

Metagenomika – Vyhodnocení dat (16S rRNA)

Petra Vídeňská Ph.D.

NGS formáty

- Samotná sekvence + informace o kvalitě jednotlivých nukleotidů
- 454 → .sff – lze rozdělit na dva podsoubory .fasta a .qual (kvalita)
- Illumina, IonTorrent - FastQ

FastQ formát

```
@HWUSI-EAS582_157:6:1:1:1501/1
NCACAGACACACACGAACACACAAAGACATGCCCATATGAAGAT
+
%.7786867:778556858746575058873/347777476035
@HWUSI-EAS582_157:6:1:1:1606/1
NCTGGCACCTTGATTTTGGACTTCCCAGCCTCCAGAACTGTGAG
+
%1948988888798988366898888648998788898888588
```

- Line 1 : begins with @ plus sequence identifier
- Line 2 : sequence("read")
- Line 3 : begins +
- Line 4 : quality values

FastQ formát

@HWUSI-EAS100R:1:2:99:88#0/1

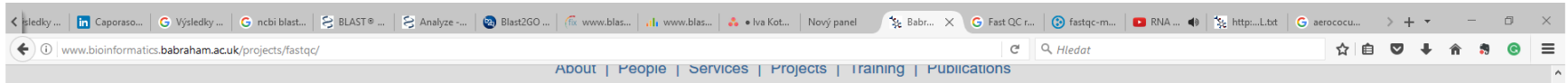
- 1: flowcell number
- 2: tile number
- 99: "x" - coordinate
- 88: "y" - coordinate
- #0: number of index

FastQ formát – quality score

Q	P	Base was call right	
10	0.1	90%	(1 from 10)
20	0.01	99%	(1 from 100)
30	0.001	99.9%	(1 from 1000)
40	0.0001	99.99%	(1 from 10 000)

- $Q = -10 \log_{10} P$
P...probability base call is incorrect

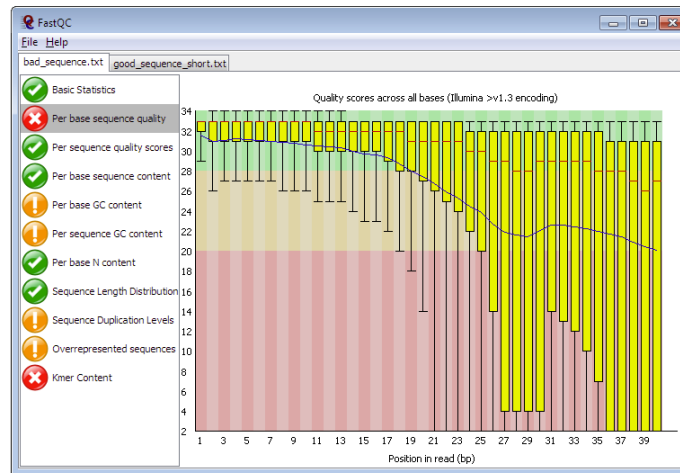
Fast QC



FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later .
Initial Contact	Simon Andrews

[Download Now](#)



[View our tutorial video](#)

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

Blast2go

Blast2GO Basic

File Analysis Tools View Help

start genefind blast Interpro mapping annot charts graphs select diff-expr

Table: 10_S1012 587 of 587

Nr	Tags	SeqName	Description	Length	#Hits	e-Value	sim mean	#GO	GO IDs	GO Names	Enzyme Codes	Enzyme Na...	InterPro IDs	InterPro GO IDs	InterPro GO Names
1	BLASTED	M04232_81_000000000-AY8BF_1_1101_4738_6391_1_N_0_10	Uncultured Kle...	301	2	1.2E-100	92%								
2	BLASTED	M04232_81_000000000-AY8BF_1_1101_10866_19466_1_N_0_10	Uncultured bac...	301	2	5.67E-41	86%								
3	BLASTED	M04232_81_000000000-AY8BF_1_1102_5223_5896_1_N_0_10	Uncultured bac...	301	2	8.91E-114	93%								
4	BLASTED	M04232_81_000000000-AY8BF_1_1102_2195_10821_1_N_0_10	Uncultured Kle...	301	2	6.63E-127	96%								
5	BLASTED	M04232_81_000000000-AY8BF_1_1102_23492_13143_1_N_0_10	Uncultured Kle...	301	2	3.94E-122	95%								
6	BLASTED	M04232_81_000000000-AY8BF_1_1102_29280_15457_1_N_0_10	Klebsiella pneu...	301	2	1.35E-146	98%								
7		M04232_81_000000000-AY8BF_1_1102_28387_16310_1_N_0_10		301											
8		M04232_81_000000000-AY8BF_1_1102_5274_16337_1_N_0_10		301											
9		M04232_81_000000000-AY8BF_1_1102_7788_18401_1_N_0_10		301											
10	BLASTED	M04232_81_000000000-AY8BF_1_1102_17817_19181_1_N_0_10	Klebsiella pneu...	301	2	2.82E-132	95.5%								
11	BLASTED	M04232_81_000000000-AY8BF_1_1102_21078_19529_1_N_0_10	Uncultured Kle...	301	2	1.81E-133	96%								
12	BLASTED	M04232_81_000000000-AY8BF_1_1102_22031_19912_1_N_0_10	Klebsiella pneu...	301	2	1.68E-127	96%								
13	BLASTED	M04232_81_000000000-AY8BF_1_1102_5732_21131_1_N_0_10	Uncultured Kle...	301	2	5.52E-149	98%								
14	BLASTED	M04232_81_000000000-AY8BF_1_1102_12950_22634_1_N_0_10	Uncultured bac...	301	2	1.44E-115	93.5%								
15	BLASTED	M04232_81_000000000-AY8BF_1_1102_14082_22825_1_N_0_10	Uncultured Kle...	301	2	1.68E-127	95%								
16		M04232_81_000000000-AY8BF_1_1102_14693_25114_1_N_0_10		301											
17		M04232_81_000000000-AY8BF_1_1103_8875_3598_1_N_0_10		301											
18		M04232_81_000000000-AY8BF_1_1103_8595_6418_1_N_0_10		301											
19		M04232_81_000000000-AY8BF_1_1103_4193_8478_1_N_0_10		301											

Progress File Manager Application Messages QBLAST

Welcome Message Blast Result of M04232_81_00000... Blast Result of M04232_81_00000... Blast Result of M04232_81_00000... Blast Result of M04232_81_00000... Blast Result of M04232_81_00000...

Query Name (Length): M04232_81_000000000-AY8BF_1_1102_12950_22634_1_N_0_10 Blast Version: BLASTN 2.6.0+ Database: nt

E-Value Cutoff: 0.001 Filters: L; Blast Program: blast

Annotation: - Enzyme: - References: Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Sequences Producing Significant Alignments	Scientific Taxonomy Name	E-Value	Hit length	Align length	Pos	Sim	Hsp/Hit	Hsp/Query	Hsps	Frame	Gene Identifier	Acc	Gene Name (Taxa ID)	Xref (DB)	Mapping
gi 322083597 gb JF126227.1 Uncultured bacterium clone ncd1465e03c1 16S ribosomal RNA gene, partial sequence	uncultured bacterium	1.44E-115	1361	278	0	94.2%	20.4%	92.4%	1	0	322083658	JF126227.1		JF126245.1	
gi 322083658 gb JF126245.1 Uncultured bacterium clone ncd1465e06c1 16S ribosomal RNA gene, partial sequence															
gi 1120768694 gb CP018671.1 Klebsiella pneumoniae strain CAV1042, complete genome	Klebsiella pneumoniae	3.52E-113	5424949	278	0	93.9%	0.0%	92.4%	8	0	1120768694	CP018671.1			

Europe, Germany; DE3 Version: b2g_jan17

Analýza dat (Qiime)

- Rozřazení vzorků podle značek, spojení sekvencí z Illuminy

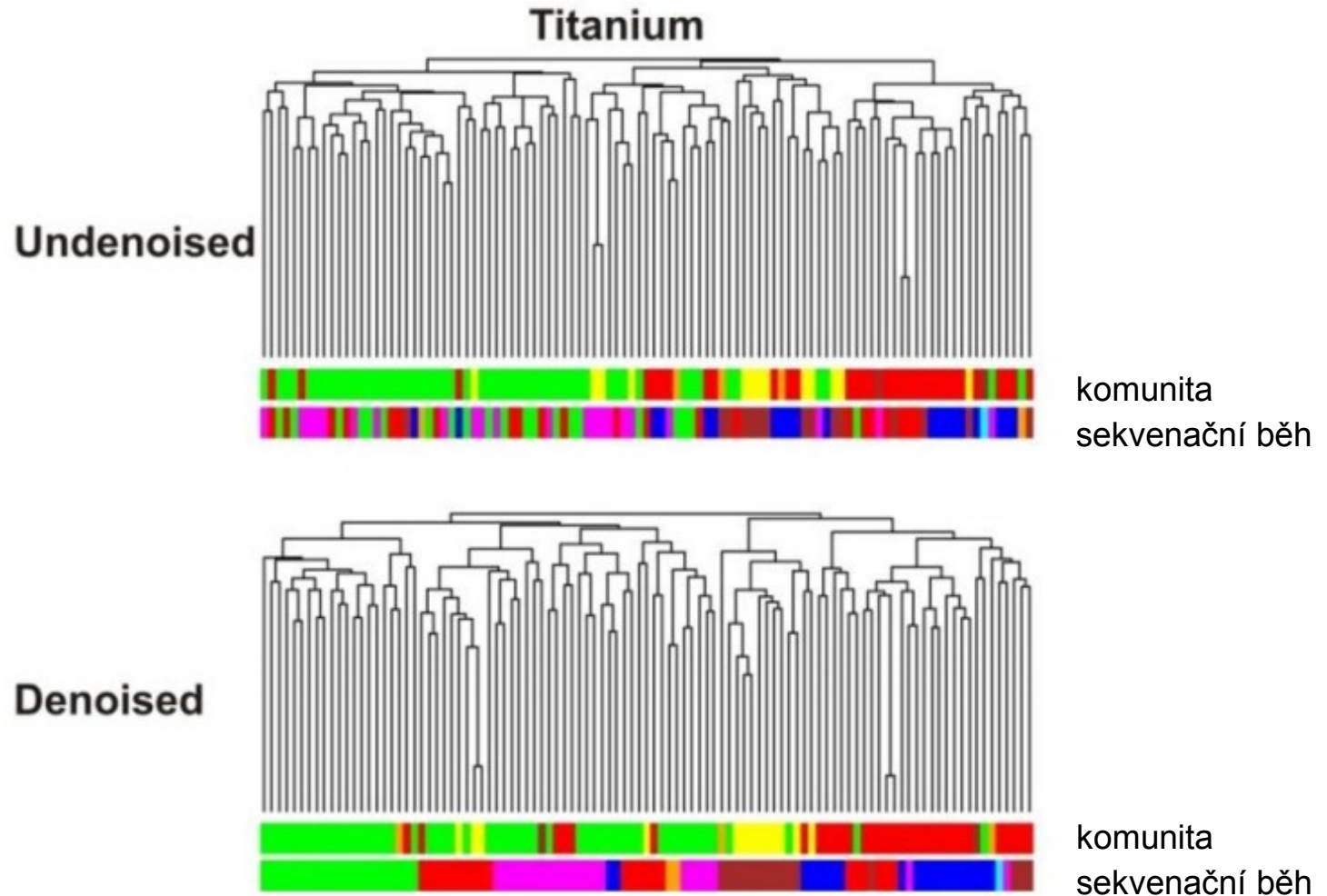
Rozřazení vzorků podle značek

```
>16_5000037 HN7PVAR01BC90H orig_bc=TAGTATCAGC new_bc=TAGTATCAGC
bc_diffs=0
TGTTTGCTACCCATGCTTTCGAGCCTCAGCGTCAGTTAGTGCCAGTAGGCCGCTTCGCCACTGGTGTTCTT
CCCGATATCTACGCATTCACCGCTACACCGGAATTCCGCCTACCTCTACACCCTCAAGATCCACAGTTTT
GAAAGCAGTTCATGGGTTGAGCCCATGTATTTCACTTCCAACCTGCATACCCGCTGCGCTCCCTTACACCC
AGTAATCCGGACAACGCTTGTGACCTACGTTTTACC GCGGCTGCTGGCACGTAGTTAGCCGCTACTTCCTT
TTGGGTACCATCCGTATTTCCCAACAACAGGAGTTTACAATCCGAAGACCTTCTTCTCCACGCGCGTTCG
CTGCATCAGGGTTTCCCCATTGTGCAAT
>17_5000038 HN7PVAR01ALS7Q orig_bc=CGTGTCTCTA new_bc=CGTGTCTCTA
bc_diffs=0
TGTTTGCTCCCCACGCTTTCGCGCCTCACCGTCAGTTGCCGTCAGTTATCCGCCTTCGCCACTGGTGTTCTT
CCTTATATCTACGCATTTACCGCTACACAAGGAATTCGATAACCTCTCCGGTACTCAAGACCAGCAGTTTC
AAATGCAGTTCGTGGGTTAAGCCACCGCATTTACATCTGACTTGCCAGCCCGGCTGCACGCCCTTACACCC
AGTAAATCCGGACAACGCTTGCCACCTACGTATTACC GCGGCTGCTGGCACGTAGTTAGCCGTGGCTTATTCG
TCAGGTACCCTCTTCTACTGTTCCCTGACAAAAGAAGTTTACAACCCGAGGGCCTTCTTCTTACGCGGCGT
TGCTGGGTACGCTTGCGCCCATGCCCCAAT
>16_5000039 HN7PVAR01AGDPN orig_bc=TAGTATCAGC new_bc=TAGTATCAGC
bc_diffs=0
TGTTTGATACCCACGCTTTCGAGCTTCAATGTCAGTCGCGGCTTGGTGGACTGCCTGCGCAATCGGAGTTCTT
CGTGATATCTATGCATTTACCGCTACACCAGGAATTCATCCACCGCAAACGCACTCAAGACTGCCAGTTTC
AACTGCAGCCGACGCTTGGAGCCGCGGATTTCAACTGACTTAACAGCCATCTACGCTCCCTTAAACCC
AATAAATCCGGATAACGCTCGCATCCTACGTATTACC GCGGCTGCTGGCACGTAGTTAGCCGTGGCTTCCTCG
TACACTACCGTCATACCAGCCATTTACACAACCGGCACATTCGTCATGTAACAACAGAGCTTTACAATCCGA
AGACCTTCTTACTACGCGGCGTTGCTCCGTCAGGCTTTCGTCCCACTTGCGGAAGATTCCCTACCTGCCTG
CCTCCGCTGATACTA
>16_5000040 HN7PVAR01AVMTO orig_bc=TAGTATCAGC new_bc=TAGTATCAGC
bc_diffs=0
TGTTTGATACCCGCACTTTCGAGCATCAACGTCAGTTACGGTCCAGCAAGCTGCCTTCGCAATCGGAGTTCTT
CGTGATATCTAGGCATTTACCGCTACACTAGGAATTCGCGCTGCCTTCCAAGTACTCAAGAAGTACAGTTTC
AAATGCAGTTCGGGGTTGAGCCCGGAATTTACATCTGACTTGCAATCCCGCTACACGCCCTTACACCC
AGTAAATCCGGACAACGCTCGCTCCCTACGTATTACC GCGGCTGCTGGCACGTAGTTAGCCGGAGCTTATTCG
TCAGGTACCCTCTTCTACTGTTCCCTGACAAAAGAAGTTTACAACCCGAACGCTTCTTCTTACGCGGCG
TTGCTGGGTACGCTTGCGCCCATGCCCCAATCTTCCCCACCTGCTGCCTCCGCTGATACTA
>11_5000041 HN7PVAR01AC3MJ orig_bc=CTCGCGTGTC new_bc=CTCGCGTGTC
bc_diffs=0
TGTTTGATACCCACGCTTTCGAGCTTCAATGTCAGTTGCGGCTTGGTGGACTGCCTTCGCAATCGGAGTTCTT
CGTGATATCTAAGCATTTACCGCTACACCAGGAATTCATCCACCGCAAACGCACTCAAGGCTACCAGTTTC
AATTGCAGTTCACCGGTTGAGCCGCTGACTTTCAACTGACTTAACAGCCATCTACGCTCCCTTAAACCC
AATAAATCCGGATAACGCTCGCATCCTCCGTATTACC GCGGCTGCTGGCACGGAGTTAGCCGATGCTTATTCG
TACCGGTACATGCACACGACCACAGTGGACGCGGTTATTCCCGTACAAAAGGAGTTTACAACCCGTAAGCCGT
CTTCTCCACGCTACTTGGCTGGTTACGGCTCAGCCCATGACCAAT
>7_5000042 HN7PVAR01A8NE0 orig_bc=AGCACTGTAG new_bc=AGCACTGTAG bc_diffs=0
```

Analýza dat (Qiime)

- Rozřazení vzorků podle značek, spojení sekvencí z Illuminy
- Denoising – odstranění chyb vzniklých sekvenací – pouze 454 data

Denoising – pouze 454



Analýza dat (Qiime)

- Rozřazení vzorků podle značek, spojení sekvencí z Illuminy
- Denoising – odstranění chyb vzniklých sekvenací
- Úprava sekvencí - délka, kvalita

Úprava sekvencí

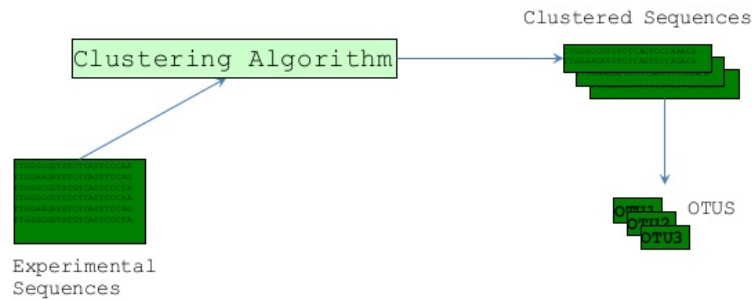
```
>0 17_6006268
GGCCCTCGGGTTGTAAACTTCGTTTTATCAGGGACGAAGCAAGTGACGGTACCTGTATGAATAAGCCACGGCTA
ACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCGAGCGTTATCCGGATTACTGGGTGTAAAGGGCGTGT
AGGCGGGAAGGCAAGTCAGACGTGAAACCACGGGCTCAACCTGTGGCCTGCATTTGAAACTGTAGTTCTTGA
GTGCTGGAGAGGCAATCGGAATCCGTGTGTAGCGGTGAAATGCGTAGATATACGGAGGAACACCAGTGGCGA
AGGCGGATTGCTGGACGATAACTGACGCTGAGGCGCGAAAGCGTGGGG
>1 8_4033127
GTGCTTCAGCGTCCGTCCTGGCCTGGTATGCTGCCTTCGCAATCGGGGTTCTGCGTGATATCTATGCATTTCA
CCGCTACACCATACATTCGCTTACCGCCAACTACTCTCTAGCCCAACAGTATGGAGGCAGTTTCAGGGTTAA
GCCCTAACATTTACCTCCAACCTTATCGAACCGCCTACGCACCCCTTAAACC CAATAAAATCCGGATAACGCTT
GGATCCTCCGTATTACCGCGGCTGCTGGCACGGAGTTAGCCGATCCTTACTCTGACGATACTTTACAGACAGAT
ACACGTATCTGCGTTTACCCTCGTACAAAAGCAGTTTACAACCTCATAG
>10 14_1017803
CTGCCTTCGCGATCGGAGTCTTCGTGATATCTAAGCATTTCACCGCTACACTCGGAATCCGCCAACCTCTA
CTTCACTCAAGAAAACCAGTTTCAACTGCAGTCTACAGGTTAAGCCCGTAGTTTTTACAGCTGACTTGGCTCC
CCGCCTGCGCTCCCTTTACACCAGTAATTCGGGACAACGCTTGCCACCTACGTATTACCGCGGCTGCTGGCA
CGTAGTTAGCCGTGGCTTGCTCCTTAGCTACCGTCACTATCTTCACTAAGAAACAGAAGTTTACAATCCGAAA
CCGTCTTCCTCACGCGCGTTGCTGCATCAGGGTTTCCCCATTGTG
>100 17_6006779
AAGCGTTTCGGGTTGTAAACTTCTTTTCTCAGGGACGAAGAAAATGACGGTACCTGAGGAATAAGCCACGGC
TAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTGTCCGGATTTACTGGGTGTAAAGGGCGC
GTAGGCGGGGAGACAAGTCAGATGTGAAAACCAGGGGCTCAACCTCTGGCCTGCATTTGAAACTGTAGCTCTT
GAGTGTGCGGAGAGGCAATCGGAATCCGTGTGTAGCGGTGAAATGCGTAGATATACGGAGGAACACCAGTGGC
GAAGGCGGATTGCTGGACGATAACTGACGCTGAGGCGCGAAAGCGTGG
>1000 15_1026250
CTGCCTTCGCAATCGGGGTTCTTCGTGATATCTATGCATTTACCGCTACACCACGAATTCGCGATGCCGCGA
CCGTACTCAAGCCCCACAGTTTCAACTGCAATTTTACGGTTGAGCCGCAAACCTTTCACAGCTGACTTAAGGGG
CCGTCTGCGCTCCCTTTAAACCAATAAACTCCGGATAAACGCTCGCATCCTCCGTATTACCGCGGCTGCTGGC
ACGGAGTTAGCCGATGCTTTTTCTTCGGATACTTGCAATACGCTACACGTAGCGCACTTTACTCTCCGACAAA
ACGAAGTTTACAACCCGTAGGGCCGTCTTCCTTACGCGACTTGGCTG
>10000 K3.1_20000229
GTATCTCGGTATGTAAAGCTCTATCAGCAGGGAAGAAAATGACGGTACCTGACTAAGAAGCACCGGCTAAATA
CGTGCCAGCAGCCGCGGTAATACGTATGGTGAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGAC
GGAGAAGCAAGTCTGGAGTGAAAACCCGGGCTCAACCCCGGACTGCTTTGAAACTGTTTTTCTAGAGTGC
CGGAGAGGTAAGCGGAATTCCTAGTGTAGCGGTGAAATGCGTAGATATTAGGAGGAACACCAGTGGCGAAGGC
GGCTTACTGGACGTTAACTGACGTTGAGGCTCGAAAGCGTGGGAGCAA
~10001 14_1025021
```

Analýza dat (Qiime)

- Rozřazení vzorků podle značek, spojení sekvencí z Illuminy
- Denoising – odstranění chyb vzniklých sekvenací
- Úprava sekvencí - délka, kvalita
- Tvorba OTUs na základě 97% podobnosti
- Tvorba reprezentativních sekvencí a jejich taxonomické zařazení (RDP)

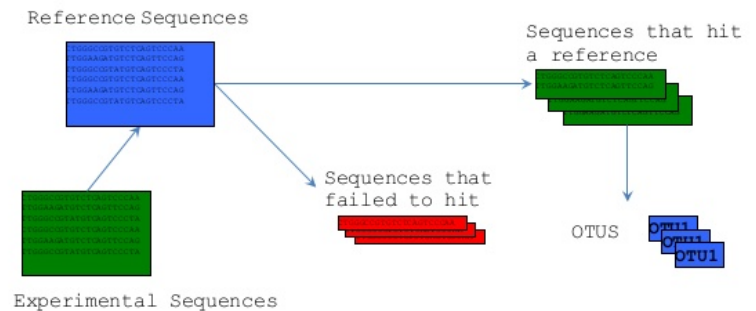
Tvorba OTUs

OTU Picking - “de-novo”



- Pros
 - Vast majority of reads are clustered
 - No reference database bias
- Cons
 - Speed; not easily parallelizable
 - Erroneous reads get clustered

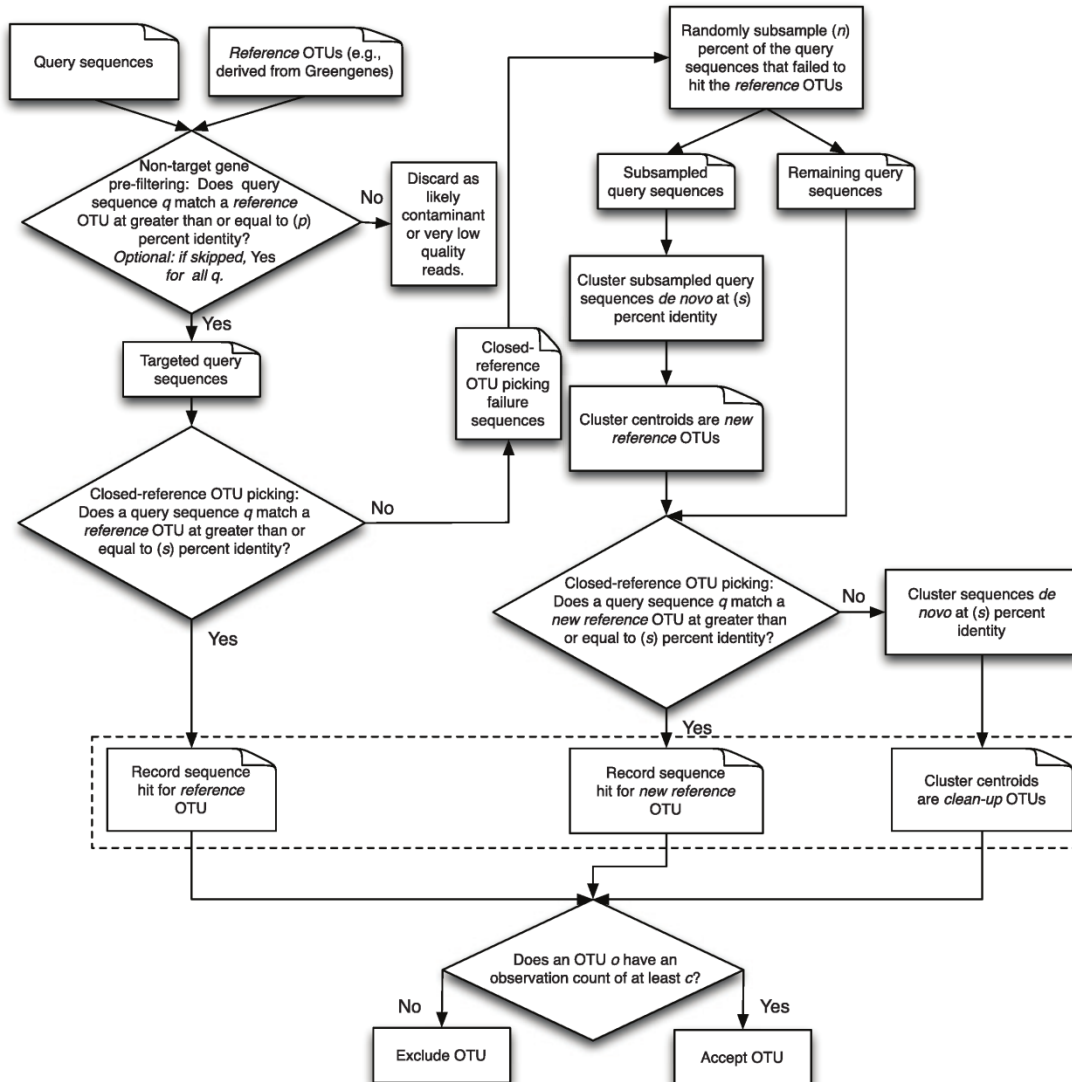
OTU Picking - “closed-reference”



- Pros
 - Reference database is a quality filter
 - Speed; easily parallelizable
- Cons
 - No new OTUs can be observed
 - Reference database bias

Tvorba OTUs

Subsampled open-reference OTU picking workflow



Legend



Data file (input, intermediate, or output)



Per-sequence decision. These are applied individually to all sequences provided as input.



Process applied to all sequences provided as input as a collection.



Output OTUs

(*p*): percent sequence identity threshold used for pre-filtering of sequences (default: 60%)

(*s*): percent sequence identity threshold used when clustering sequences either *de novo* or closed-reference (default: 97%)

(*n*): percentage of sequences that are randomly subsampled from sequences that failed to hit *reference* OTUs (default: 0.1%)

(*c*): minimum observation count for an OTU to be accepted during post-OTU picking processing (default: 2)

Tvorba OTUs

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	# QIIME v1.4.0 OTU table																		
2	#OTU ID	1	2	3	4	5	6	7	8	9	Consensus Lineage								
3	0	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae								
4	1	0	0	0	0	0	0	0	1	0	0 Root;k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae								
5	2	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Lactobacillaceae								
6	4	0	0	0	0	0	0	0	1	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae								
7	7	0	0	0	0	0	0	3	2	0	0 Root;k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f__								
8	10	0	0	0	0	1	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae								
9	11	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae								
10	12	0	0	0	0	0	0	0	1	0	0 Root;k_Bacteria								
11	14	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteria								
12	15	0	0	0	0	0	0	0	1	0	0 Root;k_Bacteria;p_Proteobacteria;c_Epsilonproteobacteria;o_Campylobacterales;f_Helicobactera								
13	16	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae								
14	18	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae								
15	23	0	0	2	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae								
16	25	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Actinobacteria;c_Actinobacteria;o_Bifidobacteriales;f__								
17	32	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f__								
18	34	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales								
19	35	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae								
20	36	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f__								
21	38	0	0	0	0	0	0	16	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f__								
22	40	0	0	0	0	0	0	2	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae								
23	41	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae								
24	42	0	2	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae								
25	45	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Tenericutes;c_Erysipelotrichi;o_Erysipelotrichales;f_Erysipelotrichaceae								
26	46	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Bacilli;o_Lactobacillales;f_Lactobacillaceae								
27	49	0	0	0	0	0	0	0	1	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae								
28	50	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae								
29	51	0	1	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae								
30	52	0	0	1	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae								
31	53	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Catabacteriaceae								
32	55	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Rikenellaceae								
33	56	1	1	0	0	1	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae								
34	59	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Ruminococcaceae								
35	66	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Bacteroidaceae								
36	68	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f__								
37	70	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f__								
38	73	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Firmicutes;c_Clostridia;o_Clostridiales;f_Lachnospiraceae								
39	77	0	0	0	0	0	0	0	0	0	0 Root;k_Bacteria;p_Bacteroidetes;c_Bacteroidia;o_Bacteroidales;f_Porphyromonadaceae								

Analýza dat (Qiime)

- Rozřazení vzorků podle značek, spojení sekvencí z Illuminy
- Denoising – odstranění chyb vzniklých sekvenací
- Úprava sekvencí - délka, kvalita
- Tvorba OTUs na základě 97% podobnosti
- Tvorba reprezentativních sekvencí a jejich taxonomické zařazení (RDP)
- Odstranění chimér
- Vizualizace
- Výpočet indikátorů diverzity, rarefakční křivky

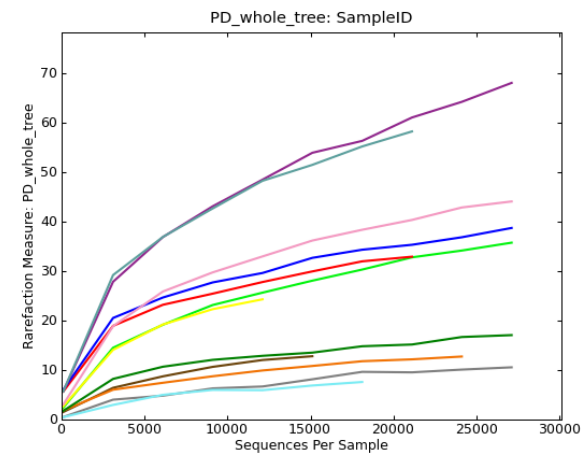
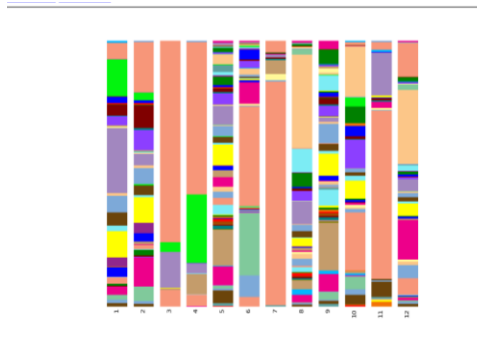
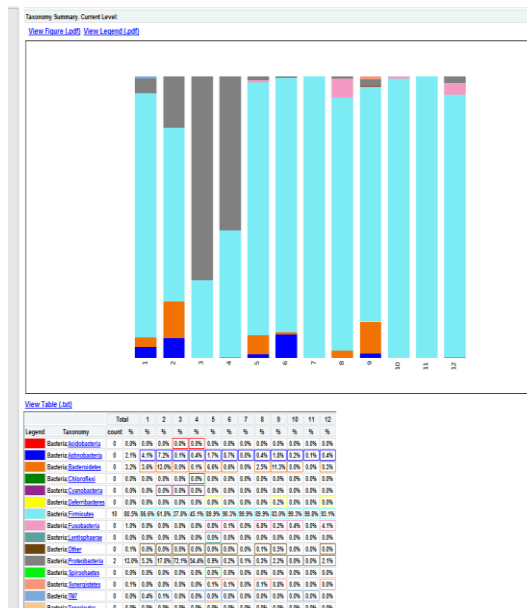
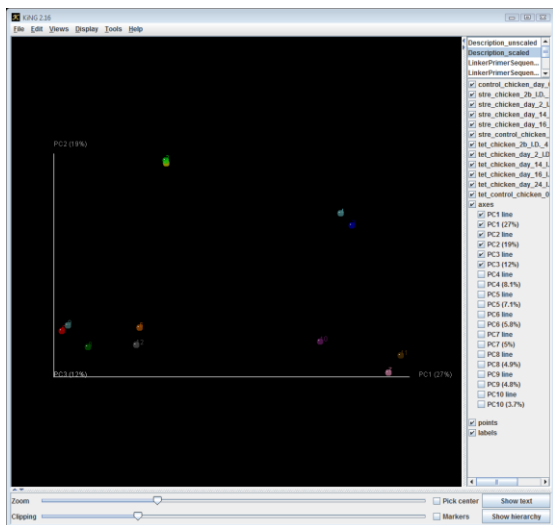
Třídění sekvencí

- U 454 - potřeba fna a qual soubor nebo sff file
- U Illumina - FastQ
- Mapping file – nutné zadat jméno vzorků, barcode, primer, název
- Lze zadat mnoho dalších kritérií (minimální a maximální délku, počet chyb v primeru, počet chyb v MIDu - značící sekvenci, minimum quality score,...)

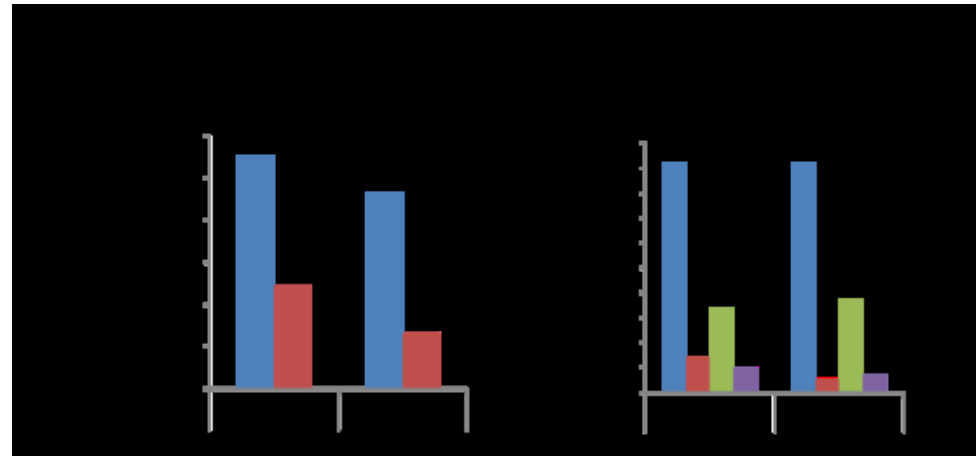
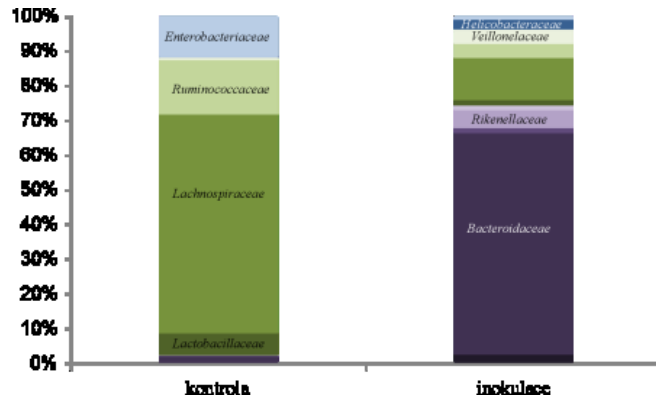
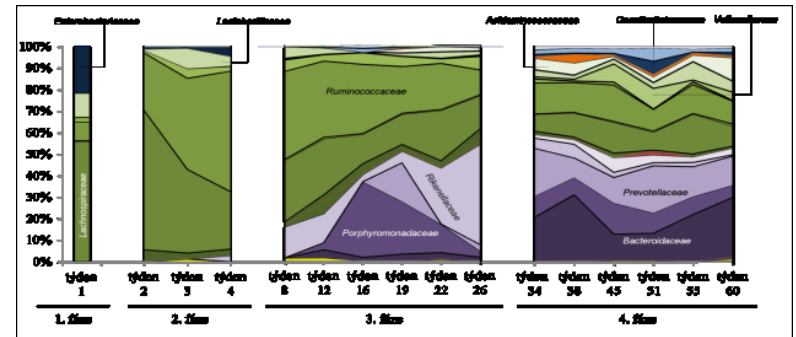
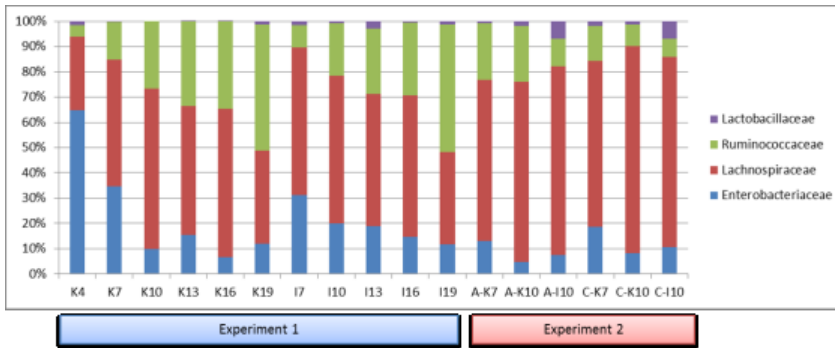
Ukázka mapping file

#SampleID	BarcodeSequence	LinkerPrimerSequence	ReversePrimer	Treatment	Description
1	ACGCTCGACA	GGAGGCAGCAGTRRGAAT	CTACCRGGGTATCTAATCC	Control_1	stre_control_chicken_0
2	AGACGCACTC	GGAGGCAGCAGTRRGAAT	CTACCRGGGTATCTAATCC	Control_2	tet_control_chicken_0
3	AGCACTGTAG	GGAGGCAGCAGTRRGAAT	CTACCRGGGTATCTAATCC	atb_1	stre_chicken_2b_I.D._3
4	ATCAGACACG	GGAGGCAGCAGTRRGAAT	CTACCRGGGTATCTAATCC	atb_2	tet_chicken_2b_I.D._4
#SampleID	BarcodeSequence	LinkerPrimerSequence	ReversePrimer	Treatment	Description
1	ACGCTCGACA	CTACCRGGGTATCTAATCC	GGAGGCAGCAGTRRGAAT	Control_1	stre_control_chicken_0
2	AGACGCACTC	CTACCRGGGTATCTAATCC	GGAGGCAGCAGTRRGAAT	Control_2	tet_control_chicken_0_I
3	AGCACTGTAG	CTACCRGGGTATCTAATCC	GGAGGCAGCAGTRRGAAT	atb_1	stre_chicken_2b_I.D._3
4	ATCAGACACG	CTACCRGGGTATCTAATCC	GGAGGCAGCAGTRRGAAT	atb_2	tet_chicken_2b_I.D._4

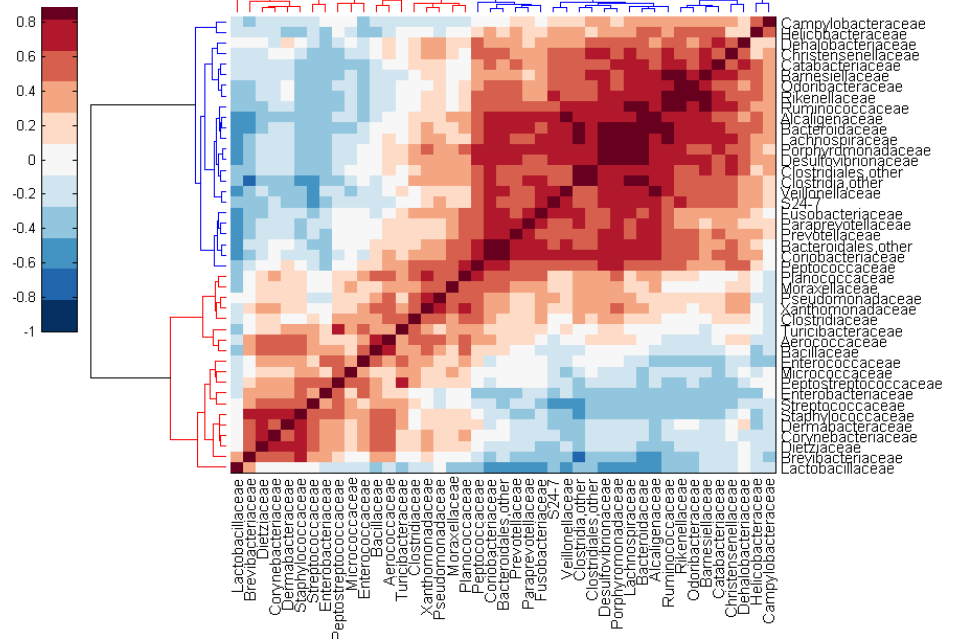
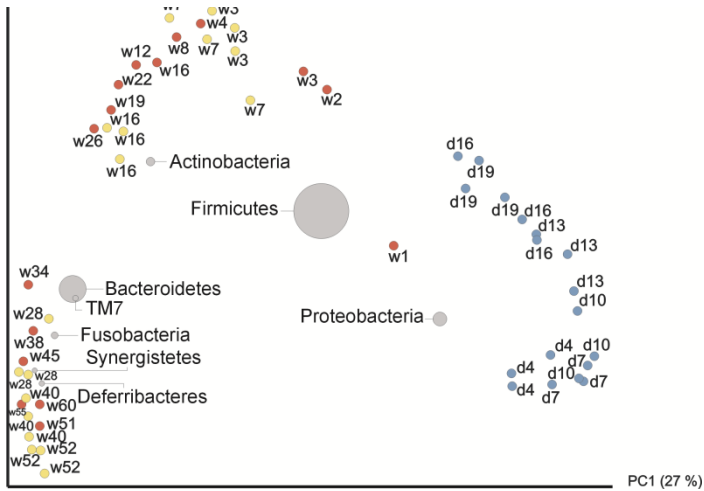
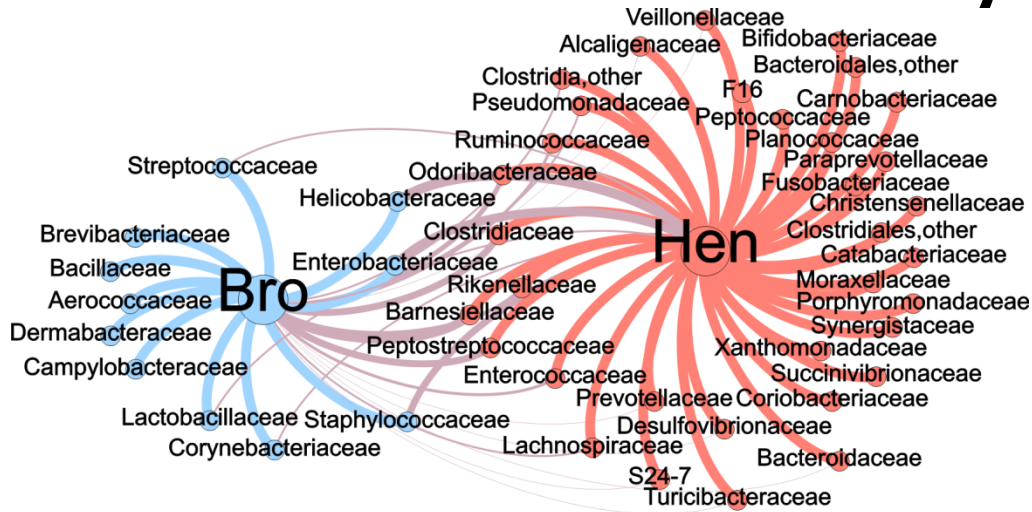
Ukázka výstupů - grafy



Analýza



Analyza



Vyhodnocování

Qiime

- http://www.qiime.org/svn_documentation/index.html
- Pracuje se v příkazovém řádku, nutné znát základní příkazy

index

QIIME Tutorials

The QIIME tutorials are documents that illustrate how to use various features of the QIIME. We recommend that all users begin with the [QIIME overview tutorial](#) which takes the user through a full analysis of sequencing data. After you've begun analyzing your own data, you'll want to move on to the special-purpose tutorials as needed.

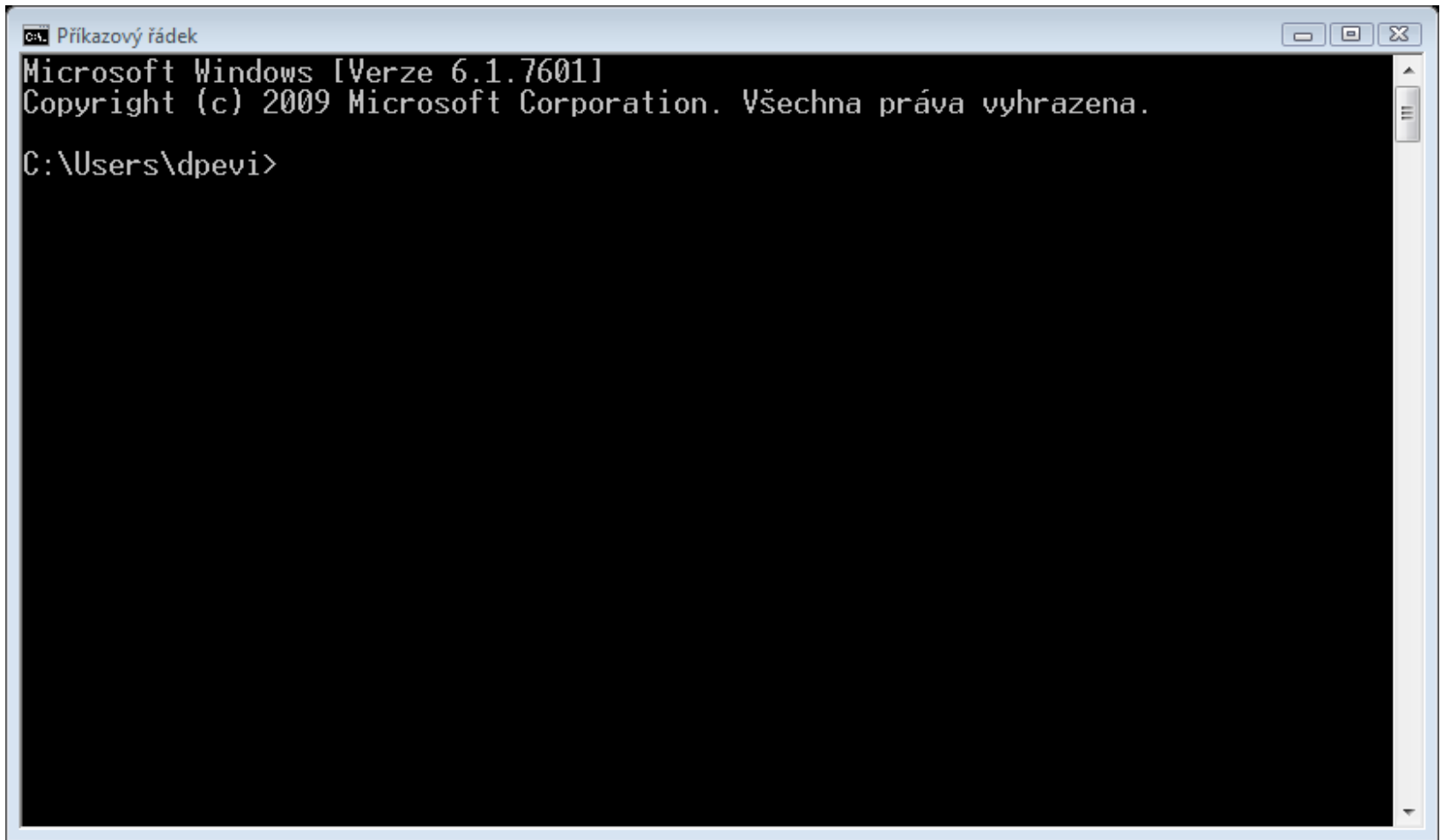
The QIIME development team is very interested in incorporating community-contributed tutorials. Please contact qiime.help@colorado.edu if you're interested in contributing.

Contents

- [QIIME Overview Tutorial](#)
- [Multi-step OTU picking](#)
- [Chimera checking sequences with QIIME](#)
- [Creating Distance Comparison Plots](#)
- [Denoising of 454 Data Sets](#)
- [Using OTUpipeline with QIIME](#)
- [Using parallel QIIME](#)
- [Processing 18S data](#)
- [Processing Illumina Data](#)
- [Performing Procrustes Analysis](#)
- [Re-training the RDP Classifier](#)
- [Running Supervised Learning](#)
- [Basic Unix/Linux/OS X commands](#)
- [Working with QIIME on Amazon Web Services EC2](#)



Příkazový řádek

A screenshot of the Windows Command Prompt window. The title bar at the top reads "Příkazový řádek" (Command Prompt) and includes standard window control buttons (minimize, maximize, close). The main area is black with white text. The text displayed is: "Microsoft Windows [Verze 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Všechna práva vyhrazena.
C:\Users\dpevi>". The prompt character ">" is positioned at the end of the line, indicating the system is ready for a command.

```
Microsoft Windows [Verze 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. Všechna práva vyhrazena.
C:\Users\dpevi>
```

Příkazy

```
check_id_map.py -m mapa.txt -o mapping_output -v  
quality_scores_plot.py -q seqs.qual -o quality_histograms/ -s 20
```

```
split_libraries.py -m mapa.txt -f IR -q IR -n 1000000 -o output_split_lib2/ -M 1 -b 13 -z truncate_only -l
```

```
sed '/^[A-Z]/s/[A-Za-z]{50}\{([A-Za-z]{340}\)}.*\1/' output_split_lib/seqs.fna > cutseq.fna
```

```
pick_otus.py -i cutseq.fna -o picked_otus/
```

```
pick_rep_set.py -i picked_otus/cutseq_otus.txt -f cutseq.fna
```

```
assign_taxonomy.py -i cutseq.fna_rep_set.fasta -c 0.5 -o assigned_taxonomy/
```

```
parallel_align_seqs_pynast.py -i cutseq.fna_rep_set.fasta -t core_set_aligned.fasta -O 6 -o parallel_align
```

```
parallel_identify_chimeric_seqs.py -m ChimeraSlayer -i parallel_align/cutseq.fna_rep_set_aligned.fasta -a  
core_set_aligned.fasta -o chimeric_seqs.txt -v
```

```
filter_fasta.py -f parallel_align/cutseq.fna_rep_set_aligned.fasta -o non_chimeric_rep_set_aligned.fasta -s  
chimeric_seqs.txt -n
```

```
make_otu_table.py -i otu_map.txt -o otu_table.biom -e chimeric_seqs.txt -t taxonomy.txt
```

Ukázka příkazů

split_libraries.py - Split libraries according to barcodes specified in mapping file

Description:

Since newer sequencing technologies provide many reads per run (e.g. the 454 GS FLX Titanium series can produce 400-600 million base pairs with 400-500 base pair read lengths) researchers are now finding it useful to combine multiple samples into a single 454 run. This multiplexing is achieved through the application of a pyrosequencing-tailored nucleotide barcode design (described in (Parameswaran et al., 2007)). By assigning individual, unique sample specific barcodes, multiple sequencing runs may be performed in parallel and the resulting reads can later be binned according to sample. The script `split_libraries.py` performs this task, in addition to several quality filtering steps including user defined cut-offs for: sequence lengths; end-trimming; minimum quality score. To summarize, by using the `fasta`, `mapping`, and `quality` files, the program `split_libraries.py` will parse sequences that meet user defined quality thresholds and then rename each read with the appropriate Sample ID, thus formatting the sequence data for downstream analysis. If a combination of different sequencing technologies are used in any particular study, `split_libraries.py` can be used to perform the quality-filtering for each library individually and the output may then be combined.

Sequences from samples that are not found in the mapping file (no corresponding barcode) and sequences without the correct primer sequence will be excluded. Additional scripts can be used to exclude sequences that match a given reference sequence (e.g. the human genome; `exclude_seqs_by_blast.py`) and/or sequences that are flagged as chimeras (`identify_chimeric_seqs.py`).

Usage: `split_libraries.py` [options]

Input Arguments:

```
[REQUIRED]
-m, --map
    Name of mapping file. NOTE: Must contain a header line indicating SampleID in the first column and BarcodeSequence in the second, LinkerPrimerSequence in the third.
-f, --fasta
    Names of fasta files, comma-delimited

[OPTIONAL]
-q, --qual
    Names of qual files, comma-delimited [default: None]
-r, --remove_unassigned
    DEPRECATED: pass --retain_unassigned_reads to keep unassigned reads [default: None]
-l, --min-seq-length
    Minimum sequence length, in nucleotides [default: 200]
-L, --max-seq-length
    Maximum sequence length, in nucleotides [default: 1000]
-t, --trim-seq-length
    Calculate sequence lengths after trimming primers and barcodes [default: False]
-s, --min-qual-score
    Min average qual score allowed in read [default: 25]
-k, --keep-primer
    Do not remove primer from sequences
```

Ukázka příkazů

Output:

Three files are generated by split_libraries.py:

1. .fna file (e.g. seqs.fna) - This is a FASTA file containing all sequences which meet the user-defined parameters, where each sequence identifier now contains its corresponding sample id from mapping file.
2. histograms.txt- This contains the counts of sequences with a particular length.
3. split_library_log.txt - This file contains a summary of the split_libraries.py analysis. Specifically, this file includes information regarding the number of sequences that pass quality control (number of seqs written) and how these are distributed across the different samples which, through the use of bar-coding technology, would have been pooled into a single 454 run. The number of sequences that pass quality control will depend on length restrictions, number of ambiguous bases, max homopolymer runs, barcode check, etc. All of these parameters are summarized in this file. If raw sequences do not meet the specified quality thresholds they will be omitted from downstream analysis. Since we never see a perfect 454 sequencing run, the number of sequences written should always be less than the number of raw sequences. The number of sequences that are retained for analysis will depend on the quality of the 454 run itself in addition to the default data filtering thresholds in the split_libraries.py script. The default parameters (minimum quality score = 25, minimum/maximum length = 200/1000, no ambiguous bases allowed, no mismatches allowed in primer sequence) can be adjusted to meet the user's needs.

Standard Example:

Using a single 454 run, which contains a single FASTA, QUAL, and mapping file while using default parameters and outputting the data into the Directory "Split_Library_Output":

```
split_libraries.py -m Mapping_File.txt -f 1.TCA.454Reads.fna -q 1.TCA.454Reads.qual -o Split_Library_Output/
```

Multiple FASTA and QUAL Files Example:

For the case where there are multiple FASTA and QUAL files, the user can run the following comma-separated command as long as there are not duplicate barcodes listed in the mapping file:

```
split_libraries.py -m Mapping_File.txt -f 1.TCA.454Reads.fna,2.TCA.454Reads.fna -q 1.TCA.454Reads.qual,2.TCA.454Reads.qual -o Split_Library_Output_comma_separated/
```

Duplicate Barcode Example:

An example of this situation would be a study with 1200 samples. You wish to have 400 samples per run, so you split the analysis into three runs and reuse barcoded primers (you only have 600). After initial analysis you determine a small subset is underrepresented (<500 sequences per samples) and you boost the number of sequences per sample for this subset by running a fourth run. Since the same sample IDs are in more than one run, it is likely that some sequences will be assigned the same unique identifier by split_libraries.py when it is run separately on the four different runs, each with their own barcode file. This will cause a problem in file concatenation of the four different runs into a single large file. To avoid this, you can use the '-n' parameter which defines a start index for split_libraries.py. From experience, most FLX runs (when combining both files for a single plate) will have 350,000 to 650,000 sequences. Thus, if Run 1 for split_libraries.py uses '-n 1000000', Run 2 uses '-n 2000000', etc., then you are guaranteed to have unique identifiers after concatenating the results of multiple FLX runs. With newer technologies you will just need to make sure that your start index spacing is greater than the potential number of sequences.

To run split_libraries.py, you will need two or more (depending on the number of times the barcodes were reused) separate mapping files (one for each Run, for example one for Run1 and another one for Run2), then you can run split_libraries.py using the FASTA and mapping file for Run1 and FASTA and mapping file for Run2. Once you have run split_libraries on each file independently, you can concatenate (e.g. using the 'cat' command) the sequence files that were generated by split_libraries.py. You can also concatenate the mapping files, since the barcodes are not necessary for downstream analyses, unless the same sample IDs are found in multiple mapping files.

Run split_libraries.py on Run 1:

```
split_libraries.py -m Mapping_File.txt -f 1.TCA.454Reads.fna -q 1.TCA.454Reads.qual -o Split_Library_Run1_Output/ -n 1000000
```

mothur

- <http://www.mothur.org/>



The screenshot shows the homepage of the mothur project. At the top, the word "mothur" is displayed in a bold, white font on a dark blue background. Below this, there is a navigation bar with four links: "Download", "Wiki", "Forum", and "facebook", each in white text on a dark blue background. The main content area has a light blue background. On the left, there is a paragraph of text welcoming visitors to the website, mentioning Dr. Patrick Schloss and the Department of Microbiology & Immunology at The University of Michigan. On the right, there is a circular logo featuring a stylized woman's face with a red headscarf and a white shirt with DNA sequence letters (A, T, C, G) on it. The logo is surrounded by a colorful, swirling border. At the bottom of the page, there is a dark blue footer containing contact information for the Department of Microbiology & Immunology at The University of Michigan Medical School, and a note that the site is maintained by Pat Schloss, with a copyright notice for 2008-2009.

mothur

Download **Wiki** **Forum** **facebook**

Welcome to the website for the mothur project, initiated by [Dr. Patrick Schloss](#) and his software development team in the [Department of Microbiology & Immunology](#) at [The University of Michigan](#). This project seeks to develop a single piece of open-source, expandable software to fill the bioinformatics needs of the microbial ecology community. In February 2009 we released the first version of mothur, which had accelerated versions of the popular DOTUR and SONS programs. Since then we have added the functionality of a number of other popular tools. mothur is currently the most cited bioinformatics tool for analyzing 16S rRNA gene sequences. Step inside the wiki and user forum and learn how you can use mothur to process data generated by Sanger, PacBio, IonTorrent, 454, and Illumina (MiSeq/HiSeq). If you would like to contribute code to the project feel free to download the source code and make your own improvements. Alternatively, if you have an idea or a need, but lack the programming expertise, let us know and we'll add it to the queue of features we would like to add. Our current goal is to release a new iteration of the project every couple of months.

Department of Microbiology & Immunology
The University of Michigan Medical School
The University of Michigan

This site is maintained by Pat Schloss
© 2008-2009

<https://www.youtube.com/watch?v=X4aV4J8FkEU&nohtml5=False>

Qiime vs. mothur

- <http://blog.mothur.org/2016/01/12/mothur-and-qiime/>

mothur and QIIME

Jan 12, 2016 • PD Schloss • 32 min read

Despite their differences in philosophy, *most* of the differences in mothur and QIIME are cosmetic. Both packages have been successful. Having both of them around is good for microbial ecology. Within both packages there are warts - inconveniences to the users and antiquated/bad ideas. Within both packages there are strengths. If you are going to criticize someone for their choice of software, do it for some specific point. If you are going to campaign for mothur or QIIME, do your best to accurately represent the strengths of your pet package.

RDP

<http://rdp.cme.msu.edu>

[ABOUT RDP](#) | [ASSIGNMENT GENERATOR](#) | [CITATION](#) | [CONTACTS](#) | [RELATED SITES](#) | [RESOURCES](#) | [TUTORIALS](#) | [USER WIKI](#)



ANNOUNCEMENTS

RDP News

[10/07/2015 Xander assembler article is published.](#)

Xander: Employing a Novel Method for Efficient Gene-Targeted Metagenomic Assembly

[10/07/2015 Warcup Fungal ITS article is accepted!](#)

Fungal identification using a Bayesian Classifier and the 'Warcup' training set of Internal Transcribed Spacer sequences.

[07/08/2015 *** Pyro Job Submission up ***](#)

Hardware Issues causing pyro Issues now fixed

[05/28/2015 RDP Staff attending ASM Meeting In New Orleans](#)

RDP staff will be attending the ASM General Meeting in New Orleans in the coming week. Two RDP posters will be presented: first on Tuesday morning:...

[05/26/2015 RDP Release 11.4 available](#)

Updated 16S rRNA hierarchy model to training set No. 14.

[03/27/2015 FrameBot new option Add de novo to references available](#)

Unique abundant query sequences will be added to the starting reference set if qualifications are met.

[02/23/2015 WARNING -- RDP unavailable Sat., March 7th](#)

Building network infrastructure upgrades planned 8 A.M. through 6 P.M.

[02/16/2015 Introducing Xander assembler](#)

RDP's new gene-target metagenomic assembler, Xander, is released

[10/21/2014 Classifier provides gene copy number adjustment](#)

RDP Classifier provides gene copy number adjustment for 16S gene sequences.

[09/17/2014 Using RDPTools Output with Phyloseq](#)

A comprehensive tutorial using RDPTools output with Phyloseq package released

RDP Release 11, Update 4 :: May 26, 2015

3,224,600 16S rRNAs :: 108,901 Fungal 28S rRNAs
Find out what's new in RDP Release 11.4 [here](#).



[Cite RDP's latest tool articles.](#)

RDP provides quality-controlled, aligned and annotated Bacterial and Archaeal 16S rRNA sequences, and Fungal 28S rRNA sequences, and a suite of analysis tools to the scientific community. New to RDP release 11:

- RDP tools have been updated to work with the new fungal 28S rRNA sequence collection.
- A new Fungal 28S Aligner and updated Bacterial and Archaeal 16S Aligner. We optimized the parameters for these secondary-structure based Infernal aligners to provide improved handling for partial sequences.
- Updated RDPipeline offers extended processing and analysis tools to process high-throughput sequencing data, including single-strand and paired-end reads.
- Most of the RDP tools are now available as open source packages for users to incorporate in their local workflow.



RDP's mission and funding:

Part of RDP's mission is to provide support to our users. Email and phone contacts are available on the [contacts page](#).



Questions/comments: rdpstaff@msu.edu



http://rdp.cme.msu.edu/tutorials/init_process/RDPtutorial_INITIAL-PROCESS.html

http://rdp.cme.msu.edu/tutorials/init_process/RDPtutorial_INITIAL-PROCESS_pe.html

http://rdp.cme.msu.edu/tutorials/Submission_Tools/fastq.html

RDP

RDP'S PYROSEQUENCING PIPELINE

[[Help](#) | [FunGene Home](#) | [RDP Home](#)]

About the RDP's Pyrosequencing Pipeline

The Ribosomal Database Project's Pyrosequencing Pipeline aims to simplify the processing of large 16S rRNA sequence libraries obtained through pyrosequencing. This site processes and converts the data to formats suitable for common ecological and statistical packages such as SPADE, EstimateS, and R.

HOVER over tool menu item for a popup description;
CLICK on the tool menu item to begin working with it.

Note: If you experience problems with the pyrosequencing pipeline, please contact us and we will help you get your sequences processed.

[Pyro News](#)

NCBI/EBI Submission Tools:

ENA SEQUENCE READ ARCHIVE

FASTQ

Data Processing Steps:

PIPELINE INITIAL PROCESS

ALIGNER

COMPLETE LINKAGE CLUSTERING

Formats for Common Programs:

SPADE FORMATTER

CIUSTER TO R FORMATTER

ESTIMATE S FORMATTER

MOTHUR: COLUMN DISTANCE MATRIX

MOTHUR: PHYLIP DISTANCE MATRIX

Analysis Tools:

SHANNON & CHAO1 INDEX

JACCARD & SORENSEN INDEX

RAREFACTION

RDP CLASSIFIER

RDP LIB COMPARE

MOCK COMMUNITY ANALYSIS

Miscellaneous Utilities:

ALIGNMENT MERGER

REPRESENTATIVE SEQUENCE

SEQUENCE SELECTION

10/09/2013 FunGene article published
The article describing our FunGene data and tools is published in *Frontier in Microbiology*.

10/09/2013 RDP FrameBot article published
The article describing RDP FrameBot (a frameshift correction tool) is published in the journal *mBio*

10/01/2013 RDP Staff and Poster
5th Argonne Soil Metagenomics Meeting

09/25/2013 Campus internet interruptions
RDP is back online.

08/16/2013 Power outage alert
RDP sites are now back online

06/06/2013 Amplicon chimera checking with uchime
The functional gene pipeline now offers a tool to check amplicon sequencing datasets for chimeras powered by 03/02/2013 System Maintenance 2013-03-06
All websites will be taken offline Wednesday 5-7pm eastern time for regular system maintenance

12/11/2012 RDP Hardware Upgrades
To keep up with increasing demands on our sites we've added 5 new hosts to the Pyro, Fungene, and CME website cluster.

Additionally we...

11/26/2012 Scheduled site maintenance
The pyro, fungene, and cme websites will be briefly unavailable this afternoon while we perform some routine hardware maintenance. The downtime wi...

11/11/2012 Announcing the new RDP Wiki!
We are pleased to announce the release of the RDP Wiki, available at <http://rdp.cme.msu.edu/wiki>

GreenGenes

<http://greengenes.lbl.gov>

May 2013 Notice: The most recent Greengenes database and taxonomy updates are now found at greengenes.secondgenome.com. Taxonomic information on this site is deprecated and should be used with caution.



16S rRNA gene database and
workbench compatible with ARB
greengenes.lbl.gov



Functions

Home
Browse
Export
Slice
Consensus
Compare
Search
Probe
Align
Trim
Download
Curate
More Tools...

About

Citation
Tutorial
FAQ
Objectives
Methods
Contact

My Interest List

remove all
collapse all
show marked

My Taxonomy

greengenes
Activate
Changing
taxonomy will
empty My
Interest List.

greengenes: 16S rDNA data and tools

The greengenes web application provides access to the 2011 version of the greengenes 16S rRNA gene sequence alignment for browsing, blasting, probing, and downloading. The data and tools presented by greengenes probes, interpreting microarray results, and aligning/annotating novel sequences. If you are an **ARB** user, you can use greengenes to keep your own local database current.

The May 2013 greengenes taxonomy now available!

The trees, tables and sequence data organized as described in McDonald et al. ISME J article can all be found in **one place**.

News:

- Search is now possible using new **Simrank** (developed by Niels Larsen) for similarity searching against the 2011 greengenes sequences.
- New **import filter template** posted for slurping greengenes exports into ARB.
- Looking for Hugenholtz or PHPR taxonomy? It is now the greengenes taxonomy.
- Dr. Mike Dyall-Smith has graciously made available his tutorial for **installing Arb on MacOSX**. Thanks Mike.
- The greengenes taxonomy for the Cyanobacteria is now consistent with **cyanoDB** using cyanoDB type species as a guide to map cyanoDB taxonomy to the greengenes reference 16S tree.
- Thanks to Greg Caporaso and Rob Knight for posting **OTU reference and utility files** for use with QIIME software.
- **The Wall Street Journal picks the Berkeley PhyloChip as the top advance in environmental technology of 2008 and 3rd best innovation overall.**
- **Pollution Engineering Magazine selects Berkeley PhyloChip as most likely to aid pollution control and abatement in the near future.**
- **The Berkeley PhyloChip wins R&D100 award as one of the 100 most significant technological advances of the year.**
- Are you the world expert on the taxonomy of a particular phylogenetic lineage? Have you checked this database and nobody has got it right? **Tell us!** - we will fix it.
- We thank Jakob Fredslund for developing a tool, **Gexcellent**, to convert XML trees to Newick format!
- We thank **J.P. Euzéby** and Hans Trüper for expert **etymological advice**.



Browse taxonomic tree of your choice and mark nodes.



Export sequence records of your choice.



Specify a **Slice** (sub-alignment) of the prokMSA to view/download.



Calculate **Consensus** sequences from My Interest List (soon!).



Compare **my local sequences/probes** against the prokMSA using BI AST or Simrank.

http://greengenes.lbl.gov/cgi-bin/JD_Tutorial/nph-Tutorial_2Main2.cgi

Megan

- <http://ab.inf.uni-tuebingen.de/software/megan5/>

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

MATHEMATISCH-
NATURWISSENSCHAFTLICHE FAKULTÄT
Algorithms in Bioinformatics

Search →

Uni A-Z →
Login →

HOME NEWS PEOPLE RESEARCH PUBLICATIONS **SOFTWARE** TALKS TEACHING THESES/PROJECTS ADDRESS DOWNLOADS

Home > Software > MEGAN

ALGORITHMS IN BIOINFORMATICS

CGViz

Copycat

Darwin Rocks!

Crosslink

Dendroscope

LOCAS

MEGAN

How to use BLAST

Old datasets

comparative

MEG2DIST

FESIN 2010

Review

datasets

download

Metasim

microHARVESTER

NRPSpredictor

OSLay

PAT

PAUDA- High-throughput protein aligner

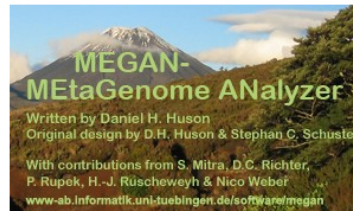
ReadSim

MEGAN 4 - MEtaGenome Analyzer

Software for analyzing metagenomes.

[\(Download here\)](#)

Over 7000 registered users.



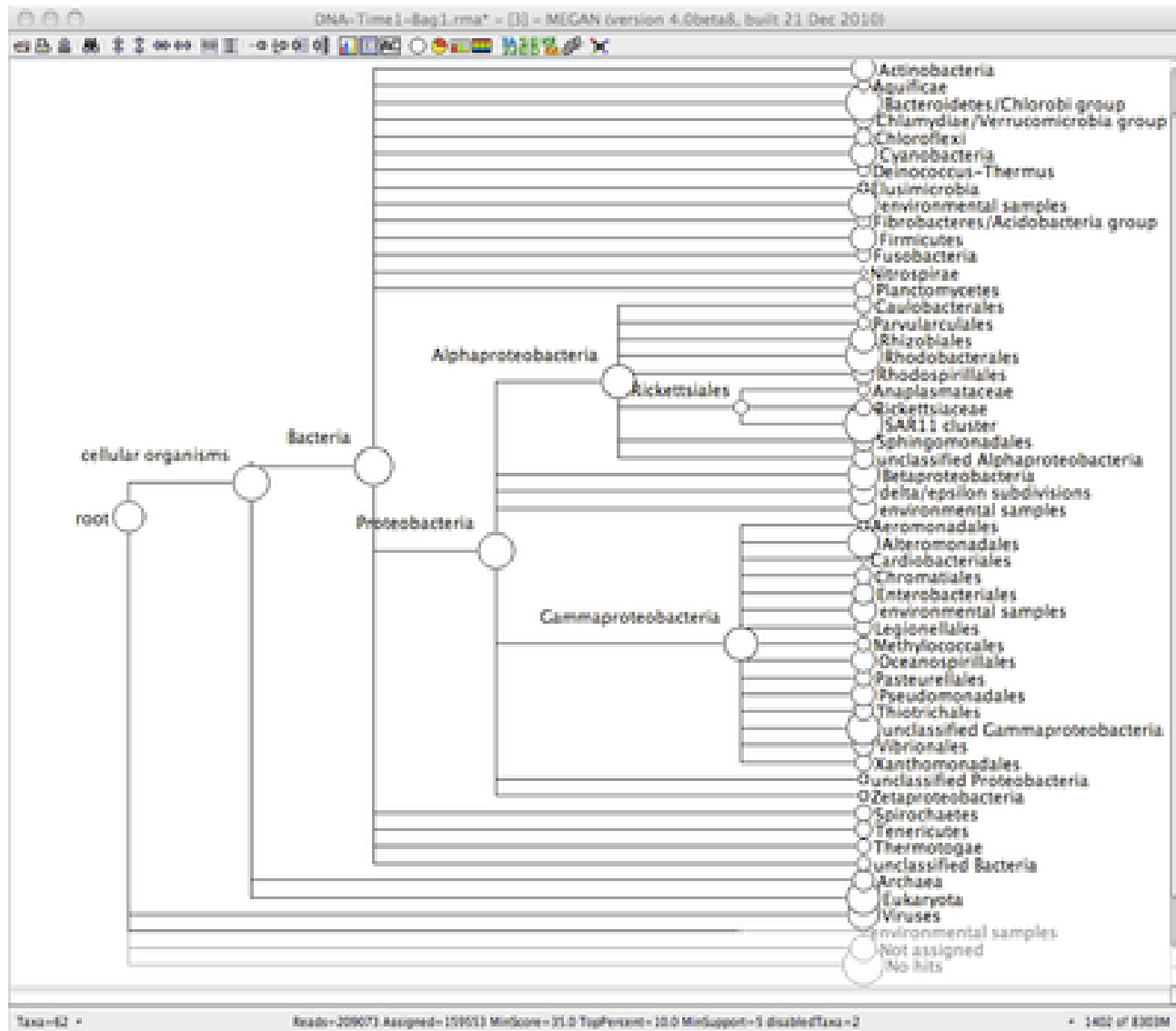
MEGAN 4 written by D. H. Huson, original design by D. H. Huson and S.C. Schuster, with contributions from S. Mitra, D.C. Richter, P. Rupek, H.-J. Ruscheweyh and N. Weber.

85 people like this. Sign Up to see what your friends like.

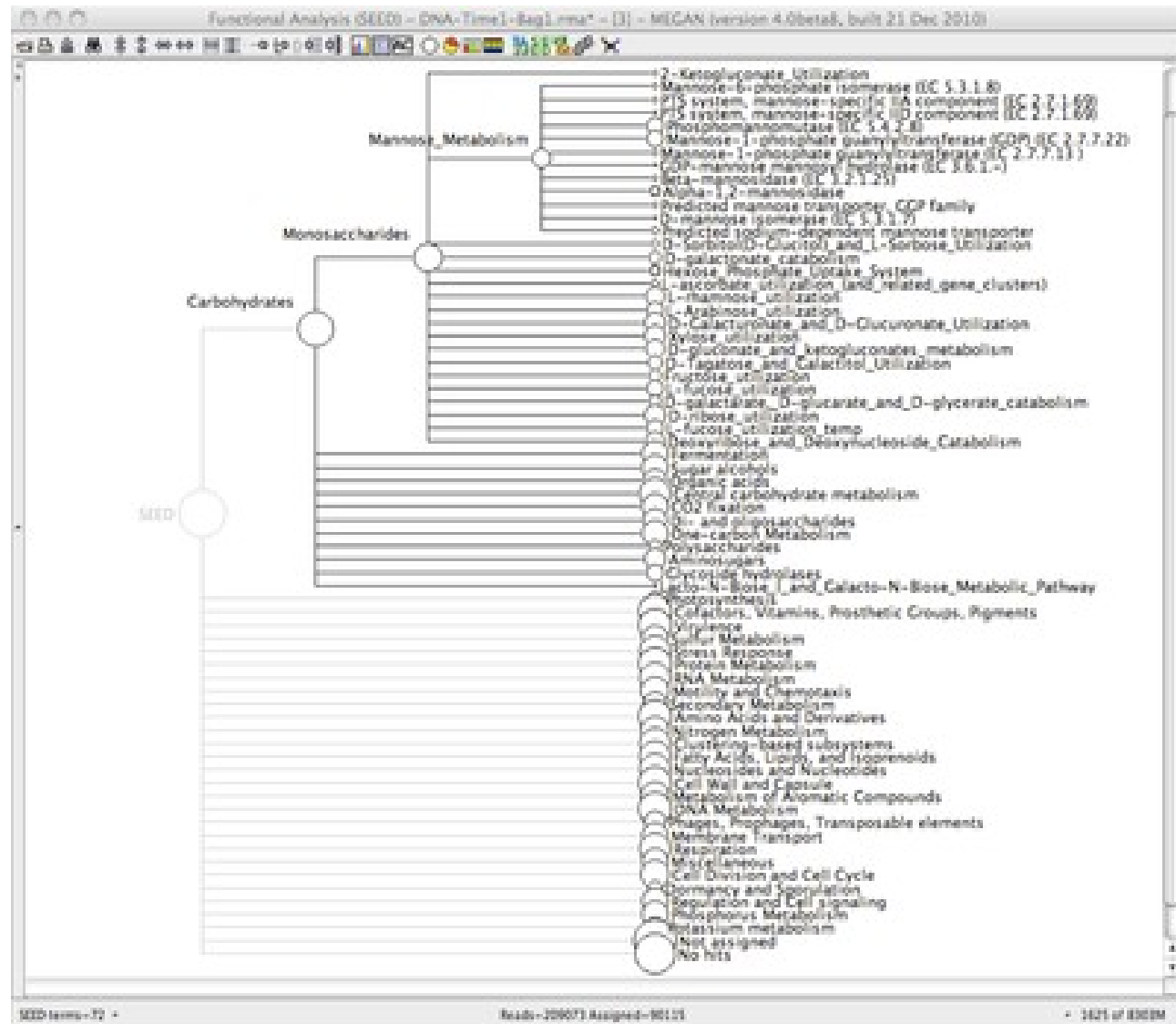
Introduction

In metagenomics, the aim is to understand the composition and operation of complex microbial consortia in environmental samples through sequencing and analysis of their DNA. Similarly, metatranscriptomics and metaproteomics target the RNA and proteins obtained from such samples. Technological advances in next-generation sequencing methods are fueling a rapid increase in the number and scope of environmental sequencing projects. In consequence, there is a dramatic increase in the volume of sequence data to be analyzed.

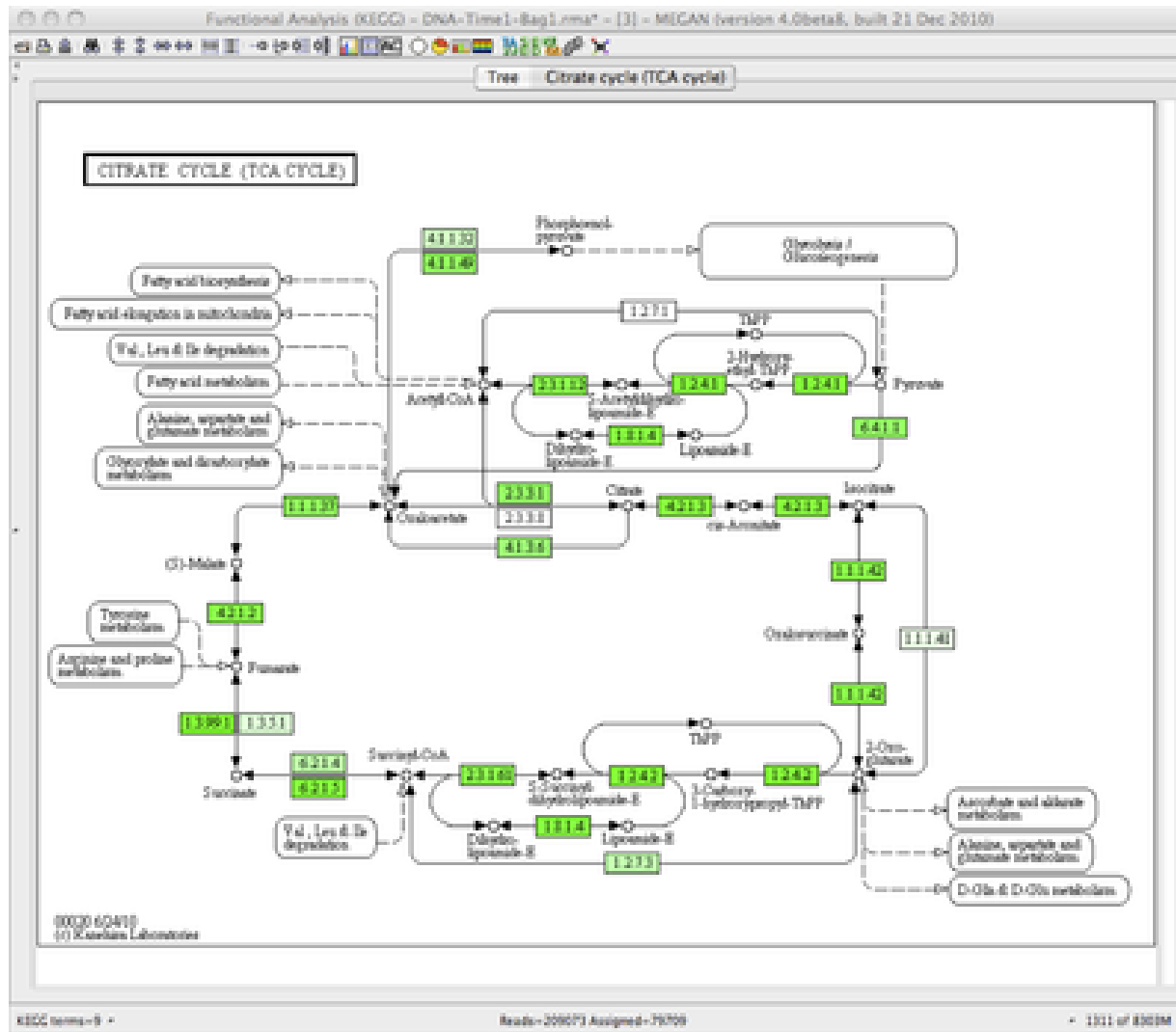
Taxonomická analýza



SEED analýza



KEGG analýza



EBI Metagenomic

<https://www.ebi.ac.uk/metagenomics/>

The screenshot shows the EBI Metagenomics website. At the top, there is a navigation bar with 'Services', 'Research', 'Training', and 'About us'. Below this is the EBI Metagenomics logo and a search bar. A secondary navigation bar includes 'Home', 'Submit data', 'Projects', 'Samples', 'Comparison tool', 'About', and 'Contact'. A large banner area contains the text 'Submit, analyse, visualize and compare your data.' with a prominent 'SUBMIT DATA' button. Below the banner, there are statistics for data sets, public and private runs, and samples. The 'Browse projects' section is divided into 'By selected biomes' and 'Latest projects'. The 'By selected biomes' section lists various biomes with their respective sample counts. The 'Latest projects' section features three featured projects with brief descriptions and links to view more details.

EMBL-EBI Services Research Training About us

EBI Metagenomics

Home Submit data Projects Samples Comparison tool About Contact Not logged in Login

Submit, analyse, visualize and compare your data.

SUBMIT DATA

31122 data sets

8706 metagenomes
884 metatranscriptomes
21463 amplicons
69 assemblies

28323 runs
18983 samples
205 projects

2799 runs
2711 samples
112 projects

Browse projects

By selected biomes

- Soil (34)
- Marine (37)
- Forest (9)
- Non-human host (51)
- Engineered (17)
- Freshwater (14)
- Grassland (6)
- Human gut (29)
- Air (1)
- Wastewater (2)

[View all biomes](#)

Latest projects 205

- Microbiome Quality Control Project - 16S**
The MBQC is a collaborative effort to comprehensively evaluate methods for measuring the human microbiome. This pilot phase includes 16S rDNA amplicon based surveys of the human microbiome. ...
[View more - 22 samples - compare](#)
- Human Microbiome Project (HMP) 16S rRNA Gene Diversity, the diversity of 16S ribosomal RNA genes in the human microbiome: 454 Protocol Validation - Mock**
This HMP Centers' Evaluation of the standard 454 SOP represents the pyrosequencing of 16S rRNA genes amplified from HMP even Mock community distributed to each of the four HMP sequencing ...
[View more - 1 sample - compare](#)
- The POU/Oct transcription factor Pdm1/nub is necessary for a beneficial gut microbiota and normal lifespan of Drosophila**
Maintenance of a stable gut microbial community relies on a delicate balance between immune defense and immune tolerance. We have used Drosophila to study how the microbial gut flora is ...
[View more - 18 samples - compare](#)

[View all projects](#)

Spotlight TARA ocean project **Tools** Functional sample comparison


<https://www.ebi.ac.uk/metagenomics/projects/SRP000319/samples/SRS000998/runs/SRR029687/results/versions/1.0>

EBI pipeline



Taxonomická analýza

EMBL-EBI [Services](#) [Research](#) [Training](#) [Industry](#) [About us](#)

 EBI Metagenomics

[Home](#) [Submit data](#) [Projects](#) **[Samples](#)** [About Metagenomics](#) [Contact](#) Not logged in [Login](#)





EBI Metagenomics > Project: Developing infant gut microbiome > Sample: 100 day old Infant gut microbiome

Sample (SRS086444)
100 day old Infant gut microbiome

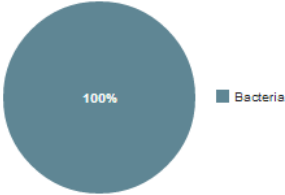
[Overview](#) [Quality control](#) **[Taxonomy analysis](#)** [Functional analysis](#) [Download](#)

These are the results from the taxonomic analysis steps of our pipeline. You can switch between different views of the data using the menu of icons below (pie, bar, stacked and interactive krona charts). If you wish to download the full set of results, all files are listed under the "Download" tab.

Top taxonomy Hits

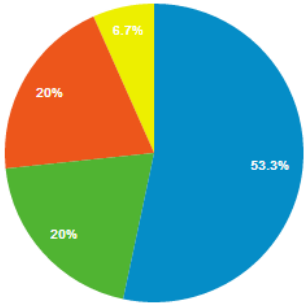
Switch view:    

Domain composition



100% Bacteria

Phylum composition (Total: 15 OTUs)

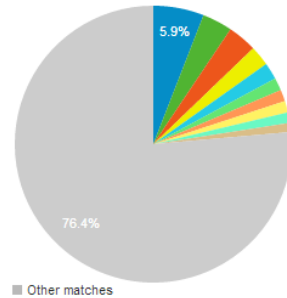


	Phylum	Domain	Unique OTUs	%
1	Actinobacteria	Bacteria	8	53.33
2	Firmicutes	Bacteria	3	20
3	Unassigned Bacteria	Bacteria	3	20
4	Proteobacteria	Bacteria	1	6.67

Funkční analýza

InterPro match summary

InterPro matches summary (Total: 781)



Filter table:

Entry name	ID	pCDS matched	%
NAD(P)-binding domain	IPR016040	184	5.91
Regulator of K+ conductance, N-terminal	IPR003148	109	3.5
ABC transporter-like	IPR003439	106	3.41
Potassium uptake protein TrkA	IPR006036	73	2.35
Glycoside hydrolase, family 3, N-terminal	IPR001764	61	1.96
Extracellular solute-binding protein, family 3	IPR001638	46	1.48
Winged helix-turn-helix transcription repressor DNA-binding	IPR011991	42	1.35
Aldolase-type TIM barrel	IPR013785	42	1.35
Lacto-N-biose phosphorylase	IPR012711	38	1.22
Serpin domain	IPR023796	33	1.06

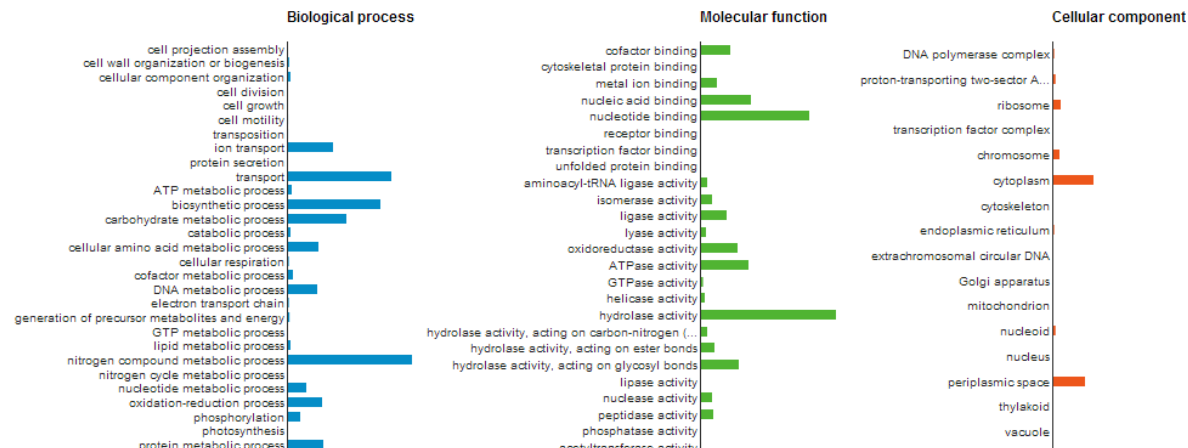
prev next

1 2 10 70 79

GO Terms annotation

A summary of Gene Ontology (GO) terms derived from InterPro matches to your sample is provided in the charts below.

Switch view:



MG - RAST

<http://metagenomics.anl.gov/>

The screenshot shows the MG-RAST website interface. At the top, the logo 'MG-RAST metagenomics analysis server' is displayed. Below the logo, a red banner contains a warning: 'fox browser. Since you are using Chrome, many features will not be available and / or behave incorrectly.' The main content area features a navigation bar with links for 'Browse Metagenomes', 'Register', 'Contact', 'Help', 'Upload', and 'News'. A search bar is also present. The 'About' section provides a description of the server and a table of statistics.

# of metagenomes	94,996
# base pairs	35.6 Tbp
# of sequences	313.12 billion
# of public metagenomes	13,609

The server primarily provides upload, quality control, automated annotation and analysis for prokaryotic metagenomic shotgun samples. MG-RAST was launched in 2007 and has over 8000 registered users and 94,996 data sets. The current server version is 3.3.7.3. We suggest users take a look at MG-RAST for the impatient. Also available for download is the technical manual.

- MG-RAST API Release of Version 1
- MG-RAST 3.3.6 release notes (API changes and new Search implementation) [July 2013]
- MG-RAST v3 tech-report and manual available
- MG-RAST 3.3 release notes [December 12, 2012]

* login required

This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900040C.

This work was supported in part by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11537.

[cite MG-RAST](#)

<http://metagenomics.anl.gov/metagenomics.cgi?page=MetagenomeOverview&metagenome=4447943.3>

Taxonomická analýza

Metagenome Analysis

1 Data Type

ORGANISM ABUNDANCE

- Representative Hit Classification
- »Best Hit Classification**
- Lowest Common Ancestor

FUNCTIONAL ABUNDANCE

- Hierarchical Classification
- All Annotations

OTHER

- Recruitment Plot

2 Data Selection

Metagenomes 4440283.3

Annotation Sources MSNR

Max. e-Value Cutoff 1e-5

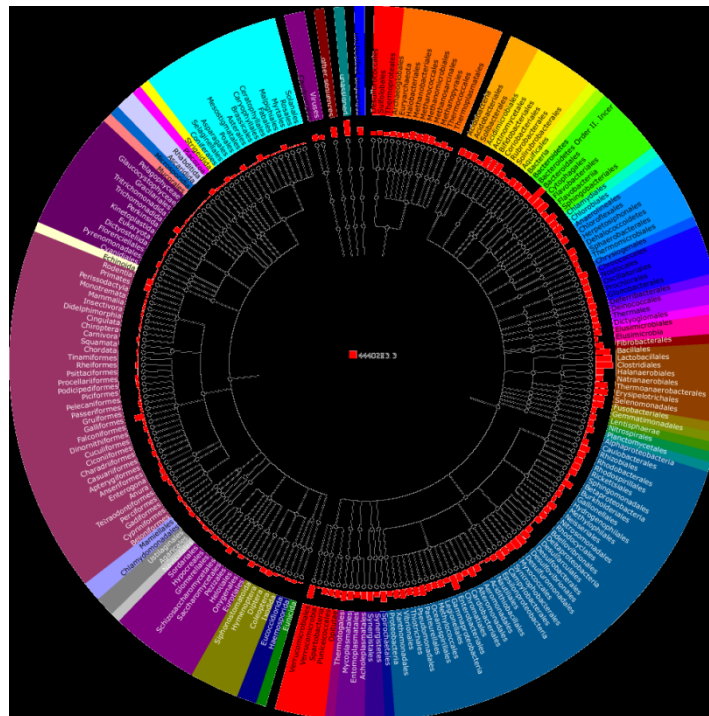
Min. % Identity Cutoff 60 %

Min. Alignment Length Cutoff 15

Workbench use features from workbench

3 Data Visualization

barchart tree table heatmap PCoA rarefaction



KEGG analýza

KeggMapper

DATA SELECTION

Target Buffer

Data B

Metagenomes

+

Max. e-Value Cutoff

1e-5

+

Min. % Identity Cutoff

60 %

+

Min. Alignment Length Cutoff

15

+

load data

DATA A



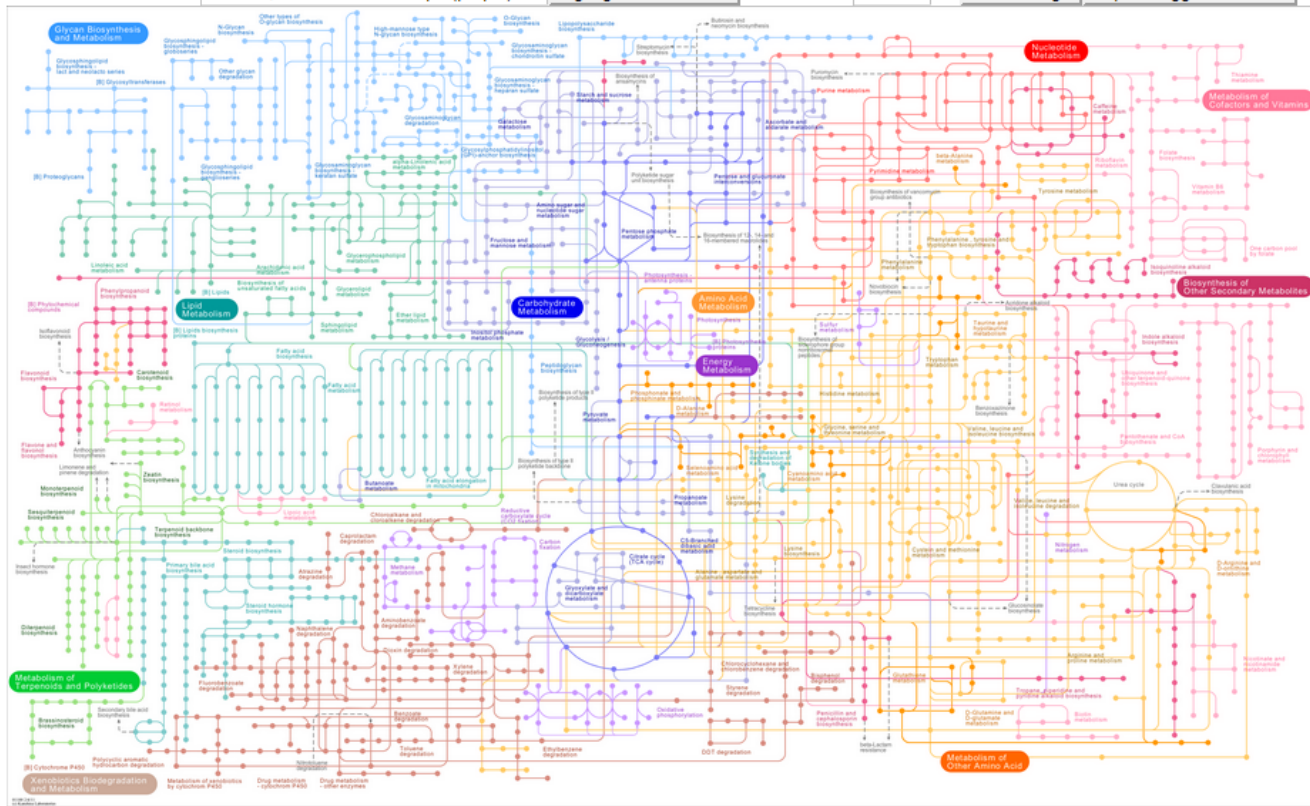
clear

DATA B



clear

Show unique data from Data A, Data B and overlaps (purple) highlight loaded data image size 25 % scale image export kegg abundance



- [http://metagenomics.anl.gov/mgmain.html?
mgpage=overview&metagenome=mgm47231](http://metagenomics.anl.gov/mgmain.html?mgpage=overview&metagenome=mgm47231)
10.3

Silva

- <http://www.arb-silva.de/>



Home SILVAngs Browser Search Aligner Download Documentation Projects FISH & Probes Shop Jobs Contact

SILVA

Welcome to the SILVA rRNA database project

A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.

SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (*Bacteria*, *Archaea* and *Eukarya*).

SILVA are the official databases of the software package ARB. For more background information → [Click here](#)

SILVA 126 - web release (Ref datasets and ARB files not updated)

	SSU Parc	LSU Parc
Release date	04.04.2016	04.04.2016
Aligned rRNA sequences	5,366,469	645,932

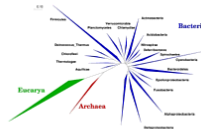
SILVAngs



Check out our new service for Next Generation Amplicon data

ARB

The software package ARB represents a graphically-oriented, fully-integrated package of cooperating software tools for handling and analysis of sequence information.



The ARB project has been started more than 15 years ago by Wolfgang Ludwig at the Technical University in Munich, Germany, see www.arb-home.de.

Citations

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl. Acids Res.* 41 (D1): D590-D596.

If you use SINA please cite:

Pruesse, E, Peplies, J and Glöckner, FO (2012) SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes



News

04.04.2016

SILVA 126 released as a web release

The SILVA webpage has been updated to represent the sequences of EMBL-EBI/ENA release 126. This release includes all taxonomic bugfixes done for SILVA release

123.1.

29.03.2016

SILVA 123.1 released

SILVA SSU 123.1 is an update of the SSU 123 full release providing corrections to the SILVA SSU taxonomy. The sequence data have not been changed.

02.03.2016

The 3rd de.NBI Quarterly Newsletter celebrates the 1st anniversary of the de.NBI network.

With this 3rd de.NBI Quarterly Newsletter (Edition February 2016) we are going to celebrate the 1st anniversary of the de.NBI network, which was officially started 1 March 2015.

During the last 12 months de.NBI was very active and able to reach nearly all the goals on the agenda for the first year. In particular, the Central Coordination Unit (CCU) as well as five Special Interest Groups (SIGs) have immediately been established. They structured the work carried out by the eight service centers.

10.02.2016

SILVAngs 1.5 released

The following improvements have been implemented: 1. Refactored Sequence Upload and 2. Request Project Execution

[go to Archive ->](#)

SILVA SSU / LSU 123 - full release

	SSU Parc	SSU Ref	SSU Ref NR	LSU Parc	LSU Ref
Minimal length	300	1200/900	1200/900	300	1900
Quality filtering	basic	strong	strong	basic	strong
Guide Tree	no	no	yes	no	yes
Release date	23.07.15	23.07.15	23.07.15	23.07.15	23.07.15
Aligned rRNA sequences	4,985,791	1,756,783	597,607	563,332	96,642

Differences between releases SSU 123 and 123.1

SILVA SSU 123.1 is an update of the SSU 123 full release providing corrections to the SILVA SSU taxonomy. The sequence data have not been changed. Updated ARB, FASTA, RAST and taxonomy files are provided in the [download](#) section. The changes in the taxonomy can be found [here](#).

The LSU 123 datasets have not been updated, therefore, the LSU 123

Srovnání databází

The ISME Journal (2011), 1–10
© 2011 International Society for Microbial Ecology. All rights reserved 1751-7362/11
www.nature.com/ismej



ORIGINAL ARTICLE

Impact of training sets on classification of high-throughput bacterial 16S rRNA gene surveys

Jeffrey J Werner^{1,7}, Omry Koren^{2,7}, Philip Hugenholtz³, Todd Z DeSantis⁴, William A Walters⁵, J Gregory Caporaso⁵, Largus T Angenent¹, Rob Knight^{5,6} and Ruth E Ley²

¹Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, USA; ²Department of Microbiology, Cornell University, Ithaca, NY, USA; ³Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD, Australia; ⁴Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; ⁵Department of Biochemistry and Chemistry, University of Colorado, Boulder, CO, USA and ⁶Howard Hughes Medical Institute, University of Colorado, Boulder, CO, USA

Taxonomic classification of the thousands–millions of 16S rRNA gene sequences generated in microbiome studies is often achieved using a naive Bayesian classifier (for example, the Ribosomal Database Project II (RDP) classifier), due to favorable trade-offs among automation, speed and accuracy. The resulting classification depends on the reference sequences and taxonomic hierarchy used to train the model; although the influence of primer sets and classification algorithms have been explored in detail, the influence of training set has not been characterized. We compared classification results obtained using three different publicly available databases as training sets, applied to five different bacterial 16S rRNA gene pyrosequencing data sets generated (from human body, mouse gut, python gut, soil and anaerobic digester samples). We observed numerous advantages to using the largest, most diverse training set available, that we constructed from the Greengenes (GG) bacterial/archaeal 16S rRNA gene sequence database and the latest GG taxonomy. Phylogenetic clusters of previously unclassified experimental sequences were identified with notable improvements (for example, 50% reduction in reads unclassified at the phylum level in mouse gut, soil and anaerobic digester samples), especially for phylotypes belonging to specific phyla (Tenericutes, Chloroflexi, Synergistetes and Candidate phyla TM6, TM7). Trimming the reference sequences to the primer region resulted in systematic improvements in classification depth, and greatest gains at higher confidence thresholds. Phylotypes unclassified at the genus level represented a greater proportion of the total community variation than classified operational taxonomic units in mouse gut and anaerobic digester samples, underscoring the need for greater diversity in existing reference databases.

The ISME Journal advance online publication, 30 June 2011; doi:10.1038/ismej.2011.82

Subject Category: integrated genomics and post-genomics approaches in microbial ecology

Keywords: Greengenes; microbiome; naive Bayesian classifier; pyrosequencing; taxonomy

Srovnání databází

Improving high-throughput taxonomic classification
JJ Wemer *et al*



5

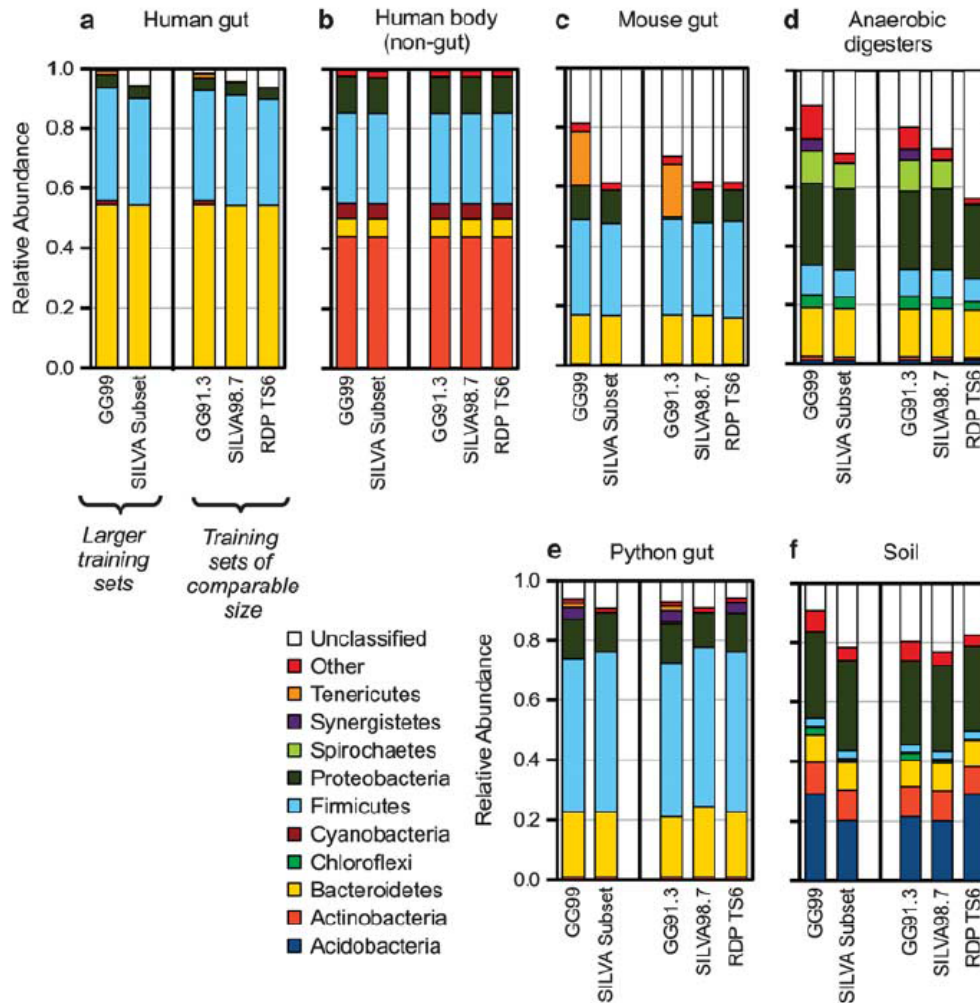


Figure 1 Relative abundance of the 10 major phyla identified by naïve Bayesian classification using five different training sets: three of approximately the same size: GG91.3, SILVA98.7, and RDP TS6, and two larger training sets: GG99 and the SILVA subset for Mothur. Relative abundances were averaged for samples of five different studies (note that human gut is shown apart from non-gut samples from the sample study): (a) human gut, (b) non-gut human body locations, (c) mouse gut, (d) anaerobic digester, (e) python gut and (f) soils.

Srovnání databází

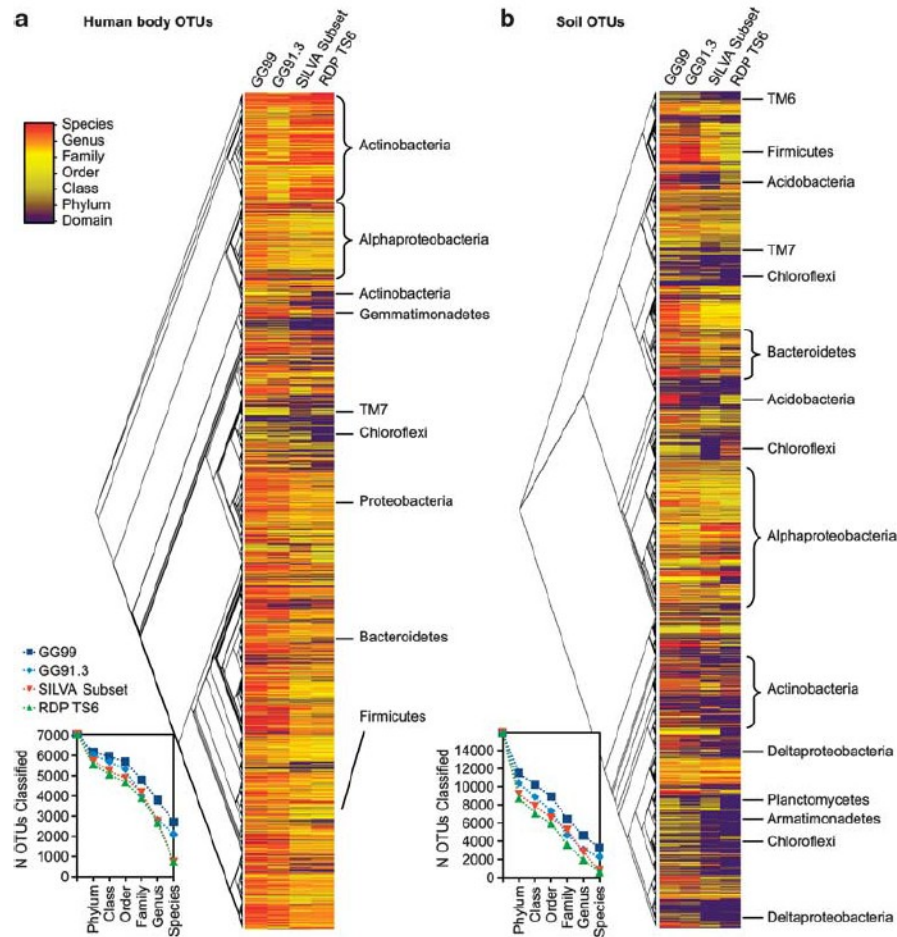


Figure 2 Summary of OTU classification depth using each of the three training sets for two of the four studies: (a) human body OTUs, and (b) soil OTUs (other three data sets shown in Supplementary Figure S6). OTUs are organized according to evolutionary history, as determined by the FastTree approximately-maximum-likelihood tree constructed in the default QIIME pipeline. Inset charts summarize the total number of OTUs classified at each taxonomic level (GG99 = dark blue, GG91.3 = light blue, SILVA = green, RDP TS6 = orange).

Metagenomic Surveys of Gut Microbiota

Rahul Shubhra Mandal ^{1,a}, Sudipto Saha ^{2,a,b}, Santasabuj Das ^{1,3,a,c}¹ Biomedical Informatics Centre, National Institute of Cholera and Enteric Diseases, Kolkata 700010, India² Bioinformatics Centre, Bose Institute, Kolkata 700054, India³ Division of Clinical Medicine, National Institute of Cholera and Enteric Diseases, Kolkata 700010, India

Received 8 July 2014; revised 10 February 2015; accepted 26 February 2015

Available online 13 July 2015

Handled by Fangqing Zhao

Další vyhodnocovací programy

Table 3 Tools/webservers related to gut microbiota studies

Name	Platform	Website	Main features	Ref.
QIIME	Stand alone	http://qiime.sourceforge.net/	Network analysis, histograms of within- or between-sample diversity	[82]
mothur	Stand alone	http://www.mothur.org/	Fast processing of large sequence data	[83]
RAMMCAP	Stand alone	http://weizhonglab.ucsd.edu/rammcap/cgi-bin/rammcap.cgi	Ultra fast sequence clustering and protein family annotation	[84]
MEGAN	Stand alone	http://www-ab.informatik.unituebingen.de/software/megan/	Laptop analysis of large metagenomic shotgun sequencing data sets	[85]
MetaPhlan	Stand alone	http://huttenhower.sph.harvard.edu/metaphlan	Faster profiling of the composition of microbial communities using unique clade-specific marker genes	[86]
MetaVelvet	Stand alone	http://metavelvet.dna.bio.keio.ac.jp/	High quality metagenomic assembler	[87]
SOAPdenovo2	Stand alone	http://soap.genomics.org.cn/soapdenovo.html	Metagenomic assembler, specifically for Illumina GA short reads	[88]
MOCAT	Stand alone	http://vmlux.embl.de/~kultima/MOCAT/	Generate taxonomic profiles and assemble metagenomes	[89]
SmashCommunity	Stand alone	http://www.bork.embl.de/software/smash/	Performs assembly and gene prediction mainly for data from Sanger and 454 sequencing technologies	[90]
HUMAnN	Stand alone	http://huttenhower.sph.harvard.edu/humann	Analysis of metagenomic shotgun data from the Human Microbiome Project	[91]
FANTOM	Stand alone	http://www.sysbio.se/Fantom/	Comparative analysis of metagenomics abundance data integrated with databases like KEGG Orthology, COG, PFAM and TIGRFAM, etc.	[92]
MetaCV	Stand alone	http://metacv.sourceforge.net/	Classification short metagenomic reads (75–100 bp) into specific taxonomic	[94]
Phymm	Stand alone	http://www.cbcb.umd.edu/software/phymm/	Phylogenetic classification of metagenomic short reads using interpolated Markov models	[97]
PhyloPythiaS	Web server	http://binning.bioinf.mpiinf.mpg.de/	Fast and accurate sequence composition-based classifier that utilizes the hierarchical relationships between clades	[96]
TETRA	Web server	http://www.megx.net/tetra	Correlation of tetranucleotide usage patterns in DNA	[93]
METAREP	Web server	http://www.jcvi.org/metarep/	Flexible comparative metagenomics framework	[98]
CD-HIT	Web server	http://weizhonglab.ucsd.edu/cd-hit/	Identity-based clustering of sequences	[99]
METAGENassist	Web server	http://www.metagenassist.ca/	Performs comprehensive multivariate statistical analyses on the data from different host and environment sites	[100]
CoMet	Web server	http://comet.gobics.de/	ORF finding and subsequent Pfam domain assignment to protein sequences	[101]
WebCARMA	Web server	http://webcarma.cebitec.unibielefeld.de/	Unassembled reads as short as 35 bp can be used for the taxonomic classification with less false positive prediction	[102]
MG-RAST	Web server	https://metagenomics.anl.gov/	High-throughput pipeline for functional metagenomic analysis	[103]
CAMERA	Web server	https://portal.camera.calit2.net/gridsphere/gridsphere	Provides list of workflows for WGS data analysis	[104]
WebMGA	Web server	http://weizhonglilab.org/metagenomic-analysis/	Implemented to run in parallel on local computer cluster	[105]

PICRUSt

Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences

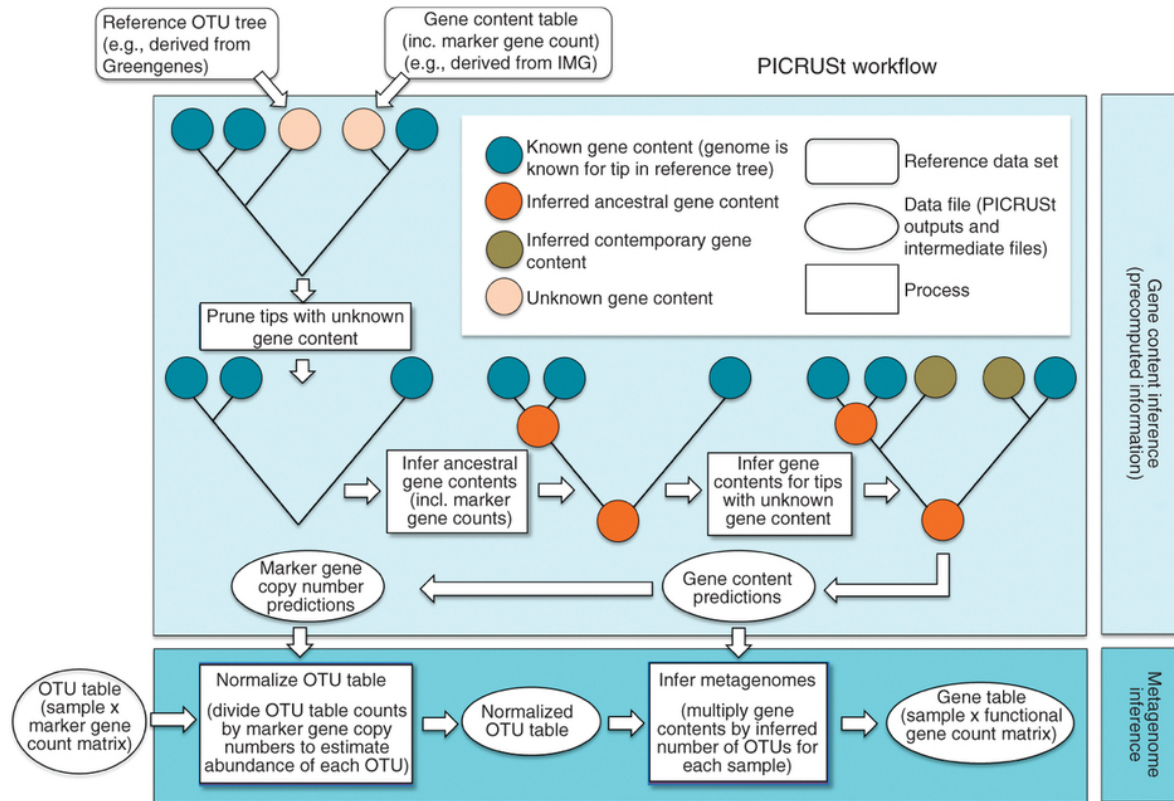
Morgan G I Langille^{1,14}, Jesse Zaneveld^{2,14}, J Gregory Caporaso^{3,4}, Daniel McDonald^{5,6}, Dan Knights^{7,8}, Joshua A Reyes⁹, Jose C Clemente¹⁰, Deron E Burkepile¹¹, Rebecca L Vega Thurber², Rob Knight^{10,12}, Robert G Beiko¹ & Curtis Huttenhower^{9,13}

Profiling phylogenetic marker genes, such as the 16S rRNA gene, is a key tool for studies of microbial communities but does not provide direct evidence of a community's functional capabilities. Here we describe PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states), a computational approach to predict the functional composition of a metagenome using marker gene data and a database of reference genomes. PICRUSt uses an extended ancestral-state reconstruction algorithm to predict which gene families are present and then combines gene families to estimate the composite metagenome. Using 16S information, PICRUSt recaptures key findings from the Human Microbiome Project and accurately predicts the abundance of gene families in host-associated and environmental communities, with quantifiable uncertainty. Our results demonstrate that phylogeny and function are sufficiently linked that this 'predictive metagenomic' approach should provide useful insights into the thousands of uncultivated microbial communities for which only marker gene surveys are currently available.

phylogenetic diversity of environmental samples. Because marker gene studies focus on one or a few universal genes, they cannot directly identify metabolic or other functional capabilities of the microorganisms under study. Conversely, metagenomic sequencing aims to sample all genes from a community and can produce detailed metabolic and functional profiles. Although relatively little sequencing is needed to characterize the diversity of a sample^{3,4}, deep, and therefore costly, metagenomic sequencing is required to access rare organisms and genes⁵. Thus, marker gene profiling of large sample collections is now routine, but deep metagenomic sequencing across many samples is prohibitively expensive.

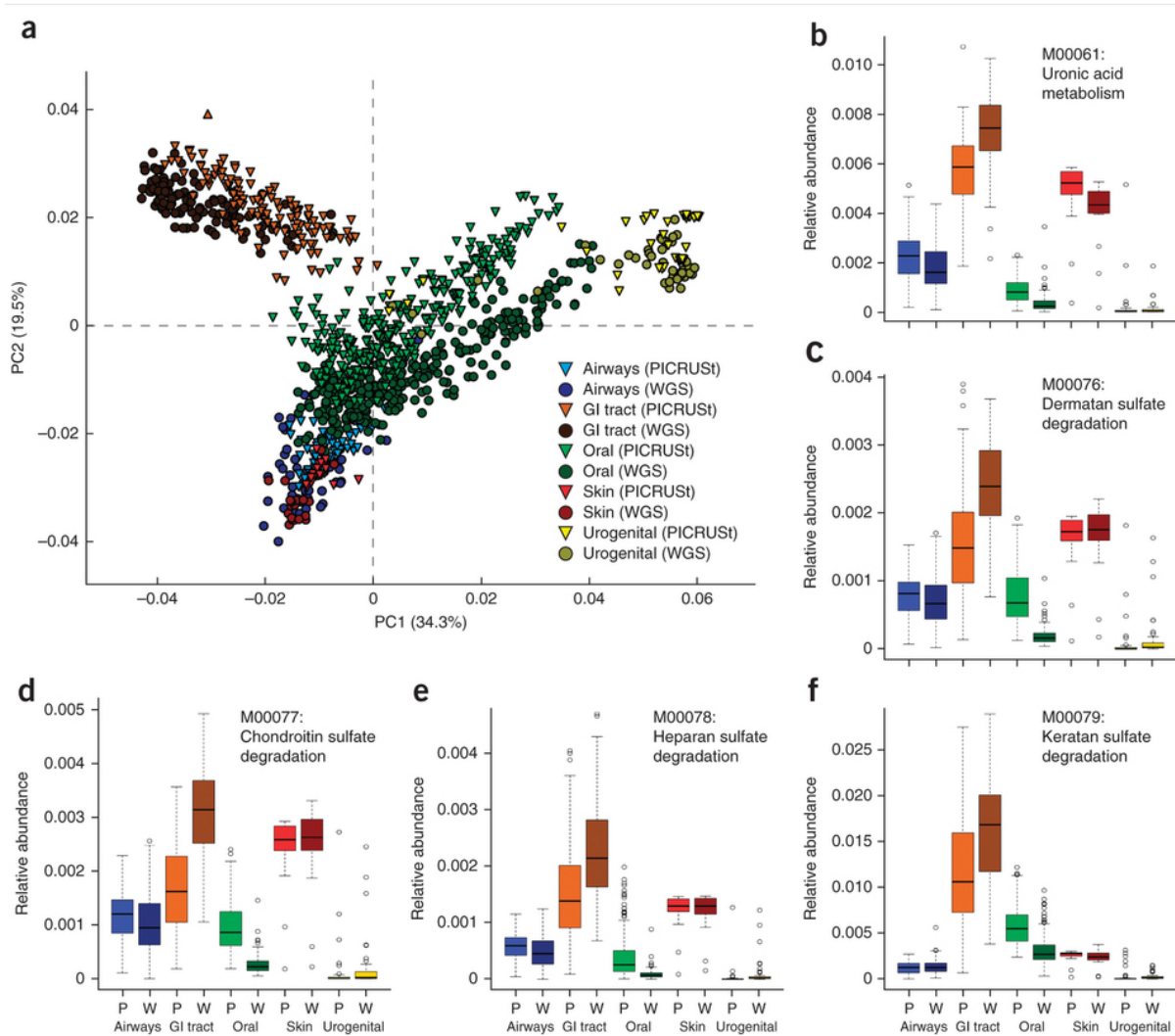
Although marker gene and shotgun sequencing strategies differ in the type of information produced, phylogeny and biomolecular function are strongly, if imperfectly, correlated. Phylogenetic trees based on 16S closely resemble clusters obtained on the basis of shared gene content⁶⁻⁹, and researchers often infer properties of uncultured organisms from cultured relatives. For example, the genome of a *Bacteroides* spp. might reasonably be inferred to contain many genes encoding glycoside hydrolase activity, based on the commonality of these activities in sequenced *Bacteroides* isolates¹⁰. This associa-

PICRUSt



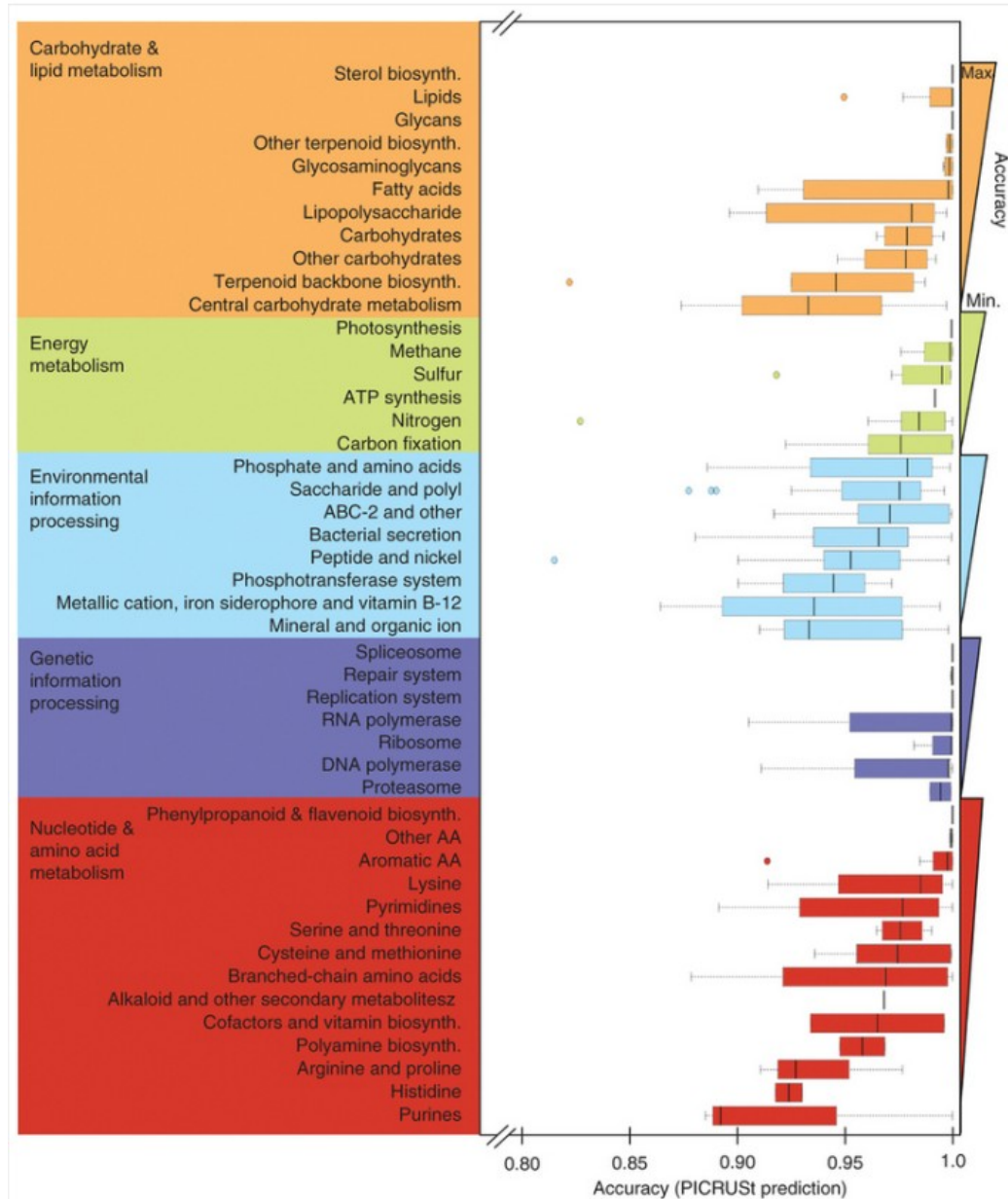
PICRUSt is composed of two high-level workflows: gene content inference (top box) and metagenome inference (bottom box). Beginning with a reference OTU tree and a gene content table (i.e., counts of genes for reference OTUs with known gene content), the gene content inference workflow predicts gene content for each OTU with unknown gene content, including predictions of marker gene copy number. This information is precomputed for 16S based on Greengenes²⁹ and IMG²⁶, but all functionality is accessible in PICRUSt for use with other marker genes and reference genomes. The metagenome inference workflow takes an OTU table (i.e., counts of OTUs on a per sample basis), where OTU identifiers correspond to tips in the reference OTU tree, as well as the copy number of the marker gene in each OTU and the gene content of each OTU (as generated by the gene content inference workflow), and outputs a metagenome table (i.e., counts of gene families on a per-sample basis).

PICRUST



(a) Principal component analysis (PCA) plot comparing KEGG module predictions using 16S data with PICRUST (lighter colored triangles) and sequenced shotgun metagenome (darker colored circles) along with relative abundances for five specific KEGG modules: (b) M00061: Uronic acid metabolism. (c) M00076: Dermatane sulfate degradation. (d) M00077: Chondroitin sulfate degradation. (e) M00078: Heparan sulfate degradation. (f) M00079: Keratan sulfate degradation. All KEGG modules are involved in glycosaminoglycan degradation (KEGG pathway ko00531) using 16S with PICRUST (P) and whole genome sequencing (W) across human body sites. Color key is the same as in a.

PICRUSt



Results are colored by functional category and sorted in decreasing order of accuracy within each category (indicated by triangular bars, right margin). Note that accuracy was >0.80 for all, and therefore the region 0.80–1.0 is displayed for clearer visualization of differences between modules.

PICRUSt

<http://picrust.github.io/picrust/>

[PICRUSt 1.0.0-dev documentation](#) »

[index](#)

PICRUSt: Phylogenetic Investigation of Communities by Reconstruction of Unobserved States

PICRUSt (pronounced "pie crust") is a bioinformatics software package designed to predict metagenome functional content from marker gene (e.g., 16S rRNA) surveys and full genomes.

PICRUSt is freely available under the [GPL](#).

Using PICRUSt

If you're new to PICRUSt, you'll want to work through these documents in order:

1. [Installing PICRUSt](#) OR [Use online Galaxy version](#).
2. [Quickstart Guide](#)
3. [Metagenome Prediction Tutorial](#)
4. [Analyzing PICRUSt predicted metagenomes](#)
5. [Quality Control of PICRUSt Predictions](#)

More advanced users may be interested in the following (in no particular order):

- [How PICRUSt Works](#)
- [Genome Prediction Tutorial](#)
- [Installing PICRUSt in Galaxy](#)
- [PICRUSt Script Index](#)
- [PICRUSt Methods](#)

Contact

For PICRUSt announcements and questions, including notification of new releases, you should subscribe to the [PICRUSt users list](#).

Citing PICRUSt

On this page we've compiled both the PICRUSt citation and links to several tools that PICRUSt is built on to make citing these various software packages easier.

Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Langille, M. G. I.*; Zaneveld, J.*; Caporaso, J. G.; McDonald, D.; Knights, D.; Reyes, J.; Clemente, J. C.; Burkepile, D. E.; Vega Thurber, R. L.; Knight, R.; Beiko, R. G.; and Huttenhower, C. *Nature Biotechnology*, 1-10. 8 2013.

The manuscript describing PICRUSt can be found [here](#)

- Additional citation resources: [here](#)

News & Announcements

- 09-03-13: Official release of PICRUSt 1.0.0
- 08-25-13: PICRUSt published in [Nature Biotechnology](#)
- 07-10-13: PICRUSt presented by Morgan Langille at [Gordon Research Conference: Applied & Environmental Microbiology](#)
- More [News & Announcements](#)

[PICRUSt 1.0.0-dev documentation](#) »

[index](#)

Table Of Contents

PICRUSt: Phylogenetic Investigation of Communities by Reconstruction of Unobserved States

- Using PICRUSt
- Contact
- Citing PICRUSt
- News & Announcements

This Page

Show Source

Quick search

Enter search terms or a module, class or function name.

PICRUST

http://picrust.github.io/picrust/tutorials/qiime_tutorial.html#qiime-tutorial

Analyzing metagenomes with QIIME

Because the metagenomes are provided in BIOM format by default, these can be plugged into many of the downstream analysis tools available in QIIME. QIIME's [Shotgun Metagenome Analysis tutorial](#) illustrates a couple of the steps that can be applied. The steps that will primarily be of interest in that tutorial are the ones that begin with a `.biom` file. For example, [computing beta diversity and PCoA plots](#) and [generating summaries of samples by KO categories](#).

Many of [QIIME's tutorials that describe diversity analyses](#) are applicable to PICRUST-predicted metagenome tables. Specific analysis tools that may be useful include:

- [alpha_diversity.py](#)
- [beta_diversity.py](#)
- [compute_core_microbiome.py](#)
- [jackknifed_beta_diversity.py](#)
- [make_distance_boxplots.py](#)
- [alpha_rarefaction.py](#)
- [beta_diversity_through_plots.py](#)
- [group_significance.py](#)
- [shared_phylotypes.py](#)

Plots of functional categories at various levels can be created using [summarize_taxa_through_plots.py](#)

- Since KEGG Orthologs belong to several pathways you should collapse your PICRUST predictions to the desired hierarchy level using [categorize_by_function.py](#)

```
categorize_by_function.py -i metagenome_predictions.biom -c "KEGG_Pathways" -l 2 -o metagenome_at_level2.biom
```

- Then add the following lines to a [qiime parameter file](#) (e.g. `qiime_params.txt`) ensuring that the level you collapsed at is the same in your config file

```
summarize_taxa:md_identifier "KEGG_Pathways"  
summarize_taxa:absolute_abundance True  
summarize_taxa:level 2
```

- Lastly, run [summarize_taxa_through_plots.py](#)

```
summarize_taxa_through_plots.py -i metagenome_at_level2.biom -p qiime_params.txt -o plots_at_level2
```

There are also a number of scripts in QIIME that may be useful for more general processing of your BIOM table. These include the following:

- [single_rarefaction.py](#)
- [filter_otus_from_otu_table.py](#)
- [filter_samples_from_otu_table.py](#)
- [per_library_stats.py](#)
- [filter_taxa_from_otu_table.py](#)
- [merge_otu_tables.py](#)
- [sort_otu_table.py](#)
- [split_otu_table.py](#)
- [split_otu_table_by_taxonomy.py](#)

Note that while many of these refer to OTU table, it's just a nomenclature issue. These are generally applicable to `.biom` tables.

Finally, if you're interested in comparing real to predicted metagenomes, or predicted metagenomes to 16S, you'll be interested in the [Procrustes Analysis tutorial](#) and the [Comparing Distance Matrices tutorial](#).

PRMT

Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset

Peter E. Larsen^{1*}, Frank R. Collart¹, Dawn Field², Folker Meyer¹, Kevin P. Keegan¹, Christopher S. Henry^{1,4}, John McGrath³, John Quinn³, Jack A. Gilbert^{1,5}

¹ Argonne National Laboratory, 9700, S. Cass Ave, Argonne, Illinois, USA

² NERC Centre for Ecology and Hydrology, CEH Oxford, Mansfield Road, Oxford, OX1 3SR, UK

³ School of Biological Science, Queens University, Medical Biology Centre, 97 Lisburn Road, Belfast, BT9 7BL, Northern Ireland, UK

⁴ Computation Institute, University of Chicago, 5735 South Ellis Avenue, Chicago, IL 60637, U.S.A

⁵ Department of Ecology and Evolution, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, U.S.A

* Corresponding author

Email PEL: plarsen@anl.gov
FRC: fcollart@anl.gov
DF: dfield@ceh.ac.uk
FM: folker@mcs.anl.gov
KPK: kkeegan@anl.gov
CSH: chrisshenry@gmail.com
JM: j.mcgrath@qub.ac.uk
JQ: j.quinn@qub.ac.uk
JAG: gilbertjack@gmail.com

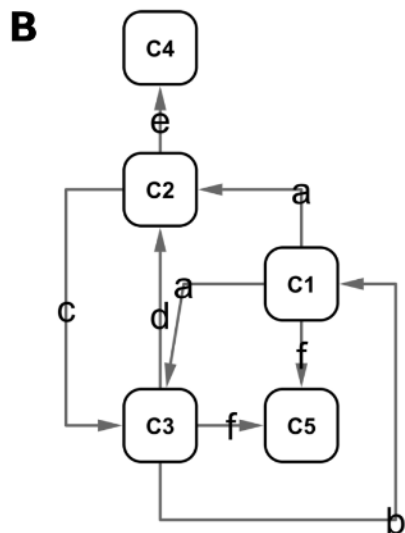
PRMT

PRMT analysis approach

PRMT scores predict the change in turnover of metabolites (defined as the potential for consumption or production) in an environmental metabolome, given the relative abundance of genes for unique enzyme functions detected in different metagenomes. In this manuscript, we use the term “unique enzyme function” to describe a specific annotation applied to an enzyme, i.e. “Phosphotransferases with an alcohol group as acceptor”. We use “enzyme reactions” to refer to metabolite transformations catalyzed by an enzyme function, i.e. “ATP + D-Glycerate \leftrightarrow ADP + 3-Phospho-D-glycerate”. A unique enzyme function may catalyze more than one enzyme reaction and an enzyme reaction may be catalyzed by more than one unique enzyme function. A metabolite is a molecular compound that is a reactant or product in an enzyme reaction. In PRMT, a metabolite is never the protein product of a gene in the metagenome.

PRMT

A $C1 \xrightarrow{a} C2$
 $C1 \xrightarrow{a} C3$
 $C3 \xrightarrow{b} C1$
 $C2 \xrightarrow{c} C3$
 $C3 \xrightarrow{d} C2$
 $C2 \xrightarrow{e} C4$
 $C1 \xrightarrow{f} C5$
 $C3 \xrightarrow{f} C5$



C

	v_a	v_b	v_c	v_d	v_e	v_f
C1	-2	1	0	0	0	-1
C2	1	0	-1	1	-1	0
C3	1	-1	1	-1	0	-1
C4	0	0	0	0	1	0
C5	0	0	0	0	0	2

D

	v_a	v_b	v_c	v_d	v_e	v_f
C1	-0.66	1	0	0	0	-0.33
C2	0.5	0	-0.5	0.5	-0.5	0
C3	0.5	-0.33	0.5	-0.33	0	-0.33
C4	0	0	0	0	1	0
C5	0	0	0	0	0	1

Figure 1

Example of generating an EMM from metagenomic data. This figure is an example of generating a simple EMM with hypothetical data. Letters a-f represent unique enzyme functions identified in the annotation of a hypothetical set of metagenomes. In (A), the set of all enzyme reactions for enzyme functions a-f between compounds C1-C5 from a database of possible reactions is listed. In (B), a metabolome is constructed from the reactions identified in A. (C) Shows the connectivity matrix of the network in B. (D) Is the complete EMM for metagenomic annotated enzyme functions a-f, normalizing values in C such that the sum of all inputs to a compound is 1 and the sum of all outputs from a compound is -1.

PRMT

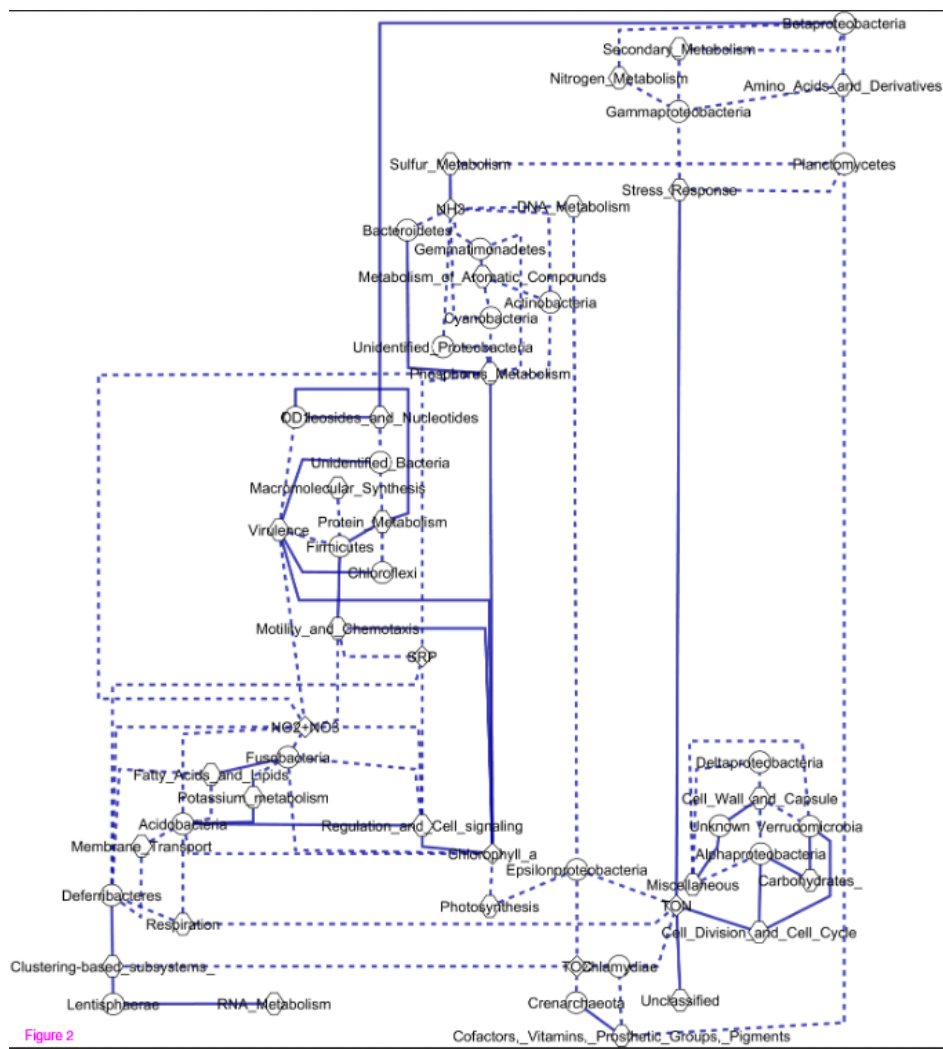


Figure 2

Strong correlations between environmental metabolites, metabolic subsystems, and bacterial population structure. This network is a graphical representation of strong (i.e. in the top or bottom 5th percentile of randomized resamples) correlations between relative abundance of measured environmental metabolites (diamonds), relative abundance of metagenomic reads annotated to metagenomic SEED subsystems (hexagons), and relative abundance of bacterial taxa (circles) across seasonal variation for the Western English Channel L4 station. Strong positive correlations are represented by solid lines and strong negative correlations by dashed lines.

PRMT

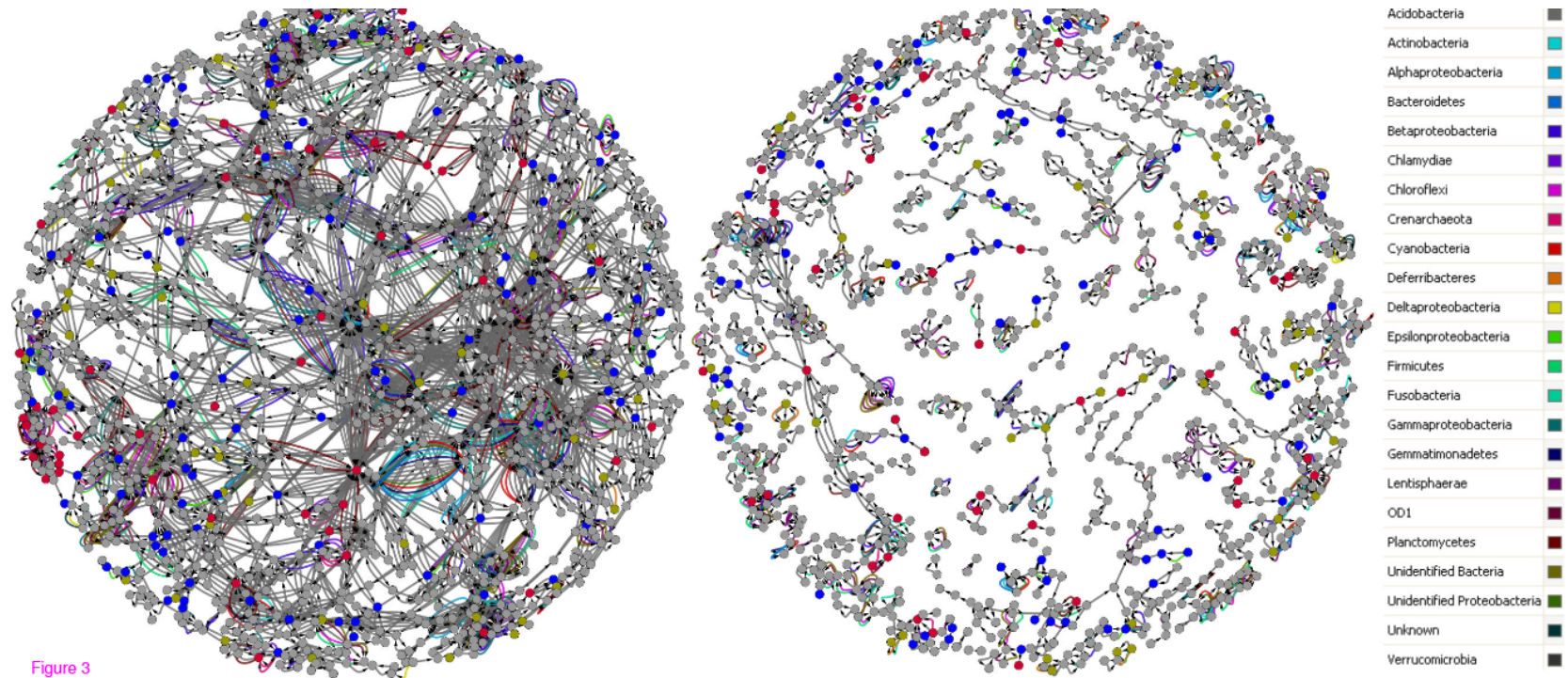


Figure 3

L4 Environmental Metabolome

In the figure, edges represent enzyme functions identified in annotated metagenomes. Nodes are predicted metabolites, inferred by the reactions catalyzed by detected enzyme functions. Nodes are highlighted if calculated PRMT scores for seasonal metagenomes correlate strongly (i.e. in the top or bottom 5th percentile of randomized resamples) with relative abundance of measured environmental parameters (Red for Total Organic Carbon, blue for Total Organic Nitrogen, and gold for Soluble Reactive Phosphorus). Edges are highlighted in one of 23 colors if they connect nodes that correlate with relative abundance of a bacterial phylum. Figure was generated using Cytoscape v2.6.1. The network and calculated PRMT-scores in this figure is available for download as additional file 3, figure S1.