GALAXY

GUT-AND-LIVER AXIS IN ALCOHOLIC LIVER FIBROSIS
GRANT NUMBER 668031

DELIVERABLE NUMBER: D4.2
DELIVERABLE DUE DATE: 30 June 2019
COMPLETION DATE OF DELIVERABLE: 27 June 2019
DISSEMINATION LEVEL: PUBLIC

DOCUMENT MAIN AUTHOR: WP4, Peer Bork (EMBL)
DOCUMENT SIGNED OFF BY: Project manager Louise Skovborg Just (University of Southern Denmark)

TOOL DEVELOPED FOR INTEGRATIVE METAGENOMICS-BASED METATRANSCRIPTOMICS DATA ANALYSIS

We have developed a tool for an integrative analysis of metagenomic and metatranscriptomic data. Metagenomics readouts are the current standard for microbiome analysis. Based on microbial DNA recovered from a (e.g. fecal) sample, the composition of the community and its functional potential can be determined. However, species detected by their DNA may not be metabolically active anymore. In contrast, RNA-based metatranscriptomics readouts quantify the actual genes expressed by a community, giving an indication of the metabolic activity of community members and a much more accurate functional readout. In addition, these readouts allow us to distinguish between species that are currently metabolically active and those that are not (i.e. dormant, dead, etc.). RNA decays quickly after cell death. This is an advantage as only recently active species are detected with metatranscriptomics, but also a disadvantage as samples have to be frozen very quickly to avoid RNA degradation. For this reason, metatranscriptomics data may suffer from more noise than metagenomics data.

To integrate metatranscriptomic and metagenomic data, we have developed a tool called “MetaGTi” that enables exploration of the relationships between these datasets in the context of species abundances. This tool produces gene and species abundance estimates from metatranscriptomic data and compares these values to species abundance estimates from metagenomic data (when both are available). This tool knits together new and existing software that has been developed in our group.

MetaGTi performs all analytical steps for the user with one command, which streamlines analysis and ensures metagenomic and metatranscriptomic datasets are processed in a way that makes their results comparable. The main steps of the pipeline are: quality control, gene abundance profiling, species abundance profiling, species abundance prediction, and integrative analysis of metagenomic and metatranscriptomic results. The products of MetaGTi are predicted species and gene abundances, and an analysis report comparing species abundances estimated from metagenomic and metatranscriptomic. This tool is open-source and publicly available at https://git.embl.de/grp-bork/metaGTi .

Input to this tool are raw sequencing results (fastq files). These samples are first preprocessed using NGLess (Coelho et al. 2019), which includes quality-based trimming and removing human genetic sequences. High quality reads are then mapped against a reference database of bacterial genomes. These genomes were obtained from proGenomes (Mende et al. 2017), filtered to remove contigs smaller than 100 base-pairs and reduced to a subset of species found to be associated with the human gut. The mappings are then used to produce species and gene abundance profiles. Orthologous gene count matrices (at bactNOG level) are generated using orthology and functional annotations produced by eggNOG-mapper (Huerta-Cepas et al. 2017) and NGLess. Marker-gene based species abundance profiles are generated using mOTU profiler (version 2) (Milanese et al. 2019). A model to predict species abundances based on metatranscriptomic data was trained and benchmarked using data from the

iHMP project on inflammatory bowel disease (Integrative HMP (iHMP) Research Network Consortium 2014). Other datasets with paired metagenomic and metatranscriptomic data were assessed but not used due to poor quality or low sample sizes. See modelling documentation for full description (https://git.embl.de/grp-bork/metaGTi/tree/master/documentation/buildingModel).

Metagenomic and metatranscriptomic species abundances are compared using an Rmarkdown script that produces a self-contained report. This includes data on the correlations between metagenomic and metatranscriptomic data, both overall and when broken down by sample and by taxonomic levels (from phylum to species). Often, species detected in metagenomic data are not detected in metatranscriptomic data, and vice-versa. This situation is investigated directly and discrepancies are broken down by species and by taxonomic order. As an illustrative example, we provide an example of such a report that compares the species abundance results of the mOTU profiler, using a subset of data from the iHMP project (see Appendix 1).

References

Coelho, Luis Pedro, Renato Alves, Paulo Monteiro, Jaime Huerta-Cepas, Ana Teresa Freitas, and Peer Bork. 2019. "NG-Meta-Profiler: Fast Processing of Metagenomes Using NGLess, a Domain-Specific Language." *Microbiome* 7 (1): 84. https://doi.org/10.1186/s40168-019-0684-8.

Huerta-Cepas, Jaime, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, and Peer Bork. 2017. "Fast Genome-Wide Functional Annotation through Orthology Assignment by EggNOG-Mapper." *Molecular Biology and Evolution* 34 (8): 2115–22. https://doi.org/10.1093/molbev/msx148.

Integrative HMP (iHMP) Research Network Consortium. 2014. "The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease." *Cell Host & Microbe* 16 (3): 276–89. https://doi.org/10.1016/j.chom.2014.08.014.

Mende, Daniel R., Ivica Letunic, Jaime Huerta-Cepas, Simone S. Li, Kristoffer Forslund, Shinichi Sunagawa, and Peer Bork. 2017. "ProGenomes: A Resource for Consistent Functional and Taxonomic Annotations of Prokaryotic Genomes." *Nucleic Acids Research* 45 (D1): D529–34. https://doi.org/10.1093/nar/gkw989.

Milanese, Alessio, Daniel R. Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Pascal Hingamp, et al. 2019. "Microbial Abundance, Activity and Population Genomic Profiling with MOTUs2." *Nature Communications* 10 (1): 1014. https://doi.org/10.1038/s41467-019-08844-4.

Comparison of species abundances derived from metagenomic and metatranscriptomic data

Contents

1 Data sources

This document presents a comparison of the species abundances predicted for microbiome samples that have been sequenced using metagenomic and metatranscriptomic approaches. The metagenomic and metatranscriptomic species abundance profiles have been compiled using the mOTUsv2 profiler.

File containing metagenomic (metaG) species abundance predictions: /output/all_samples.motusv2.relabund.metaG.tsv

File containing metatranscriptomic (metaT) species abundance predictions: /output/all_samples.motusv2.relabund.metaT.tsv

2 Basic statistics

Number of samples in metaG data: 1447

Number of samples in metaT data: 1447

Number of samples pairs: 1440

Number of species tested for presence: 7726

Number of species observed in metagenomic data at any abundance: 1665

Number of species observed in metatranscriptomic data at any abundance: 1440

2.1 Sample-species pairs

The abundance of a "sample-species pair" is the basic unit of observation in this analysis. It is the abundance of a particular species in a particular sample (e.g. SampleA-Species1, SampleA-Species2, etc.).

Number of sample-species pairs observed with abundance > 0 in metagenomic data: 178918

Number of sample-species pairs observed with abundance > 0 in metatranscriptomic data: 83775

2.2 Range of species abundance values

2.2.1 Metagenomic data:

All data:

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0000000 0.0000000 0.0000000 0.0001291 0.0000000 1.0000000
```

Metagenomic data where abundances > 0:

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0000279 0.0005821 0.0013927 0.0080260 0.0048096 1.0000000
```

2.2.2 Metatranscriptomic data:

All data:

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
0.0000000 0.0000000 0.0000000 0.0001288 0.0000000 1.0000000
```

Metatranscriptomic data where abundances > 0:

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000019 0.001523 0.004786 0.017105 0.014351 1.000000
```

2.3 Detection threshold

2.3.1 By Sample

This table summarises the distribution of the minimum species abundance values across all samples.

dataType	minMinAbund	medianMinAbund	meanMinAbunde	maxMinAbund
metaG	0.000028	0.00022	0.0014	1
metaT	0.000019	0.00228	0.0138	1

Based on the above values, the maximum of the median of the minimum values was selected as the detection threshold.

Anything less than 0.2% abundance might be absent in one of the datasets due to different read depths, so we will consider values less than this as 0 (i.e. detection threshold is 0.002).

Number of species observed in metagenomic data above detection threshold: 1315

Number of species observed in metatranscriptomic data above detection threshold: 1189

2.4 Sample-species pairs above detection threshold

Number of sample-species pairs that were detected by either metaT or metaG or both: 92253

Number of sample-species pairs that were detected by both datasets: 40274 (44% of 92253)

Number of sample-species pairs that were detected in metaG but were not detected in metaT: 33921 (37% of 92253)

Number of sample-species pairs that were detected in metaT but were detected in metaG: 18058 (20% of 92253)

3 Comparison of number of species detected by metagenomic vs metatranscriptomic data

Out of a total of 7726 species that could have been detected, 1108 species were detected in **both** the metagenomic and metatranscriptomic data.

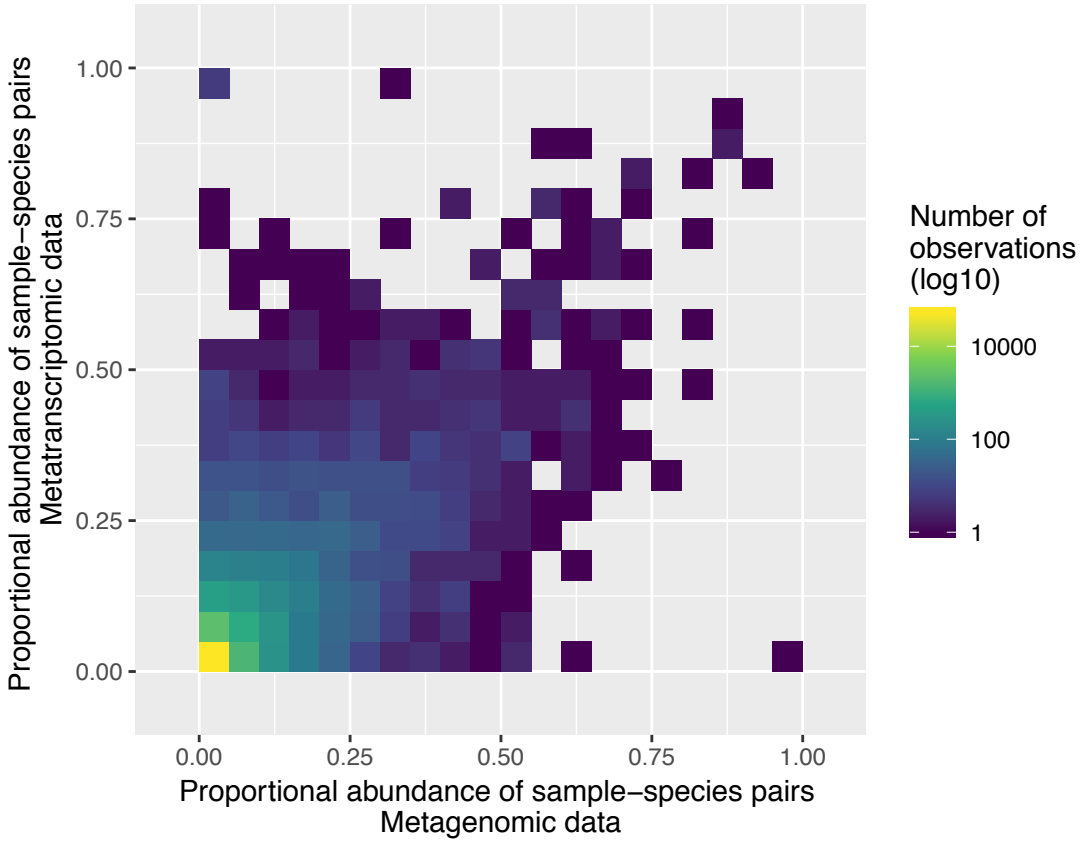
Out of 1315 species detected in the metagenomic data, 207 (16%) were not detected in the metatranscriptomic data.

Out of 1189 species detected in the metatranscriptomic data, 81 (7%) were not detected in the metagenomic data.

4 Agreement between metagenomic and metatranscriptomic abundances

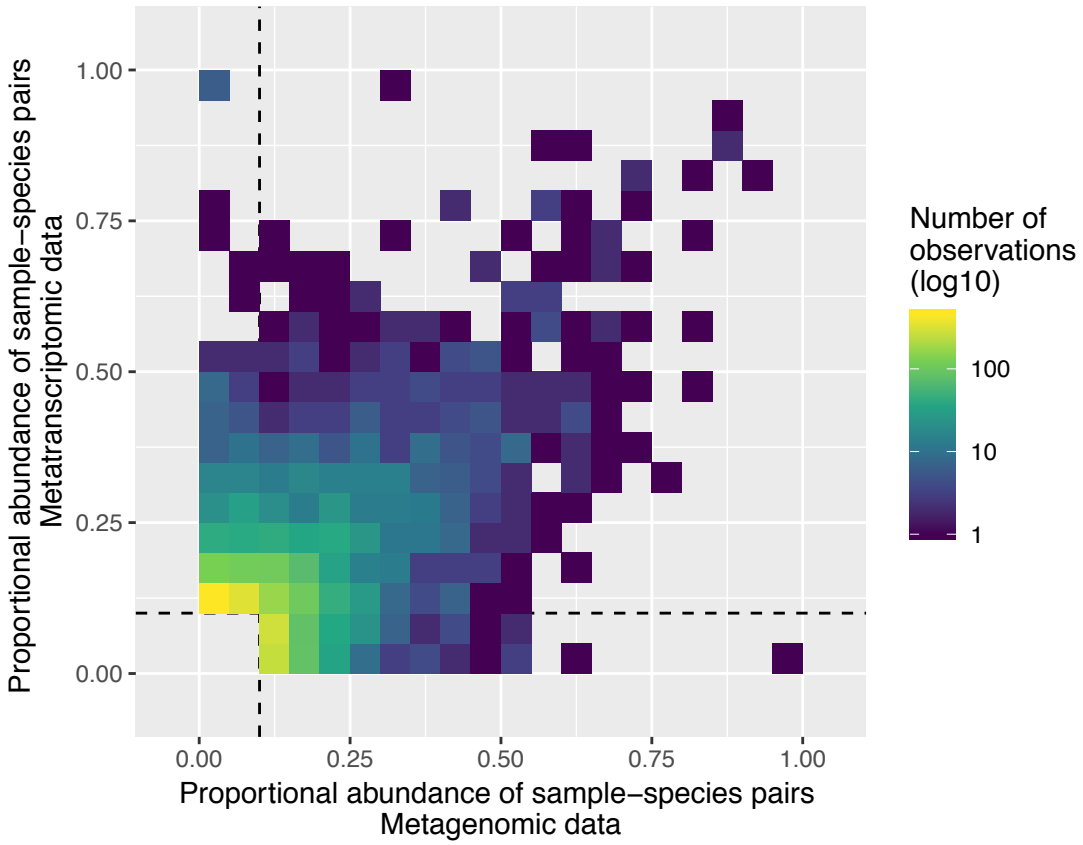
4.1 Visualisation of abundance value distributions

Sample-species pairs observed in at least one dataset above detection threshold. Note log scale in colour scale used figure below.



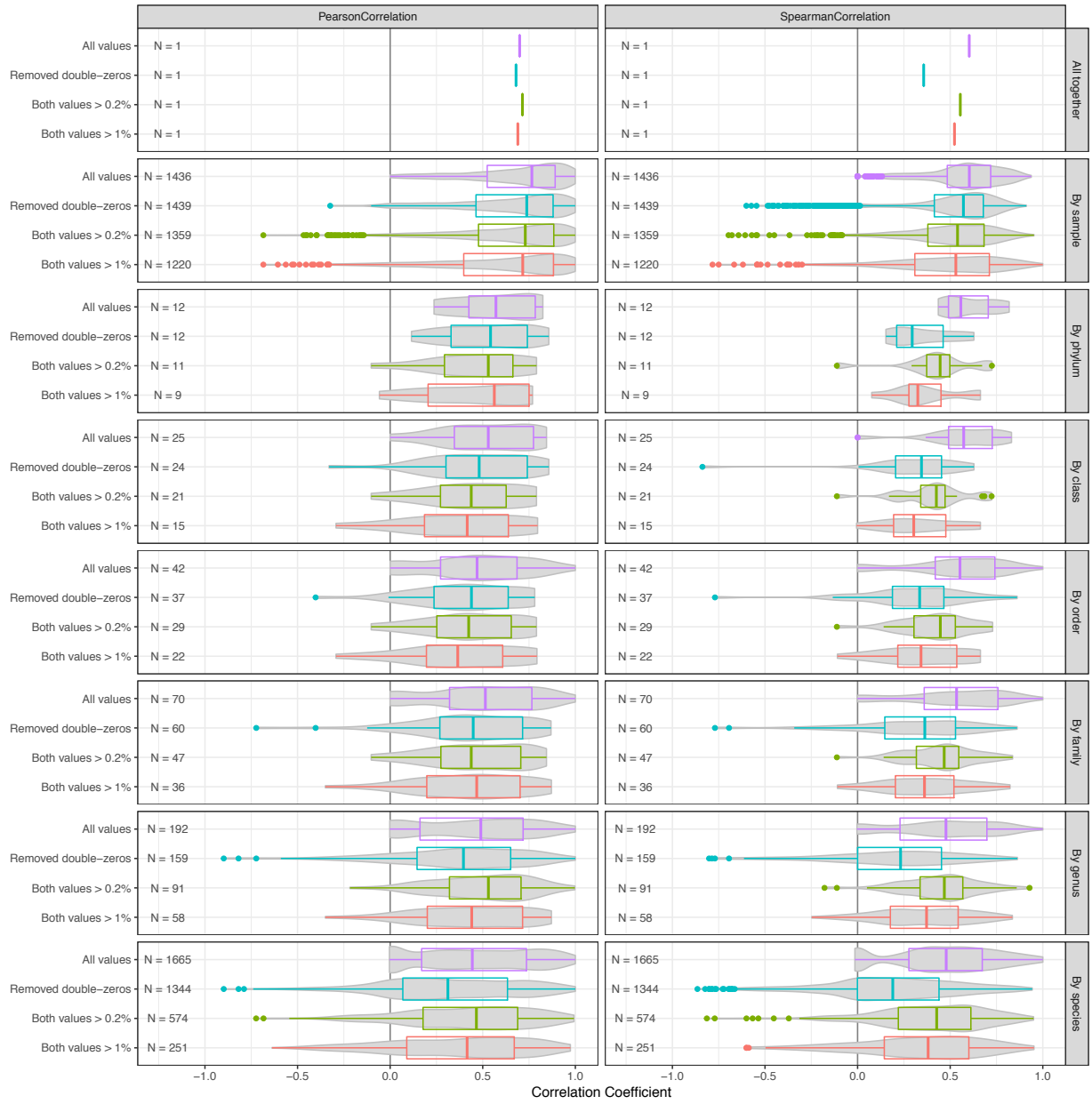
4.2 Visualisation of very high abundance values

Sample-species pair observed in at least one dataset at a high abundance (>10%). Note log scale in colour scale used figure below.



4.3 Correlation between metagenomic and metatranscriptomic values

Here, samples-species pairs are only kept if **both** metaG and metaT have abundances greater than the indicated threshold. Groups (e.g. samples, species, phyla, etc) are only considered if metaG and metaT both have more than 5 non-zero observations.



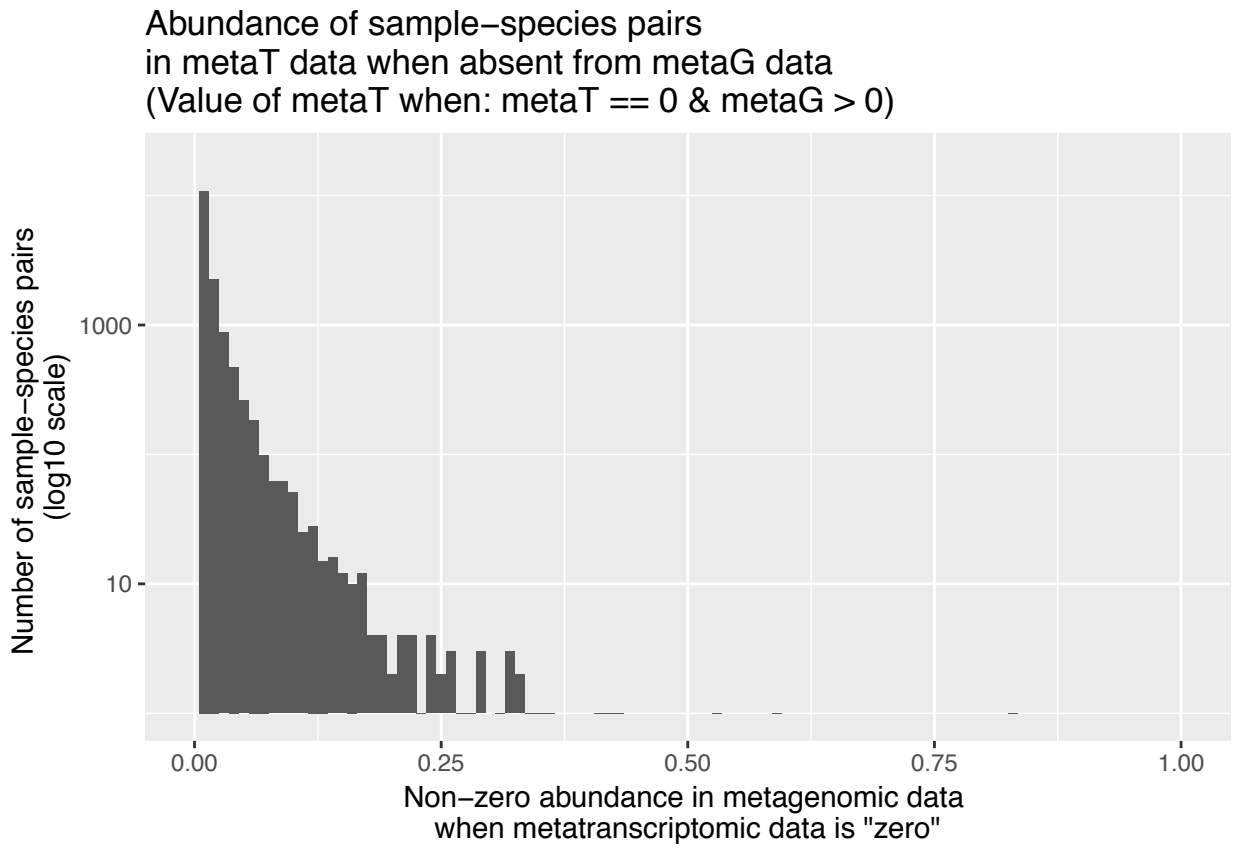
5 Mismatched "zeros"

Here we investigate cases where a species-sample pair is present according to metaG or metaT data, but is absent in the other dataset (abundance is below detection threshold).

5.1 Present in metagenomic data, absent in metatranscriptomic data

Distribution of non-zero metagenomic abundances when metatranscriptomic abundance is < 0.002.

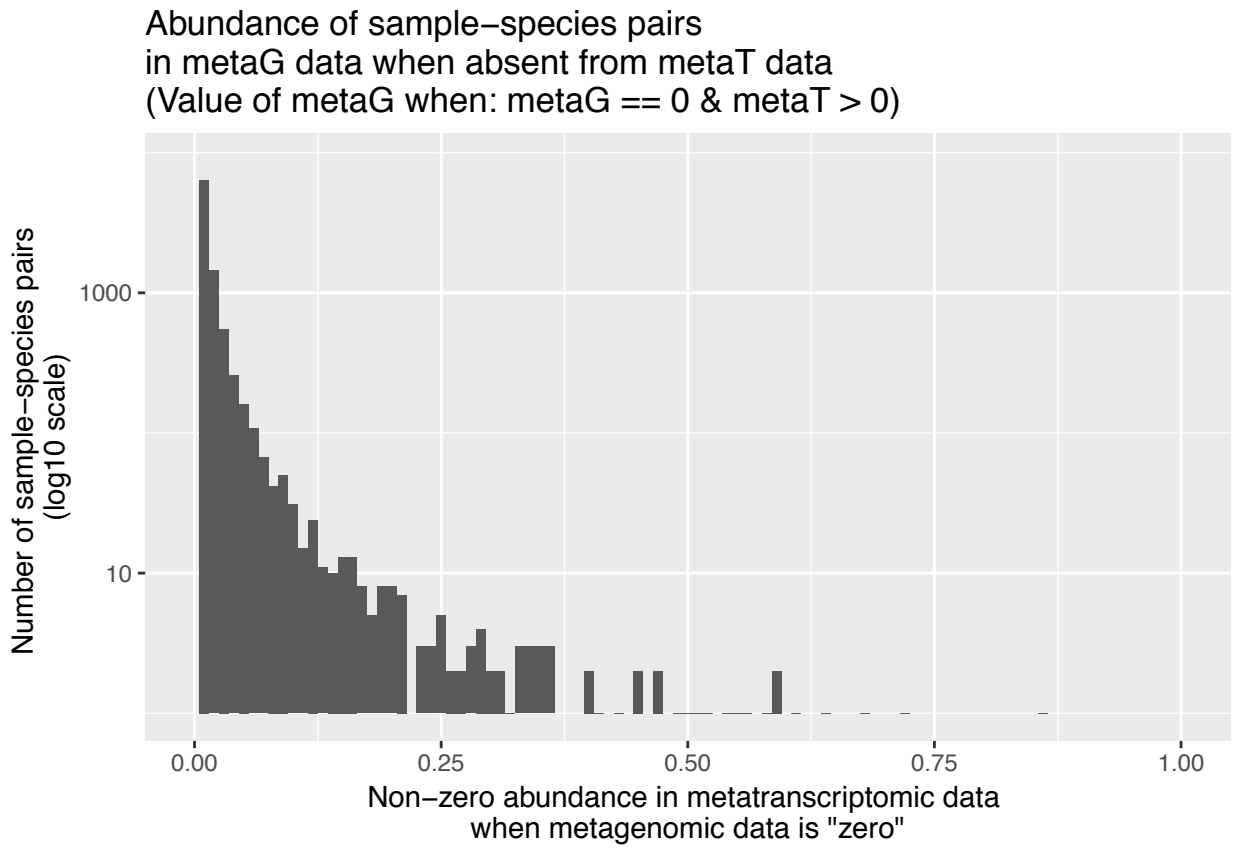
```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002000 0.002850 0.004506 0.009207 0.008854 0.830148
```



5.2 Present in metatranscriptomic data, absent in metagenomic data

Distribution of non-zero metatranscriptomic abundances when metagenomic abundance is < 0.002.

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002000 0.003127 0.005161 0.011641 0.010204 1.000000
```



5.3 Large differences between mismatched zeros

Some of this may be due to noise around the detection level. To focus on cases that are more likely to be biologically valid, we will look at cases where the non-zero abundance is > 0.02 (2 percentage points higher than 0)

whichMissing	numberOfSampleSpeciesPairsWithBigDifferences
metaG=0_metaT>0	2226
metaG>0_metaT=0	7864

whichMissing	source	meanAbund	medianAbund
metaG=0_metaT>0	metaG	0.0000000	0.0000000
metaG=0_metaT>0	metaT	0.0699009	0.0389765
metaG>0_metaT=0	metaG	0.0356019	0.0272266
metaG>0_metaT=0	metaT	0.0087736	0.0000000

5.3.1 Patterns by taxonomy

The tables below display the proportion of sample-species value pairs that either:

- match as non-zeros (`matched[G>0_T>0]`)
- mismatch and one value is big ($> 2\%$ abundance) (`bigDiff[G=0_T>0]` and `bigDiff[G>0_T=0]`)

The proportion of sample-species pairs that mismatch and one value is small (< 2% abundance) are not shown.

The `nObs` value ("number of observations") is the number of sample-species pairs that fell into one of the above categories (i.e. not matched zeros).

Results are only shown for groups that had at least 10 observations.

5.3.1.1 By species

Only displaying top 20, sorted by proportion of observations that have a "big" difference.

species	nObs	matched[G>0_T>0]	bigDiff[G=0_T>0]	bigDiff[G>0_T=0]
Butyrivibrio crossotus [ref_4383]	76	0.20	0.01	0.39
unknown Clostridiales [meta_5804]	14	0.21	0.29	0.00
Butyrivibrio sp. CAG:318 [meta_5702]	61	0.28	0.00	0.28
unknown Clostridiales [meta_6465]	11	0.45	0.00	0.27
Eubacterium sp. CAG:202 [meta_7449]	127	0.50	0.02	0.26
Actinomyces viscosus [ref_1444]	71	0.03	0.00	0.25
unknown Faecalibacterium [meta_7492]	17	0.35	0.00	0.24
unknown Firmicutes [meta_6909]	30	0.33	0.10	0.23
unknown Bacteroidales [meta_6591]	13	0.54	0.23	0.08
unknown Firmicutes [meta_6091]	115	0.52	0.01	0.23
unknown Ruminococcus [meta_5392]	77	0.30	0.13	0.22
Actinomyces sp. [ref_5209]	83	0.11	0.00	0.22
Niameybacter massiliensis [meta_7610]	14	0.50	0.21	0.07
unknown Clostridium [meta_7253]	19	0.11	0.00	0.21
unknown Flavobacteriia [meta_6771]	19	0.16	0.21	0.00
unknown Prevotella [meta_5555]	34	0.59	0.21	0.18
Enterococcus faecium [ref_0372]	15	0.13	0.00	0.20
unknown Bacteroidales [meta_5655]	21	0.29	0.14	0.19
Eubacterium rectale [ref_1416]	1028	0.55	0.00	0.19
Staphylococcus sp. CAG:324 [meta_7772]	37	0.27	0.19	0.05

5.3.1.2 By order

The value `NA` appears for meta_mOTUs and for the `-1` fraction. Sorted by proportion of observations that have a "big" difference.

order	nObs	matched[G>0_T>0]	bigDiff[G=0_T>0]	bigDiff[G>0_T=0]
1235850 Methanomassiliicoccales	39	0.31	0.15	0.00
2037 Actinomycetales	858	0.08	0.00	0.12
649776 Synergistales	27	0.37	0.11	0.00
85009 Propionibacteriales	19	0.00	0.00	0.11
2158 Methanobacteriales	172	0.40	0.09	0.01
1843488 Acidaminococcales	716	0.43	0.00	0.08
1843489 Veillonellales	1379	0.45	0.00	0.08
85004 Bifidobacteriales	766	0.21	0.00	0.07
48461 Verrucomicrobiales	279	0.58	0.03	0.07
91347 Enterobacteriales	677	0.43	0.07	0.04
72274 Pseudomonadales	16	0.25	0.06	0.00
186802 Clostridiales	20518	0.40	0.02	0.05
213115 Desulfovibrionales	691	0.29	0.04	0.00

order	nObs	matched[G>0_T>0]	bigDiff[G=0_T>0]	bigDiff[G>0_T=0]
909929 Selenomonadales	162	0.29	0.01	0.04
80840 Burkholderiales	1106	0.62	0.04	0.00
85006 Micrococcales	220	0.53	0.00	0.04
171549 Bacteroidales	18248	0.69	0.03	0.03
1737405 Tissierellales	64	0.41	0.00	0.03
meta_mOTU_v2	37628	0.33	0.02	0.03
203491 Fusobacteriales	524	0.57	0.02	0.00
1643822 Eggerthellales	240	0.12	0.02	0.00
526525 Erysipelotrichales	390	0.22	0.02	0.02
213849 Campylobacterales	112	0.42	0.02	0.02
1385 Bacillales	362	0.63	0.02	0.00
84999 Coriobacteriales	609	0.23	0.00	0.02
85007 Corynebacteriales	68	0.07	0.00	0.01
200644 Flavobacteriales	112	0.46	0.01	0.00
135625 Pasteurellales	474	0.45	0.00	0.01
186826 Lactobacillales	3472	0.47	0.01	0.01
206351 Neisseriales	498	0.40	0.00	0.01
NA	1436	0.95	0.01	0.00
NA Bacteria order incertae sedis	342	0.23	0.01	0.01