MetadataAnalysis

```
library(RColorBrewer)
```

```
## Warning: package 'RColorBrewer' was built under R version 4.3.3
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.0     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(ggplot2)
library(networkD3)
library(Polychrome)
library(lisa)
library(paletteer)
```

```
newmeta = read.csv("/Users/riahcul/Downloads/20241106_allmetadata.csv")
newmeta[newmeta == ""] <- NA
```

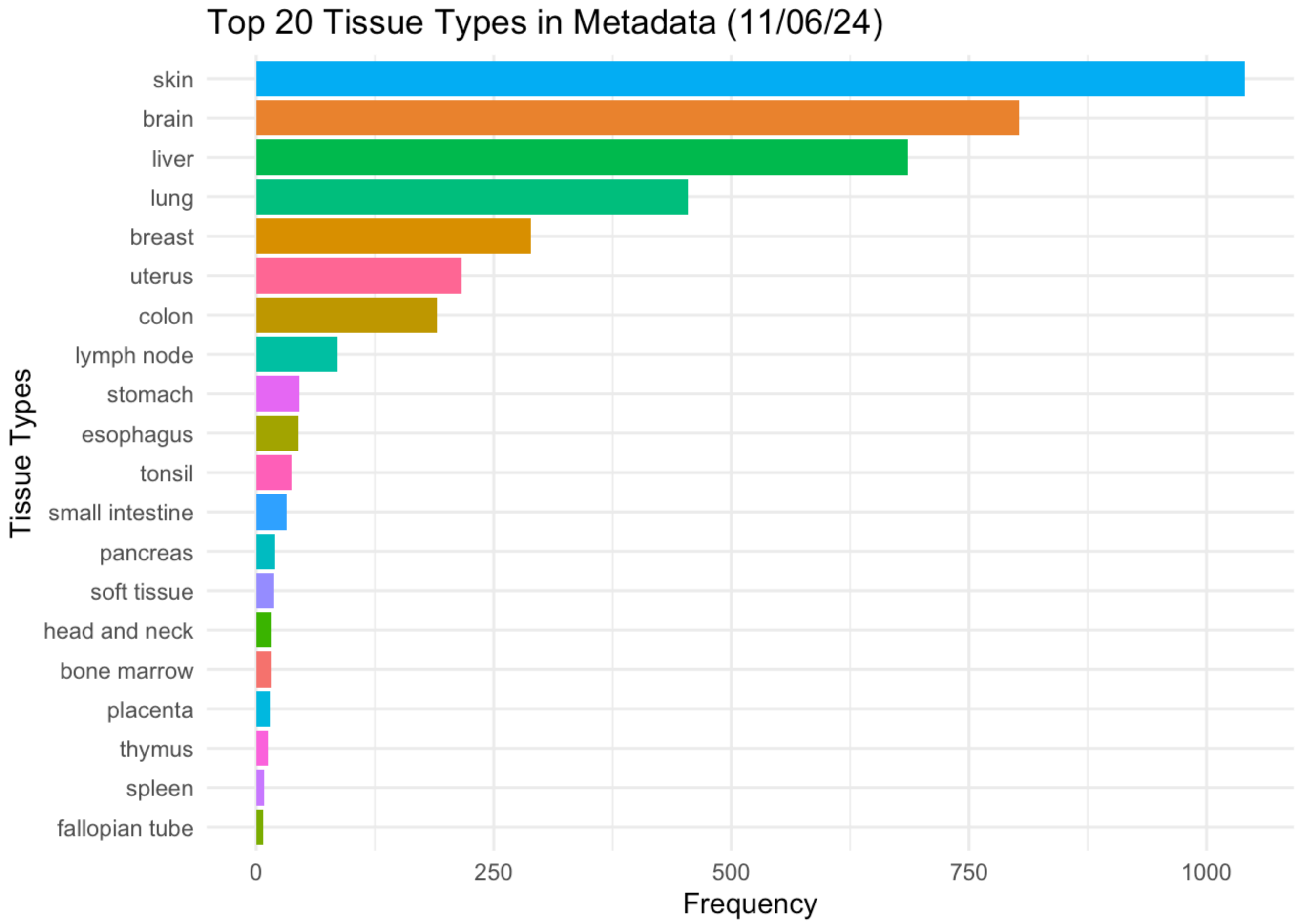
```
tissue_freq <- newmeta |>
  group_by(renamed_tissue) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

print(tissue_freq)
```

```
## # A tibble: 37 × 2
##    renamed_tissue Frequency
##    <chr>              <int>
##  1 skin                1040
##  2 <NA>                 827
##  3 brain                803
##  4 liver                686
##  5 lung                 454
##  6 breast               289
##  7 uterus               216
##  8 colon                190
##  9 lymph node            85
## 10 stomach               45
## # ℹ 27 more rows
```

```
tissue_top <- head(tissue_freq,11)
tissue_bigtop <- head(tissue_freq, 21)
print(tissue_top)
```

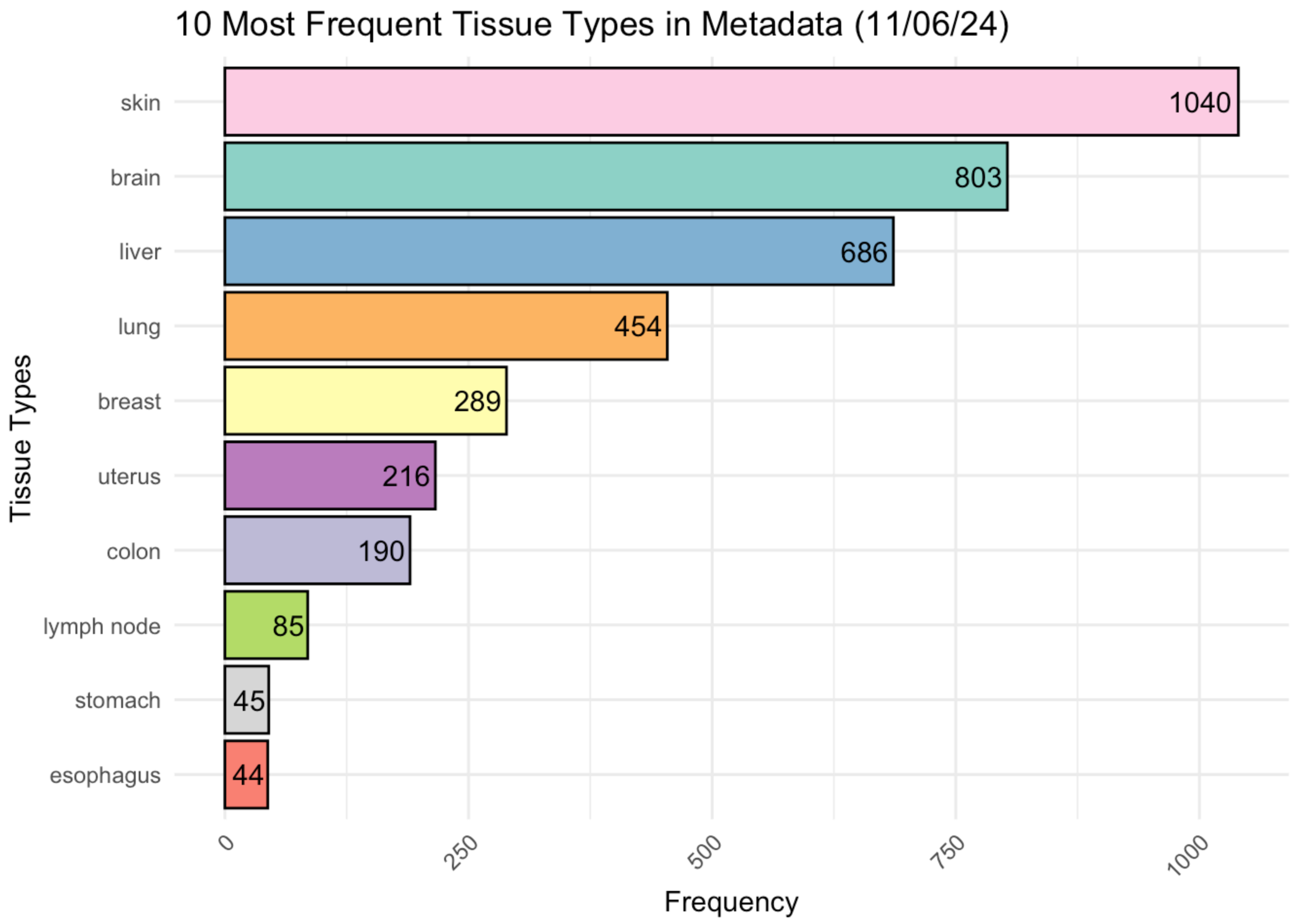
```
## # A tibble: 11 × 2
##    renamed_tissue Frequency
##    <chr>              <int>
##  1 skin                1040
##  2 <NA>                 827
##  3 brain                803
##  4 liver                686
##  5 lung                 454
##  6 breast               289
##  7 uterus               216
##  8 colon                190
##  9 lymph node            85
## 10 stomach               45
## 11 esophagus             44
```

```
tissue_bigtop |>
  filter(renamed_tissue != "NA")|>
  ggplot( aes(x= reorder(renamed_tissue, Frequency), y = Frequency, fill = renamed_tissue) ) + geom_bar(stat="identity") +
  coord_flip() +
  theme_minimal() +
  theme(legend.position = "none") +
  labs(title = "Top 20 Tissue Types in Metadata (11/06/24)",
       x ="Tissue Types")
```



Top 20 Tissue Types in Metadata (11/06/24)

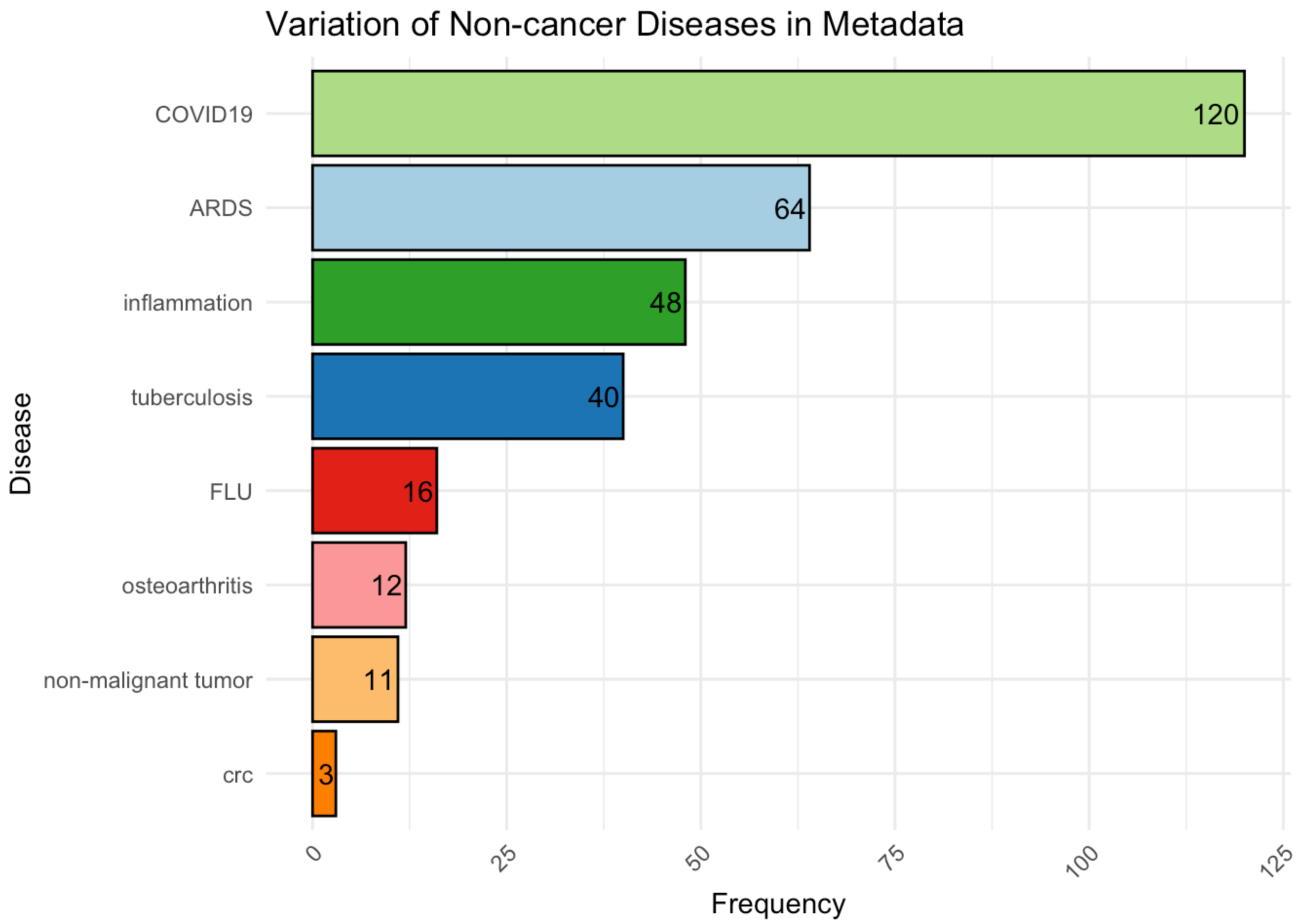
```
mypalette <- brewer.pal(10, "Set3")

tissue_top |> filter(renamed_tissue !="NA")|>
  ggplot( aes(x= reorder(renamed_tissue, Frequency), y = Frequency, fill = renamed_tissue) ) +
  geom_bar(stat ="identity", color ="black") +
  coord_flip() +
  scale_fill_manual(values=mypalette) +
  theme_minimal() +
  theme(legend.position ="none") +
  theme(axis.text.x = element_text(angle =45, vjust =1, hjust =1)) +
  geom_text(aes(label = Frequency), position = position_dodge(width=.9), hjust=1.1) +
  labs( title ="10 Most Frequent Tissue Types in Metadata (11/06/24)",
        x ="Tissue Types")
```



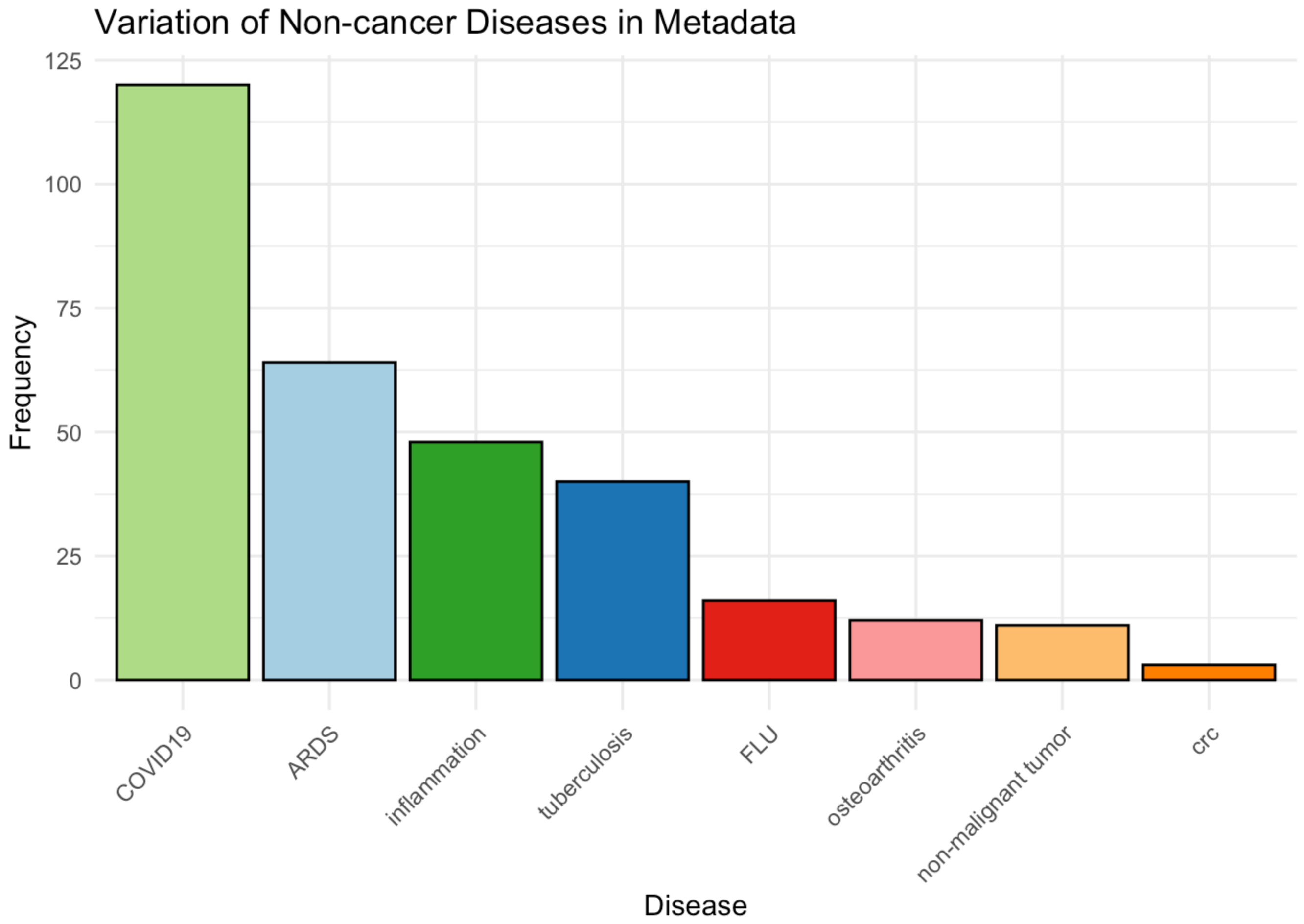
```
disease_freqs <- newmeta |>
  group_by(disease) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

print(disease_freqs)
```

```
## # A tibble: 12 × 2
##    disease             Frequency
##    <chr>                   <int>
##  1 cancer                   1929
##  2 normal                   1605
##  3 <NA>                     1014
##  4 COVID19                   120
##  5 ARDS                       64
##  6 inflammation               48
##  7 tuberculosis               40
##  8 Healthy                    37
##  9 FLU                        16
## 10 osteoarthritis             12
## 11 non-malignant tumor        11
## 12 crc                         3
```

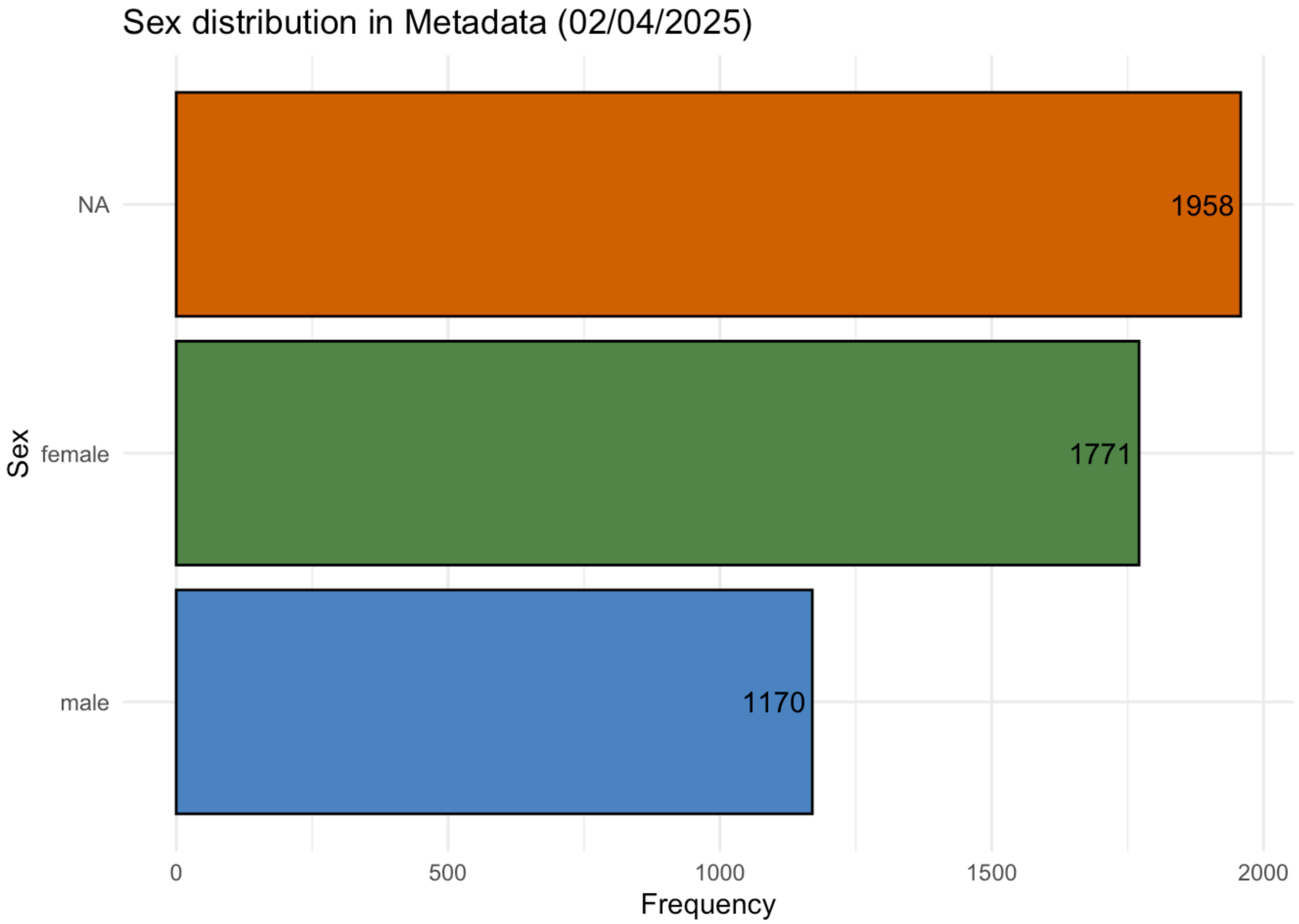
```
disease_freqs <- disease_freqs |>
  mutate(disease = if_else(disease == "Healthy", "normal", disease))

disease_long <- disease_freqs |>
  mutate(disease_split = strsplit(disease," ")) |>
  unnest(disease_split)

diseased_counts <- disease_long |>
  group_by(disease_split) |>
  summarize(Frequency = sum(Frequency)) |>
  arrange(desc(Frequency))

mypalette <- brewer.pal(8, "Paired")

# Filtering out 'cancer' before passing to ggplot
disease_freqs |>
  filter(disease != "cancer") |>
  filter(disease != "normal") |>
  ggplot(aes(x = reorder(disease, Frequency), y = Frequency, fill = mypalette)) +
  geom_bar(stat = "identity", color = "black") +
  scale_fill_manual(values = mypalette) +
  theme_minimal() +
  theme(legend.position = "none") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  geom_text(aes(label = Frequency), position = position_dodge(width=.9), hjust=1.1) +
  labs(
    title = "Variation of Non-cancer Diseases in Metadata",
    x = "Disease"
  ) +
  coord_flip()
```



```
mypalette <- brewer.pal(8, "Paired")

# Filtering out 'cancer' before passing to ggplot
disease_freqs |>
  filter(disease != "cancer") |>
  filter(disease != "normal") |>
  ggplot(aes(x = reorder(disease, -Frequency), y = Frequency, fill = mypalette)) +
  geom_bar(stat = "identity", color = "black") +
  scale_fill_manual(values = mypalette) +
  theme_minimal() +
  theme(legend.position = "none") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  labs(
    title = "Variation of Non-cancer Diseases in Metadata",
    x = "Disease"
  )
```



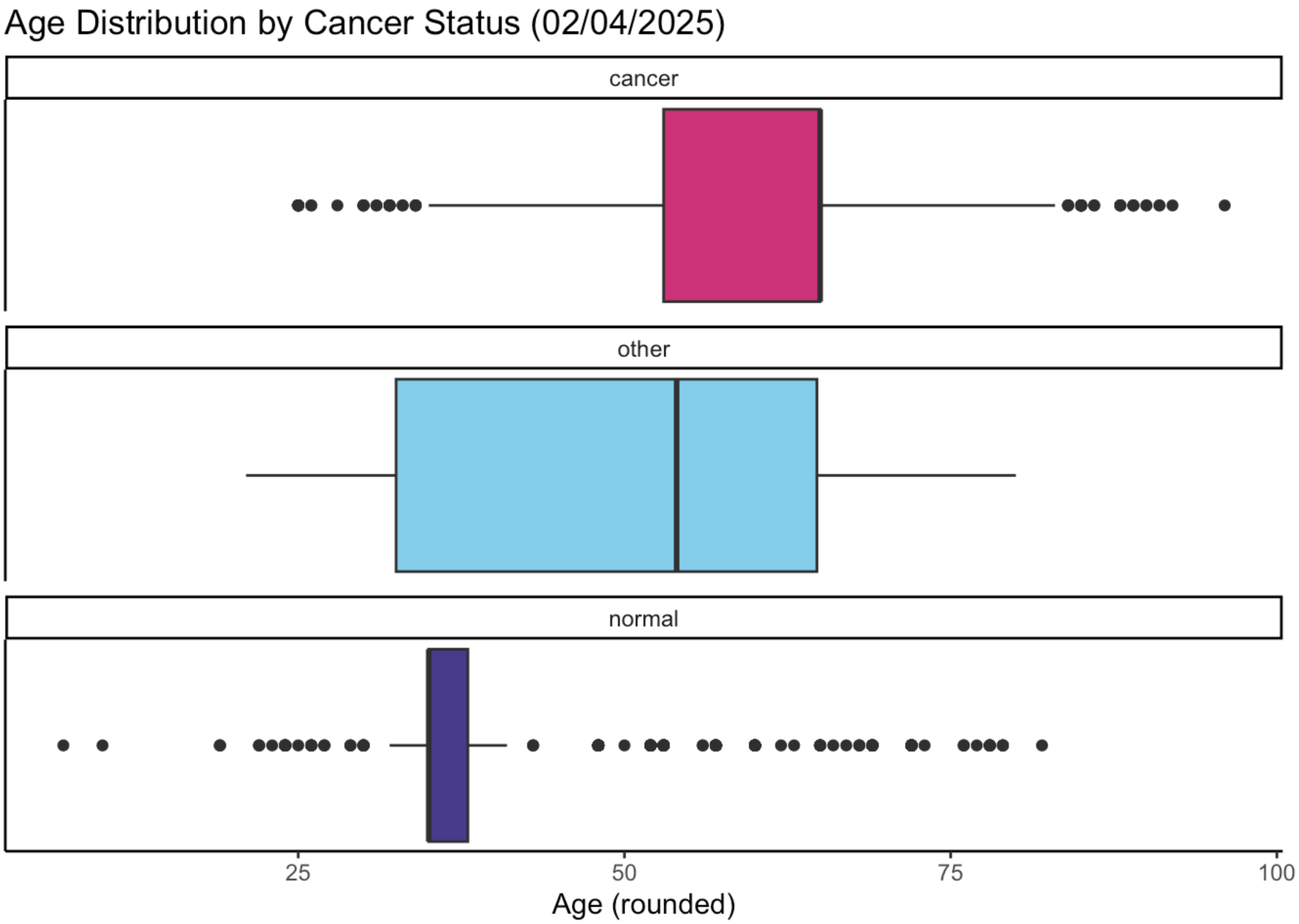
```
sex_freq <- newmeta |>
  group_by(sex) |>
  mutate(
    sex = if_else(sex == "unknown", NA_character_, sex)
    ) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))

sex_freq |>
  ggplot(
    aes(x= reorder(sex, Frequency), y = Frequency, fill = sex)
  ) + geom_bar(stat= "identity", color = "black") +
    scale_fill_manual(breaks = c("female", "male"),
                     values = c("#52854C", "#4E84C4"), na.value = "#D16103") +
   geom_text(aes(label = Frequency), position = position_dodge(width=.9), hjust=1.1) +
  theme_minimal() +
  theme(legend.position = "none") + coord_flip() +
  labs(
    title = "Sex distribution in Metadata (02/04/2025)",
    x = "Sex"
  )
```



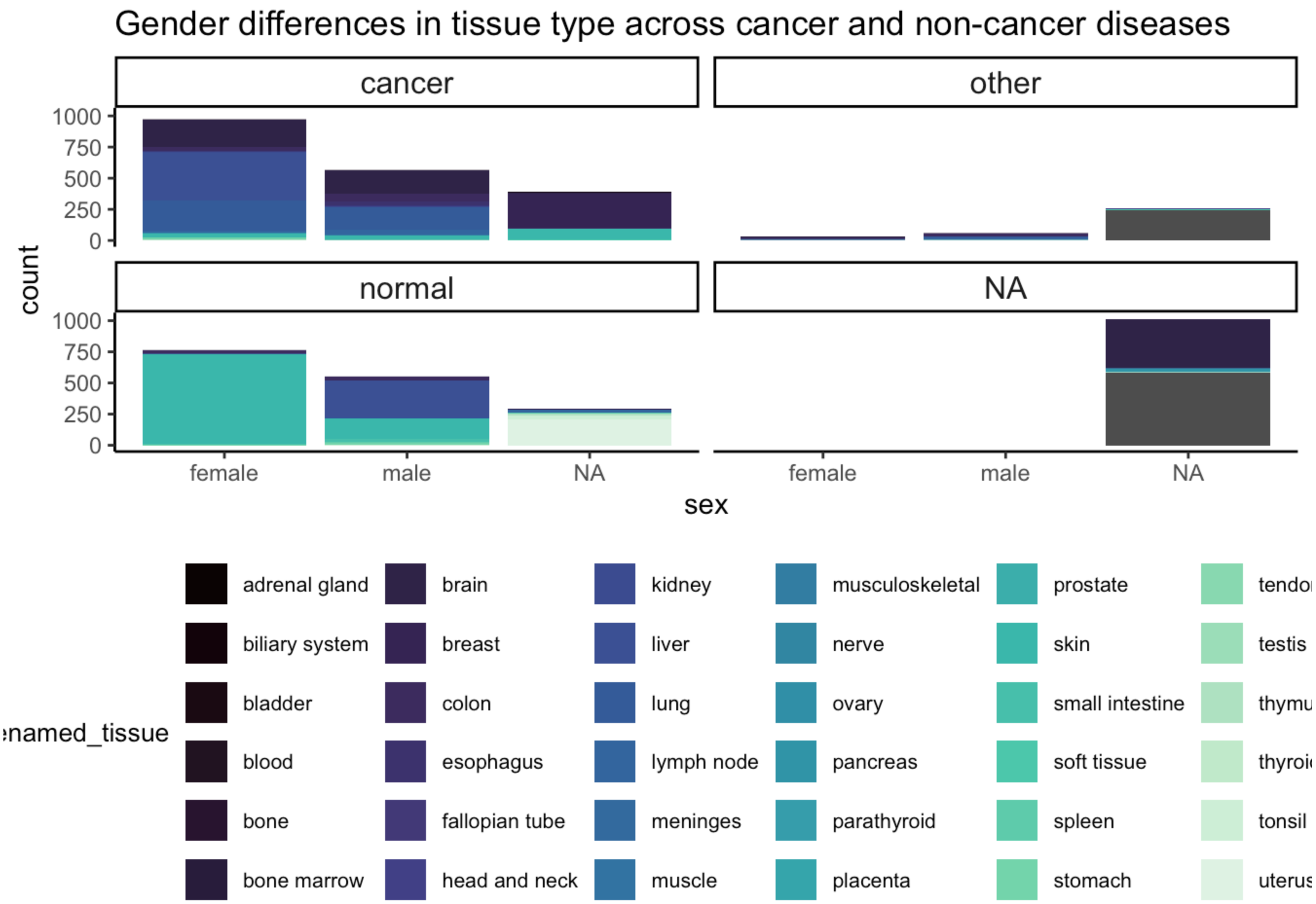
```
#newmetaupdate represents cancer or other and standardizes age
newmeta_update <- newmeta |>
  mutate(newDisease = case_when(
    disease == "cancer" ~ "cancer",
    disease == "normal" ~ "normal",
    is.na(disease) ~ NA_character_,
    TRUE ~ "other"))

newmeta_update <- newmeta_update |>
  mutate(disease = if_else(disease =="Healthy", "normal", disease))

newmeta_update <- newmeta_update |>
  mutate(newAge = round(as.numeric(age)), newDisease = factor(newDisease, levels = c("ca
ncer", "other", "normal")))
```

```
## Warning: There was 1 warning in `mutate()`.
## ℹ In argument: `newAge = round(as.numeric(age))`.
## Caused by warning:
## ! NAs introduced by coercion
```

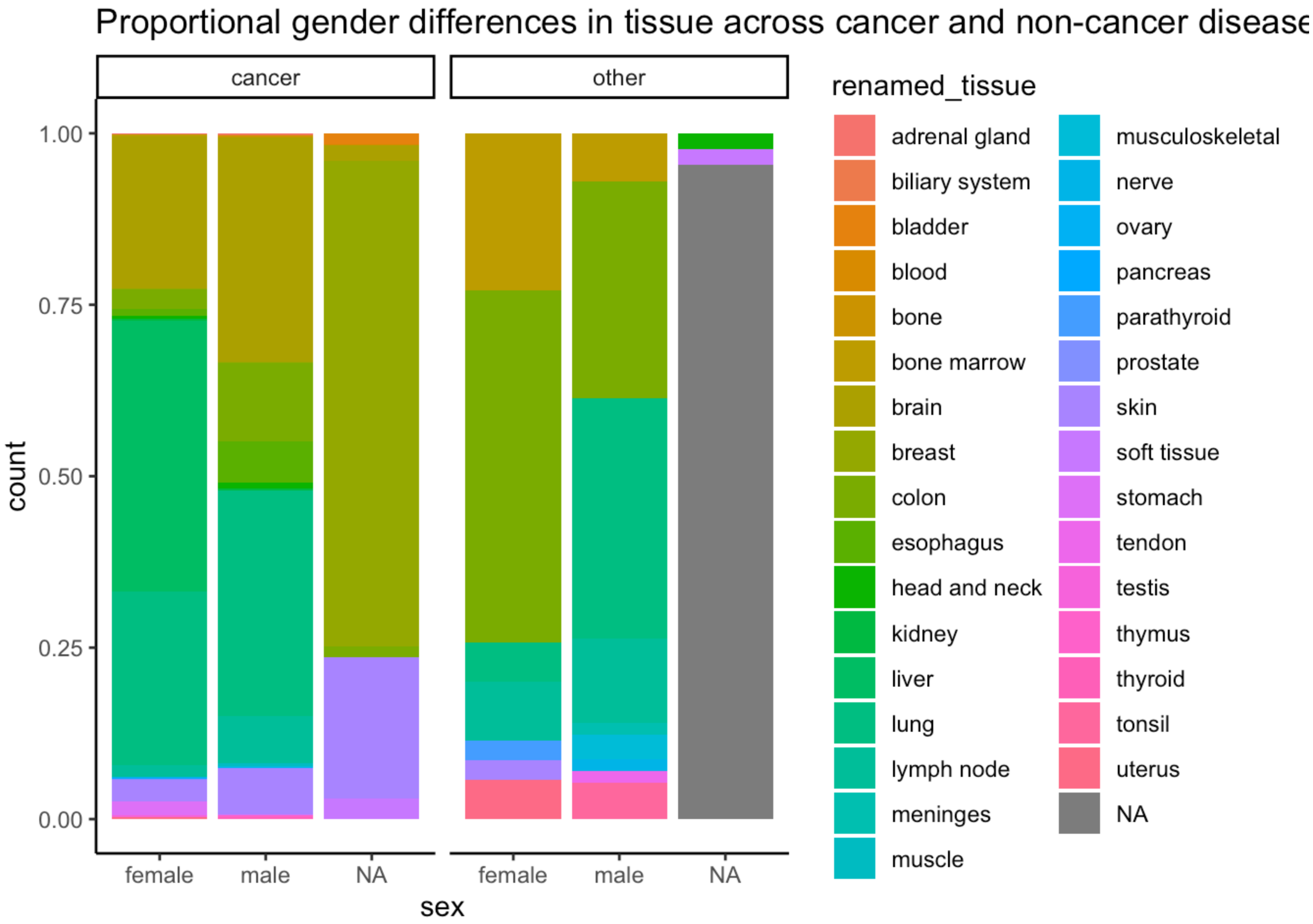
```
newmeta_update |>
  filter(!is.na(newDisease)) |>
  filter(!is.na(newAge)) |>
  ggplot(
    aes(x = newAge, fill = newDisease)) +
    geom_boxplot() +
  scale_fill_manual(breaks = c("cancer", "other", "normal"),
                     values = c("violetred3", "skyblue", "darkslateblue")) +
  facet_wrap(~newDisease, ncol =1) +
  labs(
    title = "Age Distribution by Cancer Status (02/04/2025)",
    x = "Age (rounded)"
  ) +
  theme_classic() +
  theme(legend.position = "none", axis.ticks.y = element_blank(), axis.text.y = element_
blank())
```



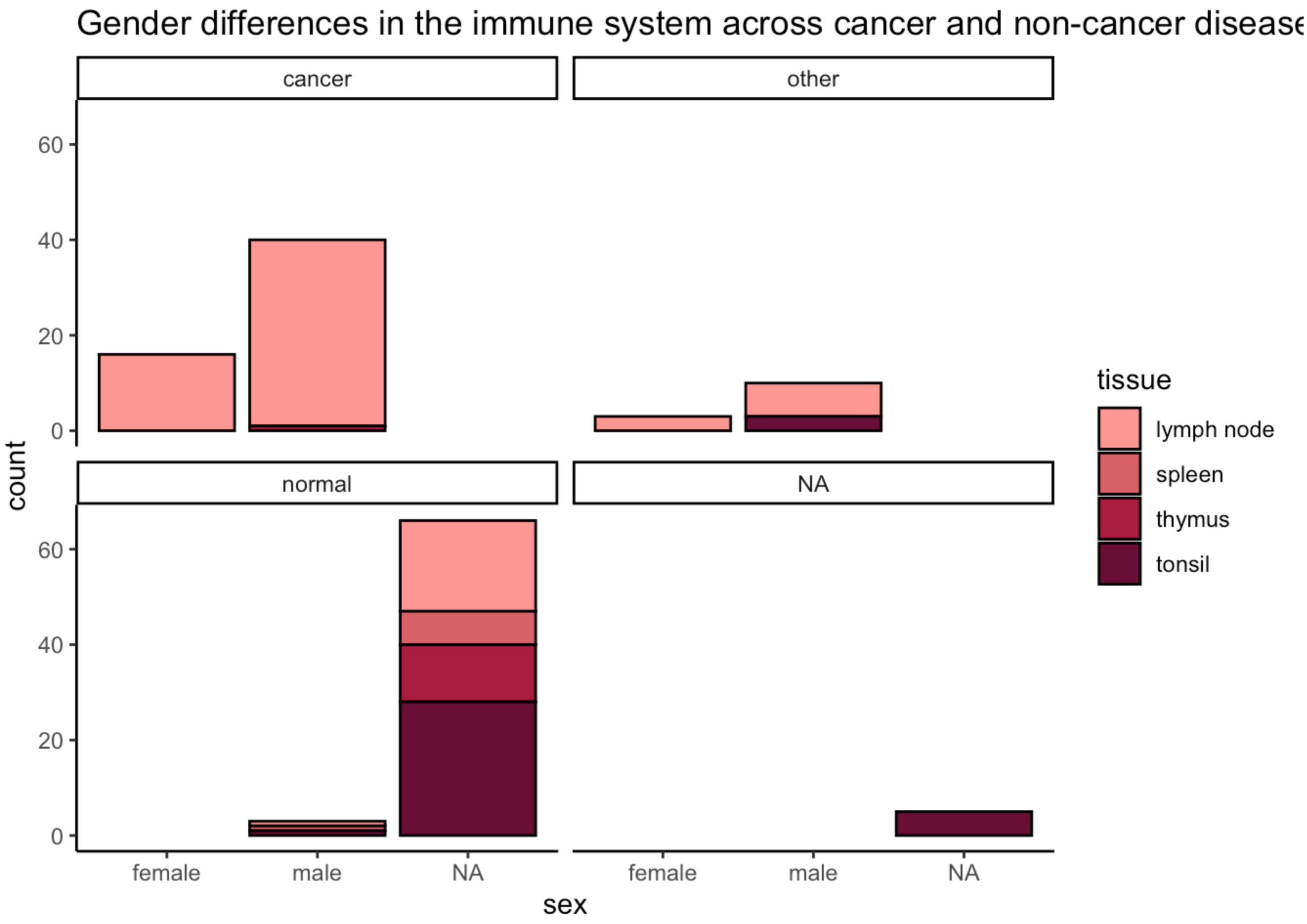
```
newmeta_update <- newmeta_update |>
   mutate(
    sex = if_else(sex == "unknown", NA_character_, sex)
    )

newmeta_update |>
  ggplot(
    aes(x = sex, fill = renamed_tissue)) +
  geom_bar(position = "stack") +
  facet_wrap(~newDisease) +
  labs(
    title = "Gender differences in tissue type across cancer and non-cancer diseases"
  ) +
  scale_fill_viridis_d(option= "mako", na.value = "grey30") +
  theme_classic() +
  guides(fill = guide_legend(nrow = 6)) +
  theme(legend.position = "bottom",
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 8),
        strip.text = element_text(size = 12))
```



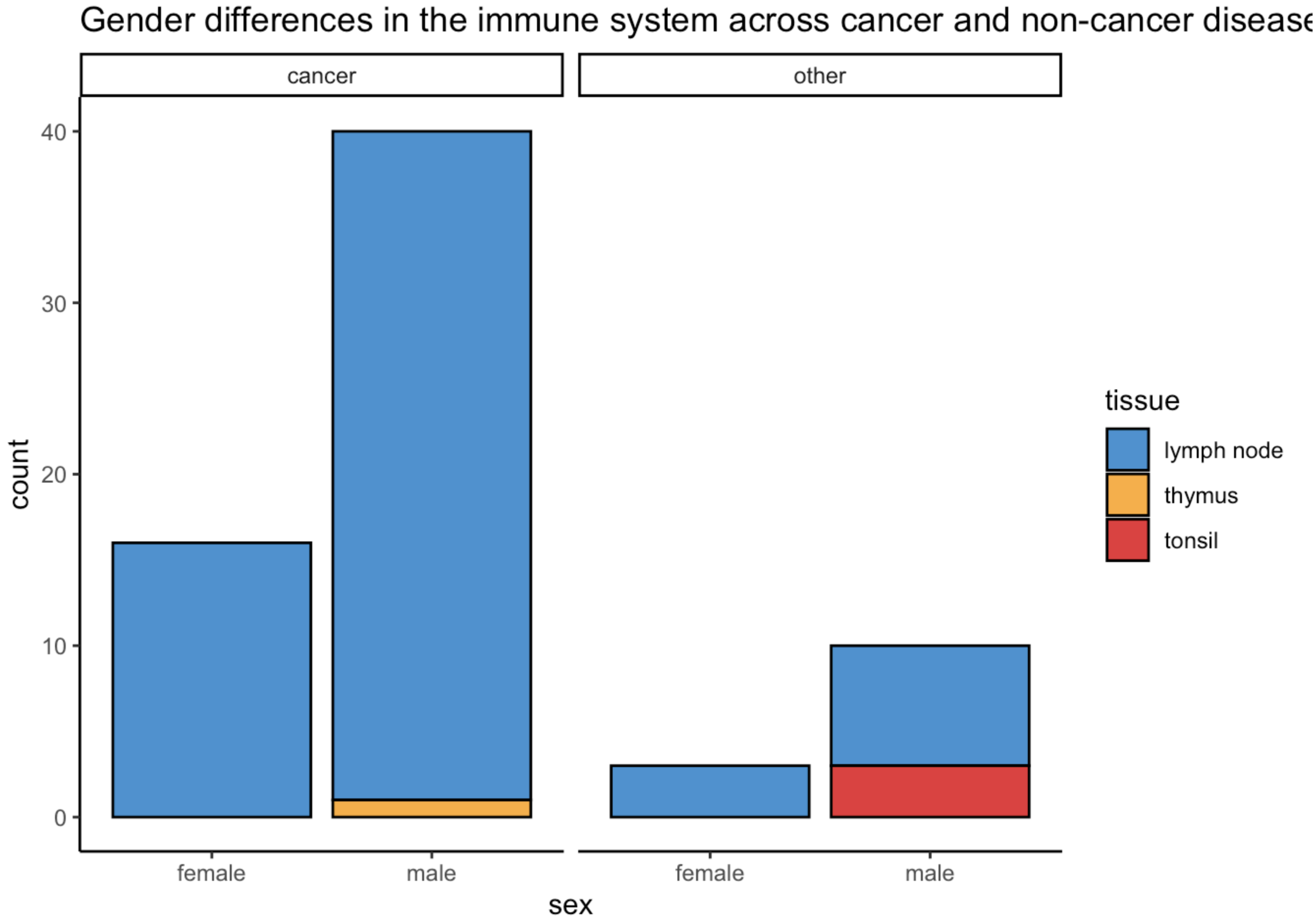
```
newmeta_update <- newmeta_update |> mutate(
  sex = if_else(sex == "unknown", NA_character_, sex)
  )

newmeta_update |>
  filter(newDisease != "normal") |>
  ggplot(
    aes(x = sex, fill = renamed_tissue)) +
  geom_bar(position = "fill") +
  facet_wrap(~newDisease) +
  labs(
    title ="Proportional gender differences in tissue across cancer and non-cancer disea
ses"
) +
theme_classic()
```



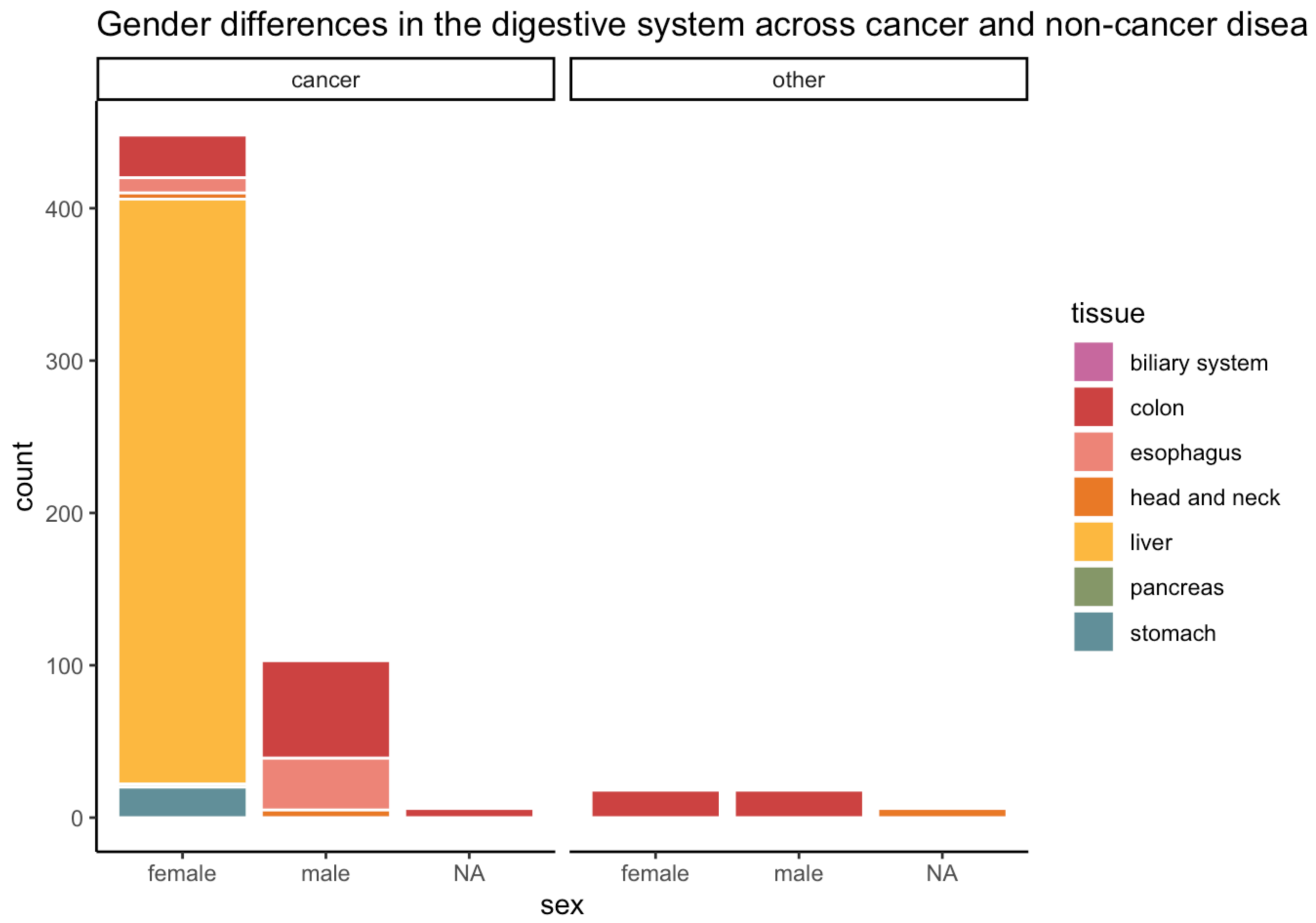
```
library(wesanderson)

newmeta_update |> filter(large_tissue_unit =="Immune System") |>
  ggplot( aes(x = sex, fill = tissue)) + geom_bar(position ="stack", color ="black") +
# scale_fill_manual(values = wes_palette("GrandBudapest2", n = 4)) +
scale_fill_paletteer_d("MoMAColors::Althoff") +
  facet_wrap(~newDisease) +
  labs( title ="Gender differences in the immune system across cancer and non-cancer dis
eases") +
  theme_classic()
```



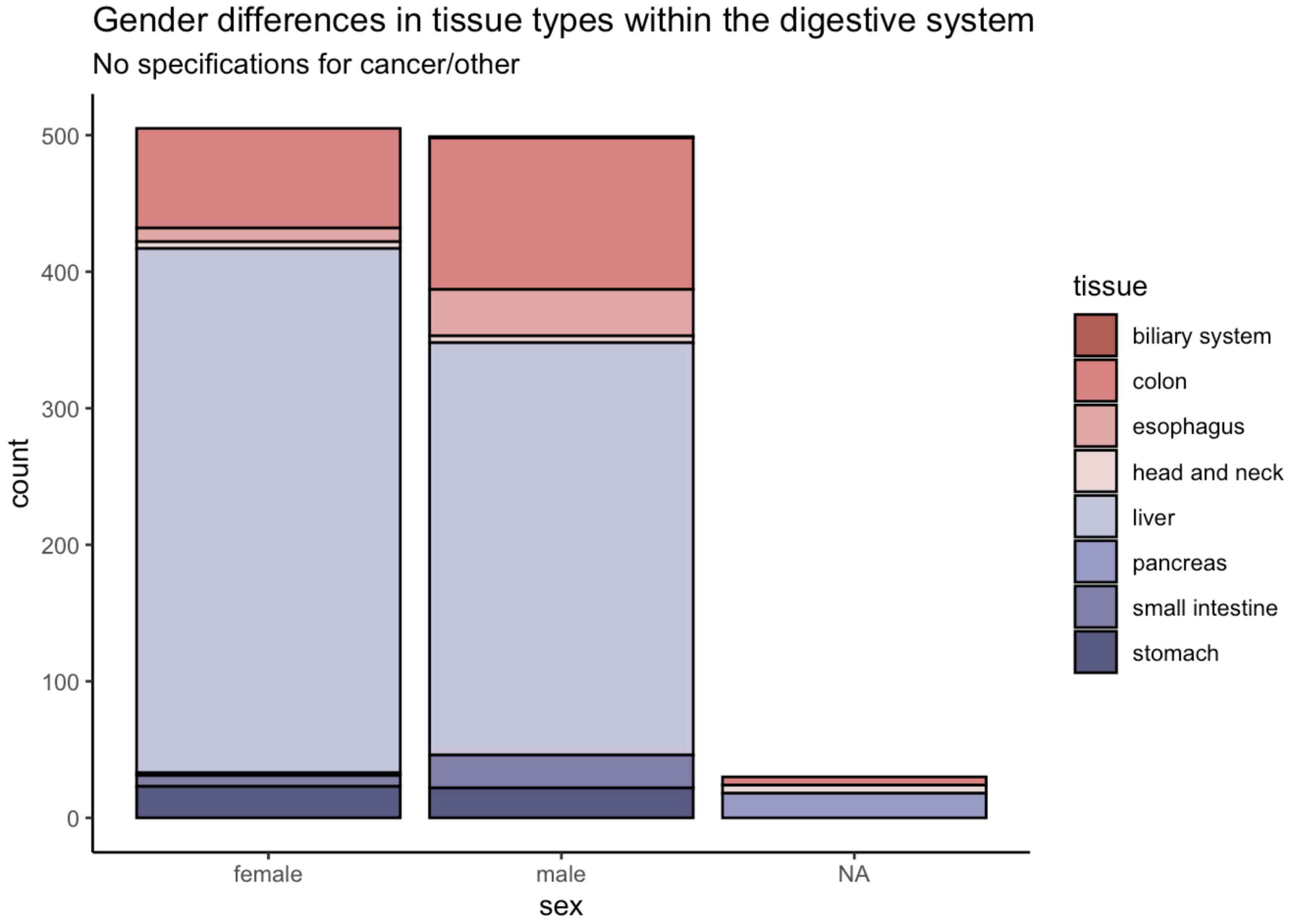
```
newmeta_update |>
  filter(large_tissue_unit =="Immune System") |>
  filter(newDisease !="normal") |>
  ggplot( aes(x = sex, fill = tissue)) +
  geom_bar(position ="stack", color ="black") +
# scale_fill_manual(values = wes_palette("GrandBudapest2", n = 4)) +
scale_fill_paletteer_d("nationalparkcolors::Badlands") +
  facet_wrap(~newDisease) +
  labs( title ="Gender differences in the immune system across cancer and non-cancer dis
eases") +
  theme_classic()
```



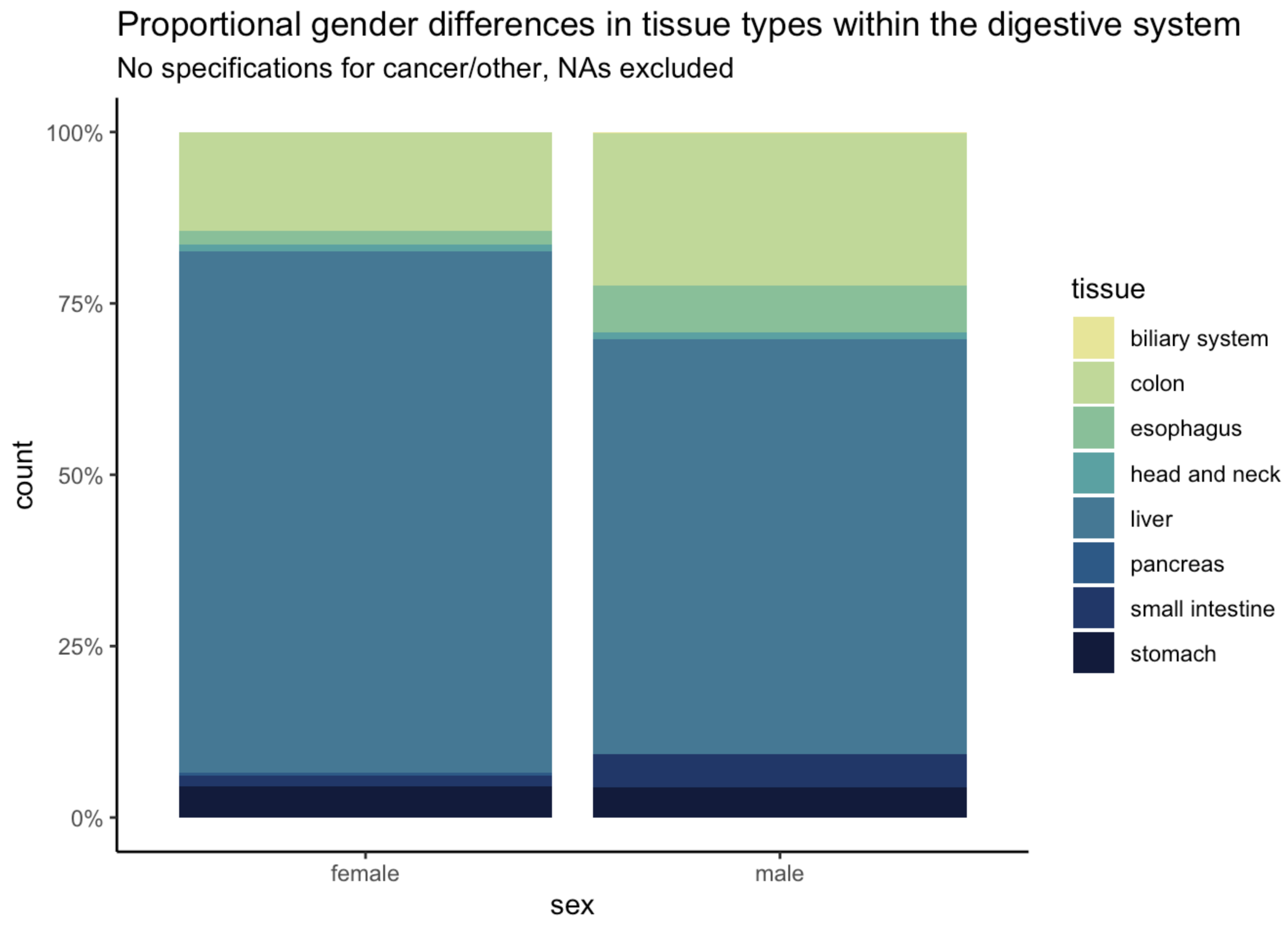
```
newmeta_update |>
  filter(large_tissue_unit =="Digestive System") |>
  filter(newDisease !="normal") |>
  ggplot( aes(x = sex, fill = tissue)) + geom_bar(position ="stack", color ="white") +
# scale_fill_manual(values = brewer.pal(8, "Paired")) +
scale_fill_paletteer_d("MetBrewer::Cross") +
  facet_wrap(~newDisease) +
  labs( title ="Gender differences in the digestive system across cancer and non-cancer diseases") +
  theme_classic()
```



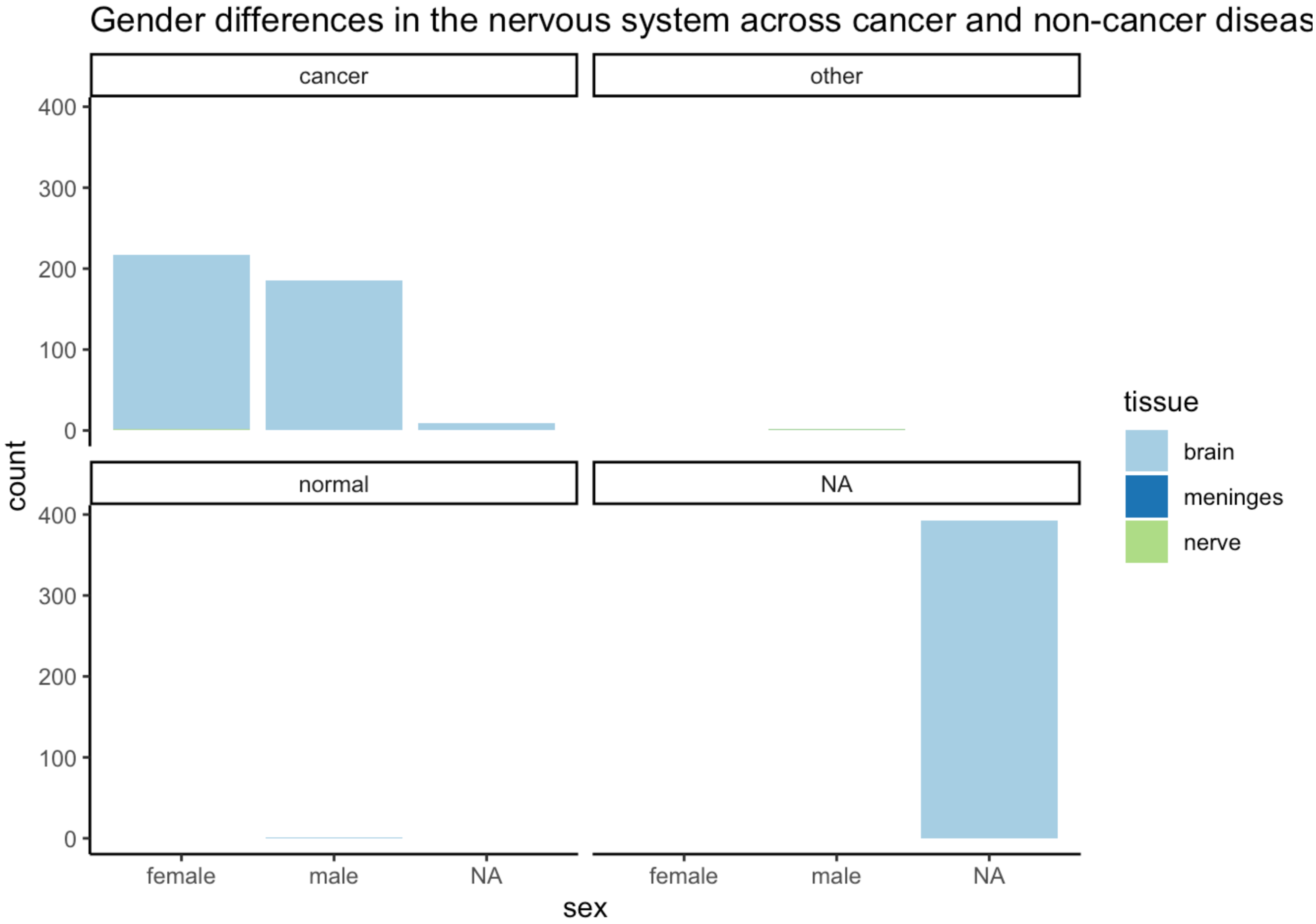
```
newmeta_update |>
  filter(large_tissue_unit =="Digestive System") |>
  ggplot( aes(x = sex, fill = tissue)) + geom_bar(color="black") +
#scale_fill_manual(values =brewer.pal(8, "Paired")) +
scale_fill_paletteer_d("MetBrewer::Cassatt1") +
  labs(
  title ="Gender differences in tissue types within the digestive system",
  subtitle = "No specifications for cancer/other") +
  theme_classic()
```



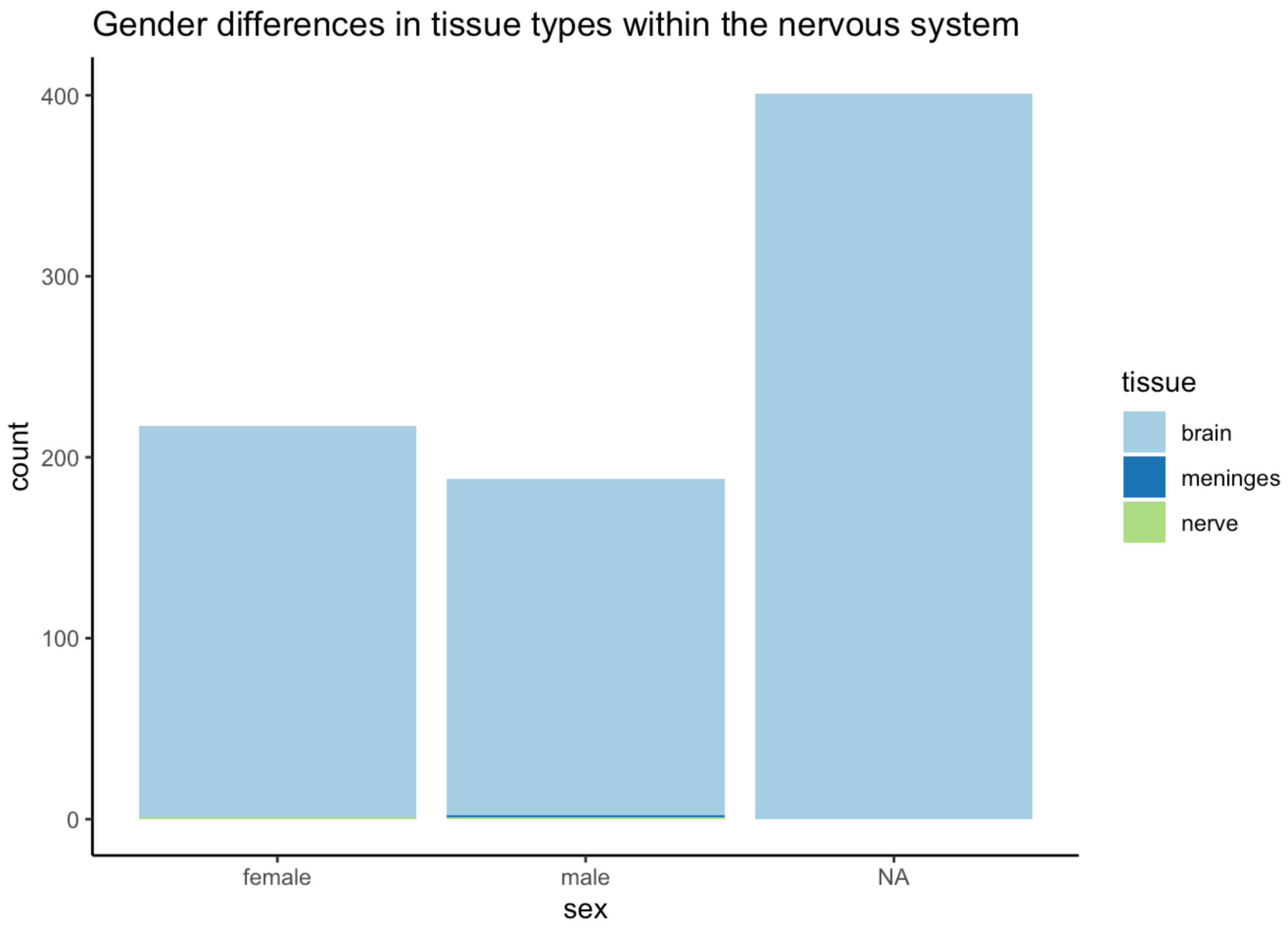
```
nona_newmeta <- newmeta_update |>
#filter(large_tissue_unit == "Digestive System") |>
filter(!is.na(renamed_tissue)) |>
filter(!is.na(sex))

nona_newmeta |>
  filter(large_tissue_unit =="Digestive System") |>
  ggplot( aes(x = sex, fill = tissue)) + geom_bar(position ="fill") +
  scale_y_continuous(labels = scales::percent) +
#scale_fill_manual(values =brewer.pal(8, "Paired")) +
scale_fill_paletteer_d("MoMAColors::Ernst") +
  labs(
    title ="Proportional gender differences in tissue types within the digestive syste
m",
    subtitle ="No specifications for cancer/other, NAs excluded") +
  theme_classic()
```



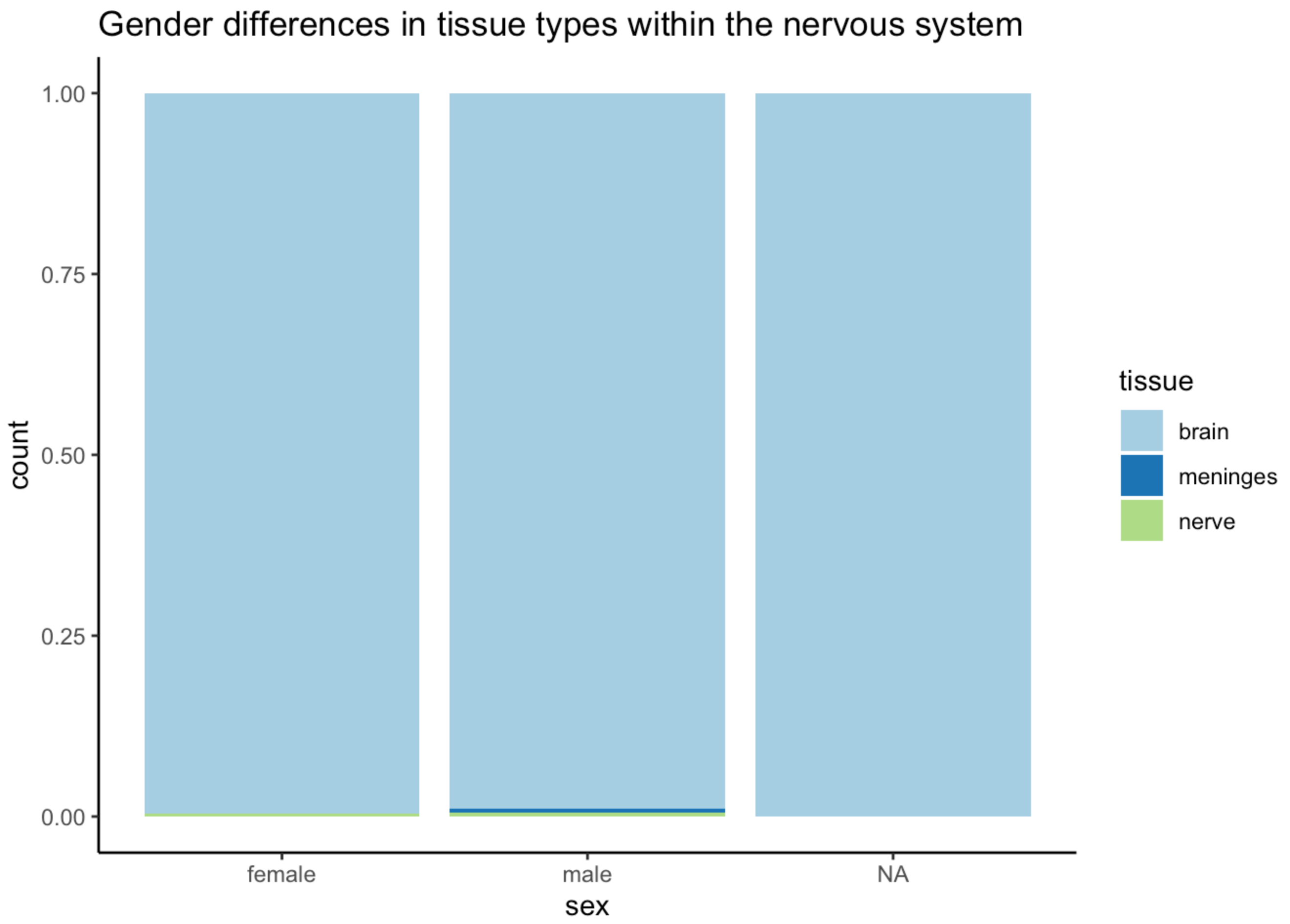
```
newmeta_update |>
  filter(large_tissue_unit =="Nervous System") |>
  ggplot( aes(x = sex, fill = tissue)) + geom_bar(position ="stack") +
  scale_fill_manual(values = brewer.pal(8,"Paired")) +
  facet_wrap(~newDisease) +
  labs(
    title ="Gender differences in the nervous system across cancer and non-cancer diseas
es") +
  theme_classic()
```



```
newmeta_update |>
  filter(large_tissue_unit =="Nervous System") |>
  ggplot( aes(x = sex, fill = tissue)) +
  geom_bar() +
  scale_fill_manual(values = brewer.pal(8, "Paired")) +
  labs(
    title ="Gender differences in tissue types within the nervous system") +
  theme_classic()
```



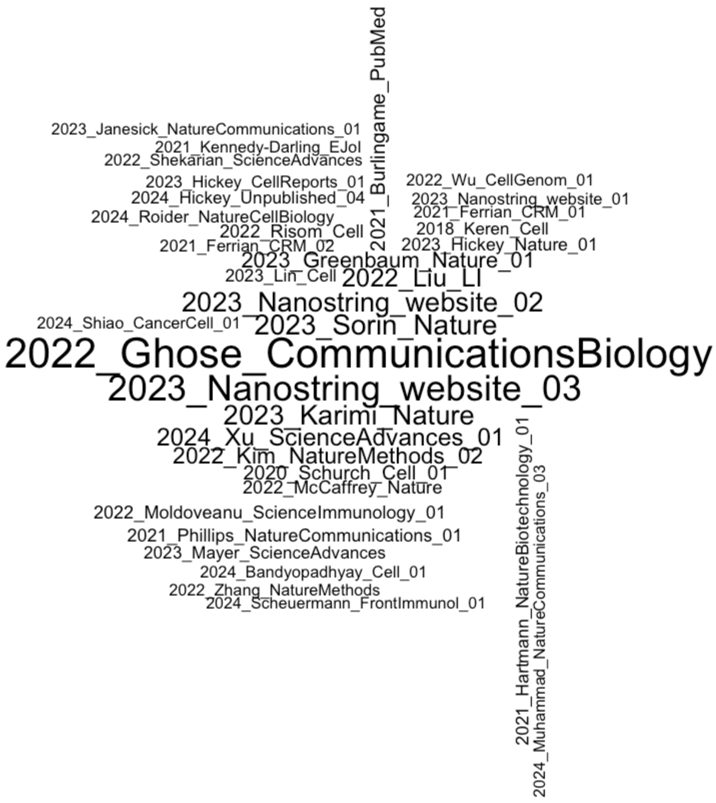
```r
newmeta_update |>
  filter(large_tissue_unit =="Nervous System") |>
  ggplot( aes(x = sex, fill = tissue)) +
  geom_bar(position ="fill") +
  scale_fill_manual(values = brewer.pal(8,"Paired")) +
  labs( title =
"Gender differences in tissue types within the nervous system"
) + theme_classic()
```



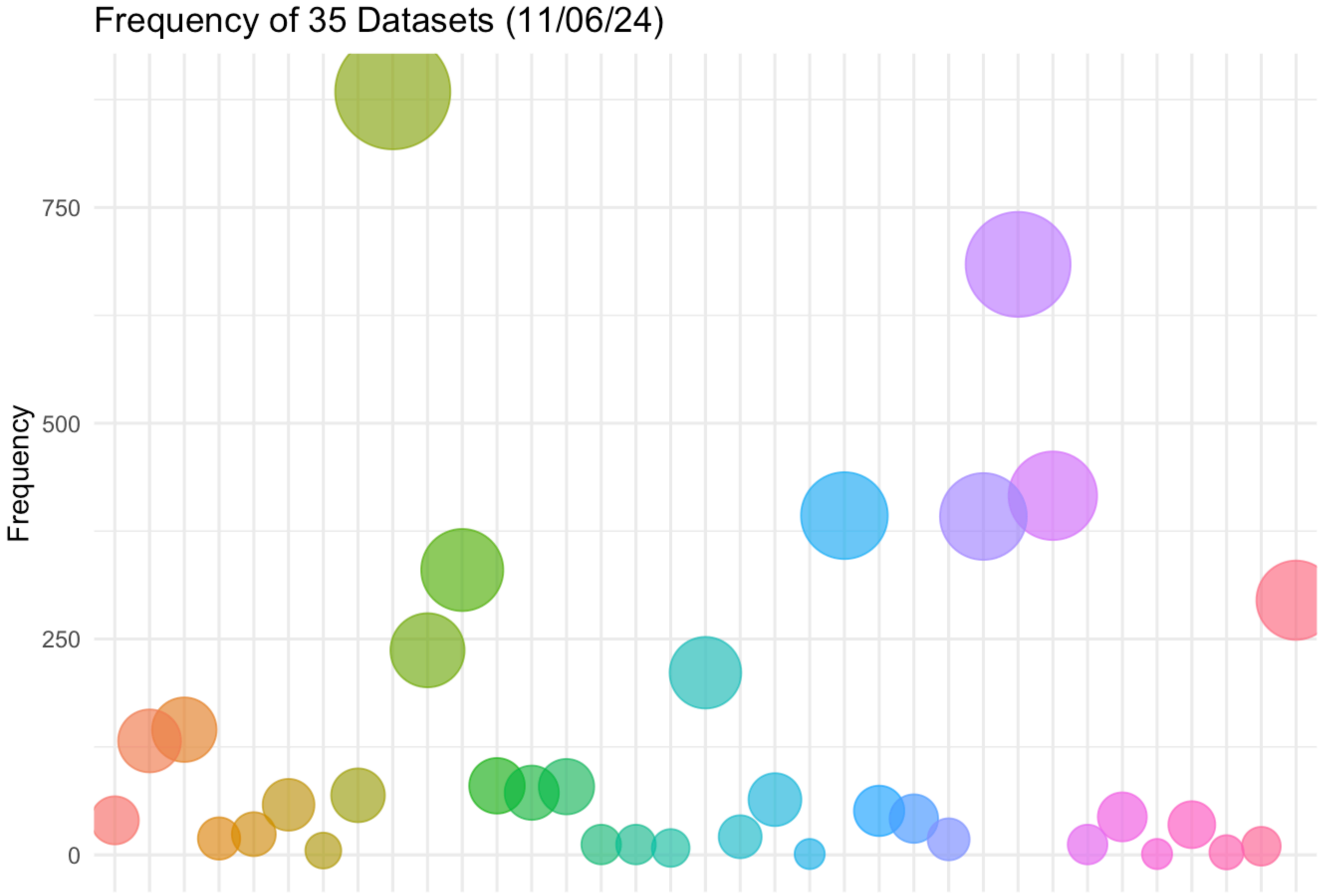
```
library(wordcloud)
word_freqs <- newmeta_update |>
#mutate(dataset_name = str_replace(dataset_name, "^([\\d]{4}_[A-Za-z]+)_.*", "\\1")) |>
group_by(dataset_name) |>
  summarize(Frequency = n()) |>
  arrange(desc(Frequency))
print(word_freqs)
```

```
## # A tibble: 35 × 2
##    dataset_name                     Frequency
##    <chr>                                <int>
##  1 2022_Ghose_CommunicationsBiology       884
##  2 2023_Nanostring_website_03             684
##  3 2023_Sorin_Nature                      416
##  4 2023_Karimi_Nature                     393
##  5 2023_Nanostring_website_02             392
##  6 2022_Liu_LI                            330
##  7 2024_Xu_ScienceAdvances_01             295
##  8 2022_Kim_NatureMethods_02              237
##  9 2023_Greenbaum_Nature_01               211
## 10 2021_Burlingame_PubMed                 145
## # ℹ 25 more rows
```

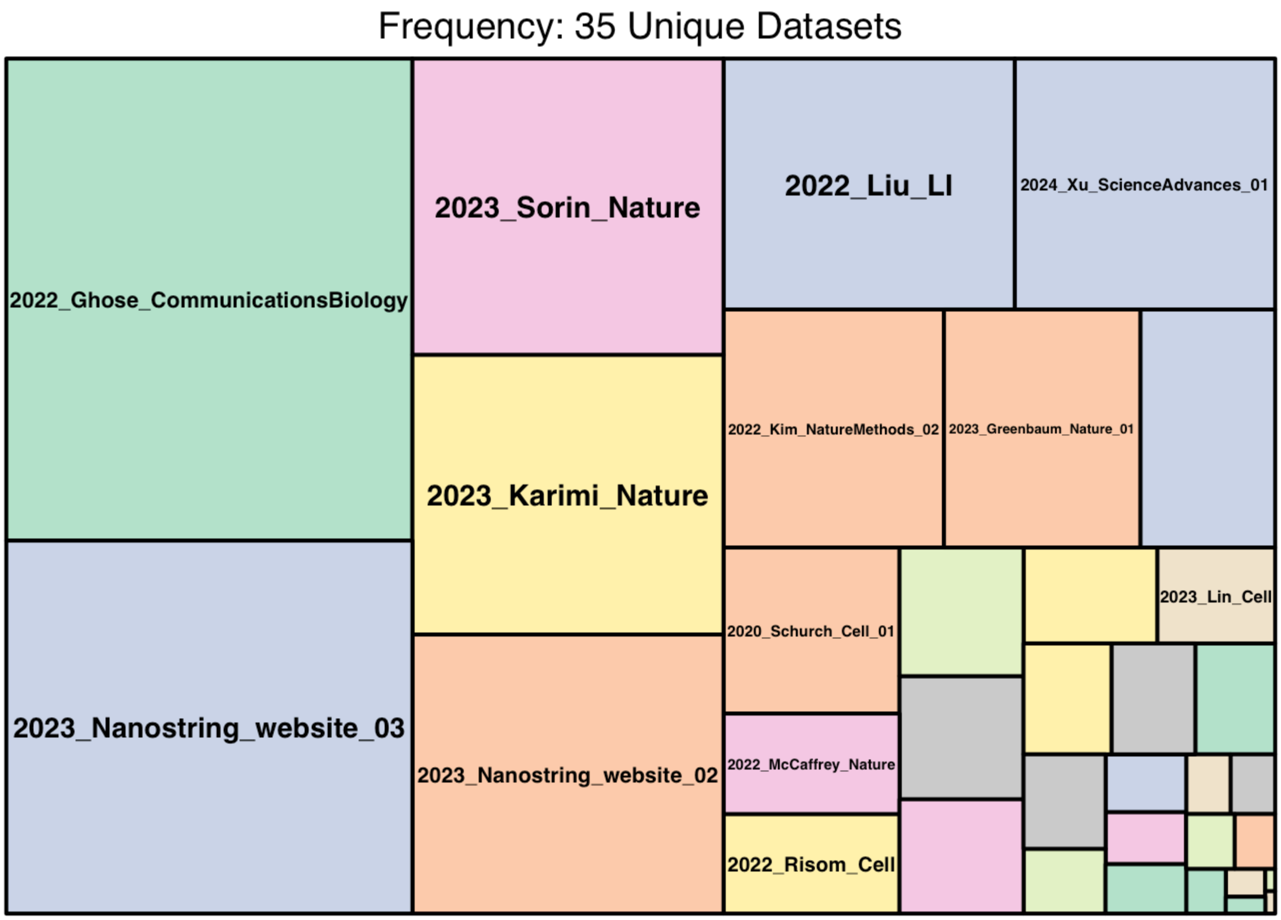
```
wordcloud(words = word_freqs$dataset_name, freq = word_freqs$Frequency, min.freq =1, sca
le = c(1.3,0.5), random.order =FALSE, rot.per =.25)
```



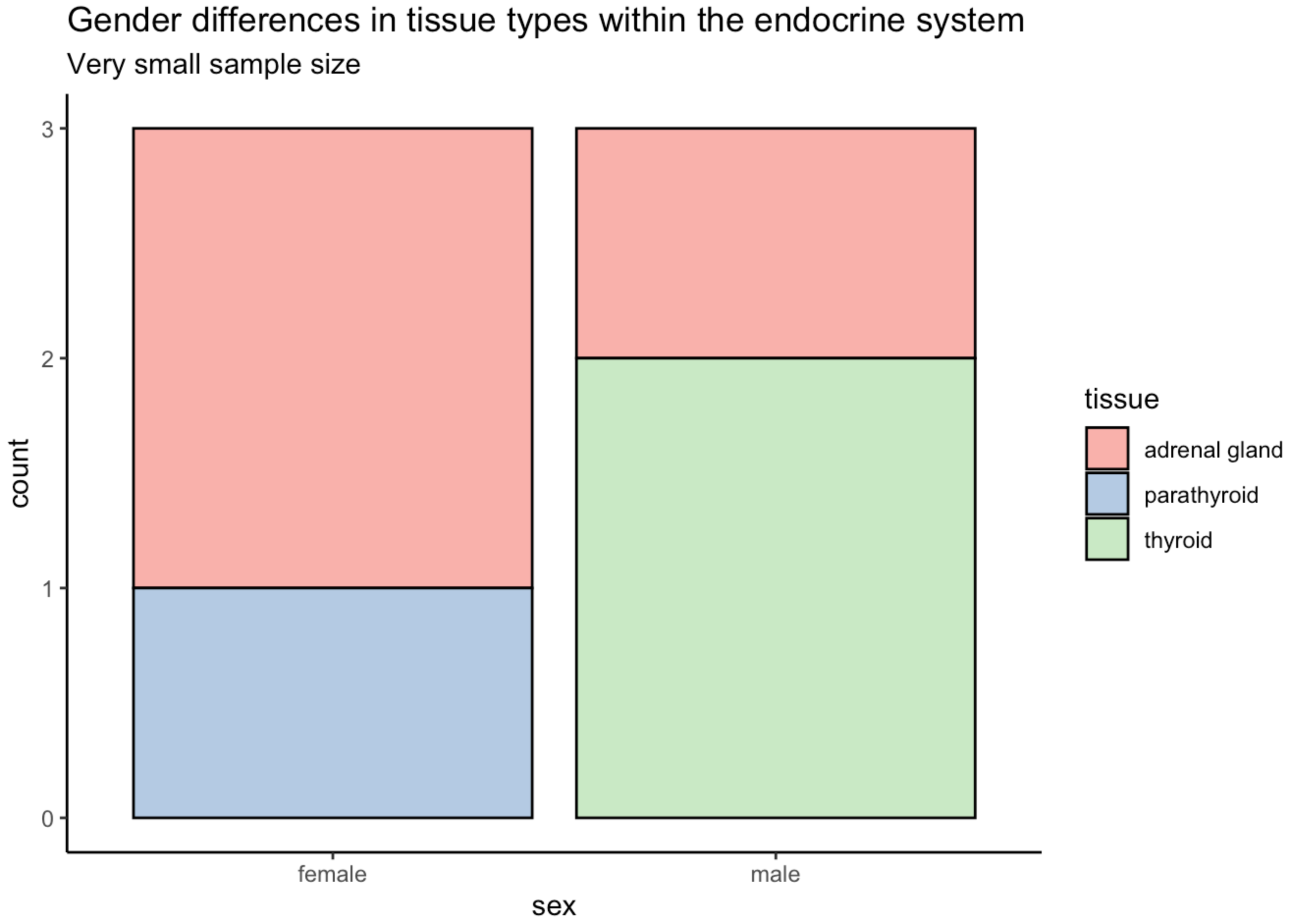
```
word_freqs |>
  ggplot(
    aes(x = dataset_name, y = Frequency, size = Frequency, color = dataset_name)) +
  geom_point(alpha =.7) +
  scale_size(range = c(5,20)) +
  theme_minimal() +
  theme(legend.position ="none", axis.text.x = element_blank(), axis.ticks.x = element_b
lank()) +
  labs(
    title ="Frequency of 35 Datasets (11/06/24)",
    x ="")
```



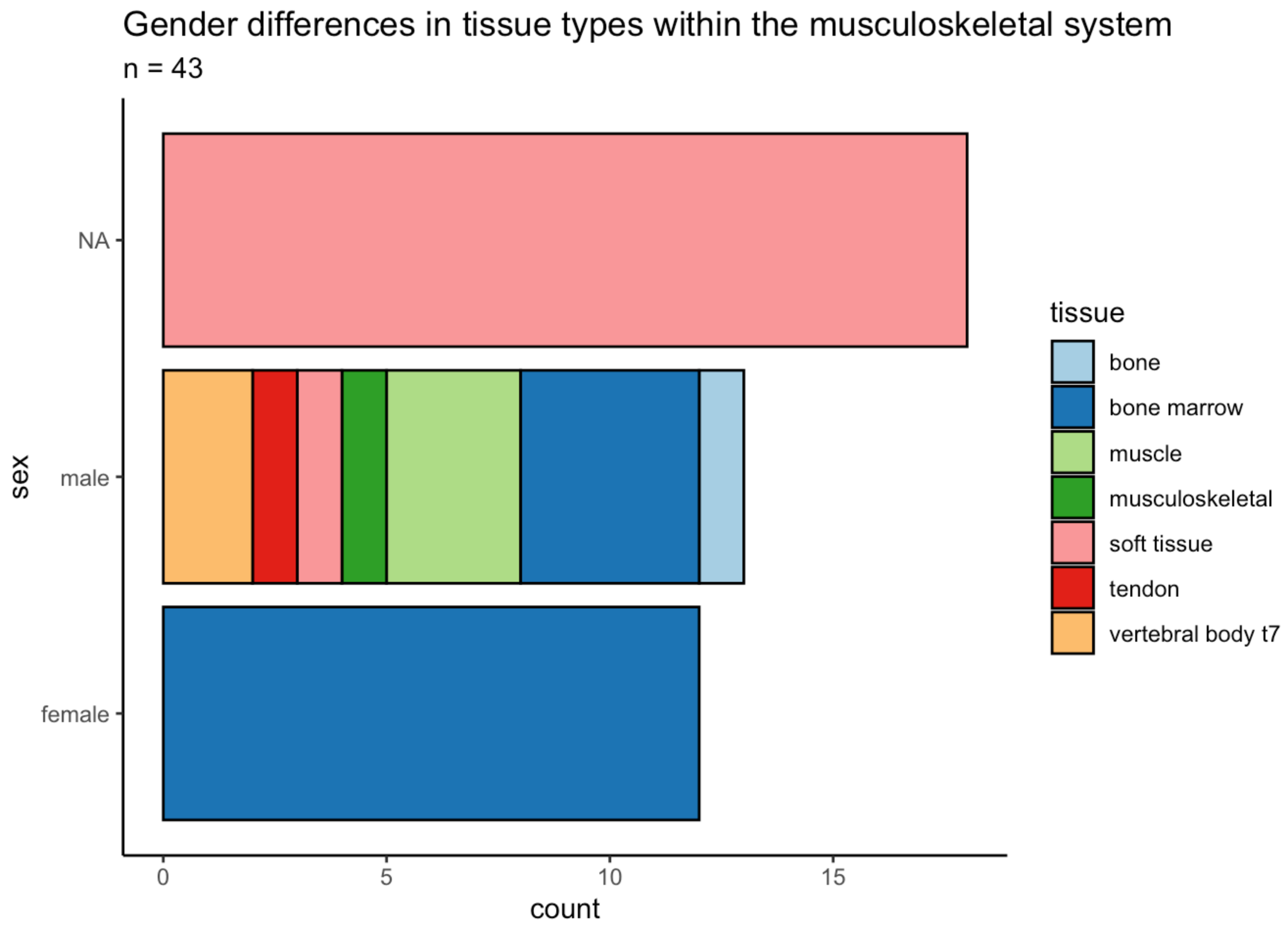
```
library(treemap)
treemap(word_freqs, index = c("dataset_name"), vSize ="Frequency", vColor ="Frequency",
title ="Frequency: 35 Unique Datasets", palette ="Pastel2", draw =TRUE)
```



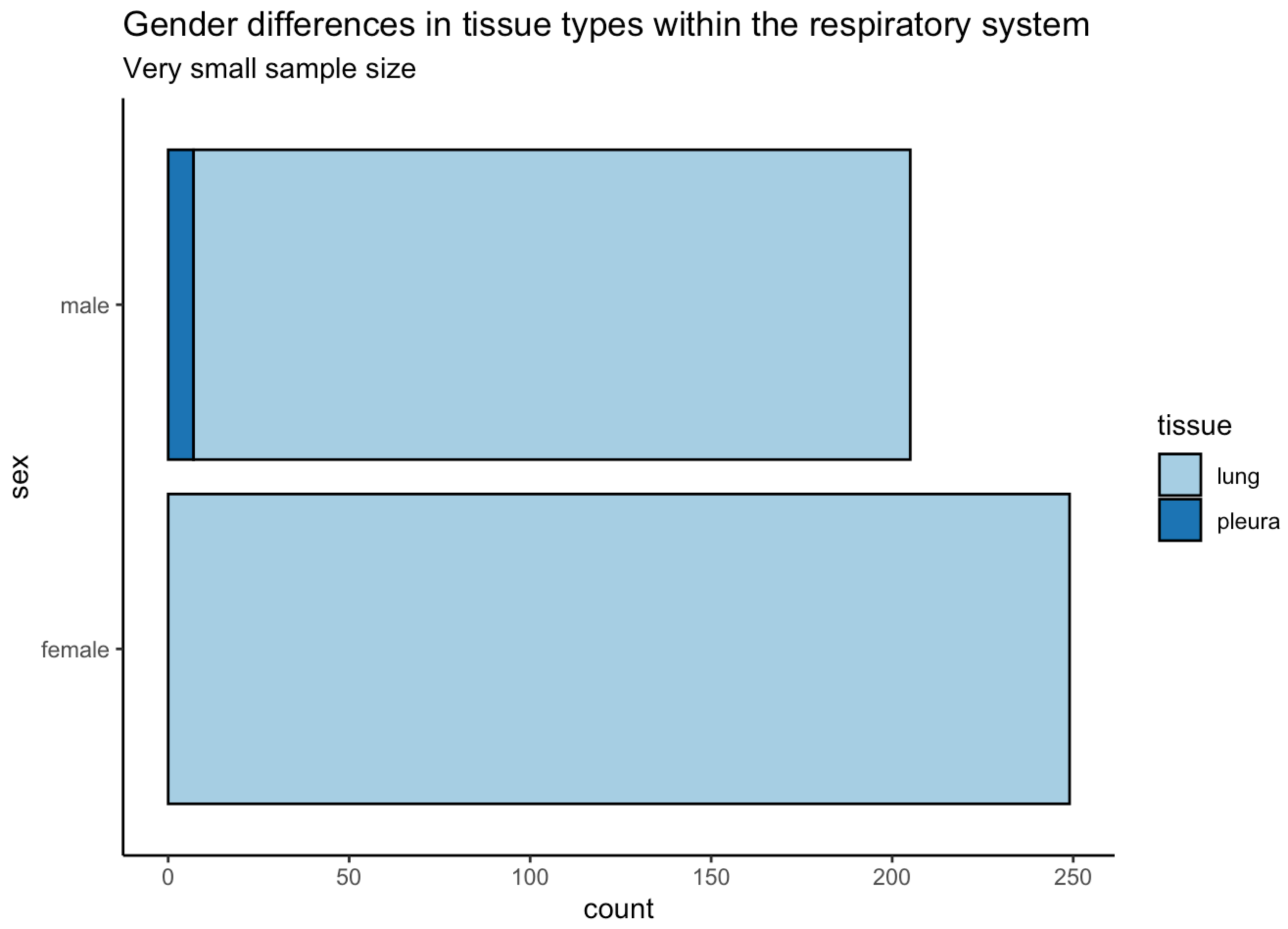
```
newmeta_update |>
  filter(large_tissue_unit =="Endocrine System") |>
  ggplot( aes(x = sex, fill = tissue)) +
  geom_bar(color="black") +
#scale_fill_manual(values =brewer.pal(8, "Paired")) +
scale_fill_brewer(palette ="Pastel1") +
  labs(
    title ="Gender differences in tissue types within the endocrine system",
    subtitle ="Very small sample size") + theme_classic()
```



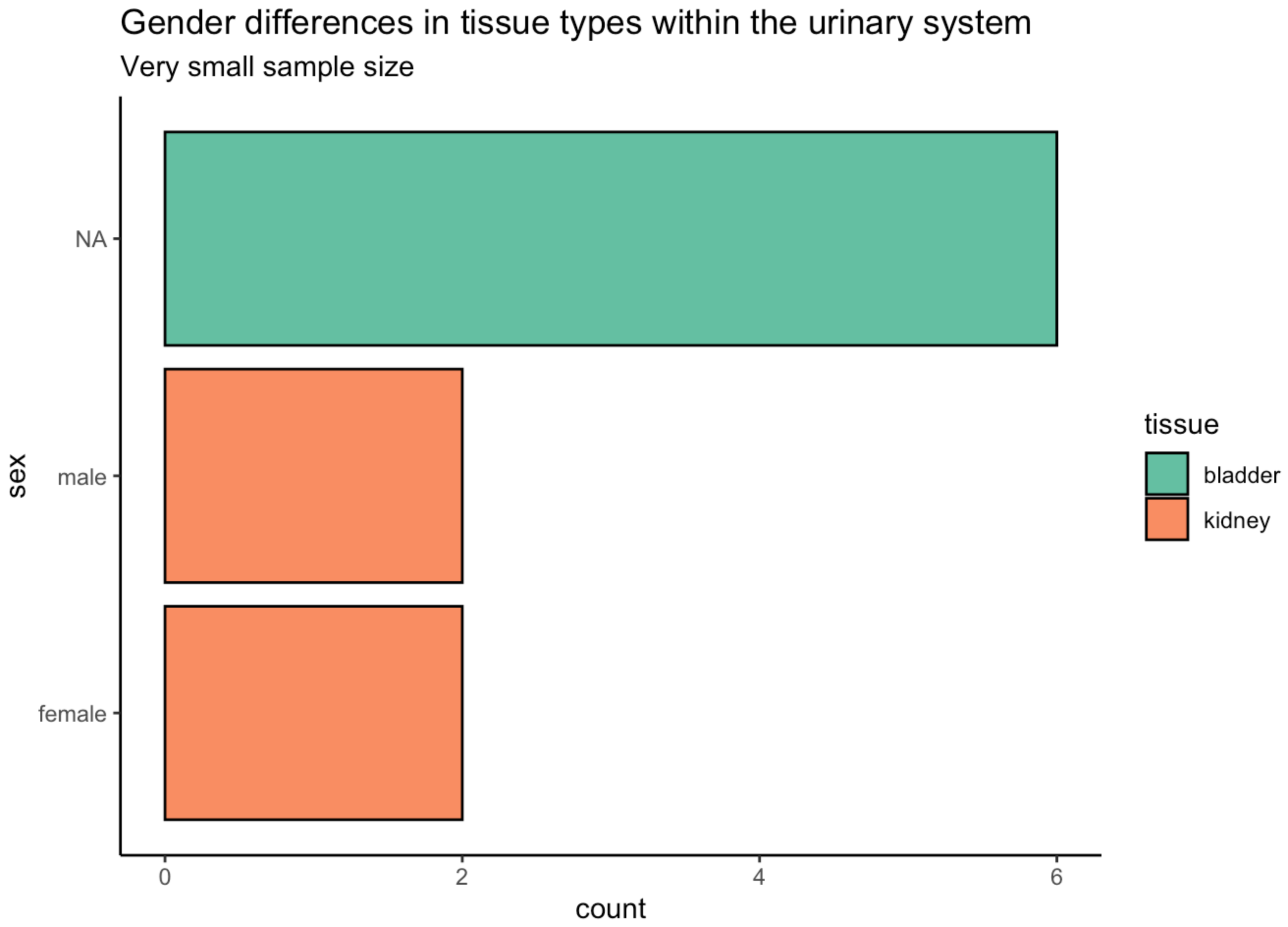
```
newmeta_update |>
  filter(large_tissue_unit =="Musculoskeletal System") |>
  ggplot( aes(x = sex, fill = tissue)) +
  geom_bar(color="black") +
#scale_fill_manual(values =brewer.pal(8, "Paired")) +
scale_fill_brewer(palette ="Paired") +
  labs(
    title ="Gender differences in tissue types within the musculoskeletal system",
    subtitle ="n = 43") +
  theme_classic() +
  coord_flip()
```



```
newmeta_update |>
  filter(large_tissue_unit =="Respiratory System") |>
  ggplot( aes(x = sex, fill = tissue)) +
  geom_bar(color="black") +
#scale_fill_manual(values =brewer.pal(8, "Paired")) +
scale_fill_brewer(palette ="Paired") +
  labs(
    title ="Gender differences in tissue types within the respiratory system",
    subtitle ="Very small sample size") +
  theme_classic() +
  coord_flip()
```



```
newmeta_update |>
  filter(large_tissue_unit =="Urinary System") |>
  ggplot( aes(x = sex, fill = tissue)) +
  geom_bar(color="black") +
#scale_fill_manual(values =brewer.pal(8, "Paired")) +
scale_fill_brewer(palette ="Set2") +
  labs(
    title ="Gender differences in tissue types within the urinary system",
    subtitle ="Very small sample size") +
  theme_classic() +
  coord_flip()
```


Gender differences in tissue types within the urinary system
Very small sample size
sex
NA
male
female
count
0
2
4
6
tissue
bladder
kidney