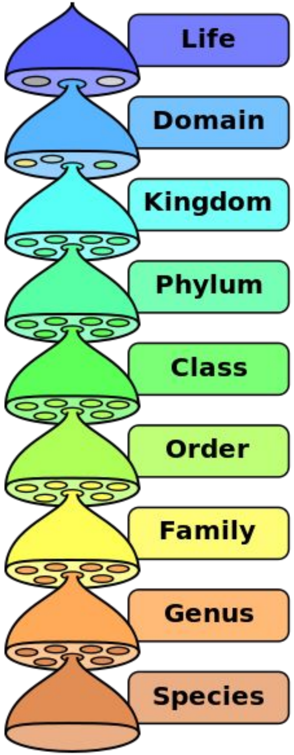
MetaPhlAn2

Konur Bayrak & Qiulin Zhang

Introduction

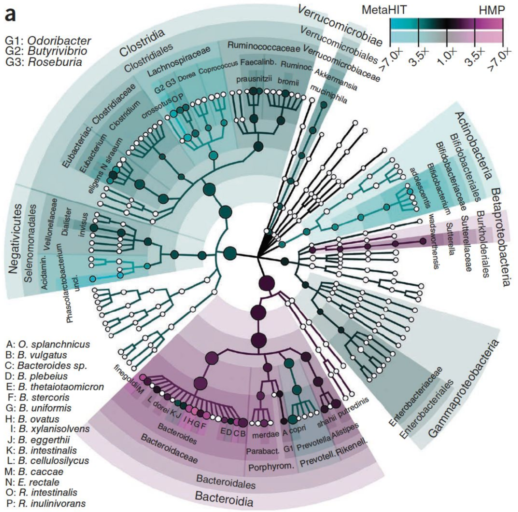


- Metagenomic shotgun sequencing provides a uniquely rich profile of microbial communities
  - Each data set yields billions of short reads sampled from DNA in the community
- Taxonomic composition can be estimated from such data by assigning each read to the most plausible microbial lineage (better taxonomic resolution than profiling 16S ribosomal RNA marker gene)
- Existing taxonomic profiling methods are inefficient for increasingly large data sets

Introduction

- Both alignment and composition based approaches (and hybrids) have been developed for this task of taxonomic classification
- None of these methods achieve both the efficiency and species level accuracy required by high complexity data sets due to computational limitations:
  - Weak accuracy for short reads (<400 nucleotides)
  - Need for clade-specific normalization
- These existing taxonomic profiling methods are inefficient for increasingly large and complex data sets

MetaPhlAn - An Overview



- Estimates microbial relative abundances by mapping metagenomic reads against a catalog of clade specific marker sequences
  - These clades can be species specific or as broad as phyla
- The clade specific markers are coding sequences that must satisfy the stringent conditions:
  - Being strongly conserved within the clade's genomes
  - Not possessing substantial local similarity with an sequence outside the clade

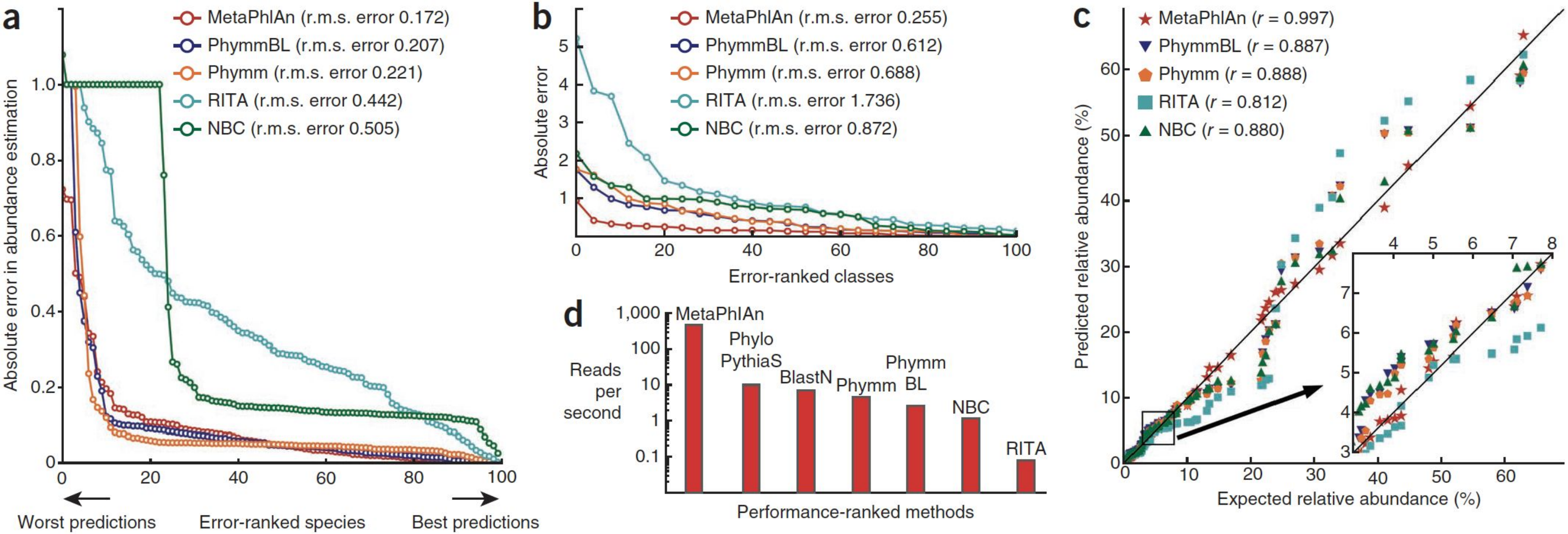
The Reference Marker Catalog

- The marker catalog spans 1,221 species with an average of 231 markers per species (2012)
- In total, 375 of 652 genera, 80 of 278 families, and 22 of 130 orders have >250 markers
  - This allows MetaPhlAn to recover relative abundances within broader clades even in the absence of sequenced genomes for all organisms in a community
- The catalog culmination process is an offline procedure and is updated regularly as newly sequenced microbial genomes become available
- The catalog downloads automatically with the associated classifier

MetaPhlan Classifier - Overview

- The classifier compares metagenomic reads against the marker catalog using nucleotide BLAST searches
    - Provides clade abundances for one or more sequenced metagenomes
- This achieves a 2-fold speedup compared to applying BLAST to the full catalog of microbial genomes
    - Because of the reduce reference catalog size
- To estimate clade relative abundance in terms of cell counts, the classifier normalizes the total number of reads in each clade by the nucleotide length of its markers

Comparison to other methods



Installation

1. **Through Bioconda:**

```
$ conda install metaphlan2
```

OR

2. **Through Docker:**

```
$ hg clone https://bitbucket.org/biobakery/metaphlan2

$ docker pull segatalab/metaphlan2
```

Using MetaPhlAn

Metaphlan2.py <options> [arguments] [input file]

options:

--input_type {fastq,fasta,multifasta,multifastq,bowtie2out,sam}

--bowtie2db The BowTie2 database file of the MetaPhlAn database.

Used if --input_type is fastq, fasta, multifasta, or multifastq

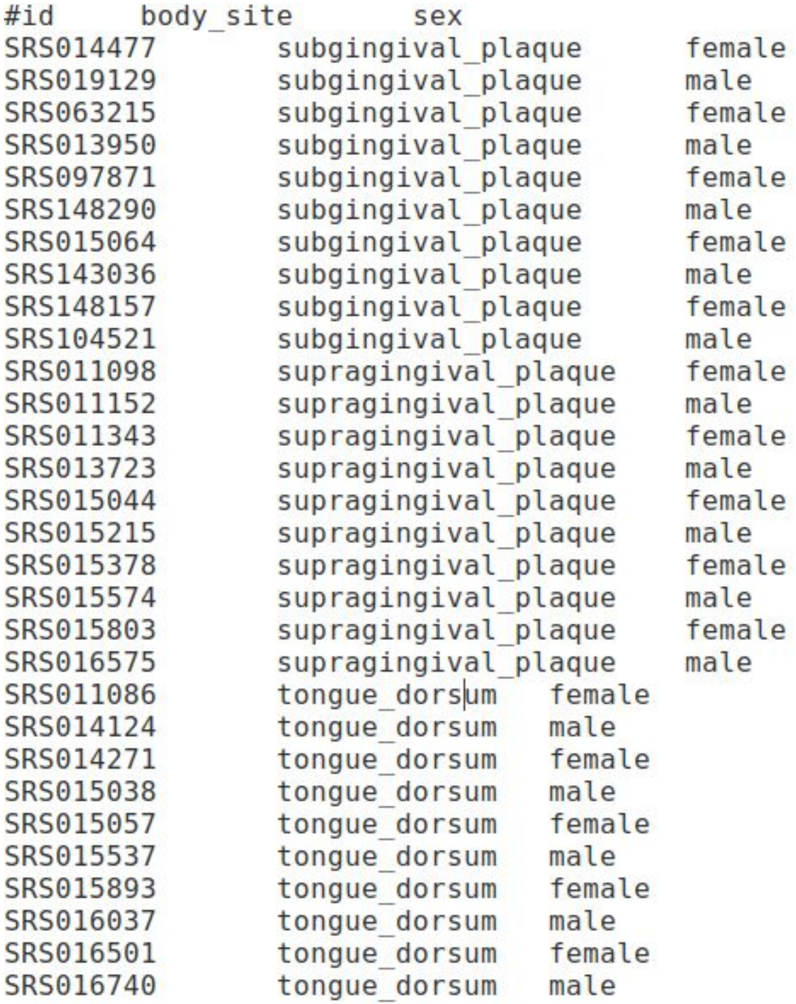
Using MetaPhlAn (Cout.)

--bowtie2out FILE_NAME The file for saving the output of BowTie2

--no_map Avoid storing the --bowtie2out map file

--nproc N The number of CPUs to use for parallelizing the mapping

-o output file The output file



```
#id	body_site	sex
SRS014477	subgingival_plaque	female
SRS019129	subgingival_plaque	male
SRS063215	subgingival_plaque	female
SRS013950	subgingival_plaque	male
SRS097871	subgingival_plaque	female
SRS148290	subgingival_plaque	male
SRS015064	subgingival_plaque	female
SRS143036	subgingival_plaque	male
SRS148157	subgingival_plaque	female
SRS104521	subgingival_plaque	male
SRS011098	supragingival_plaque	female
SRS011152	supragingival_plaque	male
SRS011343	supragingival_plaque	female
SRS013723	supragingival_plaque	male
SRS015044	supragingival_plaque	female
SRS015215	supragingival_plaque	male
SRS015378	supragingival_plaque	female
SRS015574	supragingival_plaque	male
SRS015803	supragingival_plaque	female
SRS016575	supragingival_plaque	male
SRS011086	tongue_dorsum	female
SRS014124	tongue_dorsum	male
SRS014271	tongue_dorsum	female
SRS015038	tongue_dorsum	male
SRS015057	tongue_dorsum	female
SRS015537	tongue_dorsum	male
SRS015893	tongue_dorsum	female
SRS016037	tongue_dorsum	male
SRS016501	tongue_dorsum	female
SRS016740	tongue_dorsum	male
```

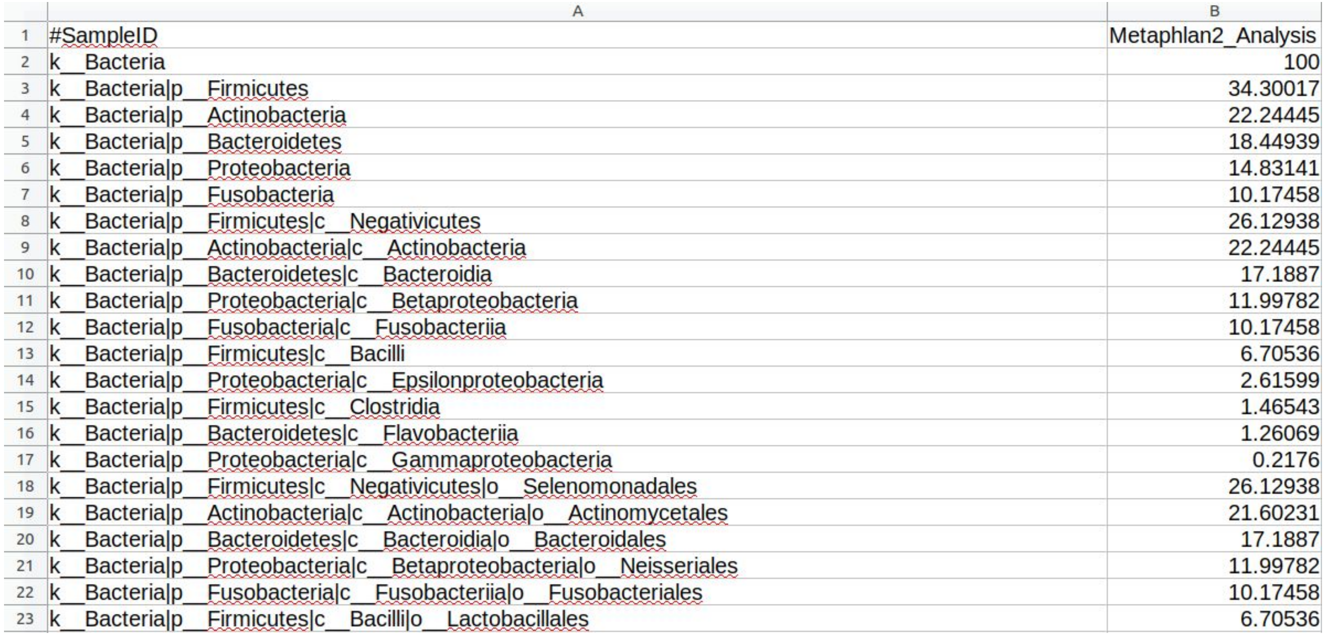
Data

1. Fastq files from 30 samples
2. Metadata

Downloaded from Human Microbiome Project

https://www.hmpdacc.org/HMIWGS/all/

Results



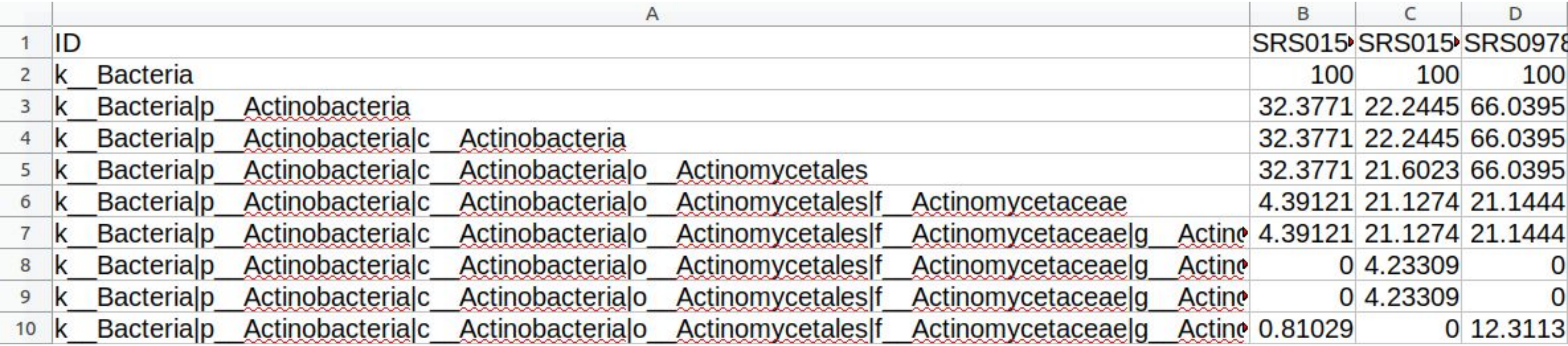
	A	B
1	#SampleID	Metaphlan2_Analysis
2	k__Bacteria	100
3	k__Bacteria|p__Firmicutes	34.30017
4	k__Bacteria|p__Actinobacteria	22.24445
5	k__Bacteria|p__Bacteroidetes	18.44939
6	k__Bacteria|p__Proteobacteria	14.83141
7	k__Bacteria|p__Fusobacteria	10.17458
8	k__Bacteria|p__Firmicutes|c__Negativicutes	26.12938
9	k__Bacteria|p__Actinobacteria|c__Actinobacteria	22.24445
10	k__Bacteria|p__Bacteroidetes|c__Bacteroidia	17.1887
11	k__Bacteria|p__Proteobacteria|c__Betaproteobacteria	11.99782
12	k__Bacteria|p__Fusobacteria|c__Fusobacteriia	10.17458
13	k__Bacteria|p__Firmicutes|c__Bacilli	6.70536
14	k__Bacteria|p__Proteobacteria|c__Epsilonproteobacteria	2.61599
15	k__Bacteria|p__Firmicutes|c__Clostridia	1.46543
16	k__Bacteria|p__Bacteroidetes|c__Flavobacteriia	1.26069
17	k__Bacteria|p__Proteobacteria|c__Gammaproteobacteria	0.2176
18	k__Bacteria|p__Firmicutes|c__Negativicutes|o__Selenomonadales	26.12938
19	k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales	21.60231
20	k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales	17.1887
21	k__Bacteria|p__Proteobacteria|c__Betaproteobacteria|o__Neisseriales	11.99782
22	k__Bacteria|p__Fusobacteria|c__Fusobacteriia|o__Fusobacteriales	10.17458
23	k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillales	6.70536

/mnt/HA/groups/rosenclassGrp/Students_SP19/metaphlan2_tutorial.tgz

metaphlan2_tutorial/hmp_metagenomics/metaphlan_precalculate/

Results

```
$ python ./utils/merge_metaphlan_tables.py [outputs] > [merged file]
```

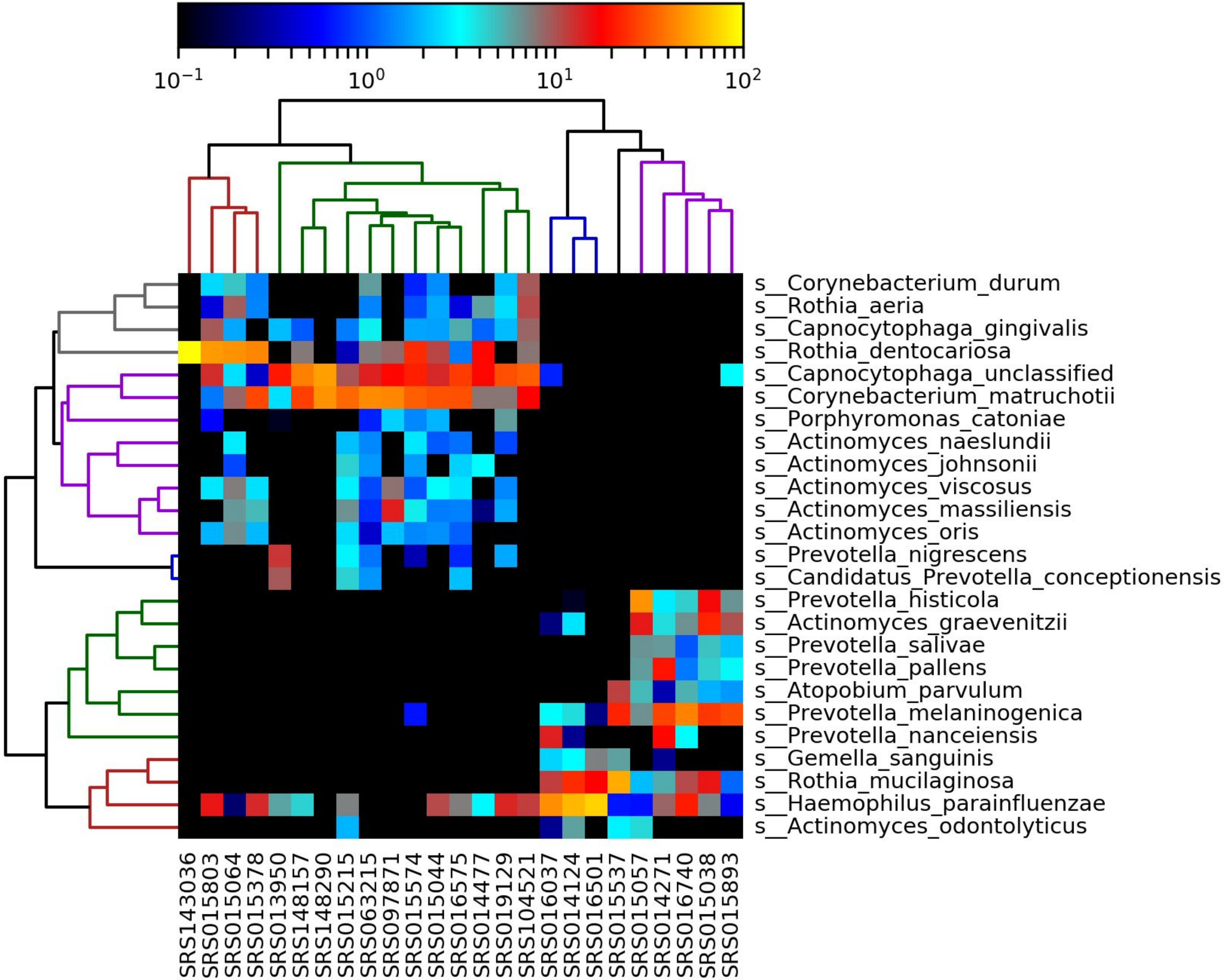


	A	B	C	D
1	ID	SRS015	SRS015	SRS0978
2	k__Bacteria	100	100	100
3	k__Bacteria|p__Actinobacteria	32.3771	22.2445	66.0395
4	k__Bacteria|p__Actinobacteria|c__Actinobacteria	32.3771	22.2445	66.0395
5	k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales	32.3771	21.6023	66.0395
6	k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae	4.39121	21.1274	21.1444
7	k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actino	4.39121	21.1274	21.1444
8	k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actino	0	4.23309	0
9	k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actino	0	4.23309	0
10	k__Bacteria|p__Actinobacteria|c__Actinobacteria|o__Actinomycetales|f__Actinomycetaceae|g__Actino	0.81029	0	12.3113

Results

Heat map

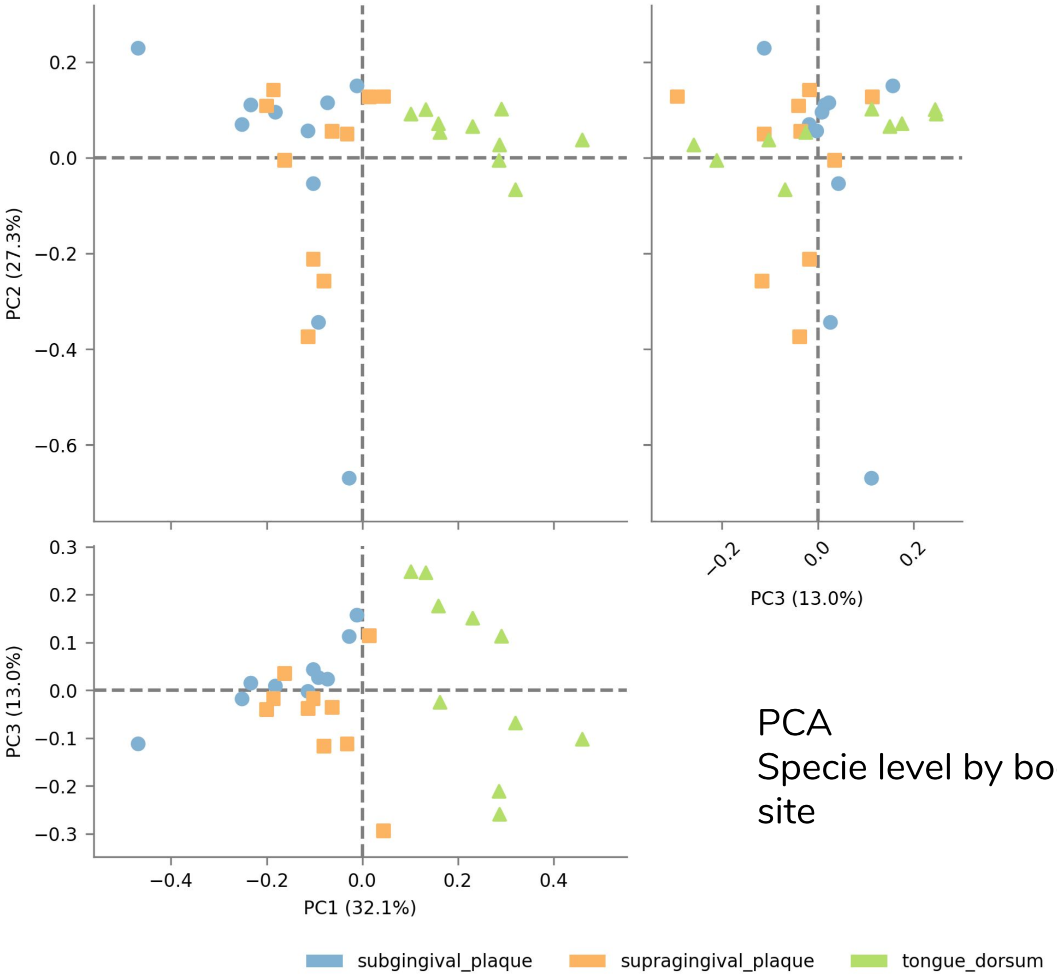
Based on the Euclidian Distance of the abundance



Results

STAMP(Statistical Analysis of Metagenomic Profiles)

Website: http://kiwi.cs.dal.ca/Software/STAMP

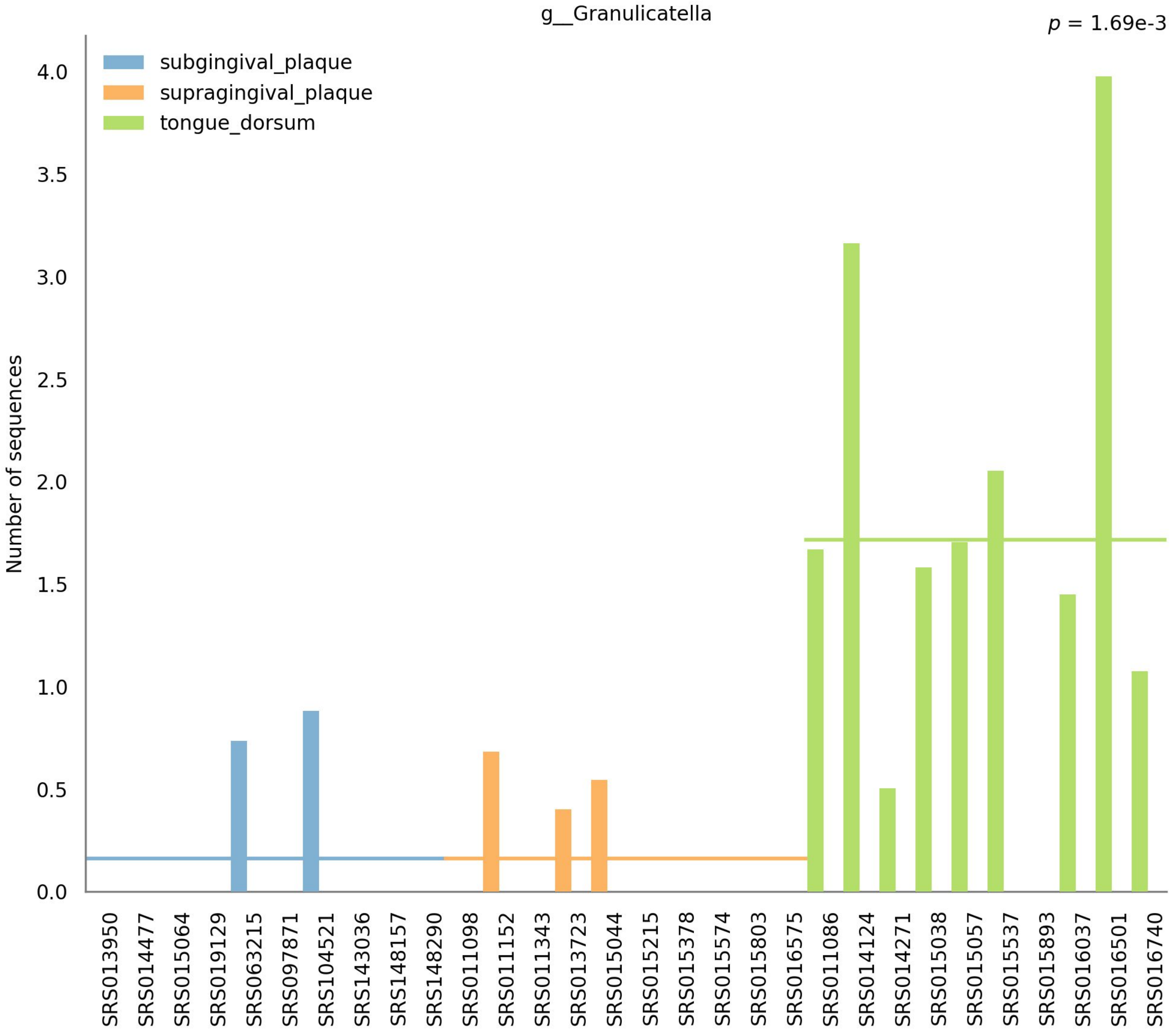


PCA
Specie level by body site

Results

STAMP(Statistical Analysis of Metagenomic Profiles)

Bar chart of the Granulicatella genus abundance



References

https://www.nature.com/articles/nmeth.2066

https://bitbucket.org/biobakery/metaphlan2/src/default/

http://kiwi.cs.dal.ca/Software/STAMP