Griffiths et al., Aneuploidy 2017: Replication of Mouse 8-cell Stage G&T-seq and Human Trisomy 21 Neuron G&T-seq Analyses

Margaret R. Starostik

Goal: Replicate results from Griffiths et al., 2017 using their data for (1) mouse 8-cell stage G&T-seq and (2) human Trisomy 21 neuron G&T-seq. The same analytical methods were used as described at https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-017-4253-x/MediaObjects/12864_2017_4253_MOESM2_ESM.html#analysis-on-mouse-8-cell-stage-cells-gt-seq.

1. Mouse 8-cell G&T-seq analysis

Table 1: Data Summary

Embryo	Cells	Treatment
A	7	Reversine
B	8	Reversine
C	7	Reversine
D	8	Reversine
E	8	Reversine
G	8	Control
H	8	Control

This data set contains a total of 54 samples that passed quality control metrics (Table 1) with a total of 32330 genes.

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). Results from the scploid package plotPCA function were confirmed.

```r
# (1) scploid package plotPCA
scploid_pca <- plotPCA(emb8, cols = emb8_meta$embryo)  # 3,414 genes and 54 samples

# (2) Validation
gene_filter <- apply(emb8@cpm, 1, median) > 50
filtered_cpm <- emb8@cpm[gene_filter, ]
# dim(filtered_cpm) # 3,414 genes and 54 samples
pca <- prcomp(t(log10(filtered_cpm + 1)))
variance <- pca$sdev^2

pca_data <- as.data.frame(pca$x[, 1:2])
pca_data$Treatment <- emb8_meta$treatment
pca_data$Embryo <- emb8_meta$embryo

pca_plot01 <- ggplot(pca_data, aes(x = PC1, y = PC2, col = Embryo,
  shape = Treatment)) + geom_point()
pca_plot02 <- pca_plot01 + labs(x = paste0("PC1 ", format(variance[1]/sum(variance) *
```

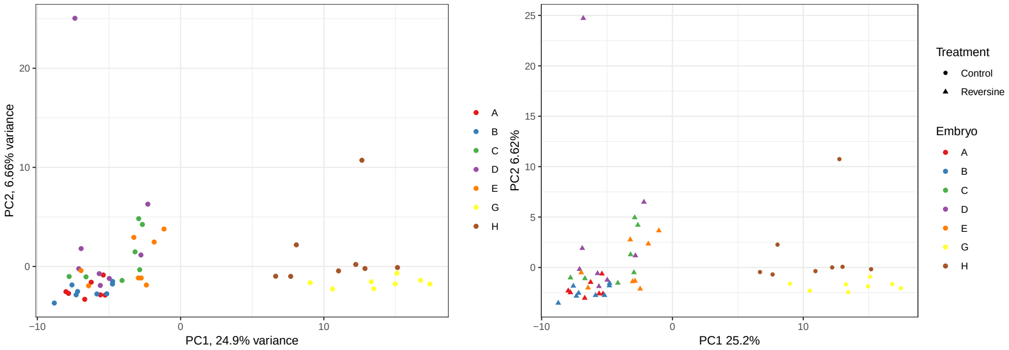

Figure 1: PCA on 3,414 genes (left: scploid; right: validation)

```
    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, "mouse_8cell_GTseq_PCA.pdf"), plot = pca_plot03,
    device = "pdf", width = 6, height = 4, units = "in")
return(pca_plot03)
```

After applying a gene filter of CPM > 50, PCA was performed on 3414 genes. This identified two clusters on PC1 (25.2%) that split by reversine treatment status (Figure 1). Due to this data substructure, subsequent analyses were performed according to reversine treatment status to identify differences based on aneuploidy rather than reversine treatment status.

1.2 Detecting aneuploidy

```
# Aneuploidy assessment. Confirm results from scploid
# package.

# (1)
emb8 <- doAneu(emb8)
mouse_hits <- getHits(emb8)
mouse_false_pos <- getFP(emb8)

mouse_test <- testPerformance(emb8)
kable(as.data.frame(t(format(mouse_test, digits = 3))), col.names = c("Sensitivity",
    "Precision", "FDR", "Specificity", "Accuracy", "F1", "FPR"),
    caption = "Mouse 8-cell G&T-seq Method Performance", align = rep("c",
        7))
```

Table 2: Mouse 8-cell G&T-seq Method Performance

Sensitivity	Precision	FDR	Specificity	Accuracy	F1	FPR
0.78000	0.88636	0.11364	0.99488	0.98441	0.82979	0.00512

1.3 Testing Model Assumptions

1.3.1 Contribution of true aneuploidy status to aneuploidy scores

If this method of aneuploidy detection works well, then a large fraction of the variance in aneuploidy scores can be explained by the true presence of an aneuploidy.

```
# Calculate Z-scores
predictions <- getScores(emb8)  # make sure to run doAneu before running this

# Denote which chromosomes are truly aneuploid
predictions$truth = FALSE
predictions$truth[which(paste0(predictions$cell, predictions$chr) %in%
   paste0(getKnownAneu(emb8)$cell, getKnownAneu(emb8)$chr))] = TRUE
# table(predictions$truth) # FALSE = 976 and TRUE = 50

# Calculate fraction of Z-score variance that is explained by
# true presence of an aneuploidy
z_model <- lm(abs(predictions$z) ~ predictions$truth)
# summary(z_model)$r.squared # 0.5434608
```

The model used to detect ploidy explains over half of the absolute Z-score variance (R^2 = 0.543).

1.3.2. Investigation of off-chromosome aneuploidy effects

Aneuploid chromosomes have abnormal gene expression levels that may disregulate gene expression on other chromosomes of the same cell, thereby counfounding the prediction of true ploidy status since these off-chromosome effects will lead to overprediction of the actual degree of aneupoloidy. To check for this issue, the variance of aneuploidy scores is compared across known normal-ploidy chromosomes in cells with aneuploidies against the variance of aneuploidy scores in cells without aneuploidies.

```
# Exclude truly aneuploid chromosomes
scrubbed_ratios <- getSijMat(emb8)
# dim(scrubbed_ratios) # 19 and 54

known <- emb8@knownAneu

for (row in 1:nrow(known)) {
   scrubbed_ratios[as.character(known$chr[row]), as.character(known$cell[row])] = NA

}

# Calculate cell-wise variance for cells that either contain
# an aneuploidy or do not contain an aneuploidy
cvar <- apply(scrubbed_ratios, 2, var, na.rm = TRUE)
aneu_var <- cvar[names(cvar) %in% known$cell]  # 19
non_aneu_var <- cvar[!names(cvar) %in% known$cell]  # 35

ks_test <- ks.test(aneu_var, non_aneu_var)  # p-value 0.2061
wilcox_test <- wilcox.test(aneu_var, non_aneu_var)  # p-value 0.12

# Plot results
var_compiled <- data.frame(aneu_status = c(rep("aneu_var", length(aneu_var)),
   rep("non_aneu_var", length(non_aneu_var))), variance = c(aneu_var,
   non_aneu_var))
```

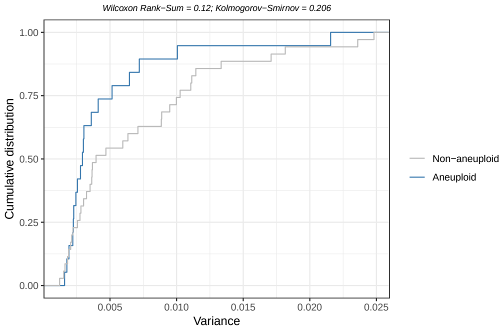

Figure 2: Cumulative distribution of ploidy variance

```
ecdf_plot01 <- ggplot(var_compiled, aes(variance)) + stat_ecdf(geom = "step",
   aes(colour = aneu_status))
ecdf_plot02 <- ecdf_plot01 + labs(title = "", subtitle = paste0("Wilcoxon Rank-Sum = ",
   round(wilcox_test$p.value, digits = 3), "; Kolmogorov-Smirnov = ",
   round(ks_test$p.value, digits = 3)), x = "Variance", y = "Cumulative distribution") +
   scale_colour_manual(name = (""), values = c("steel blue",
      "grey"), labels = c("Non-aneuploid", "Aneuploid"), breaks = c("non_aneu_var",
      "aneu_var"))
ecdf_plot03 <- ecdf_plot02 + theme_bw() + theme(plot.title = element_text(hjust = 0.5),
   plot.subtitle = element_text(size = 8, hjust = 0.5, face = "italic"))
ggsave(paste0(output_folder, "mouse_8cell_GTseq_ECDF.pdf"), plot = ecdf_plot03,
   device = "pdf", width = 6, height = 4, units = "in")
return(ecdf_plot03)
```

1.3.3. Testing for properly distributed z-scores

Chromosome scores are assumed to have a normal distribution with median absolute deviation-estimated variance, giving a set of Z-scores with t-distribution. To check for this, (1) the distribution of normal-ploidy z-scores is compared to the normal distribution in a Q-Q plot and (2) the distribution of p-values is examined in a histogram.

```
# (1) Compare distribution of normal-ploidy z-scores to the
# normal distribution in a Q-Q plot Exclude the known 50
# aneuploid data
normal_cells <- predictions[!predictions$truth, ]  # 976

## Perform Kolmogorov-Smirnov test and plot results (tie
## warning appears)
kolsmir_test <- ks.test(normal_cells$z, pt, df = length(emb8_meta$treatment[emb8_meta$treatment ==
   "Control"]))

qq_plot01 <- ggplot(normal_cells, aes(sample = z)) + stat_qq(distribution = qnorm) +
   stat_qq_line(colour = "black")
qq_plot02 <- qq_plot01 + labs(title = "z-score Q-Q plot (normal distribution)",
   subtitle = paste0("Kolmogorov-Smirnov = ", round(kolsmir_test$p.value,
      digits = 3)), x = "Theoretical Quantiles", y = "Sample Quantiles")
qq_plot03 <- qq_plot02 + scale_x_continuous(limits = c(-3.5,
   3.5), breaks = c(-3, -2, -1, 0, 1, 2, 3)) + theme_bw() +
   theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(size = 8,
      hjust = 0.5, face = "italic"))
```

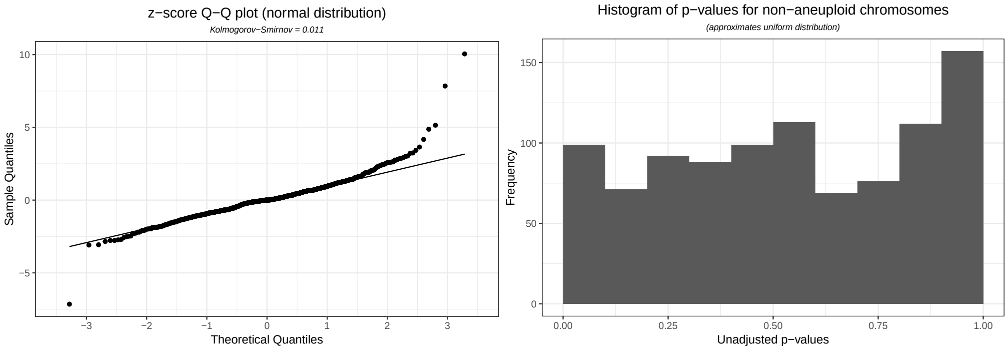

Figure 3: Distribution of z-scores and p-values

```
ggsave(paste0(output_folder, "mouse_8cell_GTseq_QQPlot.pdf"),
    plot = qq_plot03, device = "pdf", width = 6, height = 4,
    units = "in")
return(qq_plot03)

# (2) Compare the distribution of p-values in a histogram
histogram01 <- ggplot(normal_cells, aes(x = p)) + geom_histogram(binwidth = 0.1,
    boundary = 0)
histogram02 <- histogram01 + labs(title = "Histogram of p-values for non-aneuploid chromosomes",
    subtitle = paste0("(approximates uniform distribution)"),
    x = "Unadjusted p-values", y = "Frequency")
histogram03 <- histogram02 + theme_bw() + theme(plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(size = 8, hjust = 0.5, face = "italic"))
ggsave(paste0(output_folder, "mouse_8cell_GTseq_Histogram.pdf"),
    plot = histogram03, device = "pdf", width = 6, height = 4,
    units = "in")
return(histogram03)
```

1.4 False predictions

```
## Summarize all failed calls
false_neg <- paste(getFN(emb8)$cell, getFN(emb8)$chr)  # 11
false_neg <- gsub(" ", " chr", false_neg)  # nicer output

false_pos <- paste(getFP(emb8)$cell, getFP(emb8)$chr)  # 5
false_pos <- gsub(" ", " chr", false_pos)  # nicer output

## Look specifically into false negative calls from Embryo E
false_neg_E <- getFN(emb8)[grepl("E", getFN(emb8)$cell), ]  # 8/11 false negatives
truly_aneuploid_E <- getKnownAneu(emb8)[grepl("E", getKnownAneu(emb8)$cell),
    ]  # 22/50 truly aneuploid
```

1.4.1 False negatives

A total of 11 false negative predictions were made (A5 chr5, D7 chr13, D8 chr13, E1 chr7, E1 chr8, E2 chr15, E3 chr4, E4 chr10, E4 chr19, E6 chr8, E6 chr9), and most of these arose in Embryo E (72.73%).

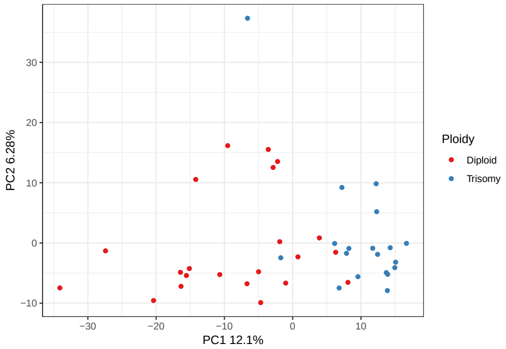

Figure 4: PCA on 1,685 genes

1.4.2 False positives

A total of 5 false positive predictions were made (A8 chr6, B8 chr11, D8 chr7, E4 chr12, E4 chr6), and these appear to be randomly distributed across all embryos.

2. Human Trisomy 21 G&T-seq analysis

Table 3: Data Summary

Ploidy	Cell Number
Diploid	22
T21	19

This data set contains a total of 41 samples that passed quality control metrics (Table 3) with a total of 22981 genes.

2.1 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) as described in *Section 1.1*.

After applying a gene filter of CPM > 50, PCA was performed on 1685 genes. Unlike in the mouse 8-cell stage G&T-seq analysis, the two clusters on PC1 (12.1%) that split by ploidy status (Figure 4) are not completely separated from each other.

2.2 Assessing variance in gene expression