MOMS-PI

2018-11-24

Settings

HMP2 Data

https://ibdmdb.org/tunnel/public/summary.html - 178 samples

16S

- **Summary:** mapping statistics, https://ibdmdb.org/tunnel/dataset_summary/HMP2/16S/1806/summary/summary.html
- **Raw files:** `.tar` files with bgzipped FASTq paired-end files, https://ibdmdb.org/tunnel/public/HMP2/16S/1806/rawfiles
- **Products:**
  - **Taxonomic profiles (BIOM):** - 178 BIOM files
  - **Taxonomic Profiles (Text):** - 178 `.tsv` files. Example:

Taxonomy	206719
Bacteria; ___Firmicutes; ___Bacilli; ___Bacillales; ___Alicyclobacillaceae; ___Tumebacillus	0
Bacteria; ___Firmicutes; ___Clostridia; ___Clostridiales; ___Lachnospiraceae; ___Lachnospiraceae_FCS020_group	0
Bacteria; ___Proteobacteria; ___Betaproteobacteria; ___Burkholderiales; ___Comamonadaceae; ___Ramlibacter	0
Bacteria; ___Cyanobacteria; ___Melainabacteria; ___Obscuribacterales; ___f; ___g	0
Bacteria; ___Proteobacteria; ___Alphaproteobacteria; ___Rhizobiales; ___Rhizobiaceae; ___Rhizobium	0

```
- **Merged Tables:** `taxonomic_profiles.biom.gz` and `taxonomic_profiles.tsv.gz`. Content is the same,
```

```
[1] "Data dimensions"
```

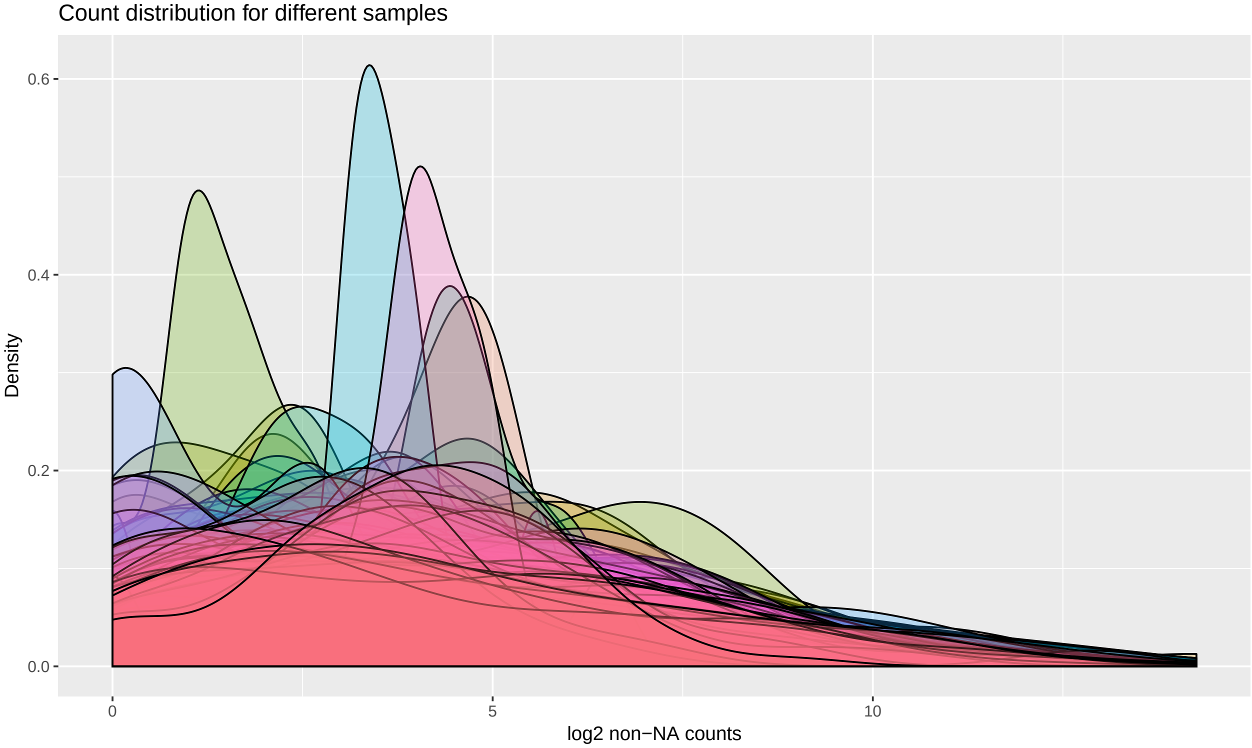
```
[1] 982 178
```

	206646	224324	206619	224326	206624
IP8BSoli	0	0	0	0	17
UncTepi3	0	0	0	0	0
Unc004ii	0	0	0	0	0
Unc00re8	0	0	0	0	0
Unc018j2	1	0	0	0	0

Count distribution

```
[1] "Summary of count data"
```

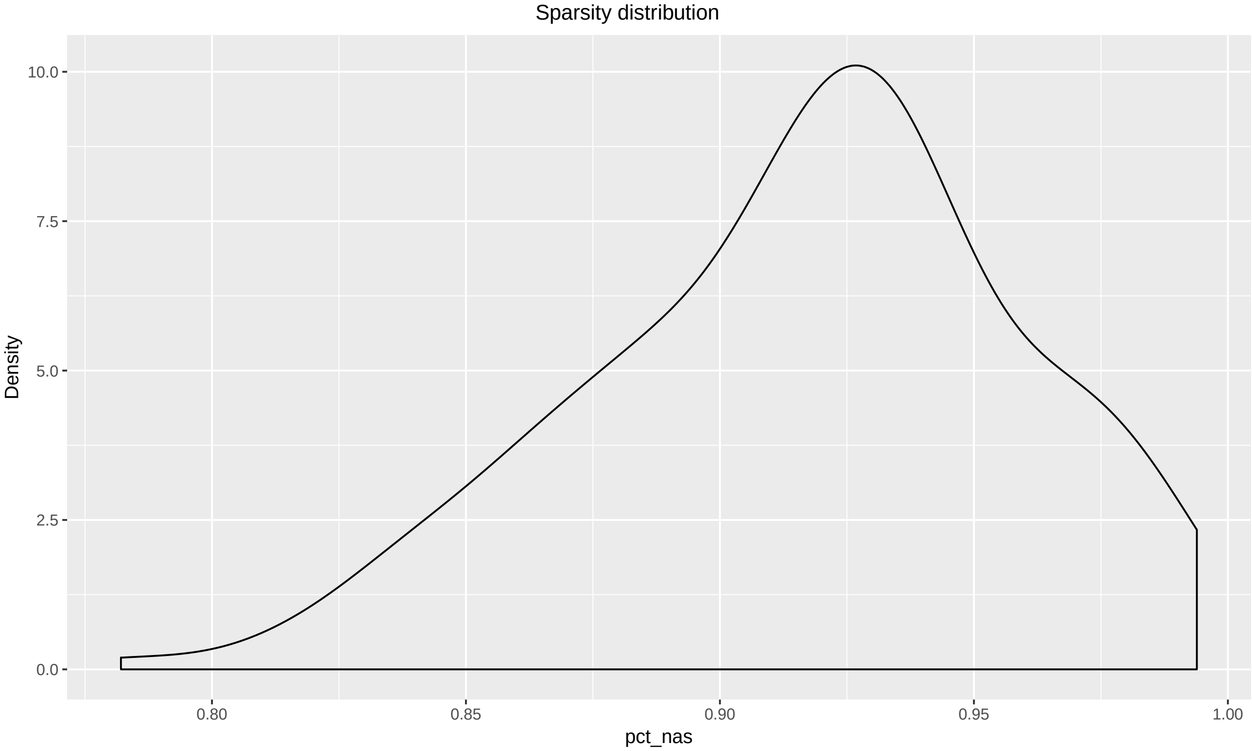
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0     0.0     0.0    15.4     0.0 19572.0 
```



Distribution of sample medians

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0       0       0       0       0       0 
```

Sparsity



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7821  0.8872  0.9221  0.9166  0.9440  0.9939
```

Questions

- What are the row IDs?
- Most sample counts are very small, median = 0. Is it expected?
- There are outliers, like a sample with maximum count = 19572 (while mean is ~15). Shall we remove such samples?
- The data is very sparse (~92% zeros) - is it expected?

Serology

`hmp2_serology_Compiled_ELISA_Data.tsv`, columns are samples

```
[1] "Data dimensions"
```

```
[1]  14 212
```

Serum ID	206454	206458	206459	206460
Site	Harvard	Harvard	Harvard	Harvard
Plate	1	1	1	1
Sample	1	2	3	4
IgA ASCA EU	0	0	0	49
IgA ASCA Pos.	0	0	0	1

Metagenomes

Lots of files. Look at Products, Merged Tables, https://ibdmdb.org/tunnel/public/HMP2/WGS/1818/products

taxonomic_profiles.tsv

```
[1] "Data dimensions"
```

```
[1] 1479 1639
```

#SampleID	CSM5FZ4M	CSM5MCUO	CSM5MCVL	CSM5MCVN
k__Archaea	0	0	0	0
k__Archaea|p__Euryarchaeota	0	0	0	0
k__Archaea|p__Euryarchaeota|c__Methanobacteria	0	0	0	0
k__Archaea|p__Euryarchaeota|c__Methanobacteria|o__Methanobacteriales	0	0	0	0
k__Archaea|p__Euryarchaeota|c__Methanobacteria|o__Methanobacteriales|f__Methanobacteriaceae	0	0	0	0

pathabundances.tsv

```
[1] "Data dimensions"
```

```
[1] 10884  1639
```

# Pathway	CSM5FZ4M	CSM5MCUO	CSM5MCVL	CSM5MCVN
1CMET2-PWY: N10-formyl-tetrahydrofolate biosynthesis	0.01581	0.01017	0.01674	0.018
1CMET2-PWY: N10-formyl-tetrahydrofolate biosynthesis|g___Akkermansia.s___Akkermansia_muciniphila	0	0	0	0
1CMET2-PWY: N10-formyl-tetrahydrofolate biosynthesis|g___Bacteroides.s___Bacteroides_barnesiae	0	0	0	0
1CMET2-PWY: N10-formyl-tetrahydrofolate biosynthesis|g___Bacteroides.s___Bacteroides_caccae	0	0.0001569	0	0
1CMET2-PWY: N10-formyl-tetrahydrofolate biosynthesis|g___Bacteroides.s___Bacteroides_cellulosilyticus	0	0	0	0

ecs.tsv

```
[1] "Data dimensions"
```

```
[1] 108433   1639
```

# Gene Family	CSM5FZ4M	CSM5MCUO	CSM5MCVL	CSM5MCVN
1.1.1.100: 3-oxoacyl-[acyl-carrier-protein] reductase	0.001835	0.001472	0.001571	0.001173
1.1.1.100: 3-oxoacyl-[acyl-carrier-protein] reductase|g___Aggregatibacter.s___Aggregatibacter_segnis	0	0	0	0
1.1.1.100: 3-oxoacyl-[acyl-carrier-protein] reductase|g___Alloprevotella.s___Alloprevotella_tannerae	0	0	0	0
1.1.1.100: 3-oxoacyl-[acyl-carrier-protein] reductase|g___Anaerococcus.s___Anaerococcus_obesiensis	0	0	0	0
1.1.1.100: 3-oxoacyl-[acyl-carrier-protein] reductase|g___Anaerococcus.s___Anaerococcus_vaginalis	0	0	0	0

Proteomics

https://ibdmdb.org/tunnel/public/HMP2/Proteomics/1633/products

HMP2_proteomics_ecs.tsv

```
[1] "Data dimensions"
```

```
[1] 910 451
```

Gene	CSM5FZ3N	CSM5FZ3T	CSM5FZ44	CSM5FZ48
UNGROUPED	1769	1128	2765	1183
1.1.1.1: Alcohol dehydrogenase	0	0	0	0
1.1.1.100: 3-oxoacyl-[acyl-carrier-protein] reductase	0	0	0	0
1.1.1.103: L-threonine 3-dehydrogenase	0	0	0	0
1.1.1.127: 2-dehydro-3-deoxy-D-gluconate 5-dehydrogenase	0	0	0	0

HMP2__proteomics__kos.tsv.gz

```
[1] "Data dimensions"
```

```
[1] 1823  451
```

KO	CSM5FZ3N	CSM5FZ3T	CSM5FZ44	CSM5FZ48
UNGROUPED	1595	1134	2771	954
K00003: homoserine dehydrogenase [EC:1.1.1.3]	0	0	0	0
K00005: glycerol dehydrogenase [EC:1.1.1.6]	0	0	0	0
K00008: L-iditol 2-dehydrogenase [EC:1.1.1.14]	0	0	0	0
K00009: mannitol-1-phosphate 5-dehydrogenase [EC:1.1.1.17]	0	0	0	0

Viromics

https://ibdmdb.org/tunnel/public/HMP2/Viromics/1732/products

taxonomic__profiles.tsv.gz

```
[1] "Data dimensions"
```

```
[1]  56 330
```

#ID	CSM5MCXD	CSM5LLMS	MSM5LLCC	CSM5MCXH
k___Viruses|p___Viruses_noname|c___Viruses_noname|o___Caudovirales|f___Myoviridae|g___Mulikevirus|s___Escherichia_	0	0	0	0
k___Viruses|p___Viruses_noname|c___Viruses_noname|o___Caudovirales|f___Myoviridae|g___Mulikevirus|s___Mulikevirus_	0	0	0	0
k___Viruses|p___Viruses_noname|c___Viruses_noname|o___Caudovirales|f___Myoviridae|g___T4likevirus|s___Klebsiella_p	0	0	0	0
k___Viruses|p___Viruses_noname|c___Viruses_noname|o___Caudovirales|f___Podoviridae|g___Epsilon15likevirus|s___Epsil	0	0	0	0
k___Viruses|p___Viruses_noname|c___Viruses_noname|o___Caudovirales|f___Siphoviridae|g___C2likevirus|s___C2likevirus_	0	0	0	0

virome__virmap__analysis.tsv.gz

```
[1] "Data dimensions"
```

```
[1] 260 704
```

Virus	MSM5LLDS	CSM5FZ4M	HSM5FZBE	CSM5MD5B
superkingdom=Viruses;dsDNA viruses, no RNA stage;family=Adenoviridae;genus=Mastadenovirus;species=Human mastadenovirus A;taxId=129875	0	0	0	0
superkingdom=Viruses;dsDNA viruses, no RNA stage;family=Baculoviridae;genus=Alphabaculovirus;species=Autographa californica multiple nucleopolyhedrovirus;taxId=307456	0	0	0	0
superkingdom=Viruses;dsDNA viruses, no RNA stage;family=Baculoviridae;genus=Alphabaculovirus;taxId=558016	0	0	0	0

Virus	MSM5LLDS	CSM5FZ4M	HSM5FZBZ	CSM5MD5B
superkingdom=Viruses;dsDNA viruses, no RNA stage;order=Caudovirales;family=Myoviridae;genus=Felixo1virus;species=Salmonella phage FelixO1;taxId=77775	0	0	0	0
superkingdom=Viruses;dsDNA viruses, no RNA stage;order=Caudovirales;family=Myoviridae;genus=Felixo1virus;unclassified FelixO1likevirus;species=Escherichia phage vB_EcoM_AYO145A;taxId=1636202	0	0	0	0

Metabolites

https://ibdmdb.org/tunnel/public/HMP2/Metabolites/1723/products

HMP2_metabolomics.csv.gz

```
[1] "Data dimensions"
```

```
[1] 81867   553
```

Method	Pooled QC sample CV	m/z	RT	HMDB (*Representative ID)
C18-neg	0.02719	313.2	9.75	HMDB04705
C18-neg	0.02641	313.2	9.95	HMDB04704
C18-neg	0.04684	115.1	5.79	HMDB00535
C18-neg	0.07408	129.1	7.43	HMDB00666
C18-neg	0.02414	149.1	5.81	HMDB00764

HMP2_Pilot Data

16S

taxonomic_profiles.biom.gz

```
[1] "Data dimensions"
```

```
[1] 503  58
```

	CSM5FZ3N	CSM5FZ3X	CSM5FZ3Z	CSM5FZ46	CSM5FZ4G
k___Bacteria; p___Firmicutes; c___Clostridia; o___Clostridiales; f___; g___; s___:390820	0	0	0	0	0
k___Bacteria; p___Firmicutes; c___Clostridia; o___Clostridiales; f___; g___; s___:369429	0	0	0	0	0

	CSM5FZ3N	CSM5FZ3X	CSM5FZ3Z	CSM5FZ46	CSM5FZ4G
k___Bacteria; p___Firmicutes; c___Clostridia; o___Clostridiales; f___Ruminococcaceae; g___Ruminococcus; s___:363646	0	0	0	0	0
k___Bacteria; p___Bacteroidetes; c___Bacteroidia; o___Bacteroidales; f___Bacteroidaceae; g___Bacteroides; s___caccae:195508	0	0	0	0	0
k___Bacteria; p___Firmicutes; c___Clostridia; o___Clostridiales; f___Lachnospiraceae; g___Dorea; s___:577406	0	0	0	0	0