HUMAnN

ECES T480/680
Tutorial 12
By: Keyur Shah, Nicole Buleza, and Bairavi Venkatesh

Agenda

I. Background

II. Workflow

III. Definition and Databases

IV. Initialization and Use

V. Outputs and Visualization

VI. HUMAnN vs. HUMAnN2

Background

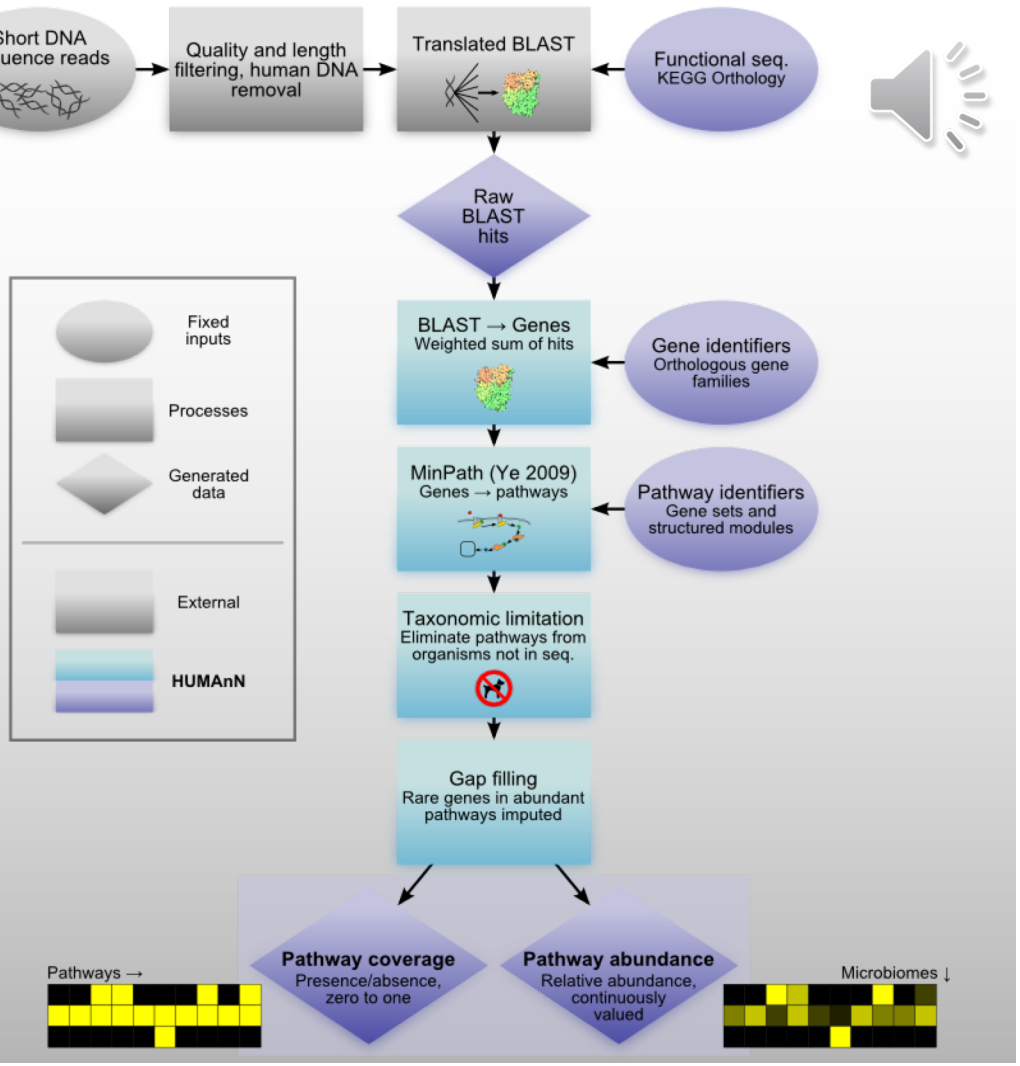
- HUMAnN2 is the next version of HUMAnN (HMP Unified Metbolic Analysis Network)
- Functional Profiling → "What are the microbes in my community-of-interest doing (or capable of doing)?"

- Urban Dataset
  - Contains sequences from train cars and subway stations across the Boston Subway.
  - 24 shotgun metagenome sequences

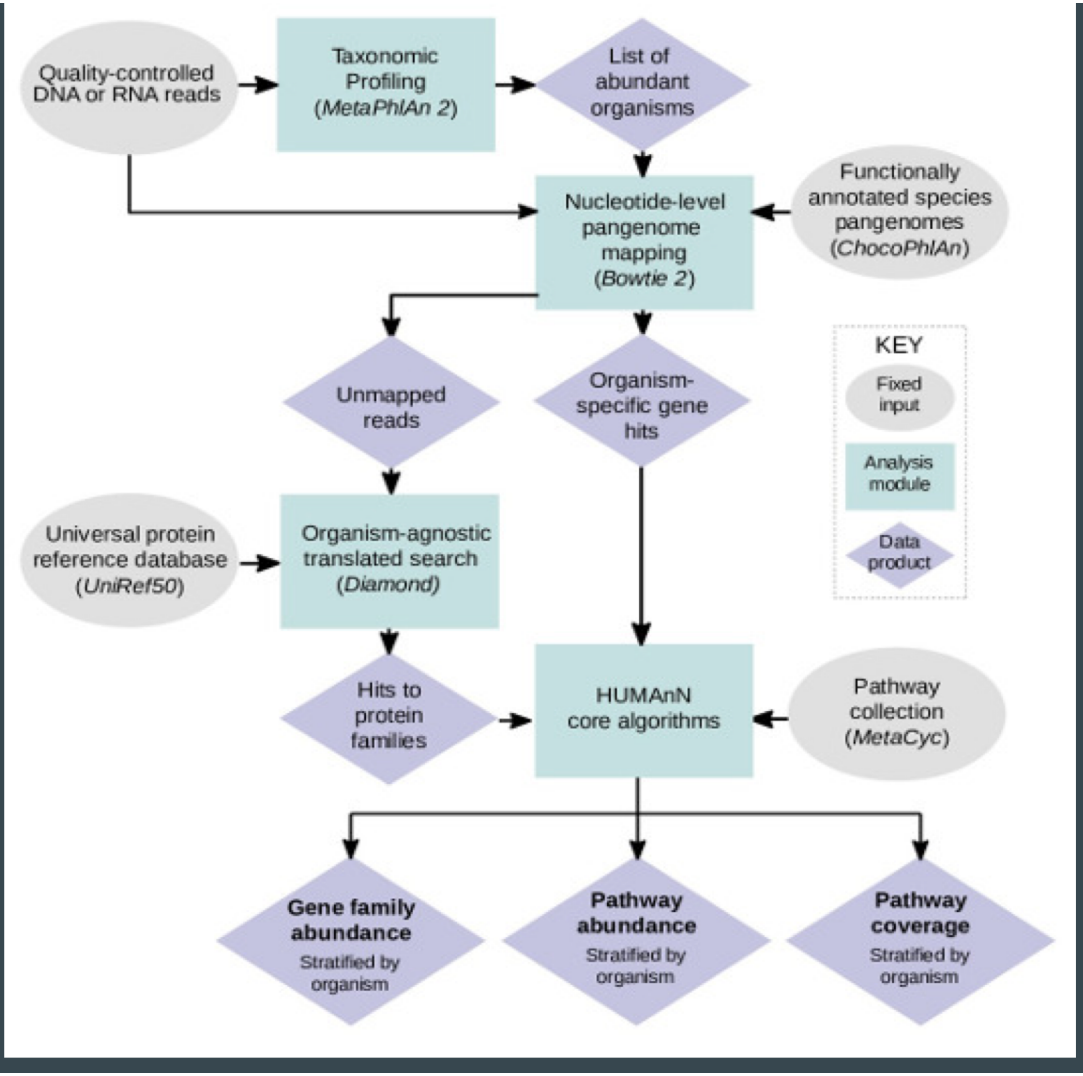
HUMAnN2 Features

- Community functional profiles stratified by known and unclassified organisms
- Considerably expanded databases of genomes, genes, and pathways
- A simple user interface (single command driven flow)
- Accelerated mapping of reads to reference databases (including run-time generated databases tailored to the input)

HUMAnN Workflow



WorkFlow


Quality-controlled DNA or RNA reads
Taxonomic Profiling (MetaPhlAn 2)
List of abundant organisms
Functionally annotated species pangenomes (ChocoPhlAn)
Nucleotide-level pangenome mapping (Bowtie 2)
Unmapped reads
Organism-specific gene hits
KEY
Fixed input
Analysis module
Data product
Universal protein reference database (UniRef50)
Organism-agnostic translated search (Diamond)
Hits to protein families
HUMAnN core algorithms
Pathway collection (MetaCyc)
Gene family abundance
Stratified by organism
Pathway abundance
Stratified by organism
Pathway coverage
Stratified by organism

Definitions and Databases

- **Ortholog**
  - Genes in different species that evolved from a common ancestral gene by speciation
  - Retain the same function throughout the course of evolution

- **COG** - "Clusters of Orthologous Groups"
  - Protein database generated by comparing predicted and known proteins in all completely sequenced microbial genomes to infer ortholog sets
  - Fast alternative for rapidly describing the functional characteristics of one microbe or a community of microbes

- **EggNOG** - "Evolutionary genealogy of genes: Non-supervised Orthologous Groups"
  - Protein database built by collecting genomes from public datasets and performing pairwise similarity matrix
  - Successor of COG protein database which extends analysis to eukaryotes

- **KEGG** - "Kyoto Encyclopedia of Genes and Genomes"
  - Database used to map molecular datasets, especially large in genomics, transcriptomics, proteomics, and metabolomics for system functions

What Is Coverage?

- Coverage (Read depth or Depth)
  - In DNA sequencing, it is the number of reads that include a given nucleotide in the reconstructed sequence.
- With HUMAnN2, the user is able to carry out metagenomic analysis based on the following classifications:
  - Orthology
    - Ex. COG
  - Structure
    - Ex. Pfam
  - Biological Role
    - Ex. KEGG

Coverage Calculation

G = Length of original genome
N = Number of reads
L = Average read length
R = Redundancy

$$R = N \times L/G$$

E(C) = The expected target coverage

$$E(C) = 1 - e^{-R}$$

Importance of Coverage

- The vast number of nucleotides in the genome means that if an individual genome is only sequenced once, there will be a significant number of sequencing errors.
- Making matters more complicated, many positions in a genome contain SNPs.
- Hence, to distinguish between sequencing errors and true SNPs, it is necessary to increase the sequencing accuracy even further by sequencing individual genomes a large number of times
- **Deep sequencing:** Refers to the general concept of aiming for high number of replicate reads of each region of a sequence.

Genome vs. Transcriptome: Functional Capacity vs Actual Capacity

- Genome: The complete set of genetic material in a single organism
- Transcriptome: For a gene's function to be carried out, DNA must be read and transcribed into RNA. The gene readouts produced are called "transcripts". A transcriptome is a collection of all the gene transcripts in a cell.
- Metatranscriptome represents the functional capacity of the microbiome
- Metagenome is the actual capacity of the microbiome
- Biological Importance: In humans and other organisms, nearly every cell contains the same genes, but different cells show different patterns of gene expression. These differences are responsible for the various properties and behaviors of cells and tissues, both in health and in disease.

Installation and Use

- HUMAnN2 can be installed on a Linux based system like Proteus.
  - **wget** https://pypi.python.org/packages/71/70/9c45436b6dab38706826a822411d6386376205d9c9fa53972e2ff3b7dda8/humann2-0.9.9.tar.gz
  - **tar zxvf** humann2-0.9.9.tar.gz
  - **module load** python/2.7-current
  - **python** setup.py install

- Dependencies:
  - Bowtie2 (version >= 2.2) - nucleotide level searches
  - MetaPhlAn2 - taxonomic profiling
  - Diamond (version >= 0.7.3) - translated searches
  - Python (version >= 2.7)

Automatically installed!

Installation and Use Continued

- For shotgun sequenced metagenome files (.fastq / .fastq.gz / .fasta):
  - ChocoPhlAn Database → ~50M genes from NCBI
  - UniRef Database → ~100M proteins from UniProt
- For paired end data:
  - Files need to be concatenated
- Basic Usage:
  - Run command
    - `$ humann2 --input $INPUT_FILE.fastq --output $OUTPUT_DIR`
  - Normalize outputs
    - Relative abundance or copies per million
    - .tsv files
  - Combine outputs
    - Merge abundance information across multiple reads

Output #1 - Gene Abundance

1	# Gene Family	SRR3545898_Abundance-RPKs	SRR3545910_Abundance-RPKs
2	UNMAPPED	0.856128	0.907145
3	UniRef90_A0A008BLW5	2.33E-07	4.28E-07
4	UniRef90_A0A008BLW5|g__Staphylococcus.s__Staphylococcus_epidermidis	0	1.07E-07
5	UniRef90_A0A008BLW5|g__Staphylococcus.s__Staphylococcus_hominis	2.33E-07	3.21E-07

- Gene Family = evolutionarily related protein-coding sequences that perform similar functions
- Average abundance of each gene family within the community (units = reads per kilobase)
- Stratified to show contributions of known species
- Community abundance = ΣSpecies abundances
- “Unmapped” = total number of reads that remain unmapped after both the nucleotide and translated search

Output #2 - Path Abundance

1	# Pathway	SRR3545898_Abundance	SRR3545910_Abundance
1076	LACTOSECAT-PWY: lactose and galactose degradation I|g__Staphylococcus.s__Staphylococcus_epidermidis	0	2.00E-06
1077	LACTOSECAT-PWY: lactose and galactose degradation I|g__Staphylococcus.s__Staphylococcus_haemolyticus	0	1.07E-06
1078	LACTOSECAT-PWY: lactose and galactose degradation I|g__Staphylococcus.s__Staphylococcus_hominis	0	2.40E-06
1079	LACTOSECAT-PWY: lactose and galactose degradation I|g__Staphylococcus.s__Staphylococcus_pettenkoferi	0	1.83E-07

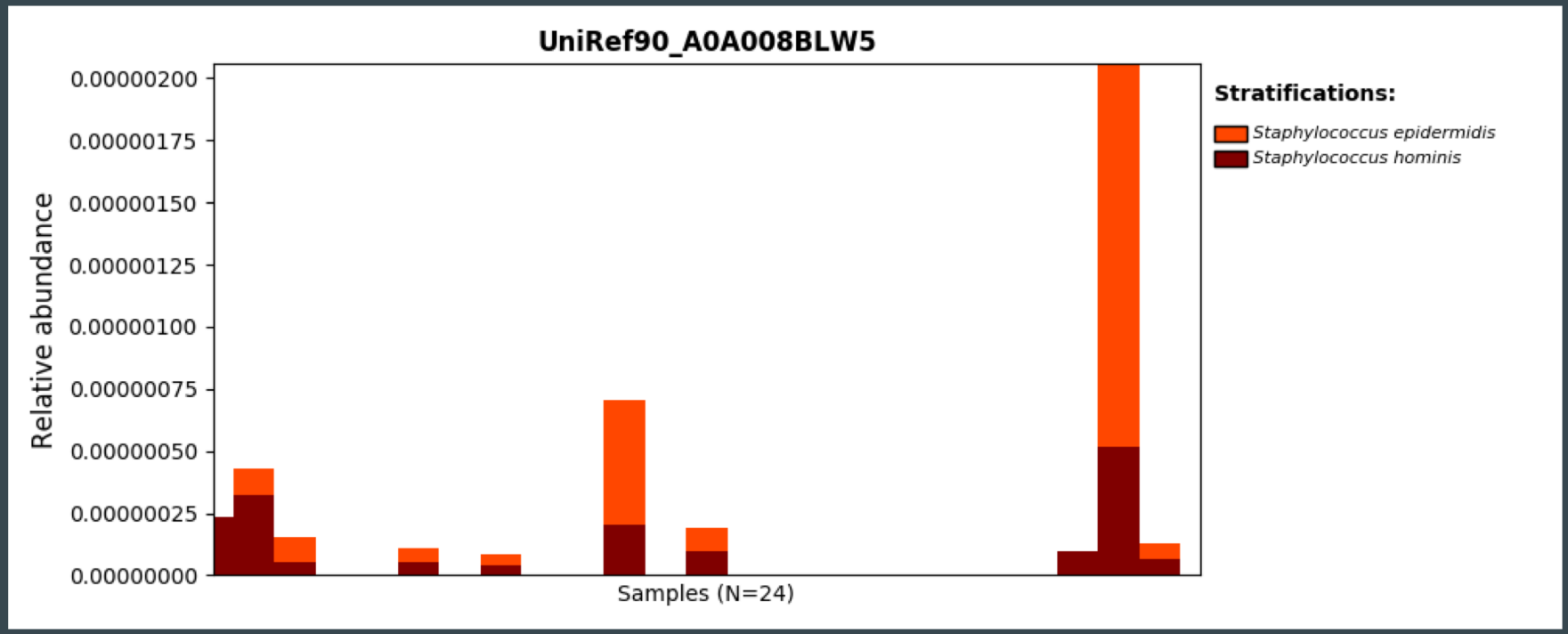
- Abundance of each pathway as a function of its component reactions
- Example Pathway) RXN1 → RXN2 → RXN3
  - Proportional to the number of complete copies (ie- bottleneck)
- Community abundance ≠ ΣSpecies abundances
- Example) Species A: [5, 5, 10] & Species B: [10, 10, 5]
  - Species level → Each species contributes 5 complete copies of the pathway
  - Community level → [15, 15, 15] and therefore 15 complete copies

Output #3 - Path Coverage

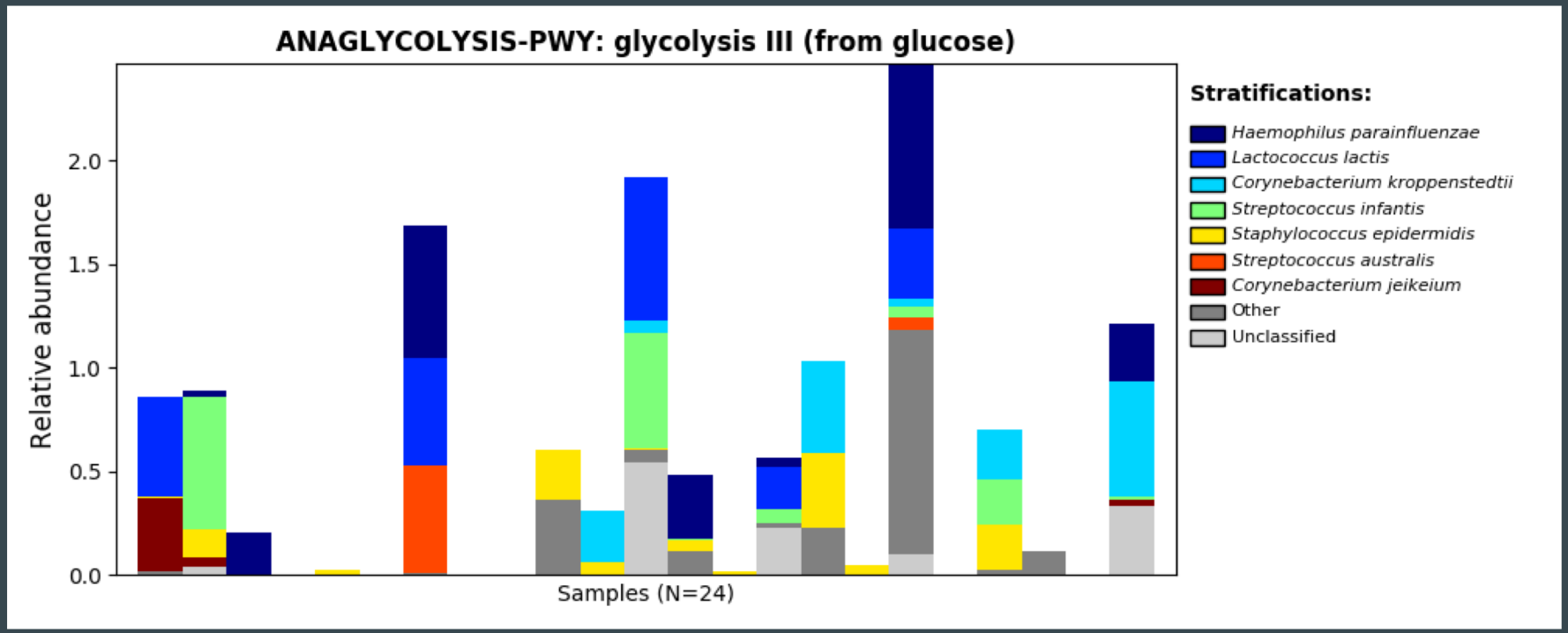
1	# Pathway	SRR3545898_Coverage	SRR3545910_Coverage
327	COA-PWY-1: coenzyme A biosynthesis II (mammalian)|g__Micrococcus.s__Micrococcus_luteus	1	0
328	COA-PWY-1: coenzyme A biosynthesis II (mammalian)|g__Modestobacter.s__Modestobacter_multiseptatus	0	1
329	COA-PWY-1: coenzyme A biosynthesis II (mammalian)|g__Neisseria.s__Neisseria_flavescens	1	1

- Coverage is independent of abundance
- Binary
  - 1 = Confidently detected
  - 0 = Confidently undetected
- Pathway may be confidently detected at a community level, but might not be confidently detected at a species level.

Outputs - Visualization



Outputs - Visualization



HUMAnN vs HUMAnN2

New analyses:

- HUMAnN2: Enzyme and pathway abundances, stratified by species
  - 1 additional output - Gene Family Abundance
  - Abundance reported as RPK instead of relative abundance units
  - Core algorithms perform both species and community level analyses
- MetaPhlAn2: Eukaryotes and viruses

References

[1] https://bitbucket.org/biobakery/humann2/wiki/Home

[2] S. Abubucker et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. Plos Computational Biology 8(6), 2012.

Algorithms - Relative Abundance

- After performing BLAST searches, HUMAnN summarizes the results based on the number of reads within each gene family, weighted by the quality of the match
- The relative abundance of gene *i* is the number of reads *j* that map to a gene sequence in the family weighted by the inverse p-value of each mapping and normalized by average gene length within the family.

$$w_i = |G_i| \sum_{g \in G_i} \frac{1}{|g|} \sum_j \frac{1 - p_{gj}}{\sum_{g'} 1 - p_{g'j}}$$