Ohnologues VS Biomart

Hussein

2025-07-09

Load the Necessary Libraries

```
library(tidyverse)
library(UpSetR)
library(grid)
library(readxl)
library(writexl)
library(VennDiagram)
```

Load the required Data sets

```
#Get a list of the protein coding genes

protein_coding_genes <- read_delim("C:/Users/HP-ssd/Desktop/Short term project/protein coding genes/gene
                                   delim = "\t", escape_double = FALSE,
                                   trim_ws = TRUE)

protein_coding_genes_list <- protein_coding_genes$symbol


#Get FUSIL file

fusil_m_gene <-  read_delim("C:/Users/HP-ssd/Desktop/Short term project2/fusil.csv")

length(unique(fusil_m_gene$gene_symbol))
```

```
## [1] 7060
```

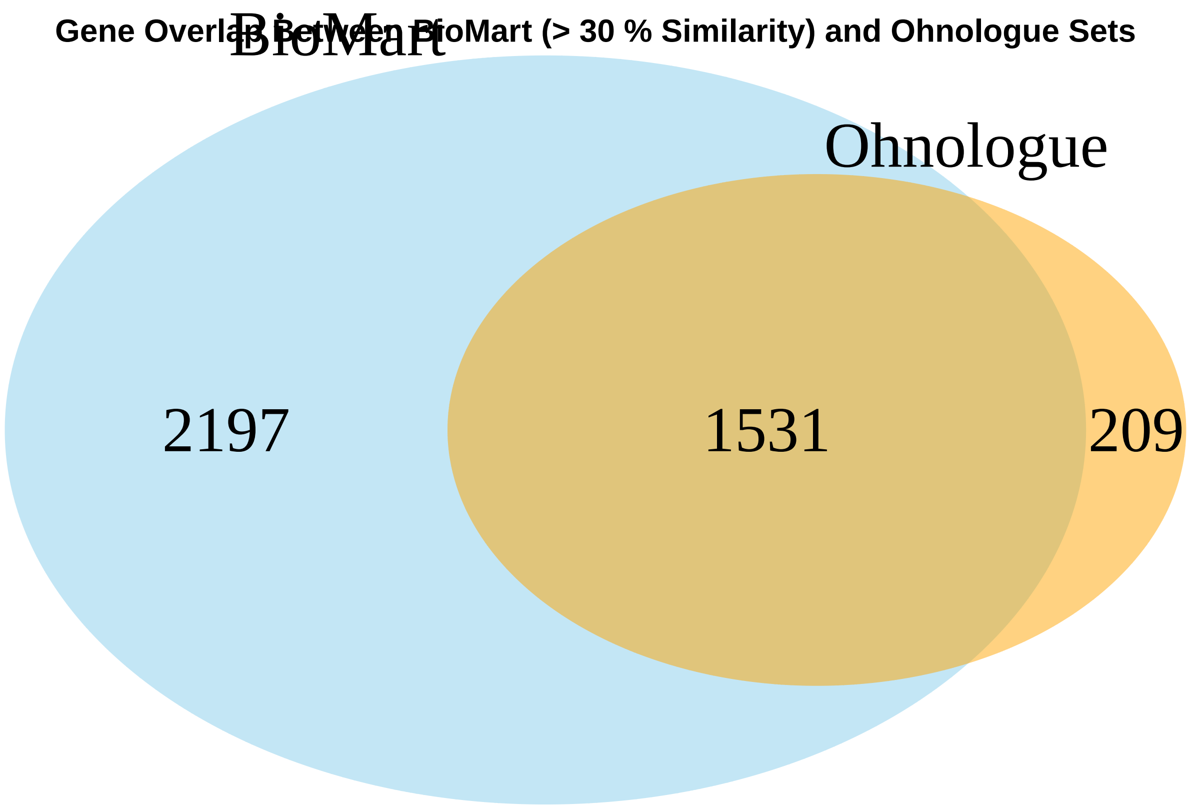
```
#Get a list of Gene Paralogues from biomart

human_gene_paralogues <- read.csv("C:/Users/HP-ssd/Desktop/Short term project2/paralogues/human_gene_pa

human_gene_paralogues <- human_gene_paralogues %>%
  dplyr::select(-1,-2,-4)%>%
  rename(gene_symbol = external_gene_name )

length(unique(human_gene_paralogues$gene_symbol))
```

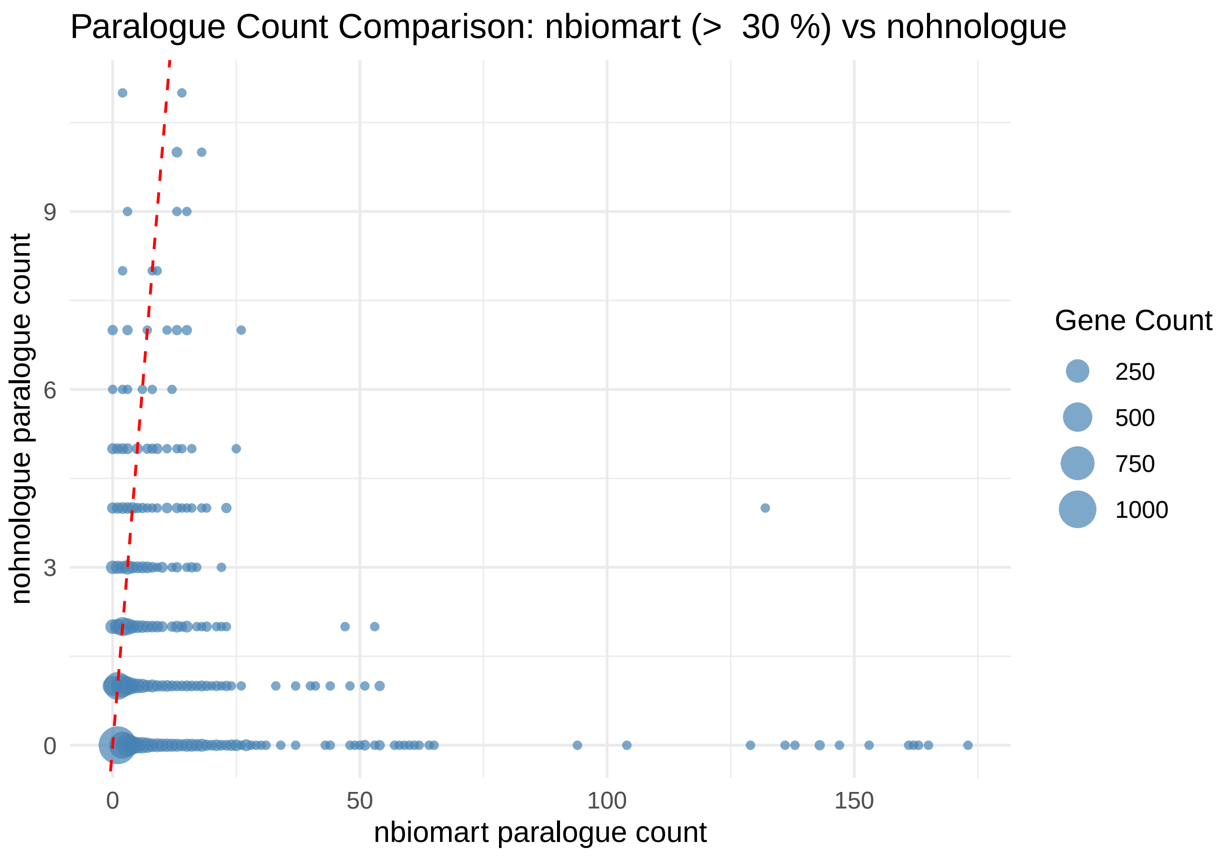
```
## [1] 7039
```

```
#Merge the FUSIL file with paralogues dataset

paralogue_fusil <- human_gene_paralogues %>%
  left_join(fusil_m_gene)%>%
  dplyr::select(-6, -7) %>%
  mutate(hsapiens_paralog_associated_gene_name = na_if(hsapiens_paralog_associated_gene_name,"")) %>%
  distinct()

ohnologs_relaxed <- read_delim("C:/Users/HP-ssd/Desktop/Short term project/ohnologs/hsapiens.Pairs.Rela
                               delim = "\t", escape_double = FALSE,
                               trim_ws = TRUE)
```

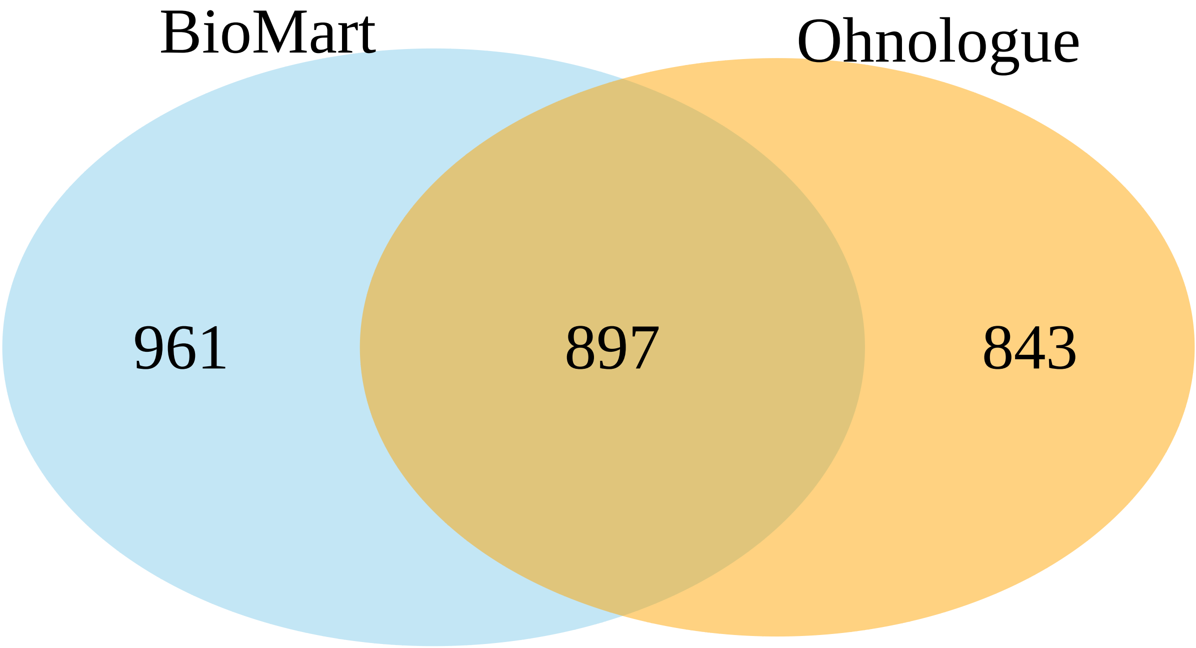
Comparing the 2 Data sets in different Similarity Thresholds for Biomart Paralogues

```
# Define thresholds to loop over
thresholds <- c(30, 50, 70)

# Loop through each similarity threshold
for (threshold in thresholds) {

  # Step 1: Filter ohnologs
  ohnologs_filtered <- ohnologs_relaxed %>%
    filter(Symbol1 %in% fusil_m_gene$gene_symbol) %>%
    dplyr::select(Symbol1, Symbol2) %>%
    rename(gene_symbol = Symbol1, gene_paralogue = Symbol2)

  # Step 2: Filter biomart by similarity threshold
  biomart_filtered <- human_gene_paralogues %>%
    filter(hsapiens_paralog_perc_id > threshold) %>%
    filter(gene_symbol %in% protein_coding_genes$symbol) %>%
    na.omit() %>%
    dplyr::select(gene_symbol, hsapiens_paralog_associated_gene_name) %>%
    rename(gene_paralogue = hsapiens_paralog_associated_gene_name)

  # Step 3: Count paralogues
  count_biomart <- biomart_filtered %>%
    group_by(gene_symbol) %>%
    tally(name = "n_biomart")

  count_ohnologue <- ohnologs_filtered %>%
    group_by(gene_symbol) %>%
    tally(name = "n_ohnologue")

  # Step 4: Merge counts
  merged_counts <- full_join(count_biomart, count_ohnologue, by = "gene_symbol") %>%
    replace_na(list(n_biomart = 0, n_ohnologue = 0)) %>%
    mutate(
      in_biomart = n_biomart > 0,
      in_ohnologue = n_ohnologue > 0
    )

  # Step 5: Venn Diagram
  biomart_genes <- merged_counts$gene_symbol[merged_counts$in_biomart]
  ohnologue_genes <- merged_counts$gene_symbol[merged_counts$in_ohnologue]

  grid.newpage()
  draw.pairwise.venn(
    area1 = length(biomart_genes),
    area2 = length(ohnologue_genes),
    cross.area = length(intersect(biomart_genes, ohnologue_genes)),
    category = c("BioMart", "Ohnologue"),
    fill = c("skyblue", "orange"),
    lty = "blank",
    cex = 2,
    cat.cex = 2,
    cat.pos = c(-20, 20)
  )
  grid.text(
    paste("Gene Overlap Between BioMart (>", threshold, "% Similarity) and Ohnologue Sets"),
    x = 0.5, y = 0.95, gp = gpar(fontsize = 12, fontface = "bold")
  )

  # Step 6: Scatter Plot
  plot_data <- merged_counts %>%
    group_by(n_biomart, n_ohnologue) %>%
    summarise(gene_count = n(), .groups = "drop")

  print(
    ggplot(plot_data, aes(x = n_biomart, y = n_ohnologue, size = gene_count)) +
      geom_point(alpha = 0.7, color = "steelblue") +
      geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red") +
      scale_size_continuous(name = "Gene Count") +
      labs(
        title = paste("Paralogue Count Comparison: nbiomart (> ", threshold, "%) vs nohnologue"),
        x = "nbiomart paralogue count",
        y = "nohnologue paralogue count"
      ) +
      theme_minimal()
  )
}
```

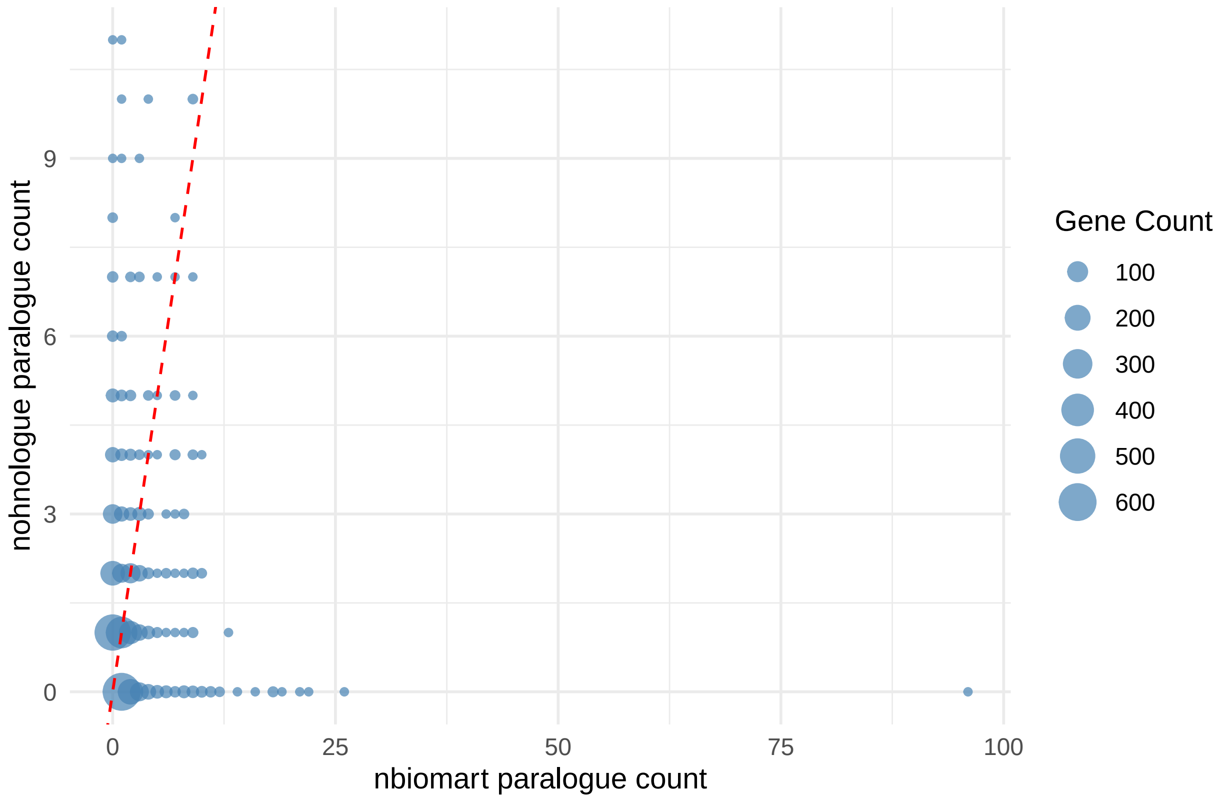

Gene Overlap Between BioMart (> 30 % Similarity) and Ohnologue Sets
BioMart
Ohnologue
2197
1531
209

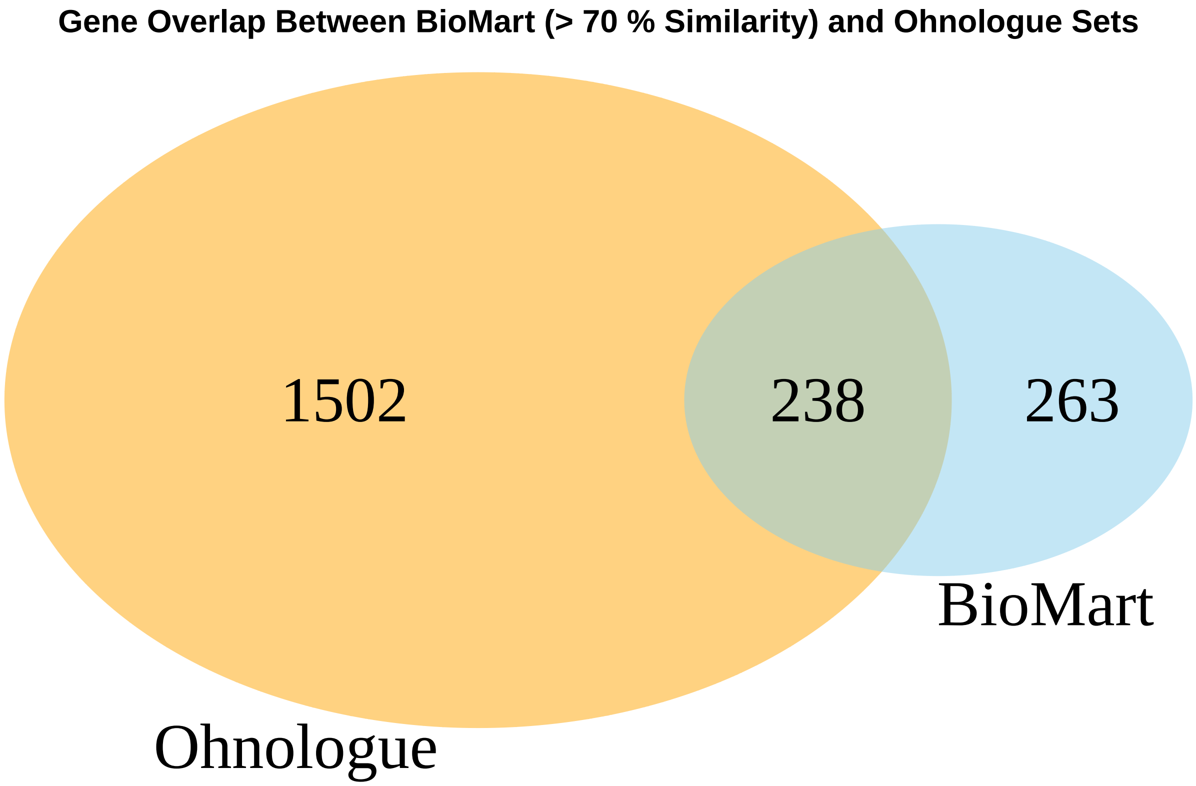

Paralogue Count Comparison: nbiomart (> 30 %) vs nohnologue
nohnologue paralogue count
0
3
6
9
nbiomart paralogue count
0
50
100
150
Gene Count
250
500
750
1000

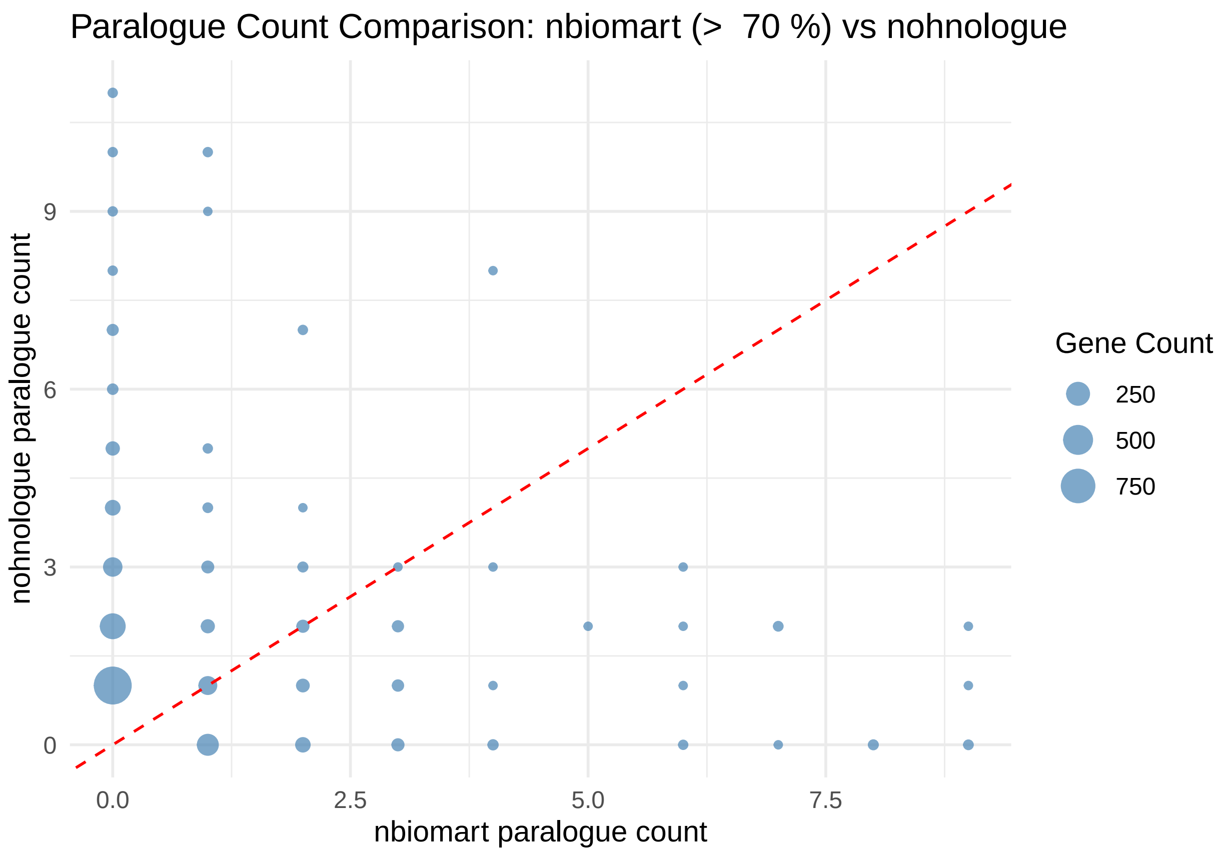
Gene Overlap Between BioMart (> 50 % Similarity) and Ohnologue Sets



Paralogue Count Comparison: nbiomart (> 50 %) vs nohnologue




Gene Overlap Between BioMart (> 70 % Similarity) and Ohnologue Sets
1502
238
263
BioMart
Ohnologue



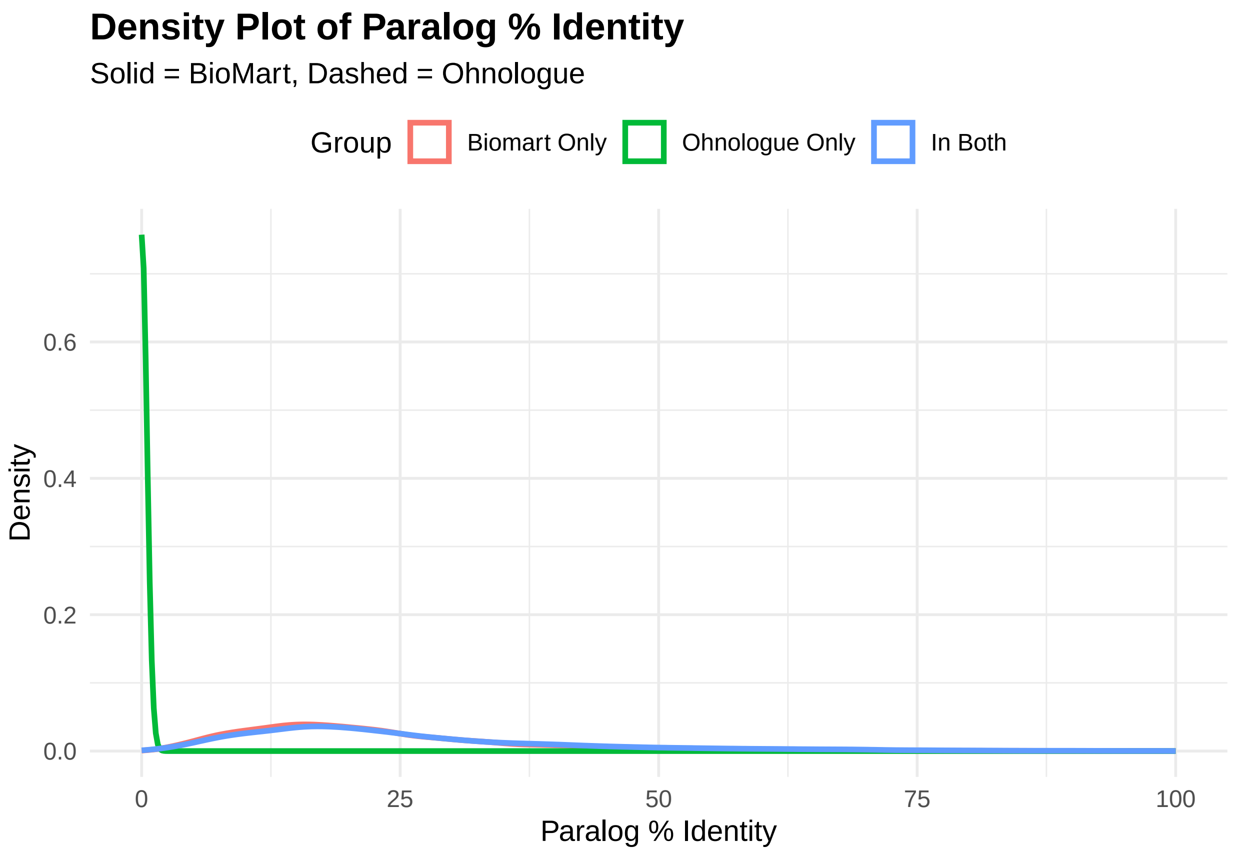
Plotting to see the distribution of Paralogues in both Data sets according to their % Similarity

```
ohnologs_relaxed_filtered <- ohnologs_relaxed %>%
  filter(Symbol1 %in% fusil_m_gene$gene_symbol)%>%  #Should we filter for the fusil genes?
  dplyr::select(3,4) %>%
  rename("gene_symbol" = "Symbol1" ) %>%
  rename( "gene_paralogue" = "Symbol2")


biomart_paralogue <- human_gene_paralogues %>%
  filter(gene_symbol %in% protein_coding_genes$symbol) %>%
  na.omit() %>%
  dplyr::select(1,2,3) %>%
  rename("gene_paralogue" = "hsapiens_paralog_associated_gene_name")

#write.csv(biomart_paralogue, "C:/Users/HP-ssd/Desktop/biomart_paralogue.csv")


count_only_in_biomart <- biomart_paralogue %>%
  group_by(gene_symbol, hsapiens_paralog_perc_id) %>%
  tally() %>%
  rename("n_biomart" = "n")
```

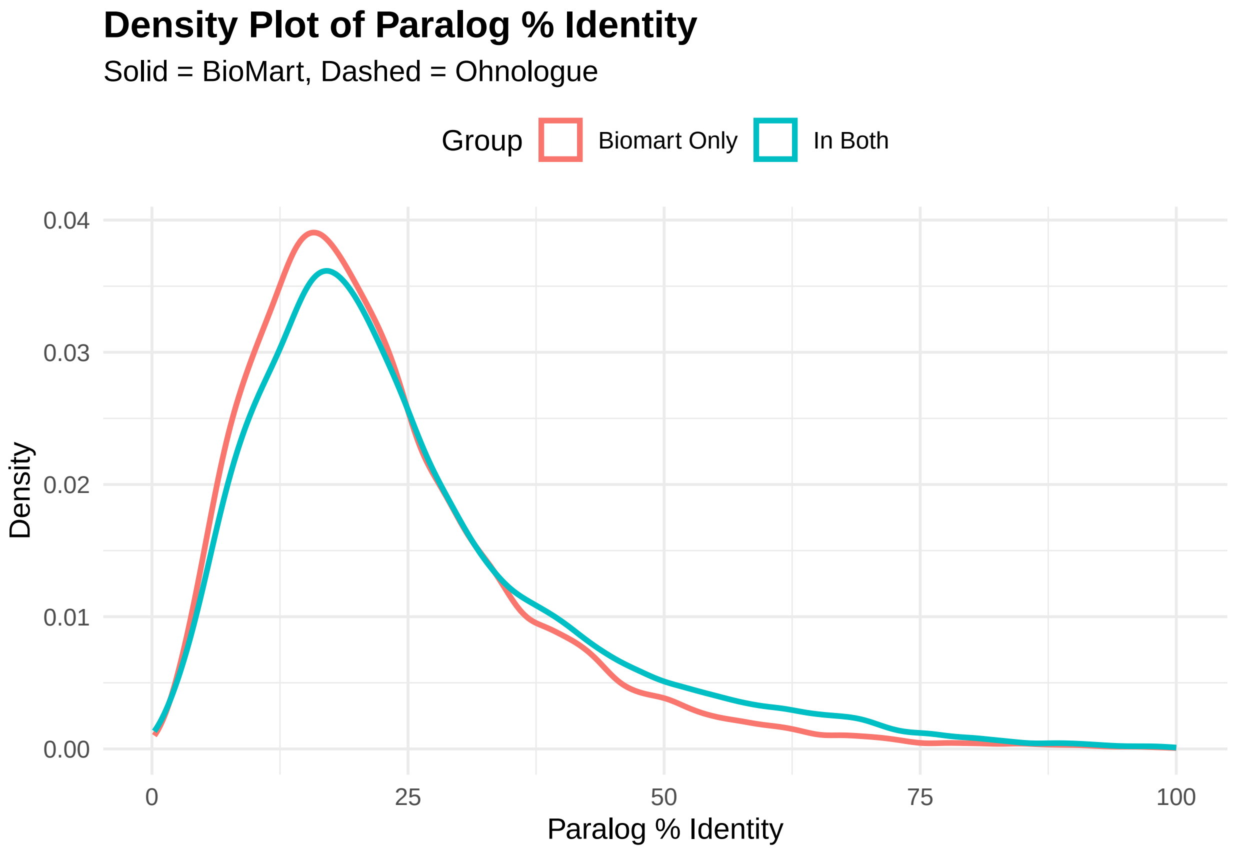
```r
count_only_in_ohnologues <- ohnologs_relaxed_filtered %>%
  group_by(gene_symbol)%>%
  tally() %>%
  rename("n_ohnologue" = "n")


# joining the 2 dataset

joined_paralogue <- count_only_in_biomart %>%
  full_join(count_only_in_ohnologues, by = "gene_symbol") %>%
  mutate_all(~replace_na(.,0))
```

```
## `mutate_all()` ignored the following grouping variables:
## * Column `gene_symbol`
## i Use `mutate_at(df, vars(-group_cols()), myoperation)` to silence the message.
```

```r
joined_paralogue$in_biomart <- as.numeric(as.integer(joined_paralogue$n_biomart >0 ))
joined_paralogue$in_ohnologue <- as.numeric(as.integer(joined_paralogue$n_ohnologue>0))


#write.csv(joined_paralogue, "C:/Users/HP-ssd/Desktop/joined_paralogue.csv")




#Density plot

#making sure the 0 and 1 are factors

joined_paralogue$in_biomart <- factor(joined_paralogue$in_biomart, levels =
                                        c(0,1), labels = c ("Not in Biomart", "In Biomart"))

joined_paralogue$in_ohnologue <- factor(joined_paralogue$in_ohnologue, levels =
                                        c(0,1), labels = c ("Not in Ohnologues", "In Ohnologues"))




joined_paralogue$group <- case_when(
  joined_paralogue$in_biomart == "In Biomart" & joined_paralogue$in_ohnologue == "In Ohnologues" ~ "In
  joined_paralogue$in_biomart == "In Biomart" & joined_paralogue$in_ohnologue == "Not in Ohnologues" ~
  joined_paralogue$in_biomart == "Not in Biomart" & joined_paralogue$in_ohnologue == "In Ohnologues" ~
  TRUE ~ "In Neither"
)

# Convert to factor with a meaningful order
joined_paralogue$group <- factor(joined_paralogue$group,
                                 levels = c("In Neither", "Biomart Only", "Ohnologue Only", "In Both"))



ggplot( joined_paralogue , aes(x = hsapiens_paralog_perc_id))+
  geom_density(aes(color = group),  linewidth = 1)+
  labs(
    title = "Density Plot of Paralog % Identity",
    subtitle = "Solid = BioMart, Dashed = Ohnologue",
    x = "Paralog % Identity",
    y = "Density",
    color = "Group"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    legend.position = "top"
  )
```



```
#Removing the 0 %

joined_paralogue_clean <- joined_paralogue %>%
  filter(hsapiens_paralog_perc_id>0)



ggplot( joined_paralogue_clean , aes(x = hsapiens_paralog_perc_id))+
  geom_density(aes(color = group),  linewidth = 1)+
  labs(
    title = "Density Plot of Paralog % Identity",
    subtitle = "Solid = BioMart, Dashed = Ohnologue",
    x = "Paralog % Identity",
    y = "Density",
    color = "Group"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    legend.position = "top"
  )
```


Density Plot of Paralog % Identity
Solid = BioMart, Dashed = Ohnologue
Group
□ Biomart Only
□ In Both
Density
Paralog % Identity