Bioinformatische Charakterisierung kurzer DNA-Fragmente aus bakteriellen Gemeinschaften

Jens Stoye

AG Genominformatik, Faculty of Technology

Institute for Bioinformatics, Center for Biotechnology

Bielefeld University, Germany

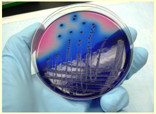
Overview

Overview

Metagenomics

Motivation

- $\approx$ 99% microbes not cultivable



- have to be taken from natural environment

Foto: Naturschutzhaus e.V.

- study whole microbial communities

illustration: Don Smith

Metagenomics

Motivation

- ≈ 99% microbes not cultivable
- have to
- communities

NATURE|Vol 453|29 May 2008

TECHNOLOGY FEATURE **METAGENOMICS**

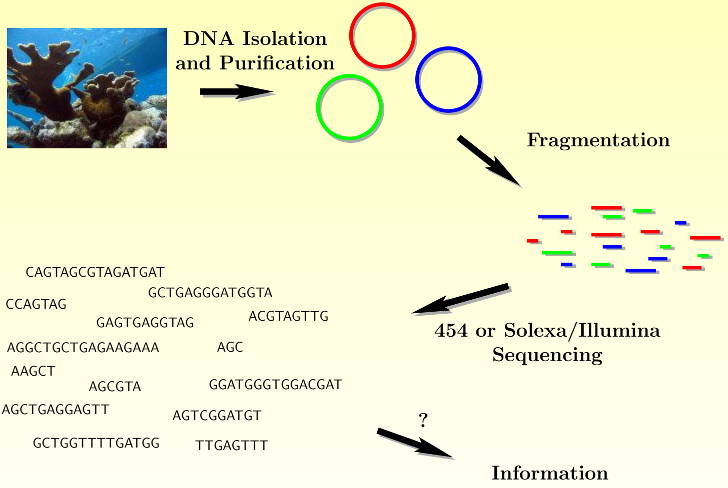
Exploring unseen communities

Advances in sequencing technology and tools for analysis are allowing researchers to unravel the environmental diversity of microbes faster and in greater detail than ever before. **Nathan Blow** reports.

Foto: Naturschutzhaus e.V.

illustration: Don Smith

Metagenomics



DNA Isolation and Purification

Fragmentation

454 or Solexa/Illumina Sequencing

CAGTAGCGTAGATGAT

GCTGAGGGATGGTA

CCAGTAG

ACGTAGTTG

GAGTGAGGTAG

AGGCTGCTGAGAAGAAA

AGC

AAGCT

AGCGTA

GGATGGGTGGACGAT

AGCTGAGGAGTT

AGTCGGATGT

GCTGGTTTTGATGG

TTGAGTTT

?

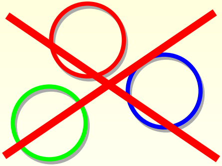
Information

Short Read Metagenomics: Difficulties

Problems with de novo assembly:

- short read length (35-250 bp)
- sequencing errors
- mixture of reads from different species

→ DNA barcoding is not possible.



Short Read Metagenomics: Our Approach

Retrieving information: our approach

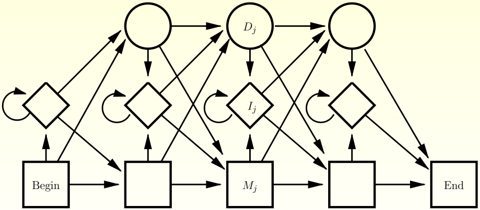
- Protein domains are highly conserved
- Find reads that encode for known proteins
  → environmental gene tags (EGTs)
- Provide information about:
  - taxonomical origin and function of EGTs
  - taxonomical composition
  - habitat-specific proteins
  - differences between communities from different environments

Pfam

Environmental gene tags are found by comparison of translated reads with Pfam domains

Pfam Database

- 10340 protein families (July 2008)
- multiple alignments
- Pfam profile hidden markov model (pHMM)
- used to match reads against protein families



Overview

CARMA

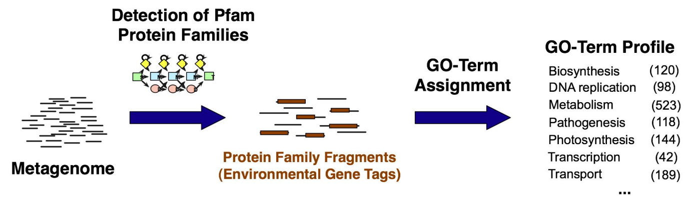
Pipeline for Characterizing Short Read Metagenomes

Two goals:

- (A) Functional Analysis
- (B) Taxonomical Classification

Krause *et al.* (2008), Phylogenetic classification of short environmental DNA fragments, *Nucleic Acids Res.* **36**(7): 2230-2239

(A) Functional Analysis

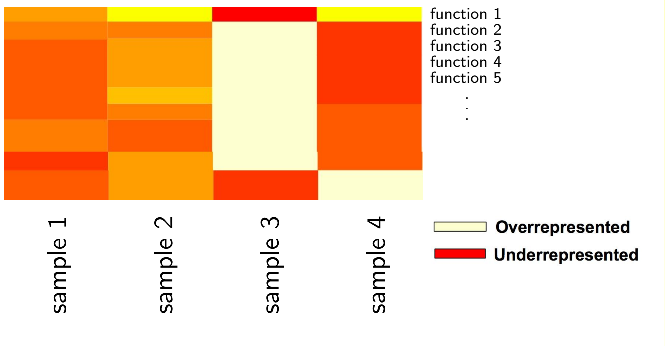


GO-Term Profile	
Biosynthesis	(120)
DNA replication	(98)
Metabolism	(523)
Pathogenesis	(118)
Photosynthesis	(144)
Transcription	(42)
Transport	(189)
...	

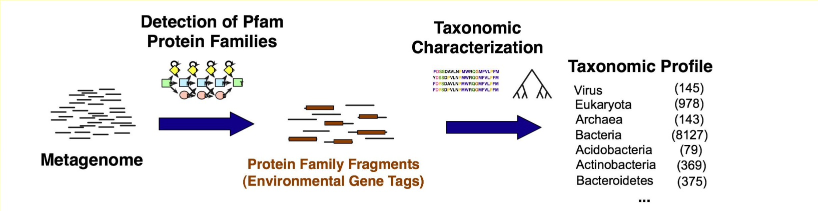
- Find environmental gene tags:
    - filter with BLASTX
    - alignment using HMM
- GO-term profile:
    - characterizes genetic diversity
    - potential metabolism of underlying community

(A) Functional Analysis: Result

Comparative analysis reveals genetic and metabolic trends



(B) Taxonomical Classification



Detection of Pfam Protein Families

Taxonomic Characterization

Metagenome

Protein Family Fragments (Environmental Gene Tags)

Taxonomic Profile

Virus	(145)
Eukaryota	(978)
Archaea	(143)
Bacteria	(8127)
Acidobacteria	(79)
Actinobacteria	(369)
Bacteroidetes	(375)
...	

Taxonomic Profile

- assignment to taxonomic groups based on a phylogenetic analysis (binning)
- characterizes species composition
- unlike 16S rDNA analysis takes all available sequence data into account

(B) Taxonomical Classification

Details of the procedure:

Multiple alignment of known family members:
(Pfam family: PF01973)

```
PF1  ...INSHSSYKAIVIGSGPSLEEHYDYLQQISSS--SHRPLIIAADTALRGLLHNNIKPDIVICI--DGLIG...
PF2  ...-----GREIFVIGTGPTLEQHFERLAVIRER--ADRPLYICVDTAYRPLREHGIVPDYVVSI--D----...
PF3  ...---IKGREAYVLATGPTLAGHFERLKQVREQ--AERPLFICVDTALRPLLEHGIRPDIV-VT-------...
PF4  ...V---PRQQVIVVGAGPSLEAALSQLECISKR--KNRPLIIAVDTALKPLLQNGIRPNIVVSI--D-ADI...
PF5  ...KSIAKNKEAFVIGAGPSLEEHYSFLASISKQPLHHRPIIIAVDAAAKGLLAHGVKLDIV-VTIDE-MID...
```

(B) Taxonomical Classification

Details of the procedure:

Multiple alignment of known family members:
(Pfam family: PF01973)

```
PF1  ...INSHSSYKAIVIGSGPSLEEHYDYLQQISSS--SHRPLIIAADTALRGLLHNNIKPDIVICI--DGLIG...
PF2  ...-----GREIFVIGTGPTLEQHFERLAVIRER--ADRPLYICVDTAYRPLREHGIVPDYVVSI--D----...
PF3  ...---IKGREAYVLATGPTLAGHFERLKQVREQ--AERPLFICVDTALRPLLEHGIRPDIV-VT-------...
PF4  ...V---PRQQVIVVGAGPSLEAALSQLECISKR--KNRPLIIAVDTALKPLLQNGIRPNIVVSI--D-ADI...
PF5  ...KSIAKNKEAFVIGAGPSLEEHYSFLASISKQPLHHRPIIIAVDAAAKGLLAHGVKLDIV-VTIDE-MID...
```

- Add gene tag to full multiple alignment using hmmalign.

(B) Taxonomical Classification

Details of the procedure:

Multiple alignment of known family members:
(Pfam family: PF01973)

```
PF1  ...INSHSSYKAIVIGSGPSLEEHYDYLQQISSS--SHRPLIIAADTALRGLLHNNIKPDIVICI--DGLIG...
PF2  ...-----GREIFVIGTGPTLEQHFERLAVIRER--ADRPLYICVDTAYRPLREHGIVPDYVVSI--D----...
PF3  ...---IKGREAYVLATGPTLAGHFERLKQVREQ--AERPLFICVDTALRPLLEHGIRPDIV-VT-------...
PF4  ...V---PRQQVIVVGAGPSLEAALSQLECISKR--KNRPLIIAVDTALKPLLQNGIRPNIVVSI--D-ADI...
PF5  ...KSIAKNKEAFVIGAGPSLEEHYSFLASISKQPLHHRPIIIAVDAAAKGLLAHGVKLDIV-VTIDE-MID...
EGT             KAIVIGSGPSLEEHYDYLQQIS
```

↑ gene tag

- Add gene tag to full multiple alignment using hmmalign.

(B) Taxonomical Classification

Details of the procedure:

```
PF1  ...INSHSSYKAIVIGSGPSLEEHYDYLQQISSS--SHRPLIIAADTALRGLLHNNIKPDIVICI--DGLIG...
PF2  ...-----GREIFVIGTGPTLEQHFERLAVIRER--ADRPLYICVDTAYRPLREHGIVPDYVVSI--D----...
PF3  ...---IKGREAYVLATGPTLAGHFERLKQVREQ--AERPLFICVDTALRPLLEHGIRPDIV-VT-------...
PF4  ...V---PRQQVIVVGAGPSLEAALSQLECISKR--KNRPLIIAVDTALKPLLQNGIRPNIVVSI--D-ADI...
PF5  ...KSIAKNKEAFVIGAGPSLEEHYSFLASISKQPLHHRPIIIAVDAAAKGLLAHGVKLDIV-VTIDE-MID...
EGT            KAIVIGSGPSLEEHYDYLQQIS
```

$$\Downarrow$$

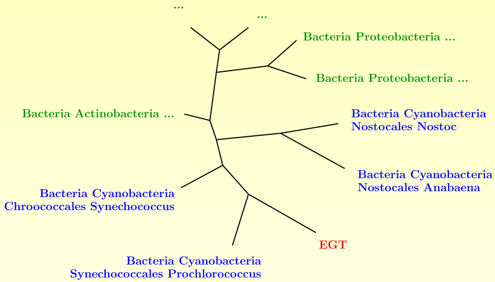
	PF1	PF2	...	EGT
PF1	0	0.3	...	0.6
PF2	0.3	0	...	0.2
...	...	...	...	...
EGT	0.6	0.2	...	0

(B) Taxonomical Classification

Details of the procedure:

	PF1	PF2	...	EGT
PF1	0	0.3	...	0.6
PF2	0.3	0	...	0.2
...	...	...	...	...
EGT	0.6	0.2	...	0

⟶



Phylogenetic tree

- Neighbor Joining
- Gene tags classified based on their location in tree

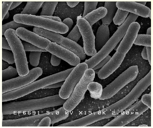
(B) Taxonomical Classification

Classification by longest common prefix of neighbor taxa:

- Bacteria Cyanobacteria Synechococcales Prochlorococcus
- Bacteria Cyanobacteria Chroococcales Synechococcus
- Bacteria Cyanobacteria Nostocales ...

→ 'Bacteria Cyanobacteria'

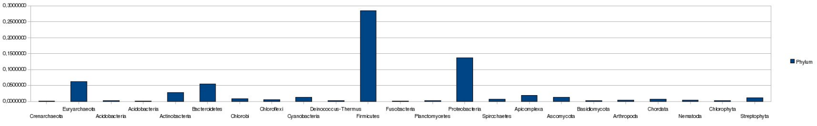
Reminder: taxonomic ranks



Level	Taxonomic Rank	Example: *E. coli*
1	Superkingdom	Bacteria
2	Phylum	Proteobacteria
3	Class	Gammaproteobacteria
4	Order	Enterobacteriales
5	Genus	Escherichia
6	Species	*Escherichia coli*

(B) Taxonomical Classification: Result

Typical output:



Overview

Evaluation: Creating a Standard of Truth

Test set: 77 complete genomes

- 2 superkingdoms (archeae and bacteria)
- 10 phyla
- 29 classes
- 62 genera
- 77 species

Test set excluded from reference set:

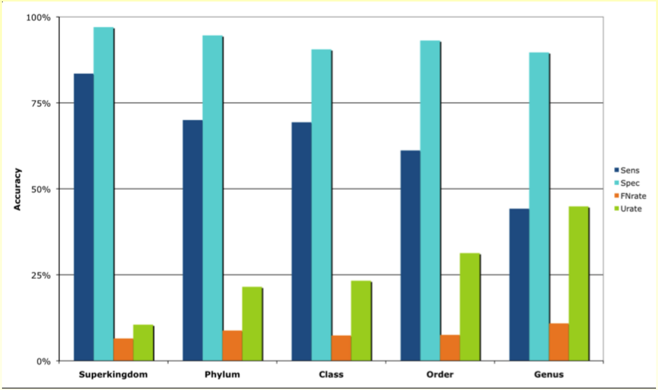
Pfam members from any of the 77 species were omitted from full multiple alignments

Evaluation: Creating a Standard of Truth

Simulation of a typical 454-sequenced metagenomic data set:

- 77 genomes fragmentized with ReadSim (Schmid *et al.* 2008)
- simulates sequencing using 454 pyrosequencing
- fragments randomly sampled (2x)
- fragment length: 80-120bp, mean 100bp
- simulates sequencing errors at homopolymers

Evaluation: Classification Accuracy



Sens: Sensitivity, fraction of correctly classified EGTs
Spec: Specificity, reliability of predictions
FNrate: False negative rate, proportion of wrongly classified EGTs
Urate: Unknown rate, proportion of EGTs not assigned to any taxonomic group

Application Example 1

Comparative Analysis of Four Microbial Coral Reef Communities

In cooperation with Rob Edwards and Forest Rohwer
(San Diego State University, California)

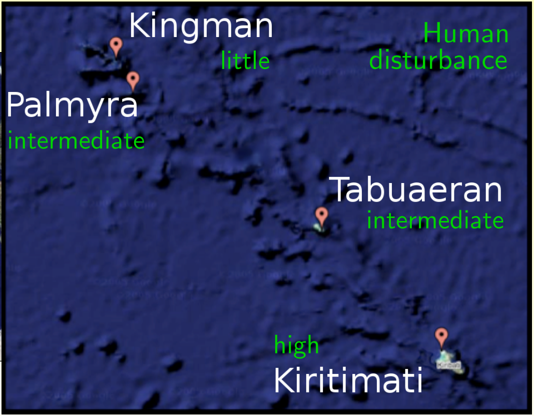
Dinsdale *et al.* (2008), Microbial Ecology of Four Coral Atolls in the Northern Line Islands, *PLoS ONE* **3**(2), e1584

Application Example 1

Influence of Human Activities on Coral Reef Microbial Communities

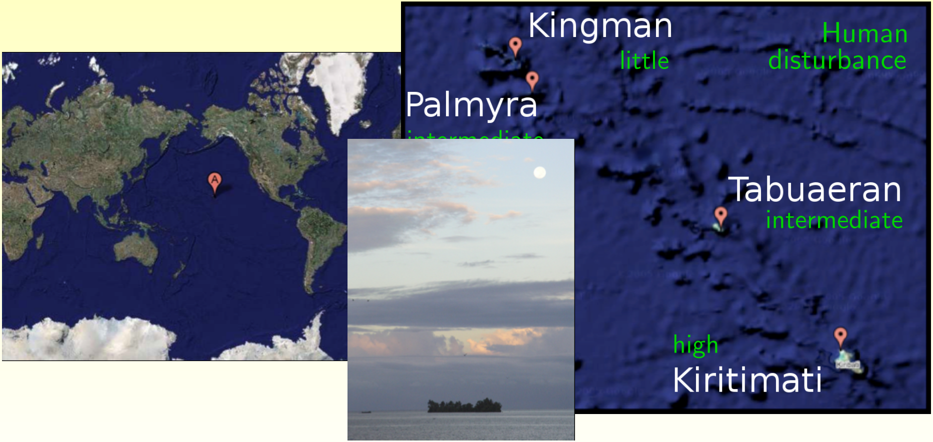
Application Example 1

Influence of Human Activities on Coral Reef Microbial Communities



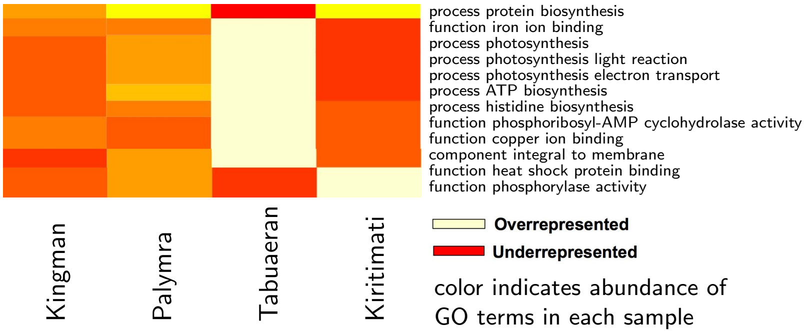
Application Example 1

Influence of Human Activities on Coral Reef Microbial Communities



Application Example 1

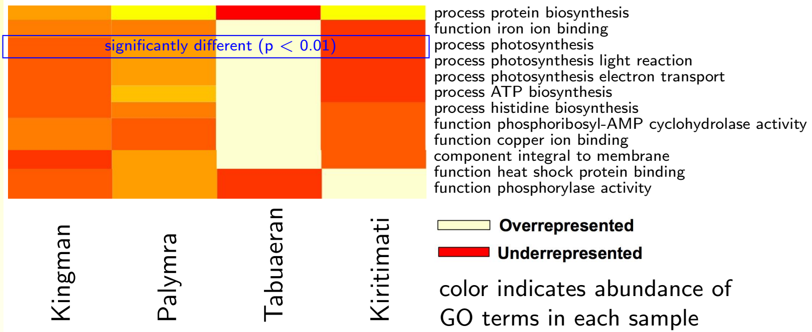
GO-term profiles indicate transition in metabolic activities



color indicates abundance of GO terms in each sample

Application Example 1

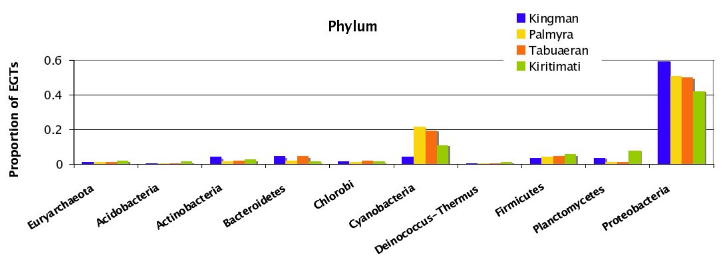
GO-term profiles indicate transition in metabolic activities



color indicates abundance of GO terms in each sample

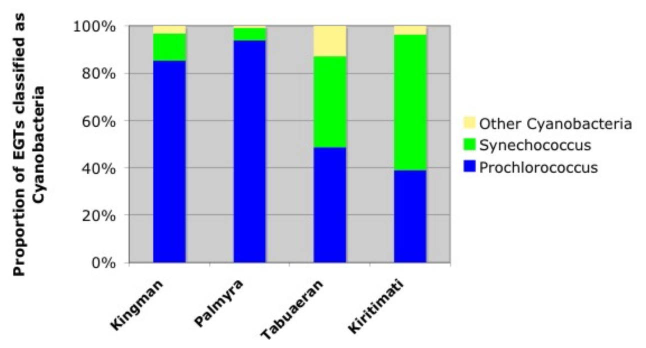
Application Example 1

Coral reef community structure



Application Example 1

Taxonomic profiles indicate transition from Prochlorococcus to Synechococcus (most abundant marine Cyanobacteria)



Application Example 2

Comparative Analysis of Three Aquatic Microbial Communities

In cooperation with Rob Edwards and Forest Rohwer
(San Diego State University, California)

Krause *et al.* (2008), Phylogenetic classification of short environmental DNA fragments, *Nucleic Acids Res.* **36**(7), 2230-2239

Application Example 2: Sampling Locations

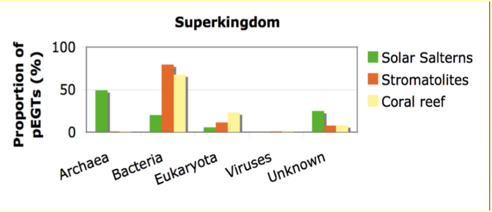
San Diego solar salterns, USA

Rios Mesquites stromatolites, Mexico

Kingman coral reef, Northern Line Islands

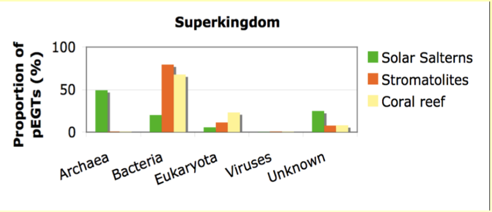
Sample data provided by Forest Rohwer and Robert Edwards

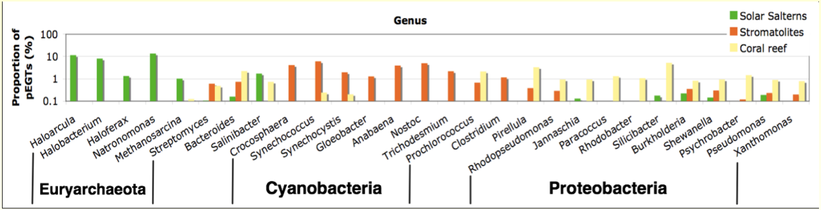
Application Example 2: Community Structure



pEGTs: prokaryotic fraction of EGTs

Application Example 2: Community Structure





pEGTs: prokaryotic fraction of EGTs

Application Example 2: Taxonomic Diversity

Sample	Phylum		Class		Order		Genus	
	H'	J	H'	J	H'	J	H'	J
Solar salterns	0.8	0.31	1.0	0.32	1.4	0.28	2.6	0.45
Stromatolites	1.2	0.42	1.2	0.37	2.7	0.55	3.6	0.70
Coral reef	1.2	0.46	1.7	0.55	3.9	0.81	4.2	0.83

H': Diversity, including richness and evenness (Shannon index)

J: Evenness, relative commonness and rarity of organisms

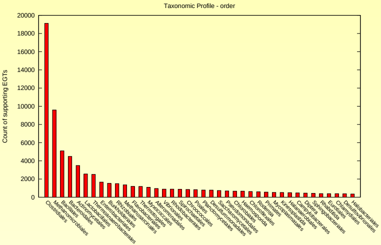
Further Applications of CARMA

Diversity of coral reef viruses
(in cooperation with Stuart Sandin, Scripps Institution of Oceanography, San Diego, USA)

Waste Water Treatment Plant plasmid sample
(in cooperation with Andreas Schlüter, Bielefeld University)

Latest development: WebCARMA

⇒



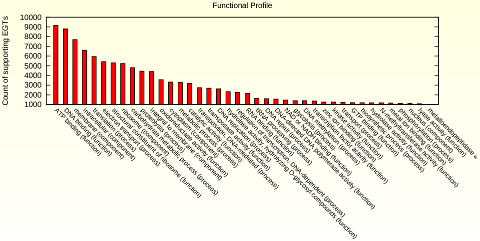
http://webcarma.cebitec.uni-bielefeld.de

⇘

⇓

controlled vocabulary:

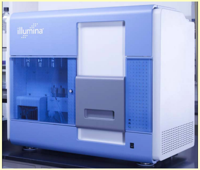
Gene Ontology
NCBI taxonomy



Overview

1. Metagenomics
2. The CARMA Pipeline for Short Read Metagenome Analysis
3. Evaluation and Applications
4. Algorithmic Issues for Very Large Datasets
5. Further Applications
6. Conclusion

Challenges of Very Large Data Sets



Illumina/Solexa technology provides millions of ultra-short reads.

- Pro: cost per base ca. $10-100\times$ lower
- Contra: reads are quite short ($35-100\,$bp)

Necessary improvements

- Optimization of repeated tree reconstruction (caching)
- fast pHMM filtration (q-gram index and bucketing)
- incorporation of mate-pair information

Speeding up CARMA: q-Gram Index

Index: **q=3**

AAA →
AAC →
AAD →
... →
DDV → 1
... →
DVV → 2,6
... →
VDC → 9
... →
VDV → 5
... →
VVD → 4,8
... →
VVV → 3,7

Protein:	**D**	**D**	**V**	**V**	**V**	**D**	**V**	**V**	**V**	**D**	**C**
Position:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**

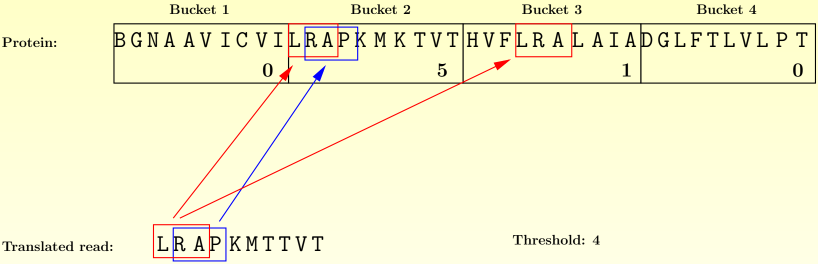
Speeding up CARMA: q-Gram Index

	Bucket 1	Bucket 2	Bucket 3	Bucket 4
Protein:	BGNAAVICVI	LRAPKMKTVT	HVFLRALAIA	DGLFTLVLPT
	0	**5**	**1**	**0**

Translated read: LRAPKMTTVT

Threshold: 4

Speeding up CARMA: q-Gram Index



	Bucket 1	Bucket 2	Bucket 3	Bucket 4
Protein:	B G N A A V I C V I	L R A P K M K T V T	H V F L R A L A I A	D G L F T L V L P T
	0	**5**	**1**	**0**

Translated read: L R A P K M T T V T

Threshold: 4

Speeding up CARMA: q-Gram Index

Adaptation for multiple alignment: avoid redundant pointers

```
PF1   INSHSSYKAIVIGSGPSLEEHYDYLQQISSS--SHRPLIIAADTALRGLLHNNIKPDIVICI--DGLIG
PF2   -----GREIFVIGTGPTLEQHFERLAVIRER--ADRPLYICVDTAYRPLREHGPSPDYVVSI--D----
PF3   ---IKGREAYVLATGPTLAGHFERLKQVREQ--AERPLFICVDTALRPLLEHGIRPDIV-VT-------
PF4   V---PRQQVIVVGAGPSLEAALSQLECISKR--KNRPLIIAVDTALKPLLQNGIRPNIVVSI--D-ADI
PF5   KSIAKNKEAFVIGAGPSLEEHYSFLASISKQPLHHRPIIIAVDAAAKGLLAHGVKLDIV-VTIDE-MID
```

Index:

...

`AAD` → 41

...

`CVD` → 41

...

`AVD` → 41

...

`GPS` → 15, 53 (only one pointer per column)

...

Speeding up CARMA: q-Gram Index

For protein q-grams, also consider neighborhood
(similar to BLAST)

Index:

. . .

`AAD` → 41; `CVD`, `AVD`

. . .

`CVD` → 41; `AAD`, `AVD`

. . .

`AVD` → 41; `AAD`, `CVD`

. . .

`GPS` → 15, 53; `GPT`

. . .

Speeding up CARMA: Results

Data set

- 5.5M Solexa/Illumina reads (50bp)
- 246,083 EGTs extracted

Runtimes

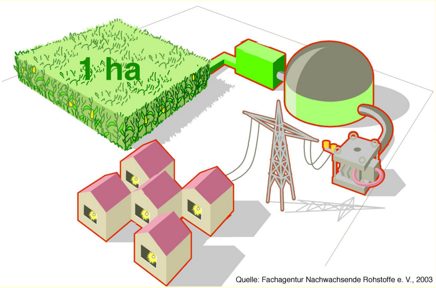
- CPU hours
  790h 26m → 631h 38m
- On Bielefeld University Bioinformatics Resource Facility Compute Cluster (ca. 300 CPU cores)
  36h 19m → 28h 13m

Overview

Overview

1. Metagenomics
2. The CARMA Pipeline for Short Read Metagenome Analysis
3. Evaluation and Applications
4. Algorithmic Issues for Very Large Datasets
5. Further Applications
   - Short vs. Ultra-Short Reads
   - Number of Reads
   - Incorporation of Mate Pairs
6. Conclusion

What lives in a biogas reactor?



Quelle: Fachagentur Nachwachsende Rohstoffe e. V., 2003

454 data set:

Schlüter, *et al.* (2008), The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology, *J. Biotechnol.* **136**(1-2): 77-90

Krause, *et al.* (2008), Taxonomic composition and gene content of a methane-producing microbial community isolated from a biogas reactor, *J. Biotechnol.* **136**(1-2): 91-101

Solexa/Illumina Appraoch to Metagenomics

Sequencing

- Cooperation of Illumina and Bielefeld University
- Sequenced the same biogas sample

Adaptation of sequencing protocol

- Maximal read length (50bp in spring 2008)
- Minimal insert size for paired-end reads (mate pairs)

Data obtained

35 bp reads: 68 M
50 bp reads: 77 M (most experiments done on a 5.5 M subset)

Data Sets

Solexa/Illumina

- $5,475,293$ reads (50 bp)
- $246,083$ EGTs (4.5%)

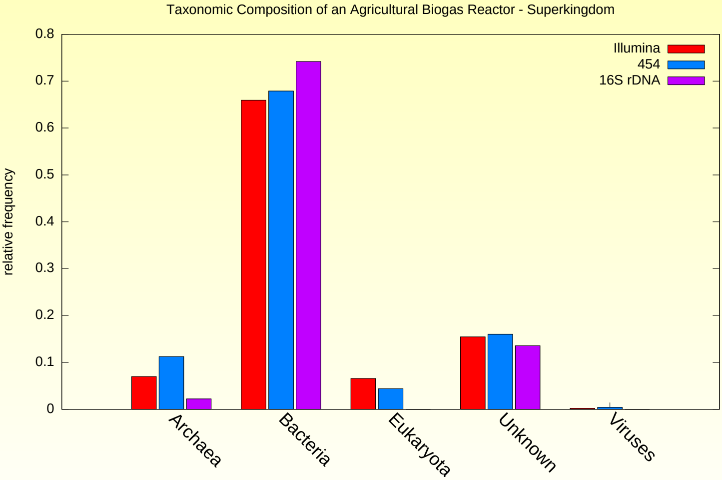
454/Roche

- $616,072$ reads (avg. 230 bp)
- $160,284$ EGTs (26.0%)

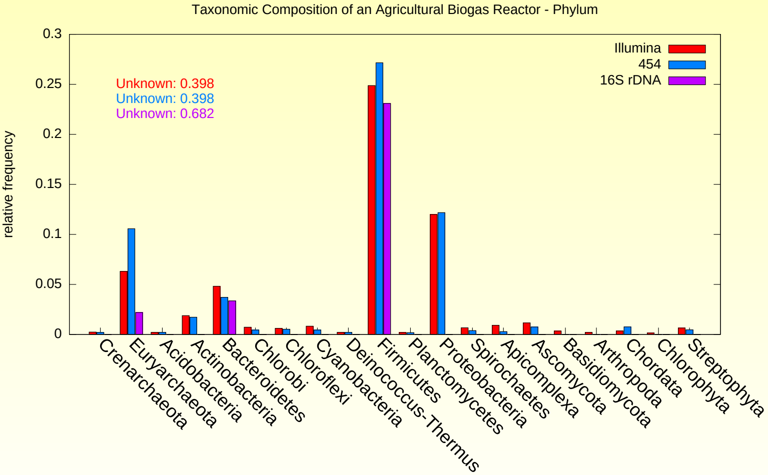
16S rDNA (from 454 reads)

- Procedure:
  1. Detection of reads that encode for 16S rDNA genes using BLAST and an rRNA database
  2. Ribosomal Database Project (RDP) Classifier, a naive Bayesian classifier
- 1,930 reads that encode for 16S rDNAs

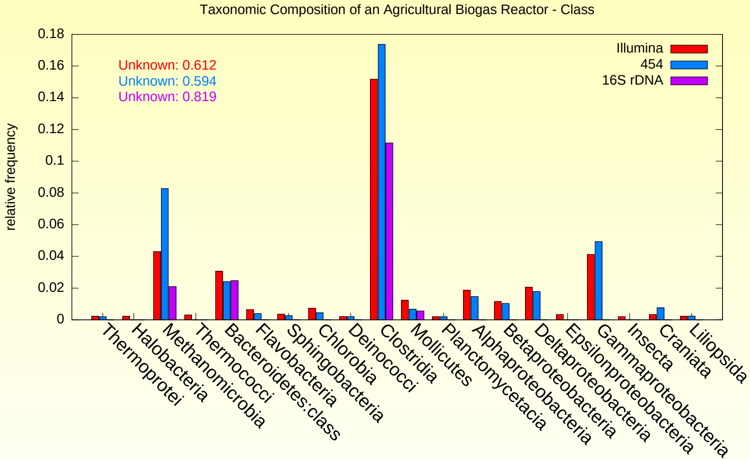
Results: (1) Superkingdom



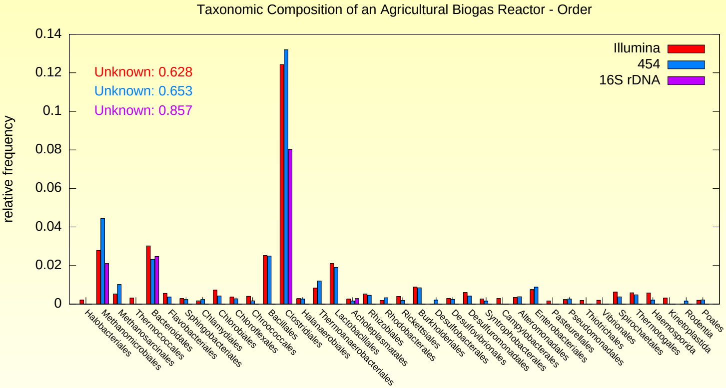
Results: (2) Phylum



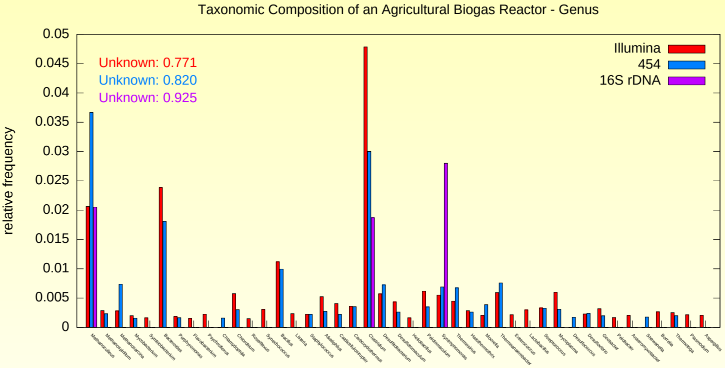
Results: (3) Class



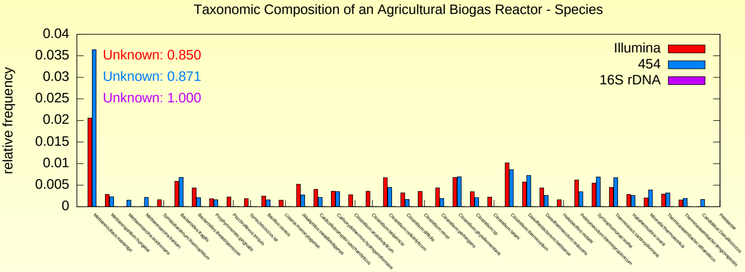
Results: (4) Order



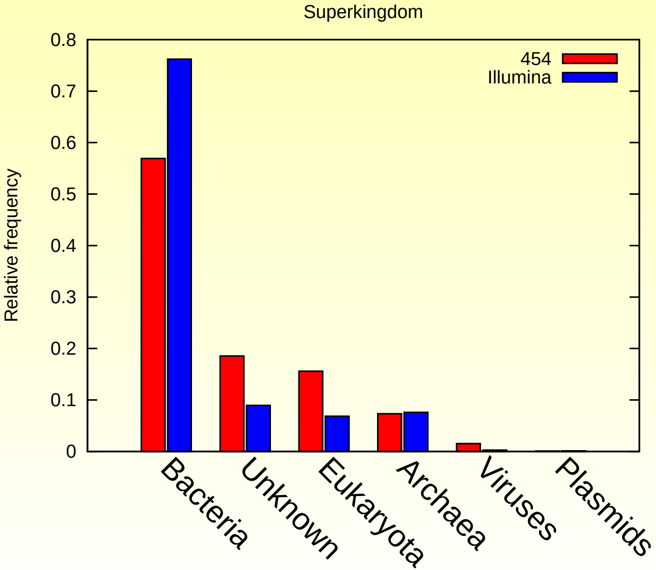
Results: (5) Genus



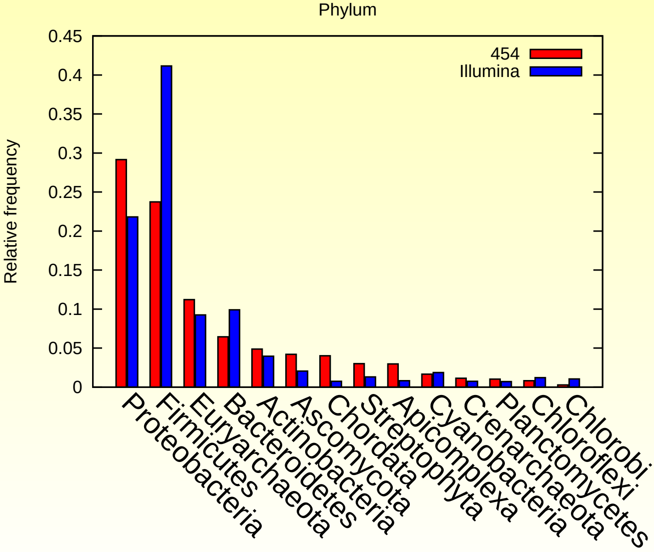
Results: (6) Species



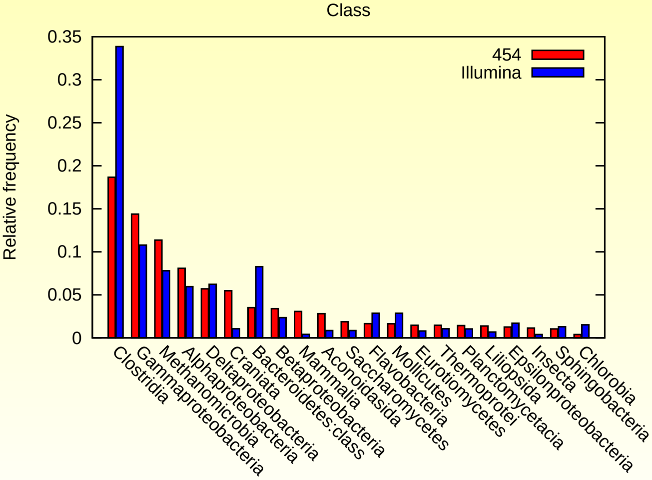
Closer look: (1) Superkingdom



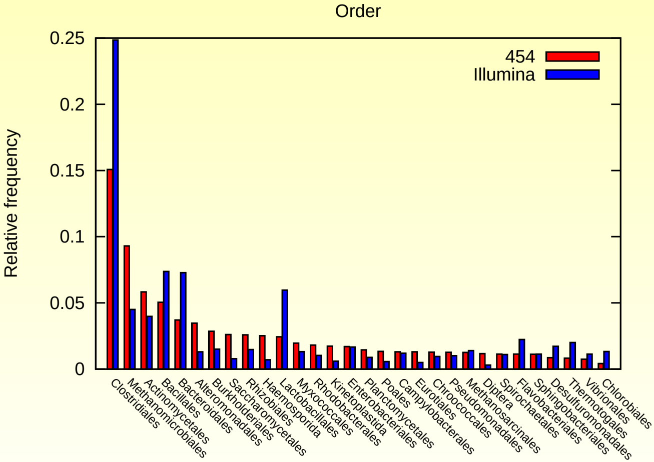
Closer look: (2) Phylum



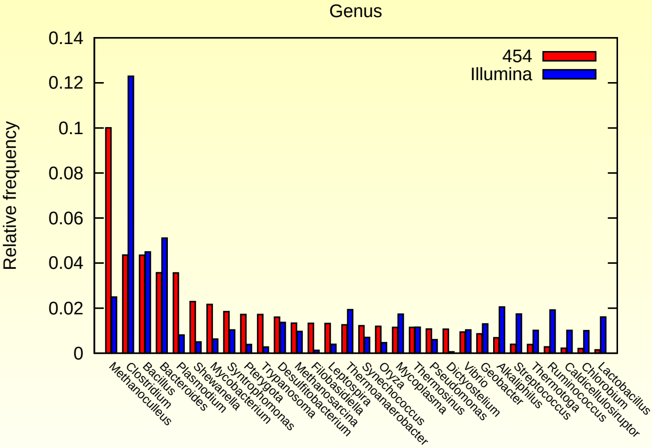
Closer look: (3) Class



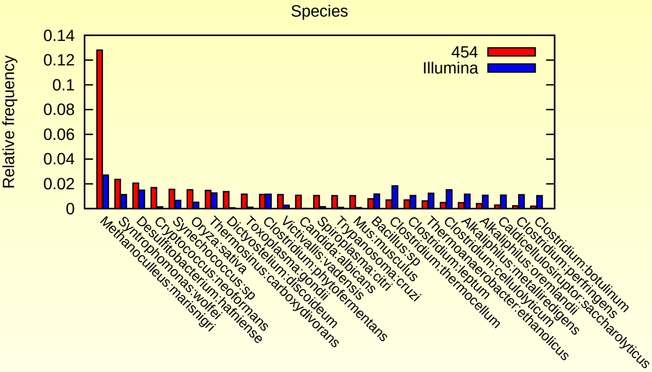
Closer look: (4) Order



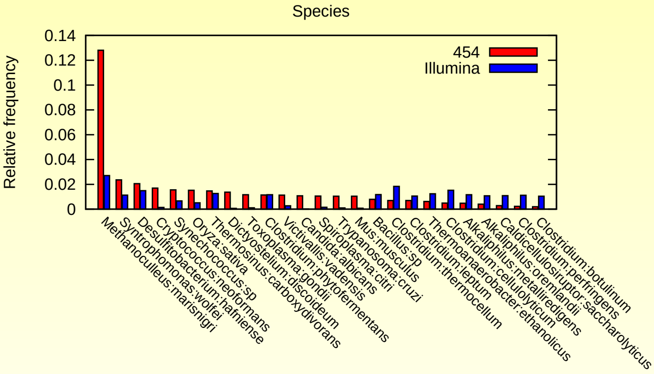
Closer look: (5) Genus



Closer look: (6) Species



Closer look: (6) Species

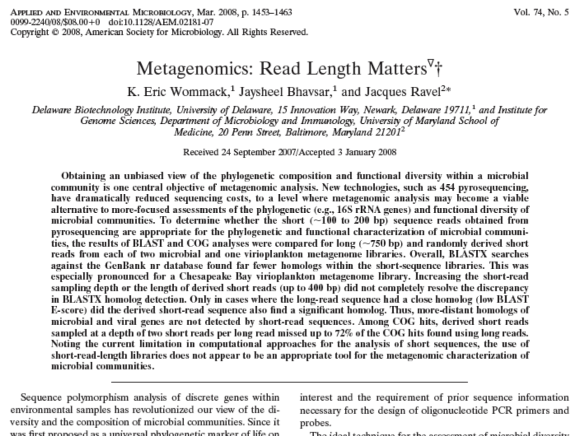


Comparison at this level is critical: GC Bias?

- 454 data: High GC content (52-55%)
- Illumina data: Low GC content (45%)

Experiment 2: Read Length Matters?

Another hypothesis: Ultra-short read length?



Applied and Environmental Microbiology, Mar. 2008, p. 1453–1463 Vol. 74, No. 5
0099-2240/08/$08.00+0 doi:10.1128/AEM.02181-07

Metagenomics: Read Length Matters ▽ †

K. Eric Wommack,[1] Jaysheel Bhavsar,[1] and Jacques Ravel[2]*

Delaware Biotechnology Institute, University of Delaware, 15 Innovation Way, Newark, Delaware 19711,[1] and Institute for Genome Sciences, Department of Microbiology and Immunology, University of Maryland School of Medicine, 20 Penn Street, Baltimore, Maryland 21201[2]

Received 24 September 2007/Accepted 3 January 2008

Obtaining an unbiased view of the phylogenetic composition and functional diversity within a microbial community is one central objective of phylogenomic analysis. New technologies, such as 454 pyrosequencing, have dramatically reduced sequencing costs, to a level where metagenomic analysis may become a viable alternative to more-focused assessments of the phylogenetic (e.g., 16S rRNA genes) and functional diversity of microbial communities. To determine whether the short (~100 to 200 bp) sequence reads obtained from pyrosequencing are appropriate for the phylogenetic and functional characterization of microbial communities, the results of BLAST and COG analyses were compared for long (~750 bp) and randomly derived short reads from each of two microbial and one virioplankton metagenome libraries. Overall, BLASTX searches against the GenBank nr database found far fewer homologs within the short-sequence libraries. This was especially pronounced for a Chesapeake Bay virioplankton metagenome library. Increasing the short-read sampling depth or the length of derived short reads (up to 400 bp) did not completely resolve the discrepancy in BLASTX homolog detection. Only in cases where the long-read sequence had a close homolog (low BLAST E-score) did the derived short-read sequence also find a significant homolog. Thus, more-distant homologs of microbial and viral genes are not detected by short-read sequences. Among COG hits, derived short reads sampled at a depth of two short reads per long read missed up to 72% of the COG hits found using long reads. Noting the current limitation in computational approaches for the analysis of short sequences, the use of short-read-length libraries does not appear to be an appropriate tool for the metagenomic characterization of microbial communities.

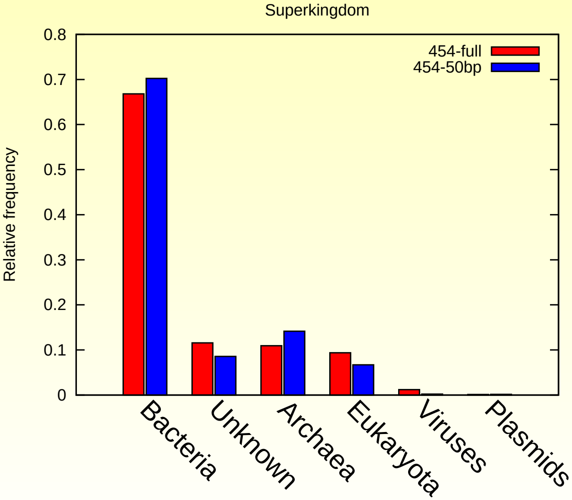
Sequence polymorphism analysis of discrete genes within environmental samples has revolutionized our view of the diversity and the composition of microbial communities. Since it was first proposed as a universal phylogenetic marker of life on

interest and the requirement of prior sequence information necessary for the design of oligonucleotide PCR primers and probes.

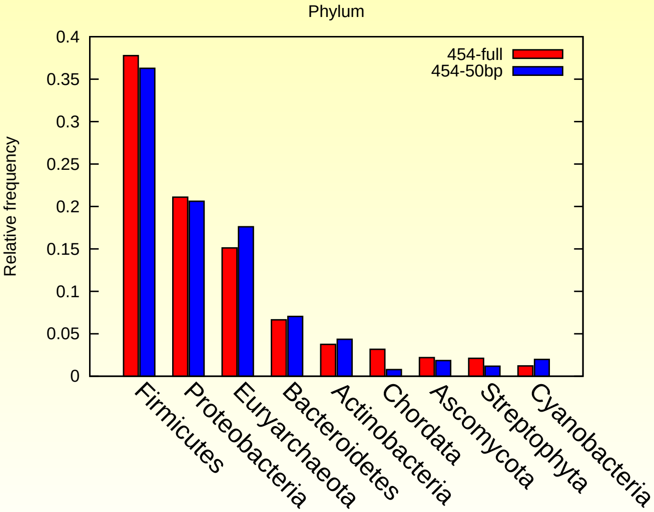
The ideal technique for the assessment of microbial diversity

Experiment 2: Read Length Matters?

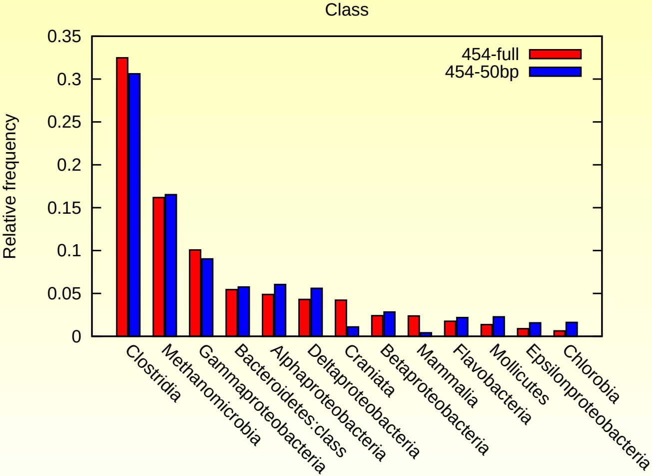
Simulated Ultra-Short Reads: 50bp extracts from 454 reads



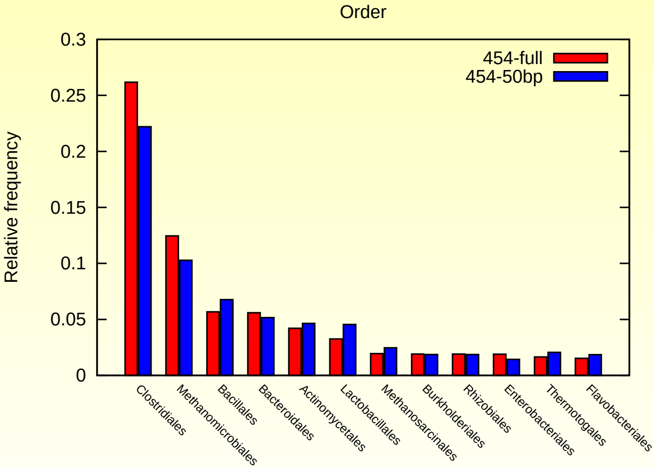
Experiment 2: Read Length Matters?



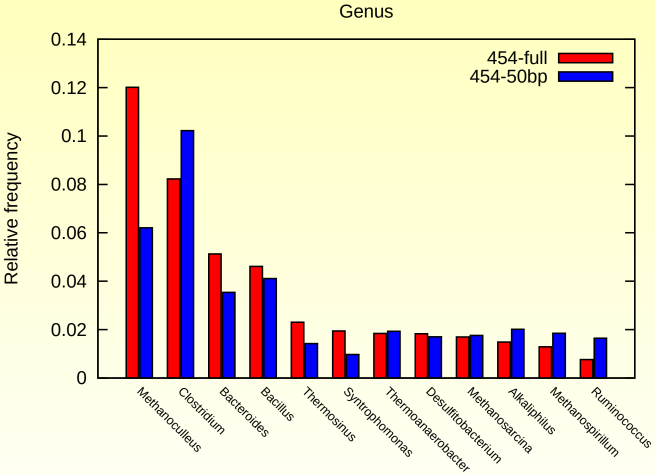
Experiment 2: Read Length Matters?



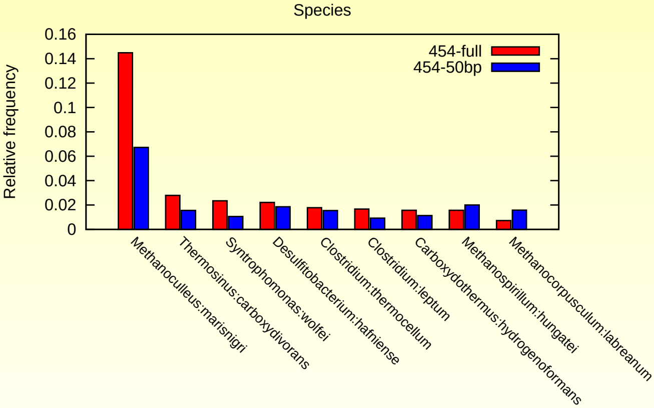
Experiment 2: Read Length Matters?



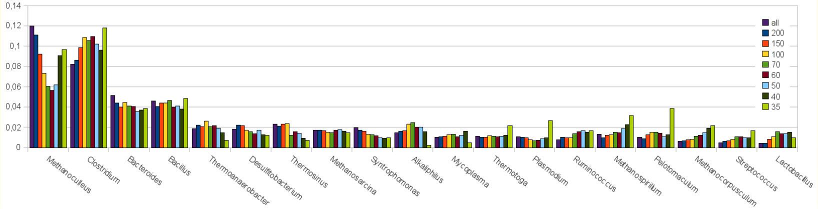
Experiment 2: Read Length Matters?



Experiment 2: Read Length Matters?



Experiment 2: Closer look at Genus level



Overview

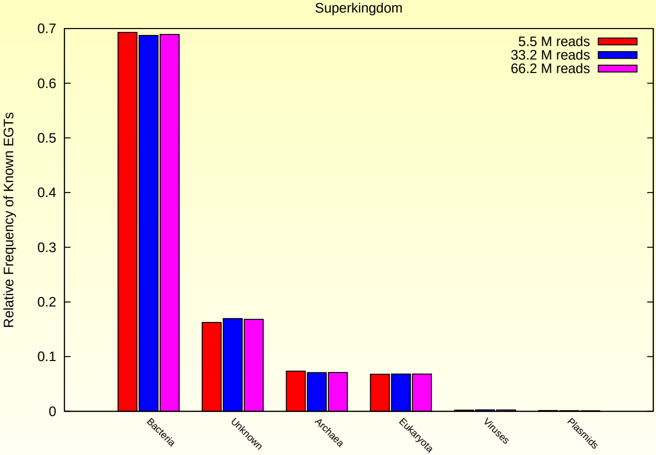
1. Metagenomics
2. The CARMA Pipeline for Short Read Metagenome Analysis
3. Evaluation and Applications
4. Algorithmic Issues for Very Large Datasets
5. Further Applications
    - Short vs. Ultra-Short Reads
    - Number of Reads
    - Incorporation of Mate Pairs
6. Conclusion

How many Ultra-Short Reads?

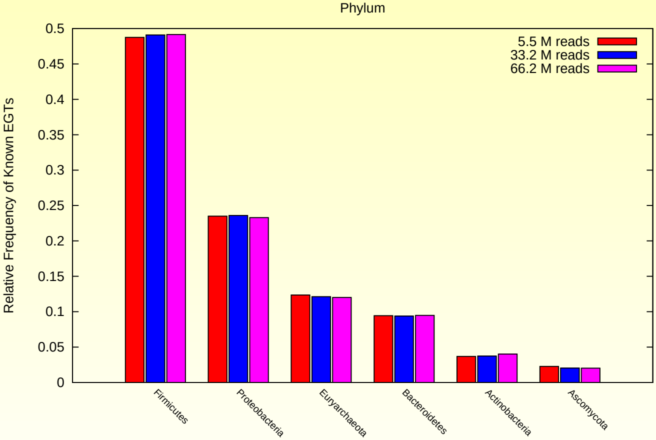
How many 50 bp-reads do we need?

- 1 million?
- 100 million?
- 1 billion?

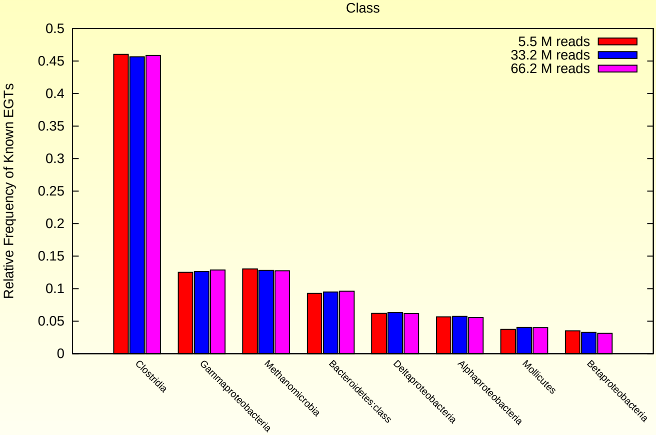
Convergence: (1) Superkingdom



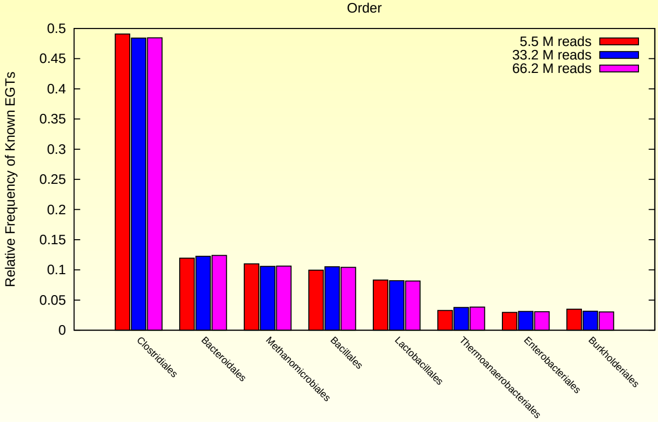
Convergence: (2) Phylum



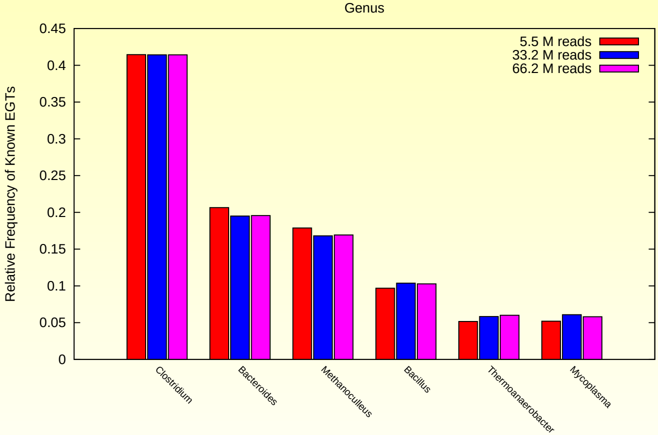
Convergence: (3) Class



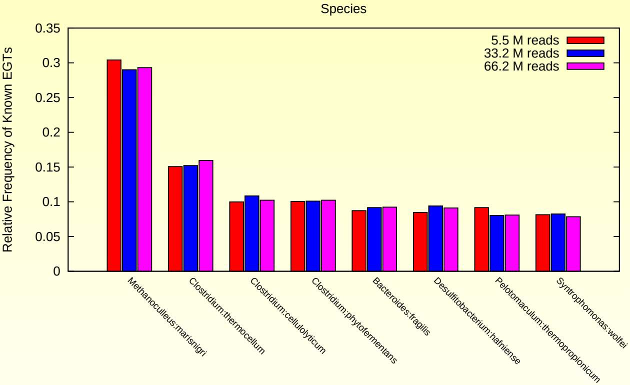
Convergence: (4) Order



Convergence: (5) Genus

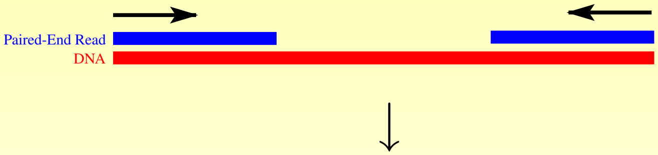


Convergence: (6) Species



Overview

Mate Pairs



Mate Pairs

- `CAGTAGGTCACACAC ... ? ... CGTAGCGGATGCTG`
- more information → better predictions

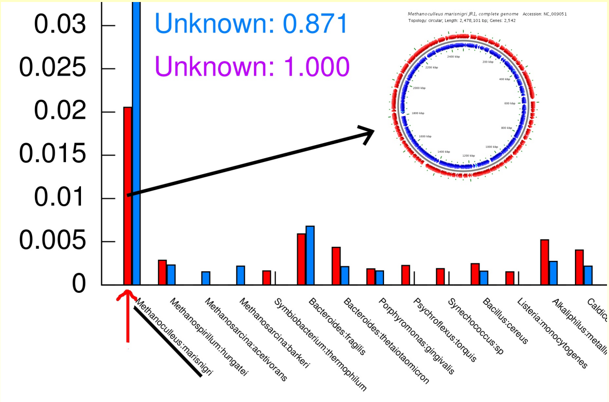
Illumina – Adaptation of protocol

Minimal insert size for paired-end reads

Including Paired-end Information

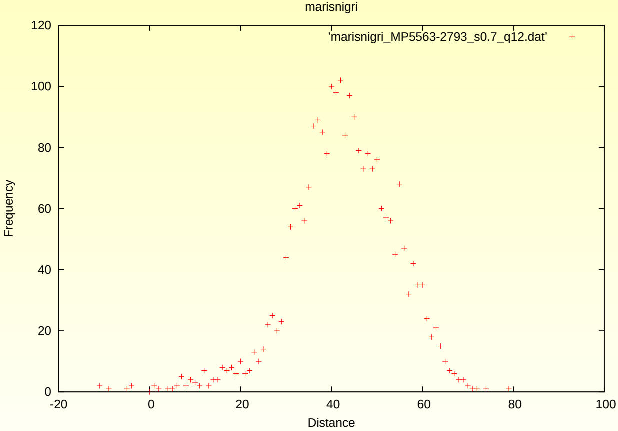
… should increase quality of results!

Question: What is the distance between mates?



Distance between Paired-end Reads

Mapped mate-pair reads classified as Methanoculleus genus to *M. marisnigri* genome.



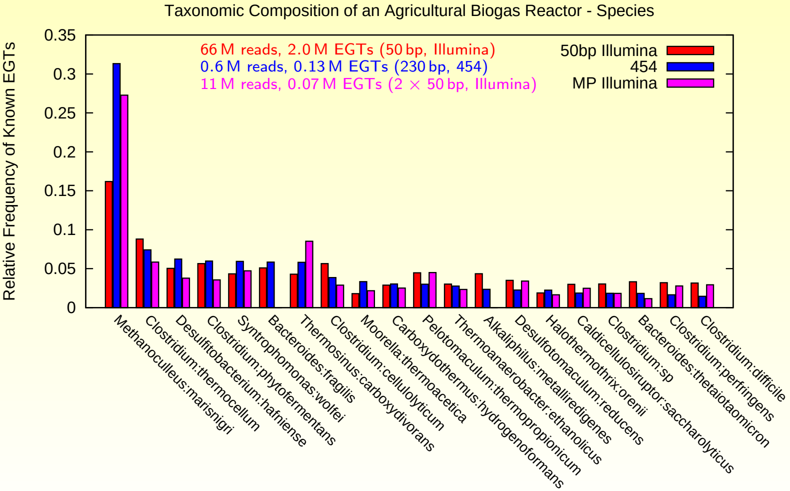
Do both mates match the same domain?

- Analysed: 11 M reads (5.5 M mate pairs)
  - found 448 702 EGTs
- Checked each mate pair (in the set of EGTs) whether both mates match the same Pfam protein family: 87 923 pairs
- $\approx$ 40% of all EGTs belong to a "usable" mate pair where both mates match

$\longrightarrow$ Integration of mate pairs should increase quality of results!

Using Mate Pairs

Early results:



Overview

1. Metagenomics
2. The CARMA Pipeline for Short Read Metagenome Analysis
3. Evaluation and Applications
4. Algorithmic Issues for Very Large Datasets
5. Further Applications
6. Conclusion

Conclusion

- Metagenomics can provide valuable information about microbial communities
- CARMA was successfully applied to ultra-short reads
- Precise quantification at lower ranks is difficult
- Algorithmic challenges
- Implemented Mate Pair usage
- Convergence using increasing number of reads (but: low-complexity data set with only few prevalent species)

Acknowledgments

Thanks to:

- Lutz Krause (CARMA 1.0)
- Wolfgang Gerlach (CARMA 2.x)

Genetics/Bielefeld:

- Andreas Schlüter, Rafael Szczepanowski, Alfred Pühler

Illumina:

- Robert Cohen, Adam Lowe, Keith Moon, Cindy Lawley, Dirk Evers

`http://webcarma.cebitec.uni-bielefeld.de`

Vielen Dank für die Aufmerksamkeit