HUMAnN4_Analysis_Protocol

1. Introduction: Beyond “Who is there?”

Welcome to the world of **Functional Metagenomics**.

In microbiome research, we often start by asking *“Who is there?”* (Taxonomic Profiling). This tells us which species are present (e.g., *E. coli*, *F. prausnitzii*). However, knowing the names of the bacteria doesn’t tell us what they are actually doing in the environment.

Functional Profiling answers the question: **“What are they capable of?”**

It allows us to quantify the metabolic potential of the community. For example:

- Do the microbes in this gut have the genes to produce Butyrate (a beneficial short-chain fatty acid)?
- Is the pathway for Antibiotic Resistance more abundant in the disease group?
- Which specific species is carrying the genes for Vitamin B12 synthesis?

This guide explains how to interpret the functional profiles generated by **HUMAnN 4** (HMP Unified Metabolic Analysis Network) and how to use them to answer your research questions.

---

2. Prerequisites & Data Origin

You do not need to run the heavy computational steps.

The computationally intensive part of the pipeline (mapping millions of DNA reads to databases) has already been performed on a High-Performance Computing (HPC) cluster by your collaborator.

What you have received:
A folder containing processed, merged, and normalized tables. These files are ready for statistical analysis on your local computer (using R or Python).

- `humann4_genefamilies-step1-merged.tsv`
- `humann4_genefamilies-step2-normalized.tsv`
- `humann4_genefamilies-step3-EggNOG.tsv`
- `humann4_genefamilies-step3-EggNOG_stratified.tsv`
- `humann4_genefamilies-step3-EggNOG_unstratified.tsv`
- `humann4_genefamilies-step3-GO.tsv`
- `humann4_genefamilies-step3-GO_stratified.tsv`
- `humann4_genefamilies-step3-GO_unstratified.tsv`
- `humann4_genefamilies-step3-KO.tsv`
- `humann4_genefamilies-step3-KO_stratified.tsv`
- `humann4_genefamilies-step3-KO_unstratified.tsv`
- `humann4_genefamilies-step3-PFAM.tsv`
- `humann4_genefamilies-step3-PFAM_stratified.tsv`
- `humann4_genefamilies-step3-PFAM_unstratified.tsv`
- `humann4_genefamilies-step3-RXN.tsv`
- `humann4_genefamilies-step3-RXN_stratified.tsv`
- `humann4_genefamilies-step3-RXN_unstratified.tsv`
- `humann4_pathabundance-step1-merged.tsv`
- `humann4_pathabundance-step2-normalized.tsv`
- `humann4_pathabundance-step2-normalized_stratified.tsv`
- `humann4_pathabundance-step2-normalized_unstratified.tsv`
- `humann4_pathway-step3-unpacked.tsv`
- `humann4_reactions-step1-merged.tsv`
- `humann4_reactions-step2-normalized.tsv`
- `humann4_reactions-step2-normalized_stratified.tsv`
- `humann4_reactions-step2-normalized_unstratified.tsv`

3. File Inventory: Understanding Your Data

You will see several types of output files. Here is how to use each one to answer specific research questions.

A. Gene Families (`genefamilies`)

- **File:** `humann4_genefamilies-step2-normalized.tsv` (and regrouped versions like `step3-KO` , `step3-GO`)
- **What is it?** The abundance of specific gene sequences or functional groups.
- **Research Question:** "Is a specific enzyme (e.g., Alcohol Dehydrogenase) more abundant in Group A vs Group B?"
- **How to use:**
  - Use the **KO (KEGG Orthology)** table for metabolic reconstruction. It is the most specific functional unit.
  - Use the **GO (Gene Ontology)** table for a high-level overview (e.g., "Is 'Metabolic Process' enriched?").
  - **Analysis:** Input these tables into **MaAsLin2** to find differentially abundant functions.

B. Pathways (`pathabundance`)

- **File:** `humann4_pathabundance-step2-normalized.tsv`
- **What is it?** The abundance of complete metabolic pathways (e.g., "Tryptophan Biosynthesis"). A pathway is made of multiple genes working together.
- **Research Question:** "Is the *entire capability* to synthesize Tryptophan enriched in the healthy gut?"
- **Why use this instead of genes?** It is more biologically interpretable. Finding a change in a whole pathway is often more robust than finding a change in a single gene.
- **How to use:** This is your **primary file** for most statistical testing. Use the **Unstratified** version for community-wide comparisons.

C. Reactions (`reactions`)

- **File:** `humann4_reactions-step2-normalized.tsv`
- **What is it?** The abundance of specific chemical reactions (MetaCyc reactions).
- **Research Question:** "Is the reaction converting Pyruvate to Acetyl-CoA enriched?"
- **How to use:** Use this when you need chemical specificity but don't care about the broader pathway context.

D. Stratified vs. Unstratified Files

For every table above, you will see two versions:

1. Unstratified (`_unstratified.tsv`)

- **Content:** The **TOTAL** abundance of a function in the entire community.
- **Usage: ALWAYS START HERE.**
- **Research Question:** "Does the *community as a whole* have more of this function?"
- **Analysis:** Perform your statistical tests (t-tests, MaAsLin2) on this file first.

2. Stratified (`_stratified.tsv`)

- **Content:** The abundance broken down by species (e.g., `g__Bacteroides.s__Bacteroides_fragilis` contributes 50 CPM).
- **Usage:** Use this **only after** you find a significant result in the unstratified table.
- **Research Question:** "I found that 'Pathway X' is enriched in Group A. **Which species is responsible for this increase?**"
- **Analysis:** Don't run stats on this whole file (it's too sparse). Instead, filter for your significant pathway and plot the species contributions as a bar chart.

4. Downstream Analysis Guide

Now that you have your files, here is the roadmap for your analysis.

Step 1: Diversity Analysis

Goal: Assess the overall functional complexity of your samples.
* **Alpha Diversity (Richness):** Calculate the Shannon Index on your **Pathabundance** table.
* *Question:* "Do healthy people have a more functionally diverse microbiome?"

* **Beta Diversity (Dissimilarity):** Calculate Bray-Curtis dissimilarity and visualize with PCoA.
* *Question:* "Do the functional profiles of patients cluster separately from healthy controls?"

Step 2: Differential Abundance Testing

Goal: Find specific biomarkers (pathways or genes) that differ between groups.
* **Tool: MaAsLin2** (Microbiome Multivariable Associations with Linear Models).
* **Why?** It handles the unique properties of microbiome data (compositionality, sparsity) and allows you to correct for confounders (e.g., Age, Sex, BMI).
* **Input:** `humann4_pathabundance-step2-normalized_unstratified.tsv`
* **Output:** A list of significant pathways (FDR < 0.05).

Step 3: Driver Analysis (Stratification)

Goal: Link function back to taxonomy.
* **Scenario:** MaAsLin2 tells you "Pathway X" is higher in Disease.
* **Action:** Go to the **Stratified** table. Extract the rows for "Pathway X".
* **Visualization:** Create a stacked bar plot.
* *Result:* You might see that in Healthy controls, *Species A* provides this function, but in Disease, *Species A* is gone and *Species B* is providing it at a much higher rate.

Step 4: Visualization

- **Heatmaps:** Show the top 50 most variable pathways across all samples to see patterns.
- **Scatterplots:** Correlate a pathway's abundance with a clinical marker (e.g., "Abundance of Butyrate Synthesis vs. Inflammation Score").

---

5. Summary Checklist

□ **Receive Data:** Get the folder of `.tsv` files from your collaborator.

□ **Check Normalization:** Ensure you are using the files with `normalized` in the name.

□ **Start High-Level:** Run diversity and differential abundance on **Unstratified Pathways**.

□ **Dig Deeper:** If you find interesting pathways, look at the **Stratified** data to find the species responsible.

□ **Validate:** Use **Gene Families (KO)** if you need to confirm specific enzymatic steps within a pathway.