Next Generation Sequencing Technologies in Microbial Ecology

Frank Oliver Glöckner

JACOBS UNIVERSITY

Max Planck Institute for Marine Microbiology

Investigation of the role, diversity and features of microorganisms

Interactions with physical and chemical processes in marine and other aquatic habitats

Founded 1992 in Bremen, Germany

Marine Microbiology at MPI – a Holistic Approach

Who is out there and

How much of which kind?

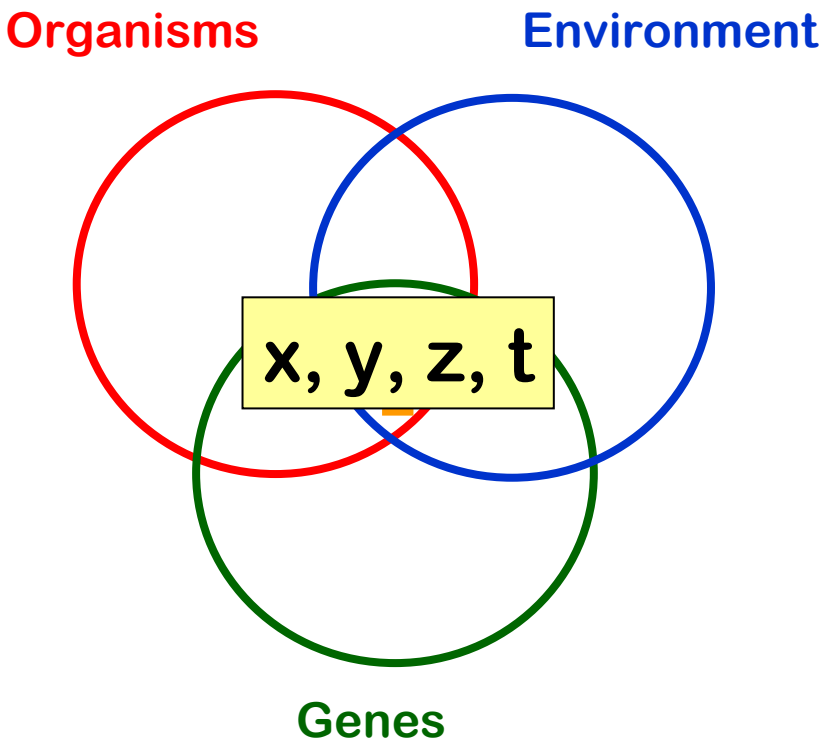
What are they doing and

Under which conditions are they doing what?

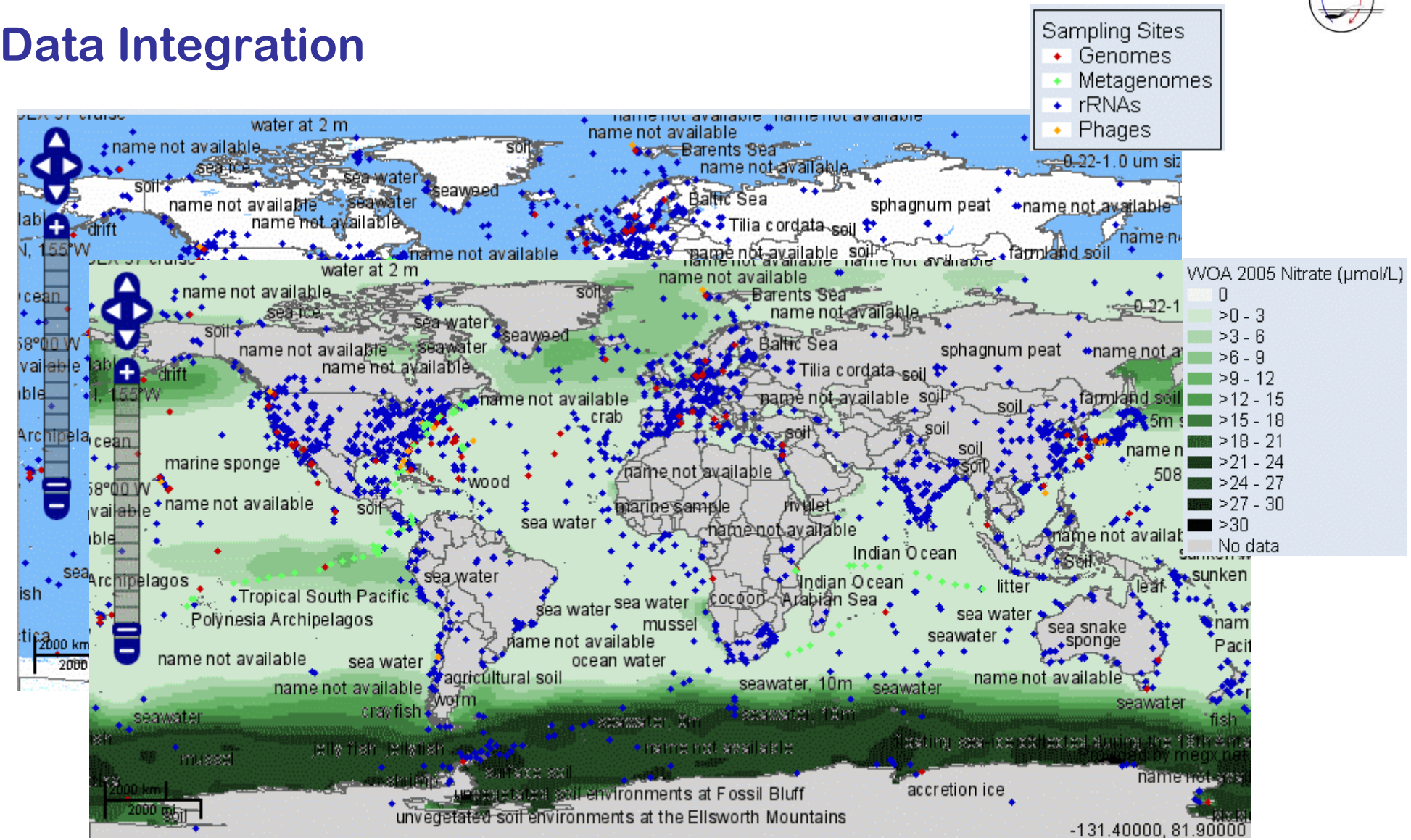
Promise of NGS: Much Denser Network of Data

- Phylogenetic diversity
  - Qualitative data
  - Quantitative data
- Functional diversity
  - Functional inventory
  - Operon structures
  - Expression profiles
- Environmental descriptors

-> Integrated datasets



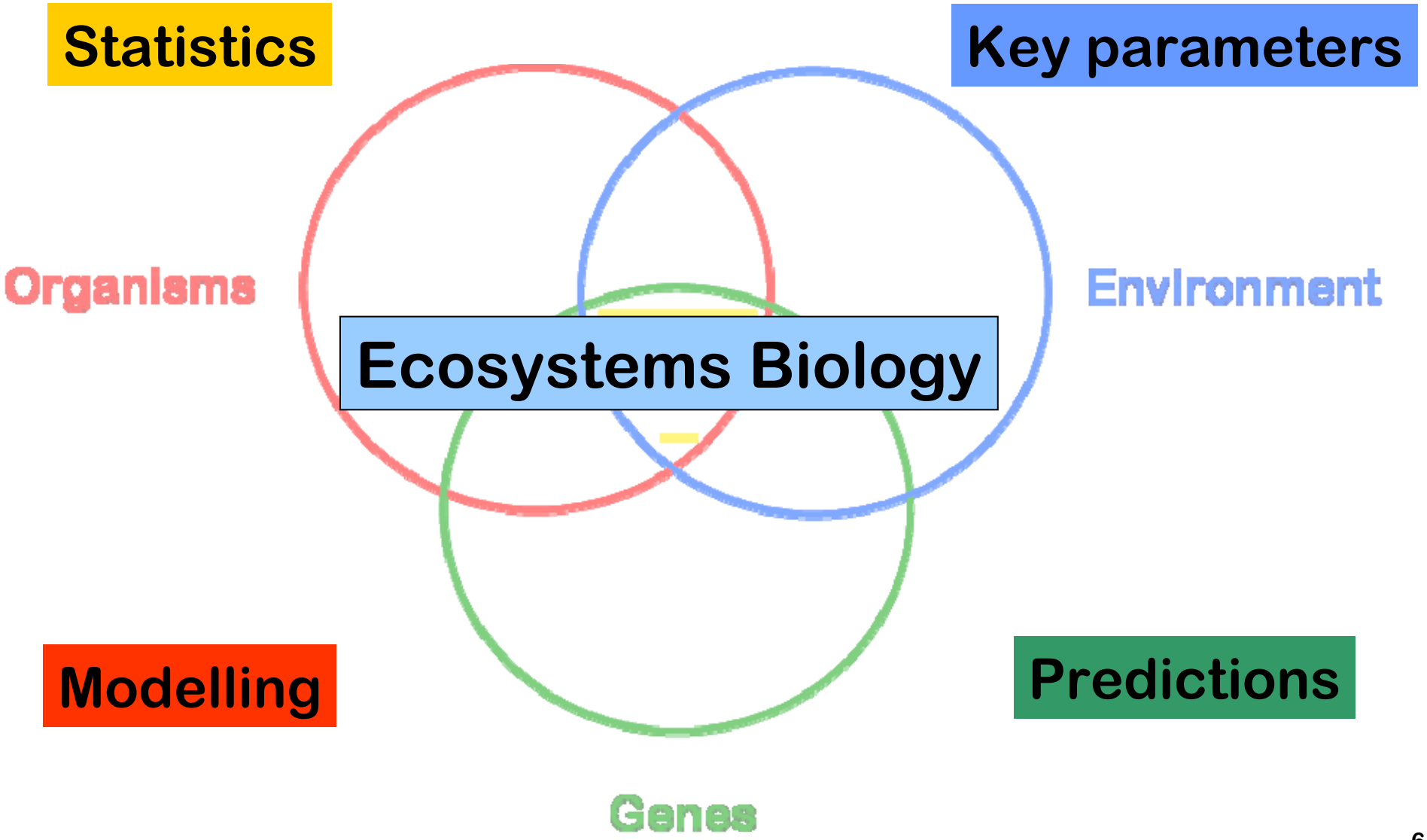
Data Integration



www.megx.net

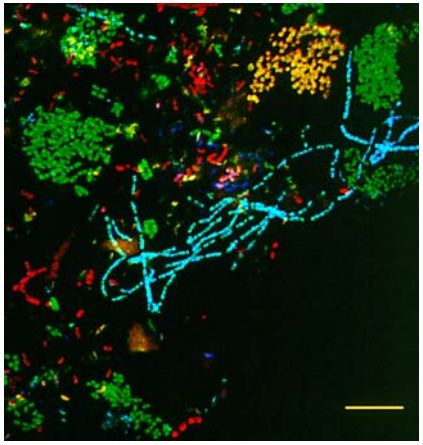
Kottmann et al., NAR, submitted

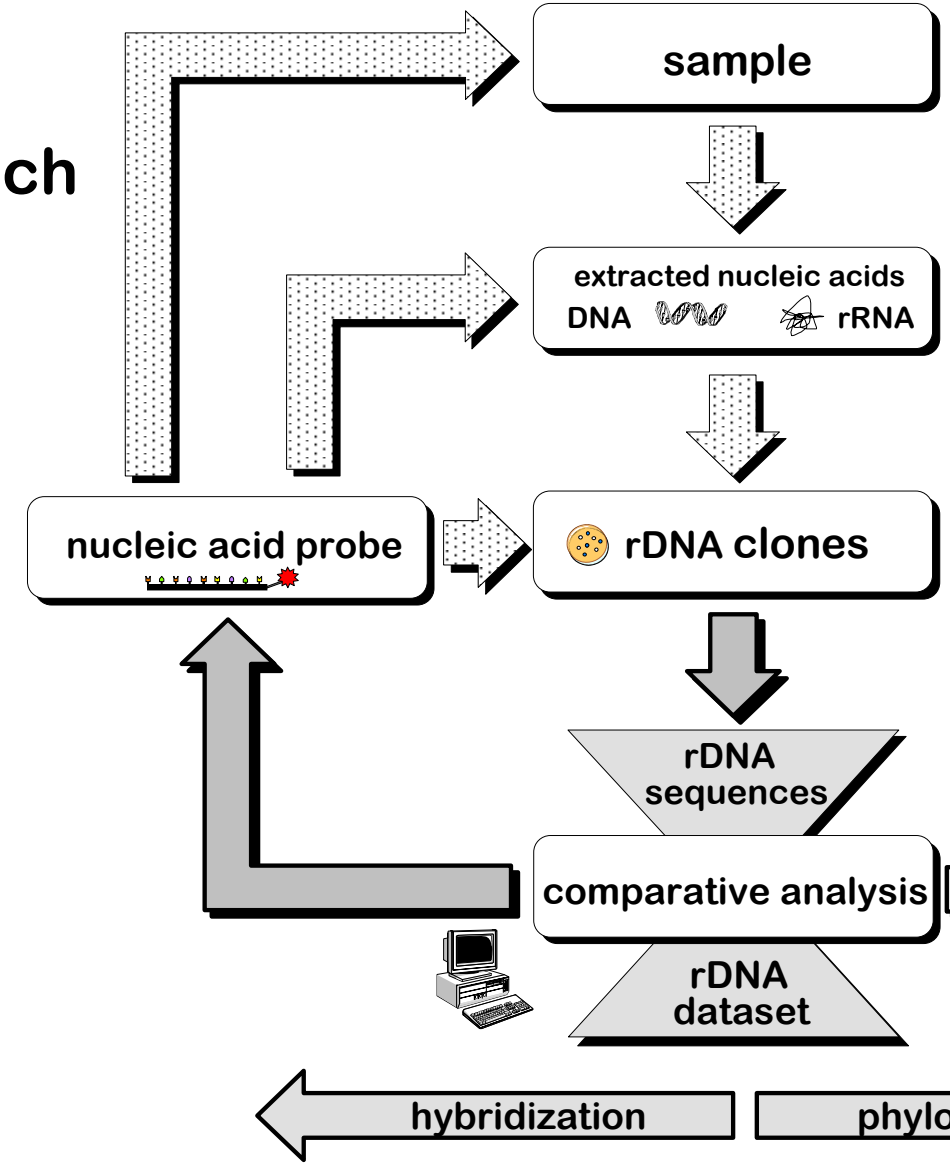
Ecological Genomics – The Vision

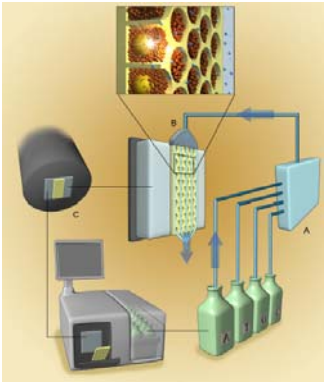


Ribosomal RNA as a universal marker gene

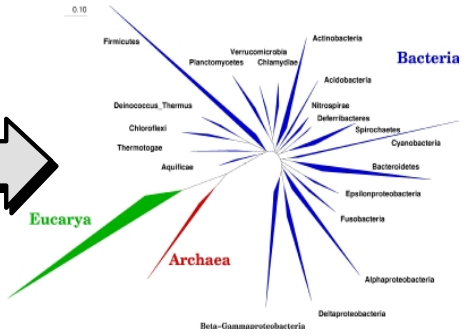
Full cycle rRNA-approach







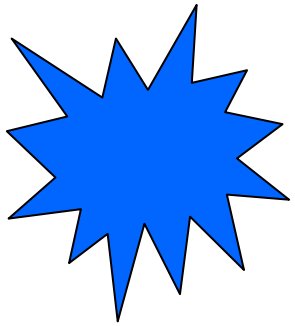
Pyrosequencing



Amann, 1995

Diversity Analysis

Sample



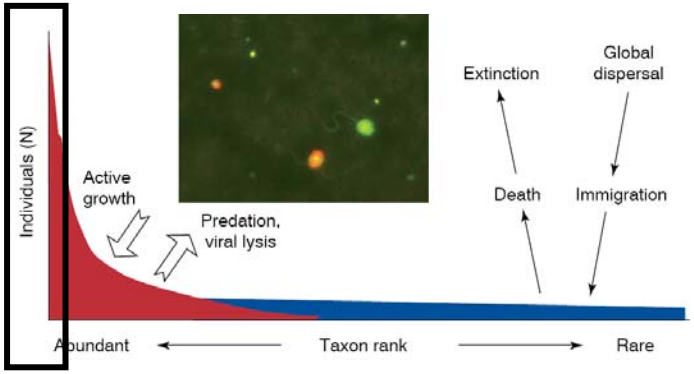
High diversity

PCR

Clone lib

100-500

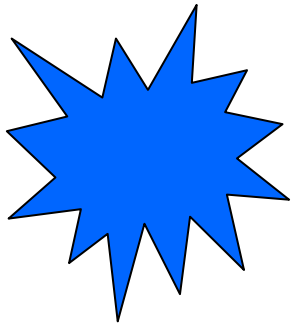
2-3 month



Pedros-Alio, Trends in Microbiology, 2006, vol. 12, issue 6, page 257

Diversity Analysis

Sample



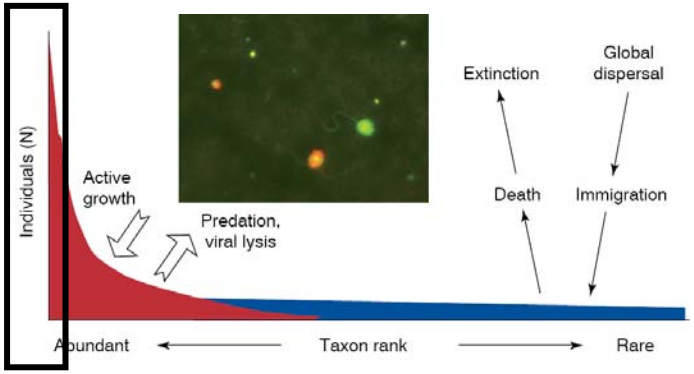
High diversity

PCR

Clone lib

100-500

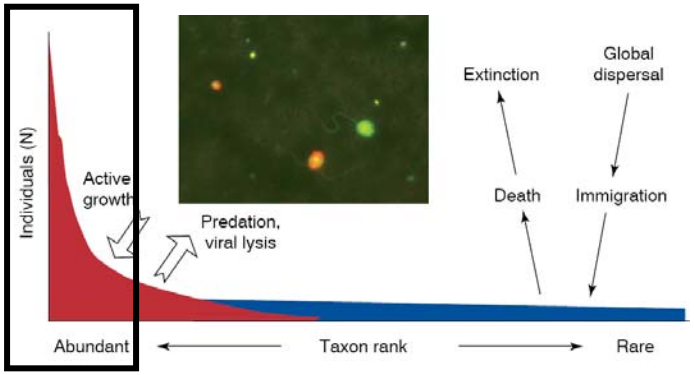
2-3 month



Tags

10,000-50,000

1 week



Problems

- **Processing the data**

Your next-generation sequencing data just arrived...

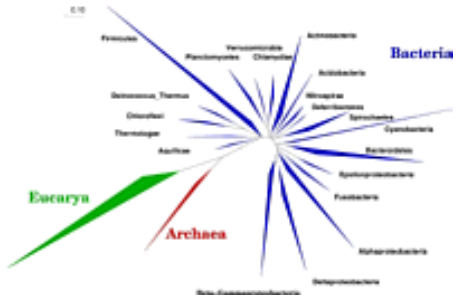
Now What?

Richard was having a great day,
until the arrival of his next-generation data files.

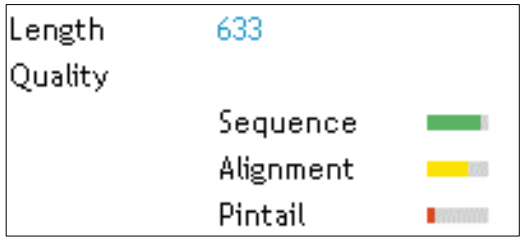
- Accuracy/Quantitative?
  - DNA/RNA extraction
  - Multiple Operons
  - Technical replicates
  - Noise (sequencing errors…)

SILVA Databases Specifications

- Comprehensive & Aligned
  - *Bacteria, Archaea, Eukarya*
  - SSU, LSU
  - Regularly updated

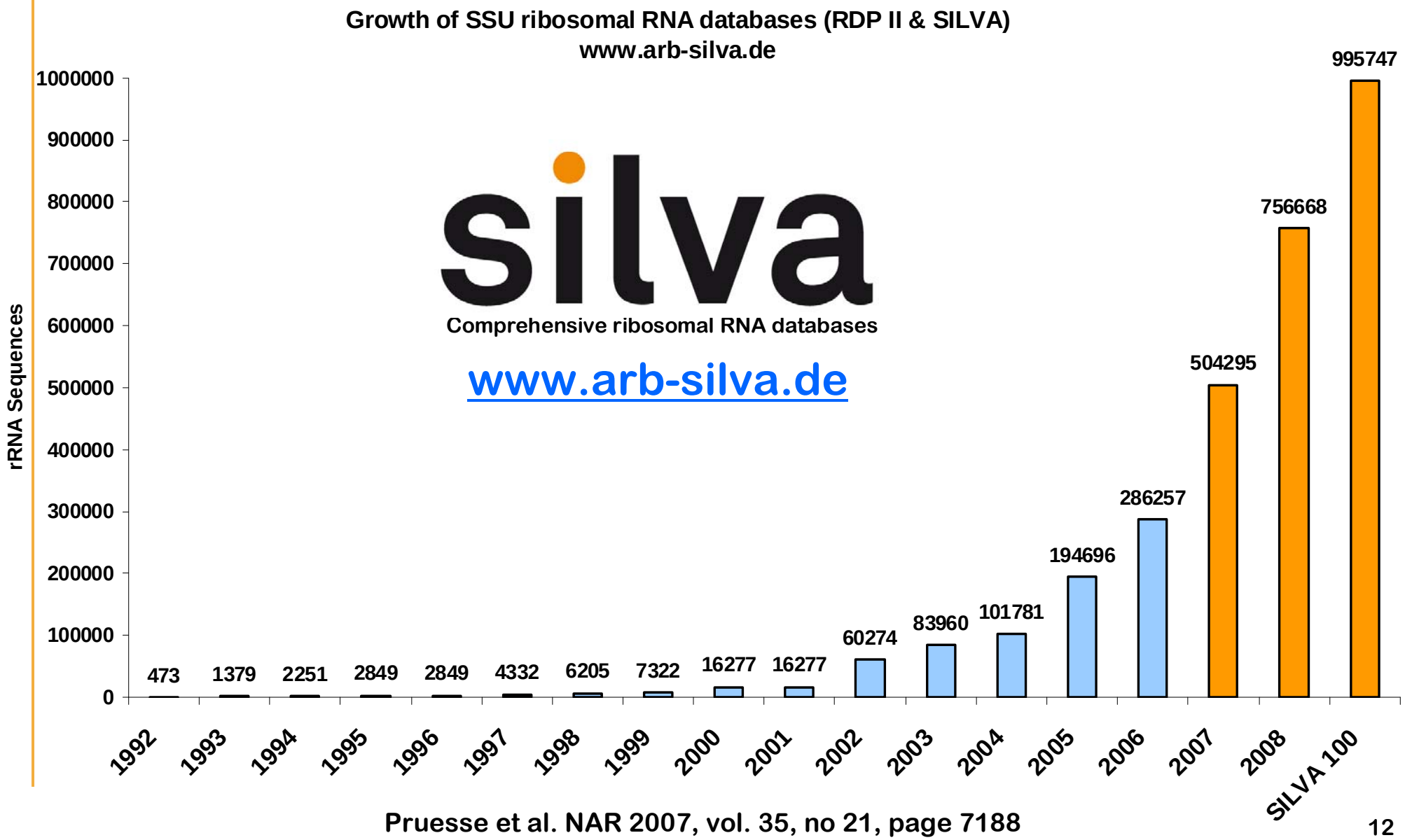


- Quality first
  - Quality management
  - Transparent process documentation



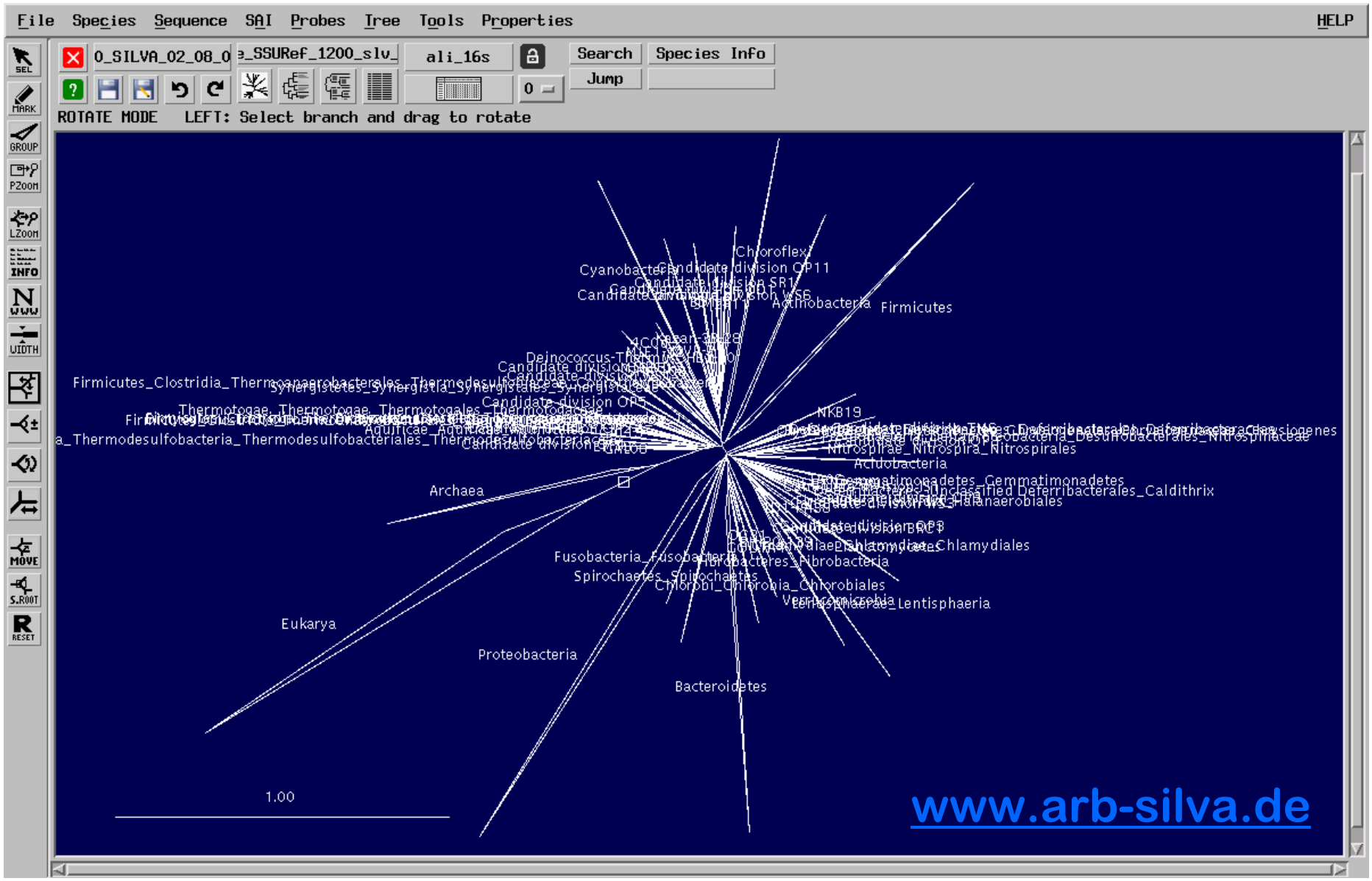
- Integrative
  - Nomenclature
  - Taxonomy
  - Cultured, Typestrains
  - Habitat (r100)

Growth of rRNA databases (RDP & SILVA)

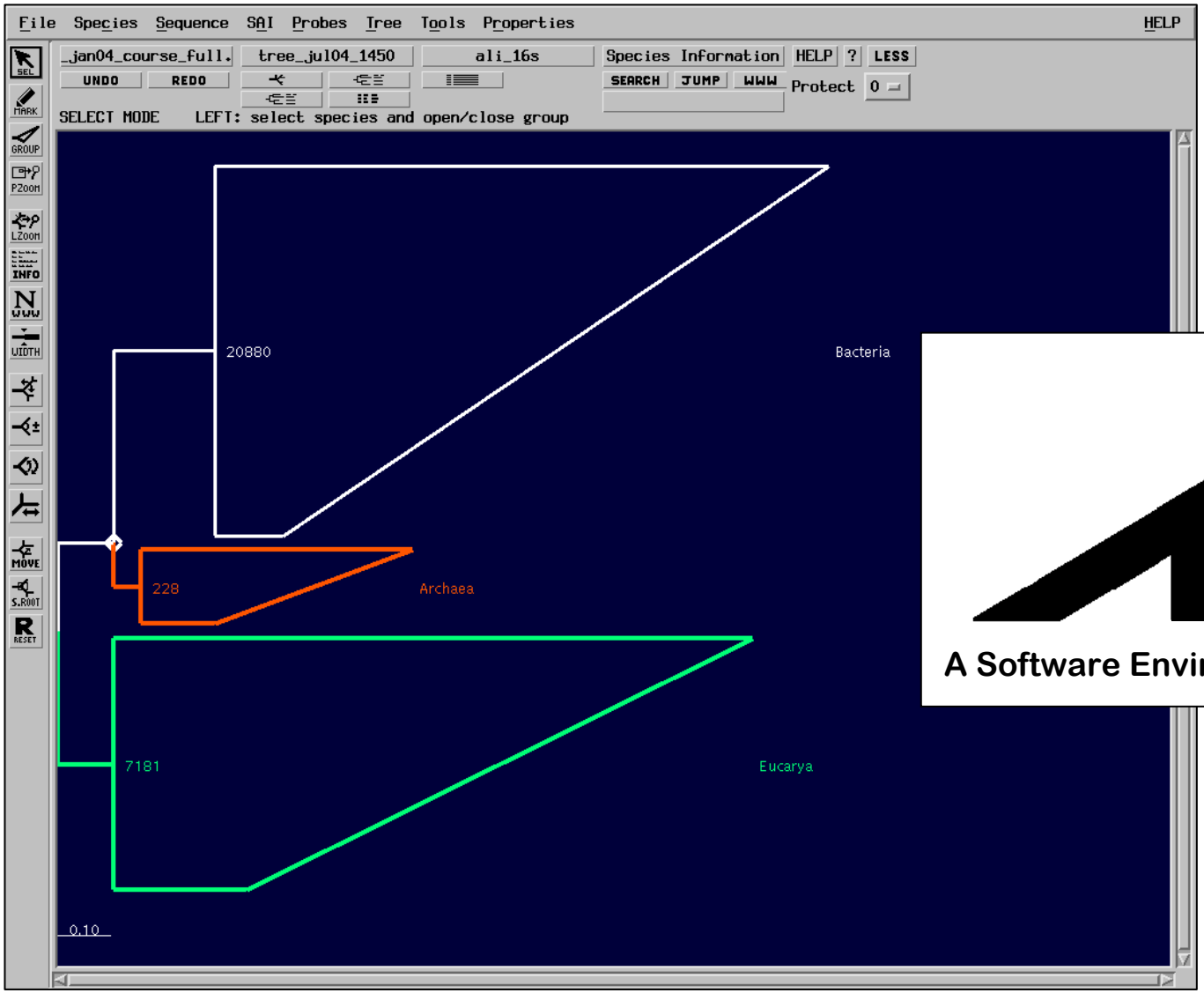


Pruesse et al. NAR 2007, vol. 35, no 21, page 7188

SILVA SSURef 100: Fully classified guide tree



ARB Software Suite



www.arb-home.de

TECHNISCHE UNIVERSITÄT MÜNCHEN

A Software Environment for Sequence Data

Ludwig et al. NAR, 2004

ARB 5.0, 64 bit version released on 04. September 2009

Problems

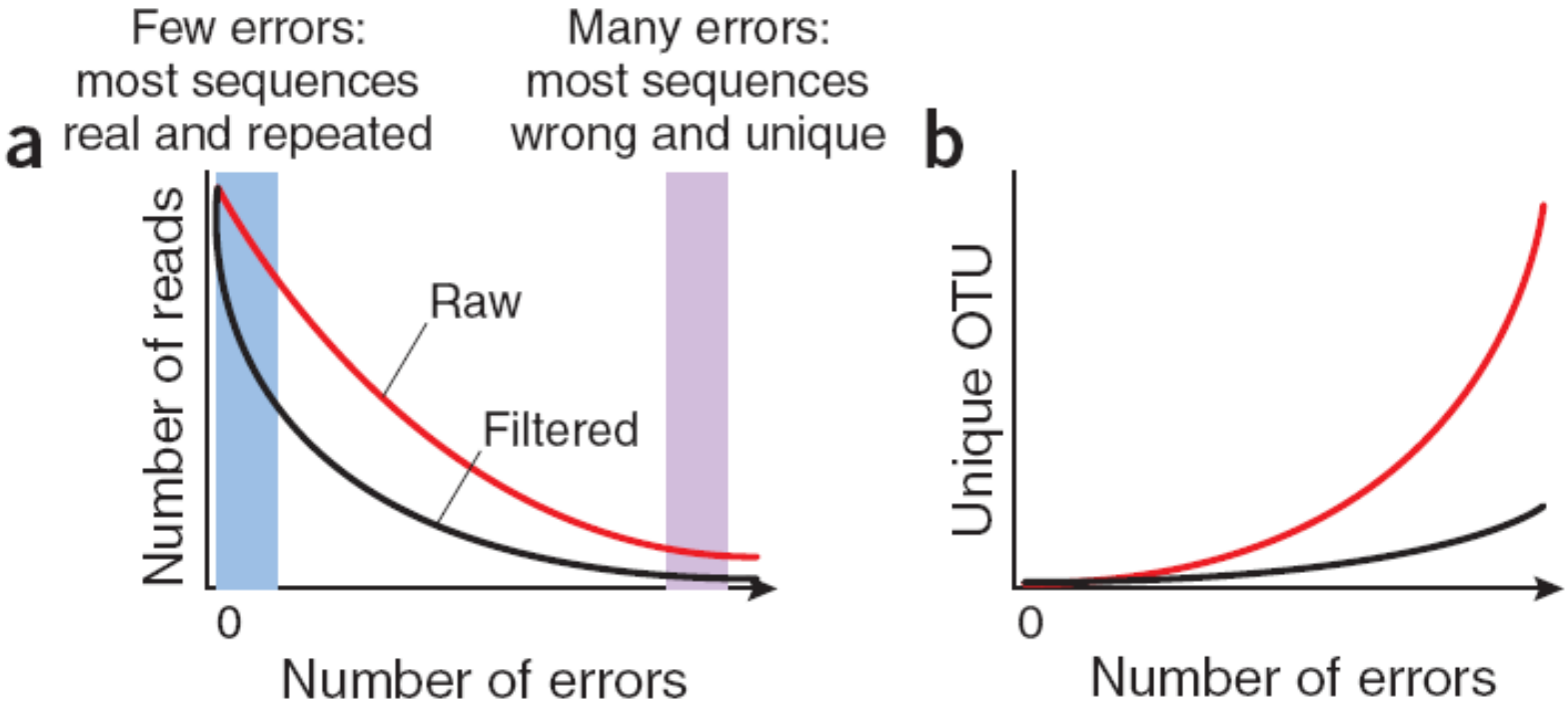
- Processing the data

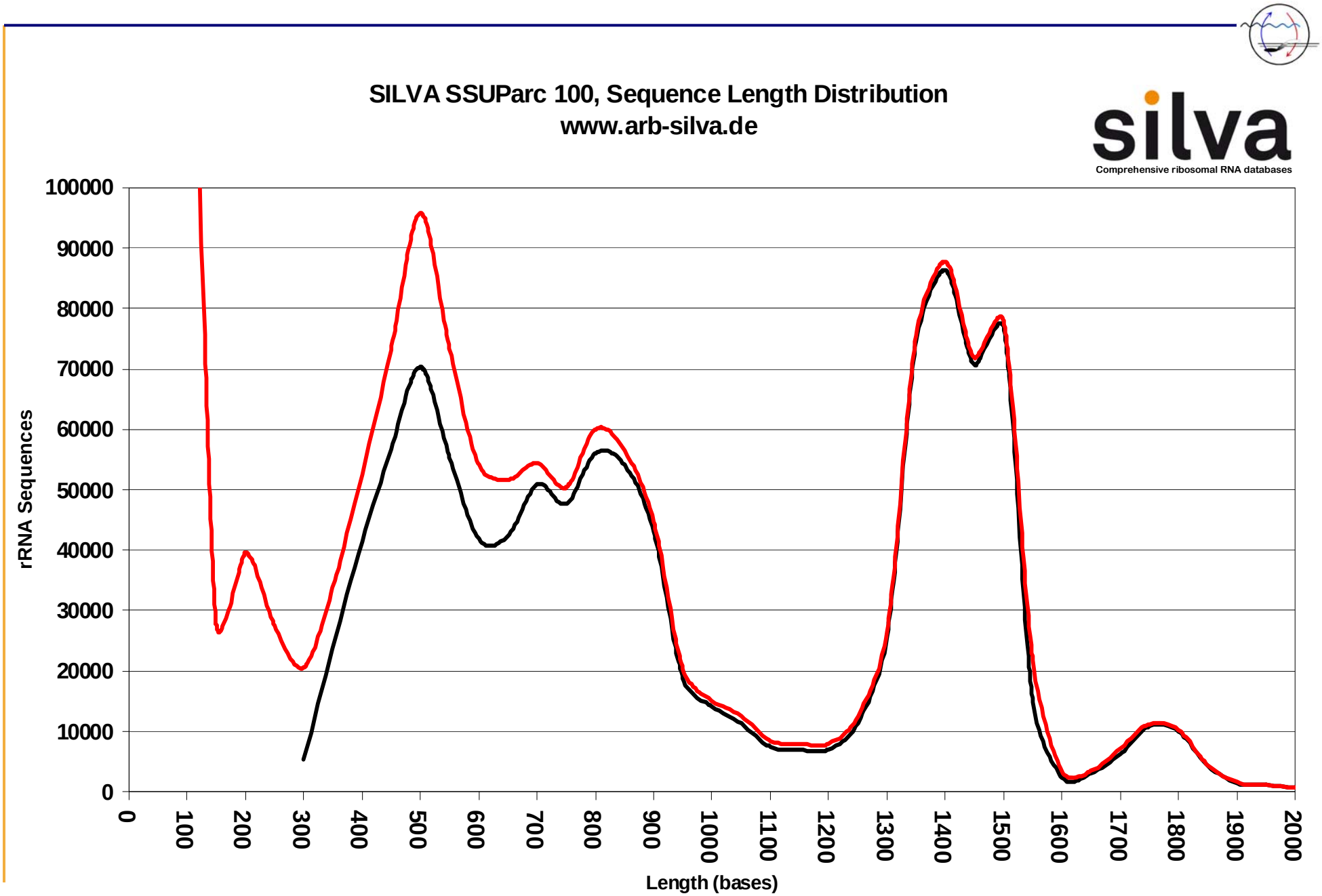
- Accuracy/Quantitative?
  - DNA/RNA extraction
  - Multiple Operons
  - Noise (sequencing errors…)
  - Technical replicates

Accuracy

The ‘rare biosphere’: a reality check

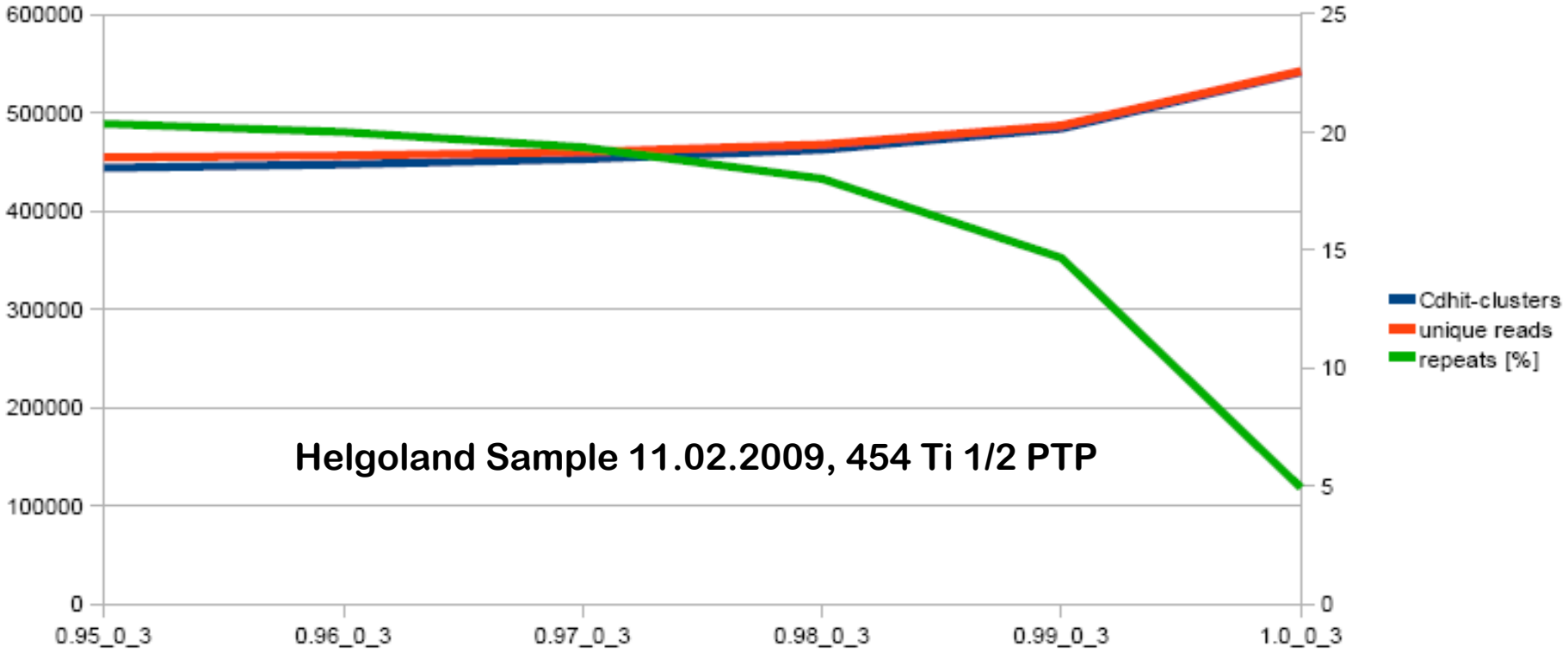
Reeder and Knight, 2009, Nature Methods vol. 6, no. 9, p. 636




SILVA SSUParc 100, Sequence Length Distribution
www.arb-silva.de
silva
Comprehensive ribosomal RNA databases
rRNA Sequences
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
Length (bases)

Technical Replicates I

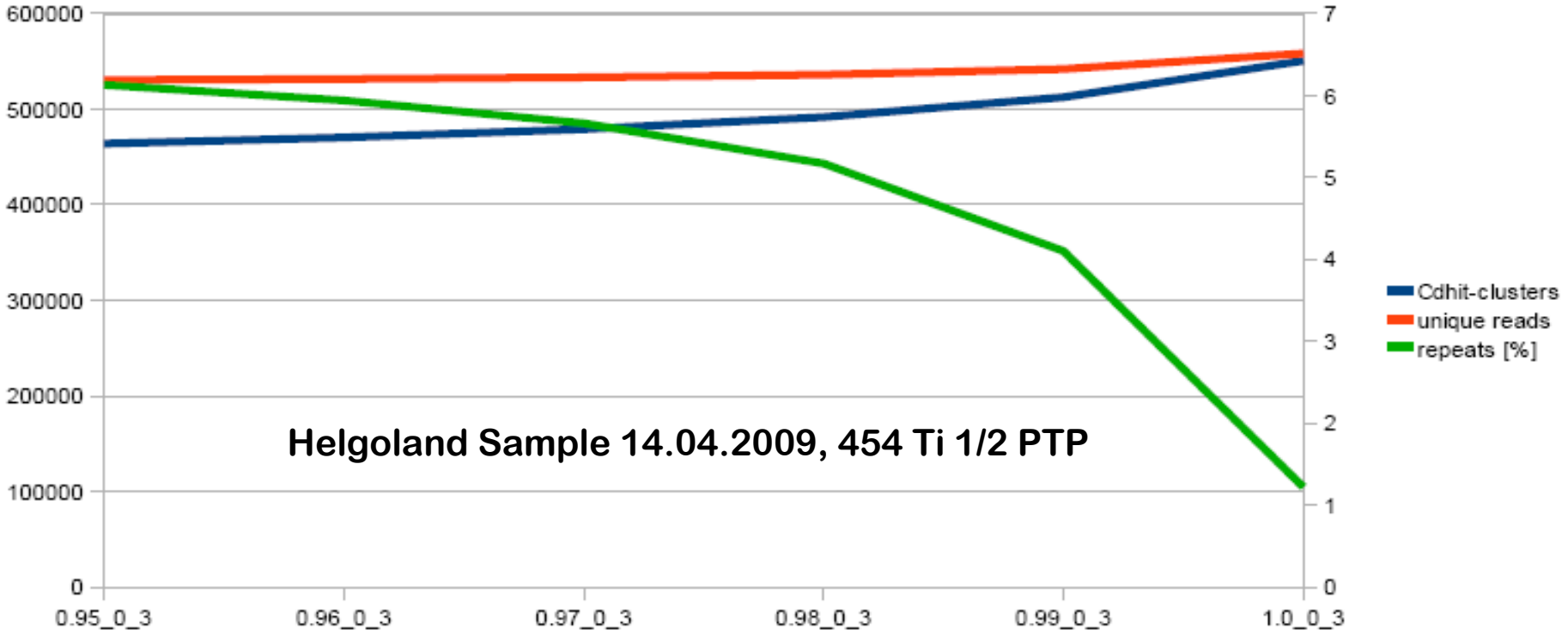
name	Cdhit-clusters	unique reads	repeats [%]	identity cutoff	length difference req.	initial base pair req.
0.95_0_3	443508	454154	20.35	0.95	0	3
0.96_0_3	447004	456150	20	0.96	0	3
0.97_0_3	452473	459824	19.35	0.97	0	3
0.98_0_3	461977	467396	18.02	0.98	0	3
0.99_0_3	483491	486594	14.66	0.99	0	3
1.0_0_3	540920	542038	4.93	1	0	3



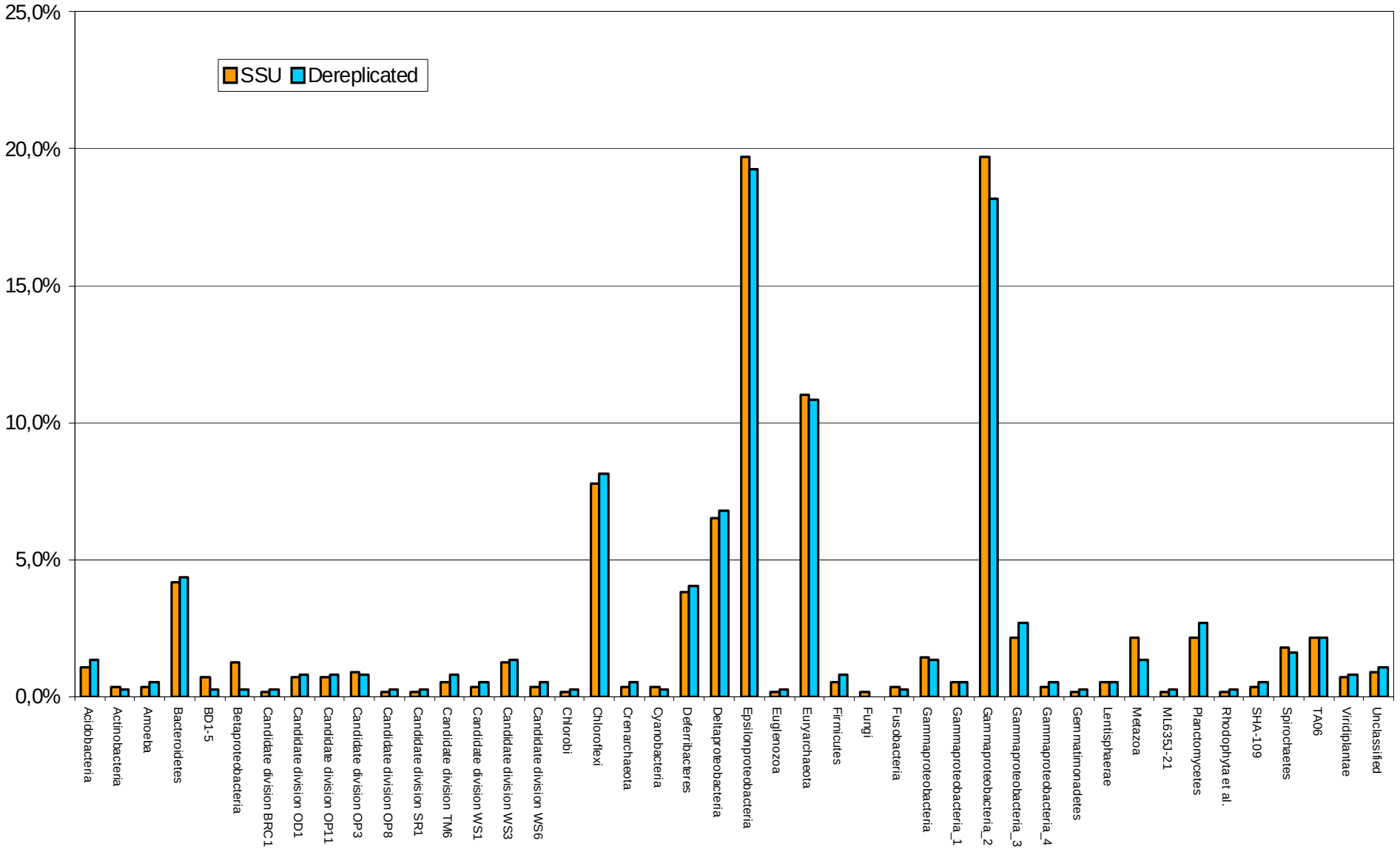
Gomez-Alvarez et al., 2009, ISME Journal, pages 1-4

Technical Replicates II

name	Cdhit-clusters	unique reads	repeats [%]	identity cutoff	length difference req.	initial base pair req.
0.95_0_3	463763	530516	6.13	0.95	0	3
0.96_0_3	470233	531581	5.94	0.96	0	3
0.97_0_3	479141	533202	5.66	0.97	0	3
0.98_0_3	491641	535978	5.17	0.98	0	3
0.99_0_3	512518	542010	4.1	0.99	0	3
1.0_0_3	550538	558287	1.22	1	0	3



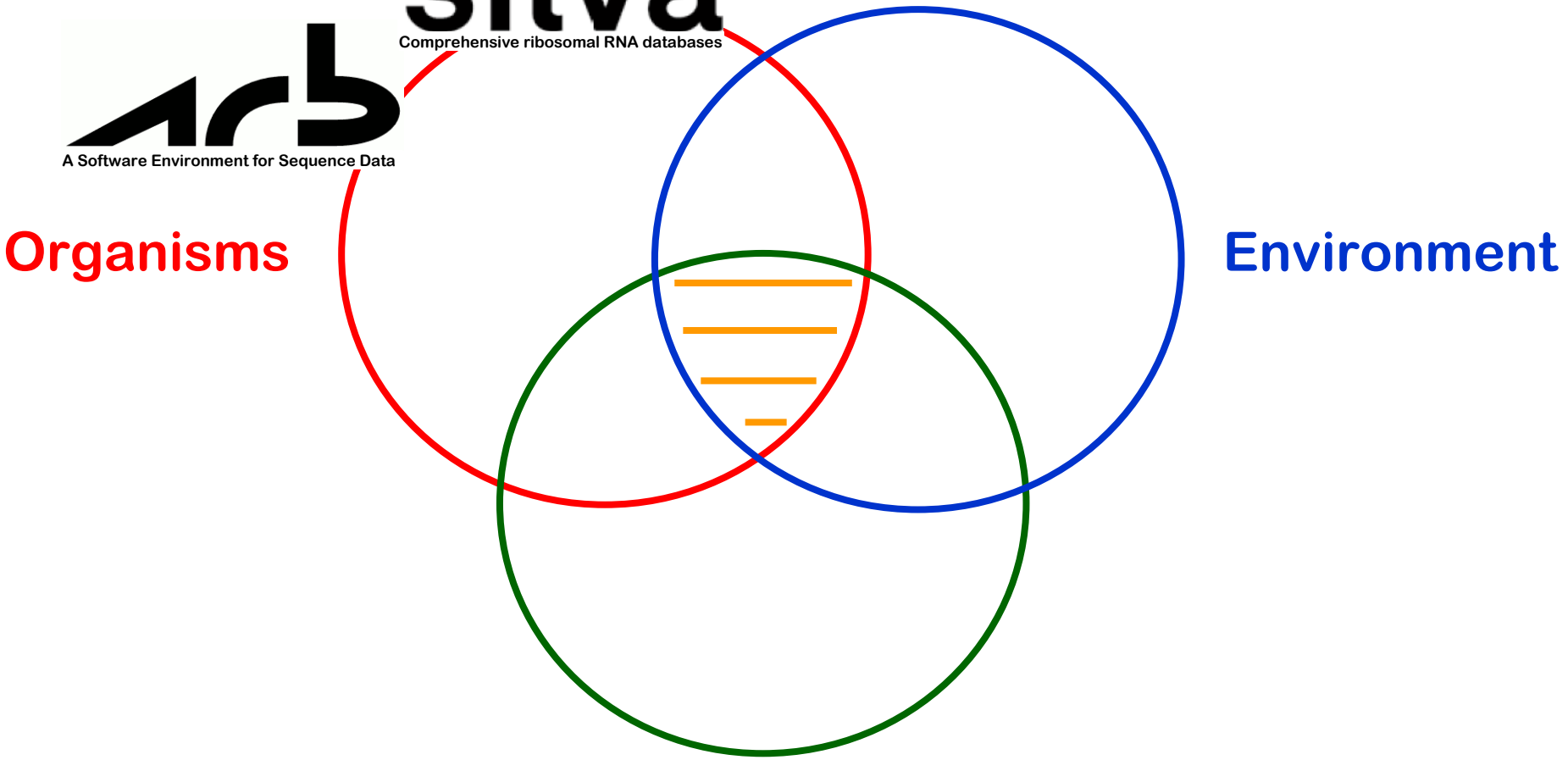
Technical Replicates III - Dereplication



A Bioinformatic Workbench for Ecological Genomics

silva
Comprehensive ribosomal RNA databases

arb
A Software Environment for Sequence Data



Functional Diversity Analysis

Sample

DNA

Fosmids

40,000

2-3 month

Random sequencing

10-100 ~ 400-4000 ORFs

NGS

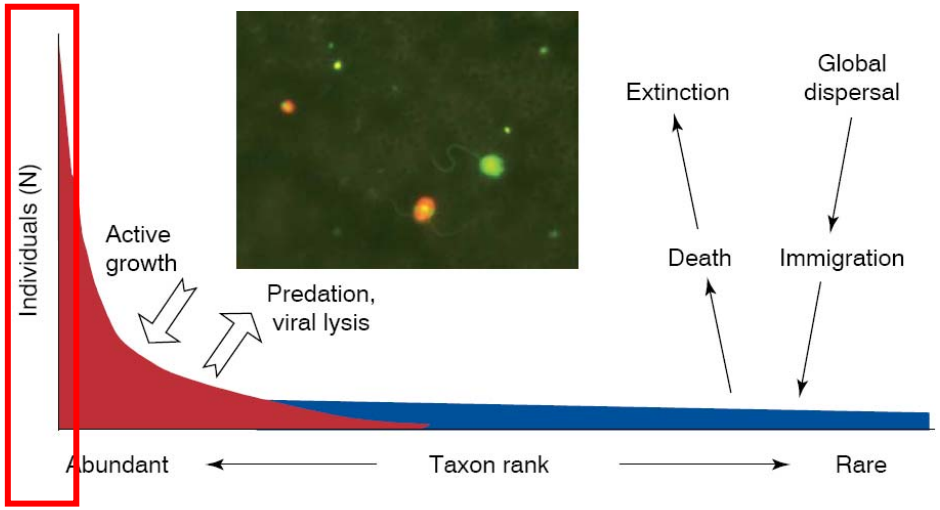
1000

40,000 ORFs

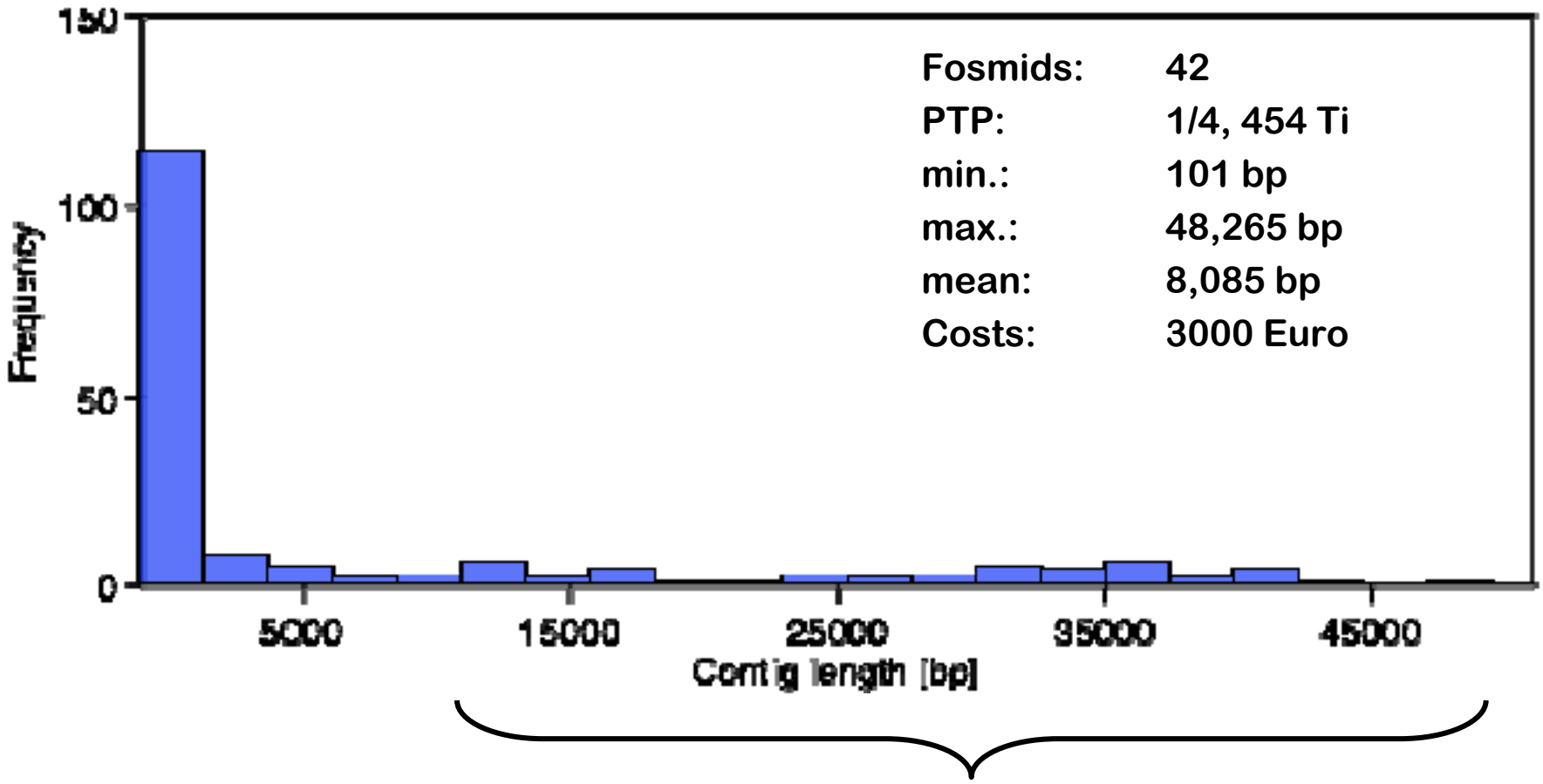
End sequencing

20,000 'ORFs'

High diversity



Fosmid Sequencing



48 contigs > 10 kb

Functional Diversity Analysis

Sample

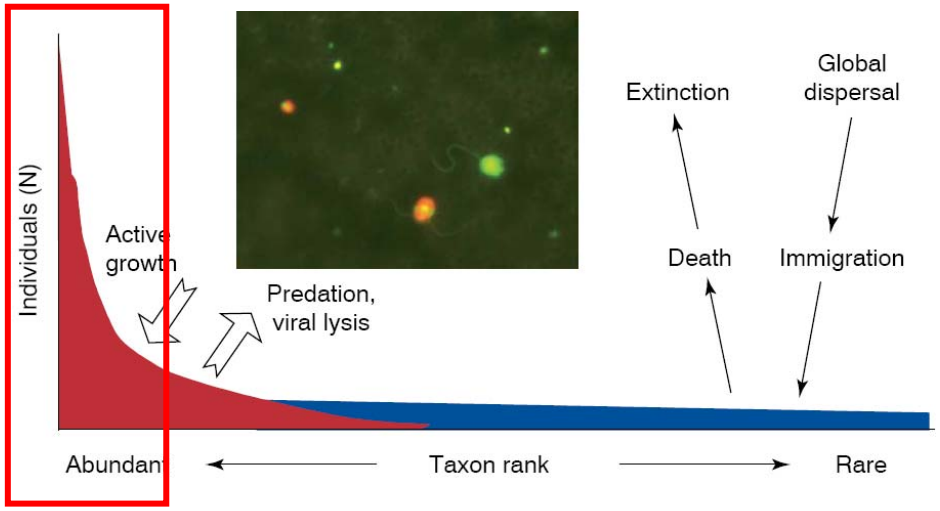
DNA

Tags

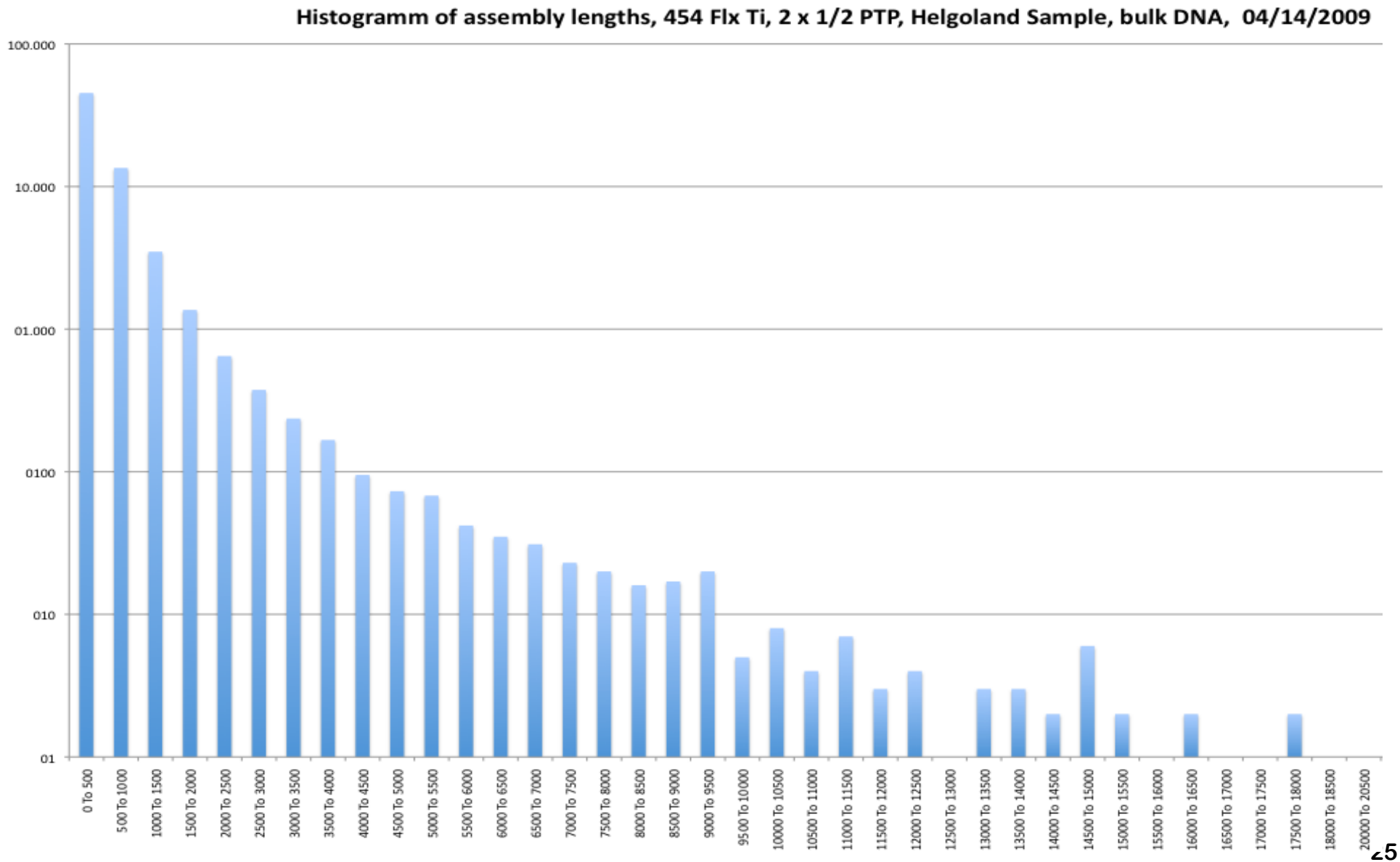
500,000 - 1 Mio
1 week

Statistics
BLAST
COGs
Classify
Assembly?

High diversity



Assembly


Histogramm of assembly lengths, 454 Flx Ti, 2 x 1/2 PTP, Helgoland Sample, bulk DNA, 04/14/2009
100.000
10.000
01.000
0100
010
01
0 To 500
500 To 1000
1000 To 1500
1500 To 2000
2000 To 2500
2500 To 3000
3000 To 3500
3500 To 4000
4000 To 4500
4500 To 5000
5000 To 5500
5500 To 6000
6000 To 6500
6500 To 7000
7000 To 7500
7500 To 8000
8000 To 8500
8500 To 9000
9000 To 9500
9500 To 10000
10000 To 10500
10500 To 11000
11000 To 11500
11500 To 12000
12000 To 12500
12500 To 13000
13000 To 13500
13500 To 14000
14000 To 14500
14500 To 15000
15000 To 15500
15500 To 16000
16000 To 16500
16500 To 17000
17000 To 17500
17500 To 18000
18000 To 18500
20000 To 20500

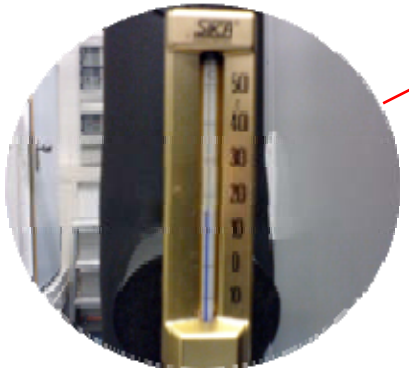
Are we prepared for the Data Flood?

Your next-generation sequencing data just arrived...

Now What?

Richard was having a great day,
until the arrival of his next-generation data files.

Computing Infrastructure



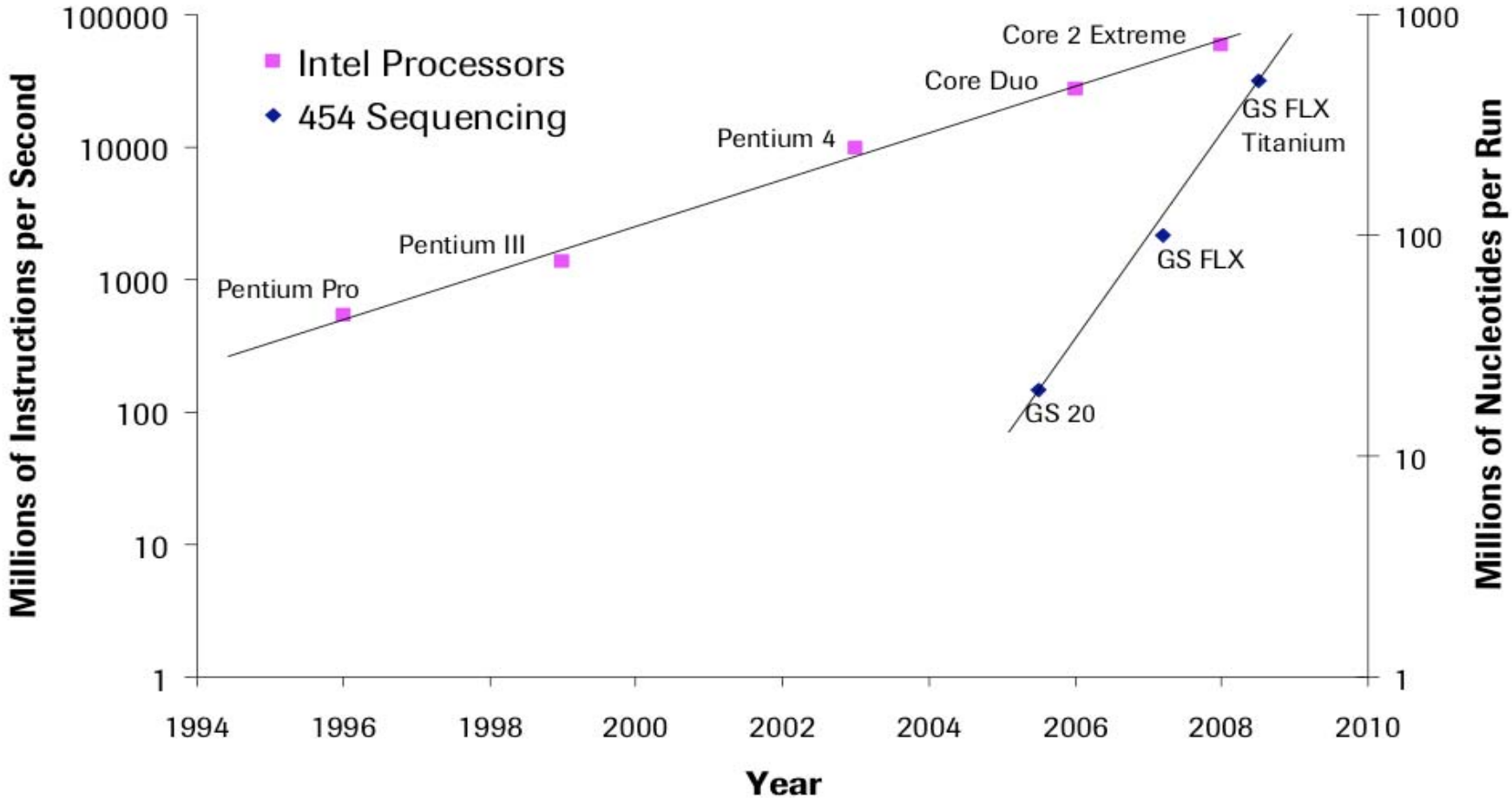
Cooling
- Liquid cooling with 5,000 L/h at 8° C

Power
- 25 - 30 kW at peak
- installed: 40 kVA

Storage
- 8 TByte RAID file server
- 4 TByte RAID database server

Computers
- 43 cluster nodes
- several larger servers
- 400 CPU cores

Moore's Law - Outcompeted



www.roche-applied-science.com

Take Home Message for the Next Generation Biologists

- Three languages!
  - Mother tongue
  - English
  - Perl, Phyton…

  - Data management
    - Garbage in -> garbage out!

  - Standardisation

Genomic Standards Consortium

www.gensc.org

The Group

http://www.microbial-genomics.de

Thanks for your attention