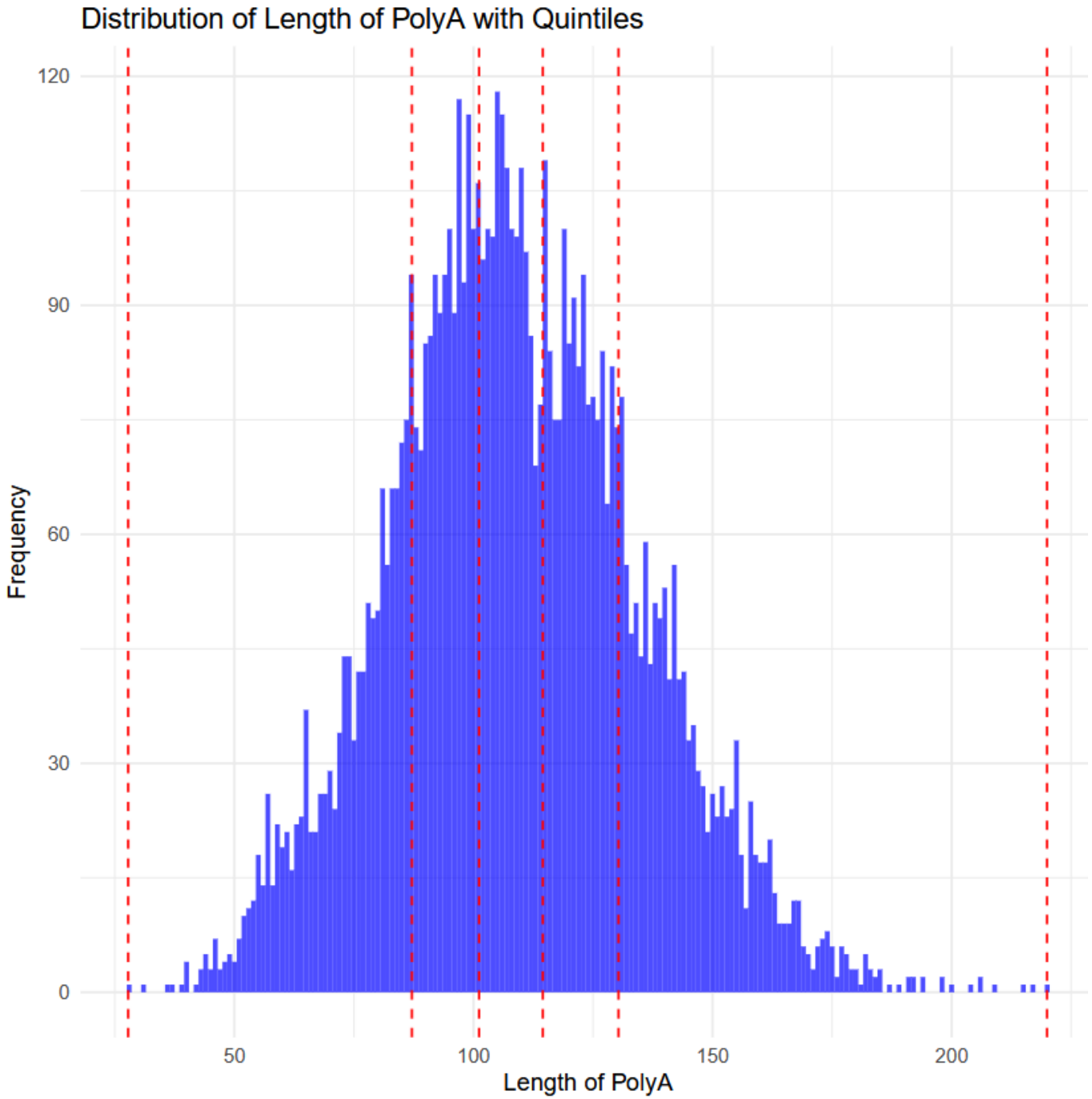
GpalEnrichment

Gene Ontology enrichment of Globodera pallida for genes in each quintiles of polyA length

1.Distribution of polyA length

Quintiles are calculated, and used for plot.

```
/data/pathology/program/Miniforge3/envs/R4.3.2/bin/R
setwd("/data/pathology/cxia/projects/Sebastian/04.GpalEnrichment")
# Load necessary libraries
library(ggplot2)
library(dplyr)

# Read the tab-delimited file
data <- read.table("polyA_nanopolish_medians.tsv", header= TRUE, sep = "\t")

# Calculate quintiles
quintiles <-
quantile(data$Newton_RNA_run1.newton_direct_RNA_newton_scaffolds_minimap_sorted_G_2500
0_no_secondary_pa_tag.bam, probs = seq(0, 1, by = 0.2))

# Print out length ranges of each quintile
print(quintiles)
#     0%     20%     40%     60%     80%    100%
# 27.780  87.120 101.183 114.514 130.338 219.960

# Create a data frame for plotting
quintile_df <- data %>%
  mutate(Quintile =
cut(Newton_RNA_run1.newton_direct_RNA_newton_scaffolds_minimap_sorted_G_25000_no_secon
dary_pa_tag.bam, breaks = quintiles, include.lowest = TRUE))

# Plot the distribution with quintiles
pdf("01.polyA_distribution.pdf")
ggplot(data, aes(x =
Newton_RNA_run1.newton_direct_RNA_newton_scaffolds_minimap_sorted_G_25000_no_secondary
_pa_tag.bam)) +
  geom_histogram(binwidth = 1, fill = "blue", alpha = 0.7) +
  geom_vline(xintercept = quintiles, linetype = "dashed", color = "red") +
  labs(title = "Distribution of Length of PolyA with Quintiles",
       x = "Length of PolyA",
       y = "Frequency") +
  theme_minimal()
dev.off()
```



2.GeneSet and GO terms

2.1 Extract gene names and associated GO terms for each quintile interval.

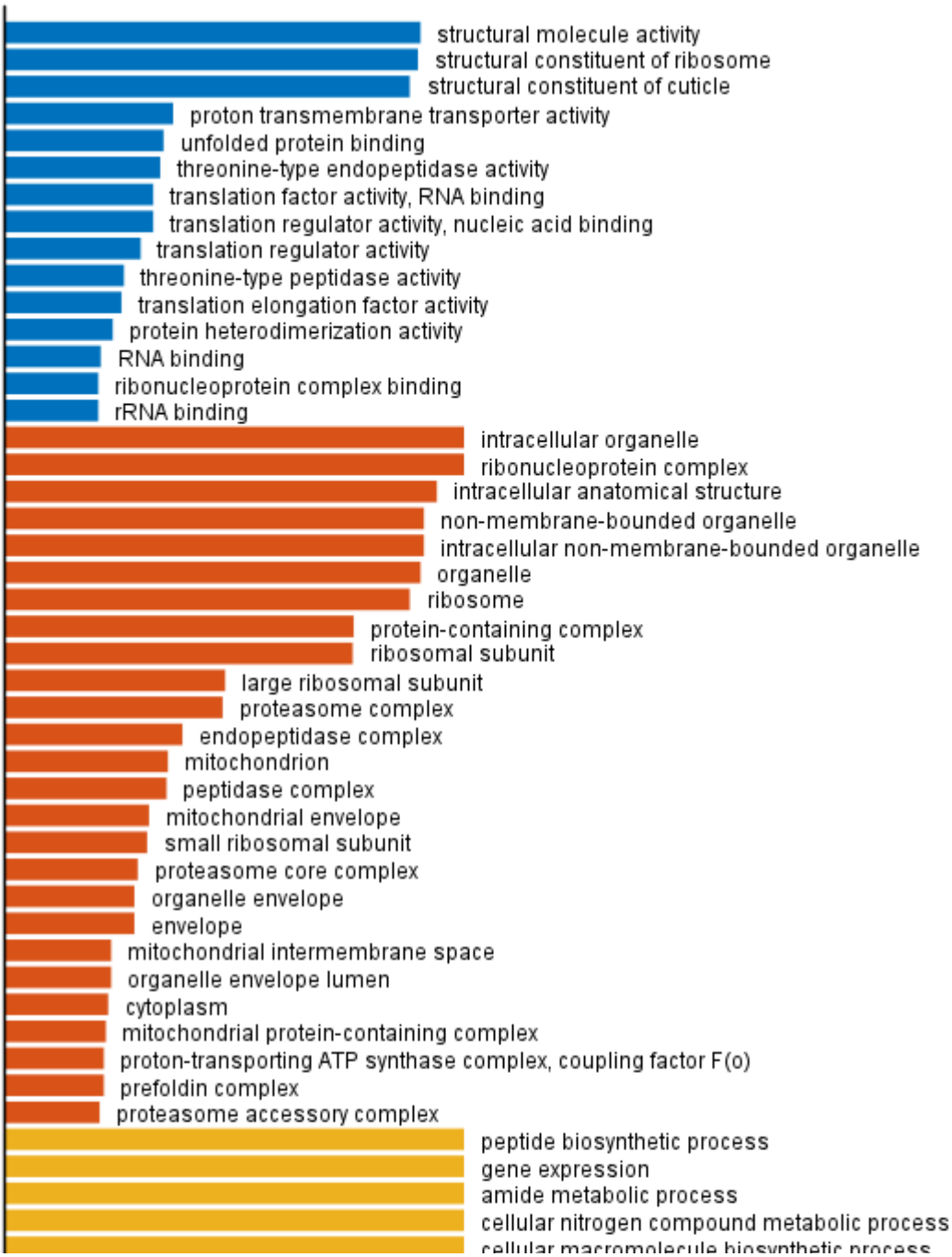
```python
import pandas as pd
import re

# Define the quintile intervals
quintile_ranges = [27.780, 87.120, 101.183, 114.514, 130.338, 219.960]

# File paths
data_file = "polyA_nanopolish_medians.tsv"
gff3_file = "Gpal_newton_newton.gff3"

# Function to extract GO terms from a GFF3 attribute field
def extract_go_terms(attributes):
    go_terms = re.findall(r"GO:\d{7}", attributes)
    return ";".join(go_terms) if go_terms else ""

# Read the data file
data = pd.read_csv(data_file, sep='\t')

# Read the GFF3 file and extract gene IDs and GO terms
gene_go_terms = {}
with open(gff3_file, 'r') as file:
    for line in file:
        if not line.startswith('#') and 'mRNA' in line.split('\t')[2]:
            attributes = line.strip().split('\t')[8]
            gene_id_match = re.search(r"Parent=([^;]+)", attributes)
            if gene_id_match:
                gene_id = gene_id_match.group(1)
                go_terms = extract_go_terms(attributes)
                if go_terms:
                    gene_go_terms[gene_id] = go_terms

# Extract the gene names and GO terms for each quintile interval
for i in range(1, len(quintile_ranges)):
    lower = quintile_ranges[i-1]
    upper = quintile_ranges[i]

    # Filter data within the current quintile interval
    quintile_data = data[(data.iloc[:, 5] >= lower) & (data.iloc[:, 5] < upper)]

    # Prepare the output for each gene in the quintile
    quintile_output = []
    for gene_id in quintile_data['gene_id']:
        go_terms = gene_go_terms.get(gene_id, "")
        if go_terms:
            quintile_output.append(f"{gene_id}\t{go_terms}")

    # Save to file and print results
    output_file = f"quintile_{i}_go_terms.txt"
    with open(output_file, 'w') as outfile:
        outfile.write("\n".join(quintile_output))
```

```
    print(f"Quintile {i} ({lower} - {upper}):")
    print("\n".join(quintile_output))
    print("")
```

3.GO enrichment

'GO & KEGG' module in TBtools was used for GO enrichment analysis. Results and a barplot for each quintile are in each folder.

Top100 of Quintile 1:




GO Term
cellular macromolecule metabolic process
organonitrogen compound biosynthetic process
translation
peptide metabolic process
amide biosynthetic process
cellular nitrogen compound biosynthetic process
macromolecule biosynthetic process
ribonucleoprotein complex biogenesis
cellular biosynthetic process
cellular metabolic process
ribosome biogenesis
biosynthetic process
organic substance biosynthetic process
organonitrogen compound metabolic process
nucleoside triphosphate biosynthetic process
ribonucleoside triphosphate biosynthetic process
rRNA processing
purine nucleoside triphosphate biosynthetic process
purine ribonucleoside triphosphate biosynthetic process
ribonucleoside triphosphate metabolic process
rRNA metabolic process
nucleoside triphosphate metabolic process
cellular component biogenesis
purine nucleoside triphosphate metabolic process
purine ribonucleoside triphosphate metabolic process
RNA processing
ribonucleotide biosynthetic process
ribose phosphate biosynthetic process
protein metabolic process
primary metabolic process
ribonucleotide metabolic process
ribose phosphate metabolic process
purine ribonucleotide biosynthetic process
purine-containing compound metabolic process
purine nucleotide metabolic process
purine ribonucleotide metabolic process
nitrogen compound metabolic process
purine nucleotide biosynthetic process
metabolic process
protein folding
purine-containing compound biosynthetic process
RNA splicing, via transesterification reactions
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
mRNA splicing, via spliceosome
ATP biosynthetic process
proton motive force-driven ATP synthesis
cellular process
organic substance metabolic process
ATP metabolic process
translational elongation
RNA splicing
ncRNA processing
organic substance catabolic process
macromolecule metabolic process
0.00
3.00
6.00
9.00
12.00
15.00
18.00
-log10(P-value)
Molecular function
Cellular component
Biological process