EMTAB3929: Data Preparation

Margaret R. Starostik

Data were acquired for the human embryo (EMTAB3929) scRNA-seq aneuploidy project as follows:
(1) EMTAB3929 data (PMID 27062923) were downloaded on 11/6/2018 from http://imlspenticton.uzh.ch:3838/conquer/.
Data included:
(a) MultiAssay Experiment (EMTAB3929.rds)
(b) MultiQC report
(c) Scater report
(d) Salmon archive (EMTAB3929_salmo.tar and EMTAB3929 folder)

(2) Supplementary files from Griffiths et al., 2017 were forked on 11/06/2018 from MarioniLab/Aneuploidy2017 on Github: https://github.com/MarioniLab/Aneuploidy2017.

(3) Data from Griffiths et al., 2017 were downloaded on 11/14/2018 using the shell script supplied in their supplementary files (sh get_data.sh)

```
# Set up folders
project_folder <- "/Users/Margaret/Desktop/JHU/COURSEWORK/2018_Fall/Rotation02_RajivMcCoy/AneuploidyProject/"

# Load EMTAB3929 data
emtab3929_meta <- readRDS(paste0(project_folder, "RawData/emtab3929/EMTAB3929.rds"))
```

Data Preparation

```
# (1) Remove version numbers from Ensembl gene IDs
rownames(emtab3929_meta@ExperimentList@listData$gene@assays$data$count) <- sapply(strsplit(rownames(emtab3929_meta@Experimen
#dim(emtab3929_meta@ExperimentList@listData$gene@assays$data$count) # 65218  1529

# (2) Obtain annotation for reference genome GRCh38.84 (which is GRCh38.p5 Ensembl 84:Mar 2016); keep autosomal genes only
human_ensembl <- useMart(biomart = "ENSEMBL_MART_ENSEMBL",
                         host = "mar2016.archive.ensembl.org",
                         path = "/biomart/martservice",
                         dataset = "hsapiens_gene_ensembl")

annotation <- getBM(
  attributes = c("ensembl_gene_id", "external_gene_name", "chromosome_name", "gene_biotype"),
  mart = human_ensembl,
  values = as.character(rownames(emtab3929_meta@ExperimentList@listData$gene@assays$data$count)),
  filters = "ensembl_gene_id"
)
annotation <- annotation [annotation $chromosome_name %in% 1:22, ] # 56400 genes

# (3) Cell lineage information. Compile sample metasheet.
cell_lineage_data <- read_xlsx(paste0(project_folder, "stirparo2018_tableS4.xlsx"), sheet = 1)
cell_lineage_data <- cell_lineage_data[cell_lineage_data$Study == "Petropoulos et al., 2016 (ERP012552)", ] # 1,481 cells
cell_lineage_data$Cell <- gsub("_", ".", cell_lineage_data$Cell)
cell_lineage_data$EStage <- sapply(strsplit(cell_lineage_data$Embryo, "_"), "[", 1)
```

```
metasheet <- cell_lineage_data[, c(2:6, 8:10)]
colnames(metasheet)[2] <- "Sample"
metasheet$Cell <- sapply(strsplit(metasheet$Sample, "\\."), tail, n = 1)
rownames(metasheet) <- metasheet$Sample
metasheet <- metasheet[, c(2, 8, 3, 1, 9, 4:7)]

salmon_summary <- emtab3929_meta@metadata$salmon_summary[, c(1, 6:8)]
salmon_summary$sample <- gsub("_", ".", salmon_summary$sample)

metasheet <- inner_join(metasheet, salmon_summary, by = c("Sample" = "sample"))
colnames(metasheet)[colnames(metasheet) == "num_processed"] <- "Processed Reads"
colnames(metasheet)[colnames(metasheet) == "num_mapped"] <- "Mapped Reads"
colnames(metasheet)[colnames(metasheet) == "percent_mapped"] <- "Percent Mapped"
metasheet$`Revised lineage (this study)` <- gsub("epiblast", "Epiblast", metasheet$`Revised lineage (this study)`)
metasheet$`Revised lineage (this study)` <- gsub("Inner cell mass", "ICM", metasheet$`Revised lineage (this study)`)
metasheet$`Revised lineage (this study)` <- gsub("intermediate", "Intermediate", metasheet$`Revised lineage (this study)`)
metasheet$`Revised lineage (this study)` <- gsub("primitive_endoderm", "Primitive Endoderm", metasheet$`Revised lineage (this study)`)
metasheet$`Revised lineage (this study)` <- gsub("trophectoderm", "Trophectoderm", metasheet$`Revised lineage (this study)`)
metasheet$`Revised lineage (this study)` <- gsub("undefined", "Undefined", metasheet$`Revised lineage (this study)`)

# (4) Modify EMTAB gene expression matrix to contain only information for genes in `annotation` and samples with cell lineage information, and ther
emtab3929_counts <- emtab3929_meta@ExperimentList@listData$gene@assays$data$count[annotation$ensembl_gene_id, ]
colnames(emtab3929_counts) <- gsub("_", ".", colnames(emtab3929_counts))
emtab3929_counts <- emtab3929_counts[, colnames(emtab3929_counts) %in% metasheet$Sample]

emtab3929_cpm <- edgeR::cpm(emtab3929_counts, normalized.lib.sizes = TRUE, log = FALSE)
emtab3929_cpm <- emtab3929_cpm[, colnames(emtab3929_cpm) %in% metasheet$Sample]
#dim(emtab3929_cpm) # 56,400 genes and 1,481 cells
emtab3929_log2cpm <- log2(emtab3929_cpm + 1)

# Apply gene filter used in Griffiths analysis.
gene_filter <- apply(emtab3929_cpm, 1, median) > 50
filtered_cpm <- emtab3929_cpm[gene_filter, ]
#dim(filtered_cpm) # 2,991 genes and 1,481 cells
filtered_log2cpm <- log2(filtered_cpm + 1)

filtered_counts <- emtab3929_counts[rownames(emtab3929_counts) %in% rownames(filtered_cpm),
                                    colnames(emtab3929_counts) %in% colnames(filtered_cpm)]
#dim(filtered_counts) # 2,991 genes and 1,481 cells
```

```
# Save objects for easy uploading in the future.
save(metasheet, emtab3929_counts, filtered_counts, emtab3929_cpm, filtered_cpm, filtered_log2cpm, annotation,
    file = paste0(project_folder, "ProcessedData/EMTAB3929_DataPrep.RData"))
```

To keep all analyses consistent, the following modifications to the original EMTAB3929 data were made:
(1) Removed version numbers from EMTAB3929 Ensembl IDs
(2) Included GRCh38.p5 Ensembl 84:Mar2016 (same as GRCh38.84) reference gene annotation. Kept only information for autosomal genes
(3) Included cell lineage information from Stirparo et al., 2018 (Table S4)
(4) Applied gene filter of median CPM > 50 used in Griffiths analysis

The final data used in downstream analyses contain 2,991 genes and 1,481 cells from a total of 88 embryos.