In House pipeline: NG-Tax

Validation based on controls

Open & reproducible microbiome data analysis spring school
Wageningen, The Netherlands, May 28-30, 2018

Gerben DA Hermes, PhD

Laboratory of Microbiology,

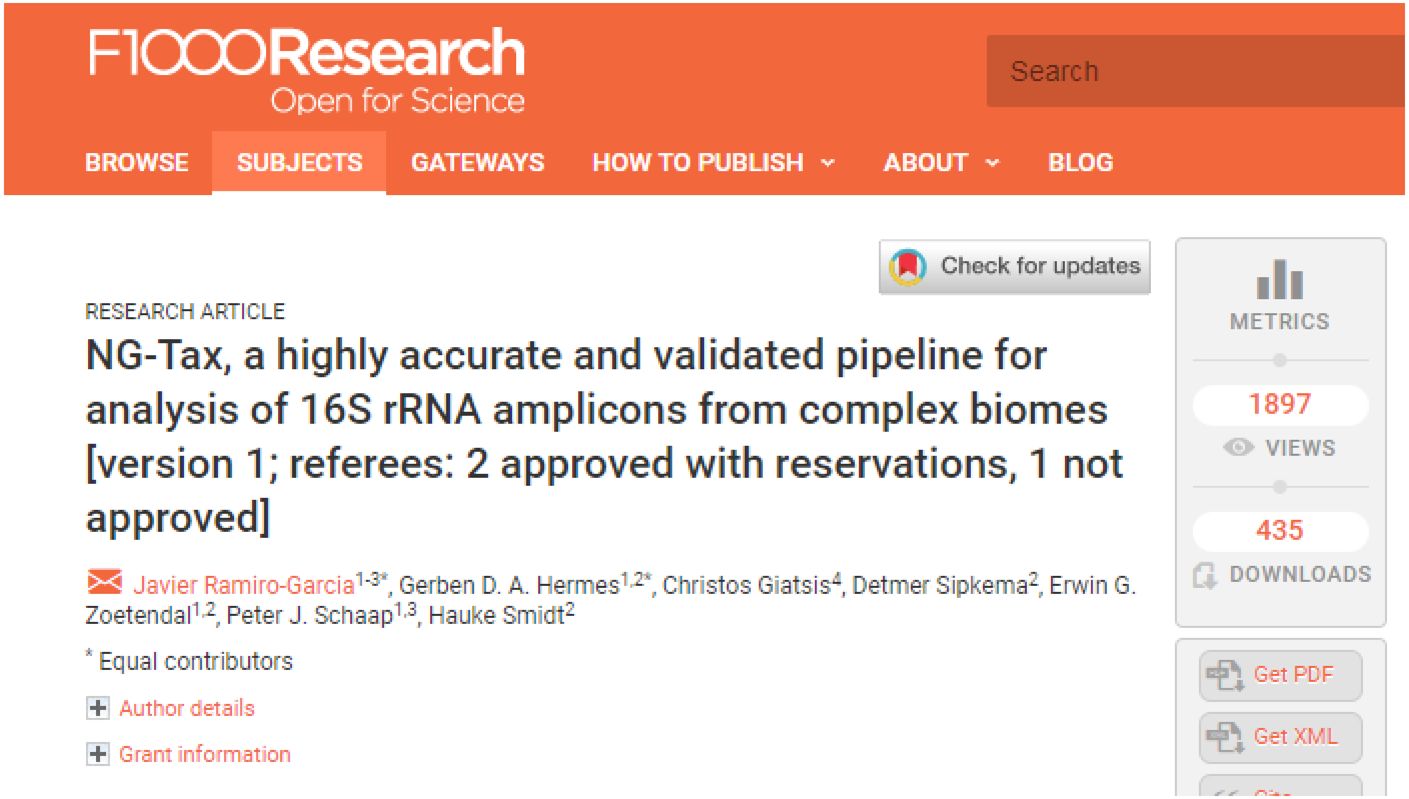
Wageningen University & Research

Tested parameters

- Primer pair
- Barcoding strategy
- Chimera filtering
- Filtering parameters
- OTU picking
- Database
- Etc…

And try to come up with an ‘optimal’ combination of each

In house analysis pipeline: NG-Tax



RESEARCH ARTICLE

NG-Tax, a highly accurate and validated pipeline for analysis of 16S rRNA amplicons from complex biomes [version 1; referees: 2 approved with reservations, 1 not approved]

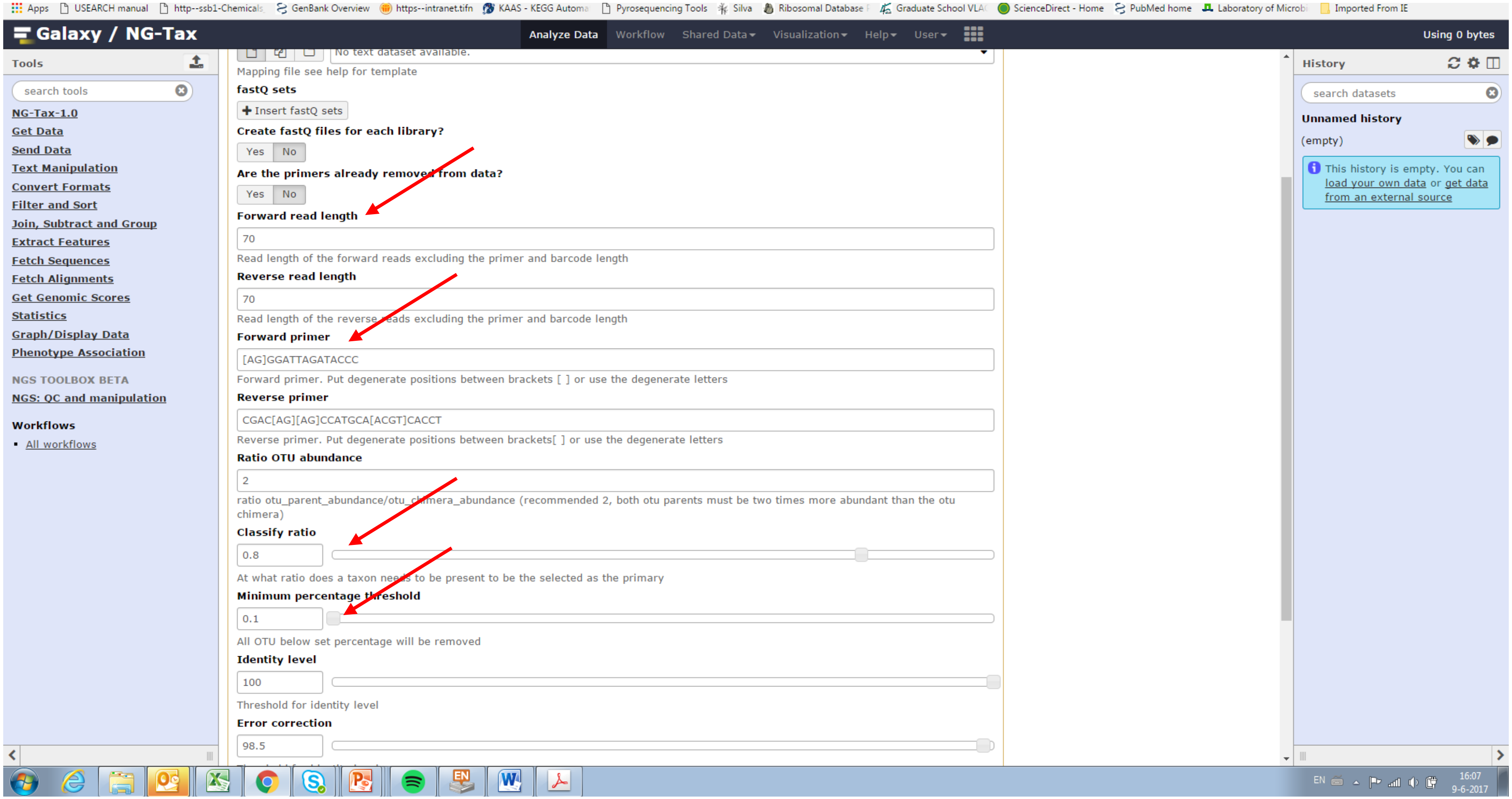
Javier Ramiro-Garcia[1-3*], Gerben D. A. Hermes[1,2*], Christos Giatsis[4], Detmer Sipkema[2], Erwin G. Zoetendal[1,2], Peter J. Schaap[1,3], Hauke Smidt[2]

[*] Equal contributors

Author details

Grant information

NG-Tax analysis & settings
Galaxy interface



No text dataset available.

Mapping file see help for template

fastQ sets

+ Insert fastQ sets

Create fastQ files for each library?

Yes [✓ No]

Are the primers already removed from data?

Yes [✓ No]

Forward read length

70

Read length of the forward reads excluding the primer and barcode length

Reverse read length

70

Read length of the reverse reads excluding the primer and barcode length

Forward primer

[AG]GGATTAGATACCC

Forward primer. Put degenerate positions between brackets [ ] or use the degenerate letters

Reverse primer

CGAC[AG][AG]CCATGCA[ACGT]CACCT

Reverse primer. Put degenerate positions between brackets[ ] or use the degenerate letters

Ratio OTU abundance

2

ratio otu_parent_abundance/otu_chimera_abundance (recommended 2, both otu parents must be two times more abundant than the otu chimera)

Classify ratio

0.8

At what ratio does a taxon needs to be present to be the selected as the primary

Minimum percentage threshold

0.1

All OTU below set percentage will be removed

Identity level

100

Threshold for identity level

Error correction

98.5

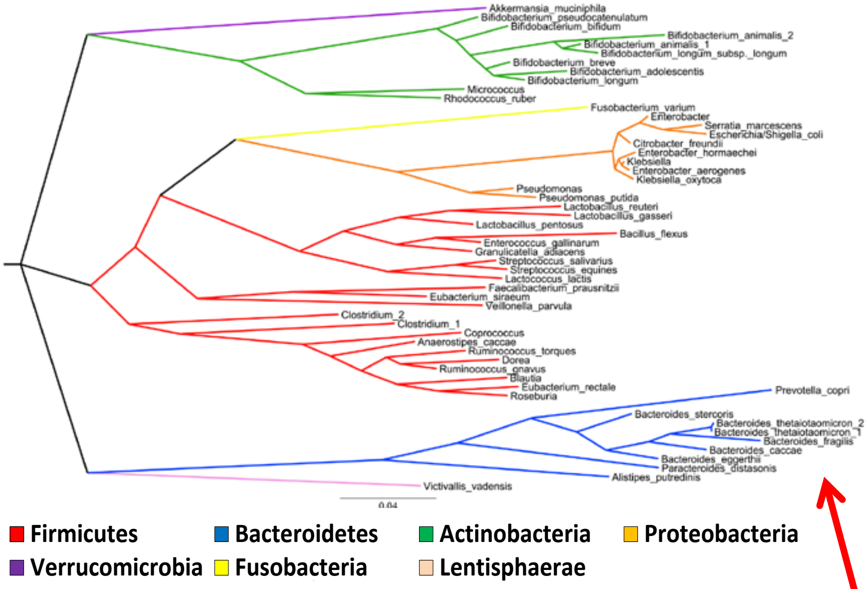
Unnamed history

(empty)

This history is empty. You can load your own data or get data from an external source

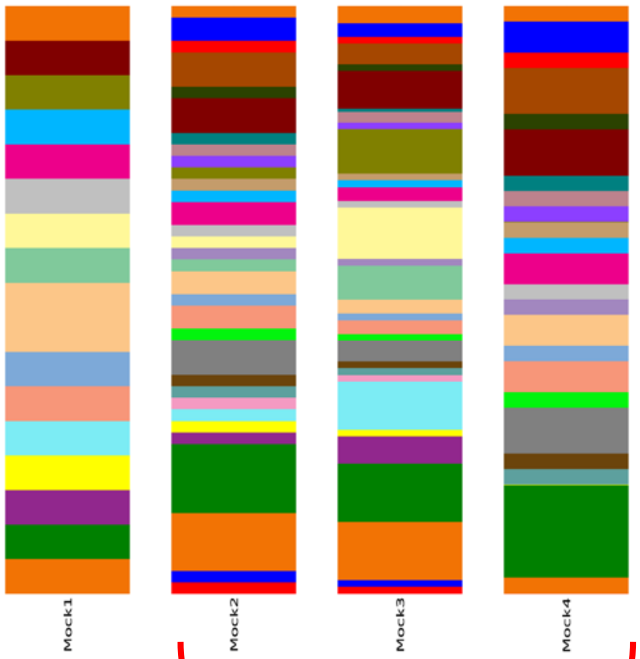
Validation:

Positive controls/Synthetic or Mock communities



Qualitative & quantitative

1 2 3 4



Resolution

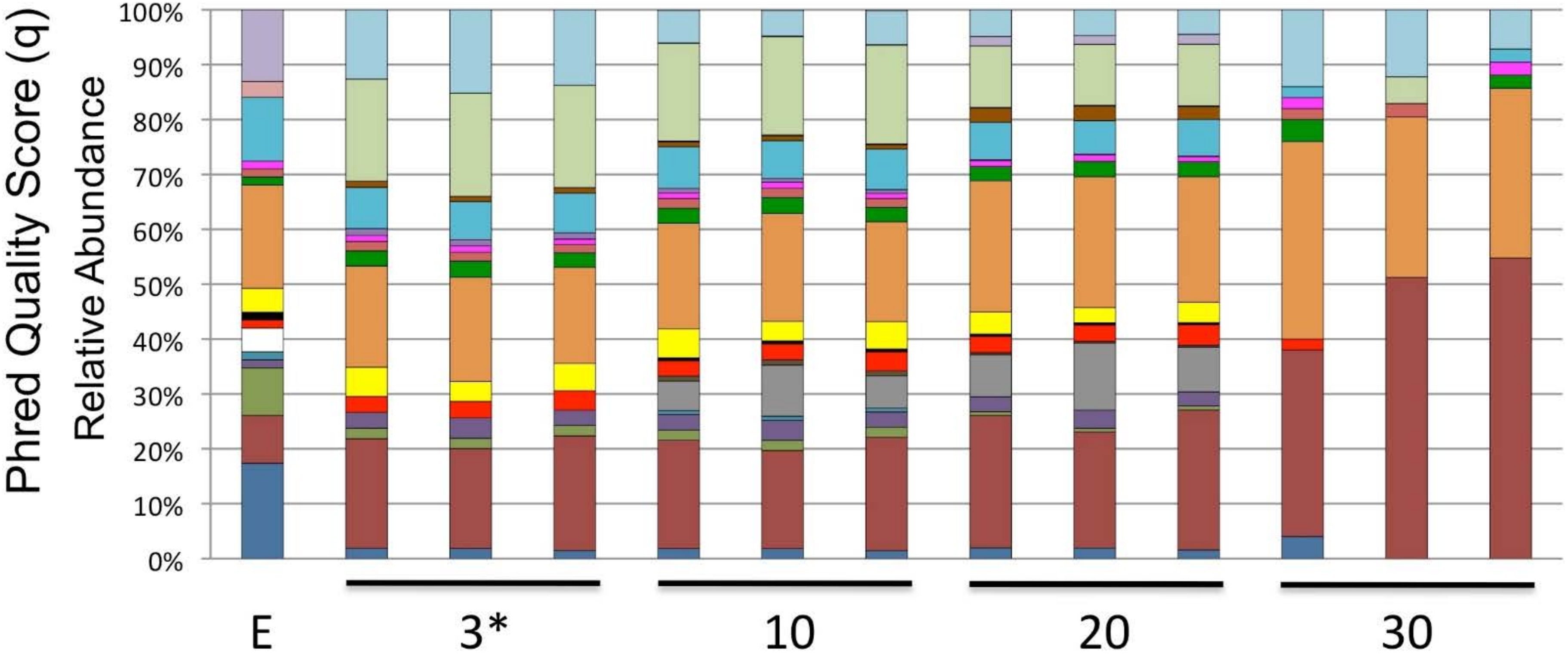
How low can we go

- 54 species
- 32 genera
- 7 phyla

WAGENINGEN
UNIVERSITY & RESEARCH

Effect of quality filtering

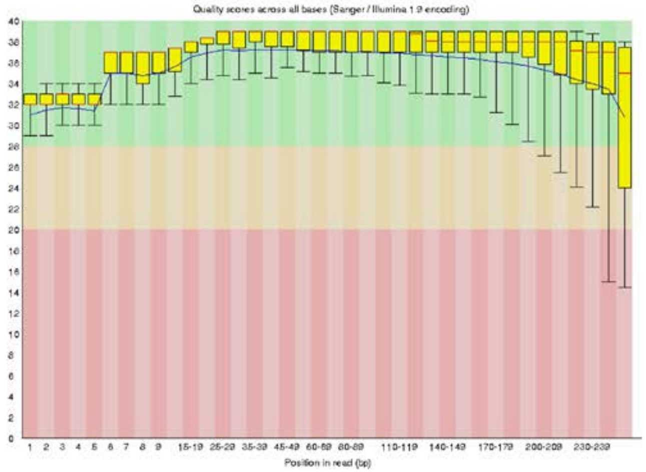
- Bokulich 2013, Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat methods. supplementals



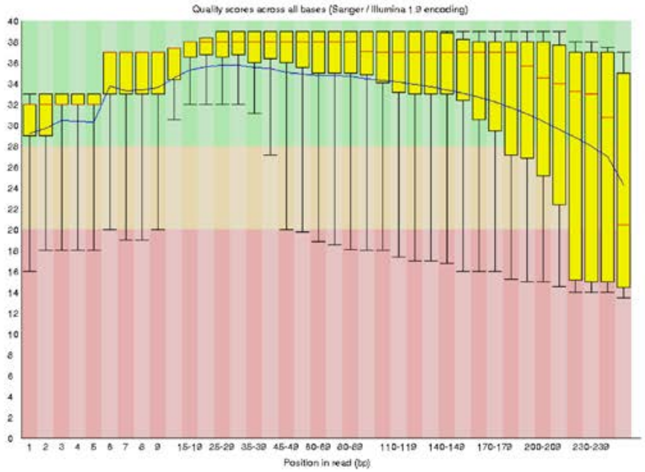
Settings optimization

- No quality filtering based on phred score

Phred 10 = 90% accuracy, 30 99.9%



Forward reads

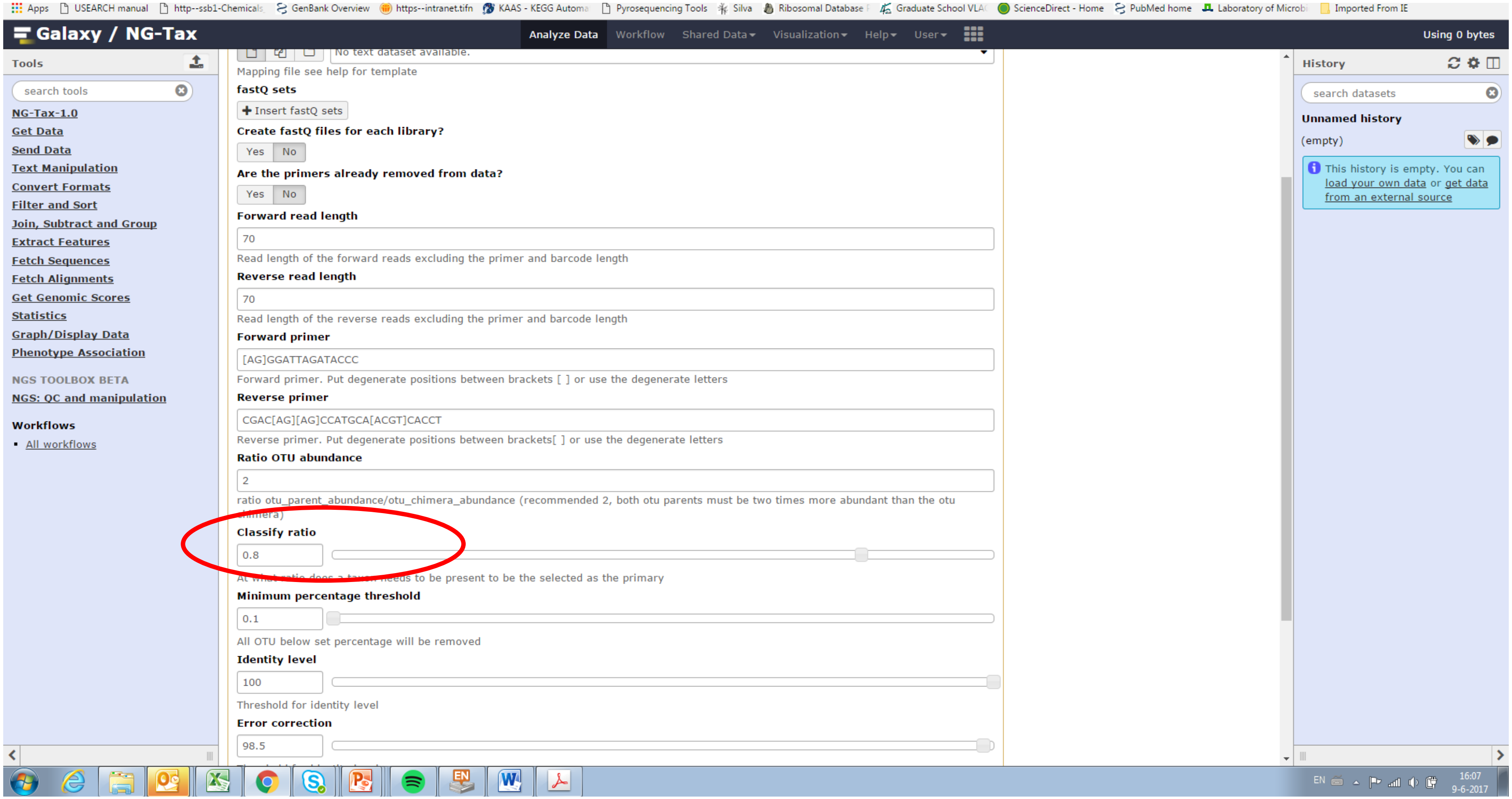


Reverse reads

- Filter by abundance -> high quality
- “QC” = demultiplexing (matching valid barcodes)
- ‘Short’ reads: 2x70nt (enough for identification) = less room for error.

WAGENINGEN
UNIVERSITY & RESEARCH

OTU classification: Classify ratio



No text dataset available.

Mapping file see help for template

fastQ sets

+ Insert fastQ sets

Create fastQ files for each library?

Yes | No

Are the primers already removed from data?

Yes | No

Forward read length

70

Read length of the forward reads excluding the primer and barcode length

Reverse read length

70

Read length of the reverse reads excluding the primer and barcode length

Forward primer

[AG]GGATTAGATACCC

Forward primer. Put degenerate positions between brackets [ ] or use the degenerate letters

Reverse primer

CGAC[AG][AG]CCATGCA[ACGT]CACCT

Reverse primer. Put degenerate positions between brackets[ ] or use the degenerate letters

Ratio OTU abundance

2

ratio otu_parent_abundance/otu_chimera_abundance (recommended 2, both otu parents must be two times more abundant than the otu chimera)

Classify ratio

0.8

At what ratio does a taxon needs to be present to be the selected as the primary

Minimum percentage threshold

0.1

All OTU below set percentage will be removed

Identity level

100

Threshold for identity level

Error correction

98.5

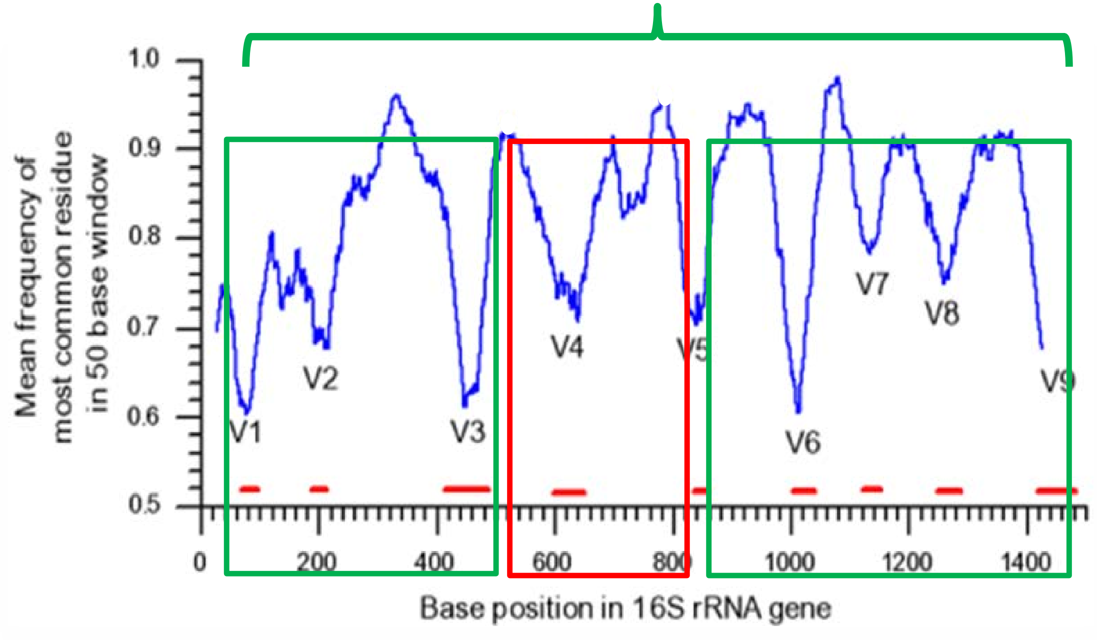
This history is empty. You can load your own data or get data from an external source

WAGENINGEN
UNIVERSITY & RESEARCH

95% sequence similarity: genus level
Based on full length 1500nt SILVA alignment

95% 100% 99%

95% sequence similarity: genus level

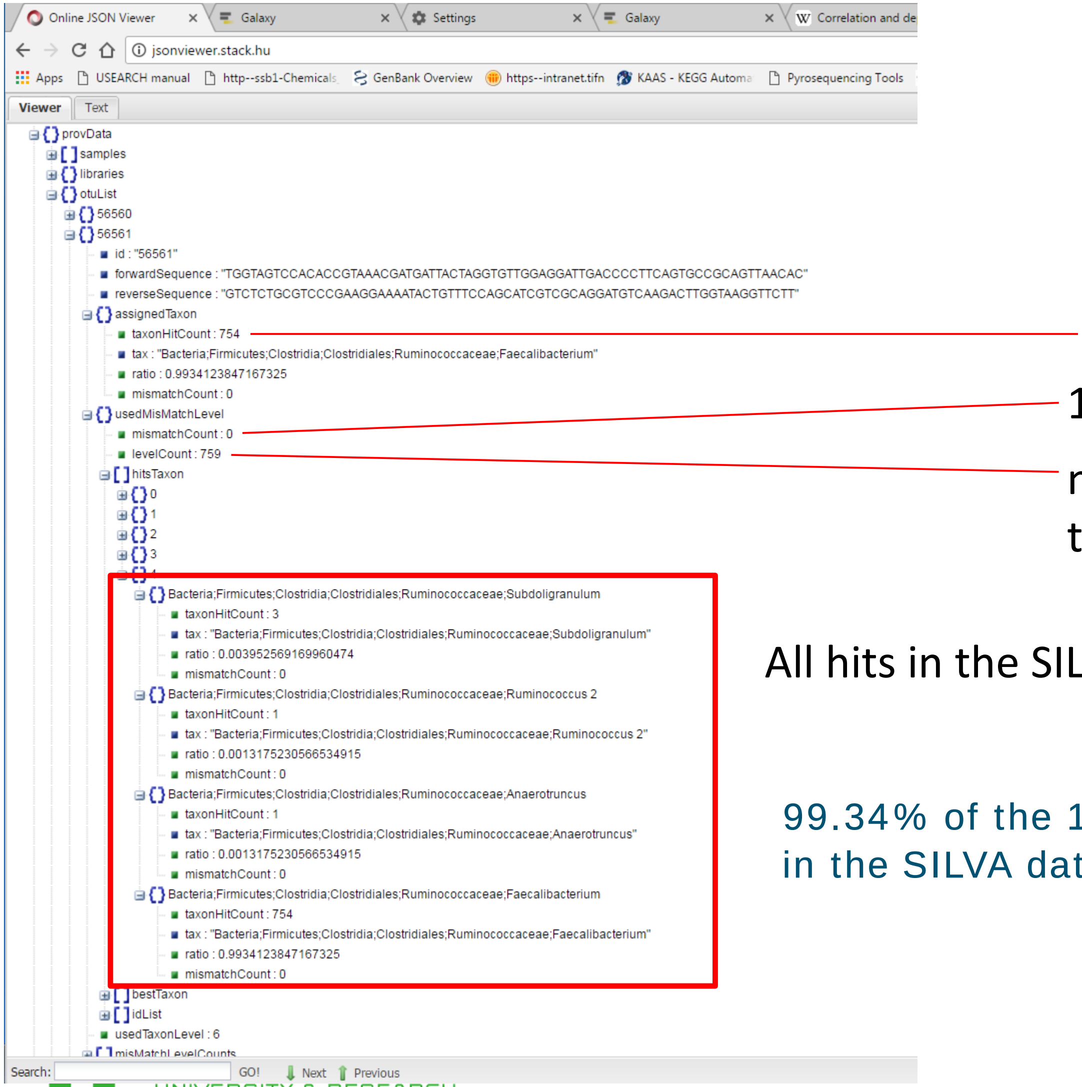


SILVA: Blautia/Faecalibacterium

Same sequence/different classification

'choose'?

Real example: very clear



```
{ } provData
  [ ] samples
  { } libraries
  { } otuList
    { } 56560
    { } 56561
      id : "56561"
      forwardSequence : "TGGTAGTCCACACCGTAAACGATGATTACTAGGTGTTGGAGGATTGACCCCTTCAGTGCCGCAGTTAACAC"
      reverseSequence : "GTCTCTGCGTCCCGAAGGAAAATACTGTTTCCAGCATCGTCGCAGGATGTCAAGACTTGGTAAGGTTCTT"
      { } assignedTaxon
        taxonHitCount : 754
        tax : "Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Faecalibacterium"
        ratio : 0.9934123847167325
        mismatchCount : 0
      { } usedMisMatchLevel
        mismatchCount : 0
        levelCount : 759
        [ ] hitsTaxon
          { } 0
          { } 1
          { } 2
          { } 3
          { } 4
            { } Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Subdoligranulum
              taxonHitCount : 3
              tax : "Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Subdoligranulum"
              ratio : 0.003952569169960474
              mismatchCount : 0
            { } Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcus 2
              taxonHitCount : 1
              tax : "Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcus 2"
              ratio : 0.0013175230566534915
              mismatchCount : 0
            { } Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Anaerotruncus
              taxonHitCount : 1
              tax : "Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Anaerotruncus"
              ratio : 0.0013175230566534915
              mismatchCount : 0
            { } Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Faecalibacterium
              taxonHitCount : 754
              tax : "Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Faecalibacterium"
              ratio : 0.9934123847167325
              mismatchCount : 0
        [ ] bestTaxon
        [ ] idList
      usedTaxonLevel : 6
      [ ] misMatchLevelCounts
```

Nr hits that are *Feacalibacterium*

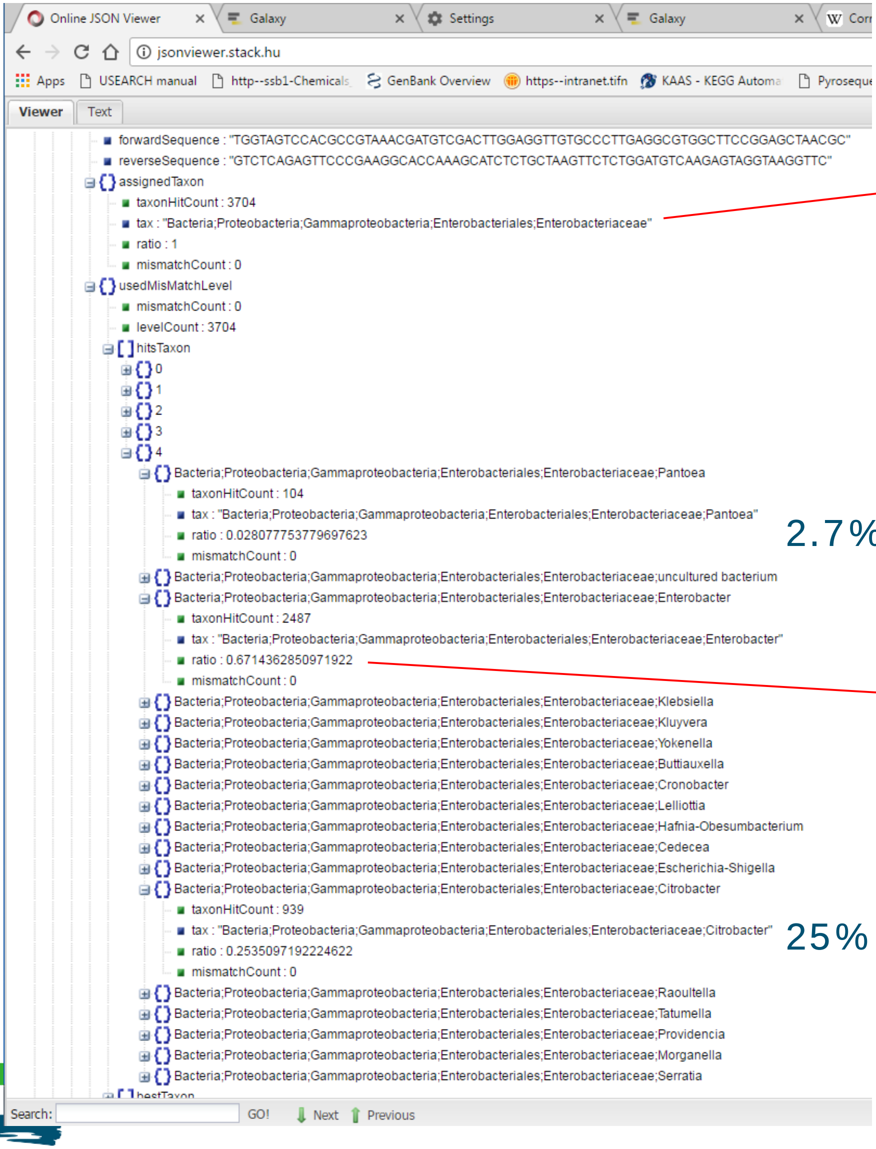
100% match in the SILVA DB

nr hits with 100% match in the SILVA DB

All hits in the SILVA DB with 100%

99.34% of the 100% matches of that sequence (OTU) in the SILVA database are *Feacalibacterium*

Real example: less clear



```
forwardSequence : "TGGTAGTCCACGCCGTAAACGATGTCGACTTGGAGGTTGTGCCCTTGAGGCGTGGCTTCCGGAGCTAACGC"
reverseSequence : "GTCTCAGAGTTCCCGAAGGCACCAAAGCATCTCTGCTAAGTTCTCTGGATGTCAAGAGTAGGTAAGGTTC"
{ } assignedTaxon
    taxonHitCount : 3704
    tax : "Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae"
    ratio : 1
    mismatchCount : 0
{ } usedMisMatchLevel
    mismatchCount : 0
    levelCount : 3704
  [ ] hitsTaxon
      { } 0
      { } 1
      { } 2
      { } 3
      { } 4
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Pantoea
            taxonHitCount : 104
            tax : "Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Pantoea"
            ratio : 0.028077753779697623
            mismatchCount : 0
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;uncultured bacterium
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Enterobacter
            taxonHitCount : 2487
            tax : "Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Enterobacter"
            ratio : 0.6714362850971922
            mismatchCount : 0
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Klebsiella
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Kluyvera
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Yokenella
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Buttiauxella
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Cronobacter
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Lelliottia
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Hafnia-Obesumbacterium
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Cedecea
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Escherichia-Shigella
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Citrobacter
            taxonHitCount : 939
            tax : "Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Citrobacter"
            ratio : 0.2535097192224622
            mismatchCount : 0
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Raoultella
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Tatumella
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Providencia
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Morganella
        { } Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Enterobacteriaceae;Serratia
  [ ] bestTaxon
```

Assigned:
Enterobacteriacaea

2.7% Pantoea

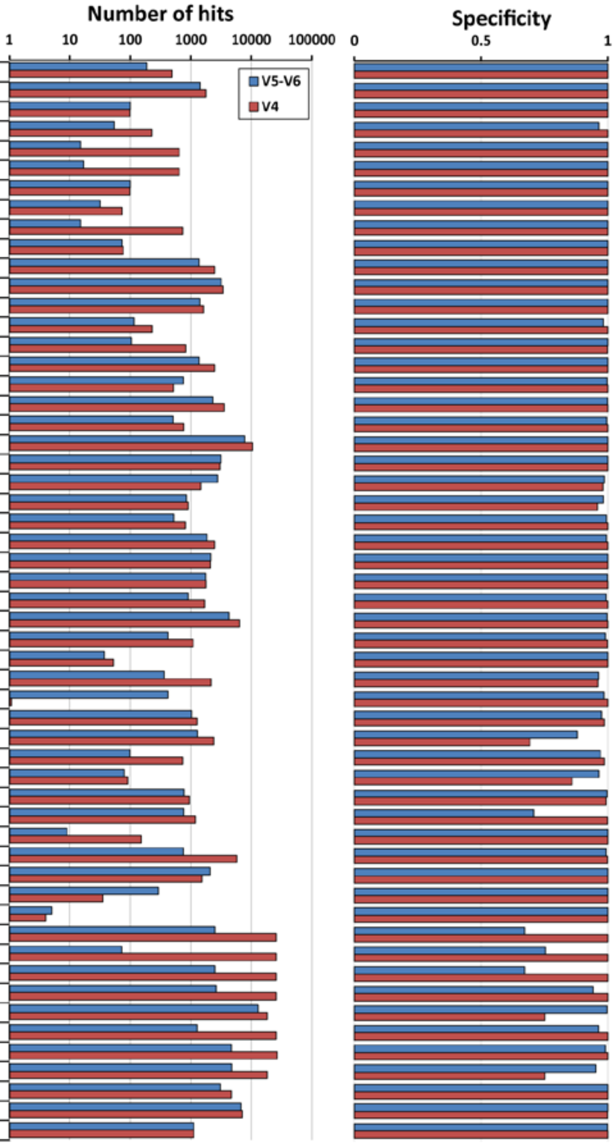
67% @100% match with 3704 hits -> back to family

25% Citrobacter

NG-Tax classification accuracy & confidence

2x70nt information

	SINA full length	NG-tax trimmed to V5-V6	NG-tax trimmed to V4
1	Rhodococcus	Rhodococcus	Rhodococcus
2	Micrococcus	Micrococcus	Micrococcus
3	Bifidobacterium	Bifidobacterium	Bifidobacterium
4	Bifidobacterium	Bifidobacterium	Bifidobacterium
5	Bifidobacterium	Bifidobacterium	Bifidobacterium
6	Bifidobacterium	Bifidobacterium	Bifidobacterium
7	Bifidobacterium	Bifidobacterium	Bifidobacterium
8	Bifidobacterium	Bifidobacterium	Bifidobacterium
9	Bifidobacterium	Bifidobacterium	Bifidobacterium
10	Bifidobacterium	Bifidobacterium	Bifidobacterium
11	Bacteroides	Bacteroides	Bacteroides
12	Bacteroides	Bacteroides	Bacteroides
13	Bacteroides	Bacteroides	Bacteroides
14	Bacteroides	Bacteroides	Bacteroides
15	Bacteroides	Bacteroides	Bacteroides
16	Bacteroides	Bacteroides	Bacteroides
17	Parabacteroides	Parabacteroides	Parabacteroides
18	Prevotella 9	Prevotella 9	Prevotella 9
19	Alistipes	Alistipes	Alistipes
20	Bacillus	Bacillus	Bacillus
21	Bacillus	Bacillus	Bacillus
22	Granulicatella	Granulicatella	Granulicatella
23	Enterococcus	Enterococcus	Enterococcus
24	Lactobacillus	Lactobacillus	Lactobacillus
25	Lactobacillus	Lactobacillus	Lactobacillus
26	Lactobacillus	Lactobacillus	Lactobacillus
27	Lactococcus	Lactococcus	Lactococcus
28	Streptococcus	Streptococcus	Streptococcus
29	Streptococcus	Streptococcus	Streptococcus
30	Clostridium sensu stricto 1	Clostridium sensu stricto 1	Clostridium sensu stricto 1
31	Anaerostipes	Anaerostipes	Anaerostipes
32	Blautia	Blautia	Blautia
33	Dorea	Dorea	Dorea
34	[Ruminococcus] gnavus group	[Ruminococcus] gnavus group	[Ruminococcus] gnavus group
35	[Eubacterium] rectale group	[Eubacterium] rectale group	[Eubacterium] rectale group
36	[Eubacterium] hallii group	[Eubacterium] hallii group	[Eubacterium] hallii group
37	Roseburia	Roseburia	Roseburia
38	[Ruminococcus] torques group	[Ruminococcus] torques group	[Ruminococcus] torques group
39	Peptostreptococcaceae*	Intestinibacter	Peptostreptococcaceae*
40	Ruminiclostridium 6	Ruminiclostridium 6	Ruminiclostridium 6
41	Faecalibacterium	Faecalibacterium	Faecalibacterium
42	Veillonella	Veillonella	Veillonella
43	Fusobacterium	Fusobacterium	Fusobacterium
44	Victivallis	Victivallis	Victivallis
45	Citrobacter	Enterobacter	Enterobacteriaceae*
46	Enterobacteriaceae*	Klebsiella	Enterobacteriaceae*
47	Enterobacter	Enterobacter	Enterobacteriaceae*
48	Enterobacter	Enterobacter	Enterobacteriaceae*
49	Escherichia-Shigella	Escherichia-Shigella	Enterobacteriaceae*
50	Klebsiella	Klebsiella	Enterobacteriaceae*
51	Enterobacteriaceae*	Escherichia-Shigella	Enterobacteriaceae*
52	Serratia	Serratia	Escherichia-Shigella
53	Pseudomonas	Pseudomonas	Pseudomonas
54	Pseudomonas	Pseudomonas	Pseudomonas
55	Akkermansia	Akkermansia	Akkermansia



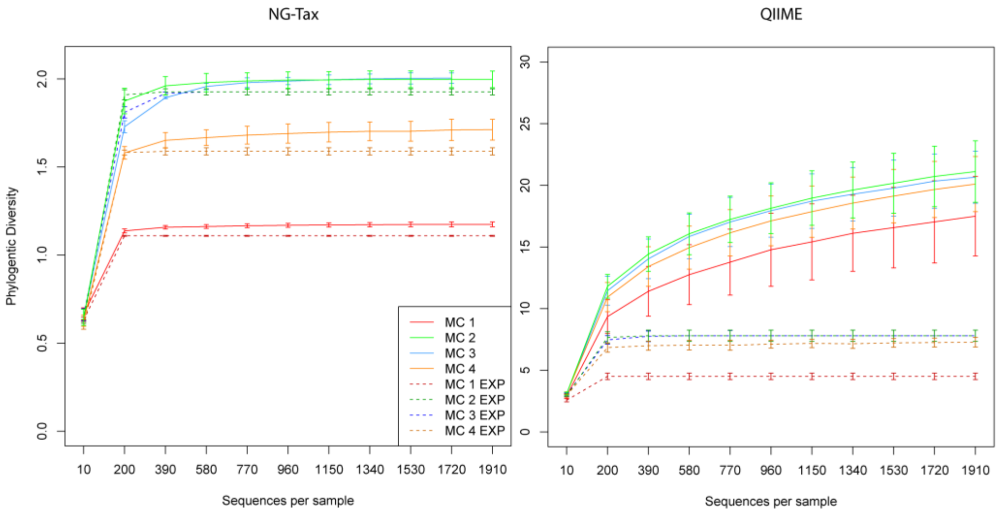
Validation

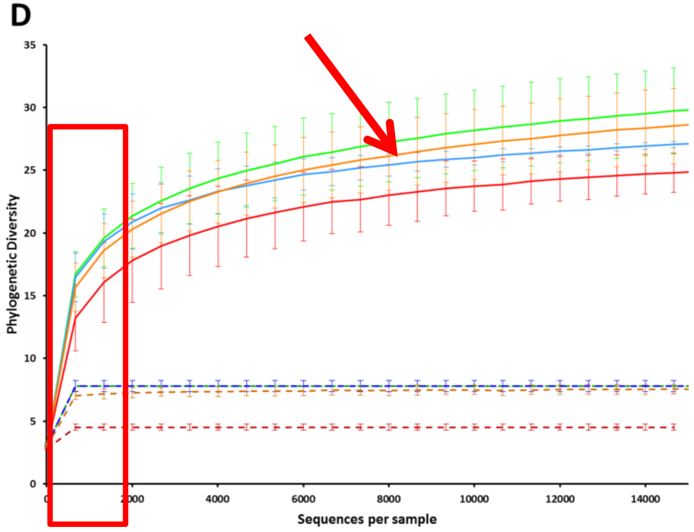
Sources of error/noise:

- 4 different synthetic community types (each in triplicate)
- 2 different variable regions V4, V5-V6
- 2 primer pairs (515F-806R, 784F-1064R)
- 25-35 PCR cycles
- 7 libraries
- 3 different sequencing runs

NG-Tax performance: alpha diversity

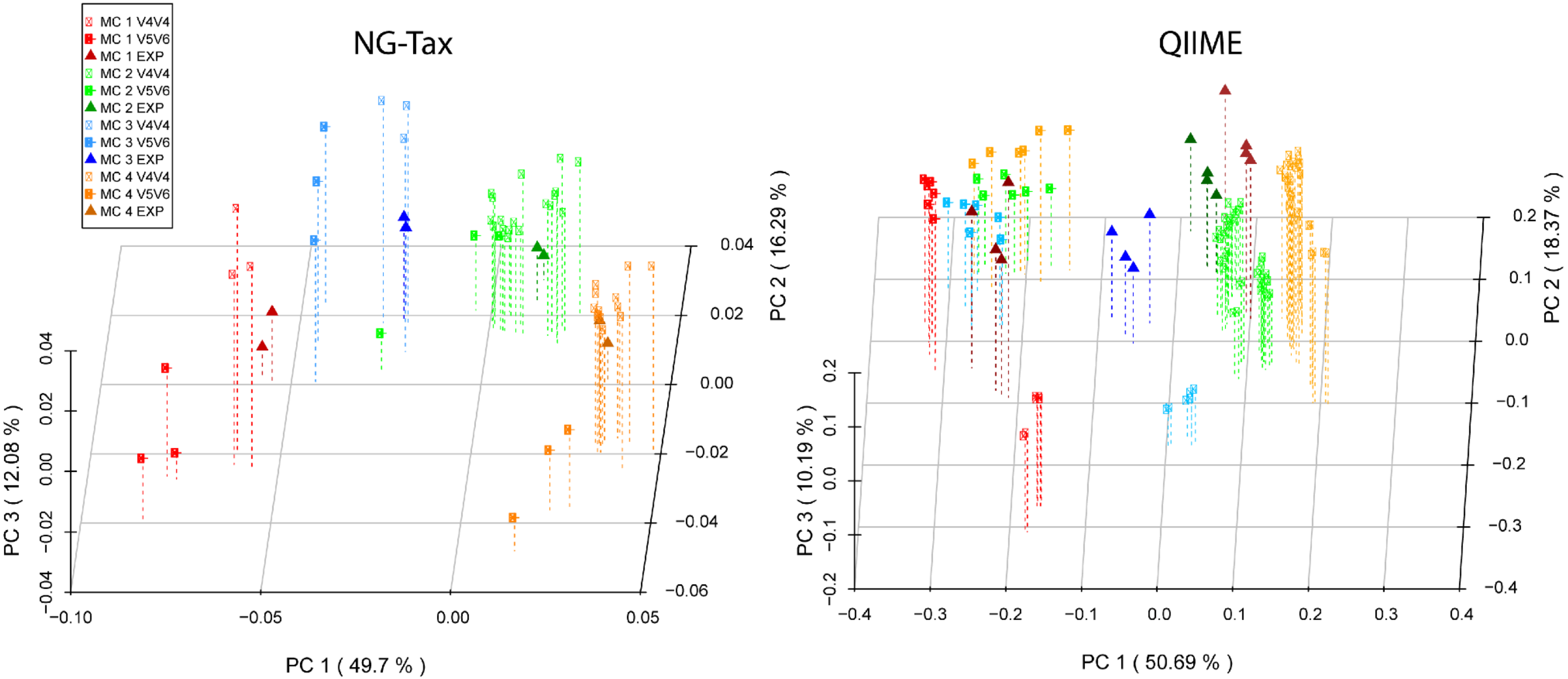
Rarefaction plots





NG-Tax performance: beta diversity

Weighted Unifrac

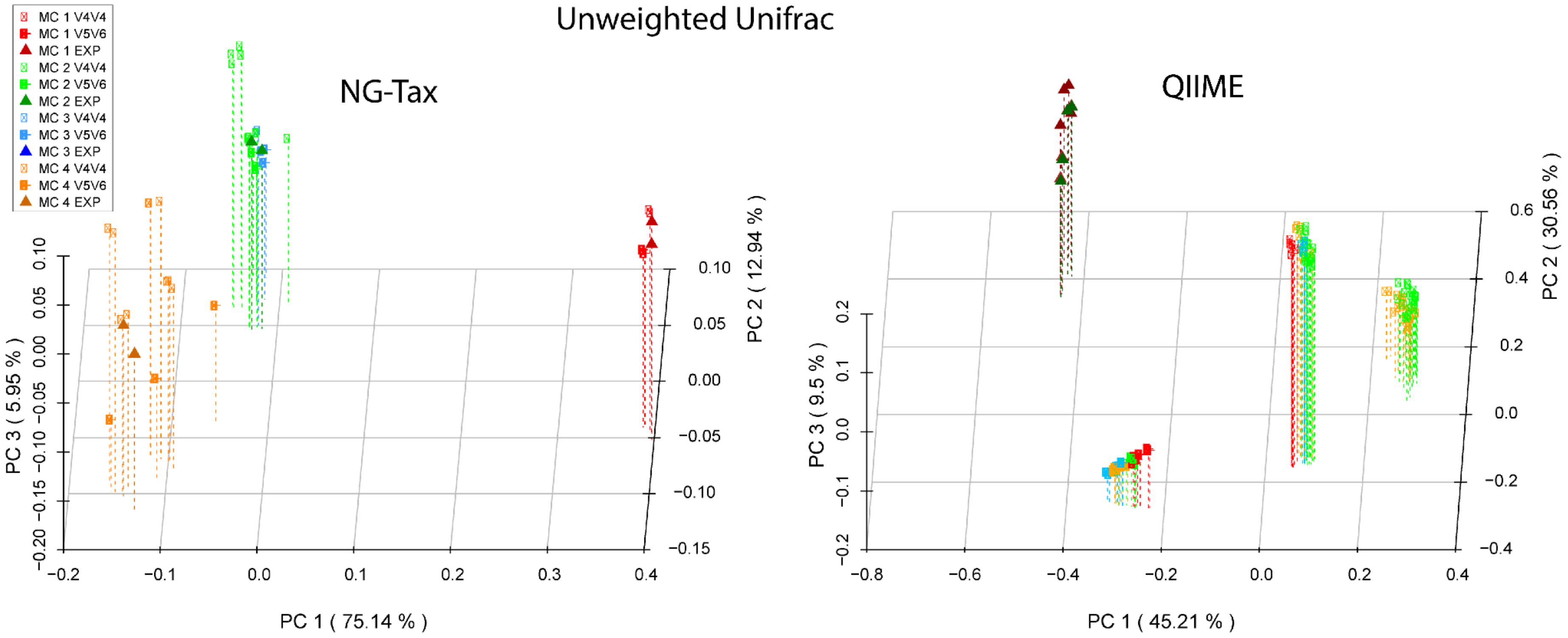


NG-Tax

QIIME

WAGENINGEN
UNIVERSITY & RESEARCH

NG-Tax performance: beta diversity



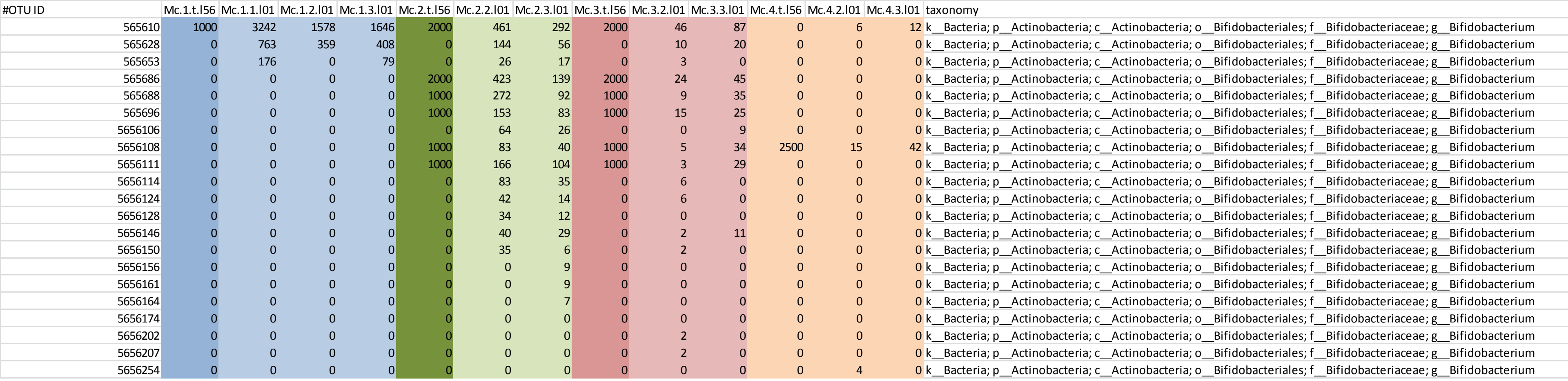
WAGENINGEN UNIVERSITY & RESEARCH

Retain or retrieve the relevant **high resolution biological information** despite high noise levels

Independent of

- primer type/variable region
- Sample diversity/complexity
- Sequencing run

Biom (OTU table)



#OTU ID	Mc.1.t.l56	Mc.1.1.l01	Mc.1.2.l01	Mc.1.3.l01	Mc.2.t.l56	Mc.2.2.l01	Mc.2.3.l01	Mc.3.t.l56	Mc.3.2.l01	Mc.3.3.l01	Mc.4.t.l56	Mc.4.2.l01	Mc.4.3.l01	taxonomy
565610	1000	3242	1578	1646	2000	461	292	2000	46	87	0	6	12	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
565628	0	763	359	408	0	144	56	0	10	20	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
565653	0	176	0	79	0	26	17	0	3	0	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
565686	0	0	0	0	2000	423	139	2000	24	45	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
565688	0	0	0	0	1000	272	92	1000	9	35	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
565696	0	0	0	0	1000	153	83	1000	15	25	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656106	0	0	0	0	0	64	26	0	0	9	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656108	0	0	0	0	1000	83	40	1000	5	34	2500	15	42	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656111	0	0	0	0	1000	166	104	1000	3	29	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656114	0	0	0	0	0	83	35	0	6	0	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656124	0	0	0	0	0	42	14	0	6	0	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656128	0	0	0	0	0	34	12	0	0	0	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656146	0	0	0	0	0	40	29	0	2	11	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656150	0	0	0	0	0	35	6	0	2	0	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656156	0	0	0	0	0	0	9	0	0	0	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656161	0	0	0	0	0	0	9	0	0	0	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656164	0	0	0	0	0	0	7	0	0	0	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656174	0	0	0	0	0	0	0	0	0	0	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656202	0	0	0	0	0	0	0	0	2	0	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656207	0	0	0	0	0	0	0	0	2	0	0	0	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium
5656254	0	0	0	0	0	0	0	0	0	0	0	4	0	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium

Data interpretation: NG-Tax

- Reads/OTU -> confidence
- ~3 reads not very confident. 30.000 reads very confident
- Classification ratio 0.8, maybe check
- Preferably at genus level, OTUs contain too much risk/error, unless the OTU is very ‘clean’
- Preferably use phylogenetic information because OTU based methods are very sensitive to noise
- Check classification of interesting OTUs
- Classify Nas using whatever method (blast is ugly)