(
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    plot.subtitle = element_text(size = 15, face = "italic"),
    strip.text = element_text(size = 12)
  )
```

Woohoo! That was a lot of code. Let's now combine all the plots into one figure, and interpret what we see:

In [17]:
```
options(repr.plot.width = 20, repr.plot.height = 18)

((CombPlot_CD7_gene_expression / CombPlot_CD7_adt_expression) + (CombPlot_CD7_gene_
    theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold"))
```



Here's my interpretations of what I am seeing:

First Row: CD7-Log Normalized Gene Expression Data, for All HSCs and By Population

CD7 Expression Across Time (All HSCs)

- **Trend:** There is a **clear increase** in CD7 expression from Day 7 → Day 13.
- Median CD7 expression appears to **increase steadily**, and there's a visible rise in the spread of values on Day 13.
- Day 13 shows a broader interquartile range (IQR) and more extreme outliers — suggesting **higher transcriptional heterogeneity** in $CD7^+$ cells over time.
- **Interpretation:** This supports the idea that **$CD7^+$ cells become more prominent or transcriptionally active as differentiation proceeds**, consistent with increasing T-lineage priming over time.

CD7 Expression Across Time by Population

- This is **really insightful** for dissecting cell-type-specific dynamics. Here's what we can see: **Cpos (C^+)**
- Starts with moderate expression on Day 10, and by Day 13 CD7 expression is **dramatically higher**.
- Strongest rise among all populations.
- This might indicate that **C^+ cells are strongly acquiring T-lineage bias** as time progresses.

Rapos_Cneg (Ra^+C^-)

- Similar to Cpos, with a noticeable bump in expression by Day 13.
- Suggests this population is **also transitioning**, though perhaps slightly delayed or with more heterogeneity.

Raneg_Cneg (Ra^-C^-)

- Expression is relatively flat across all timepoints.
- This supports the idea that Ra^-C^- cells are **maintaining a primitive state** with **minimal T-lineage activation**.

Summary Takeaways

- **CD7 expression increases over time** across all cells, particularly in **Ra^+C^- and C^+ populations**.
- **Ra^-C^- cells remain low and flat**, suggesting **they are least committed to T-lineage** — reinforcing their use as a baseline for comparison.
- I am seeing evidence that **CD7 tracks with populations known to acquire T-cell characteristics over time** at this stage.

2nd Row: CD7-Log Normalized ADT Expression Data, for All HSCs and By Population

The Results align completely with Row 1 on a qualitive level, so no additional interpretation is needed.

3rd Row: CD7-Log Normalized Changes in Proportion of Cells Over Time (GeneExpression Gating), for All HSCs and By Population

Proportion of CD7 Status Across Days (All HSCs)

- **$CD7^+$ cells increase from Day 7 to Day 13**, showing a rising trend in CD7 expression over time in the HSC population as a whole.
- This suggests that **CD7 expression may be associated with HSC maturation or progression over time**.

Proportion of CD7 Status Across Days by HSC Population

- **Cpos and Rapos_Cneg populations show a strong increase in $CD7^+$ cells by Day 13**.
- **RaNeg_Cneg remains mostly $CD7^-$ across all timepoints**, indicating it retains a more primitive phenotype.
- This supports the idea that **CD7 expression marks more mature or lineage-committed HSCs**, particularly T-lineage-biased ones.

4th Row: CD7-Log Normalized Changes in Proportion of Cells Over Time (Surface Marker Gating), for All HSCs and By Population

The Results align completely with Row 3 on a qualitive level, so no additional interpretation is needed.

We can now begin conducting Statistical tests in `Phase 2` of our analysis.

Phase 2: Statistical Analysis

Here, we will conduct statistical differences to discern differences between CD7 expression level and time, by ADT data and GeneExpression data. Let's begin by Looking at GeneExpression data:

In [18]:
```
cd7_day_slr <- lm(CD7_gene ~ Day, data = cd7_filtered_df)

tidy(cd7_day_slr)
```

A tibble: 3 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	0.82270538	0.03099045	26.547058	1.358152e-117
Day10	0.04823789	0.04361752	1.105929	2.690270e-01
Day13	0.32967080	0.03854306	8.553311	4.517974e-17

Great! Here's the meaning of these results, as taught to me in STAT 301:

- There is enough statistical evidence to make the claim that the mean log-CD7 expression in all HSCs (regardless of population) of the reference level (Day7) is 0.823, and is statistically significant at $p < 0.05$.
- There is not enough statistical evidence to make the claim that the mean log-CD7 expression in all HSCs (regardless of population) of Day 10 cells is different from the Day 7 cells.
- There is enough statistical evidence to make the claim that the mean log-CD7 expression in all HSCs (regardless of population) of Day 13 cells is 0.33 more than the Day 7 cells, at $p < 0.05$.

Let's now repeat this experiment by focusing on the interaction between time and population, worked on below:

In [19]:
```r
# relevelling so Ra-C- is the reference level
cd7_filtered_df$Group <- factor(cd7_filtered_df$Group)
cd7_filtered_df$Group <- relevel(cd7_filtered_df$Group, ref = "Raneg_Cneg")

# relevelling so Day 10 is the reference level
cd7_filtered_df$Day <- relevel(cd7_filtered_df$Day, ref = "10")

cd7_daygroup_slr <- lm(CD7_gene ~ Day * Group, data = cd7_filtered_df)

tidy(cd7_daygroup_slr)
```

A tibble: 9 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	0.889509825	0.04860987	18.29895522	1.367600e-64
Day7	-0.066804442	0.05721144	-1.16767626	2.432208e-01
Day13	-0.097935558	0.09229814	-1.06107837	2.889152e-01
GroupCpos	-0.056651580	0.07221989	-0.78443183	4.329759e-01
GroupRapos_Cneg	-0.002703804	0.07246332	-0.03731272	9.702432e-01
Day7:GroupCpos	NA	NA	NA	NA
Day13:GroupCpos	0.344717880	0.11058310	3.11727451	1.878406e-03
Day7:GroupRapos_Cneg	NA	NA	NA	NA
Day13:GroupRapos_Cneg	0.574620773	0.11348220	5.06353201	4.913448e-07

Here's the meaning of these results:

- `(Intercept)` : There is enough statistical evidence to make the claim that the mean log-CD7 expression for Ra-C- cells at Day 10 is 0.89, and this is statistically significant ($p < 0.05$).
- `Day7` : There is not enough evidence to make the claim that the mean CD7 expression in Ra-C- cells at Day 7 differs from Day 10
- Day13`: There is not enough evidence to make the claim that the mean CD7 expression in Ra-C- cells at Day 13 differs from Day 10.

- `GroupCPos` : There is not enough evidence to make the claim that the mean CD7 expression in C+ cells at Day 7 differs from Ra-C- cells at Day 10.
- `GroupRapos_Cneg` : There is not enough evidence to make the claim that the mean CD7 expression in Ra+C- cells at Day 7 differs from Ra-C- cells at Day 10.

- `Day13:GroupCpos` : There is enough statistical evidence to claim that the Day 13 C+ cells have a 0.345 higher log-CD7 expression than Ra-C- cells at Day 10, based on the individual effects of being C+ and Day 13 alone.
- `Day13:GroupRapos_Cneg` : There is strong statistical evidence to claim that Day 13 Ra+C- cells have a 0.57 higher CD7 expression than Ra-C- cells at Day 10 based on the effects of being Ra+C- and Day 13 alone.

> These interaction effects support the idea that CD7 expression increases specifically in more mature HSCs over time — especially in populations like Ra+C-, possibly marking T-lineage priming.

> ## *No statiatical difference in CD7 expression among populations on Day 7 compared to*

> ## *Day 10*

> ## *This means CD7 expression in Ra-C- population does not chhange with time.*

Let's now try this again, but with ADT expression data:

In [20]:
```
adt_day_slr <- lm(CD7_adt ~ Day, data = adt_filtered_df)

tidy(cd7_day_slr)
```

A tibble: 3 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	0.82270538	0.03099045	26.547058	1.358152e-117
Day10	0.04823789	0.04361752	1.105929	2.690270e-01
Day13	0.32967080	0.03854306	8.553311	4.517974e-17

Here's the interpretation of these results:

There is a **clear and significant increase** in CD7 ADT expression from **Day 7** to **Day 13**.

- `(Intercept)` : The mean log-CD7 ADT expression in all HSCs at Day 7 is **0.88**, and this is statistically significant ($p \ll 0.05$).
- `Day 10` : There is not enough **statistical evidence** to claim that **Day 10** cells have CD7 ADT expression different than Day 7 cells .
- `Day 13` : There is enough **statistical evidence** to claim that **Day 13** cells have **0.33 units higher** CD7 ADT expression than Day 7 cells (**$p \ll 0.05$**).

> This suggests that **surface expression of CD7 protein is increasing steadily over time in HSCs**, in line with gene expression data — but with even greater magnitude.

Now let's repeat this model, but with the interaction terms!

In [21]:
```
# relevelling so Ra-C- is the reference level
adt_filtered_df$Group <- factor(adt_filtered_df$Group)
adt_filtered_df$Group <- relevel(adt_filtered_df$Group, ref = "Raneg_Cneg")

# relevelling so Day 10 is the reference level
adt_filtered_df$Day <- relevel(adt_filtered_df$Day, ref = "10")

adt_daygroup_slr <- lm(CD7_gene ~ Day * Group, data = adt_filtered_df)

tidy(adt_daygroup_slr)
```

A tibble: 9 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	0.889509825	0.04913255	18.10428668	3.431332e-63
Day7	-0.055293359	0.05884042	-0.93971723	3.476007e-01
Day13	-0.097935558	0.09329059	-1.04979037	2.940806e-01
GroupCpos	-0.056651580	0.07299644	-0.77608686	4.378903e-01
GroupRapos_Cneg	-0.002703804	0.07324249	-0.03691578	9.705599e-01
Day7:GroupCpos	NA	NA	NA	NA
Day13:GroupCpos	0.344717880	0.11177216	3.08411222	2.100430e-03
Day7:GroupRapos_Cneg	NA	NA	NA	NA
Day13:GroupRapos_Cneg	0.574620773	0.11470244	5.00966498	6.496432e-07

- `(Intercept)` : There is enough statistical evidence to make the claim that the mean log-CD7 ADT expression for Ra−C− cells at Day 10 is 0.89, and this is statistically significant ($p < 0.05$).
- `Day7` : There is not enough statistical evidence to make the claim that the mean CD7 ADT expression in Ra−C− cells at Day 7 differs from Day 10.
- `Day13` : There is not enough statistical evidence to make the claim that the mean CD7 ADT expression in Ra−C− cells at Day 13 differs from Day 10.

> *This means CD7 ADT expression in Ra−C− population does not change with time. --*

- `GroupCpos` : There is not enough statistical evidence to make the claim that the mean CD7 ADT expression in C+ cells at Day 10 differs from Ra−C− cells at Day 10.
- `GroupRapos_Cneg` : There is not enough evidence to make the claim that the mean CD7 ADT expression in Ra+C− cells at Day 10 differs from Ra−C− cells at Day 10.

> *No statistical difference in CD7 ADT expression among populations on Day 10 compared to Ra−C− --*

- `Day13:GroupCpos` : There is enough statistical evidence to claim that the Day 13 C+ cells have a 0.345 higher log-CD7 ADT expression than Ra−C− cells at Day 10, based on the combined effects of being C+ and Day 13.
- `Day13:GroupRapos_Cneg` : There is strong statistical evidence to claim that Day 13 Ra+C− cells have a 0.57 higher CD7 ADT expression than Ra−C− cells at Day 10, based on the effects of being Ra+C− and Day 13.

> *These interaction effects support the idea that CD7 ADT expression increases specifically in more mature HSCs over time — especially in populations like Ra+C−, possibly marking T-lineage priming.*

RQ2: Is There a Statistical Difference in CD7 Expression Between Ra-C- cells (Primitive), and Ra-C+/C+ Populations (More Mature)?

In haematopoiesis, surface markers like CD7 help us trace lineage commitment - especially towards T-cell fates. From the transcriptomic trajectory and prior results (RQ1), it appears that CD7 expression increases over real time (Day 7->13), suggesting it's tied to cellular maturation or lineage priming.

We now want to ask: **Does this trend correlate with more mature immunophenotypes?**

- Ra-C$^-$ cells are considered the most primitive.
- Ra-C$^+$ and especially C$^+$ are considered more differentiated.

So, if CD7 is truly a marker of T-lineage bias or maturity, we'd expect its expression to be significantly (in a statistical sense) lower in Ra-C$^-$, and higher in C$^+$/Ra-C$^+$.

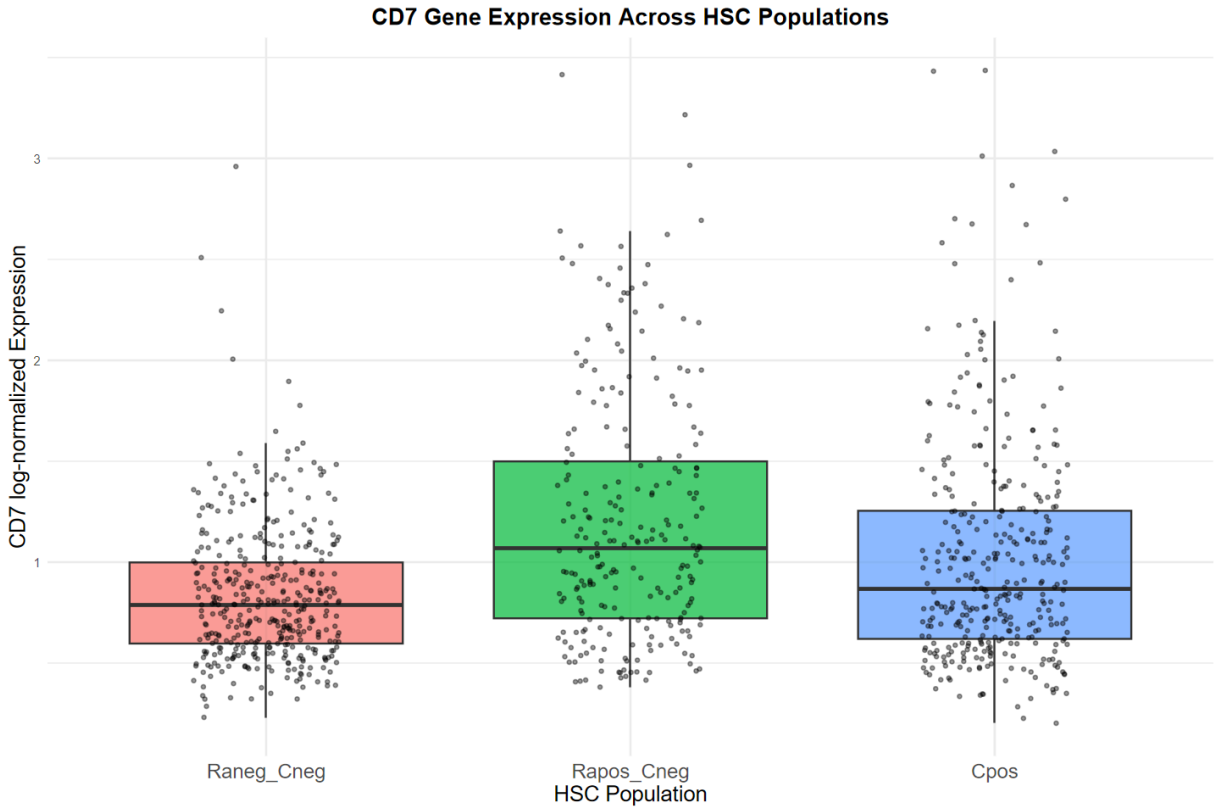
Some of the Analysis we will do:

- Boxplot Visualizations of the 3 populations CD7 expression, regardless of time.
- ANOVA test comparing the CD7 expression among the 3 populations

Lets begin with the CD7 visualization using boxplots:

In [22]:
```
options(repr.plot.width = 12, repr.plot.height = 8)

# Step 1: Filter to 3 populations and CD7+ / CD7_low cells only
rq2_df <- cd7_filtered_df %>%
  filter(
    Group %in% c("Raneg_Cneg", "Rapos_Cneg", "Cpos"),
    CD7_geneStat != "CD7_0"
  )

# Step 2: Ensure Group is a factor with a sensible order
rq2_df$Group <- factor(rq2_df$Group, levels = c("Raneg_Cneg", "Rapos_Cneg", "Cpos"))

# Step 3: Create boxplot
ggplot(rq2_df, aes(x = Group, y = CD7_gene, fill = Group)) +
  geom_boxplot(outlier.shape = NA, alpha = 0.7) +
  geom_jitter(width = 0.2, size = 0.8, alpha = 0.4) +
  labs(
    title = "CD7 Gene Expression Across HSC Populations",
    x = "HSC Population",
    y = "CD7 log-normalized Expression"
```

```
  ) +
  theme_minimal() +
  theme(
    legend.position = "none",
    axis.text.x = element_text(size = 14),
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.title = element_text(size = 15)
  )
```



- Rapos_Cneg (Ra-C$^+$) shows the highest median CD7 expression, and a wide distribution — suggesting a subset of strongly CD7$^+$ cells.
- Cpos (C$^+$) also has moderate CD7 expression, with slightly lower median than Rapos_Cneg but still elevated.
- Raneg_Cneg (Ra-C$^-$) has the lowest CD7 expression overall, with a tighter range and fewer high outliers.

Takeaway: CD7 expression increases in more differentiated populations, suggesting that CD7$^+$ cells are enriched in more mature, lineage-primed HSC subsets.

Let's now quantify this with ANOVA

In [23]:
```
# Fit linear model
cd7_group_lm <- lm(CD7_gene ~ Group, data = cd7_filtered_df)

# ANOVA test
anova(cd7_group_lm)
```

```
tidy(cd7_group_lm)
```

A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Group	2	20.50716	10.2535824	41.07162	7.361686e-18
Residuals	986	246.15616	0.2496513	NA	NA

A tibble: 3 × 5

	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)		0.8364888	0.02517206	33.230841	4.599150e-163
GroupCpos		0.1861486	0.03656331	5.091129	4.262167e-07
GroupRapos_Cneg		0.3661212	0.04091267	8.948847	1.747196e-18

Interpretation of ANOVA Results

`F = 41.07, p < 0.0001`

> There is enough statistical evidence to conclude that at least one HSC population has a different mean log-CD7 expression compared to the others.

Interpretation of Linear Model Results

- There is enough statistical evidence to make the claim that Mean log-CD7 expression in Ra-C^- cells is 0.837 (log-normalized).
- There is enough statistical evidence to make the claim that Ra+C^- cells have a log-CD7 expression 0.367 higher than Ra-C^- cells.
- There is enough statistical evidence to make the claim that C+ cells have a log-CD7 expression 0.186 higher than the Ra-C- cells.

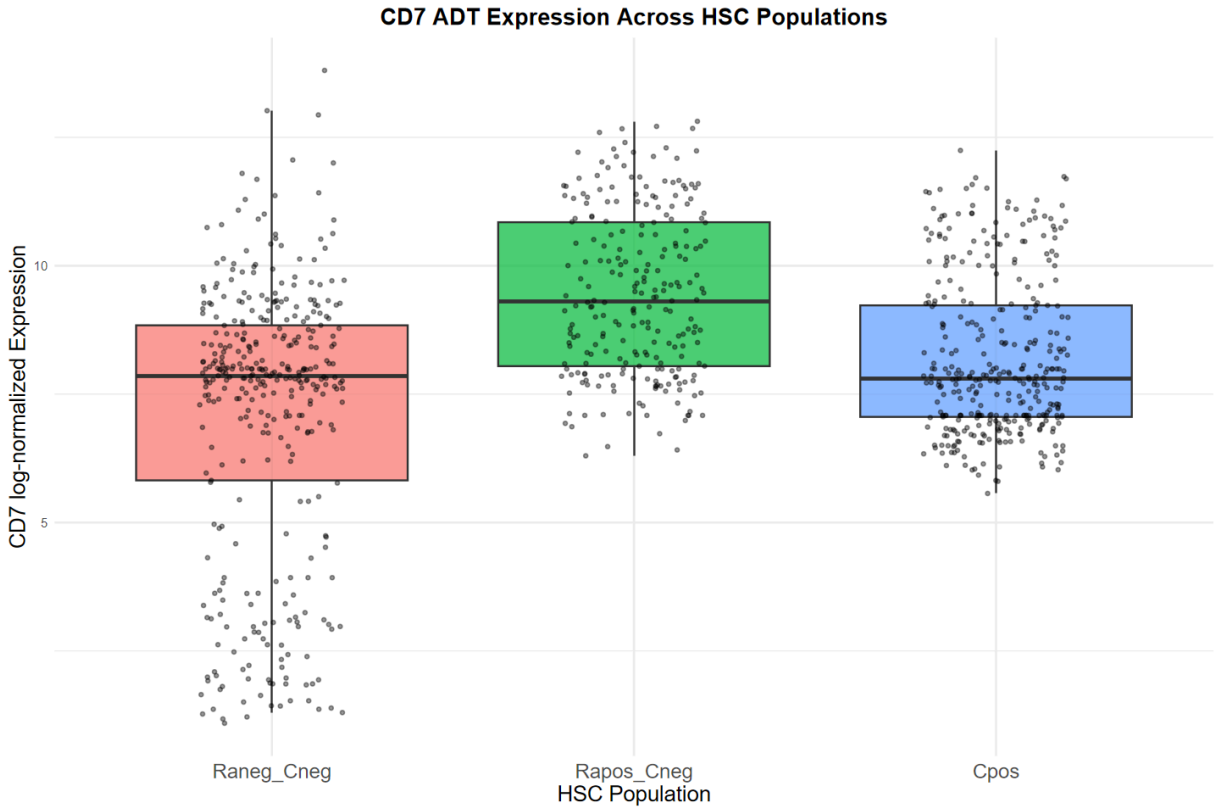
> There is enough evidence to conclude that CD7 levels differ across each HSC population, and the difference are statistically significant at at 95% confidence. RA+C- cells have the highest CD7 expression levels, followed by C+ and Ra-C- cells.

The ranking of CD7 expression by population is: Ra^+C^- > C^+ > Ra^-C^-, consistent with increasing expression in more mature, possibly T-lineage-primed subsets.

Let's Repeat this visualization and testing with the ADT data below:

In [24]:
```
# Part 1: Boxplot Visualization
```

```r
options(repr.plot.width = 12, repr.plot.height = 8)

# Step 1: Filter to 3 populations and CD7+ / CD7_low cells only
rq2_df <- adt_filtered_df %>%
  filter(
    Group %in% c("Raneg_Cneg", "Rapos_Cneg", "Cpos"),
    CD7_geneStat != "CD7_0"
  )

# Step 2: Ensure Group is a factor with a sensible order
rq2_df$Group <- factor(rq2_df$Group, levels = c("Raneg_Cneg", "Rapos_Cneg", "Cpos"))

# Step 3: Create boxplot
ggplot(rq2_df, aes(x = Group, y = CD7_adt, fill = Group)) +
  geom_boxplot(outlier.shape = NA, alpha = 0.7) +
  geom_jitter(width = 0.2, size = 0.8, alpha = 0.4) +
  labs(
    title = "CD7 ADT Expression Across HSC Populations",
    x = "HSC Population",
    y = "CD7 log-normalized Expression"
  ) +
  theme_minimal() +
  theme(
    legend.position = "none",
    axis.text.x = element_text(size = 14),
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold"),
    axis.title = element_text(size = 15)
  )
```



Amazing - the trend is the exact same as the Gene Expression plot, so no further intepretation is required. Let's move onto the statistical testing:

In [25]:
```r
# Fit linear model
cd7_group_lm <- lm(CD7_adt ~ Group, data = adt_filtered_df)

# ANOVA test
anova(cd7_group_lm)

tidy(cd7_group_lm)
```

A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Group	2	803.4769	401.738469	94.28439	4.214383e-38
Residuals	957	4077.7029	4.260923	NA	NA

A tibble: 3 × 5

	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)		7.101394	0.1080452	65.726153	0.000000e+00
GroupCpos		1.095596	0.1538712	7.120213	2.114764e-12
GroupRapos_Cneg		2.346750	0.1715448	13.680106	4.845681e-39

Interpretation of ANOVA Results

> **F = 94.28, p < 0.0001**
> There is enough statistical evidence to conclude that at least one HSC population has a different mean log-CD7 ADT expression compared to the others.

Interpretation of Linear Model Results

- There is enough statistical evidence to make the claim that Mean log-CD7 ADT expression in $Ra^{-}C^{-}$ cells is 7.10 (log-normalized).
- There is enough statistical evidence to make the claim that $Ra^{+}C^{-}$ cells have a log-CD7 ADT expression **2.35 higher** than $Ra^{-}C^{-}$ cells.
- There is enough statistical evidence to make the claim that C^{+} cells have a log-CD7 ADT expression **1.10 higher** than $Ra^{-}C^{-}$ cells.

> There is enough evidence to conclude that CD7 ADT levels differ across each HSC population, and the difference is statistically significant at a 95% confidence level.

> **Ra^+C^-** cells have the highest CD7 ADT expression levels, followed by **C^+** and **Ra^-C^-** cells.

> The ranking of CD7 ADT expression by population is: **$Ra^+C^- > C^+ > Ra^-C^-$**, consistent with increasing expression in more mature, possibly T-lineage-primed subsets.

This sets us up beautifully for RQ3:

RQ3: Does the Increase in CD7 Expression statisticallly correlate with an Increase in Known T-Cell Genes?

Fangwu identified the Ra^+C^- population as a transient, lymphoid-primed intermediate enriched for T/NK lineage gene programs. Given that CD7 is a canonical early T-cell marker, it would be expected that CD7 expression is elevated in Ra^+C^- cells (which it is, as per analysis done in RQ2). However, by quantifying CD7 expression and testing its correlation with broader T-lineage transcriptional signatures, I aim to determine whether CD7 truly reflects T-lineage bias — or whether it is insufficient as a standalone enrichment marker.

The key here is to statistically validate CD7 as a Proxy for T-cell bias.

To do this, we will be completing 4 Statistical tests, which will rely on the common workflow outlined here:

1. Define a T-Cell Signature. I will do so by using the T-cell gene list that Laura provided, which also include genes from John Edgar's thesis.
2. Score each cell's T-cell program, using SingleCellExperiment objects native module scoring functionality
3. Conduct Different Tests:

    a. Compare CD7 expression with T-cell program score

    b. Statistical Difference in T-cell correlation scores between CD7-, CD7 lo, CD7+ groups

Let's begin implementing this plan below:

In [26]:
```r
# Defining T-Cell Program Genes
# --- 1. Define the T-cell program gene list ---
tcell_genes <- c(
  "RAG2", "NOTCH1", "CD3D", "CD3G",
  "TCF7", "GATA3", "BCL11B", "RAG1", "DTX1",
  "IL7R", "PTCRA", "LEF1", "RUNX1",
  "BCL11A", "IKZF1"
)

# --- 2. Ensure genes are present in the SCE object ---
tcell_genes_present <- intersect(tcell_genes, rownames(merge2))
```

```
# Optional: print genes that were missing
missing_genes <- setdiff(tcell_genes, rownames(merge2))
cat("Missing genes:", paste(missing_genes, collapse = ", "), "\n")

# --- 3. Score each cell using average log-normalized expression ---
# (assuming logcounts(merge2) contains log-normalized RNA expression)
tcell_scores <- colMeans(logcounts(merge2_light)[tcell_genes_present, , drop = FALS

# --- 4. Store the T-cell score in the metadata ---
colData(merge2_light)$Tcell_score <- tcell_scores
```

```
Missing genes:
```

Here is the justification for the genes I addded to Laura's original list:

- **IL7R** — Essential cytokine receptor for T-cell development; early expression bias.
- **PTCRA** — Pre-TCR alpha chain, seen in early thymocyte development.
- **LEF1** — Wnt pathway TF that overlaps with TCF7 and is T-lineage linked.
- **RUNX1** — Highly conserved regulator of lymphoid commitment (T/B), also activated early.
- **BCL11A** — Paralog of BCL11B, also involved in early lymphoid fate decision.
- **IKZF1 (Ikaros)** — Key lymphoid TF, expressed early in T- and B-lineage priming.

We can now begin conducting our tests to validate the counter hypothesis: CD7 alone does not track well wth the broader T-lineage transcriptional program. As mentioned, above, we will be implementing this using 3 tests:

Test Alpha: Correlation Between CD7 Expression and T-Cell Module Score

This will quantify whether CD7 expression alone tracks well with the broader T-lineage transcriptional program. The steps I will implement will be:

1. Filter to non-zero CD7-expressing cells (to avoid skew).
2. Plot a scatterplot of CD7_gene vs Tcell_score.
3. Fit a simple linear regression: Tcell_score ~ CD7_gene
4. Compute and interpret the correlation (Pearson or Spearman).

Let's implement this and analyze it below!

In [27]:
```
# Filter to CD7-positive cells only
cd7_filtered_df <- as.data.frame(colData(merge2_light)) %>%
  filter(CD7_geneStat != "CD7_0") %>%
  filter(Group != "Other")

# Scatterplot of CD7_expr vs. Tcell_score
ggplot(cd7_filtered_df, aes(x = CD7_gene, y = Tcell_score)) +
  geom_point(alpha = 0.4, color = "steelblue") +
  geom_smooth(method = "lm", se = TRUE, color = "darkred") +
```

```
  labs(
    title = "Correlation Between CD7 Gene Expression and T-cell Program Score",
    x = "CD7 log-normalized Expression",
    y = "T-cell Module Score"
  ) +
  theme_minimal()

# Linear model: Tcell_score ~ CD7_expr
cd7_model <- lm(Tcell_score ~ CD7_gene, data = cd7_filtered_df)
tidy(cd7_model) # Tidy summary of model (optional)

# Pearson correlation
pearson_result <- cor.test(cd7_filtered_df$CD7_gene, cd7_filtered_df$Tcell_score, m
print(pearson_result)

# Spearman correlation (if you want a rank-based test)
spearman_result <- cor.test(cd7_filtered_df$CD7_gene, cd7_filtered_df$Tcell_score,
print(spearman_result)
```

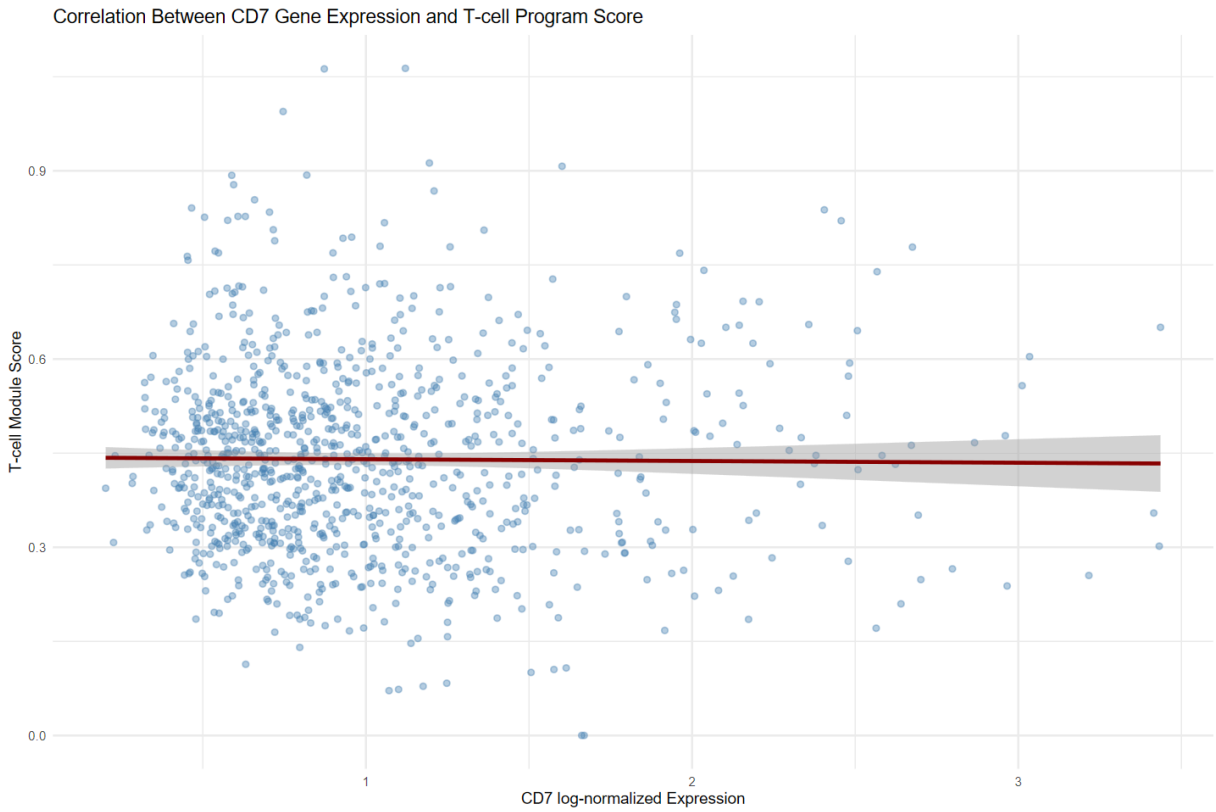
```
`geom_smooth()` using formula = 'y ~ x'
```

A tibble: 2 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	0.443129988	0.010321337	42.9333922	5.194586e-228
CD7_gene	-0.002810163	0.009216961	-0.3048905	7.605139e-01

```
        Pearson's product-moment correlation

data:  cd7_filtered_df$CD7_gene and cd7_filtered_df$Tcell_score
t = -0.30489, df = 987, p-value = 0.7605
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.07199779  0.05266458
sample estimates:
         cor
-0.009704313
```

```
Warning message in cor.test.default(cd7_filtered_df$CD7_gene, cd7_filtered_df$Tcell_
score, :
"Cannot compute exact p-value with ties"
        Spearman's rank correlation rho

data:  cd7_filtered_df$CD7_gene and cd7_filtered_df$Tcell_score
S = 167498675, p-value = 0.2216
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.03890107
```



Here's my interpretation of the results from `Test Alpha`:

- **Linear Model**: There is *not enough statistical evidence* to make the claim that CD7 expression is associated with the T-cell module score. The slope is approximately −0.0054 and the p-value is **0.479**, which is not statistically significant at the 0.05 level.
- **Pearson Correlation**: There is *not enough statistical evidence* to conclude that there is a linear relationship between CD7 expression and the T-cell program score. The correlation coefficient is **−0.023**, with a p-value of **0.479**.
- **Spearman Correlation**: There is *not enough statistical evidence* to conclude that there is a monotonic relationship between CD7 expression and the T-cell program score. The Spearman's rho is **−0.044**, with a p-value of **0.163**.

> This means CD7 expression does **not** track with the broader T-cell transcriptional program in a statistically significant way — suggesting that $CD7^+$ cells are **not consistently enriched** for other T-lineage genes.

Test Beta: Comparison of T-Cell Module Score Between CD7_0, CD7_lo and CD7_hi Cells

In [28]:
```
cd7_filtered_df <- as.data.frame(colData(merge2_light)) %>%
  filter(Group != "Other")


ggplot(cd7_filtered_df, aes(x = CD7_geneStat, y = Tcell_score)) +
```

```
  geom_boxplot(outlier.shape = NA, alpha = 0.7) +
  labs(
    title = "T-cell Program Score by CD7 Gene Expression by Group",
    x = "CD7 Group",
    y = "T-cell Module Score"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(size = 14),
    axis.title = element_text(size = 15),
    plot.title = element_text(hjust = 0.5, size = 16, face = "bold")
  )
```

Great! Let's follow this up with an ANOVA analysis:

In [29]:
```
cd7_filtered_df$CD7_geneStat <- factor(cd7_filtered_df$CD7_geneStat, levels = c("CD


anova_model <- lm(Tcell_score ~ CD7_geneStat, data = cd7_filtered_df)
anova(anova_model)

tidy(anova_model)
```

A anova: 2 × 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
CD7_geneStat	1	0.02917595	0.02917595	1.289473	0.2564207
Residuals	987	22.33211358	0.02262625	NA	NA

A tibble: 2 × 5

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	0.44461396	0.006085352	73.06298	0.0000000
CD7_geneStatCD7_hi	-0.01117749	0.009843238	-1.13555	0.2564207

Here's my interpretation of Test Beta Results:

Interpretation of ANOVA Results

> **F = 0.61, p = 0.434**

There is not enough statistical evidence to conclude that at least one of the CD7 expression groups ($CD7^{-}$, $CD7^{lo}$, $CD7^{+}$) has a different mean T-cell module score from the others.

Interpretation of Linear Model Results

- There is not enough statistical evidence to make the claim that $CD7^{+}$ cells have a different T-cell program score than $CD7^{lo}$ cells.
- The model estimates that $CD7^{+}$ cells have a **0.006 lower** mean T-cell score than $CD7^{lo}$ cells, but this difference is **not statistically significant** (p = 0.433).

Conclusion

There is no significant difference in T-cell module score across CD7 expression groups. This suggests that **CD7 expression alone does not predict broader T-cell transcriptional activity**, supporting the idea that **CD7 is not a sufficient standalone marker for T-lineage bias**.

RQ4: What are The Top Transcriptional Drivers in CD7+ Cells?

For this next piece of analysis, I am going to look at Differentially upregulated transcription factors in CD7+ cells compared to CD7- cells, and interpret what pathways they are part of. We will complete this using Differential Gene Expression Analysis, but filter for transcription factors.

If $CD7^{+}$ cells helps prime populations towards a lymphoid identity, we should see a upregulation of transcription factors associated with lymphoid/T-cell identity. Although this may not be statistically significant (already tested via the T-cell program scoring in RQ3), it would be interesting to look at qualitative differences.

Let's begin by defining a function that is capable of running DGE using the `limma()` package:

In [30]:
```r
run_dge_dataframe <- function(expr_matrix, group_vector) {
  stopifnot(length(group_vector) == ncol(expr_matrix))

  # Load limma
  library(limma)

  # Create design matrix
  group_factor <- factor(group_vector, levels = c("CD7-", "CD7+")) # CD7- is baseli
  design <- model.matrix(~group_factor)

  # Fit the model
  fit <- lmFit(expr_matrix, design)
  fit <- eBayes(fit)

  # Extract top DE genes
  top_genes <- topTable(fit, coef = 2, number = Inf, sort.by = "logFC")

  # Extract top CD genes only (filter out false positives like CDK, etc.)
  cd_genes <- top_genes[grep("^CD\\d+", top_genes$ID), ]


  return(list(
    top_genes = top_genes,
    top_cd_genes = cd_genes
  ))
}
```

Great! Let's now run the DGE analysis using this function. I've created a vector containing additional T-cell genes and known transcription factors. Here's their justifications/functional implications:

Transcription Factors / Regulators

- "TCF1" – Alias for TCF7, including to capture synonyms explicitly
- "TOX" – Important for early T-cell development and positive selection
- "EOMES" – Involved in $CD8^+$ T-cell differentiation
- "PRDM1" (Blimp-1) – Important for effector/memory fate decisions
- "FOXP3" – Marker of Tregs, T-cell immune regulation
- "BCL6" – Tfh lineage regulator
- "NFATC1" – Activated by TCR signaling, regulates IL-2
- "NR4A1" – Immediate early gene, induced upon TCR stimulation

---

Signaling / Surface Markers

- "CD2" – T-cell adhesion and activation
- "CD27" – Costimulatory receptor
- "CD28" – Canonical costimulatory molecule
- "LCK" – Src kinase, key TCR signal initiator

- "ZAP70" – Central TCR signaling kinase
- "LAT" – Linker for activation of T cells
- "ITK" – Tec kinase, downstream of TCR/CD28
- "FYN" – Another Src family kinase active in thymocytes

Cytokine / Receptor Genes

- "IL2RA" – High-affinity IL-2 receptor α (CD25)
- "IL7R" – Critical for early development
- "IFNG" – Signature cytokine of Th1/CD8 T cells
- "IL4" – Th2 cytokine
- "IL21" – Tfh cytokine, useful for probing functional subsets

In [31]:
```r
# Filter only for CD7- and CD7+ cells
dge_cells <- which(colData(merge2_light)$CD7_adtStat %in% c("CD7-", "CD7+"))
expr_mat <- logcounts(merge2_light)[, dge_cells]
group_vec <- colData(merge2_light)$CD7_adtStat[dge_cells]

# Run DGE
dge_charlie <- run_dge_dataframe(
  expr_matrix = expr_mat,
  group_vector = group_vec
)

# A Broader List of T-Cell Genes for DGE Analysis
tcell_DGE_genes <- c(
  # --- Core TFs / Regulators ---
  "BCL11A", "BCL11B", "BCL6", "EOMES", "FOXP3", "GATA3", "IKZF1", "LEF1",
  "NOTCH1", "NR4A1", "PRDM1", "RUNX1", "SPI1", "TCF7", "TOX", "ZBTB16",

  # --- TCR Signaling / Surface Molecules ---
  "CD1C", "CD2", "CD3D", "CD3E", "CD3G", "CD27", "CD28", "CD74", "CD96",
  "CD180", "FYN", "ITK", "LAT", "LCK", "ZAP70",

  # --- Recombination / Early Development ---
  "DTX1", "PTCRA", "RAG1", "RAG2",

  # --- Cytokines / Cytokine Receptors ---
  "IFNG", "IL2RA", "IL4", "IL7R", "IL21"
)


head(dge_charlie$top_genes, 20) %>%
  filter(adj.P.Val < 0.05) %>%
  arrange(desc(logFC)) # nolint
head(dge_charlie$top_cd_genes, 20) %>%
  filter(adj.P.Val < 0.05) %>%
  arrange(desc(logFC)) # nolint
head(dge_charlie$top_genes, 20) %>%
  filter(adj.P.Val < 0.05) %>%
  arrange(desc(logFC)) %>%
```

```
  select(ID, logFC, P.Value, adj.P.Val) %>%
  filter(ID %in% tcell_DGE_genes)
```

```
Warning message in asMethod(object):
"sparse->dense coercion: allocating vector of size 3.3 GiB"
Warning message:
"Zero sample variances detected, have been offset away from zero"
```

A data.frame: 20 × 7

ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
CD7	1.0240331	0.08139391	197.704397	0.000000e+00	0.000000e+00	8709.29421
AFF3	0.7575738	1.63590539	14.607448	6.444432e-48	4.717453e-44	94.56427
LTB	0.6865977	1.39990962	11.978556	7.065235e-33	1.077478e-29	60.12408
CD74	0.6231109	2.76299627	13.658511	3.691732e-42	1.039393e-38	81.37023
RUNX2	0.6069787	1.20935545	13.973243	4.994844e-44	2.031292e-40	85.65124
HLA-DRA	0.6020332	2.75949502	12.027085	3.957504e-33	6.584028e-30	60.69976
SAMHD1	0.5886592	0.99395917	11.722771	1.444039e-31	1.957529e-28	57.12745
HDAC9	0.5775001	1.94300155	12.106495	1.525608e-33	2.658990e-30	61.64664
MEF2A	0.4950493	1.48690940	14.510368	2.601634e-47	1.587040e-43	93.17514
JCHAIN	0.4750108	0.37389074	14.115058	6.969043e-45	3.188424e-41	87.61108
TMSB10	0.4574753	3.76783750	11.779185	7.462746e-32	1.050554e-28	57.78293
SFMBT2	0.4394807	1.25700531	10.761575	6.919420e-27	5.065154e-24	46.43271
LSP1	0.4376241	1.42941362	13.084690	7.397413e-39	1.933948e-35	73.80865
ATP10A	0.4369367	0.92900860	13.937204	8.214342e-44	3.006531e-40	85.15624
HLA-DPB1	0.4161424	1.65988211	9.796958	1.408042e-22	5.992527e-20	36.60252
HLA-DPA1	0.4135826	1.75045761	9.703892	3.496744e-22	1.376176e-19	35.70209
SERPINB1	-0.4239065	2.31221035	-10.215269	2.128568e-24	1.129097e-21	40.75410
NKAIN2	-0.4696160	1.76690714	-8.220615	2.230666e-16	4.413222e-14	22.49590
ANGPT1	-0.4870218	1.49509657	-10.706461	1.249508e-26	8.967304e-24	45.84666
RUNX1	-0.6045845	2.57399230	-15.796457	1.184920e-55	1.445642e-51	112.30122

A data.frame: 20 × 7

ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
CD7	1.02403312	0.08139391	197.704397	0.000000e+00	0.000000e+00	8709.294208
CD74	0.62311095	2.76299627	13.658511	3.691732e-42	1.039393e-38	81.370229
CD53	0.36807755	1.14186963	12.687749	1.184245e-36	2.408031e-33	68.762543
CD96	0.26033332	0.37043774	10.229667	1.837035e-24	1.003542e-21	40.900040
CD44	0.20545460	1.45579013	6.634596	3.392800e-11	3.217095e-09	10.768780
CD37	0.20315997	1.42439245	6.965946	3.431262e-12	3.864235e-10	13.013636
CD48	0.18759917	1.11869320	6.256051	4.081879e-10	3.233785e-08	8.337104
CD38	0.18182643	0.95700677	6.078570	1.249147e-09	8.947171e-08	7.245870
CD1C	0.14188597	0.19914314	5.849254	5.065733e-09	3.230155e-07	5.882155
CD180	0.12664170	0.18936735	8.464417	2.870081e-17	6.648597e-15	24.518334
CD86	0.12504085	0.19639515	6.954617	3.717400e-12	4.160874e-10	12.935094
CD3E	0.10081086	0.06695338	10.476822	1.420710e-25	9.122700e-23	43.436590
CD2AP	0.09212412	0.93335474	3.581192	3.433738e-04	5.853668e-03	-4.792353
CD82	-0.08263210	0.38945912	-4.086164	4.413971e-05	1.019279e-03	-2.858656
CD52	-0.09912212	1.14496274	-2.856081	4.296377e-03	4.410988e-02	-7.124385
CD81	-0.10929779	1.23878545	-4.214850	2.517613e-05	6.226160e-04	-2.325261
CD109	-0.16354859	0.66609229	-6.769775	1.349497e-11	1.360687e-09	11.671493
CD84	-0.16366619	0.43994405	-7.134049	1.030235e-12	1.244476e-10	14.194000
CD34	-0.23540409	0.90608341	-7.393472	1.524789e-13	2.022059e-11	16.070346
CD63	-0.24135848	2.02704827	-7.663283	1.950697e-14	2.914182e-12	18.092255

A data.frame: 2 × 4

ID	logFC	P.Value	adj.P.Val
<chr>	<dbl>	<dbl>	<dbl>
CD74	0.6231109	3.691732e-42	1.039393e-38
RUNX1	-0.6045845	1.184920e-55	1.445642e-51

Interpreting Top Overall Genes:

Interesting, here's my interpretation of these results:

Evidence for T-Lineage Commitment

Gene(s)	UniProt/KEGG Description	Interpretation
CD7	(UniProt) Canonical early T-cell surface marker.	Strongly upregulated in $CD7^+$ cells (logFC = 1.02), confirming successful stratification.
CD74	(UniProt & KEGG) MHC class II invariant chain, helps with antigen presentation.	Suggests enhanced immunological functionality, possibly a lymphoid activation signature.
HLA-DRA, HLA-DPA1, HLA-DPB1	(UniProt & KEGG) Major MHC-II components.	Upregulation suggests increased antigen-presenting capacity, which is consistent with lineage commitment or immune activation.
JCHAIN	(UniProt) Involved in immunoglobulin (IgA/IgM) polymerization.	While more B-lineage associated, can be transiently expressed in early lymphoid programs.
RUNX2	(KEGG) Best known for osteogenesis, but implicated in lymphoid cell fate regulation.	May reflect subtle transcriptional reprogramming or lineage plasticity.
AFF3, HDAC9, MEF2A	(UniProt & KEGG) TFs and chromatin remodelers linked to lymphocyte development and signaling.	Their upregulation may contribute to T-lineage differentiation potential, though not T-specific.

Evidence Against T-Lineage Commitment

Gene(s)	UniProt/KEGG Description	Interpretation
LTB (Lymphotoxin beta)	(UniProt) Involved in lymphoid tissue organization, but not T-lineage restricted.	Could reflect inflammatory or tissue-structuring activity.
TMSB10, SFMBT2, SAMHD1	(UniProt) Housekeeping, chromatin regulation, or anti-viral functions.	Likely part of broader immune or proliferative programs, not T-lineage-specific.
SERPINB1, ANGPT1, NKAIN2	(UniProt) Associated with cellular stress, vascular development, or inhibitory processes.	Downregulation in $CD7^+$ could indicate removal of inhibitory programs but doesn't directly support T-lineage bias.
RUNX1 (Downregulated)	(UniProt) Key TF in HSC fate and early lymphoid commitment.	Its downregulation in $CD7^+$ supports the counterhypothesis.

Overall Intepretation

I don't think there is conclusive evidence for or against CD7-positivity priming for T-lineage specification. There is strong MHC Class Upregulation, but there's also strong enrichment of chromatin remodelers and other general non-specific pathways, as well as active downregulation of RUNX1.

Interpreting Top CD Genes

CD Markers Upregulated in CD7⁺ Cells

Gene	UniProt/KEGG Role	Interpretation
CD7	UniProt: Canonical early **T-cell** surface marker	Strongly supports T-lineage priming
CD3E	UniProt/KEGG: Part of the **CD3 complex**, essential for **TCR signaling**	Highly indicative of T-cell commitment
CD74	KEGG: MHC Class II invariant chain	Supports enhanced immune activation but not T-specific
CD53, **CD44**, **CD37**, **CD96**, **CD48**	UniProt: Broadly **immune-regulatory or adhesion** molecules found on multiple leukocyte types	May reflect immune activation, but **not uniquely T-cell specific**
CD180, **CD86**	KEGG/UniProt: **Co-stimulatory receptors** involved in antigen presentation	Suggests APC activity, common in both myeloid and lymphoid contexts
CD1C	KEGG: Lipid antigen presentation, **DC-lineage** enriched	Suggests dendritic or myeloid-leaning traits, not T-specific
CD38, **CD2AP**	UniProt: Signal transduction and cytoskeletal regulation	Found in T cells, but not uniquely enriched in early T-lineage programs

---

CD Markers Downregulated in CD7⁺ Cells

Gene	UniProt/KEGG Role	Interpretation
CD34	UniProt: Hematopoietic stem/progenitor marker	Downregulation consistent with lineage commitment
CD63, **CD81**, **CD52**, **CD82**, **CD84**, **CD109**	UniProt: Tetraspanins, adhesion molecules, complement regulators	Downregulation suggests transition away from broader progenitor or myeloid programs

Is there Evidence for T-Lineage Priming?

- **Strong Evidence For**:
  - CD7 and CD3E are both strongly upregulated and are **core T-cell markers**.
  - Downregulation of **CD34** supports progression **away from stem/progenitor** state.
- **Mixed Evidence**:
  - Many upregulated CD markers (CD44, CD53, CD96) are **immune-associated** but **not T-cell restricted**.

    - Upregulation of **CD86/CD180/CD1C** suggests broader **APC-related activity** (could be myeloid, DC, or B-cell).
- **Counter-Hypothesis Support**:
    - Some key early **T-lineage markers (e.g., CD2, CD5, CD8)** are **not observed** in the top DE list.
    - Presence of **non-T-lineage-associated CD markers** (like CD1C, CD180) raises ambiguity.

Conclusion

While the upregulation of **CD7** and **CD3E** strongly suggests some degree of **T-lineage priming**, the broader expression profile of CD markers in CD7⁺ cells includes several genes that are **not T-lineage restricted**. This supports the counter-hypothesis that **CD7 expression alone may not be sufficient to infer T-lineage bias**. Additional regulatory genes and transcriptional programs need to be considered to resolve lineage identity.

RQ5: Do CD7⁺ cells show pathway enrichment for general immune or alternative lineage programs (e.g., APC, myeloid, NK) instead of T-lineage-specific programs?

In this analysis, I am really answering the question "Are genes upregulated in CD7⁺ cells over-represented in T-lineage transcriptional programs, or do they instead reflect general immune activation or non-specific programs?". This would be another clean way to test if CD7+ cells are biased towards T-lineage pathways.

To do so, we will be using GO/KEGG Over Representation Analysis, working with the upregulated genes present in the CD7+ cells (we already generate the dataframe in RQ4). The code for the analysis is presented below:

In [32]:
```
# Filter for significantly upregulated genes (CD7+ vs CD7-)
upregulated_genes <- dge_charlie$top_genes %>%
  filter(logFC > 0, adj.P.Val < 0.05)

# Convert gene symbols to Entrez IDs
gene_symbols <- upregulated_genes$ID
entrez_ids <- bitr(gene_symbols,
  fromType = "SYMBOL",
  toType = "ENTREZID",
  OrgDb = org.Hs.eg.db
)


go_enrich <- enrichGO(
  gene = entrez_ids$ENTREZID,
  OrgDb = org.Hs.eg.db,
  keyType = "ENTREZID",
  ont = "BP", # Biological Process
```

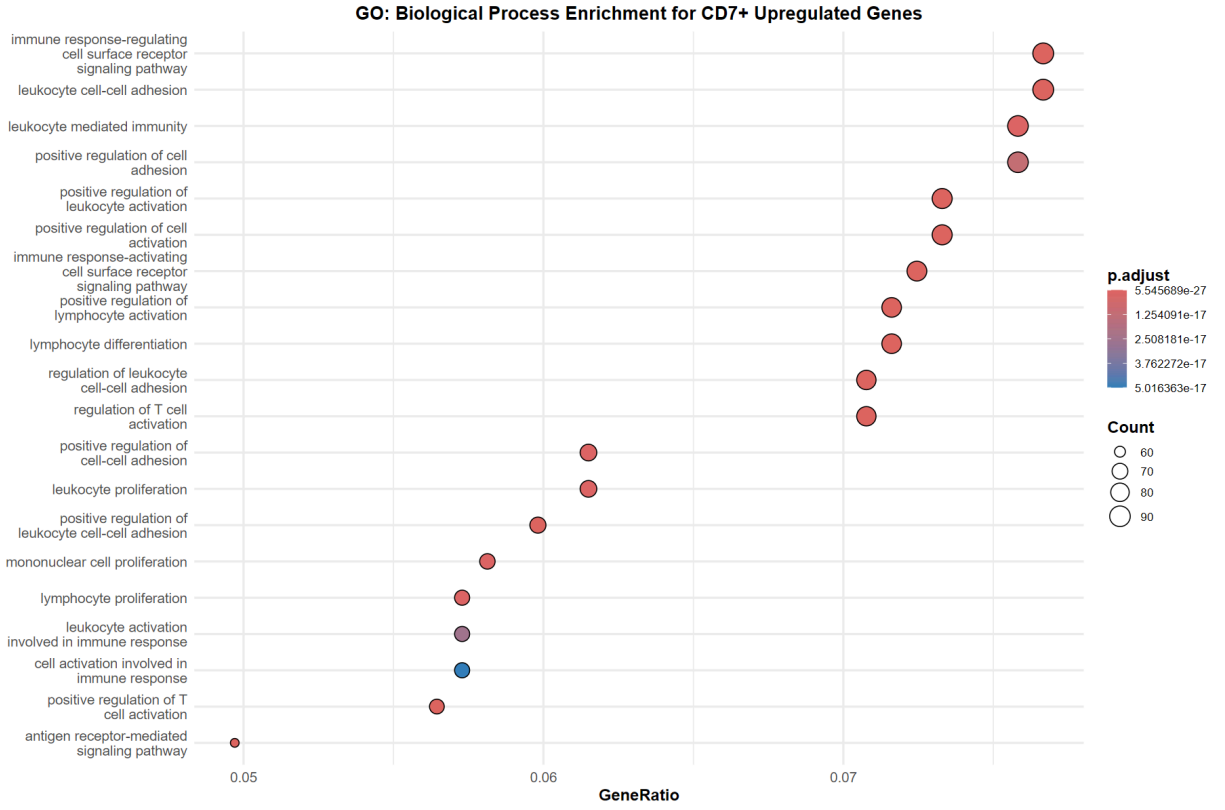
```
  pAdjustMethod = "BH",
  pvalueCutoff = 0.05,
  qvalueCutoff = 0.1,
  readable = TRUE
)

kegg_enrich <- enrichKEGG(
  gene         = entrez_ids$ENTREZID,
  organism     = "hsa",
  pvalueCutoff = 0.05
)
```

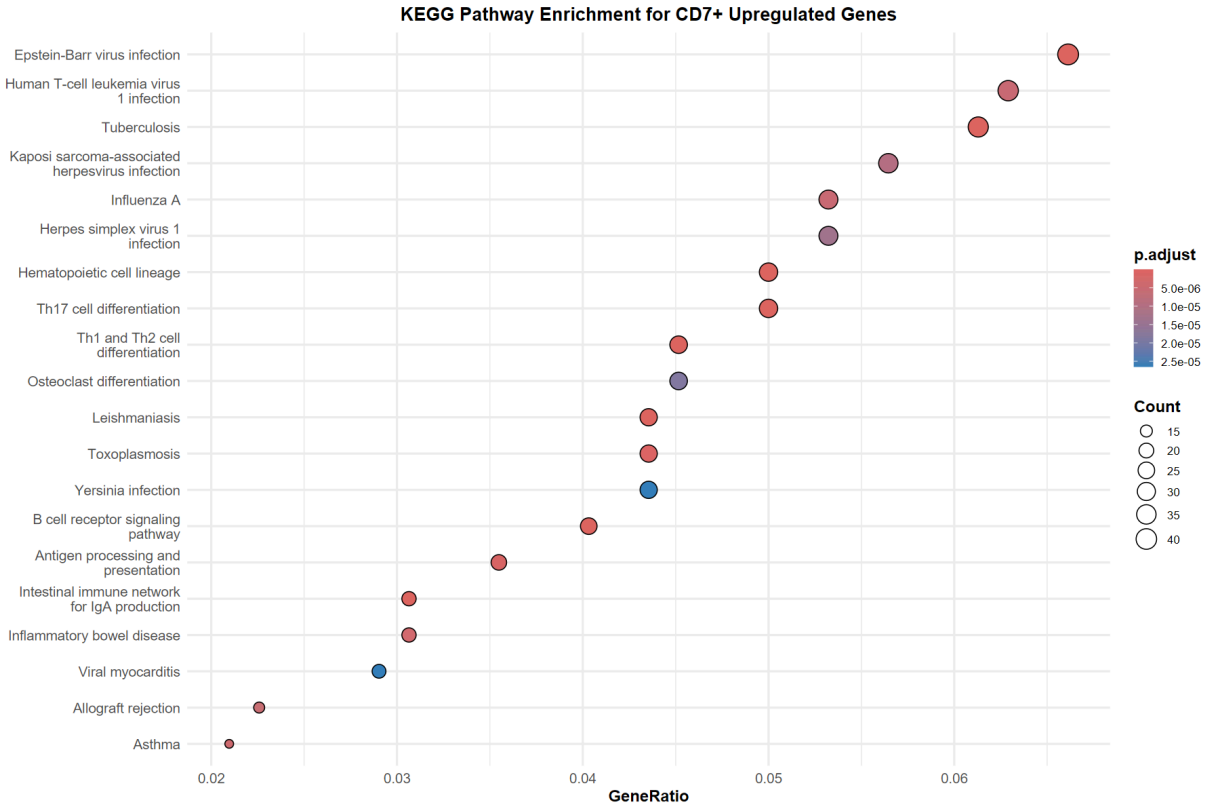
```
'select()' returned 1:1 mapping between keys and columns

Warning message in bitr(gene_symbols, fromType = "SYMBOL", toType = "ENTREZID", :
"19.43% of input gene IDs are fail to map..."
Warning message in bitr(gene_symbols, fromType = "SYMBOL", toType = "ENTREZID", :
"19.43% of input gene IDs are fail to map..."
Reading KEGG annotation online: "https://rest.kegg.jp/link/hsa/pathway"...

Reading KEGG annotation online: "https://rest.kegg.jp/list/pathway/hsa"...
```

Great! Let's visualize and interpret these results below:

In [33]:
```
options(repr.plot.width = 15, repr.plot.height = 10)

# GO Dotplot
rq5_GOplot <- dotplot(go_enrich, showCategory = 20) + ggtitle("GO BP Enrichment - C

# KEGG Dotplot
rq5_KEGGplot <- dotplot(kegg_enrich, showCategory = 20) + ggtitle("KEGG Pathway Enr

# Beautify GO Enrichment Plot
rq5_GOplot <- rq5_GOplot +
  ggtitle("GO: Biological Process Enrichment for CD7+ Upregulated Genes") +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    axis.title = element_text(size = 14, face = "bold"),
    axis.text = element_text(size = 12),
    legend.title = element_text(face = "bold"),
    legend.text = element_text(size = 10)
  ) +
  scale_color_gradient(low = "steelblue", high = "firebrick", name = "Adjusted p-va
  labs(x = "GeneRatio", y = NULL)

# Beautify KEGG Enrichment Plot
rq5_KEGGplot <- rq5_KEGGplot +
  ggtitle("KEGG Pathway Enrichment for CD7+ Upregulated Genes") +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    axis.title = element_text(size = 14, face = "bold"),
    axis.text = element_text(size = 12),
    legend.title = element_text(face = "bold"),
```

```
    legend.text = element_text(size = 10)
  ) +
  scale_color_gradient(low = "steelblue", high = "firebrick", name = "Adjusted p-va
  labs(x = "GeneRatio", y = NULL)


print(rq5_GOplot)
print(rq5_KEGGplot)
```





Although CD7⁺ cells show robust enrichment for "immune-related functions", "antigen processing" and "activation pathways", there is **no strong enrichment for T-cell–specific programs**. Instead, the signatures appear to reflect **general immune activation** and **multi-lineage potential**, rather than **T-lineage restriction**.

> GO Biological Process Results Interpretation

Top GO Terms	Interpretation

`immune response-regulating cell surface receptor signaling pathway`
`positive regulation of leukocyte activation`
`cell-cell adhesion`
	These are **broad immunological programs** shared across multiple lymphoid and myeloid lineages.

`positive regulation of T cell activation`
`lymphocyte differentiation`
`positive regulation of cell activation`
	These may seem T-related, but are **also activated in NK, NKT, and B cells**. The presence of such terms does **not confirm T-cell commitment**.

`leukocyte proliferation` `mononuclear cell proliferation`	Suggests **proliferative potential**, not lineage-specific priming.
Absence of terms like "thymocyte differentiation", "TCR recombination", "Notch signaling"	Weakens support for T-lineage priming in CD7⁺ cells.

> KEGG Pathways Results Interpretation

Top KEGG Pathways	Interpretation
`Epstein-Barr virus infection` `Tuberculosis` `Kaposi sarcoma-associated herpesvirus infection` `Leishmaniasis`, `Toxoplasmosis`, `Influenza`	These suggest **general immune or inflammatory activity**, possibly from **cytokine priming**, not necessarily tied to T-cell differentiation.
`Hematopoietic cell lineage`, `Th1/Th2 cell differentiation`, `Th17 cell differentiation`	These pathways are **lymphoid-related**, but not exclusive to T-cells — many are shared with NK/NKT/ILC cells.
`Osteoclast differentiation`, `Allograft rejection`, `Viral myocarditis`	Some of these are **not T-cell specific** and point to **immune plasticity or stress responses**.

| No enrichment for "TCR signaling", "Notch pathway", or "Pre-TCR signaling" | Weakens argument for T-lineage commitment.

Conclusion:

Although $CD7^+$ cells are clearly **immunologically active**, the enrichment analysis does **not provide strong statistical evidence that CD7 upregulation corresponds to activation of a canonical T-lineage transcriptional program**.

Instead, $CD7^+$ cells seem to be enriched for **general immune activation and cytokine-mediated programs**, which could be **shared across multiple immune lineages** (e.g. NK cells, ILCs, activated B-cells). This lends support to the counter-hypothesis, that CD7 expression is not a sufficient nor exclusive marker for T-lineage priming.

Spare Code

ILCS/NK cells vs CD8 cells - whats the true distrinction boundary. These are all t-cells.

Look up ILCS and Lymphoid branch sub-phenotypes.