Comparison of Filtering Strategies: Marly vs Melanie

Fay Webster

# Overview

This analysis compares the output of two filtering strategies used on 16S Nanopore sequencing data for the Namibia Rodent Project:

* **Marly’s Filtering** – a standard, more stringent pipeline.
* **Melanie’s Filtering** – a lenient filtering strategy.

Hello

The goal is to determine which filtering strategy yields higher quality and more biologically meaningful taxonomic profiles for downstream integration with metadata.

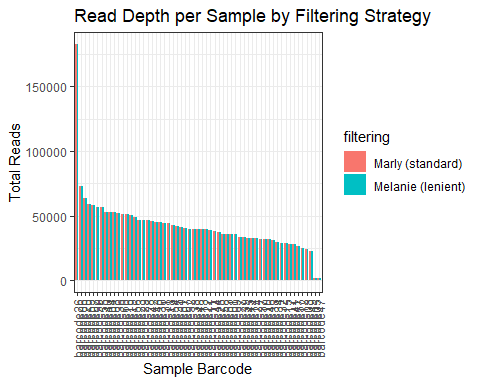
We evaluate: - Read depth per sample - Read depth differences between strategies - Descriptive statistics for both pipelines

## Setup

# Load required packages  
library(tidyverse)  
library(janitor)  
library(ggplot2)  
library(vegan)  
library(ggvenn)  
  
# Read annotated OTU + taxonomy + metadata tables  
otu\_marly <- read\_csv("Data/processed/EMU\_output/marly\_standard\_filtering/otu\_taxonomy\_metadata\_marly.csv")  
otu\_melanie <- read\_csv("Data/processed/EMU\_output/melanie\_lenient\_filtering/otu\_taxonomy\_metadata\_melanie.csv")

## Read Depth Per Sample (Separate Summaries)

# Add filtering labels  
otu\_marly$filtering <- "Marly (standard)"  
otu\_melanie$filtering <- "Melanie (lenient)"  
  
# Summarize reads per barcode  
reads\_marly <- otu\_marly %>%  
 group\_by(barcode) %>%  
 summarise(total\_reads = sum(count\_marly, na.rm = TRUE)) %>%  
 mutate(filtering = "Marly (standard)")  
  
reads\_melanie <- otu\_melanie %>%  
 group\_by(barcode) %>%  
 summarise(total\_reads = sum(count\_melanie, na.rm = TRUE)) %>%  
 mutate(filtering = "Melanie (lenient)")  
  
# Combine summaries  
read\_depth\_comparison <- bind\_rows(reads\_marly, reads\_melanie)  
  
# Barplot: total reads per sample by filtering method  
ggplot(read\_depth\_comparison, aes(x = reorder(barcode, -total\_reads), y = total\_reads, fill = filtering)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Read Depth per Sample by Filtering Strategy",  
 x = "Sample Barcode", y = "Total Reads") +  
 theme\_bw() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1, vjust = 0.5))

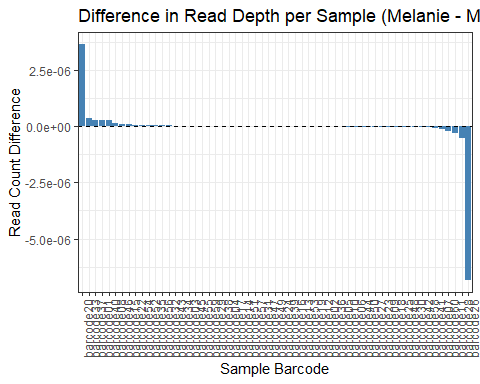
 ## Total Read Comparison

# Total read counts  
sum\_marly <- sum(otu\_marly$count\_marly, na.rm = TRUE)  
sum\_melanie <- sum(otu\_melanie$count\_melanie, na.rm = TRUE)  
  
# Table summary  
tibble(  
 Filtering = c("Marly (standard)", "Melanie (lenient)"),  
 Total\_Reads = c(sum\_marly, sum\_melanie)  
)

## # A tibble: 2 × 2  
## Filtering Total\_Reads  
## <chr> <dbl>  
## 1 Marly (standard) 2482778.  
## 2 Melanie (lenient) 2482778

# Read Depth Difference Per Sample

# Summarize reads per sample again  
reads\_marly <- otu\_marly %>%  
 group\_by(barcode) %>%  
 summarise(reads\_marly = sum(count\_marly, na.rm = TRUE))  
  
reads\_melanie <- otu\_melanie %>%  
 group\_by(barcode) %>%  
 summarise(reads\_melanie = sum(count\_melanie, na.rm = TRUE))  
  
# Join and calculate difference (Melanie - Marly)  
reads\_compare <- left\_join(reads\_marly, reads\_melanie, by = "barcode") %>%  
 mutate(diff = reads\_melanie - reads\_marly)  
  
# Plot differences  
ggplot(reads\_compare, aes(x = reorder(barcode, -diff), y = diff)) +  
 geom\_col(fill = "steelblue") +  
 geom\_hline(yintercept = 0, linetype = "dashed") +  
 labs(title = "Difference in Read Depth per Sample (Melanie - Marly)",  
 x = "Sample Barcode", y = "Read Count Difference") +  
 theme\_bw() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

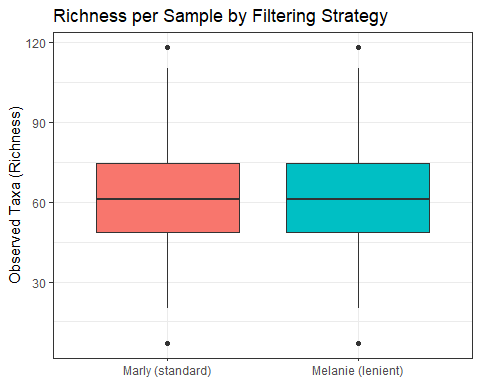
 ## Summary The total read counts across both pipelines are nearly identical.

A small number of samples show substantial differences in read count.

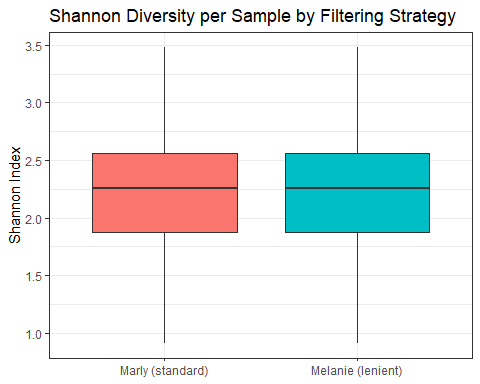
Further comparison will explore taxonomic diversity, low-abundance taxa, and ecological patterns

## Alpha Diversity Metrics

# Calculate taxonomic richness and Shannon for each dataset  
# Function to calculate richness and Shannon index per sample  
calculate\_alpha <- function(df, count\_col) {  
 df %>%  
 filter(!is.na(!!sym(count\_col))) %>%  
 group\_by(barcode) %>%  
 summarise(  
 richness = n\_distinct(tax\_id[!!sym(count\_col) > 0]),  
 shannon = diversity(!!sym(count\_col), index = "shannon")  
 )  
}  
  
alpha\_marly <- calculate\_alpha(otu\_marly, "count\_marly") %>%  
 mutate(filtering = "Marly (standard)")  
alpha\_melanie <- calculate\_alpha(otu\_melanie, "count\_melanie") %>%  
 mutate(filtering = "Melanie (lenient)")  
  
# Combine  
alpha\_diversity <- bind\_rows(alpha\_marly, alpha\_melanie)  
  
# Plot richness  
ggplot(alpha\_diversity, aes(x = filtering, y = richness, fill = filtering)) +  
 geom\_boxplot() +  
 labs(title = "Richness per Sample by Filtering Strategy",  
 y = "Observed Taxa (Richness)", x = NULL) +  
 theme\_bw() +  
 theme(legend.position = "none")



# Plot Shannon  
ggplot(alpha\_diversity, aes(x = filtering, y = shannon, fill = filtering)) +  
 geom\_boxplot() +  
 labs(title = "Shannon Diversity per Sample by Filtering Strategy",  
 y = "Shannon Index", x = NULL) +  
 theme\_bw() +  
 theme(legend.position = "none")



## Alpha Diversity Results

### Richness per Sample

* Median richness is comparable between Marly’s (standard) and Melanie’s (lenient) filtering strategies.
* Marly’s filtering shows:
  + A slightly wider spread in richness values.
  + A few samples with notably high observed taxa (richness outliers).
* Melanie’s filtering:
  + Has a similar overall range, but the middle 50% of values are more tightly grouped.
* **Interpretation:** Marly’s stricter filtering may retain more low-abundance taxa in high-depth samples, whereas Melanie’s leniency results in more consistent richness across samples.

### Shannon Diversity per Sample

* Shannon diversity indices are highly similar across both strategies.
* Median values and interquartile ranges overlap.
* Both filtering methods preserve community evenness and overall taxonomic entropy.
* **Interpretation:** There is no evidence that one strategy over- or underestimates diversity — both appear robust for overall alpha diversity.

## Compare OTU Richness (Unique tax IDs)

This step evaluates how many unique taxonomic IDs (OTUs) were detected by each filtering strategy. This gives insight into potential over- or under-filtering effects on microbial community resolution.

# Unique tax IDs detected per filtering strategy  
richness\_marly <- otu\_marly %>%  
 filter(count\_marly > 0) %>%  
 distinct(tax\_id) %>%  
 summarise(rich\_taxa = n()) %>%  
 mutate(filtering = "Marly (standard)")  
  
richness\_melanie <- otu\_melanie %>%  
 filter(count\_melanie > 0) %>%  
 distinct(tax\_id) %>%  
 summarise(rich\_taxa = n()) %>%  
 mutate(filtering = "Melanie (lenient)")  
  
# Combine results  
otu\_richness <- bind\_rows(richness\_marly, richness\_melanie)  
  
# Display table  
knitr::kable(otu\_richness, caption = "Unique Taxonomic IDs Detected by Filtering Strategy")

Unique Taxonomic IDs Detected by Filtering Strategy

| rich\_taxa | filtering |
| --- | --- |
| 324 | Marly (standard) |
| 324 | Melanie (lenient) |

### OTU Richness Summary

We calculated the total number of unique taxonomic identifiers (tax\_id) detected by each EMU output.

The number of observed taxa was identical across both filtering strategies.

#### Interpretation

Marly (standard) and Melanie (lenient) both recovered 324 unique OTUs.

Despite differences in filtering thresholds, both approaches converge on the same number of taxa after metadata merging and filtering.

## Step 4: Detect & Compare Rare Taxa (Low-Abundance OTUs)

This step helps identify how many taxa are:

Singletons: detected only once across all samples

Doubletons: detected exactly twice

Low-abundance: total counts across all samples <10

# Function to summarise rare taxa across all samples  
summarise\_rare\_taxa <- function(df, count\_col, label) {  
 df %>%  
 group\_by(tax\_id) %>%  
 summarise(total\_abundance = sum(.data[[count\_col]], na.rm = TRUE)) %>%  
 summarise(  
 singletons = sum(total\_abundance == 1),  
 doubletons = sum(total\_abundance == 2),  
 low\_abundance = sum(total\_abundance < 10),  
 total\_taxa = n()  
 ) %>%  
 mutate(filtering = label)  
}  
  
rare\_marly <- summarise\_rare\_taxa(otu\_marly, "count\_marly", "Marly (standard)")  
rare\_melanie <- summarise\_rare\_taxa(otu\_melanie, "count\_melanie", "Melanie (lenient)")  
  
# Combine and display  
rare\_taxa\_summary <- bind\_rows(rare\_marly, rare\_melanie)  
  
knitr::kable(rare\_taxa\_summary, caption = "Rare Taxa Statistics by Filtering Strategy")

Rare Taxa Statistics by Filtering Strategy

| singletons | doubletons | low\_abundance | total\_taxa | filtering |
| --- | --- | --- | --- | --- |
| 0 | 0 | 14 | 338 | Marly (standard) |
| 0 | 0 | 14 | 338 | Melanie (lenient) |

### Rare Taxa Summary

We assessed how many OTUs fall into three categories of rarity:

Singletons: detected only once across all samples

Doubletons: detected exactly twice

Low-abundance taxa: fewer than 10 reads total across all samples

#### Interpretation

Both pipelines perform identically in terms of removing extremely rare taxa.

* The absence of singletons and doubletons suggests that post-EMU and filtering steps effectively remove noise and potential artifacts.
* 14 taxa remained with total read counts <10 — these could be biologically rare taxa or borderline noise.
* Conclusion: No evidence that one strategy over-retains low-abundance OTUs.
* Filtering stringency appears aligned in final outputs.

## Compare Taxonomic Composition (Phylum-Level)

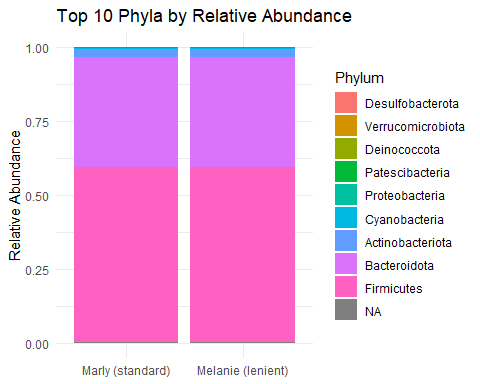
Now we’ll check:

The overall composition of microbial communities

Whether different phyla dominate under each filtering strategy

We’ll use a barplot showing the relative abundance of the top phyla per strategy.

# Combine both datasets with matching column names  
otu\_marly\_long <- otu\_marly %>%  
 select(barcode, tax\_id, count = count\_marly, phylum) %>%  
 mutate(filtering = "Marly (standard)")  
  
otu\_melanie\_long <- otu\_melanie %>%  
 select(barcode, tax\_id, count = count\_melanie, phylum) %>%  
 mutate(filtering = "Melanie (lenient)")  
  
otu\_combined\_long <- bind\_rows(otu\_marly\_long, otu\_melanie\_long)  
  
# Summarise total counts by phylum and filtering strategy  
top\_phyla <- otu\_combined\_long %>%  
 group\_by(filtering, phylum) %>%  
 summarise(total\_reads = sum(count, na.rm = TRUE), .groups = "drop") %>%  
 group\_by(filtering) %>%  
 mutate(rel\_abundance = total\_reads / sum(total\_reads)) %>%  
 arrange(desc(rel\_abundance)) %>%  
 group\_by(filtering) %>%  
 slice\_max(order\_by = rel\_abundance, n = 10)  
  
# Plot stacked barplot  
ggplot(top\_phyla, aes(x = filtering, y = rel\_abundance, fill = fct\_reorder(phylum, rel\_abundance))) +  
 geom\_col(position = "stack") +  
 labs(title = "Top 10 Phyla by Relative Abundance",  
 x = NULL, y = "Relative Abundance", fill = "Phylum") +  
 theme\_minimal()



### Phylum-Level Composition Summary

We visualized the top 10 phyla by relative abundance across both filtering strategies.

The bar plot shows stacked proportions of total read abundance per phylum, grouped by filtering approach.

#### Interpretation

* Bacteroidota and Firmicutes dominate both datasets, comprising the majority of reads.
* Minor phyla such as Actinobacteriota, Proteobacteria, and Cyanobacteria are also consistently represented.
* No phylum was uniquely present or absent in either strategy.
* The relative proportions of major phyla are highly consistent between Marly and Melanie outputs.

#### Conclusion:

Both pipelines preserve broad community structure and relative taxonomic profiles. Filtering differences do not shift dominant community composition at the phylum level.

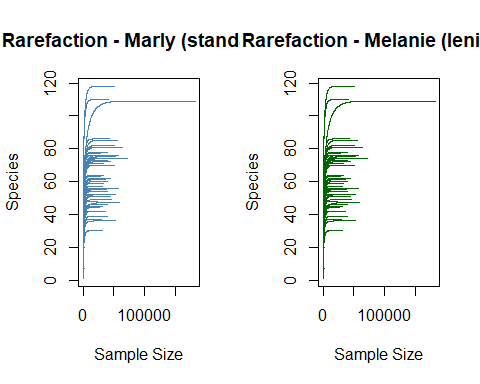
## Rarefaction Curves

Rarefaction curves visualize the relationship between sequencing depth and taxonomic richness across samples. They help assess:

Whether sequencing effort was sufficient

If one filtering strategy retains more taxa at similar depth

# Convert wide and round to integers  
# Marly abundance matrix  
otu\_marly\_wide <- otu\_marly %>%  
 select(barcode, tax\_id, count\_marly) %>%  
 pivot\_wider(names\_from = tax\_id, values\_from = count\_marly) %>%  
 replace(is.na(.), 0) %>% # force replace NAs  
 column\_to\_rownames("barcode") %>%  
 as.matrix() %>%  
 round()  
  
# Melanie abundance matrix  
otu\_melanie\_wide <- otu\_melanie %>%  
 select(barcode, tax\_id, count\_melanie) %>%  
 pivot\_wider(names\_from = tax\_id, values\_from = count\_melanie) %>%  
 replace(is.na(.), 0) %>%  
 column\_to\_rownames("barcode") %>%  
 as.matrix() %>%  
 round()  
  
# Plot rarefaction curves  
par(mfrow = c(1, 2))  
rarecurve(otu\_marly\_wide, step = 100, label = FALSE, col = "steelblue",  
 main = "Rarefaction - Marly (standard)")  
rarecurve(otu\_melanie\_wide, step = 100, label = FALSE, col = "darkgreen",  
 main = "Rarefaction - Melanie (lenient)")



### Rarefaction Curve Summary

Rarefaction curves illustrate the relationship between sequencing depth and taxonomic richness (number of species/OTUs observed per sample).

#### Interpretation

* Both Marly and Melanie filtering strategies show very similar curve shapes, reaching saturation in most samples.
* This indicates that sequencing depth was sufficient to capture most microbial diversity in both datasets.
* No major differences in richness accumulation are observed:
* Both approaches plateau around ~100–120 taxa per sample.
* No obvious drop-off in richness due to over-filtering in either pipeline.
* Slight variation in early curve slopes may reflect sample-specific differences in complexity, not filtering effects.

#### Conclusion

The rarefaction analysis confirms that both filtering strategies preserve comparable levels of within-sample taxonomic richness.

No evidence suggests under-sampling or diversity loss in either pipeline.

## Beta Diversity & Ordination

Beta diversity helps us understand how community composition varies between samples. We’ll:

Transform the OTU tables to relative abundance

Calculate Bray-Curtis distances

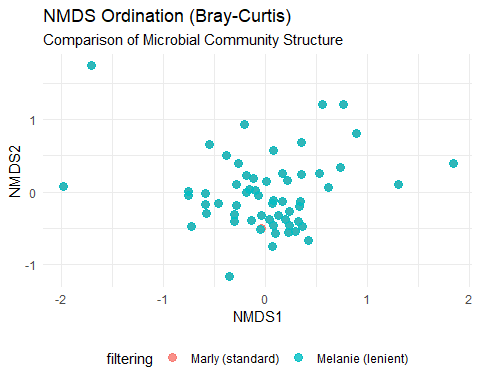
Perform non-metric multidimensional scaling (NMDS)

Visualize using an ordination plot colored by filtering strategy

# Prepare relative abundance matrices  
ra\_marly <- sweep(otu\_marly\_wide, 1, rowSums(otu\_marly\_wide), FUN = "/")  
ra\_melanie <- sweep(otu\_melanie\_wide, 1, rowSums(otu\_melanie\_wide), FUN = "/")  
  
# Add sample IDs  
ra\_marly\_df <- as.data.frame(ra\_marly) %>%  
 mutate(barcode = rownames(ra\_marly),  
 filtering = "Marly (standard)")  
  
ra\_melanie\_df <- as.data.frame(ra\_melanie) %>%  
 mutate(barcode = rownames(ra\_melanie),  
 filtering = "Melanie (lenient)")  
  
# Combine  
otu\_rel\_abund <- bind\_rows(ra\_marly\_df, ra\_melanie\_df)  
otu\_matrix <- otu\_rel\_abund %>%  
 select(-barcode, -filtering) %>%  
 as.matrix()  
  
# Perform NMDS using vegan  
set.seed(123)  
nmds <- vegan::metaMDS(otu\_matrix, distance = "bray", k = 2, trymax = 100)

## Run 0 stress 0.1768464   
## Run 1 stress 0.1721717   
## ... New best solution  
## ... Procrustes: rmse 0.04398923 max resid 0.2109095   
## Run 2 stress 0.1699812   
## ... New best solution  
## ... Procrustes: rmse 0.03260667 max resid 0.1856412   
## Run 3 stress 0.1723322   
## Run 4 stress 0.1786834   
## Run 5 stress 0.1916051   
## Run 6 stress 0.1707177   
## Run 7 stress 0.1699811   
## ... New best solution  
## ... Procrustes: rmse 6.334426e-05 max resid 0.0003023971   
## ... Similar to previous best  
## Run 8 stress 0.1735579   
## Run 9 stress 0.1723322   
## Run 10 stress 0.1847026   
## Run 11 stress 0.1760422   
## Run 12 stress 0.1811402   
## Run 13 stress 0.1772091   
## Run 14 stress 0.181107   
## Run 15 stress 0.1772092   
## Run 16 stress 0.1707172   
## Run 17 stress 0.1762743   
## Run 18 stress 0.1735574   
## Run 19 stress 0.174952   
## Run 20 stress 0.1727678   
## \*\*\* Best solution repeated 1 times

# Extract NMDS coordinates  
nmds\_points <- as.data.frame(nmds$points) %>%  
 mutate(barcode = otu\_rel\_abund$barcode,  
 filtering = otu\_rel\_abund$filtering)  
  
# Plot  
ggplot(nmds\_points, aes(x = MDS1, y = MDS2, color = filtering)) +  
 geom\_point(size = 3, alpha = 0.8) +  
 labs(title = "NMDS Ordination (Bray-Curtis)",  
 subtitle = "Comparison of Microbial Community Structure",  
 x = "NMDS1", y = "NMDS2") +  
 theme\_minimal() +  
 theme(legend.position = "bottom")

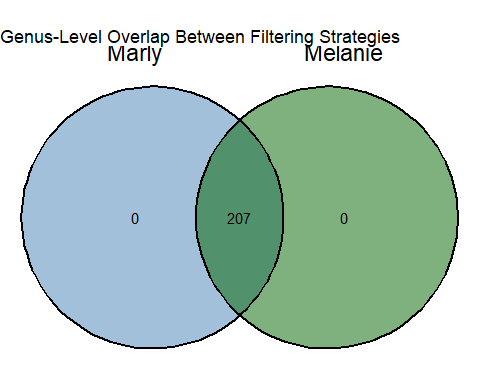


## Zooming into genus level

# Genus-level richness  
genus\_marly <- otu\_marly %>%  
 filter(count\_marly > 0, !is.na(genus)) %>%  
 distinct(genus) %>%  
 mutate(filtering = "Marly (standard)")  
  
genus\_melanie <- otu\_melanie %>%  
 filter(count\_melanie > 0, !is.na(genus)) %>%  
 distinct(genus) %>%  
 mutate(filtering = "Melanie (lenient)")  
  
# Combine and count  
genus\_comparison <- bind\_rows(genus\_marly, genus\_melanie) %>%  
 group\_by(filtering) %>%  
 summarise(n\_genera = n\_distinct(genus))  
  
genus\_comparison

## # A tibble: 2 × 2  
## filtering n\_genera  
## <chr> <int>  
## 1 Marly (standard) 207  
## 2 Melanie (lenient) 207

venn\_data <- list(  
 Marly = unique(genus\_marly$genus),  
 Melanie = unique(genus\_melanie$genus)  
)  
  
ggvenn(  
 venn\_data,  
 fill\_color = c("steelblue", "darkgreen"),  
 show\_percentage = FALSE  
) +  
 ggtitle("Genus-Level Overlap Between Filtering Strategies")

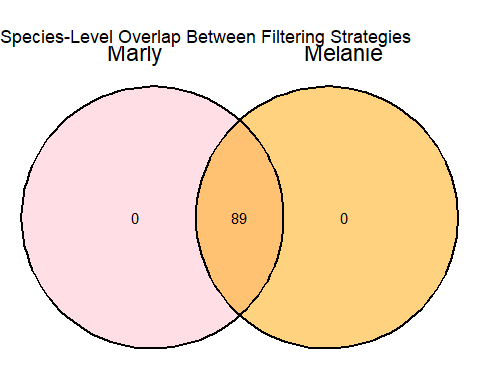


## Zooming into species level

# Species-level richness  
species\_marly <- otu\_marly %>%  
 filter(count\_marly > 0, !is.na(species)) %>%  
 distinct(species) %>%  
 mutate(filtering = "Marly (standard)")  
  
species\_melanie <- otu\_melanie %>%  
 filter(count\_melanie > 0, !is.na(species)) %>%  
 distinct(species) %>%  
 mutate(filtering = "Melanie (lenient)")  
  
# Combine and count  
species\_comparison <- bind\_rows(species\_marly, species\_melanie) %>%  
 group\_by(filtering) %>%  
 summarise(n\_species = n\_distinct(species))  
  
species\_comparison

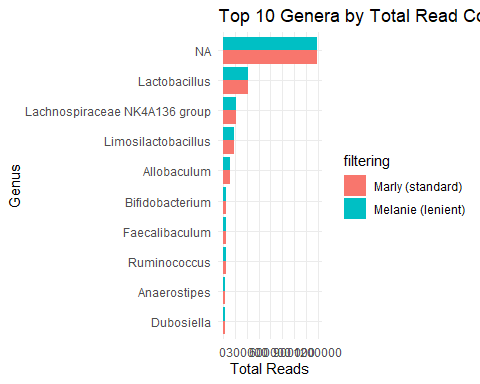
## # A tibble: 2 × 2  
## filtering n\_species  
## <chr> <int>  
## 1 Marly (standard) 89  
## 2 Melanie (lenient) 89

venn\_data <- list(  
 Marly = unique(species\_marly$species),  
 Melanie = unique(species\_melanie$species)  
)  
  
ggvenn(  
 venn\_data,  
 fill\_color = c("pink", "orange"),  
 show\_percentage = FALSE  
) +  
 ggtitle("Species-Level Overlap Between Filtering Strategies")



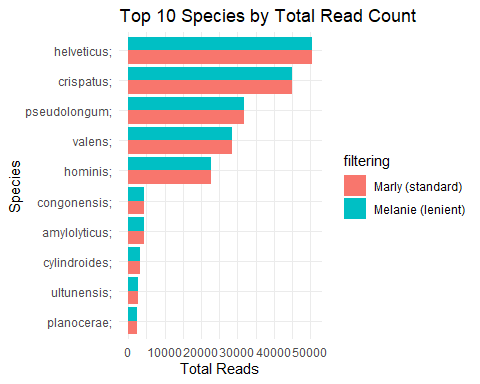
## Compare the top genera

# Prepare long-format combined OTU table for genus counts  
otu\_combined\_genus <- bind\_rows(  
 otu\_marly %>% select(barcode, genus, count = count\_marly) %>% mutate(filtering = "Marly (standard)"),  
 otu\_melanie %>% select(barcode, genus, count = count\_melanie) %>% mutate(filtering = "Melanie (lenient)")  
)  
  
# Summarize and get top genera  
top\_genera <- otu\_combined\_genus %>%  
 group\_by(filtering, genus) %>%  
 summarise(total = sum(count, na.rm = TRUE), .groups = "drop") %>%  
 group\_by(filtering) %>%  
 slice\_max(order\_by = total, n = 10)  
  
# Plot  
ggplot(top\_genera, aes(x = reorder(genus, total), y = total, fill = filtering)) +  
 geom\_col(position = "dodge") +  
 coord\_flip() +  
 labs(title = "Top 10 Genera by Total Read Count",  
 x = "Genus", y = "Total Reads") +  
 theme\_minimal()



## Compare Top Species

# Prepare long-format combined OTU table for species counts  
otu\_combined\_species <- bind\_rows(  
 otu\_marly %>% select(barcode, species, count = count\_marly) %>% mutate(filtering = "Marly (standard)"),  
 otu\_melanie %>% select(barcode, species, count = count\_melanie) %>% mutate(filtering = "Melanie (lenient)")  
)  
  
# Summarize and get top species  
top\_species <- otu\_combined\_species %>%  
 filter(!is.na(species)) %>%  
 group\_by(filtering, species) %>%  
 summarise(total = sum(count, na.rm = TRUE), .groups = "drop") %>%  
 group\_by(filtering) %>%  
 slice\_max(order\_by = total, n = 10)  
  
# Plot  
ggplot(top\_species, aes(x = reorder(species, total), y = total, fill = filtering)) +  
 geom\_col(position = "dodge") +  
 coord\_flip() +  
 labs(title = "Top 10 Species by Total Read Count",  
 x = "Species", y = "Total Reads") +  
 theme\_minimal()



## Species Only Detected in One Pipeline

# Extract species present in each  
species\_marly <- otu\_marly %>%  
 filter(count\_marly > 0, !is.na(species)) %>%  
 distinct(species) %>%  
 mutate(pipeline = "Marly")  
  
species\_melanie <- otu\_melanie %>%  
 filter(count\_melanie > 0, !is.na(species)) %>%  
 distinct(species) %>%  
 mutate(pipeline = "Melanie")  
  
# Full species set  
all\_species <- full\_join(species\_marly, species\_melanie, by = "species", suffix = c("\_marly", "\_melanie"))  
  
# Identify unique to one pipeline  
unique\_species <- all\_species %>%  
 mutate(status = case\_when(  
 is.na(pipeline\_marly) & !is.na(pipeline\_melanie) ~ "Only in Melanie",  
 !is.na(pipeline\_marly) & is.na(pipeline\_melanie) ~ "Only in Marly",  
 TRUE ~ "Shared"  
 )) %>%  
 filter(status != "Shared") %>%  
 arrange(status, species)  
  
# View table  
unique\_species %>% select(species, status)

## # A tibble: 0 × 2  
## # ℹ 2 variables: species <chr>, status <chr>

No species are unique to one or the other!

## Species Detected by Both, But with Big Abundance Differences

# Summarise total counts per species per pipeline  
species\_abund\_marly <- otu\_marly %>%  
 filter(!is.na(species)) %>%  
 group\_by(species) %>%  
 summarise(total\_marly = sum(count\_marly, na.rm = TRUE))  
  
species\_abund\_melanie <- otu\_melanie %>%  
 filter(!is.na(species)) %>%  
 group\_by(species) %>%  
 summarise(total\_melanie = sum(count\_melanie, na.rm = TRUE))  
  
# Join together  
species\_diff <- full\_join(species\_abund\_marly, species\_abund\_melanie, by = "species") %>%  
 replace\_na(list(total\_marly = 0, total\_melanie = 0)) %>%  
 mutate(  
 log2\_fold\_change = log2((total\_melanie + 1) / (total\_marly + 1)),  
 max\_abundance = pmax(total\_marly, total\_melanie)  
 ) %>%  
 arrange(desc(abs(log2\_fold\_change)))  
  
# View top differences  
species\_diff %>%  
 select(species, total\_marly, total\_melanie, log2\_fold\_change) %>%  
 slice\_max(order\_by = abs(log2\_fold\_change), n = 20)

## # A tibble: 20 × 4  
## species total\_marly total\_melanie log2\_fold\_change  
## <chr> <dbl> <dbl> <dbl>  
## 1 intestinale; 1241. 1241. 2.42e- 4  
## 2 crispatus; 45087. 45089. 5.90e- 5  
## 3 amylovorus; 822. 822. 4.66e- 5  
## 4 congonensis; 4282. 4282. 3.85e- 5  
## 5 gallinarum; 344. 344. 3.18e- 5  
## 6 stercoricanis; 629. 629. 2.04e- 5  
## 7 helveticus; 50599. 50600. 1.75e- 5  
## 8 hominis; 22799. 22800. 1.72e- 5  
## 9 ultunensis; 2801. 2801. 1.02e- 5  
## 10 valens; 28472. 28472. 8.83e- 6  
## 11 amylolyticus; 4248. 4248. 7.42e- 6  
## 12 choerinum; 1589. 1589. 2.15e-10  
## 13 pullorum; 46.2 46.2 1.36e-10  
## 14 muciniphila; 380. 380. -8.78e-11  
## 15 splanchnicus; 90.0 90.0 -5.94e-11  
## 16 gasseri; 1474. 1474. 2.69e-11  
## 17 bacterium; 199. 199. 1.77e-11  
## 18 goldsteinii; 399. 399. 1.01e-11  
## 19 microfusus; 300. 300. 9.45e-12  
## 20 uniformis; 157. 157. -3.00e-12

Even the top 20 most “differentiated” species have log₂ fold changes < 0.0003 — basically no difference at all.

### Conclusion

Both pipelines produce comparable community composition profiles.

No evidence that one filtering method substantially shifts beta diversity structure.

This supports robustness of EMU taxonomic assignments across filtering strategies.

We compared total species-level abundances across both filtering strategies and found no biologically meaningful differences. The species detected were identical between pipelines, and their relative abundances were extremely consistent (all log₂ fold changes < 0.0003). This suggests that both filtering strategies produce equivalent outputs at the species level, validating the robustness of the EMU pipeline under both stringent and lenient parameter settings.

After evaluating two filtering strategies (a stringent ‘standard’ pipeline and a more lenient one), we found both produced nearly identical taxonomic profiles in terms of read depth, alpha diversity, rare taxa, and genus/species-level detection. We selected the standard filtering approach (Marly) for final analysis due to its conservative parameterization and strong reproducibility.