

MIMARKS

**Minimum
Information
about a
MARKer gene
Sequence**

Pelin Yılmaz

Max Planck Institute for Marine Microbiology

The 'Big Data'



The 'Big Data'

nature

Community cleverness required

researchers need to adapt their institutions and practices in response to torrents of new data — and need to complement smart science with smart searching.



COMMENTARY

How do your data grow?

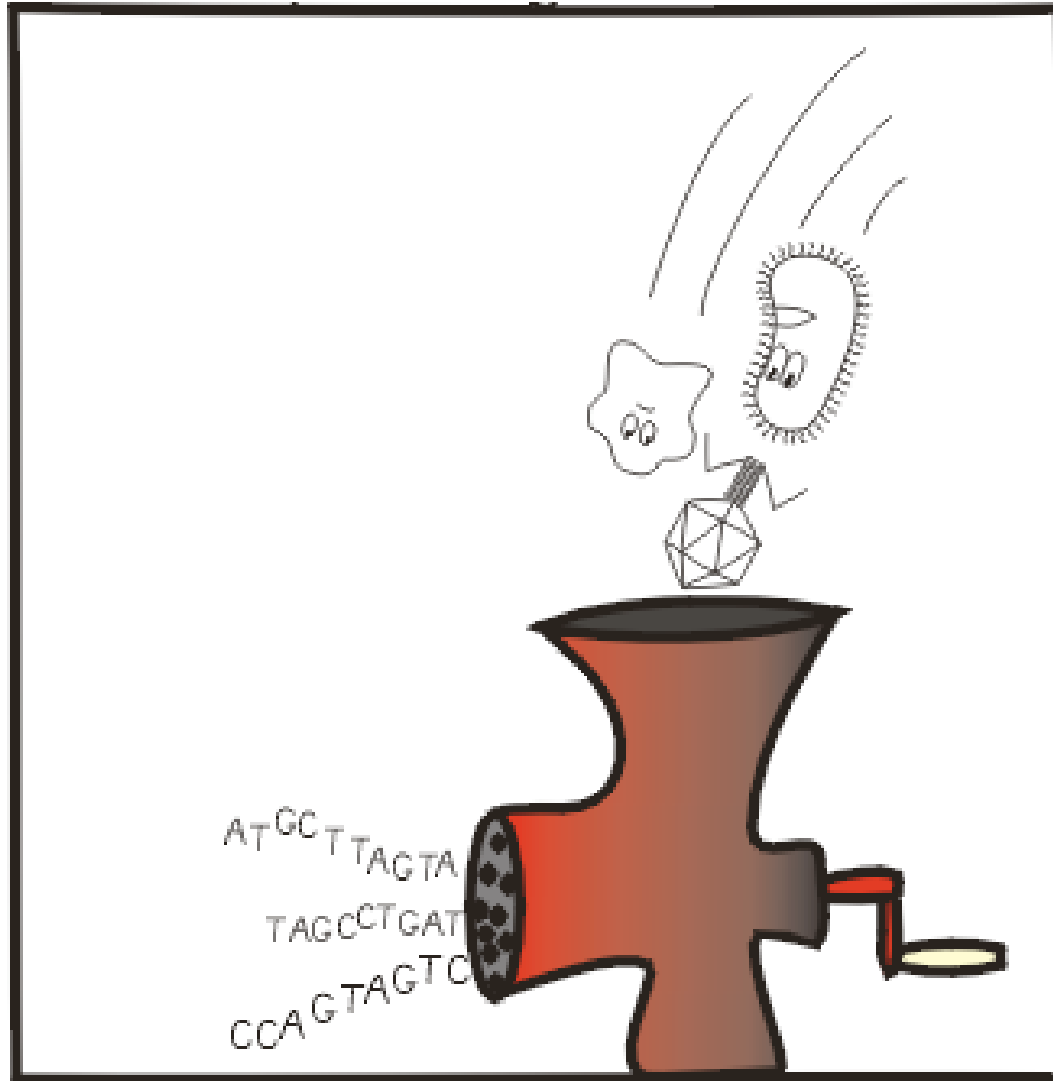
Maintaining data takes big organization, says Clifford Lynch.

FEATURE

The future of biocuration

To thrive, the field that links biologists and their data urgently needs structure, recognition and support.

Mass Sequencing



Sequence data INSDC databases

A 3D rendering of a DNA double helix structure, colored in shades of blue, set against a background of blurred DNA sequence text. The helix is shown in a perspective view, curving upwards and to the right. The background consists of multiple lines of DNA sequence characters (A, T, C, G) in a light blue color, which are out of focus, creating a sense of depth and data volume.

Contextual data



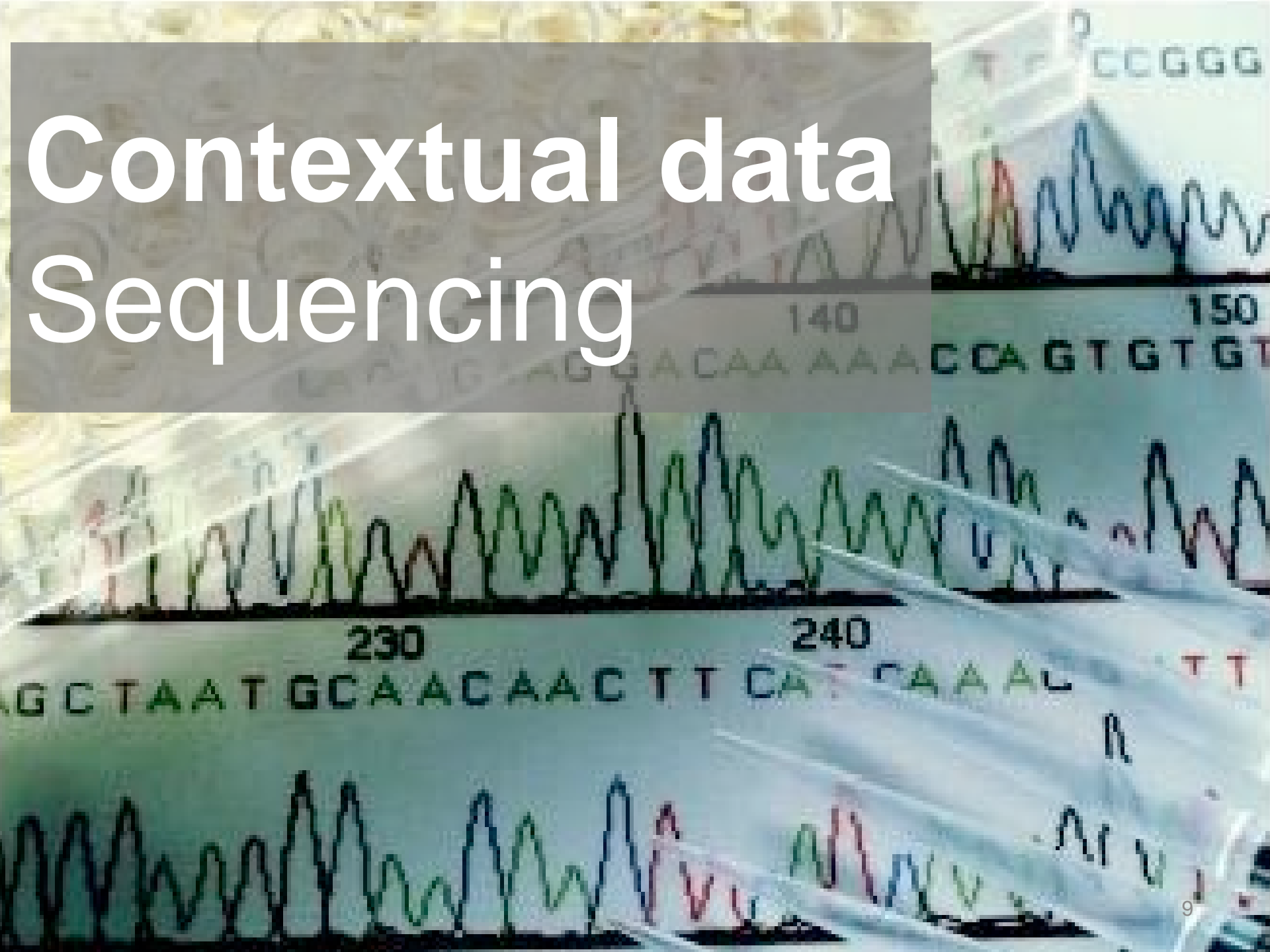
Contextual data Environment



Contextual data Experiments



Contextual data Sequencing



A Problem...

'Abandoned' sequences in INSDC databases

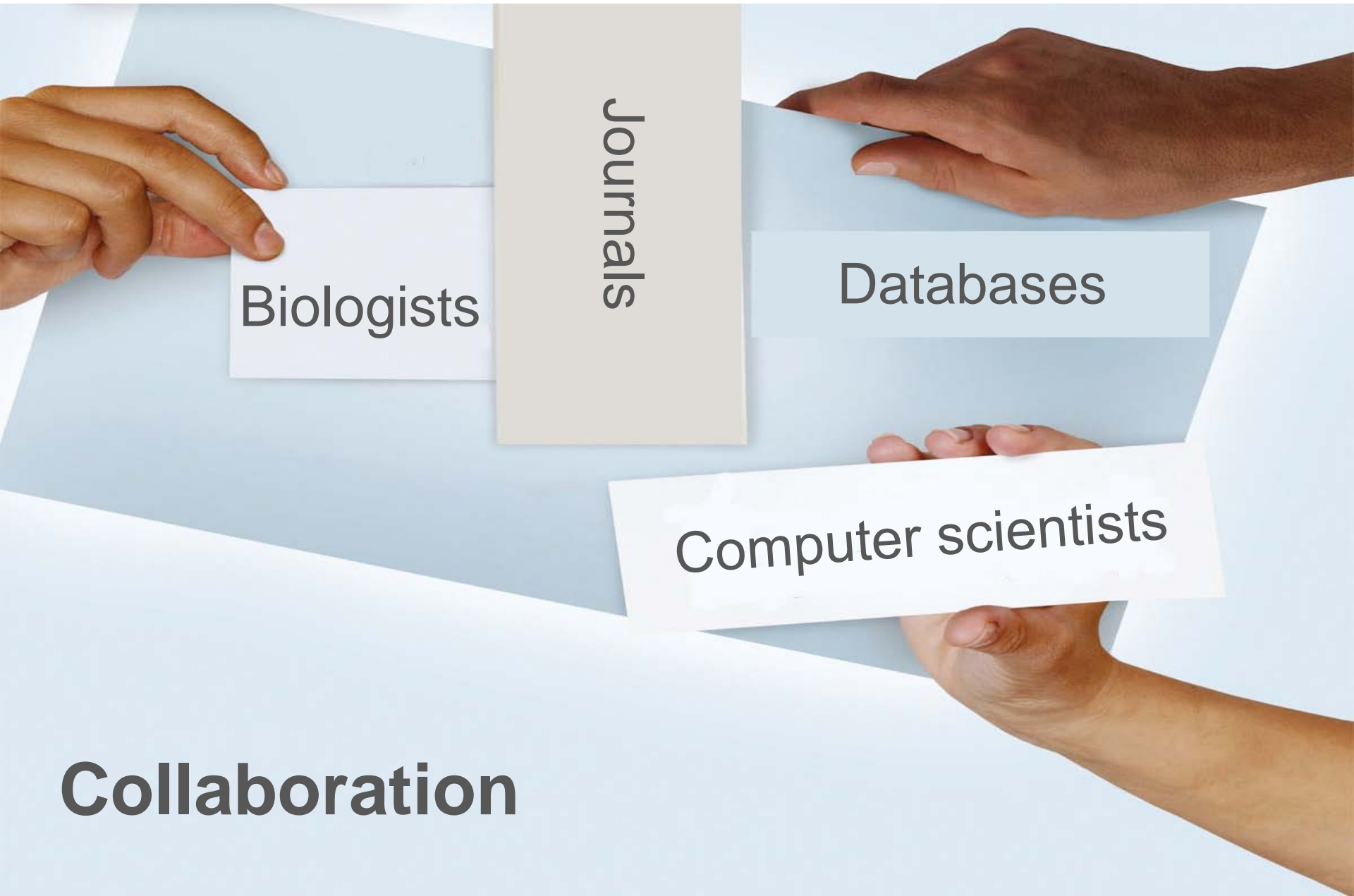
```
FEATURES                                     Location/Qualifiers
    source                                     1..1038
                                             [/organism="uncultured bacterium"]
                                             /mol_type="genomic DNA"
                                             /db_xref="taxon:77133"
                                             /clone="Ep_T1.185"
                                             /environmental_sample
    gene                                       1..1038
                                             [/gene="16S rRNA"]
    rRNA                                       1..1038
                                             /gene="16S rRNA"
                                             /product="16S ribosomal RNA"
```

8% with coordinates (latitude/longitude)

9% with collection date

41% with taxonomic assignment

The Solution



Biologists

Journals

Databases

Computer scientists

Collaboration

The Solution



Define Scope

Minimum information to be reported

XML

```
<reporting_package  
  <reporting_instance  
    <reporting_identifier="doi:10.4198/journalreporterGroup4-200-1001"  
      name="Departmental" />  
    <reporting_scope  
      <ontologyEntry category="Organism" value="Eukaryota" />  
      <ontologyReference_uri />  
    />  
  />  
</reporting_package>
```

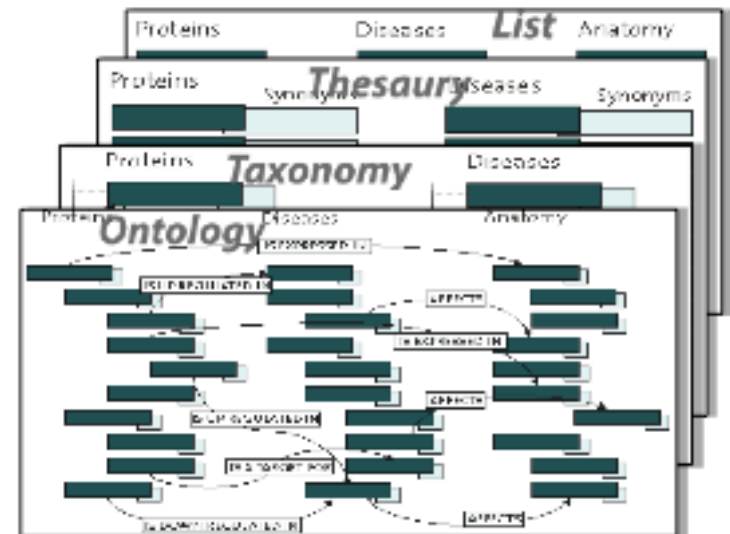
Tabular

Domain Variables for Vital Signs - Findings

Variable	Label	Type	Code	Origin	Role	Comment
USUBJID	Unique Subject Identifier	text	Sponsor Patient			Unique subject identifier with the following
USUBJID	VISIT		VSTESTCD	VSORRES		
0001	1		DIABP	70		
0001	1		SYSBP	110		
0001	1		BMI	25.3		

Define syntax

Formats for communication



Define semantics

Terminology for description¹²



Promote mechanisms that standardize

Description of genomes and metagenomes

Minimum Information about a Genome/Metagenome Sequence (MIGS/MIMS)
(Field et al. Nature Biotech 2007)

Exchange and integration of genomic data

Genomic Contextual Data Markup Language (GCDML)
(Kottmann et al. OMICS 2008)

AND NOW!

Description of marker genes

Minimum Information about a MARKer gene Sequence (MIMARKS)

MIMARKS

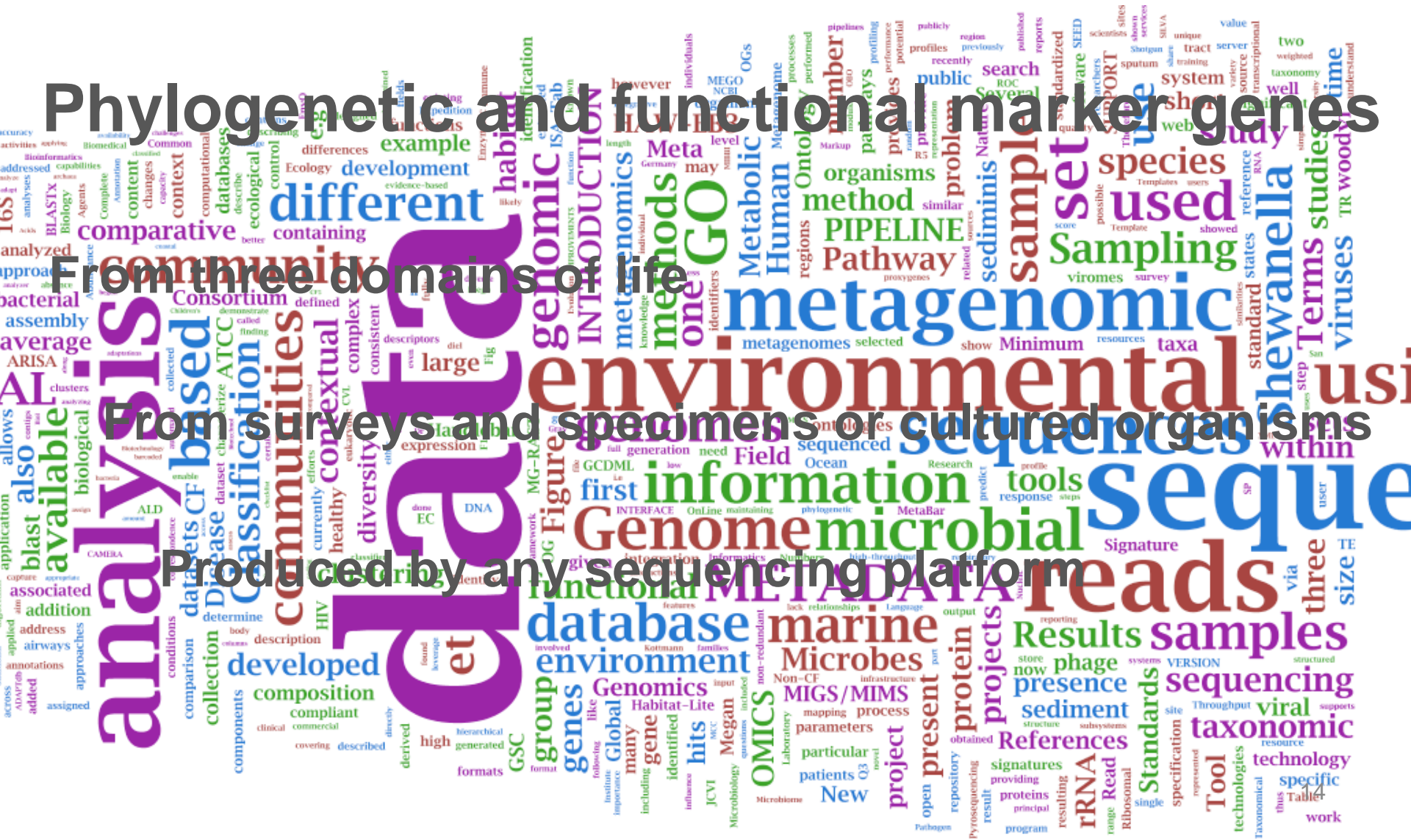
What is it good for?

Phylogenetic and functional marker genes

From three domains of life

From surveys and specimens or cultured organisms

Produced by any sequencing platform



MIGS/MIMS Checklists

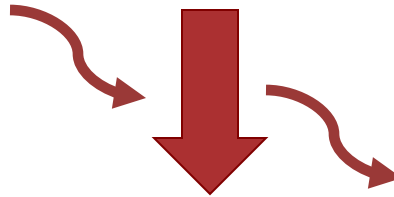
Investigation	Report type					
	EU	BA	PL	VI	OR	ME
<ul style="list-style-type: none"> • Submit to trace archives and INSDC 	M	M	M	M	M	M
<ul style="list-style-type: none"> • Investigation type (i.e., report type) 	M	M	M	M	M	M
<ul style="list-style-type: none"> • Project name² <ul style="list-style-type: none"> • Study • Environment 	M	M	M	M	M	M
<ul style="list-style-type: none"> • Geographic location (latitude and longitude^{float (point, transect and region)}, depth and altitude of sample)^(integer) 	M	M	M	M	M	M
<ul style="list-style-type: none"> • Time of sample collection^(UCT) 	M	M	M	M	M	M
<ul style="list-style-type: none"> • Habitat^{EnvO} 	M	M	M	M	M	M
MIMS extension: select to report a set of uniform measurements for a given habitat:						M
<ul style="list-style-type: none"> • Water body: (temperature, pH, salinity, pressure, chlorophyll, conductivity, light intensity, dissolved organic carbon (DOC), current, atmospheric data, density, alkalinity, dissolved oxygen, particulate organic carbon (POC), phosphate, nitrate, sulfates, sulfides, primary production)^(integer, unit) 						
<ul style="list-style-type: none"> • Nucleic acid sequence source <ul style="list-style-type: none"> • Subspecific genetic lineage (below lowest rank of NCBI taxonomy, which is subspecies) (e.g., serovar, biotype, ecotype)^(CABRI) 	M	M	M	M	M	–
<ul style="list-style-type: none"> • Ploidy (e.g., allopolyploid, polyploid)^(PATO) 	M					
<ul style="list-style-type: none"> • Number of replicons (EU, BA: chromosomes (haploid count); VI: segments)^(integer) 	M	M	–	M	–	–
<ul style="list-style-type: none"> • Extrachromosomal elements^(integer) 	X	M				
<ul style="list-style-type: none"> • Estimated size (before sequencing; to apply to all draft genomes)^(integer; base pairs) 	M	X	X	X	X	–
<ul style="list-style-type: none"> • Reference for biomaterial (primary publication if isolated before genome publication; otherwise, primary genome report)^(PMID or DOI) 	X	M	X	X	X	X
<ul style="list-style-type: none"> • Source material identifiers: (cultures of microorganisms: identifiers^(alphanumeric) for two culture collections^(OBI); specimens (e.g., organelles and Eukarya); voucher condition and 	M	M	M	M	M	M

MIGS/MIMS+MIMARKS Checklists



Surveys

MIGS/MIMS



**Publication mining
INSDC resources**

MIGS/MIMS/MIMARKS



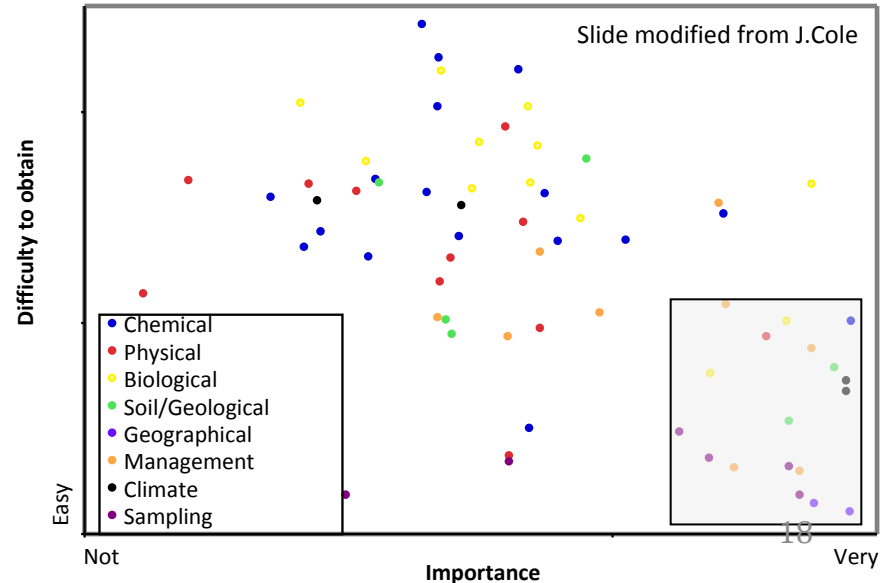
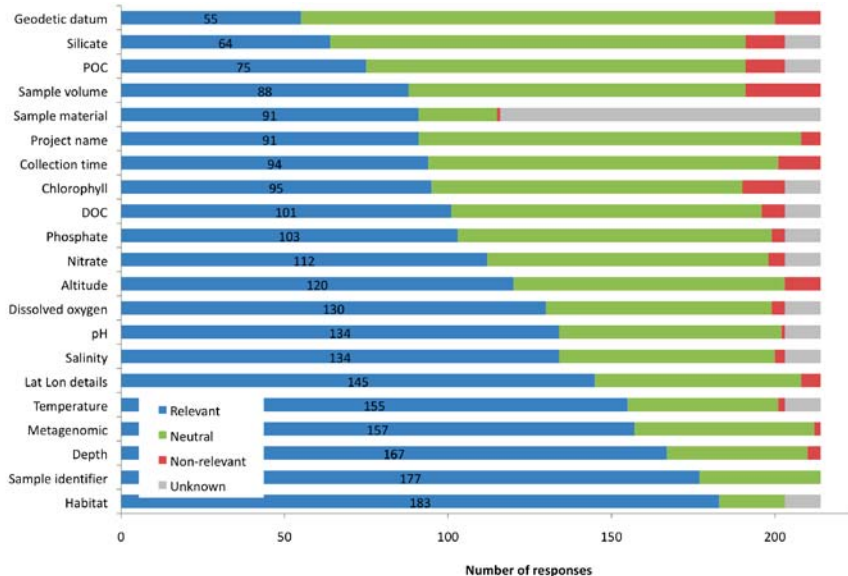
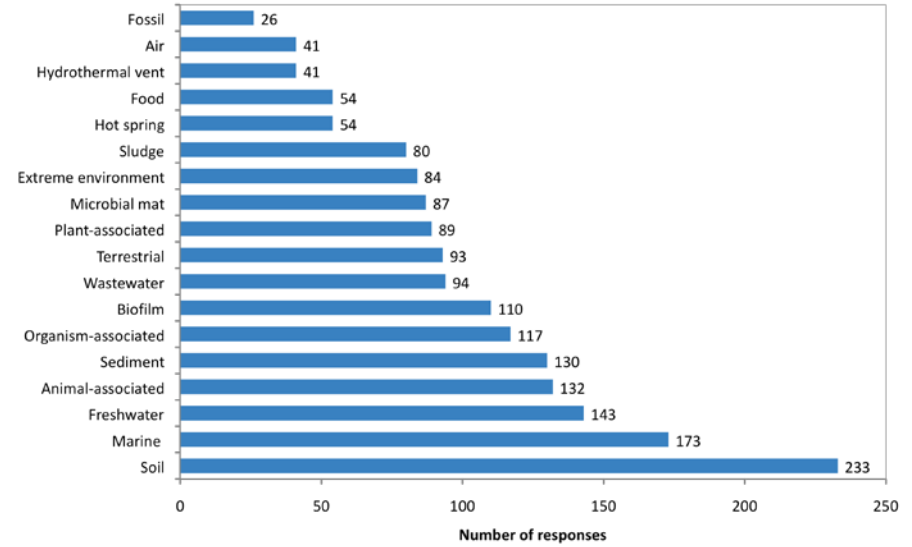
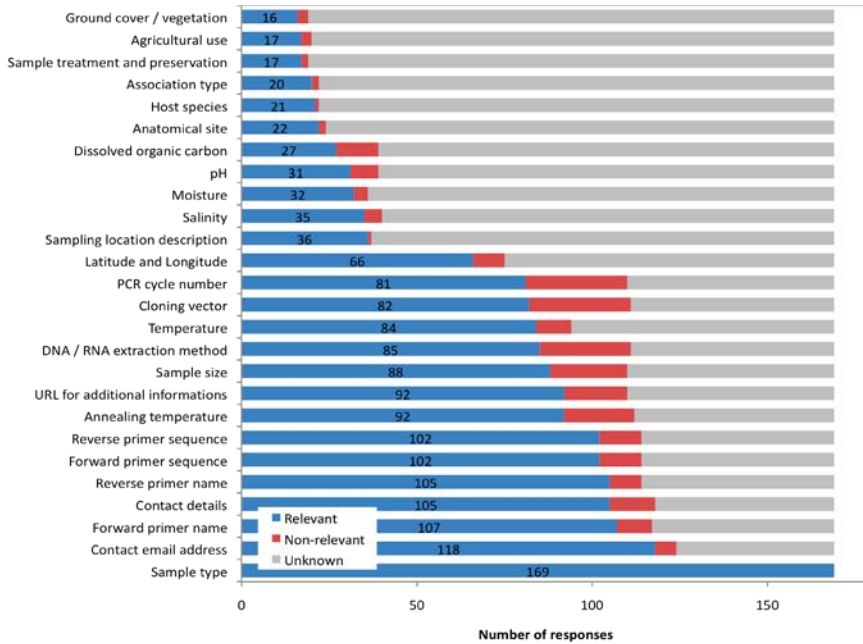
Community surveys

Four surveys:

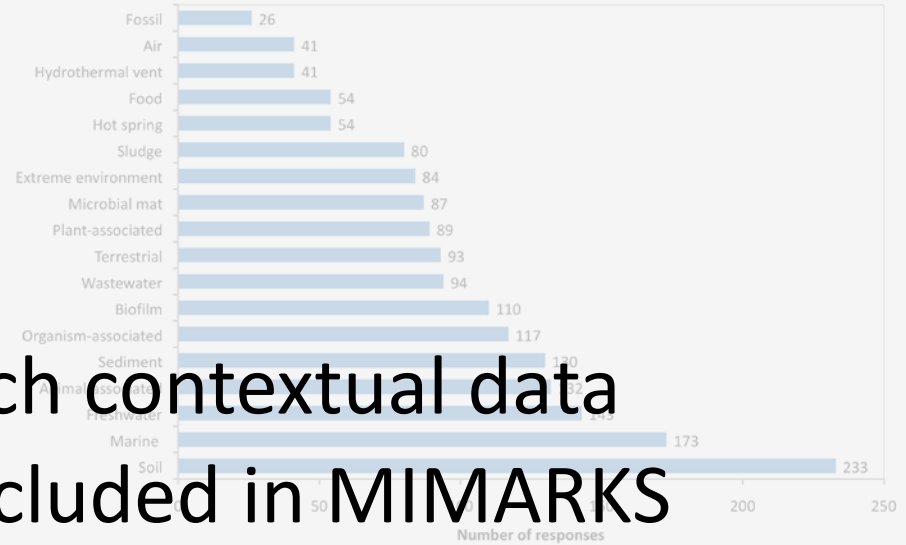
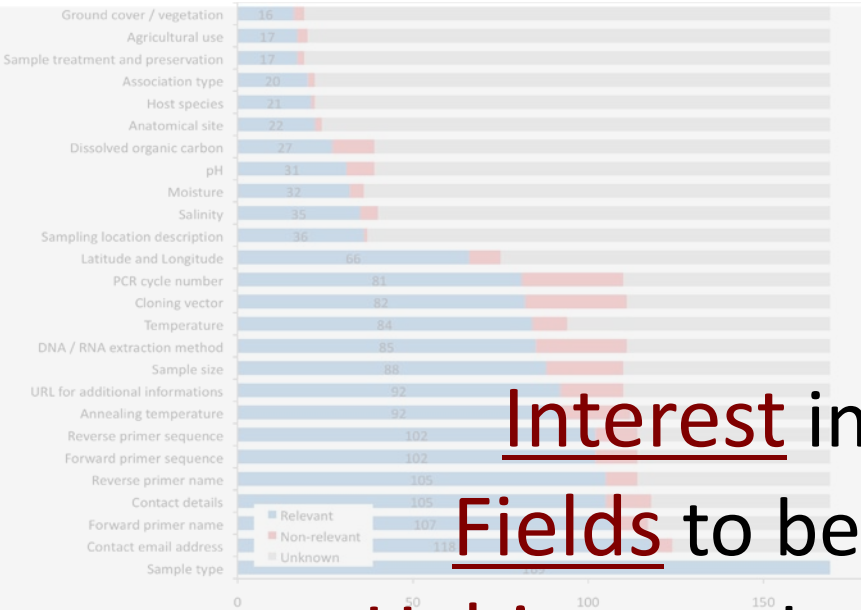
- JGI survey (focus on general description of marker genes)
- RDP habitat survey (focus on habitat terms)
- SILVA survey (focus on general description of marker genes)
- Terragenome survey (focus on soil descriptors)

Detailed information under: http://gensc.org/gc_wiki/index.php/MIMARKS_history#Metadata_surveys

What did we learn?



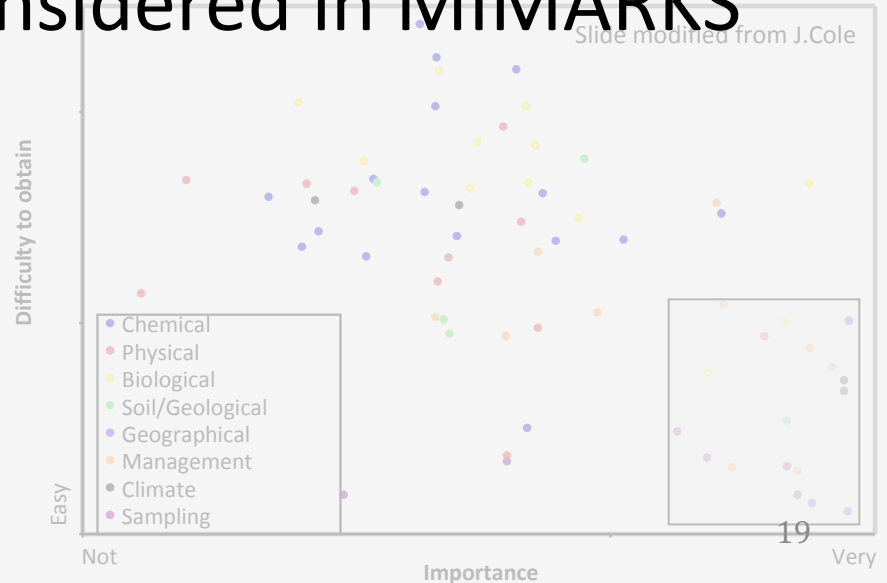
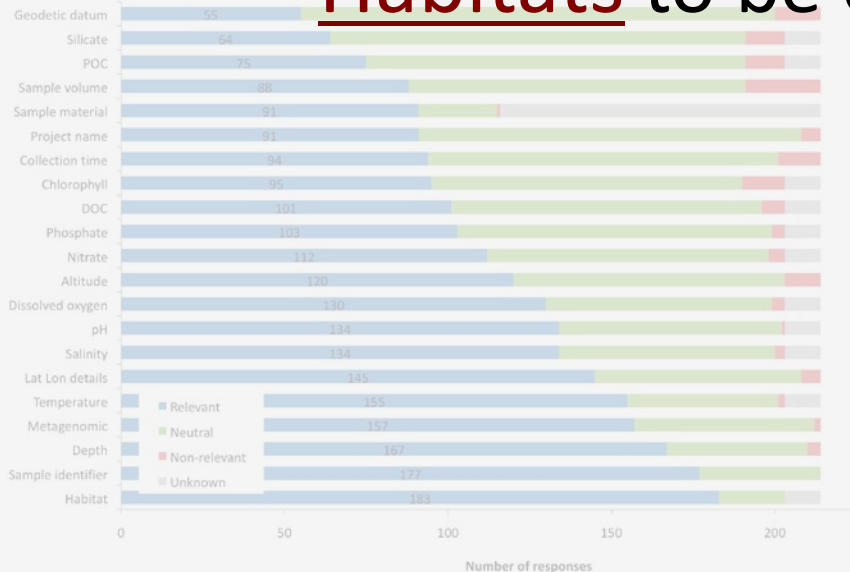
What did we learn?



Interest in rich contextual data



Fields to be included in MIMARKS

Habitats to be considered in MIMARKS



Publication mining

- Selection of publications from SILVA with the highest number of sequence submissions
 - 39 publications
 - minimum 89, maximum 149,159 sequences
 - diverse habitats; air, hydrothermal fields, microbial mats
 - 200 parameters selected

Title	DOI	Study Type 	Number of sequences 	Parameters
Pyrosequencing enumerates and contrasts soil microbial diversity	doi:10.1038/ismej.2007.53	soil	149159	depth below surface elevation pH soil type
Evolution of Mammals and Their Gut Microbes	doi: 10.1126/science.1155725	organism-associated(gut)	26160	age host species diet country
Assessment of bias associated with incomplete extraction of microbial DNA from soil	doi:10.1128/AEM.00120-09	soil	21471	depth below surface pH total organic carbon [] particle classification (silt,clay %)

More information under: http://gensc.org/gc_wiki/index.php/MIMARKS_history#Metadata_from_publications

Publication mining

- Selection of publications from SILVA with the highest number of sequence submissions
 - 39 publications
 - minimum 89, maximum 149,159 sequences
 - diverse habitats; air, hydrothermal fields, microbial mats
 - 200 parameters selected

Title	DOI	Study Type 
Pyrosequencing enumerates and contrasts soil microbial diversity	doi:10.1038/ismej.2007.53	soil
Evolution of Mammals and Their Gut Microbes	doi: 10.1126/science.1155725	organism-associated(gut)
Assessment of bias associated with incomplete extraction of microbial DNA from soil	doi:10.1128/AEM.00120-09	soil

Parameters
depth below surface elevation pH soil type
age host species diet country
depth below surface pH total organic carbon [] particle classification (silt,clay %)

INSDC resources

- Source information parsed from INSDC for each SILVA release

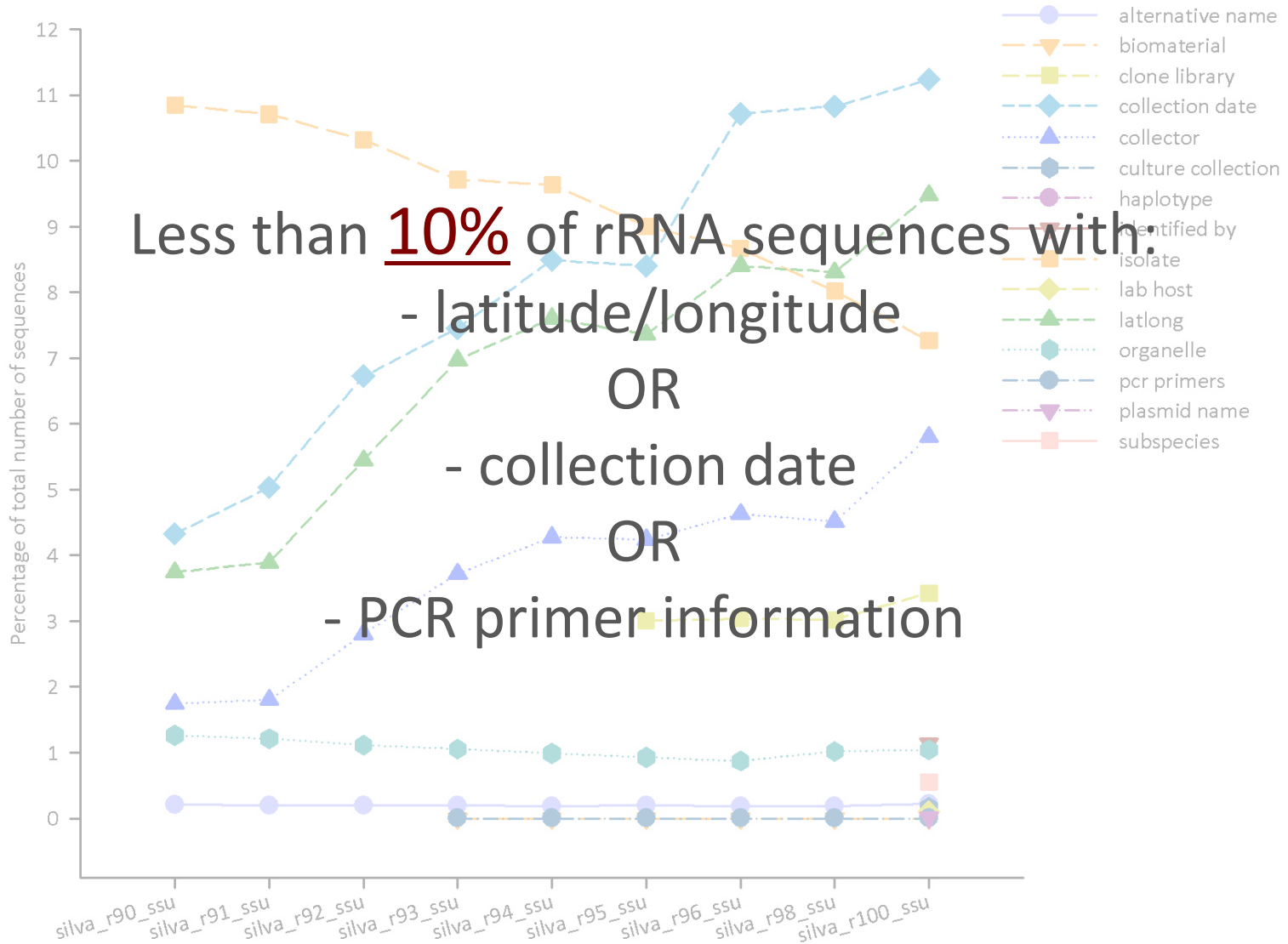
```
TITLE      Direct Submission
JOURNAL    Submitted (19-JUN-2009) Marine Sciences, University of Georgia,
           Athens, GA 30602, USA
```

FEATURES	Location/Qualifiers
→ source	1..690
qualifiers	/organism="marine bacterium SIMO-4497"
	/mol_type="genomic DNA"
	/strain="SIMO-4497"
	/isolation_source="coastal seawater"
	/db_xref="taxon: 661098 "
	/country="USA: Sapelo Island, GA"
rRNA	<1..>690
	/product="16S ribosomal RNA"

ORIGIN

```
1 taacatttct agcttgctag aagatgacga gcgggcggacg ggtgagtaat gcttgggaac
61 atgccttgag gtggggggaca accattggaa acgatggcta ataccgcata atgtctacgg
121 accaaagggg gcttcgggctc tcgcctttag attggcccaa gtgggattag ctagttgggtg
181 aggtaaaggc tcaccaaggc gacgatccct agctggtttg agaggatgat cagccacact
241 ggaactgaga cacggtccag actcctacgg gaggcagcag tggggaatat tgcacaatgg
301 gcgcaagcct gatgcagcca tgccgcgtgt gtgaagaagg ccttcggggtt gtaaagcact
361 ttcagtcagg aggaaaggtt agtagttaat acctgctagc tgtgacgtta ctgacagaag
421 aagcaccggc taactccgtg ccagcagccg cggtaatacg gaggggtgcga gcgttaatcg
481 gaattactgg gcgtaaagcg tacgcaggcg gtttgttaag cgagatgtga aagccccggg
541 cttaacctgg gaactgcatt tcgaactggc aaactagagt gtgatagagg gtggtagaat
601 ttcaggtgta gcggtgaaat gcgtagagat ctgaaggaat cccgatggcg agggcagcca
```

INSDC resources



MIMARKS checklist v2.1

“Investigation”

“M”=mandatory “X”=recommended “C”=conditional mandatory “-”=not applicable

	Survey	Specimen
submit to insdc	M	M
investigation type	M	M
project name	M	M
experimental factor	C	X

MIMARKS checklist v2.1

“Environment”

“M”=mandatory “X”=recommended “C”=conditional mandatory “E”=environment-dependent “-”=not applicable

	Survey	Specimen
collection date	M	M
geographic location (latitude and longitude)	M	M
geographic location (depth)	E	E
geographic location (altitude/elevation)	E	E
geographic location (country)	M	M
environment (biome)	M	M
environment (feature)	M	M
environment (material)	M	M
environmental package	C	X

MIMARKS checklist v2.1

“Environment”

“M”=mandatory “X”=recommended “C”=conditional mandatory “E”=environment-dependent “-”=not applicable

	Survey	Specimen
collection date	M	M
geographic location (latitude and longitude)	M	M
	Survey	Specimen
geographic location (depth) [air, host-associated, human-associated, human-skin, human-oral, human-gut, human-vaginal, microbial mat/biofilm, miscellaneous natural or artificial environment, plant-associated, sediment, soil, wastewater/sludge, water]		E
environmental package	C	X
geographic location (country)	M	M
environment (biome)	M	M
environment (feature)	M	M
environment (material)	M	M
environmental package	C	X

MIMARKS checklist v2.1

“Nucleic Acid Sequence Source”

“M”=mandatory “X”=recommended “C”=conditional mandatory “-”=not applicable

	Survey	Specimen
subspecific genetic lineage	-	C
extrachromosomal elements	-	X
source material identifiers	-	C
observed biotic relationship	-	C
trophic level	-	C
relationship to oxygen	X	C
isolation and growth condition	-	M
sample collection device or method	C	X
sample material processing	C	C
amount or size of sample collected	C	X

MIMARKS checklist v2.1

“Sequencing”

“M”=mandatory “X”=recommended “C”=conditional mandatory “-”=not applicable

	Survey	Specimen
nucleic acid extraction	C	C
nucleic acid amplification	C	C
library size	C	-
library reads sequenced	C	-
library vector	C	-
library screening strategy	C	-
target gene	M	M
target subfragment	C	C
pcr primers	C	C

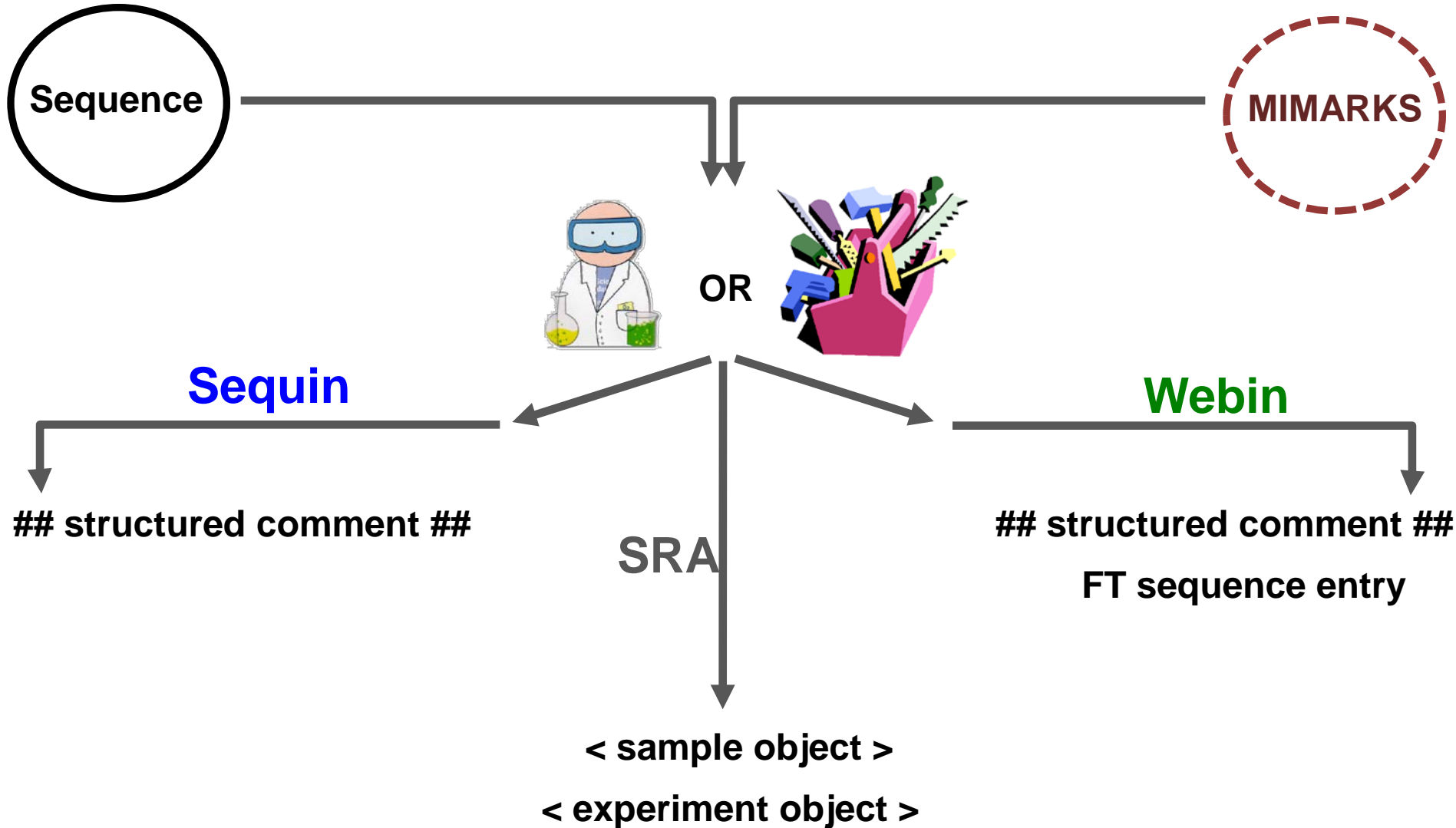
MIMARKS checklist v2.1

“Sequencing”

“M”=mandatory “X”=recommended “C”=conditional mandatory “-”=not applicable

	Survey	Specimen
multiplex identifiers	C	-
adapters	C	-
pcr conditions	C	C
sequencing method	M	M
sequence quality check	C	C
chimera check	C	C
assembly	C	-
relevant standard operating procedures	C	C
relevant electronic resources	C	C

Closing the cycle Submissions



Closing the cycle

“Submissions”

tag	value
submit_to_insd	yes
investigation_type	MIMARKS_survey
project_name	dsrA-based fingerprinting study
collection_date	2006-04-15
lat_lon	21.89845 -93.436748
depth	5.0-7.5 cm
altitude_elev	0
country	Mexico
environment	marine sediment
env_package	water
tot_depth_water_col	2900 m
samp_collect_device	ROV (Remotely Operated Vehicle) push core
samp_mat_process	Sectioned in 2.5 cm intervals; stored at -20 degC
samp_size	565.2 cm ³
nucl_acid_ext	PMID: 8593035
nucl_acid_amp	PCR
lib_size	222
lib_reads_seqd	96
lib_vector	TOPO TA;pGEM-T-Easy
lib_screen	screened
target_gene	dsrA

Closing the cycle “Submissions”

The image shows a screenshot of a web browser window with the 'Annotate' menu open. The menu items are:

- Genes and Named Regions
- Coding Regions and Transcripts
- Structural RNAs
- Bibliographic and Comments
- Sites and Bonds
- Remaining Features
- Batch Feature Apply
- Batch Feature Edit
- Publications
- Descriptors
- Generate Definition Line
- Advanced Table Readers
- Sort Unique Count By Group

The 'Advanced Table Readers' sub-menu is open, showing the following options:

- Load Structured Comments from Table

Red arrows indicate the sequence of actions:

- Arrow 1 points to the 'Annotate' menu.
- Arrow 2 points to the 'Advanced Table Readers' sub-menu.
- Arrow 3 points to the 'Load Structured Comments from Table' option.

The background shows a table with the following data:

in type	uncultured
ne	aquatic p
	35246N 5
	10m
e	25 deg C
	marine
	2 uM
	200 ppt

Below the table, there is a 'DESCRIPTION' section with the following text:

```
VERSION
KEYWORDS
SOURCE uncultured bacterium
ORGANISM uncultured bacterium
Bacteria; environmental samples.
REFERENCE 1 (bases 1 to 883)
AUTHORS Yilmaz,P.
TITLE a study
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 883)
AUTHORS Yilmaz,P.
TITLE Direct Submission
JOURNAL Submitted (27-NOV-2009) Max Planck
Location/Qualifiers
FEATURES
source 1..883
/organism="uncultured bacterium"
/mol_type="genomic DNA"
```

Closing the cycle “Submissions”



Uncultured

Target Sequence

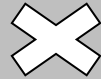
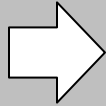
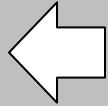
Format Mode Style

Comment: tag value submit_to_insdcc yes investigation_type MIENS_survey projec>

```
LOCUS      Uncultured                883 bp    DNA     Linear   ENV 27-NOV-2009
DEFINITION marine microorganism clone 4024AA_02 16S ribosomal RNA gene,
            partial sequence.
ACCESSION
VERSION
KEYWORDS
SOURCE     uncultured bacterium
ORGANISM   uncultured bacterium
            Bacteria; environmental samples.
REFERENCE  1 (bases 1 to 883)
AUTHORS   Yilmaz,P.
TITLE     a study
JOURNAL   Unpublished
REFERENCE  2 (bases 1 to 883)
AUTHORS   Yilmaz,P.
TITLE     Direct Submission
JOURNAL   Submitted (27 Nov 2009) Max Planck
COMMENT   ## Metadata-START##      submit_to_insdcc yes      investigation_type
MIENS_survey      project_name dsrA-based fingerprinting study
collection_date 2006-04-15 lat_lon 21.89845 -93.436748 depth
5.0-7.5 cm altitude_elev 0 country Mexico environment marine
sediment env_package water tot_depth_water_col 2900 m
samp_collect_device ROV (Remotely Operated Vehicle) push core
samp_mat_process Sectioned in 2.5 cm intervals; stored at -20 degC
samp_size 565.2 cm3 nucl_acid_ext PMID: 8593035 nucl_acid_amp PCR
lib_size 222 lib_heads_seqd 96 lib_vector TOPO TA;pGEM-T-Easy
lib_screen screened target_gene dsrA
pcr_primersFMD:acscactggaagcaag, acccaytggaagcaag, gccactggaagcaag,
accattggaacatg, actcaactggaagcaag; REV:gtgtagcagttaccgca,
gtgtaacagttaccaca,gtgtaacagttaccgca,gtgtagcagttkccgca,
gtgtagcagttaccaca,gtgtaacagttaccaca,tyttccatccaccarfcc pcr_cond
initial
denaturation:94oC_3min;denaturation:94degC_40sec;annealing:54degC_4
0sec; elongation:72degC_2min;final elongation:72degC_8 min;30
sequencing_meth dideoxy seq_quality_check none.
FEATURES             Location/Qualifiers
     source            1..883
                        /organism="uncultured bacterium"
                        /mol_type="genomic DNA"
BASE COUNT      207 a      223 c      281 g      170 t      2 others
ORIGIN
1 cattgctga ggacggaaga ggagaatgga attccagtg tagagtgaa attcgtgat
```

Closing the cycle

“Submissions”



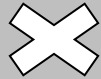
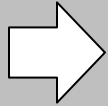
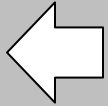
<http://www.ebi.ac.uk/embl/Submission/>

What kind of sequence(s) are you submitting?

Select	Sequence type	Description	Example
<input type="checkbox"/>	MIMARKS-compliant 16S rRNA	For the submission of 16S rRNA sequences compliant with MIMARKS standard	

Closing the cycle

“Submissions”

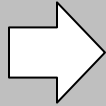
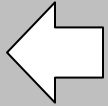


<http://www.ebi.ac.uk/embl/Submission/>

Field	Info	Common	Value
Organism		<input type="checkbox"/>	
Project name		<input checked="" type="checkbox"/>	
Experimental factor		<input checked="" type="checkbox"/>	
Latlon		<input type="checkbox"/>	
Collection date		<input type="checkbox"/>	

Closing the cycle

“Submissions”



<http://www.ebi.ac.uk/embl/Submission/>

Import your data

In order to enter your data using a spreadsheet, please download the template spreadsheet that we have prepared base on the information that you have provided so far and open in your spreadsheet application (eg. Microsoft Excel, Open Office).

[Import values in spreadsheet format](#)

Favorite Links

- http://gensc.org/gc_wiki/index.php/MIMARKS
- <http://ww.megx.net/metabar/> For GenBank and ENA submissions
- <http://www.megx.net/CDinFusion/> For GenBank submissions
- <http://isatab.sourceforge.net/> For SRA submissions
- <https://pyro.cme.msu.edu/sra/login.spr> For SRA submissions