Vinícius Bonetti Franceschi
Teste Neoprospecta

Etapa 2: Análise de dados

2.1) Gráfico de barras com a contagem absoluta das 50 bactérias mais abundantes, agrupadas por tempo (dia após o desmame);

Para esta etapa foi utilizado o pacote **phyloseq,** o qual recebe matrizes com as OTUs, os táxons em seus níveis de taxonomia e os metadados da amostra. Com essas informações conseguimos ordenar os 50 táxons mais abundantes nos níveis de taxonomia desejados e plotar o gráfico.

Foram gerados dois gráficos (Fig. 1 e Fig. 2).

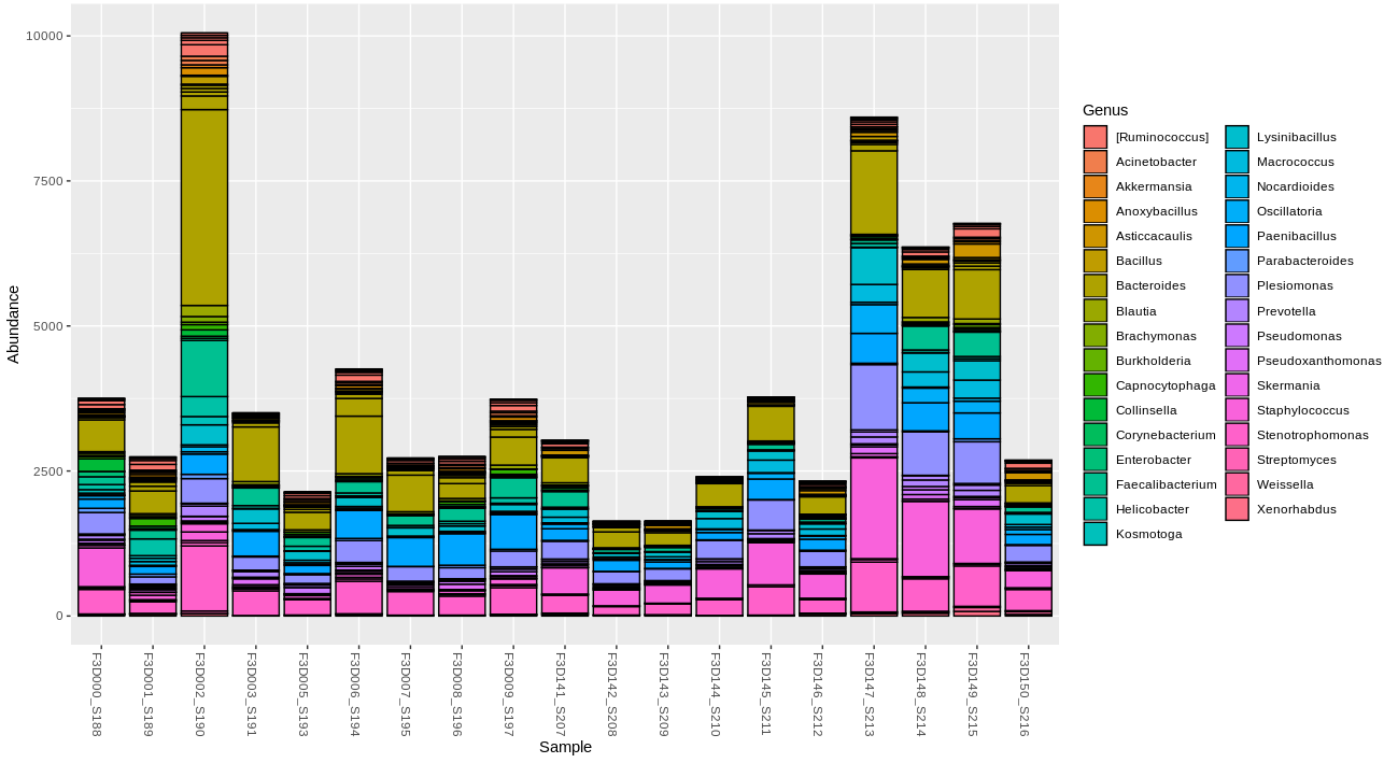

Fig. 1 – Gráfico de barras com 50 bactérias mais abundantes em nível de gênero

Como observamos, houve um aumento significativo de bactérias presentes no intestino do animal nos dias 2, 147, 148 e 149 após o desmame. As variações existentes entre a abundância das diferentes bactérias nos períodos iniciais e tardios após o desmame não parece significativas olhando este gráfico apenas, uma vez que quando há, por exemplo, aumento de abundância de determinada bactéria em determinado dia, parece que esse aumento é proporcional ao observado em outros dias. Sendo assim, a composição de bactérias parece permanecer constante, apesar do aumento de abundância de certas bactérias em determinados períodos.

Um *plot* feito anteriormente a nível de filo (img_bar_phylum na pasta plots), com os 10 filos mais abundantes, mostrou que as bactérias mais abundantes foram justamente aquelas identificadas como componentes naturais da microbiota, como B*acteroidetes*, *Firmicutes* e *Proteobacteria.*

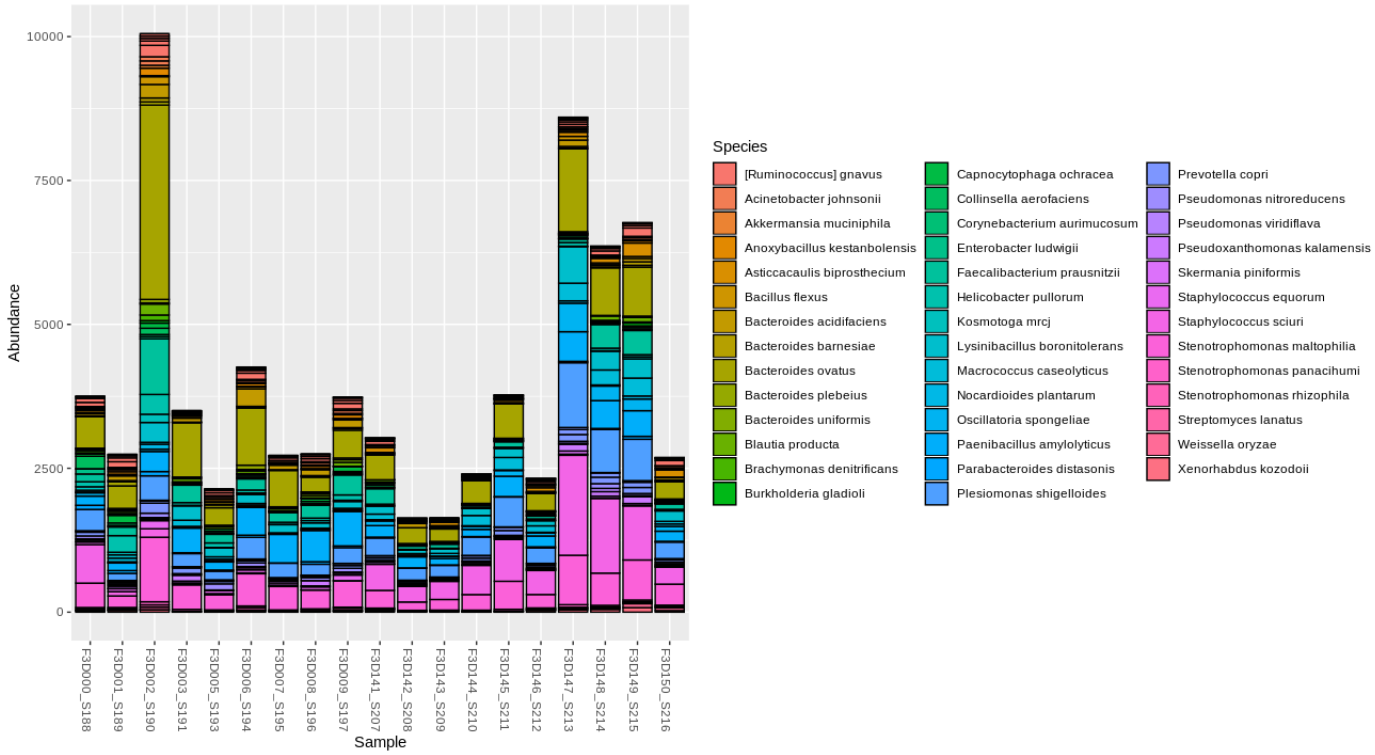

Fig. 2 – Gráfico de barras com 50 bactérias mais abundantes em nível de espécie

2.2) Gráfico de PCoA mostrando o perfil de agrupamento entre as amostras por dia após o desmame;

Em seguida, foram feitos alguns *plots* não solicitados, mas interessantes para a análise de dados, como *alpha diversity* e NMDS. Na *alpha diversity* (Fig. 3) é mostrada a riqueza de espécies, ou seja, o número de espécies diferentes em uma amostra. À direita na **Fig. 3**, podemos visualizar o índice de Shannon, que mede a distribuição uniforme dos micróbios em uma amostra. Vemos que não parece haver uma forte tendência no número de espécies nem na sua uniformidade nas amostras. O NMDS acaba sendo bem similar ao PcoA, mudando essencialmente o algoritmo.

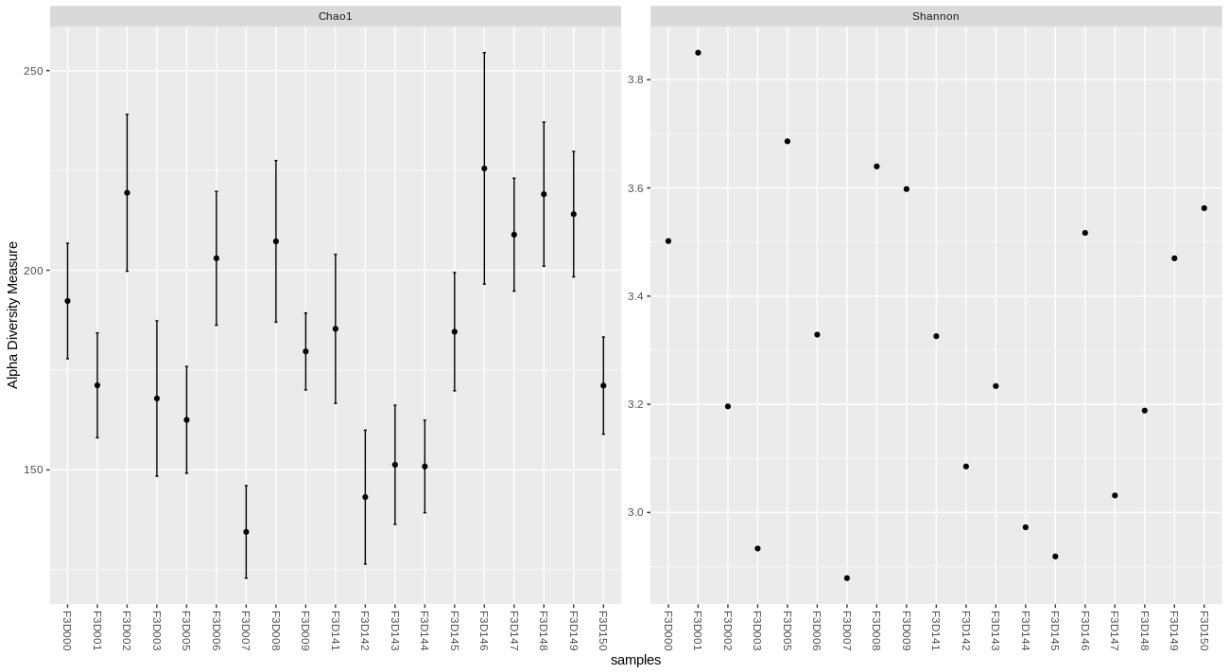

Fig. 3 – *Alpha diversity* para todos os dias de desmame

A análise de coordenadas principais (PCoA) é um método para explorar e visualizar semelhanças ou dissimilaridades de dados. Ao usar o PCoA, podemos visualizar diferenças individuais ou de grupo. No nosso gráfico de PcoA (Fig. 4), podemos ver que uma pequena porcentagem da variação entre as amostras pode ser explicada pelas componentes x (Axis1) e y (Axis2). Verificamos que, apesar de amostras de dias subsequentes não ficarem juntas no gráfico, o que indicaria ausência de diferenças, estas ficam próximas, o que indica uma similaridade entre os resultados de dias de desmame próximos. Com relação ao tempo de desmame precoce ou tardio, verificamos que as amostras de períodos precoces ficam agrupadas acima no gráfico, enquanto aquelas de períodos tardios ficam abaixo, mostrando que há uma diferença observada na composição microbiana nesses períodos.

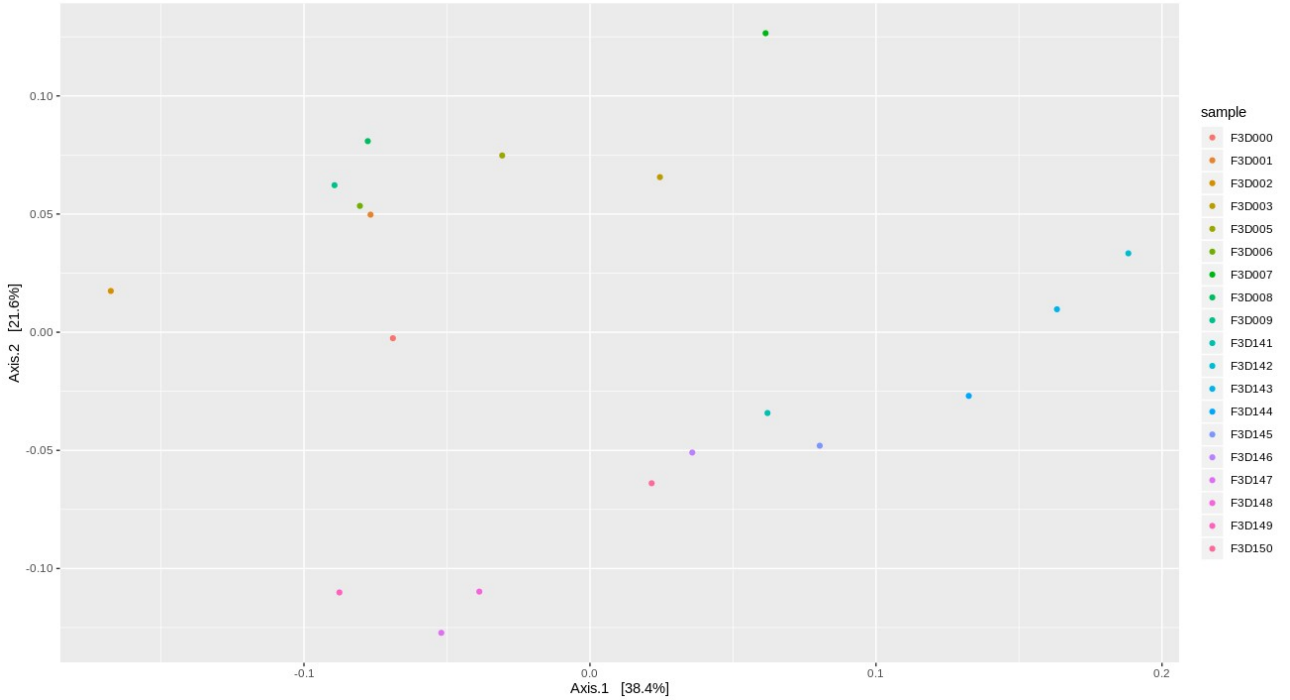

Fig. 4 - Gráfico de PCoA mostrando o perfil de agrupamento entre as amostras por dia após o desmame

2.3) Usar alguma métrica que mostre as bactérias diferencialmente abundantes entre os dias de desmame (edegR ou DESeq2, por exemplo). Em um arquivo (PDF, HTML, DOC ou similar), descreva os resultados obtidos e explique quais foram os critérios de escolha dos métodos analíticos usados.

Uma tarefa básica na análise de dados do RNA-seq é a detecção de genes diferencialmente expressos quando um organismo foi exposto a diferentes condições. Para isso, pode-se utilizar, por exemplo, o pacote em R chamado DESeq2. Esse pacote fornece vários métodos para testar a expressão diferencial usando modelos estatísticos, fazendo normalizações nos dados e testando as abundâncias relativas entre as amostras e seus controles.

Da mesma forma, podemos utilizar esse poderio estatístico do DESeq2 para analisar dados de sequenciamento 16S, podendo verificar quais as bactérias diferencialmente abundantes em diferentes períodos.

Nessa etapa, transformamos nosso objeto do **phyloseq** usado em análises anteriores para o objeto do DESeq2. Depois, obtemos as taxonomias significativamente abundantes filtrando nossos dados com um cutoff de False Discovery Rate (FDR) de 0.01, bem como utilizamos nossos metadados para promover a comparação entre amostras com dias após o desmame precoce (early) ou tardio (late). Também é aplicado um filtro para que fiquem somente taxonomias

significativamente abundantes com *p-value* ajustado < 0.1, justamente para evitar que artefatos ou dados gerados ao acaso sejam considerados.

Finalmente, podemos salvar nossa tabela (arquivo deseq2_diff_abundance.csv na pasta plots), que apresenta **log2 fold change**, erro padrão, teste estatístico, p-values and **p-values ajustados**. Log2 fold change (log2FC) é uma medida que descreve quanto uma quantidade muda do estado original para uma medição subsequente, ou seja, a mudança entre o período precoce e tardio de desmame no nosso caso. Como é dada em termos de log2, um valor de -3 log2FC nos diz que a bactéria estará 8 vezes **menos** abundante na condição precoce em relação à tardia. Geralmente, resultados estatisticamente significantes estão associados com log2 fold changes altos (negativos ou positivos) e p-values ajustados < 0.1 ou < 0.05.

Dos 309 taxons presentes no banco de referência, 50 foram detectados como diferencialmente abundantes pelo DESeq2. A partir deles, foi gerado um *plot* para visualizarmos essas espécies e suas variações de abundância entre os períodos de desmame (Fig. 5). A partir destes resultados, análises mais específicas podem ser feitas para investigar quais dessas bactérias são maléficas ou benéficas ao indivíduo e qual a influência da amamentação (ou ausência dela) na microbiota do indivíduo e, consequentemente, na sua saúde.

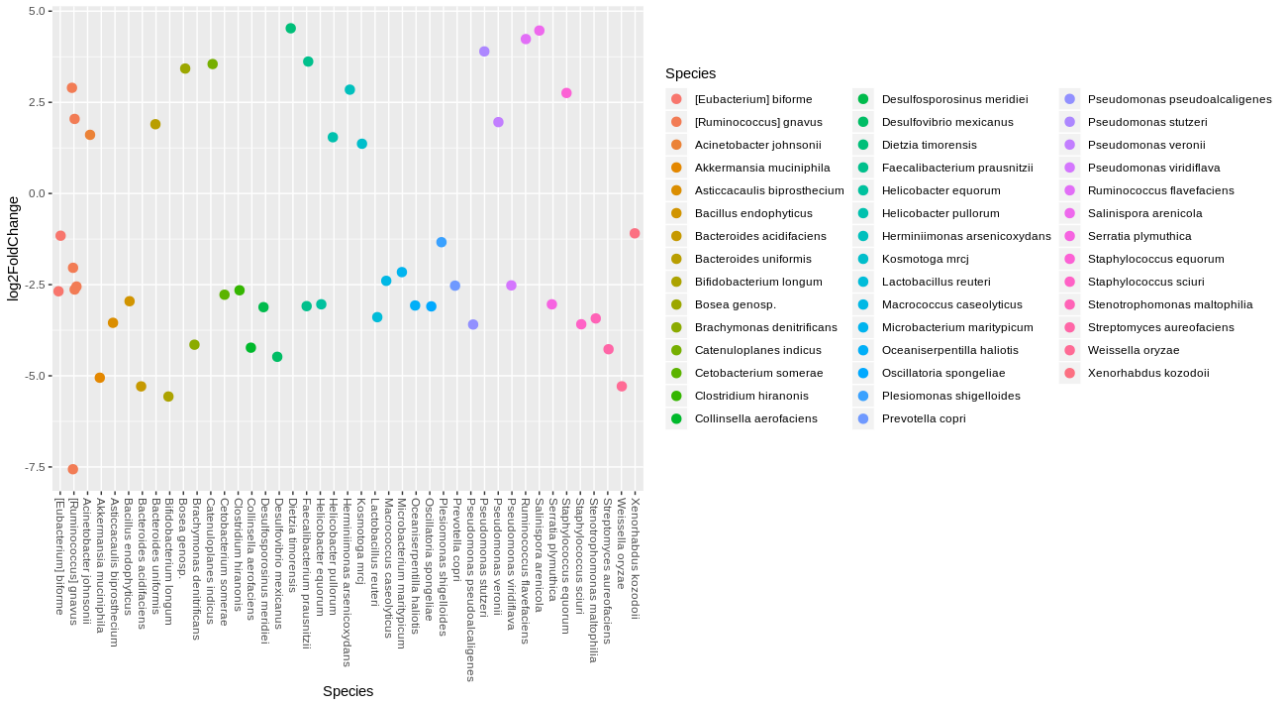

Fig. 5- Gráfico de Espécies vs log2FC utilizando o DESeq2