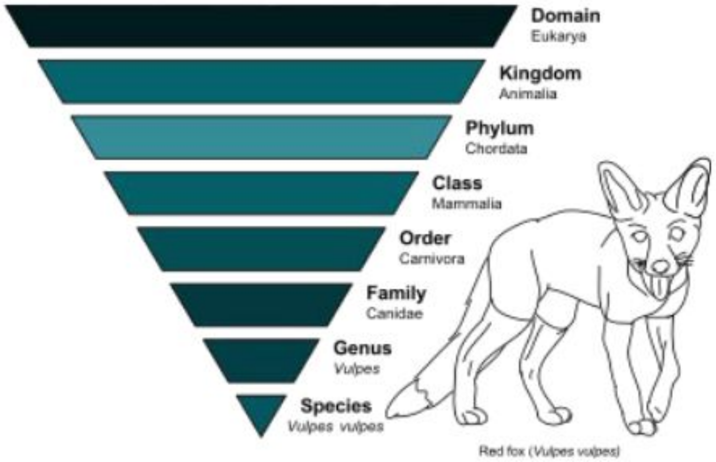
MetaPhlAn & HUMAnN

Thomas Coard

MetaPhlAn

MetaPhlAn is a program for determining the composition of microbial communities from metagenomic shotgun sequencing data.

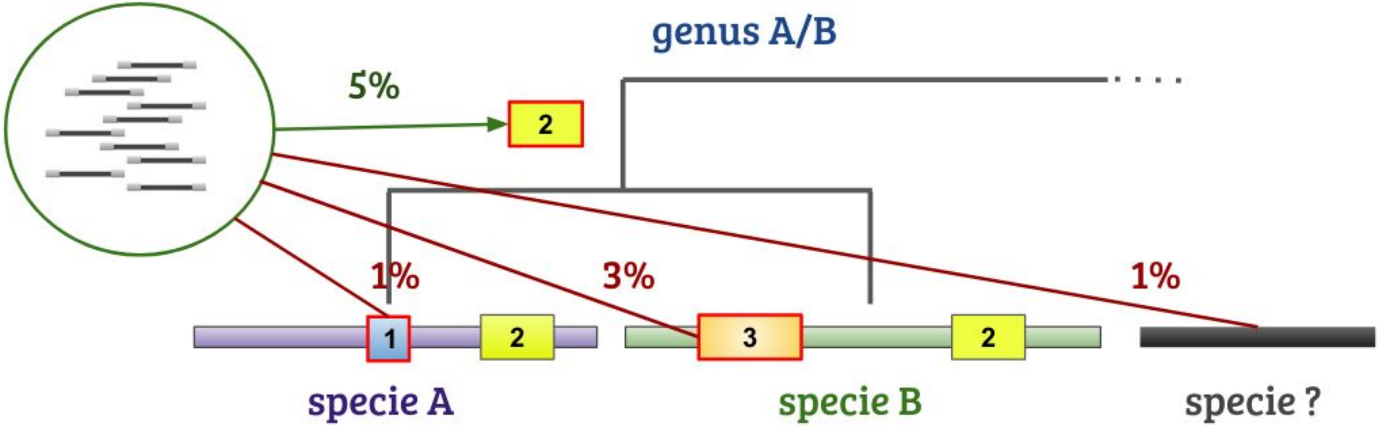


How it Works

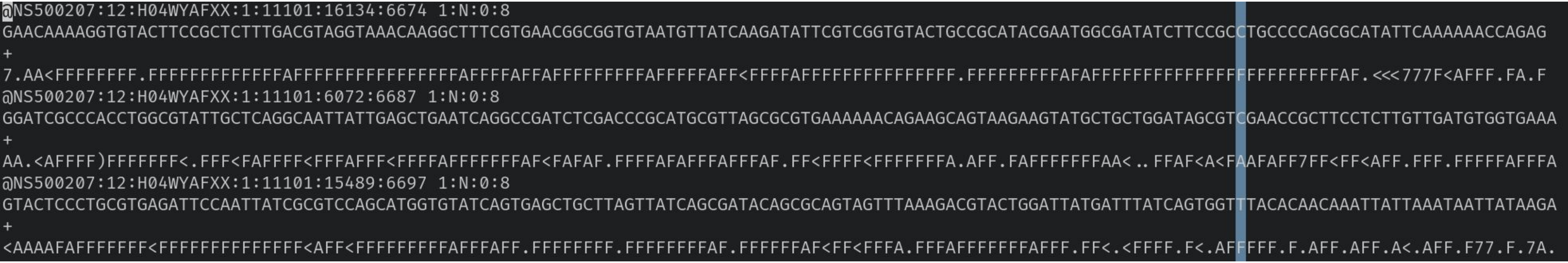
MetaPhlAn looks for highly conserved clade-specific marker genes.

Clades are groups of organisms that evolved from a common ancestor. A clad can be as specific as a species or as broad as a kingdom.

[3]

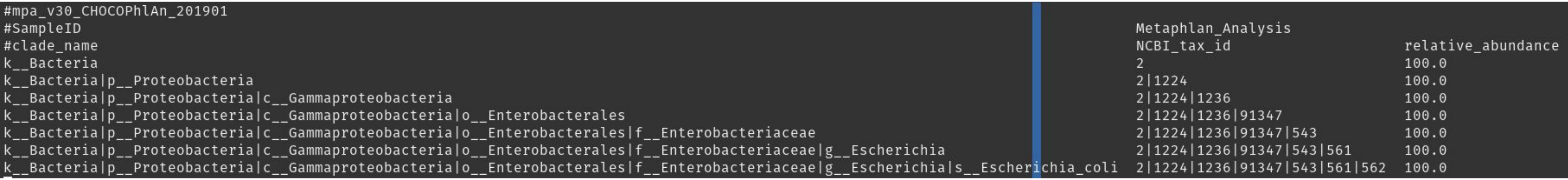


Input: 1-2 Fastq file(s)

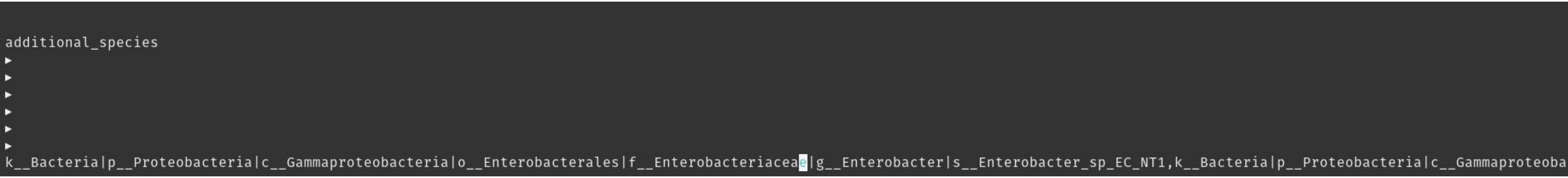


```
@NS500207:12:H04WYAFXX:1:11101:16134:6674 1:N:0:8
GAACAAAAGGTGTACTTCCGCTCTTTGACGTAGGTAAACAAGGCTTTCGTGAACGGCGGTGTAATGTTATCAAGATATTCGTCGGTGTACTGCCGCATACGAATGGCGATATCTTCCGCCTGCCCCAGCGCATATTCAAAAAACCAGAG
+
7.AA<FFFFFFFF.FFFFFFFFFFFFFAFFFFFFFFFFFFFFFFFFAFFFFAFFAFFFFFFFFFFAFFFFFAFF<FFFFAFFFFFFFFFFFFFFFFF.FFFFFFFFFAFAFFFFFFFFFFFFFFFFFFFFFFFFFFAF.<<<777F<AFFF.FA.F
@NS500207:12:H04WYAFXX:1:11101:6072:6687 1:N:0:8
GGATCGCCCACCTGGCGTATTGCTCAGGCAATTATTGAGCTGAATCAGGCCGATCTCGACCCGCATGCGTTAGCGCGTGAAAAAACAGAAGCAGTAAGAAGTATGCTGCTGGATAGCGTCGAACCGCTTCCTCTTGTTGATGTGGTGAAA
+
AA.<AFFFF)FFFFFFF<.FFF<FAFFFF<FFFAFFF<FFFFAFFFFFFFAF<FAFAF.FFFFAFAFFFAFFFAF.FF<FFFF<FFFFFFFA.AFF.FAFFFFFFFAA<..FFAF<A<FAAFAFF7FF<FF<AFF.FFF.FFFFFAFFFA
@NS500207:12:H04WYAFXX:1:11101:15489:6697 1:N:0:8
GTACTCCCTGCGTGAGATTCCAATTATCGCGTCCAGCATGGTGTATCAGTGAGCTGCTTAGTTATCAGCGATACAGCGCAGTAGTTTAAAGACGTACTGGATTATGATTTATCAGTGGTTTACACAACAAATTATTAAATAATTATAAGA
+
<AAAAFAFFFFFFF<FFFFFFFFFFFFFF<AFF<FFFFFFFFFAFFFAFF.FFFFFFFF.FFFFFFFFAF.FFFFFFAF<FF<FFFA.FFFAFFFFFFFAFFF.FF<.<FFFF.F<.AFFFFF.F.AFF.AFF.A<.AFF.F77.F.7A.
```

Output



```
#mpa_v30_CHOCOPhlAn_201901
#SampleID                                                                                                                                   Metaphlan_Analysis
#clade_name                                                                                                                                 NCBI_tax_id                       relative_abundance
k__Bacteria                                                                                                                                 2                                 100.0
k__Bacteria|p__Proteobacteria                                                                                                               2|1224                            100.0
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria                                                                                        2|1224|1236                       100.0
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacterales                                                                    2|1224|1236|91347                 100.0
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacterales|f__Enterobacteriaceae                                              2|1224|1236|91347|543             100.0
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacterales|f__Enterobacteriaceae|g__Escherichia                               2|1224|1236|91347|543|561         100.0
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacterales|f__Enterobacteriaceae|g__Escherichia|s__Escherichia_coli  2|1224|1236|91347|543|561|562     100.0
```



```
additional_species
▶
▶
▶
▶
▶
▶
k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria|o__Enterobacterales|f__Enterobacteriaceae|g__Enterobacter|s__Enterobacter_sp_EC_NT1,k__Bacteria|p__Proteobacteria|c__Gammaproteoba
```

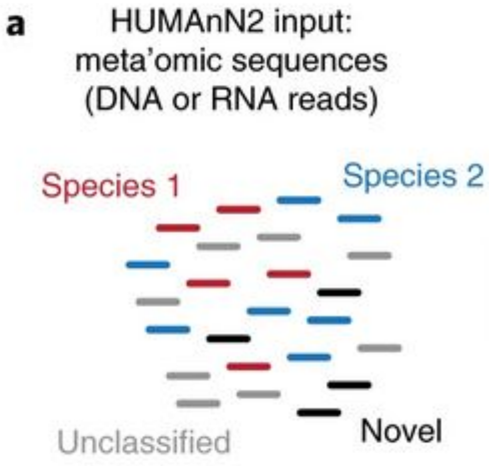
HUMAnN

HMP Unified Metabolic Analysis Network.

HUMAnN is a method for efficiently and accurately profiling the abundance of microbial metabolic pathways and other molecular functions from metagenomic or metatranscriptomic sequencing data.

How it Works Pt. 1

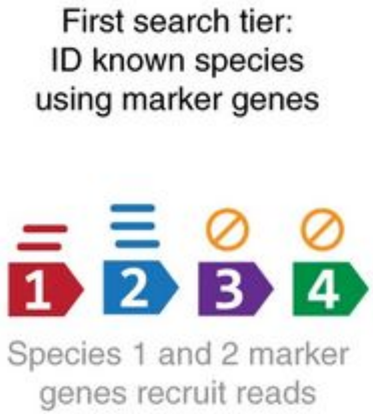
HUMAnN2 rapidly identifies known microbial species in a sample by screening DNA or RNA reads with MetaPhlAn2



(Franzosa et al.)

How it Works Pt. 2

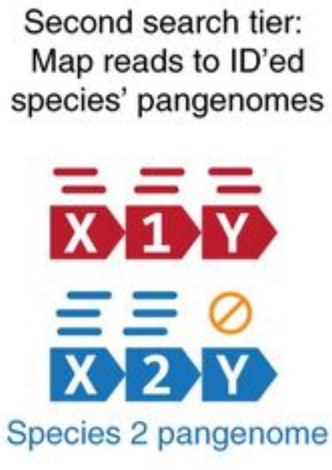
HUMAnN2 then constructs a sample-specific database by merging preconstructed, functionally annotated pangenomes of the identified species



(Franzosa et al.)

How it Works Pt. 3

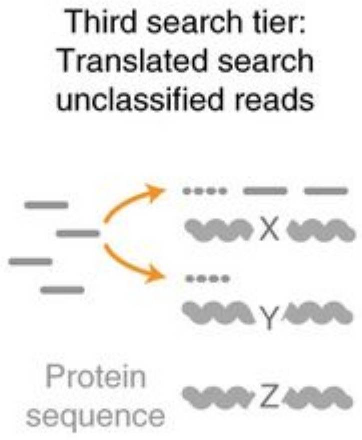
In the second tier, HUMAnN2 performs nucleotide-level mapping of all sample reads against the sample's pangenome database.



(Franzosa et al.)

How it Works Pt. 4

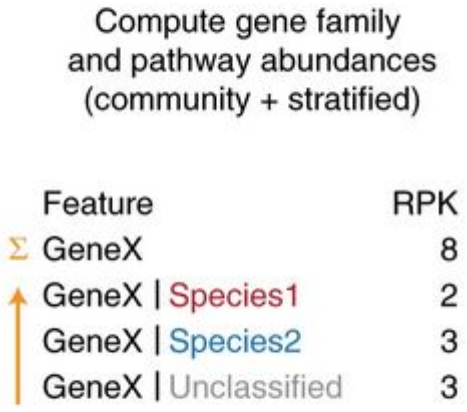
In the third and final tier, reads that do not align to identified species' pangenomes are subjected to accelerated translated search against a comprehensive protein database.



(Franzosa et al.)

How it Works Pt. 5

Finally, gene families annotated to metabolic enzymes are further analyzed to reconstruct and quantify complete metabolic pathways in the community and per organism.



Compute gene family and pathway abundances (community + stratified)

Feature	RPK
Σ GeneX	8
GeneX | Species1	2
GeneX | Species2	3
GeneX | Unclassified	3

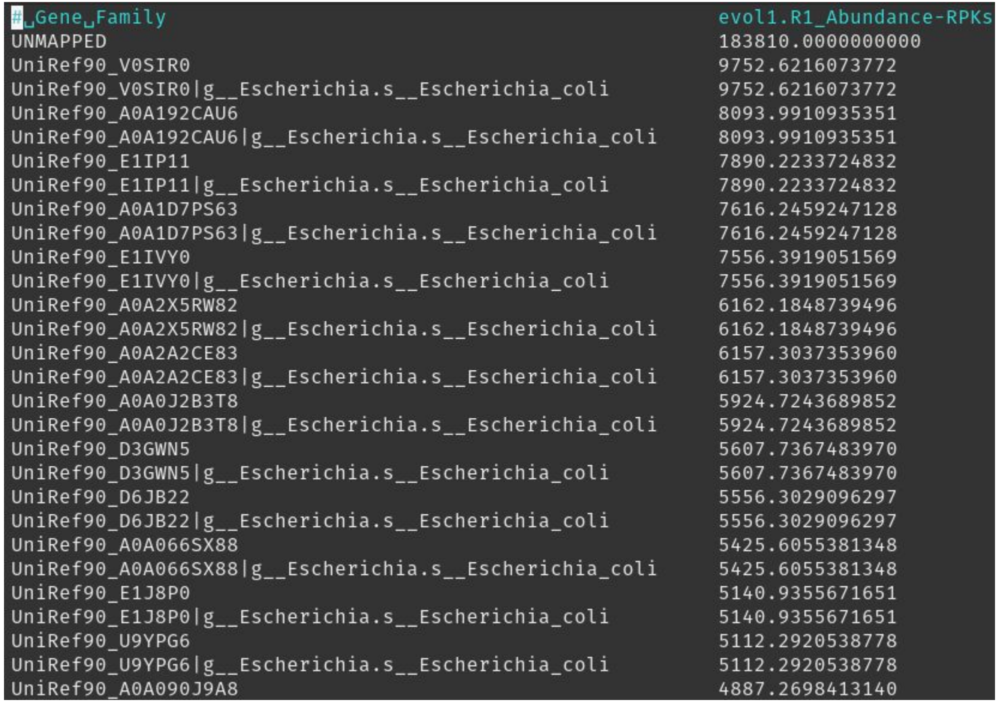
(Franzosa et al.)

Output

1. $OUTPUT_DIR/$SAMPLENAME_genefamilies.tsv
2. $OUTPUT_DIR/$SAMPLENAME_pathcoverage.tsv
3. $OUTPUT_DIR/$SAMPLENAME_pathabundance.tsv

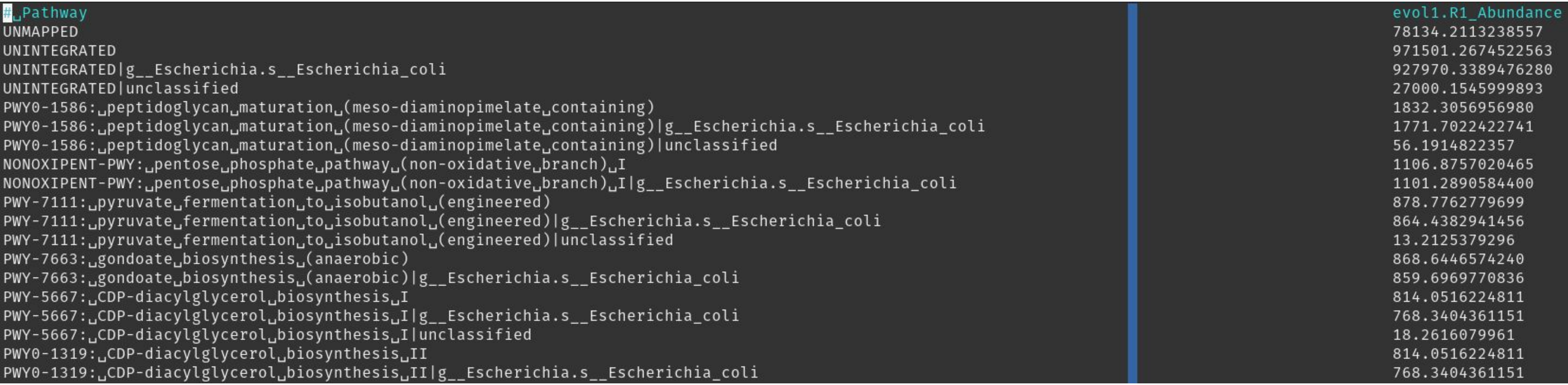
Gene Families

Gene family abundance is reported in RPK (reads per kilobase). This is computed as the sum of the scores for all alignments for a gene family. An alignment score is based on the number of matches to the reference gene for a specific sequence.



```
#␣Gene␣Family                                              evol1.R1_Abundance-RPKs
UNMAPPED                                                   183810.0000000000
UniRef90_V0SIR0                                            9752.6216073772
UniRef90_V0SIR0|g__Escherichia.s__Escherichia_coli         9752.6216073772
UniRef90_A0A192CAU6                                        8093.9910935351
UniRef90_A0A192CAU6|g__Escherichia.s__Escherichia_coli     8093.9910935351
UniRef90_E1IP11                                            7890.2233724832
UniRef90_E1IP11|g__Escherichia.s__Escherichia_coli         7890.2233724832
UniRef90_A0A1D7PS63                                        7616.2459247128
UniRef90_A0A1D7PS63|g__Escherichia.s__Escherichia_coli     7616.2459247128
UniRef90_E1IVY0                                            7556.3919051569
UniRef90_E1IVY0|g__Escherichia.s__Escherichia_coli         7556.3919051569
UniRef90_A0A2X5RW82                                        6162.1848739496
UniRef90_A0A2X5RW82|g__Escherichia.s__Escherichia_coli     6162.1848739496
UniRef90_A0A2A2CE83                                        6157.3037353960
UniRef90_A0A2A2CE83|g__Escherichia.s__Escherichia_coli     6157.3037353960
UniRef90_A0A0J2B3T8                                        5924.7243689852
UniRef90_A0A0J2B3T8|g__Escherichia.s__Escherichia_coli     5924.7243689852
UniRef90_D3GWN5                                            5607.7367483970
UniRef90_D3GWN5|g__Escherichia.s__Escherichia_coli         5607.7367483970
UniRef90_D6JB22                                            5556.3029096297
UniRef90_D6JB22|g__Escherichia.s__Escherichia_coli         5556.3029096297
UniRef90_A0A066SX88                                        5425.6055381348
UniRef90_A0A066SX88|g__Escherichia.s__Escherichia_coli     5425.6055381348
UniRef90_E1J8P0                                            5140.9355671651
UniRef90_E1J8P0|g__Escherichia.s__Escherichia_coli         5140.9355671651
UniRef90_U9YPG6                                            5112.2920538778
UniRef90_U9YPG6|g__Escherichia.s__Escherichia_coli         5112.2920538778
UniRef90_A0A090J9A8                                        4887.2698413140
```

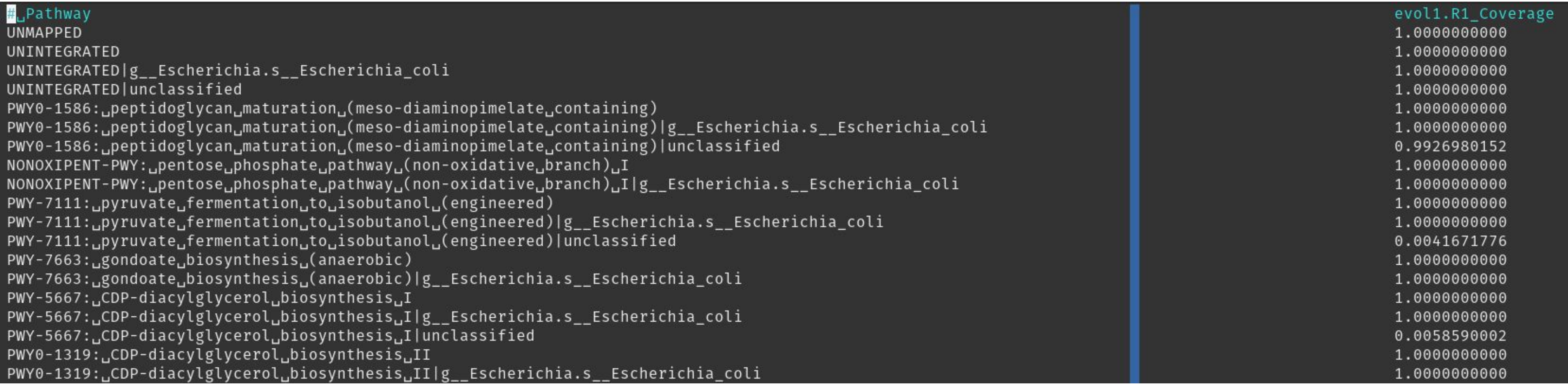
Path Abundance



```
#␣Pathway                                                                                                   evol1.R1_Abundance
UNMAPPED                                                                                                    78134.2113238557
UNINTEGRATED                                                                                                971501.2674522563
UNINTEGRATED|g__Escherichia.s__Escherichia_coli                                                             927970.3389476280
UNINTEGRATED|unclassified                                                                                   27000.1545999893
PWY0-1586:␣peptidoglycan␣maturation␣(meso-diaminopimelate␣containing)                                       1832.3056956980
PWY0-1586:␣peptidoglycan␣maturation␣(meso-diaminopimelate␣containing)|g__Escherichia.s__Escherichia_coli    1771.7022422741
PWY0-1586:␣peptidoglycan␣maturation␣(meso-diaminopimelate␣containing)|unclassified                          56.1914822357
NONOXIPENT-PWY:␣pentose␣phosphate␣pathway␣(non-oxidative␣branch)␣I                                          1106.8757020465
NONOXIPENT-PWY:␣pentose␣phosphate␣pathway␣(non-oxidative␣branch)␣I|g__Escherichia.s__Escherichia_coli      1101.2890584400
PWY-7111:␣pyruvate␣fermentation␣to␣isobutanol␣(engineered)                                                  878.7762779699
PWY-7111:␣pyruvate␣fermentation␣to␣isobutanol␣(engineered)|g__Escherichia.s__Escherichia_coli               864.4382941456
PWY-7111:␣pyruvate␣fermentation␣to␣isobutanol␣(engineered)|unclassified                                     13.2125379296
PWY-7663:␣gondoate␣biosynthesis␣(anaerobic)                                                                 868.6446574240
PWY-7663:␣gondoate␣biosynthesis␣(anaerobic)|g__Escherichia.s__Escherichia_coli                              859.6969770836
PWY-5667:␣CDP-diacylglycerol␣biosynthesis␣I                                                                 814.0516224811
PWY-5667:␣CDP-diacylglycerol␣biosynthesis␣I|g__Escherichia.s__Escherichia_coli                              768.3404361151
PWY-5667:␣CDP-diacylglycerol␣biosynthesis␣I|unclassified                                                    18.2616079961
PWY0-1319:␣CDP-diacylglycerol␣biosynthesis␣II                                                               814.0516224811
PWY0-1319:␣CDP-diacylglycerol␣biosynthesis␣II|g__Escherichia.s__Escherichia_coli                            768.3404361151
```

“The abundance for each pathway is a recursive computation of abundances of sub-pathways with paths resolved to abundances based on the relationships and abundances of the reactions contained in each.” [1]

Path Coverage



```
#␣Pathway                                                                                        evol1.R1_Coverage
UNMAPPED                                                                                         1.0000000000
UNINTEGRATED                                                                                     1.0000000000
UNINTEGRATED|g__Escherichia.s__Escherichia_coli                                                  1.0000000000
UNINTEGRATED|unclassified                                                                        1.0000000000
PWY0-1586:␣peptidoglycan␣maturation␣(meso-diaminopimelate␣containing)                            1.0000000000
PWY0-1586:␣peptidoglycan␣maturation␣(meso-diaminopimelate␣containing)|g__Escherichia.s__Escherichia_coli    1.0000000000
PWY0-1586:␣peptidoglycan␣maturation␣(meso-diaminopimelate␣containing)|unclassified               0.9926980152
NONOXIPENT-PWY:␣pentose␣phosphate␣pathway␣(non-oxidative␣branch)␣I                               1.0000000000
NONOXIPENT-PWY:␣pentose␣phosphate␣pathway␣(non-oxidative␣branch)␣I|g__Escherichia.s__Escherichia_coli    1.0000000000
PWY-7111:␣pyruvate␣fermentation␣to␣isobutanol␣(engineered)                                       1.0000000000
PWY-7111:␣pyruvate␣fermentation␣to␣isobutanol␣(engineered)|g__Escherichia.s__Escherichia_coli    1.0000000000
PWY-7111:␣pyruvate␣fermentation␣to␣isobutanol␣(engineered)|unclassified                          0.0041671776
PWY-7663:␣gondoate␣biosynthesis␣(anaerobic)                                                      1.0000000000
PWY-7663:␣gondoate␣biosynthesis␣(anaerobic)|g__Escherichia.s__Escherichia_coli                   1.0000000000
PWY-5667:␣CDP-diacylglycerol␣biosynthesis␣I                                                      1.0000000000
PWY-5667:␣CDP-diacylglycerol␣biosynthesis␣I|g__Escherichia.s__Escherichia_coli                   1.0000000000
PWY-5667:␣CDP-diacylglycerol␣biosynthesis␣I|unclassified                                         0.0058590002
PWY0-1319:␣CDP-diacylglycerol␣biosynthesis␣II                                                    1.0000000000
PWY0-1319:␣CDP-diacylglycerol␣biosynthesis␣II|g__Escherichia.s__Escherichia_coli                 1.0000000000
```

Coverage is a confidence score assigned to each reaction detected in the community.

References

[1] https://github.com/biobakery/humann

[2] https://github.com/biobakery/MetaPhlAn

[3] http://borensteinlab.com/courses/TAU_CS_3116_B_19/presentations/7_MetaPhlan.pdf

Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Schwarzberg Lipson K, Knight R, Caporaso JG, Segata N, Huttenhower C. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods 15: 962-968 (2018).

Francesco Beghini, Lauren J McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, Paolo Manghi, Matthias Scholz, Andrew Maltez Thomas, Mireia Valles-Colomer, George Weingart, Yancong Zhang, Moreno Zolfo, Curtis Huttenhower, Eric A Franzosa, Nicola Segata. eLife (2021)