

Comparison of Pipelines for Functional Gene Annotation of Sus Scrofa Gut Microbiome using EggNOG-mapper

Introduction

While species mostly differ between microbiomes, they often fill similar ecological roles. Thus, it is often valuable to not only look at the taxonomic composition, but also the gene composition and its functional capacity. EggNOG-mapper is a functional annotation tool for metagenomic data. EggNOG-mapper accepts both reads, contigs and protein sequences, but literature offers little information regarding the pros and cons of the different types of input. In this study we will investigate if the output depends on the input, and thereby affects a functional annotation analysis.

Materials and method

Data generation. The samples consist of 20 faeces samples from boars collected in Poland, and 20 faeces samples from domesticated French pigs.

DNA isolation was carried out with PowerLyzer™ PowerSoil® DNA Isolation Kit, according to the standard protocol. Library preparation was done with NEXTflex™ Rapid DNA-Seq Kit, according to the standard protocol including the optional bead size selection step. Paired-end sequencing was done with an Illumina HiSeq. Raw reads had a length of 150bp. Twenty samples were lost during sequencing, leaving six samples from boars and fourteen samples from pigs.

Data processing: First the raw data quality was assessed by creating a FastQC report. With BBtools 5bp was removed from the 5'-end along with adapters. Thereafter low-quality reads were removed together with reads shorter than 50 bp. Then PhiX contamination was removed and reads were mapped against a pig genome to remove host contamination, both done with BBtools. After pre-processing another FastQC-report was created to see the effects of the pre-processing and the quality of the data before further analysis. To investigate sequencing depth a nonpareil-curve was created for each sample. From here the pipeline was split into two parts. In one we used EggNOG directly on the pre-processed reads. In the other pipeline we first did de novo assembly with SPAdes. After this we did gene prediction with Prodigal before running EggNOG on the predicted genes. Finally, we compared the EggNOG output from the two pipelines in R.

Conclusion

- Importance of understanding different limitations in computational resources and how to utilize them optimally
- Running EggNOG on only the protein sequences are faster than on reads, but all the additional steps in this pipeline makes it more time and resource consuming
- Expected to find great overlap between the pipeline's predictions, however, at this point it does not seem to be the case

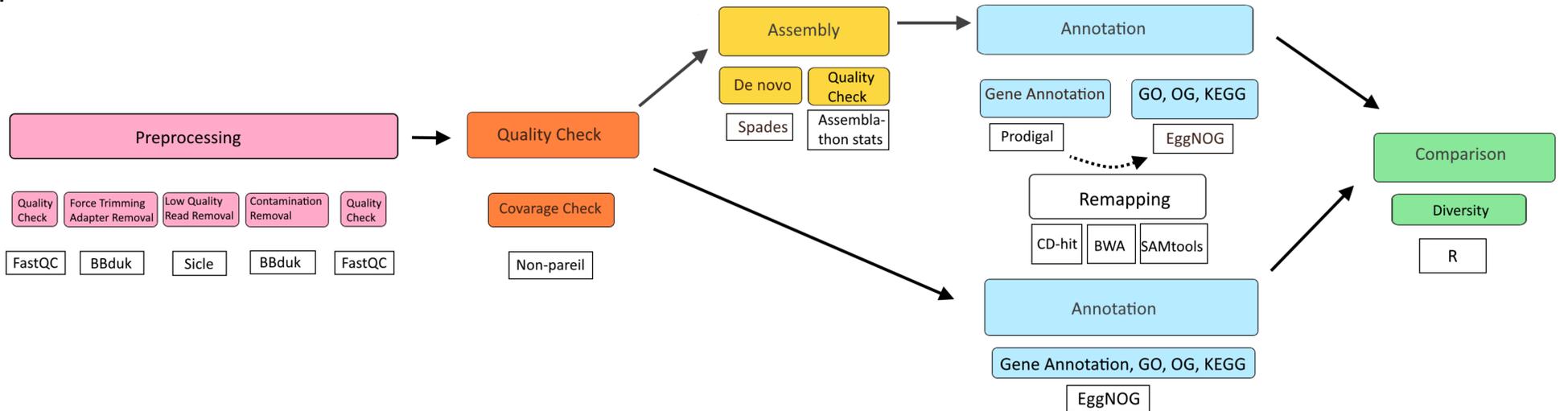
Further studies

- Finishing protein pipeline: Remapping and combining with output from EggNOG
- Make common gene set for reads-pipeline
 - Sort reads output by e-value (discard most uncertain hits)
 - Heatmap comparison on transformed data (cluster on pipeline or sample)
 - Bray-Curtis (compare sample to sample between pipelines)

The biological questions:

- Were the most significant genes the same in the two pipelines?
 - Which genes were this
 - Cluster these genes with GO terms
- Compare gene annotation of the most significant genes between the boars and pigs
 - Transform the count data
 - PCA
 - ANOVA: Choose specific genes or group of genes and test them to see if they were significantly different expressed
- Diversity between samples in preferred pipeline
 - Shannon and Bray-Curtis index

Pipeline



Results

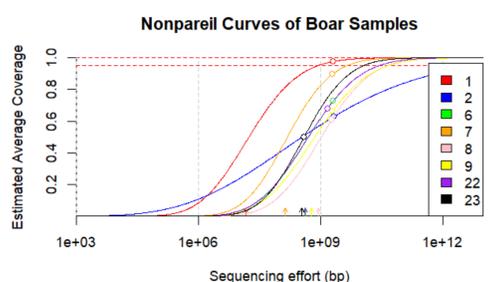
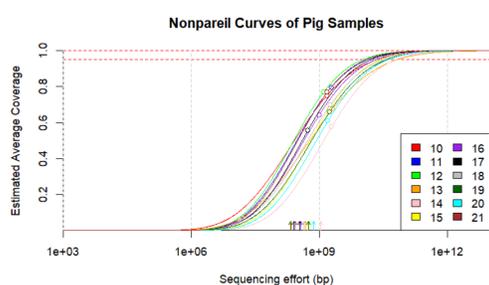


Figure 1. Nonpareil curves of pre-processed data.

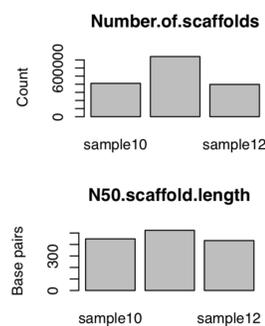


Figure 2. Assembly quality check

Table 1. Shannon Index

Sample	Protein sequences	Reads
HTAdapter_10	9.85	9.86
HTAdapter_11	10.06	10.16
HTAdapter_12	9.89	10.11

Table 2. EggNOG annotation file example

query_name	seed_eggNOG_ortholog	seed_ortholog_evalue	seed_ortholog_score	predicted_gene_name	GO_terms	KEGG_KOs	BiGG_reactions	Annotation_tax_scope	OGs	bestOG evalue score	COG cat	eggNOG annot
NODE_1_length_107463_cov_7_5.520287_3	891391.LAC305_C_04985	6.5e-64	248.4	CRCB2	GO:0005575 GO:0005623 GO:0005886 GO:0016020	K06199		bactNOG[38]	04XGY@bacNOG 06A8V@bactNOG ONR8G@firmNOG COG0239@NOG	NA NA NA	D	Protein CrcB homolog

All samples comparison

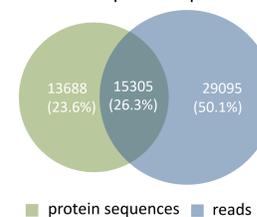


Figure 3. Venn-diagram showing seed orthologs between pipelines. Samples are pooled.

Significant genes

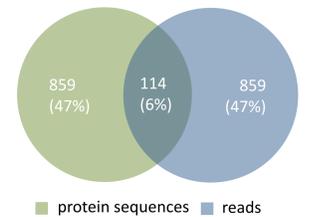


Figure 4. Venn-diagram showing 1000 most significant genes in the pipelines.

Heatmap of all Seed Orthologs in the Samples

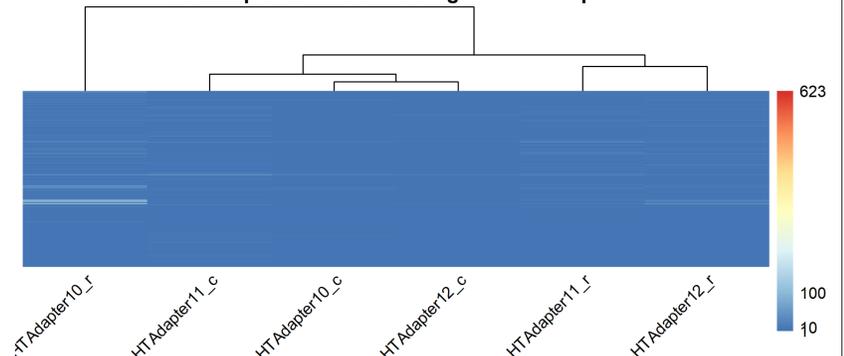


Figure 5. Heatmap clustering the samples according to seed ortholog counts.