Aneuploidy Assessment: EMTAB3929

Margaret R. Starostik

1. Human Embryo EMTAB3929 scRNA-seq analysis

1.1 Data Summary

```
# (1) Generate histogram of the number of reads that mapped
# to reference genome GRCh38.84.
emtab3929_meta@metadata$salmon_summary$Stage <- sapply(strsplit(emtab3929_meta@metadata$salmon_summary$sample,
  "\\."), `[`, 1)

mapping_plot01 <- ggplot(emtab3929_meta@metadata$salmon_summary,
  aes(x = sample, y = num_mapped, fill = Stage)) + geom_bar(stat = "identity")
mapping_plot02 <- mapping_plot01 + labs(title = "", x = "Samples",
  y = "Mapped Reads") + theme_bw() + theme(axis.text.x = element_blank(),
  axis.ticks.x = element_blank(), plot.title = element_text(hjust = 0.5))
ggsave(paste0(output_folder, "human_EMTAB3929_scRNAseq_ReadsMapped.pdf"),
  plot = mapping_plot02, device = "pdf", width = 6, height = 4,
  units = "in")
return(mapping_plot02)

# (2) Generate table with summarized read mapping information
emtab3929_summary <- data.frame(Stage = c("E3", "E4", "E5", "E6",
  "E7", "Total"), Cells = c(81, 190, 377, 415, 466, 1529),
  MeanReadsMapped = c(4736734, 5587680, 5905343, 4636041, 6331946,
    5589466), MedianReadsMapped = c(4822716, 5492478, 4282245,
    3997711, 5883858, 4912405))

kable(emtab3929_summary, caption = "Data Summary", align = rep("c",
  6))
```

Table 1: Data Summary

Stage	Cells	MeanReadsMapped	MedianReadsMapped
E3	81	4736734	4822716
E4	190	5587680	5492478
E5	377	5905343	4282245
E6	415	4636041	3997711
E7	466	6331946	5883858
Total	1529	5589466	4912405

1.1 Identifying data substructure

PCA on $\log_{10}$(counts + 1) was performed only on genes used in assessing ploidy (i.e. genes with median CPM > 50).

```
# Convert counts into CPM
emtab3929_cpm <- cpm(emtab3929_meta@ExperimentList@listData$gene@assays$data$count,
```

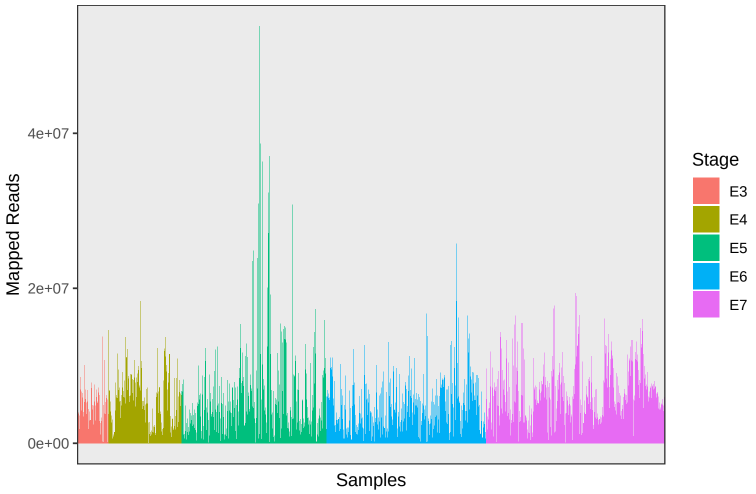

Figure 1: Reads Mapped to Reference Genome

```
    normalized.lib.sizes = TRUE, log = FALSE)
# dim(emtab3929_cpm) # 65,218 genes and 1,529 samples

# Apply gene filter and generate PCA plots
gene_filter <- apply(emtab3929_cpm, 1, median) > 50
filtered_cpm <- emtab3929_cpm[gene_filter, ]
# dim(filtered_cpm) # 2,623 genes and 1,529 samples
pca <- prcomp(t(log10(filtered_cpm + 1)))
variance <- pca$sdev^2

pca_data <- as.data.frame(pca$x[, 1:3])
pca_data$Stage <- emtab3929_meta@metadata$salmon_summary$Stage

# PCA on all data
pca_plot01 <- ggplot(pca_data, aes(x = PC1, y = PC2, col = Stage)) +
    geom_point()
pca_plot02 <- pca_plot01 + labs(x = paste0("PC1 ", format(variance[1]/sum(variance) *
    100, digits = 3), "%"), y = paste0("PC2 ", format(variance[2]/sum(variance) *
    100, digits = 3), "%"))
pca_plot03 <- pca_plot02 + theme_bw() + scale_color_brewer(palette = "Set1")
ggsave(paste0(output_folder, "human_EMTAB3929_scRNAseq_PCA01.pdf"),
    plot = pca_plot03, device = "pdf", width = 6, height = 4,
    units = "in")
return(pca_plot03)

pca_plot04 <- ggplot(pca_data, aes(x = PC1, y = PC3, col = Stage)) +
    geom_point()
pca_plot05 <- pca_plot04 + labs(x = paste0("PC1 ", format(variance[1]/sum(variance) *
    100, digits = 3), "%"), y = paste0("PC3 ", format(variance[3]/sum(variance) *
    100, digits = 3), "%"))
pca_plot06 <- pca_plot05 + theme_bw() + scale_color_brewer(palette = "Set1")
ggsave(paste0(output_folder, "human_EMTAB3929_scRNAseq_PCA02.pdf"),
    plot = pca_plot06, device = "pdf", width = 6, height = 4,
    units = "in")
```

```r
return(pca_plot06)

stages <- c("E3", "E4", "E5", "E6", "E7")
start_sample <- c(1, 82, 272, 649, 1064)
end_sample <- c(81, 271, 648, 1063, 1529)
for (i in 1:length(stages)) {
  stage_pca <- t(log10(filtered_cpm + 1))
  stage_pca <- stage_pca[start_sample[i]:end_sample[i], ]
  stage_pca <- prcomp(stage_pca)

  variance <- stage_pca$sdev^2

  stage_pca_data <- as.data.frame(stage_pca$x[, 1:2])

  pca_plot10 <- ggplot(stage_pca_data, aes(x = PC1, y = PC2)) +
    geom_point(aes(fill = "steel blue"))
  pca_plot11 <- pca_plot10 + labs(subtitle = paste0("Stage: ",
    stages[i]), x = paste0("PC1 ", format(variance[1]/sum(variance) *
    100, digits = 3), "%"), y = paste0("PC2 ", format(variance[2]/sum(variance) *
    100, digits = 3), "%"))
  pca_plot12 <- pca_plot11 + theme_bw() + guides(fill = FALSE)
  ggsave(paste0(output_folder, paste0("human_EMTAB3929_scRNAseq_PCA_",
    stages[i], ".pdf")), plot = pca_plot12, device = "pdf",
    width = 6, height = 4, units = "in")
  print(pca_plot12)
}
```

After applying a gene filter of CPM > 50, PCA was performed on 2623 genes. This identified two clusters on PC1 (18.5%) that split by embryonic stage (Figure 1).

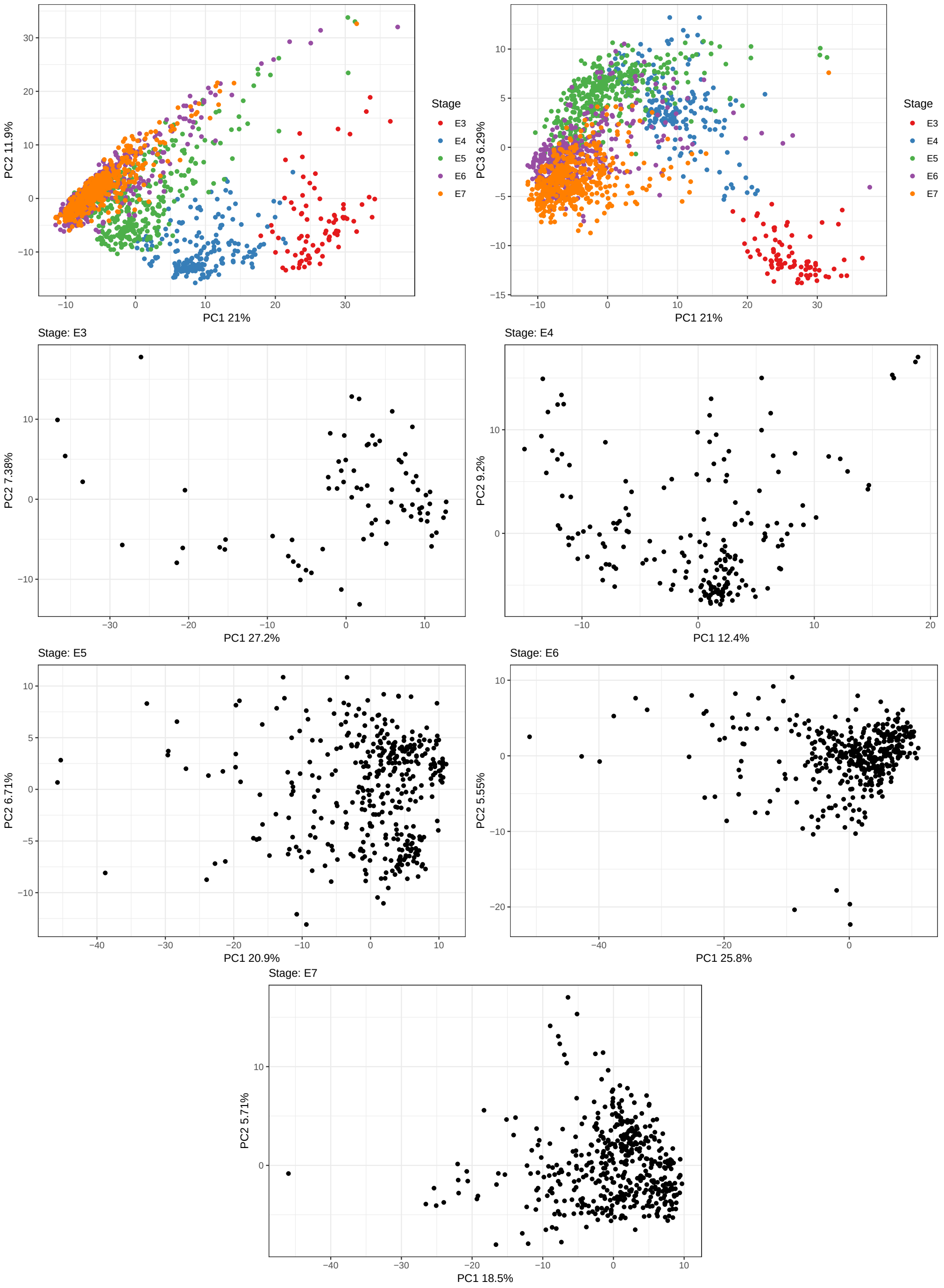

Figure 2: PCA on 2,363 genes