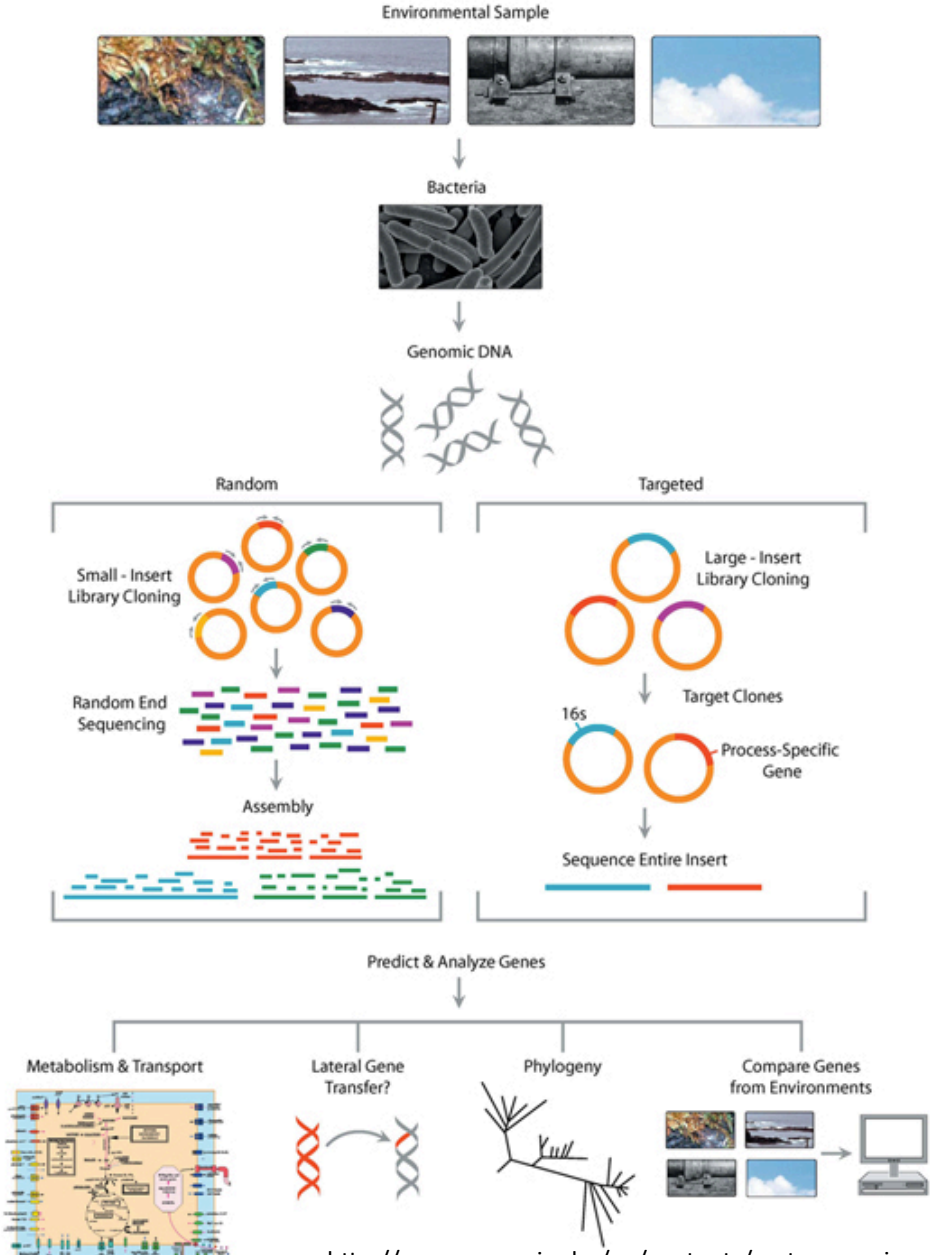
BMC Bioinformatics

RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets

Matthias Scheuch, Dirk Höper and Martin Beer

BMC Bioinformatics 2015, 16:69

Metagenomics



http://www.seqomics.hu/en/contents/metagenomics

Amplicon Sequencing – focus on specific marker genes for phylogenetic identification only (16S rRNA). Further markers may be other rRNA genes, the ITS region, RuBisCo, mcrA or other functional genes.

Metagenome Sequencing – offers a random representation of all extracted genomic sequences, giving insights into the metabolic profiles of these specific communities.

Application areas:

- Health Industry, metagenomic patterns as diagnostic markers (e.g. gut metagenome, skin surface, stool, invaded tissues, etc.)
- Food industry, food safety
- Agriculture (soil, rhizosphere communities, plant and animal-associated surface and tissue samples)
- Environmental applications, bioremediations (soil, air, water samples)
- Wastewater treatment applications
- Bioenergy (e.g. biogas consortia, anaerobic and aerobic degradation communities)

Current methods

- Web applications: MG-RAST, EBI metagenomics service
- 16S rRNA-based methods: DOTUR, SONS, LIBSHUFF, UniFrac
- Combination of taxonomic and phylogenetic data: Qiime, PhyloPythia
- Heuristic similarity-based speedup with UCLUST and USEARCH: MetaGeniE
- Require prior knowledge of user input reference sequences: SUPRI, Readscan, RINS
- Custom validated datasets: Clinical PathoScope, Kraken, MetaPhlAn

Intrigue

- Many tools are limited to 16S rRNA sequences allowing to analyze only prokaryotic sequences or skip viruses for example
- Limitation of all taxonomic and functional binning tools is the analysis duration due to the linear correlation between analysis time and data size
- Web-applications not suitable for confidential diagnostic data
- Some tools are centered on human sample analysis
- Generating a comprehensive and clear result protocol is not always the case

Aim

- Create a tool that:
  - ✓ requires information concerning the sample origin
  - ✓ requires no specific read classification (host, pathogen)
  - ✓ uses unbiased taxonomic classification based on the most similar sequence using full content of INSDC databases (DDBJ, EMBL ENA, NCBI GenBank, PDB)
  - ✓ is able to switch to amino acid sequences if DNA analysis is not fully successful
  - ✓ creates tabular summary of the classifications for sorted reads and additional information regarding the different analyses that were carried out
- RIEMS (Reliable Information Extractrion from Metagenomic Sequence datasets) pipeline

Workflow

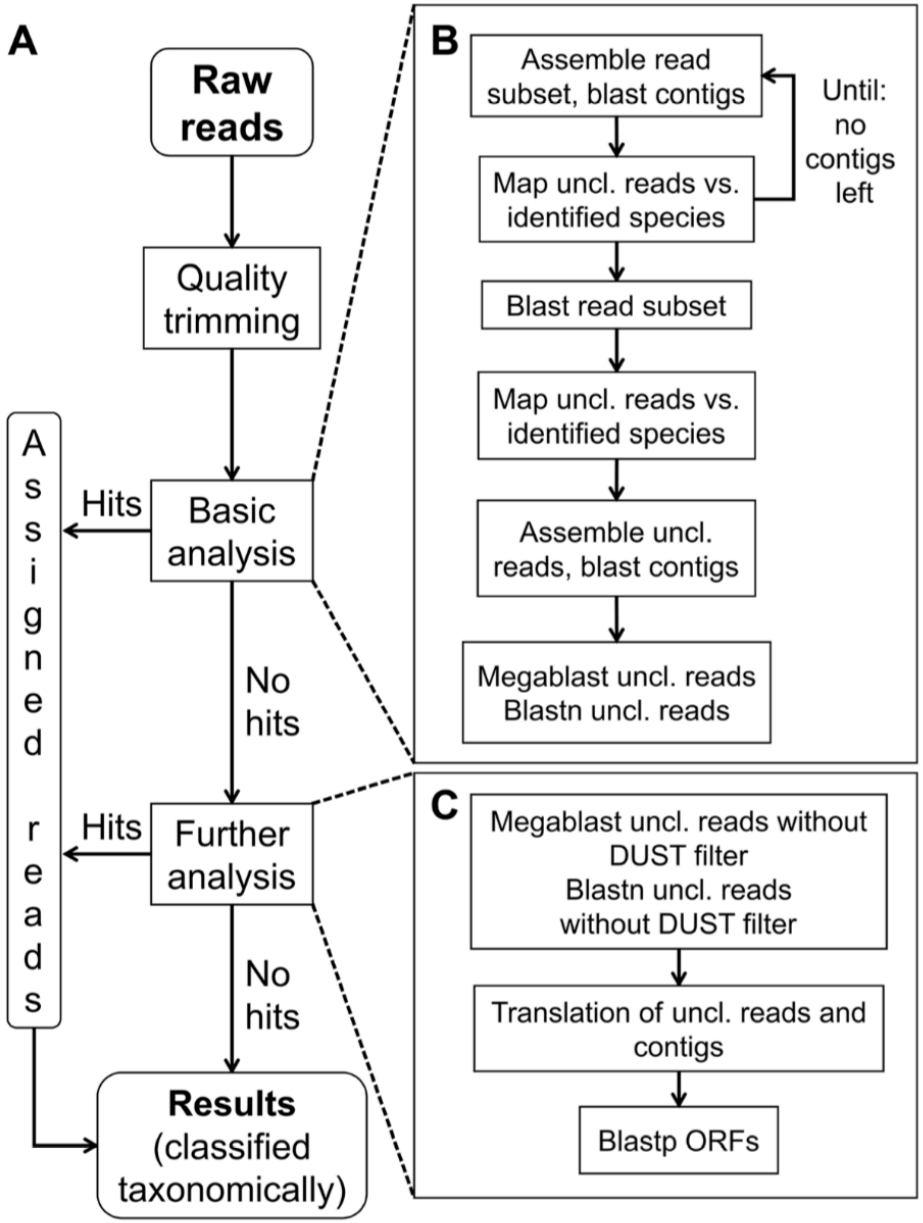


Figure 6 Flow diagram of RIEMS. (A) Main steps of RIEMS analysis. **(B)** Succession of analyses within the 'Basic analysis' **(C)** Principal steps of the 'Further analysis'. For details see text.

Used input data for validation

- Two genuine 454 sequencing datasets (13,816 and 247,833 reads)
- ~90,000 simulated reads (with length distributions equal to those from 454 pyrosequencing) from overlapping genome fragments of 11 species
  - Eucaryotes – 10,255; Viruses – 538; Bacteria – 79,592
- 10,000,000 simulated reads (100 nt) from the Clinical PathoScope project representing human, bacterial, and viral sequences

Table 1 Detailed information of sequences used to generate the simulated sequencing dataset for validation and comparison

Domain	Sequence description	Genbank identifier	Accession
Bacteria	*Bacillus anthracis* str. CDC 684 chromosome	227812678:c1-285375	NC_012581.1
	Escherichia coli O104:H4 str. 2011C-3493 chromosome	407479587:c1-457939	NC_018658.1
	Burkholderia mallei SAVP1 chromosome I	121598179:c1-586553	NC_008785.1
	Clostridium botulinum BKT015925 chromosome	331268188:c1-461407	NC_015425.1
	Staphylococcus aureus 08BA02176 chromosome	404477334:c1-460546	NC_018608.1
	Yersinia pestis A1122 chromosome	384137007:c1-457660	NC_017168.1
Viruses	Akabane virus segment M	157939617:c1-4309	NC_009895.1
	Newcastle disease virus isolate 2009_Mali_ML008	355467763:c1-15027	JF966387.1
	Influenza A virus (A/muscovy duck/Vietnam/LBM295/2012(H5N1)) viral cRNA, segment 7, complete sequence	464101994:c1-992	AB807883.1
Eukaryota	*Bos taurus* DNA sequence from clone CH240-405I18	445065096:c1-177923	FO393397.2
	Canis lupus familiaris clone rp81-289 m11	34787525:c1-193729	AC129099.6

Output (1)

```
******************** RIEMS - Rapid Extraction of Metagenomic Sequence data sets ********************

 ##############################                        Run analysis
 ##                          ##                        Start: Di 16. Jul 15:52:33 CEST 2013
 ##   Matagenomic analysis   ##                        Version: 2.0
 ##                          ##                        Sample: lib00223_reg12_RL17_run2013_07_15.sff
 ##############################                        Size: 67299 Reads -> 67284 high quality Reads

Size distribution          <50  51-100  101-150  151-200  201-250  251-300  301-350  351-400  401-450  451-500   >500

 Basic analysis
  Raw reads absolute       1617    3708     4953     6781     6132     7560     9365     8183     6993     5858   6134
  Raw reads portion       0.024   0.055    0.073    0.100    0.091    0.112    0.139    0.121    0.103    0.087  0.091
  Reads left absolute      1175    1866     1699     1602     1108     1157     1021      584      434      330    535
  Reads left portion      0.102   0.162    0.147    0.139    0.096    0.100    0.088    0.050    0.037    0.028  0.046
  Contigs left absolute       0       1        4       10       11        8       14        8       12       11     39

 Further analysis
  Reads left absolute      1143    1821     1637     1514     1021     1047      929      522      350      223     51
  Reads left portion      0.099   0.158    0.142    0.131    0.088    0.090    0.080    0.045    0.030    0.019  0.004
  Contigs left absolute       0       1        4       10       11        8       13        8       12       11     39


 ########################
 ##                    ##
 ##   Basic analysis   ##
 ##                    ##
 ########################

Identified Organisms...

Superkingdom Tax-ID
    Family Tax-ID                        Containing            Megablast     Megablast   Megablast     Megablast   Blastn      Blastn
         Tax-ID  Quantity  Mapping  viral Reads  Assembly  vs orgs      Identities  vs ntdb      identities  vs ntdb     identities  Scientific name

10239                 7                                                                                                          Viruses
    10292             1                                                                                                            Herpesviridae
         10390        1        0            0         0         0            0         0            0        1          74.31        Gallid herpesvirus 2
    10442             1                                                                                                            Baculoviridae
         70600        1        0            0         0         0            0         0            0        1          95.00        Epiphyas postvittana nucleopolyhedrovirus
    10482             1                                                                                                            Polydnaviridae
         452649       1        0            0         0         0            0         0            0        1          81.48        Cotesia sesamiae Mombasa bracovirus
    11571             2                                                                                                            Bunyaviridae
         11612        1        0            0         0         0            0         0            0        1          80.28        Impatiens necrotic spot virus
         11613        1        0            0         0         0            0         0            0        1          80.72        Tomato spotted wilt virus
    NA-10239          2                                                                                                            No name found to Tax-ID
         1229752      2        0            0         0         0            0         0            0        2          85.42        Campylobacter phage CP30A

2                   537                                                                                                          Bacteria
    119060            2                                                                                                            Burkholderiaceae
         1042878      1        0            0         0         0            0         1        77.26        0           0          Cupriavidus necator N-1
         312153       1        0            0         0         0            0         1        96.97        0           0          Polynucleobacter necessarius subsp. asymbioticus
    1268              1                                                                                                            Micrococcaceae
         861360       1        0            0         0         0            0         0            0        1          75.76        Arthrobacter arilaitensis Re117
```

Figure 1 Cut-out of a result protocol of the 'Basic analysis'. The first lines show general information concerning the analysis, followed by a read size distribution calculated before and after the 'Basic analysis' as well as calculated after the 'Further analysis'. The lower part displays the number of reads assigned to the detected species grouped at the domain and family levels. The central columns individually represent the results for mapping, assembly, and BLAST showing the number of reads detected by the respective tool. Blast results are accompanied by the range of identities of the read with the best hit in the database.

Output (2)

```
 ##########################
 ##                      ##                    Start: Mi 17. Jul 01:21:28 CEST 2013
 ##   Further analysis   ##                    Version: 3.0
 ##                      ##                    Sample: lib00223_reg12_RL17_run2013_07_15.sff
 ##########################

Identified Organisms in 11511 reads...

Superkingdom Tax-ID          Unfiltered    Unfiltered    Blastp                                
    Family Tax-ID              Blast         Blast         vs          Blastp                 
         Tax-ID  Quantity     vs ntdb     identities    protdb    identities  Scientific name

1                        1                                                    root
    NA-1                 1                                                      No name found to Tax-ID
         32630           1             1        95.83         0             0      synthetic construct

10239                    4                                                    Viruses
    10292                1                                                      Herpesviridae
         10310           1             1       100.00         0             0      Human herpesvirus 2
    10486                2                                                      Iridoviridae
         176652          1             1        87.50         0             0      Invertebrate iridescent virus 6
         262968          1             1        71.71         0             0      Singapore grouper iridovirus
    11571                1                                                      Bunyaviridae
         11612           1             1        80.95         0             0      Impatiens necrotic spot virus

2                      100                                                    Bacteria
    1162                 1                                                      Nostocaceae
         272123          1             1        94.12         0             0      Anabaena cylindrica PCC 7122
    119060               1                                                      Burkholderiaceae
         1229785         1             1        89.29         0             0      Burkholderia pseudomallei BPC006
```

Figure 2 Cut-out of a result protocol of the 'Further analysis' for reads. The first lines show general information of date, version, sample, and the number of reads used as input for the analysis. The following table displays the number of reads assigned to the detected species structured at the domain and family levels. The central columns individually display the number of reads detected by the different BLAST analyses. Blast results are accompanied by the range of identities of the read with the best hit in the database.

Output (3)

```
Identified Organisms in 118 contigs...

Superkingdom Tax-ID
    Family Tax-ID          Blastp                                                  Contig
         Tax-ID              vs       Hit       ORF       Blastp                 length
              Contig_ORF   protdb   range(aa)  length   identities   ORF(nucl)   (nucl)   Scientific name

2                             1                                                            Bacteria
    72293                     1                                                              Helicobacteraceae
         210                  1                                                                Helicobacter pylori
             contig00216_13   1        1-95      97      56.31         293-3      336         --- " " ---

2759                          2                                                            Eukaryota
    6243                      1                                                              Rhabditidae
         6238                 1                                                                Caenorhabditis briggsae
             contig00216_12   1        1-51      59      51.72        196-20      336         --- " " ---
    76657                     1                                                              Dermateaceae
         1072389              1                                                                Marssonina brunnea f. sp. 'multigermtubi' MB_m1
             contig00216_1    1        1-41      41      60.98        47-169      336         --- " " ---

117Contigs left, assembled of
2283Reads
```

Figure 3 Cut-out of a result protocol of the 'Further analysis' for contigs. The first line shows the number of contigs analysed. The following table displays the BLASTp hits for ORFs deduced from the nucleotide sequences of contigs that could not be taxonomically classified by nucleotide sequence analyses. The ORFs are assigned to species and the results are structured according to taxonomic domain and family levels. Additional columns show information about the deduced aa sequence and the alignment.

RIEMS with their simulated data

Table 2 Results of RIEMS validation using our simulated sample dataset with original (upper half) and deviating (lower half) sequences

Input species	True positive	False positive	True negative	False negative	Unclassified	Sensitivity	Specificity	Positive predictive value	Negative predicitive value	Correct classification rate	False classification rate
Original sequences											
Bacillus anthracis	12774	731	76880	0	0	100	99.06	94.59	100	99.19	0.81
Burkholderia mallei	16092	102	74129	62	0	99.62	99.86	99.37	99.92	99.82	0.18
Clostridium botulinum	12300	0	77649	436	0	96.58	100	100	99.44	99.52	0.48
Staphylococcus aureus	12425	0	77678	282	0	97.78	100	100	99.64	99.69	0.31
Escherichia coli	12289	0	77761	335	0	97.35	100	100	99.57	99.63	0.37
Yersinia pestis	12551	0	77788	46	0	99.63	100	100	99.94	99.95	0.05
Newcastle disease virus	418	0	89967	0	0	100	100	100	100	100	0
Akabane virus	119	0	90266	0	0	100	100	100	100	100	0
Influenza A virus	1	0	90384	0	0	100	100	100	100	100	0
Bos taurus	4906	0	85476	3	2	99.94	100	100	100	100	0
Canis lupus	5346	0	85039	0	0	100	100	100	100	100	0
Deviating sequences											
Bacillus anthracis	12762	633	76978	12	0	99.91	99.18	95.27	99.98	99.29	0.71
Burkholderia mallei	16055	79	74152	99	0	99.39	99.89	99.51	99.87	99.8	0.2
Clostridium botulinum	12358	1	77648	378	0	97.03	100	99.99	99.52	99.58	0.42
Staphylococcus aureus	12484	1	77677	223	0	98.25	100	99.99	99.71	99.75	0.25
Escherichia coli	12400	0	77761	224	1	98.23	100	100	99.71	99.75	0.25
Yersinia pestis	12529	0	77788	68	1	99.46	100	100	99.91	99.92	0.08
Newcastle disease virus	418	0	89967	0	0	100	100	100	100	100	0
Akabane virus	119	0	90266	0	0	100	100	100	100	100	0
Influenza A virus	1	0	90384	0	0	100	100	100	100	100	0
Bos taurus	4887	0	85476	22	5	99.55	100	100	99.97	99.98	0.02
Canis lupus	5346	0	85039	0	0	100	100	100	100	100	0

RIEMS with Clinical PathoScope data

Table 3 Results of RIEMS validation using the Clinical PathoScope simulated sample dataset

Input species	True positive	False positive	True negative	False negative	Unclassified	Sensitivity	Specificity	Positive predictive value	Negative predicitive value	Correct classification rate	False classification rate
Haemophilus influenzae	344646	17	9643530	7010	3	98.01	100	100	99.93	99.93	0.07
Homo sapiens	8975152	8541	986662	24848	7	99.72	99.14	99.9	97.54	99.67	0.33
Human mastadenovirus B	59289	0	9931563	4351	0	93.16	100	100	99.96	99.96	0.04
Human bocavirus	9	0	9995193	1	0	90	100	100	100	100	0
Human coronavirus NL63	24375	0	9970203	625	5	97.5	100	100	99.99	99.99	0.01
Enterovirus A	911	0	9994203	89	0	91.1	100	100	100	100	0
Human respiratory syncytial virus	9855	0	9985203	145	56	98.55	100	100	100	100	0
Rhinovirus C	241	0	9994953	9	0	96.4	100	100	100	100	0
Influenza A virus	15	0	9995103	85	0	15	100	100	100	100	0
Moraxella catarrhalis	169607	7	9823264	2325	1	98.65	100	100	99.98	99.98	0.02
Streptococcus pneumoniae	193919	18	9800120	1146	0	99.41	100	99.99	99.99	99.99	0.01
Streptococcus intermedius	175739	47	9818606	811	0	99.54	100	99.97	99.99	99.99	0.01

Analysis duration compared to BLAST

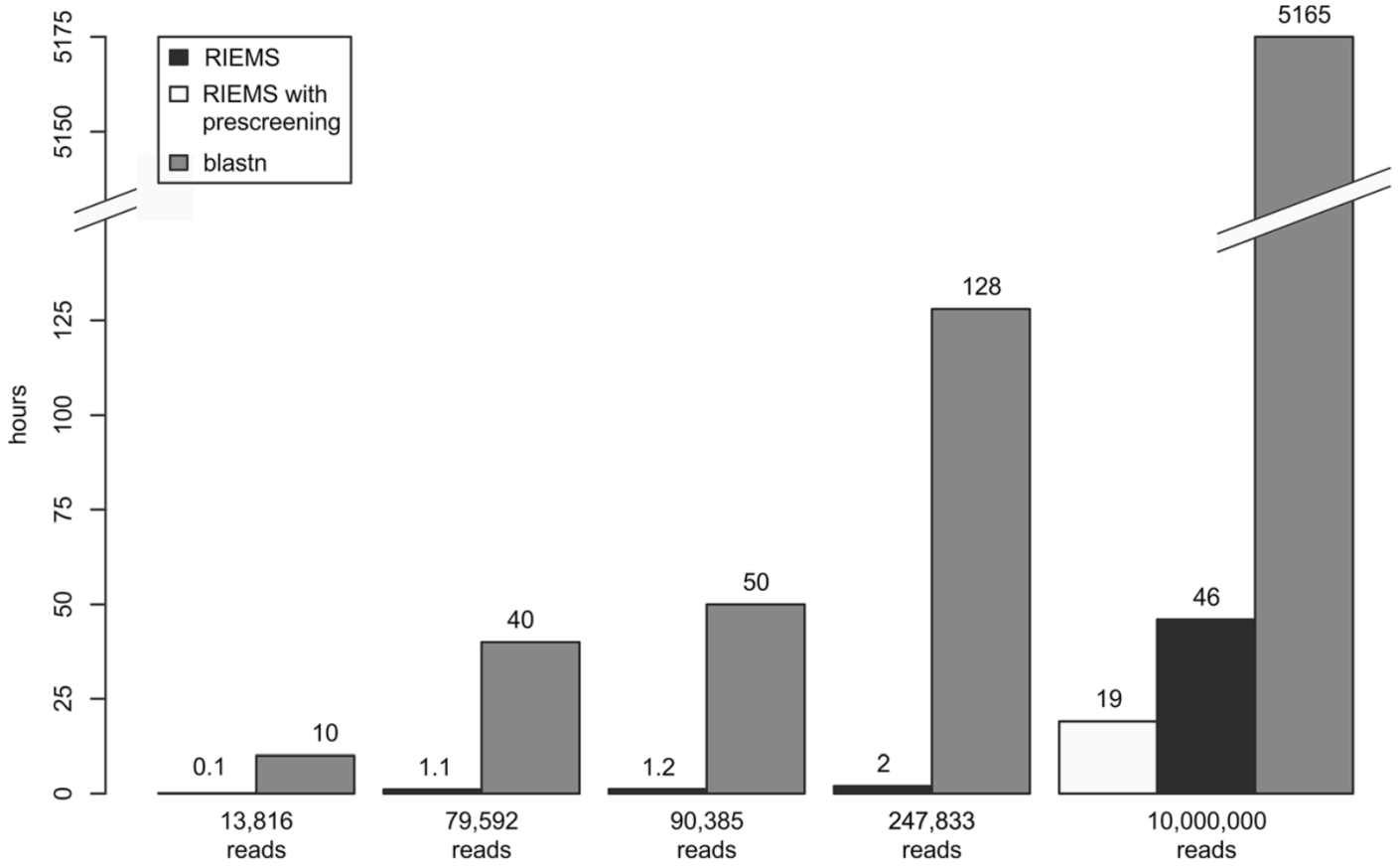


Figure 4 Comparison of the analysis duration of RIEMS and blastn analysis. All computations were performed locally using 24 cores and the NCBI nt database. For all datasets, RIEMS classified 99.9% of all reads. For the larger datasets, the duration of blastn analysis was extrapolated from the dataset with 13,816 reads.

RIEMS vs EBI metagenomics

Table 4 Comparison of RIEMS and EBI metagenomics results

Class	16S rRNA reads comprised		Assignments by RIEMS			EBI Metagenomics		
	absolute	%	Reads detected		Deviation (basis point)	OTUs detected		Deviation (basis point)
			absolute	%		absolute	%	
Bacilli	387	79.3	386	79.1	0.2	27	33.3	46.0
Clostridia	51	10.5	51	10.5	0.0	8	9.9	0.6
Unknown firmicutes	-	-	-	-	-	2	2.5	2.5
Alphaproteobacteria	-	-	-	-	-	3	3.7	3.7
Betaproteobacteria	50	10.3	51	10.5	0.2	8	9.9	0.4
Unknown proteobacteria	-	-	-	-	-	2	2.5	2.5
Unknown bacteria	-	-	-	-	-	30	37.0	37.0
Assigned			488	100		80	98.7	
Unassigned	-		0	0.0		1	1.3	
Total	488	100.0	488	100		81	100	

For this comparison a data subset of 488 16S rRNA sequences comprised in the simulated sequencing dataset was used. The deviation is presented in basis point between the percentage of the respective assignments and the percentage in the original dataset.

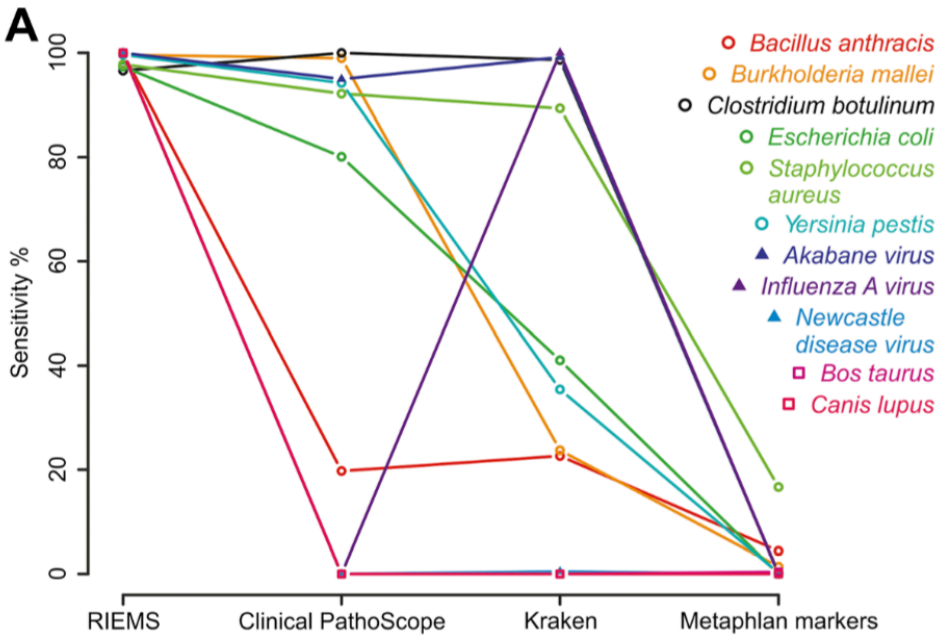
Supplementary Table 2 - Comparison of RIEMS and MG-RAST results obtained for the simulated read dataset

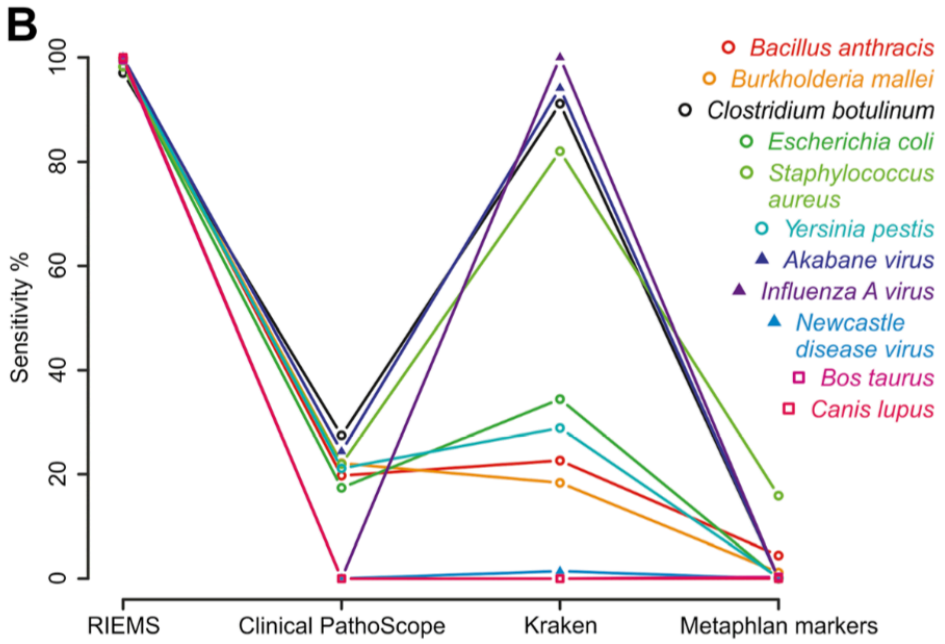
The MG-RAST results were formatted to allow the comparison with the RIEMS results (assignments were limited to the best hit; annotation source set to 'GenBank' with a maximum e-value cut-off of 10^{-4}; minimum percentage identity cut-off 60 %; minimum alignment length cut-off 15; result table grouped according to families). The columns "Deviation" hold information of the differences between the number of reads detected and the number of reads originally comprised. Note that the comparison in this table is based on results using databases as of the first quarter of 2013.

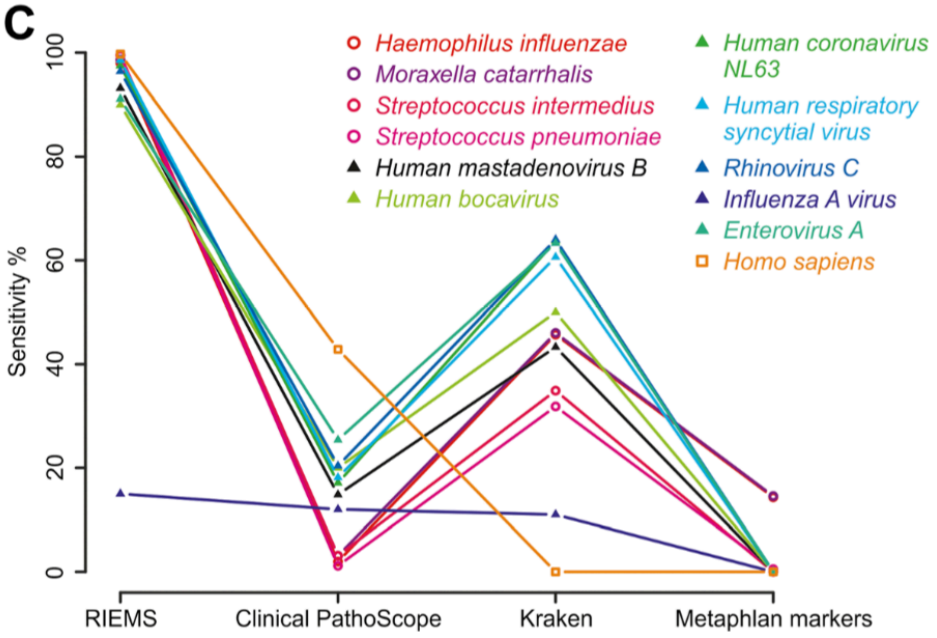
		Assignments by			
		MG-Rast normalised		RIEMS	
Family	Reads actually comprised	Reads classified	Deviation	Reads classified	Deviation
Bacillaceae	12,774	12,581	193	11,856	918
Burkholderiaceae	16,154	18,304	2,150	16,433	279
Clostridiaceae	12,736	15,819	3,083	12,506	163
Staphylococcaceae	12,707	13,321	614	12,690	17
Enterobacteriaceae	25,221	25,496	275	25,232	11
Paramyxoviridae	418	387	31	418	-
Bunyaviridae	119	97	22	119	-
Orthomyxoviridae	1	1	-	1	-
Bovidae	4,909	592	4,317	4,907	2
Canidae	5,346	-	5,346	5,346	-
Further families	-	4,701	4,701	28	28
Assigned	90,385	91,299	914	89,536	849
Unassigned	-	2		849	
Total reads analysed	90,385	91,301		90,385	

RIEMS vs MG-RAST

RIEMS vs locally installed pipelines







Comparison of the sensitivities of RIEMS, Clinical PathoScope, Kraken, and Megablast against the MetaPhlAn clade specific marker database.

The plots show the sensitivities calculated from the read to species assignments using the three simulated sample datasets. (**A**) Simulated sample comprising 90,385 reads representing original sequences derived from viral, bacterial, and eukaryotic genome sequences. (**B**) The same dataset as used in (**A**) but with 5 deviations per read. (**C**) The Clinical PathoScope simulated sample dataset.

Specificity almost 100% for all methods!

Speed

Supplementary Table 3 – Comparison of the run time (in min:sec) of the tools used on the different datasets

Dataset	RIEMS	Kraken	Clinical Pathoscope	Megablast vs. Metaphlan marker DB
RIEMS Original	151:00	0:43	5:15	0:48
RIEMS Deviating	156:00	0:44	2:27	0:36
Clinical PathoScope	2783:00	8:06	22:09	43:19