

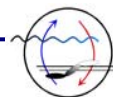
Extending MIGS/MIMS for ribosomal RNA sequences



Frank Oliver Glöckner
Metagenomics 2008

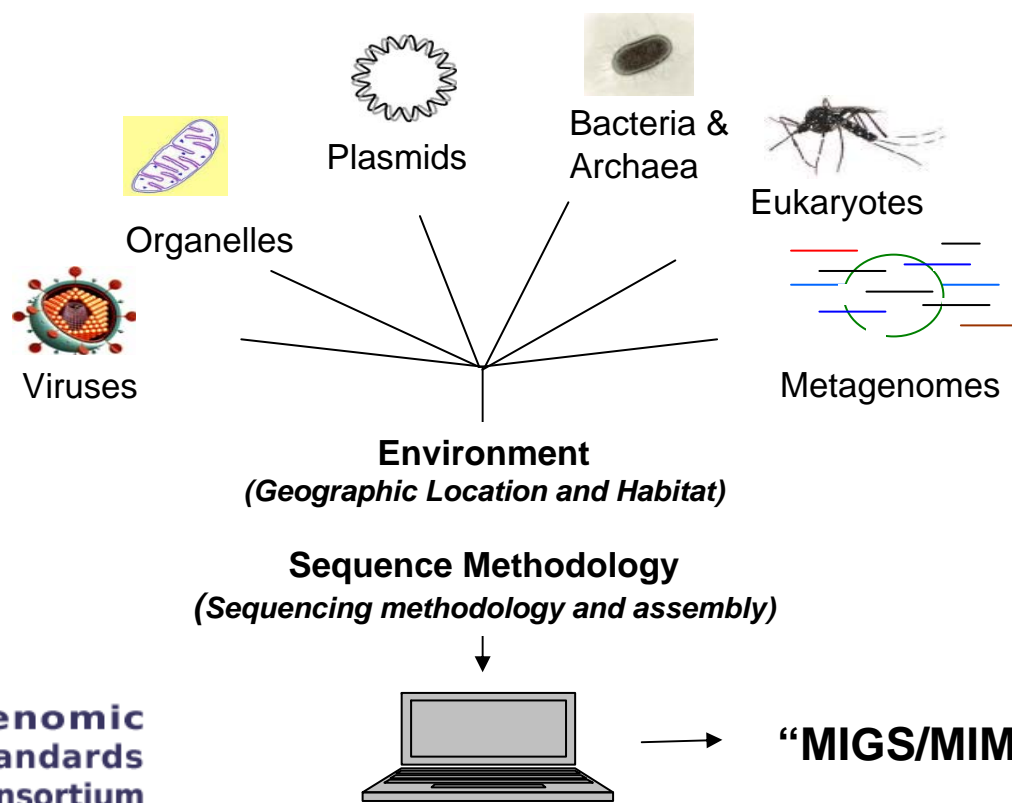


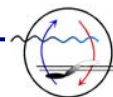
JACOBS
UNIVERSITY



GSC Goals

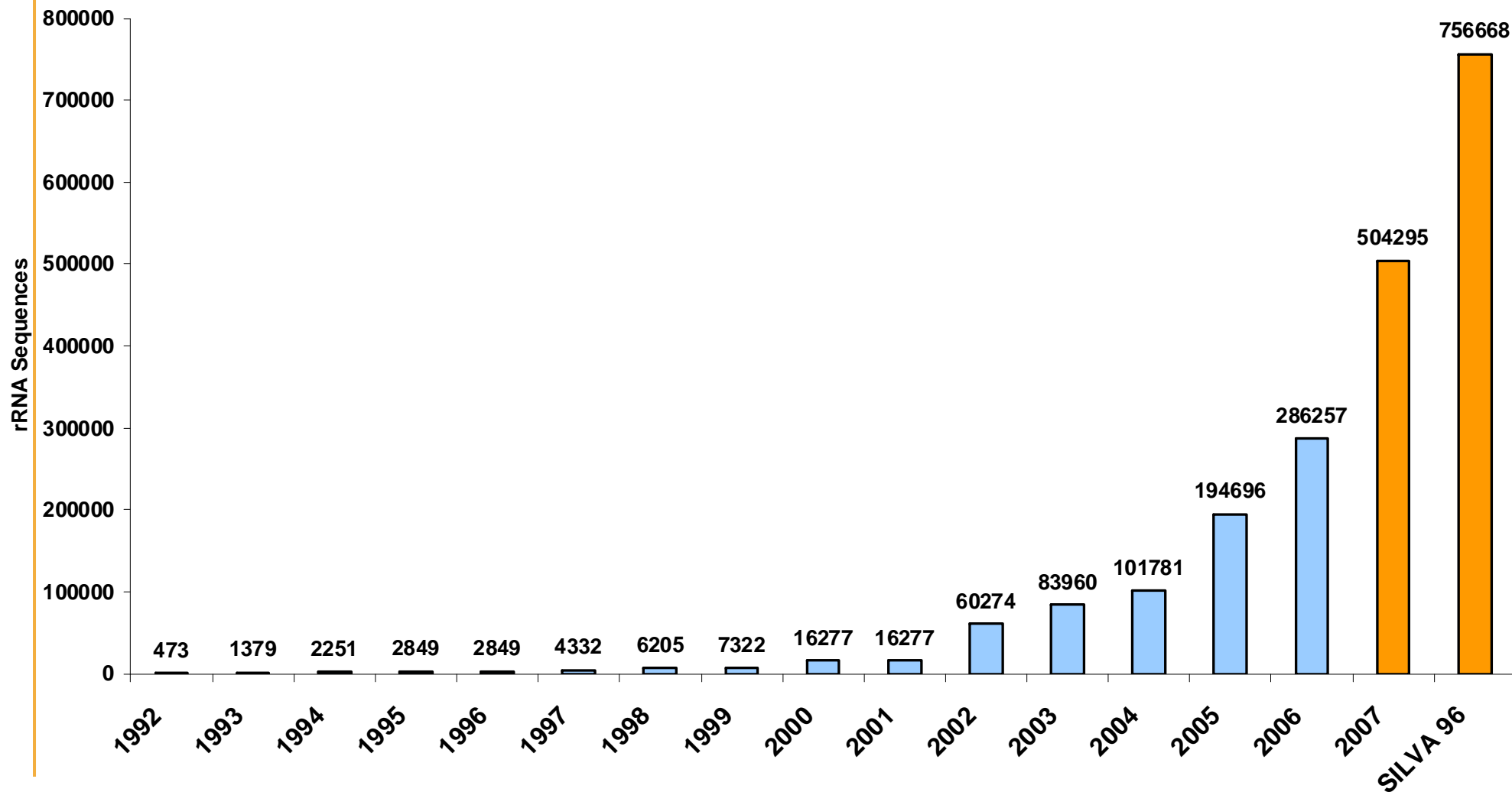
- ▶ Promote mechanisms that standardize
 - the description of **genomes and metagenomes**
 - and the exchange and integration of genomic data

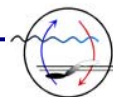




Growth of rRNA databases (RDP & SILVA)

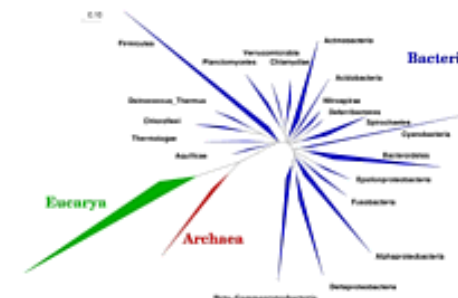
Growth of SSU ribosomal RNA databases (RDP II & SILVA)
www.arb-silva.de





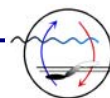
SILVA Databases Specifications

- ▶ **Comprehensive & Aligned**
 - *Bacteria, Archaea, Eukarya*
 - SSU, LSU
 - Regularly updated
- ▶ **Quality first**
 - Quality management
 - Transparent process documentation
- ▶ **Integrative**
 - Nomenclature
 - Taxonomy
 - Cultured, Typestrains



Length	633
Quality	
Sequence	
Alignment	
Pintail	





Contextual data??

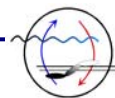
Salimicrobium album

General information

Accession number	X90834
Organism name	Salimicrobium album
Alternative name	bas: Marinococcus albus;
Strain	DSM 20748 s[T] r[T] l[T] StrainInfo.net
Isolate	--
Clone	--
Length	1489
Date created	1996-08-06
Date modified	2004-01-09
Version	1
Taxonomy EMBL	Bacteria ▶ Firmicutes ▶ Bacillales ▶ Sporolactobacillaceae ▶ Marinococcus
Taxonomy Greengenes	Bacteria ▶ Firmicutes ▶ Alicyclobacillaceae ▶ Bacilli ▶ Halobacillus ▶ Unclassified
Taxonomy RDP	Bacteria ▶ Firmicutes ▶ Bacilli ▶ Bacillales ▶ Sporolactobacillaceae ▶ Marinococcus

Environment

Country	--
Latitude/Longitude	--
Collection date	--
Isolation source	--
Specific host	--
Collected by	--

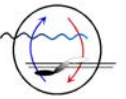


Contextual data??

Content of the SILVA 96 SSU Ref database: **324,342 entries**

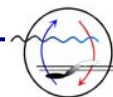
Number of entries with information in the field

“isolation_source”:	229,180	(70%), water, soil, faeces...
“country”:	93,643	(29%), Brazil, China, Canada...
“specific_host”:	69,206	(21%), rice, tomato, rabbit...
“collection_date”:	21,088	(7%), 2000, 2004, Dec-2001...
“lat_lon”:	11,768	(4%)

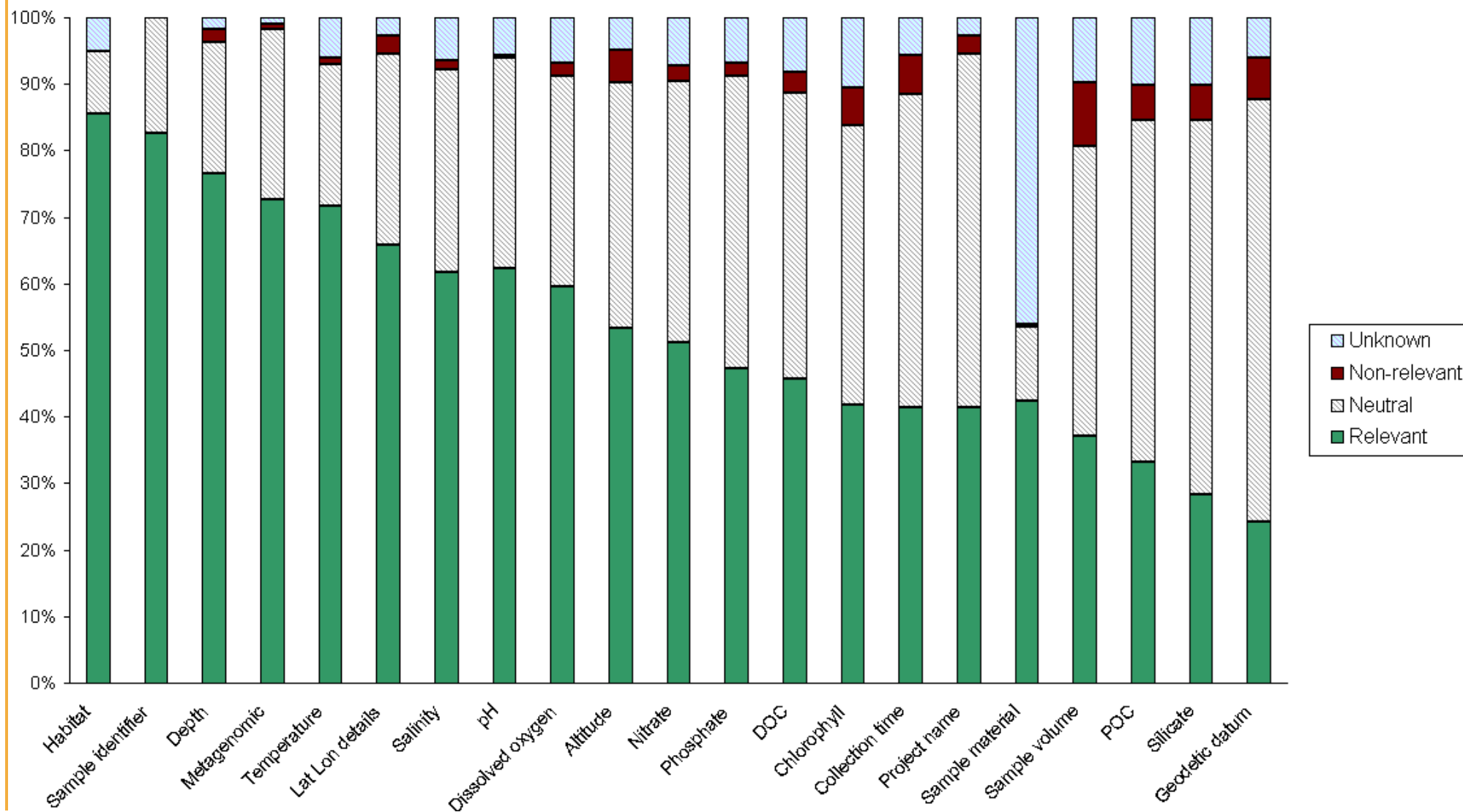


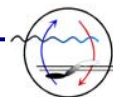
Motivation!

- ▶ **Minimal Information for an ENvironmental Sequence (MIENS)**
 - **Which attributes for environmental sequences are most relevant for the community?**
 - → surveys
 - **How can data integration and sequence submission to the INSDC (GenBank, EMBL, DDBJ) effectively be handled?**
 - → workflow

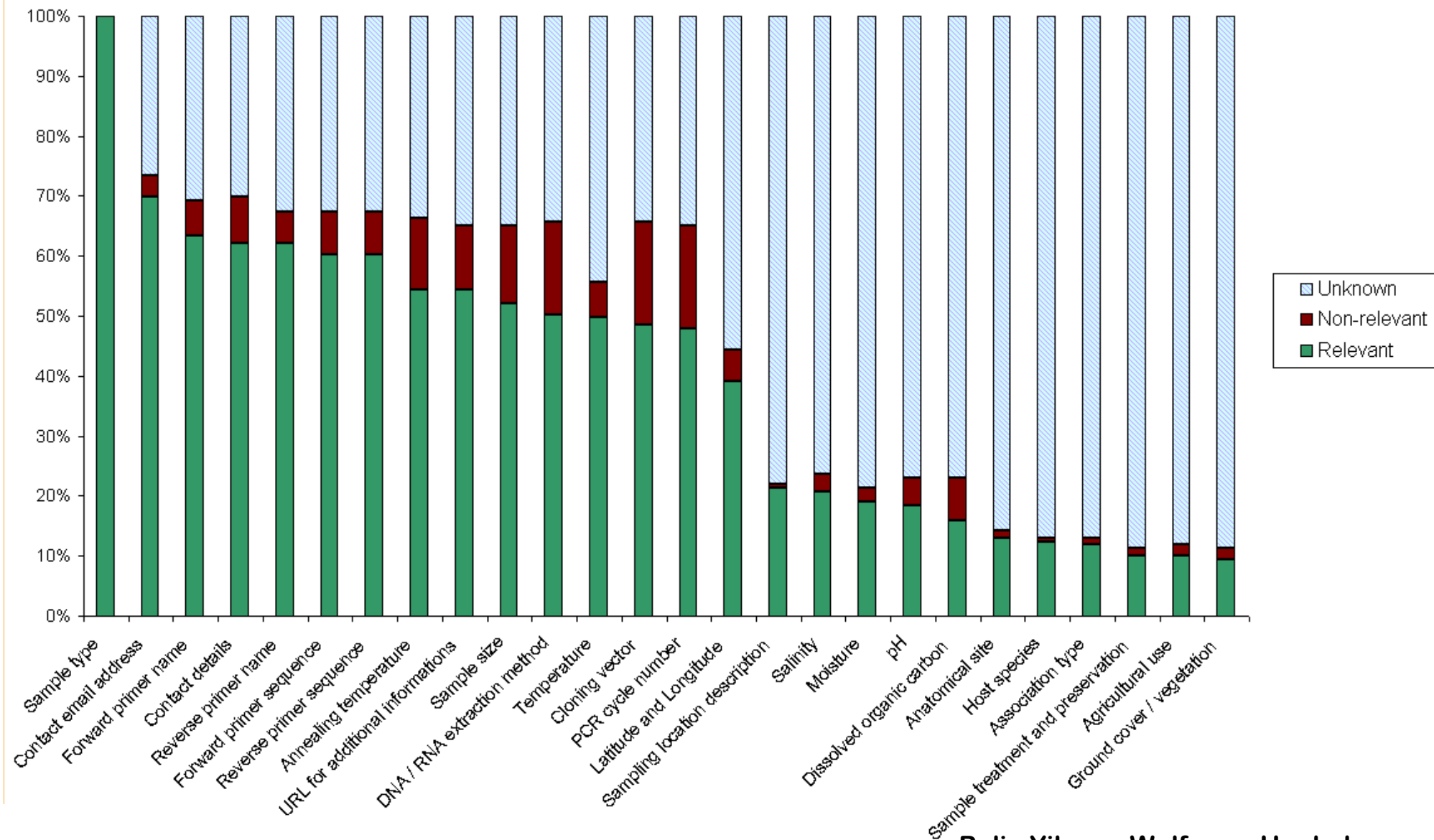


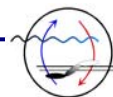
SILVA survey: relevance of fields





Hughenoltz survey: relevance of fields

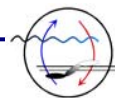




SILVA survey: user comments in free text fields

Physical environmental factors	Chemical environmental factors	Biotic factors	Sample collection and processing	Sequencing	Habitat description	Misc.
Weather conditions - rain - storm - wind	Trophic state of habitat	Rhizosphere Plantation	Filtration - size	Primers	Detailed host and tissue description - symbiotic - parasitic	Material transfer agreement
Fluid dynamics - current - flux - pressure	Ammonium concentration	Bacterial and viral abundance	Enrichment	PCR conditions	Types - soil - water body - sediment - culture	Information on other cultivated strains from same site
Moon cycle	C:N ratio	Limnology	De-aggregation	Source of sequencing template		Mass spectroscopy data
Irradiation	Heavy metal presence - contaminants	Microbiology	Sample storage method	Sequencing technology		Standard deviation and precision for measurements
Porosity	Moisture - for soil	Ecology	Sampling technique	Quality check - clones - sequence		Antibiotic susceptibility
Water potential	CaCO3 concentration - for soil		DNA extraction method	Estimate of taxon		URL
Illumination	Dissolved CO2 concentration					
	Sulfide concentration					

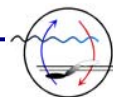
Color	# of similar comments		
			6
	2		7
	3		8
	4		9
	5		10



Motivation!

▶ Minimal Information for an ENvironmental Sequence (MIENS)

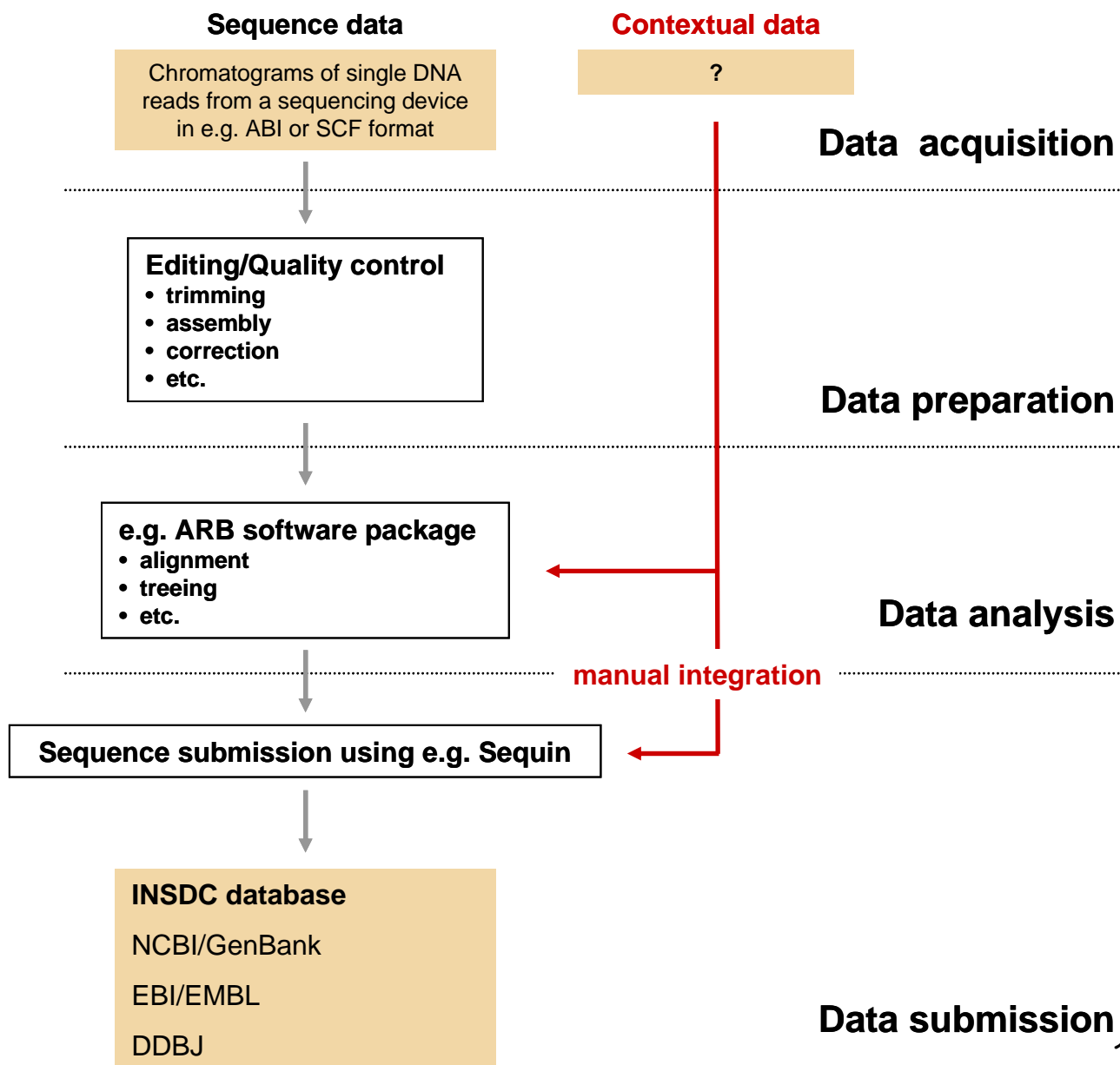
- Which attributes for environmental sequences are most relevant for the community?
- → surveys
- How can data integration and sequence submission to the INSDC effectively be handled?
- → workflow

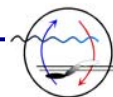


Contextual data 2 INSDC?

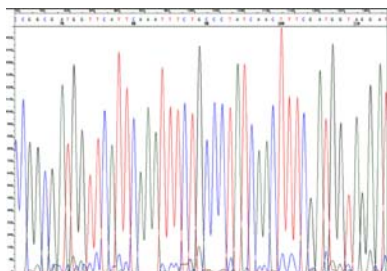
▶ There is a clear lack of solutions for

- Standardized contextual data organization/storage
- automated data integration





Workflow



Sequence Data

Sample Name	Strain	Altitude	Bio-Material	Collection Time	Collection Date	Country	Temperature
A1_fw	lib_jp_01	0 m	marine water	12:30	10-Oct-2008	Germany	23°C

Contextual Data

Metadata-FASTA

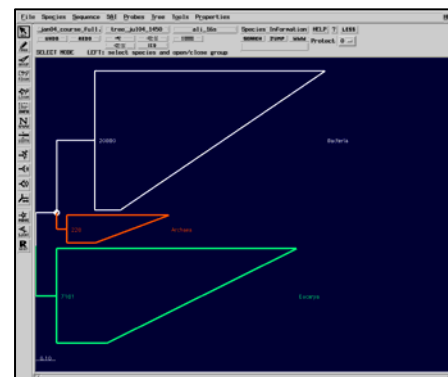
```
>A1_fw [clone-lib=lib_jp_01]
[strain=teststrain1] [altitude=0 m]
[bio-material=marine water]
[collection_time=12:30] [collection-date=10-Oct-2008] [country=Germany]
[temperature=23°C]
TGACAGNGATACTTTCCAGAGCTGAAGTTAACAAAT
GCACCTGGTTCTTTTACTAAGTGTTCAAATACC
```



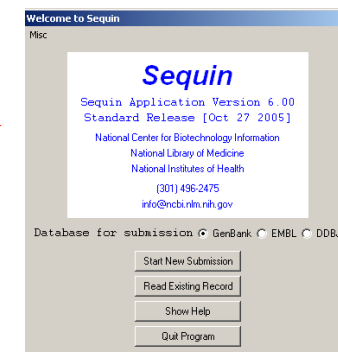
Sequence Processing & Data Integration

www.arb-silva.de/projects/contextual-data

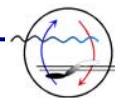
ARB & SILVA



Sequence Analysis

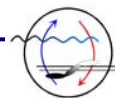


Submission



The MIGS example: Structured Comments

```
SH Authority: Genomics Standards Consortium, http://gencsc.org/
SH Definitions: http://www.insdc.org/files/structured\_comments/GSC.xml
SH
SC Collection time: 2008-10-31T09:09Z
SC Habitat: aquatic
SC Intraspecific genetic lineage: myxolydia (substrain)
SC Biomaterial: DSM 10331
SC Pathogenicity: none
SC Biotic relationship: free living
SC Trophic level: autotroph
SC Relationship to oxygen: anaerobe
SC E-resources: IMG-GEBA
SC Relevant SOPs: http://www.sopsRus.org
SC Depth: 12 metres
SC Salinity: 26.5 ppt
SC Temperature: 9.4 deg C
SC Sampling site monthly chlorophyll level: 2.2 mg/kL
SC Sampling site yearly chlorophyll level: 1.59 +/- 0.17 mg/kL
SC Lo_filter_size: 0.1 microns
SC Hi_filter_size: 0.8 microns
```



Websites

- ▶ http://gensc.org/gc_wiki/index.php/MIENS



- ▶ www.arb-silva.de/projects/contextual-data

silva

- ▶ http://tech.groups.yahoo.com/group/arb_users/
 - ARB/SILVA user group mailing list
- ▶ Thanks to
 - Jörg Peplies, Renzo Kottmann, Elmar Prüsse, Wolfgang Hankeln, Pelin Yilmaz