Data Evolution: *Next Era Biological Data Hurdles for Data Storage, Preservation and Integrity*

Richard Slayden, PhD.
Professor-Microbiology, Immunology & Pathology
Executive Director-Center for Environmental Medicine
Colorado State University

Activities

Organic & Biomolecular Chemistry

PAPER

View Article Online
View Journal

Synthesis and evaluation of new 2-aminothiophenes against *Mycobacterium tuberculosis*†

Cite this: DOI: 10.1039/c6ob00821f

Sandeep Thanna,‡[a] Susan E. Knudson,‡[b] Anna Grzegorzewicz,‡[b] Sunayana I
Christopher M. Goins,[a] Donald R. Ronning,[a] Mary Jackson,*[b] Richard A. Slay
and Steven J. Sucheck*[a]

Contents lists available at SciVerse ScienceDirect

Tuberculosis

journal homepage: http://intl.elsevierhealth.com/journals/tube

REVIEW

Updating and curating metabolic pathways of TB

Richard A. Slayden[a,*], Mary Jackson[a], Jeremy Zucker[b], Melissa V. Ramirez[a], Clinton C. Dawson[a], Rebecca Crew[a], Nicole S. Sampson[c], Suzanne T. Thomas[c], Neema Jamshidi[d], Peter Sisk[b], Ron Caspi[e], Dean C. Crick[a], Michael R. McNeil[a], Martin S. Pavelka[f], Michael Niederweis[g], Axel Siroy[g], Valentina Dona[h], Johnjoe McFadden[i], Helena Boshoff[h], Jocelyne M. Lew[j]

[a] Colorado State University, Fort Collins, CO, USA
[b] Broad Institute of MIT and Harvard, Cambridge, MA, USA
[c] Stony Brook University, Stony Brook, NY, USA
[d] University of California, San Diego, USA
[e] SRI International, Menlo Park, CA, USA
[f] University of Rochester Medical Center, Rochester, NY, USA
[g] University of Alabama at Birmingham, AL, USA
[h] NIH/NIAID, Bethesda, USA
[i] University of Surrey, Guildford, UK
[j] Global Health Institute, Ecole Polytechnique Fédérale de Lausanne, Station 19, CH-1015 Lausanne, Switzerland

O'Hara *et al. BMC Genomics* 2013, **14**:832
http://www.biomedcentral.com/1471-2164/14/832

BMC Genomics

METHODOLOGY ARTICLE Open Access

Iterative feature removal yields highly discriminative pathways

Stephen O'Hara[1], Kun Wang[1,6], Richard A Slayden[2], Alan R Schenkel[2], Greg Huber[3], Corey S O'Hern[4], Mark D Shattuck[5] and Michael Kirby[1*]

Contents lists available at ScienceDirect

Tuberculosis

journal homepage: http://intl.elsevierhealth.com/journals/tube

MOLECULAR ASPECTS

MadR1, a *Mycobacterium tuberculosis* cell cycle stress response protein that is a member of a widely conserved protein class of prokaryotic, eukaryotic and archeal origin

Rebecca Crew [a], Melissa V. Ramirez [a], Kathleen England [b], Richard A. Slayden [a,*]

[a] Mycobacteria Research Laboratories, Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, CO 80523, USA
[b] Stanford University School of Medicine, Department of Infectious Diseases, Stanford, CA 94305, USA

Perspective and thoughts about the topic

I. Storage:

- ✓ tools
- ✓ How big is big data & what's the complexity & versioning

II. Preservation:

- ✓ Lab notebook, Vocabulary & Key words
- ✓ Preservation, storage & backup & distribution

III. Integrity:

- ✓ Corruption: accidental *or intentional*
- ✓ Big data "troubles"
- ✓ Concept of "intentionally deliberately vague"

I. Storage: *Tools*

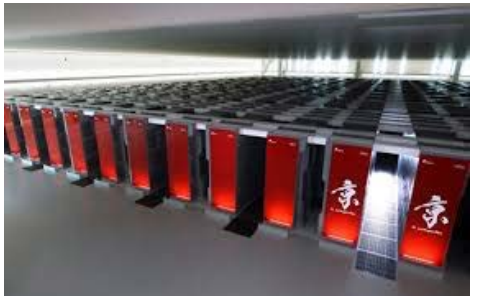
Biologists are not computer people & Computer people are not not biology people

“Biologist”

“Computationalist” or “data people”

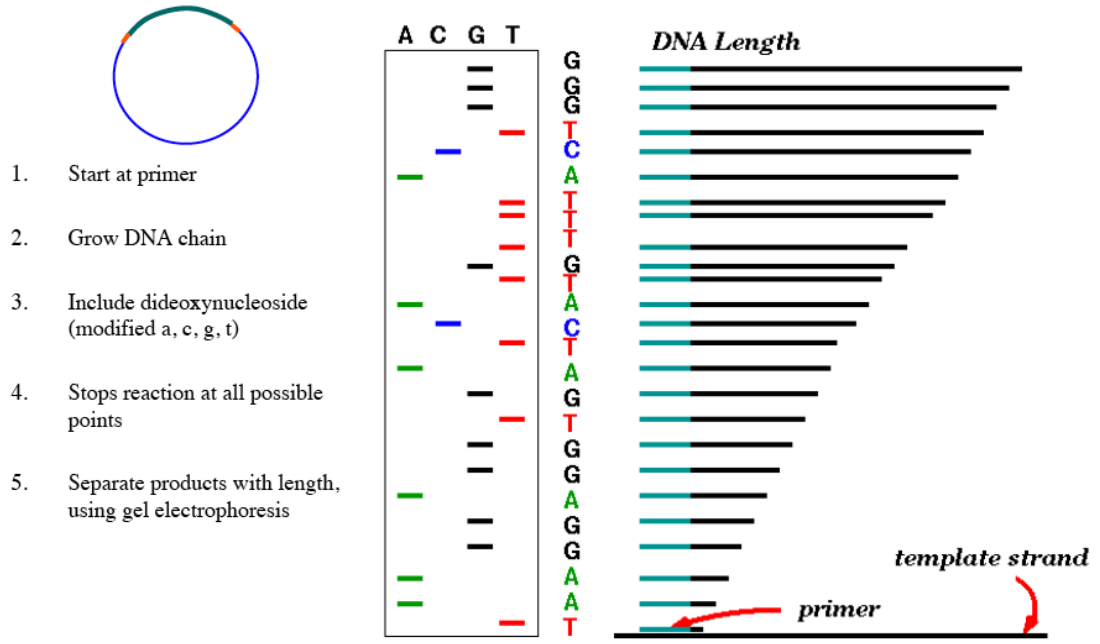
Work station

Analysis

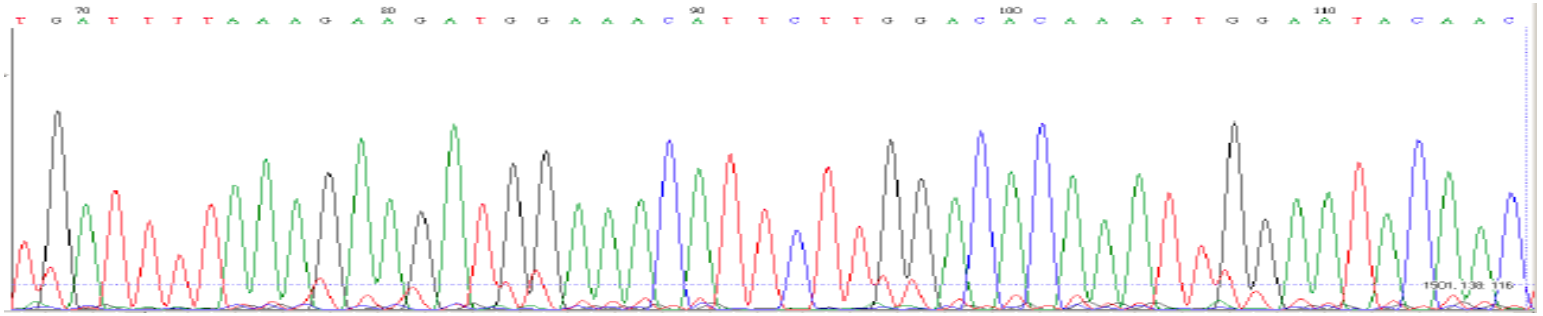


I. Storage: *Example of data explosion*

Traditional sequencing



source: robotics.stanford.edu/~serafim/cs262/Spring2003/Slides/Lecture9.ppt



I. Storage: *Example of data explosion*

Next Generation Sequencing

2001 First human genome
sequence draft: ~ 13 years and 300 million US$

Technology Review
May 2005: ~ 6 month and 20 to 30 million US$

The Scientist
(Vol. 20,2 p.67) 454: ~ 1 month and 900 000 US$ (1x coverage)

The Scientist
(Vol. 20,2 p.67) Solexa: ~ 6 month and 50 000 US$ (15x coverage)

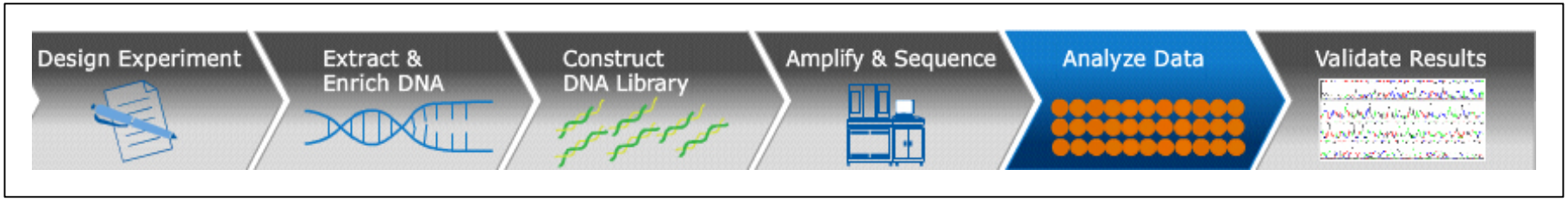
Published literature using AB SOLiD
SOLiD sequencer: 14 days and 20 000 US$ (~10x coverage)

Proton: 4 hrs- 1,000's bacteria, Human genome (~$2,000)

Terabytes of data (The prefix tera is derived from the Greek word for monster)

I. Storage: *Example of complexity of the data set*

From the Bench to the Data: *Workflow & complexity of the information required*



Information captured at each step of the process that provides context for the outcome

I. Storage: *Example of complexity of the data set*

Example of where data is coming from: *Next Generation Sequencing Technology*

<u>Platform & Data size</u>

P1: 665 million reads

P2: 1.2 billion reads

P3: 3-4 billion reads

I. Storage: *Example of data explosion*

ANALYSIS

Keep in mind that much of the data analysis software available today was not really designed for NGS-scale metagenomic datasets.

For example, simple sequence alignments for a metagenomic dataset with "only" 25M reads against a "small" database with only 1,000 records is 25 billion alignments.

On a fast server with 10 alignments per second per CPU that's about 290,000 days. If you run this on a 1,000 core cluster it's 290 days.

Substantial horsepower, or some data reduction methods, or fairly small highly targeted databases, to make runs feasible.

MEGAN is a current analysis solution and you can also install it on your workstations; it's free. However, MEGAN needs 64GB RAM and multicore (about 8-core) to *begin* to handle metagenomic-sized datasets.

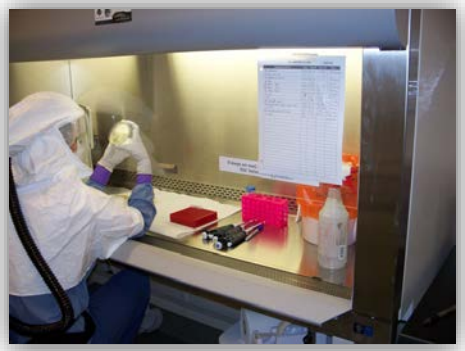
A metagenomics data analysis pipeline is in place for handling NGS sequence data.

I. Storage: *Example of complexity of the data set*

- ✓ **Scientific Applications-*Genome sequencing, whole transcriptome, modifications, structural variations***
- ✓ **Workflow: Material type (ie. DNA or RNA) & sample preparation (Total RNA vs mRNA)**
- ✓ **Workflow: library preparation & sequencing run-*mate-pair or fragment***
- ✓ **Computational Resources: *Reference or de novo sequence assembly***
- ✓ **Data reduction: Data Analysis- *What portion of the data is analyzable, condensation, biologically relevant criteria***
- ✓ **Secondary comparative analysis-*Applied analysis, incorporation with historical data, statistics, math and data structure.***

I. Storage: *Example of complexity of the data set*

What experimental data makes up information?



Study Design
Experiment
Complexity
Data Reduction
Analysis

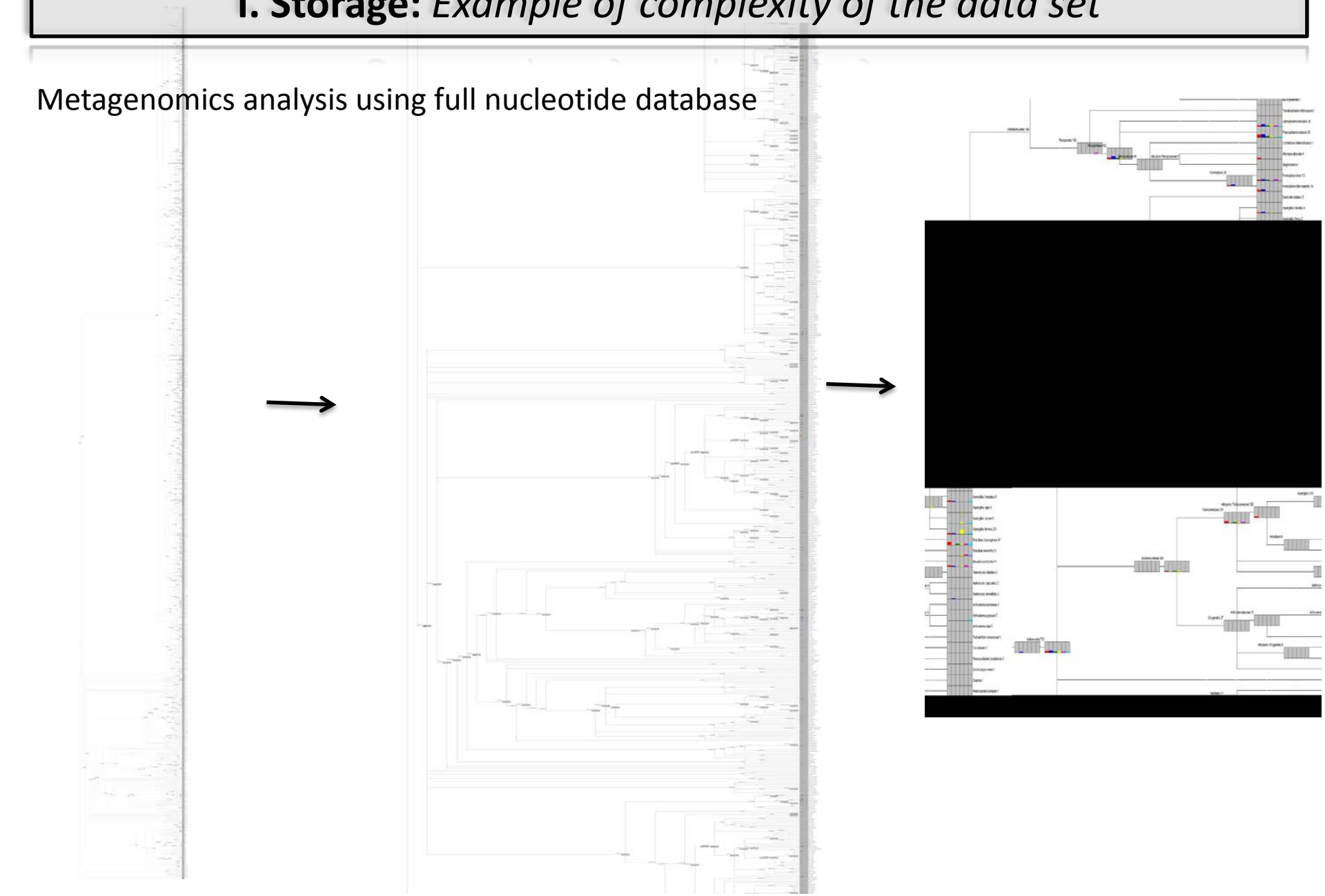
I. Storage: *Example of complexity of the data set*

Metagenomics

- Genetic material recovered from environmental samples
- NextGen sequencing => sample DNA reads
- NCBI nt (nucleotide), env (environmental), 16S databases
- Blast sample reads against NCBI databases
- MEGAN => assign reads to taxa

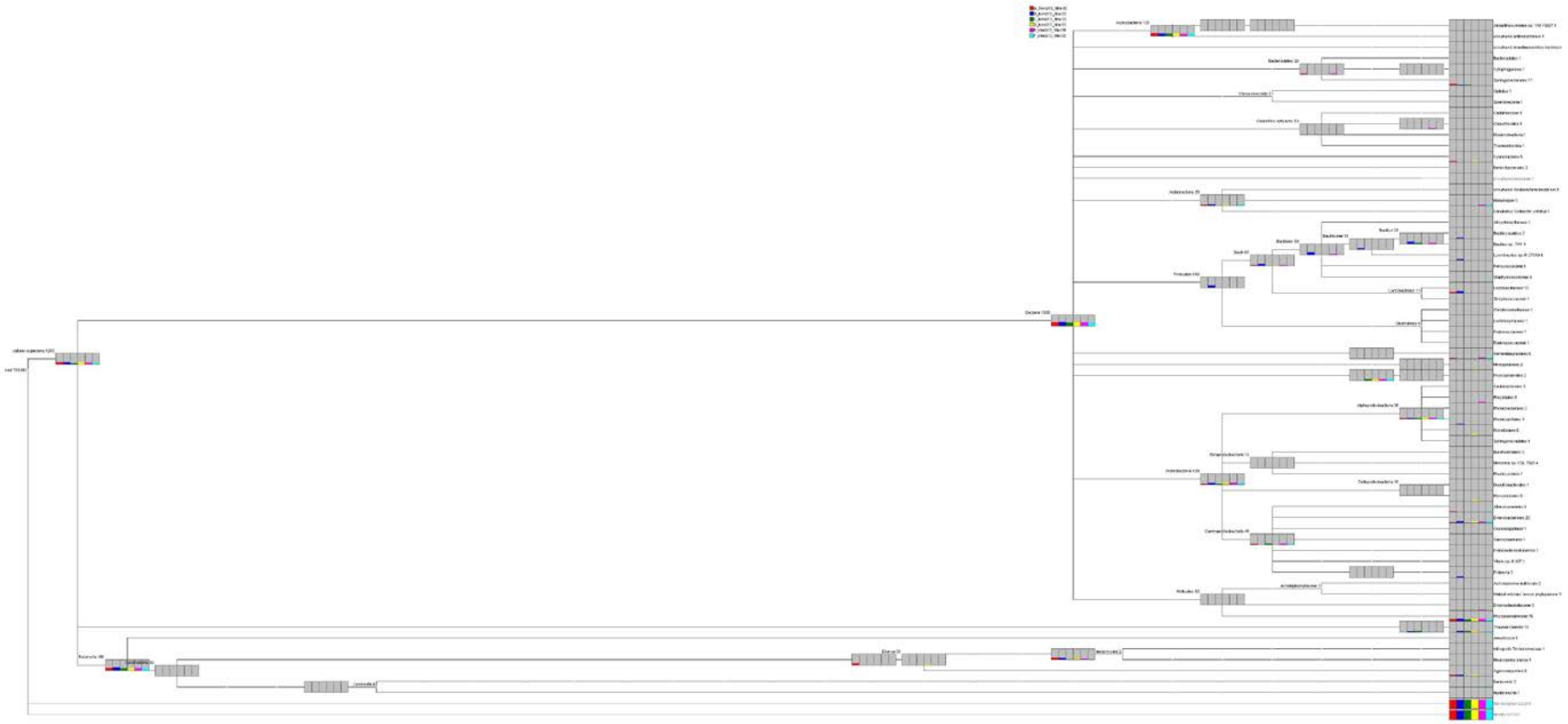
I. Storage: *Example of complexity of the data set*

Metagenomics analysis using full nucleotide database



I. Storage: *Example of complexity of the data set*

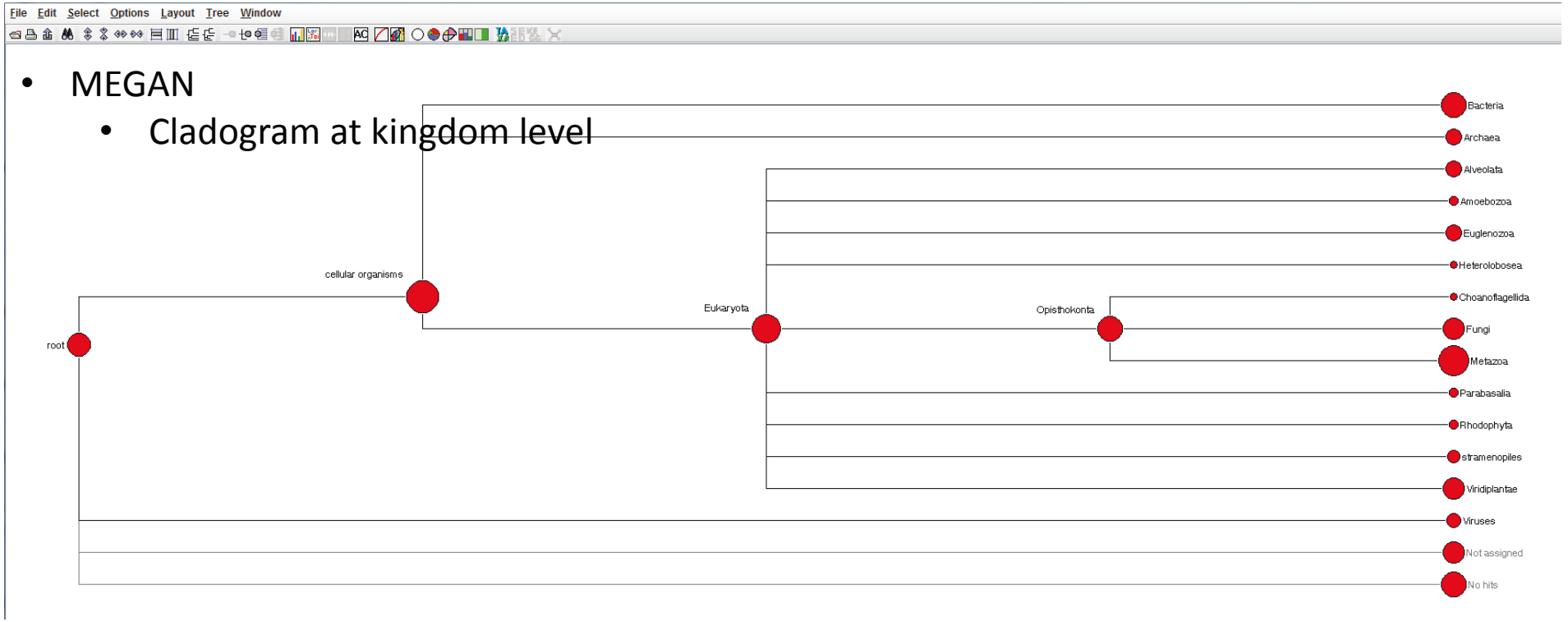
Metagenomics analysis using 16s database



I. Storage: *Example of complexity of the data set*

Metagenomics

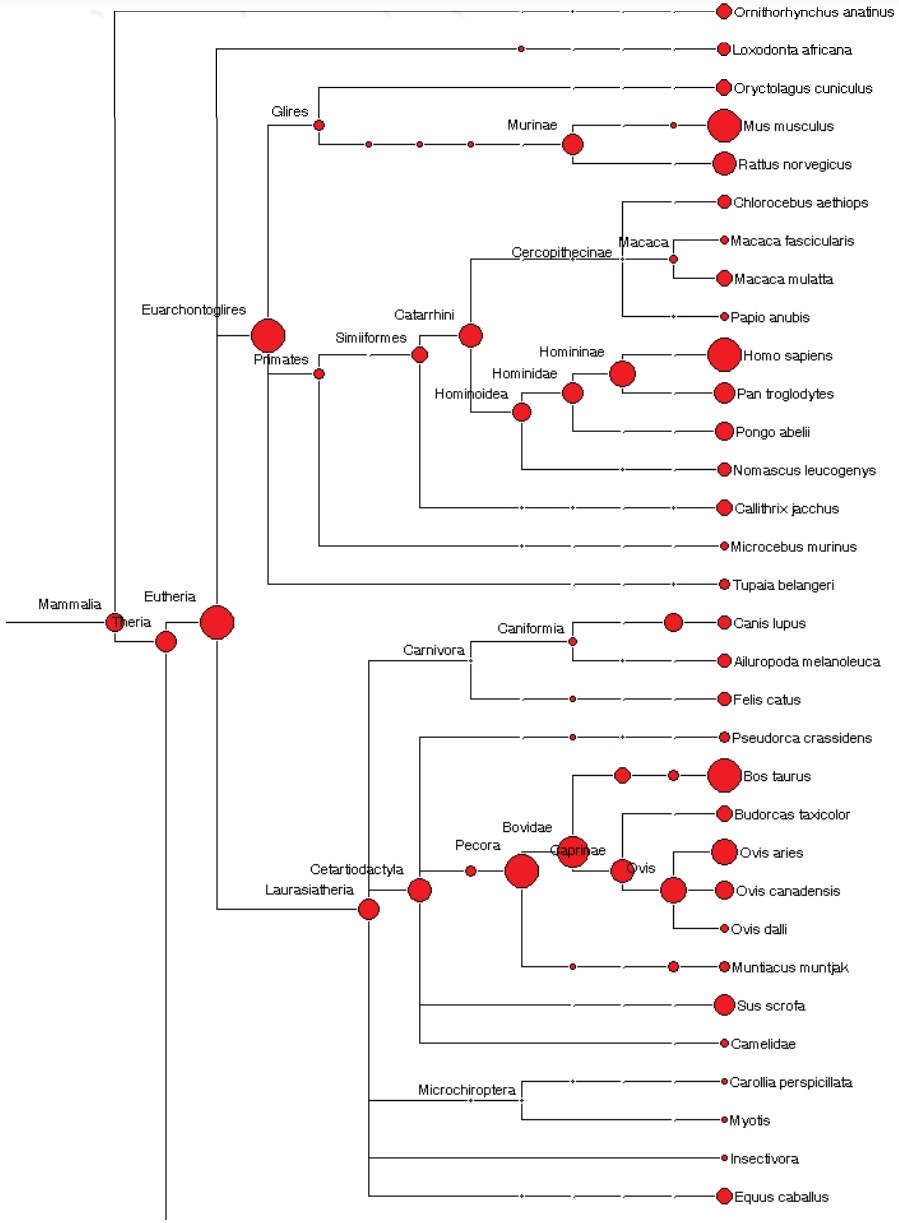
- MEGAN
  - Cladogram at kingdom level



I. Storage: *Example of complexity of the data set*

Metagenomics

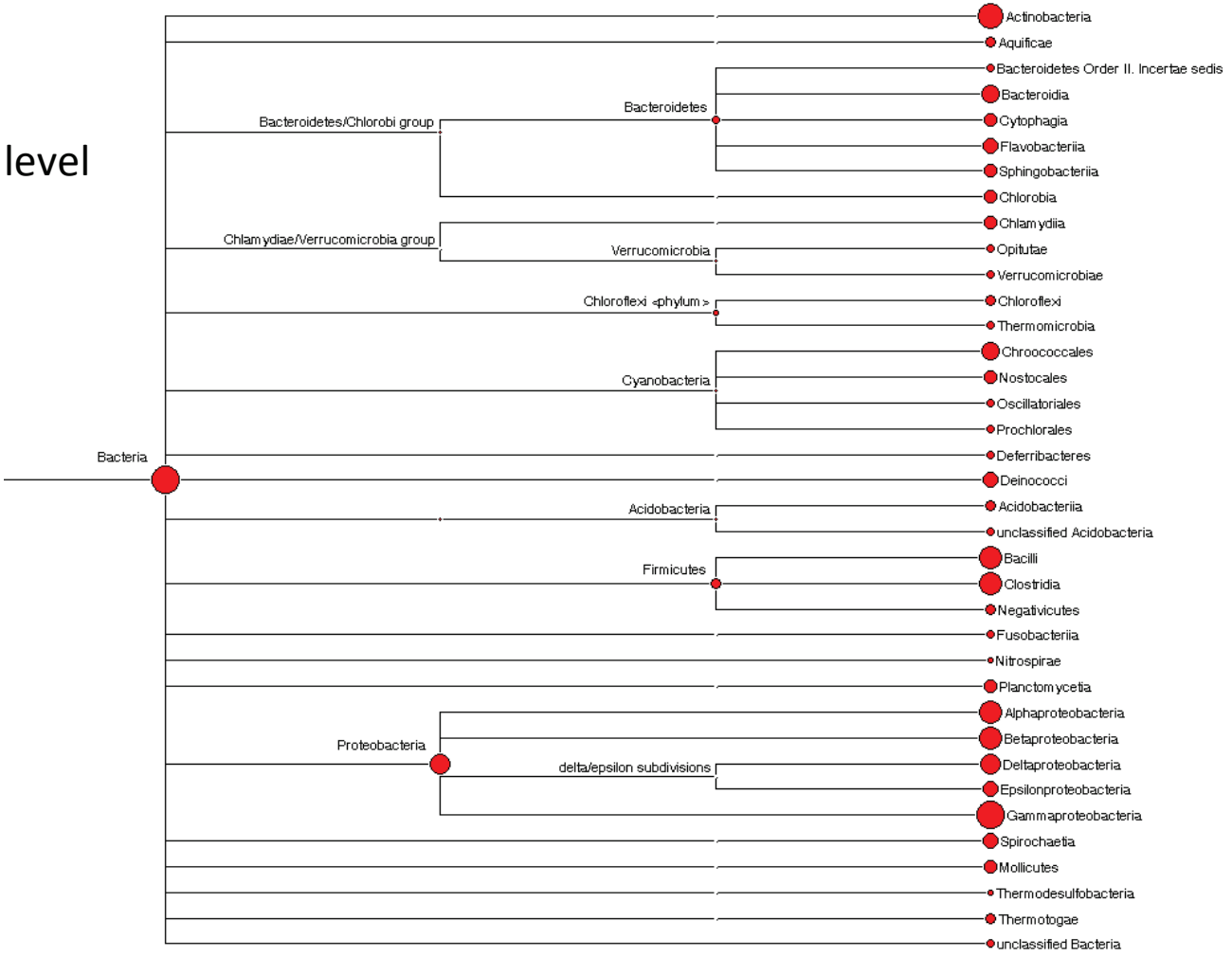
- MEGAN
  - Cladogram at species level



I. Storage: *Example of complexity of the data set*

Metagenomics

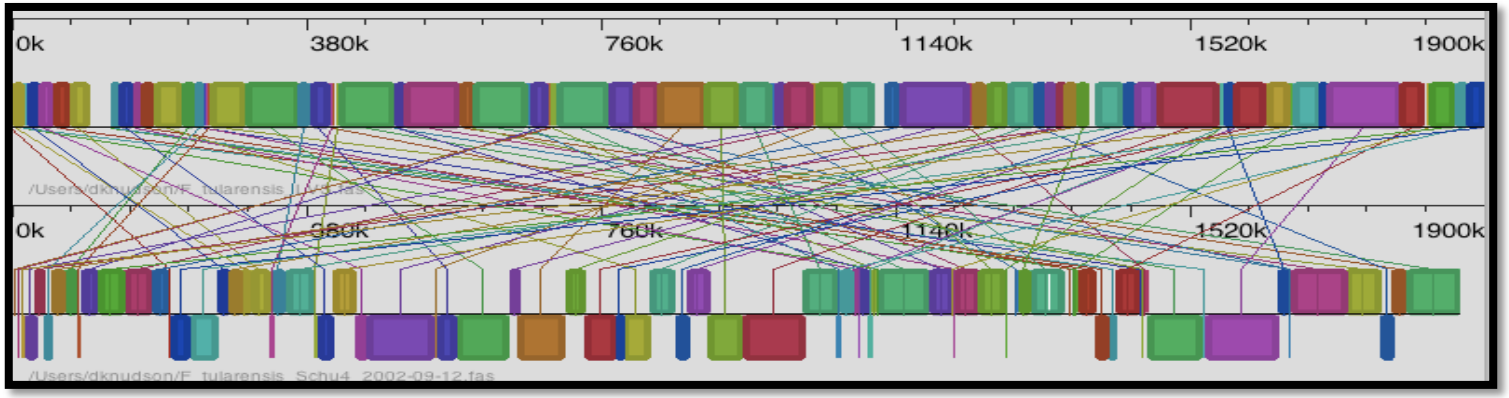
- MEGAN
  - Cladogram at class level



I. Storage: *Example of complexity of the data set*

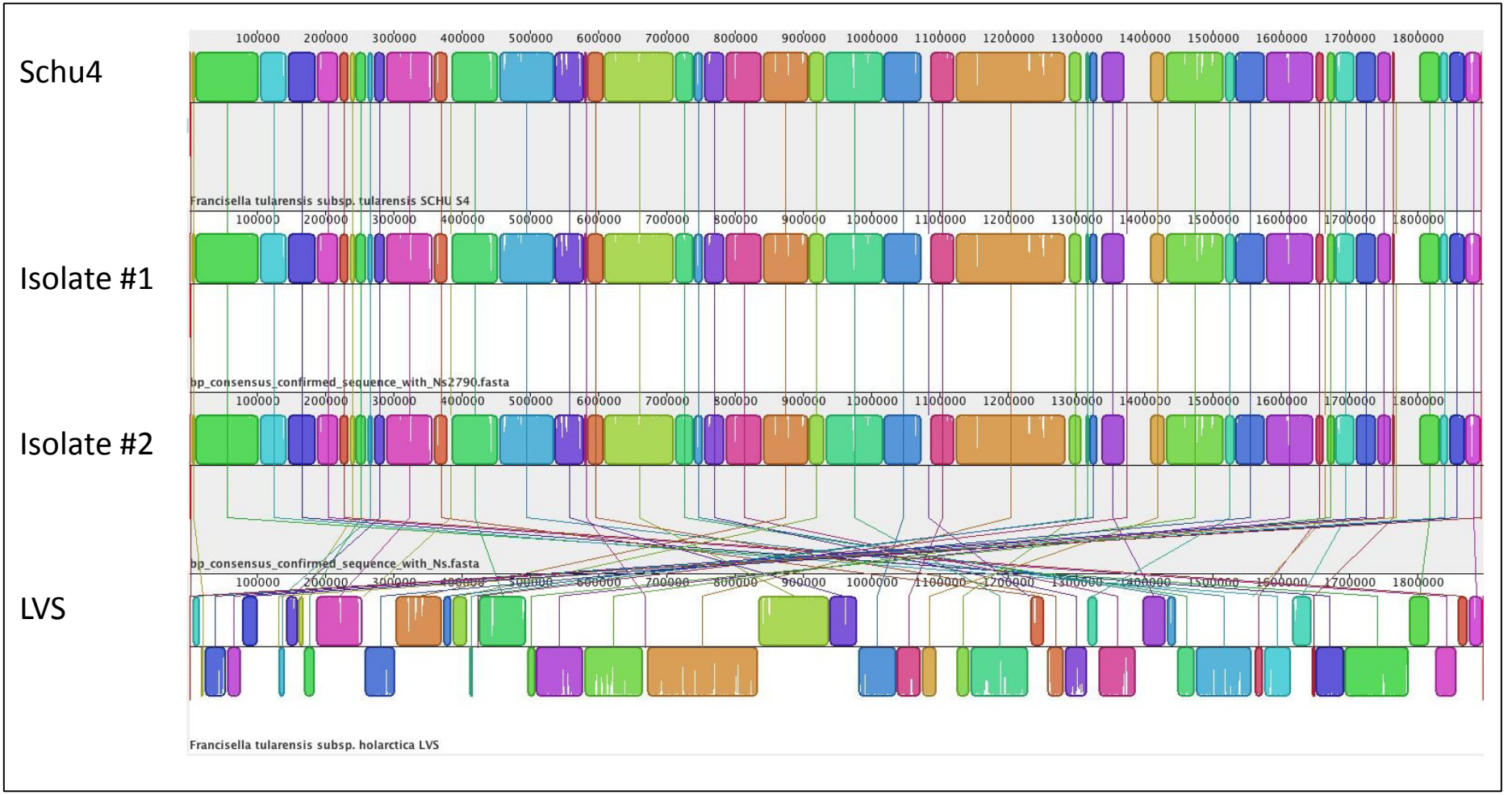
Capturing Biological information and Function

	***Francisella tularensis* Holarctica**	***Francisella tularensis***
Strain	LVS	Schu4
Accession		
Build		2002-9-12 in 37 contigs
Bases	1895998	1798384
GC%	32.15	
ORFs	2109	2056
Duplicate ORFs	132	90
Bases/Orf	899	875
Unique ORFs	1977	1966
Masking Genome		
Fraction masked		
Francisella tularensis Holarctica strain LVS	1	0.9641739
Francisella tularensis strain Schu4	0.9830348	1
Proteins at e=0		
Francisella tularensis Holarctica strain LVS	0	3
Francisella tularensis strain Schu4	2	0
Proteins at e=1e-75		
Francisella tularensis Holarctica strain LVS	0	20
Francisella tularensis strain Schu4	6	0



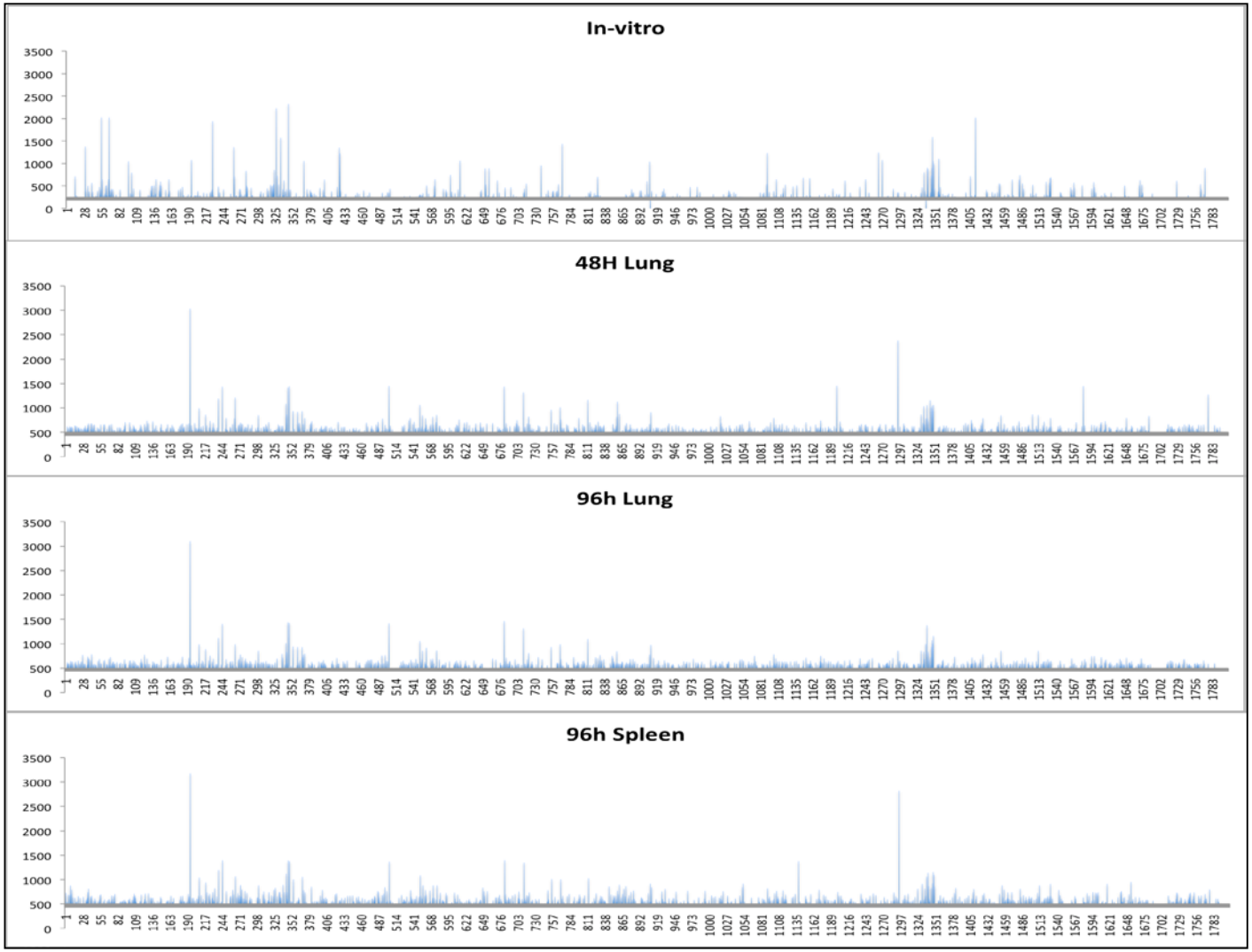
I. Storage: *Example of complexity of the data set*

Genome Analysis-Genome structure and arrangement



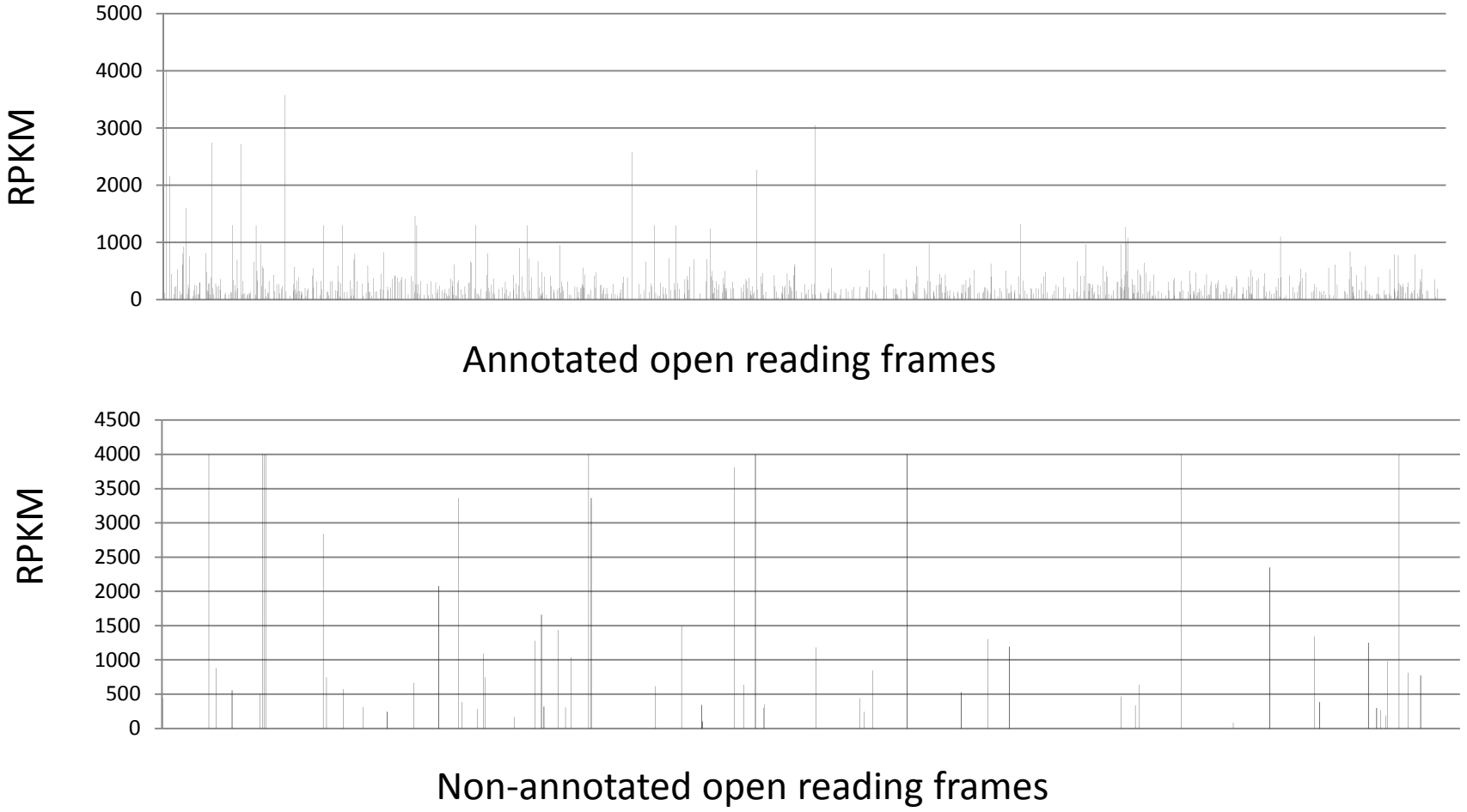
I. Storage: *Example of complexity of the data set*

Capturing Biological information and Function



I. Storage: *Example of complexity of the data set*

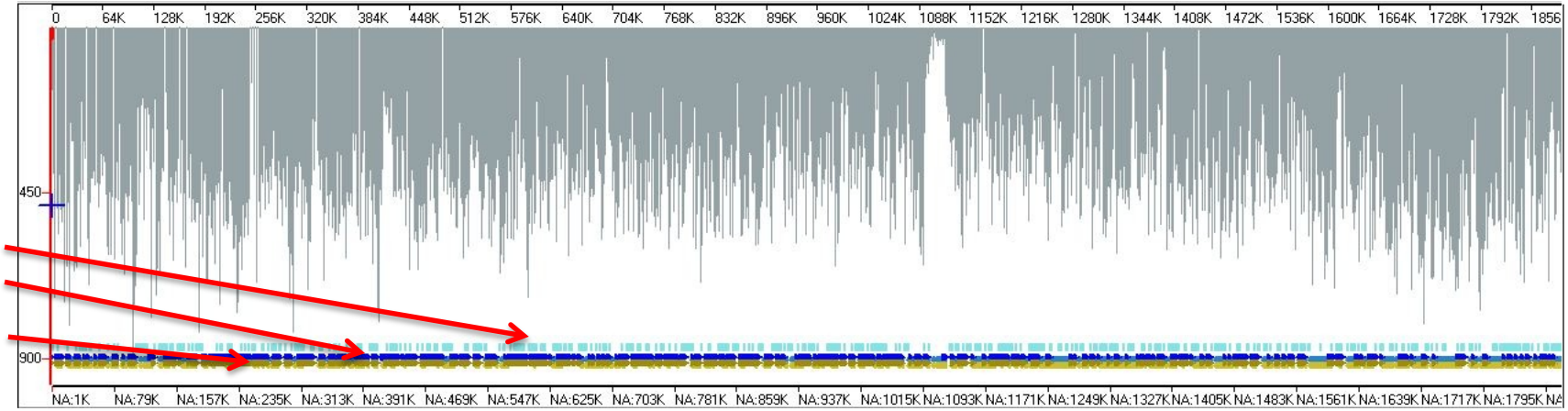
Capturing Biological information and Function



I. Storage: *INTEGRATION OF DIFFERENT SOURCES OF DATA*

Whole genome essential gene mapping

NS SNPs
+ORFs
-ORFs



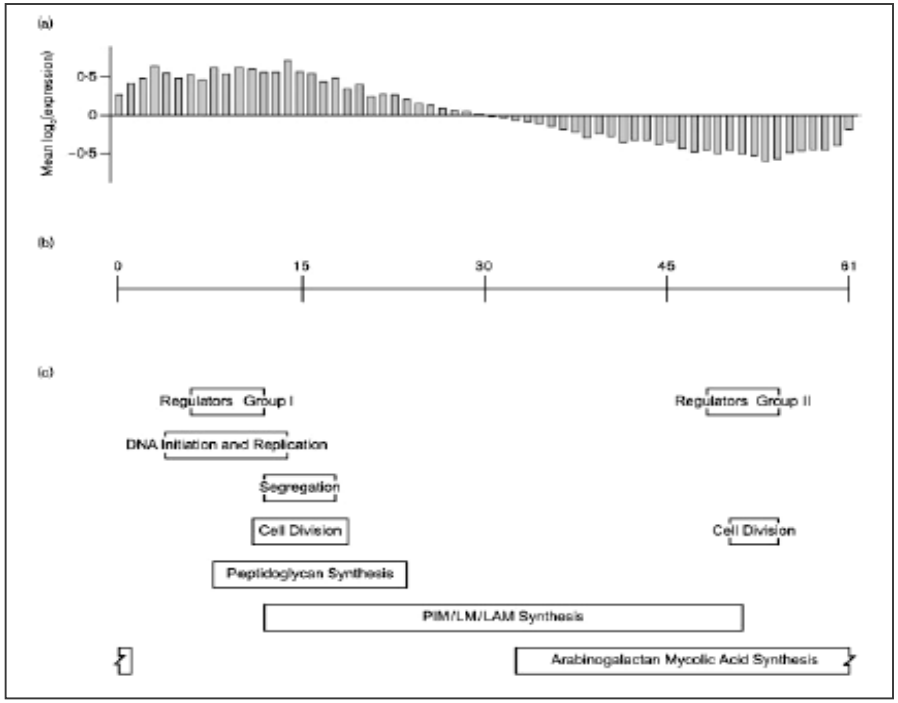
- ✓ *Genome size: ~1.9 million bases*
- ✓ *Input pool: 196,044 mutations (~10%)*
- ✓ *Bacteria from lung: 179,782 mutations*
- ✓ *Bacteria from Spleen: 77,806 mutations*

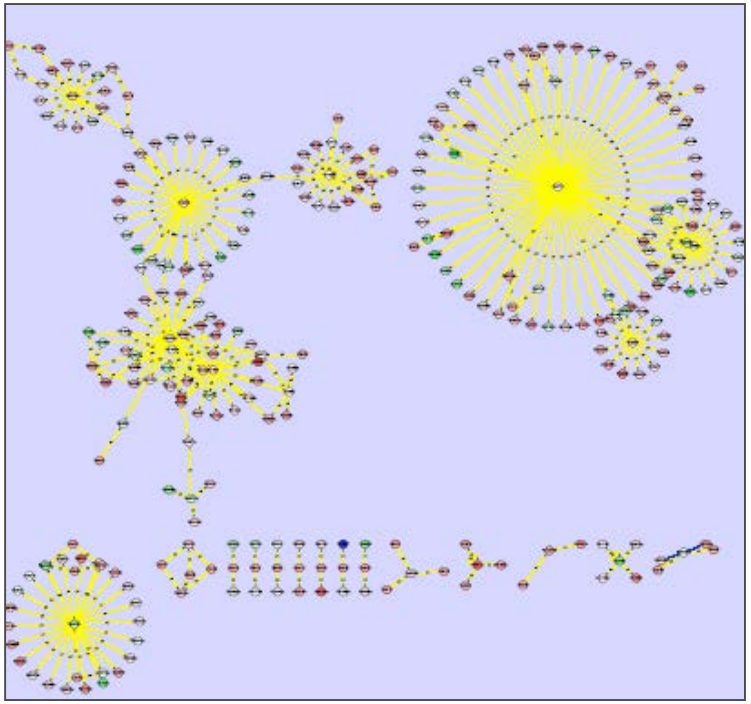
- ➢ Mapped 1,419 unique non-synonymous SNPs across the genome
- ➢ 74% are within proposed open reading frames

I. Storage: *Example of complexity of the data set*

"FUNCTIONAL" INFORMATION

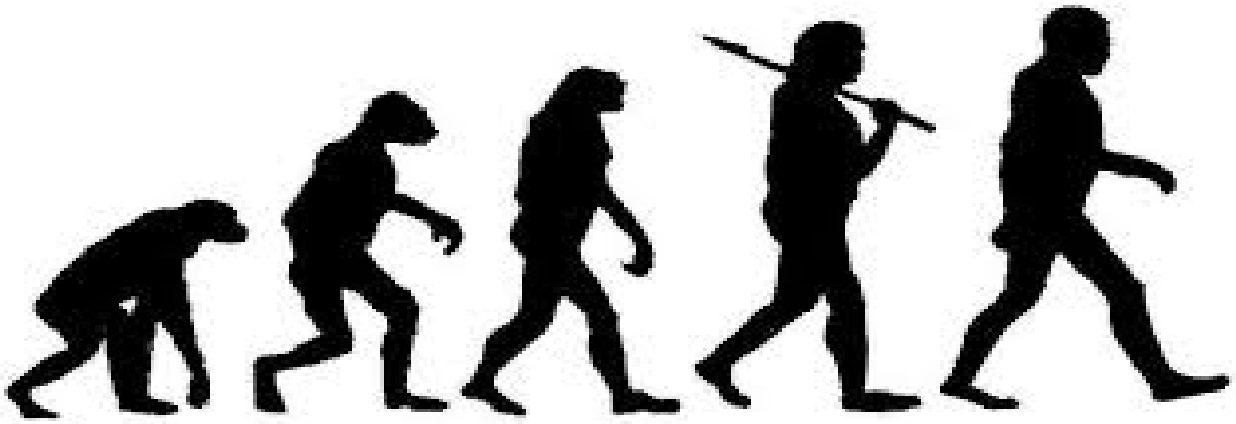
Combine bioinformatics or computational biology and large data sets

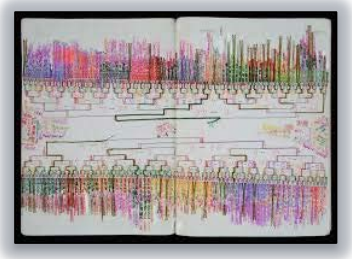


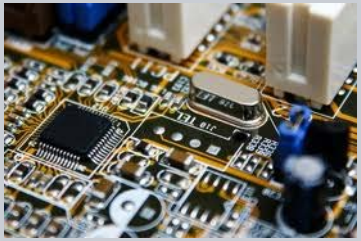


II. Preservation: *Evolution of laboratory data storage*

From recording to logging





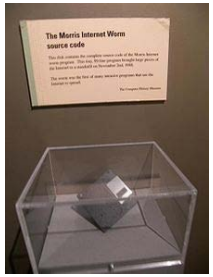


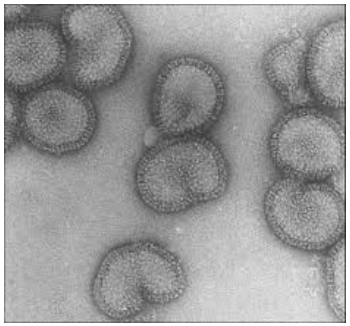
II. Preservation: *Vocabulary & data integrity*

Biologist

“Computationalist” or “data people”

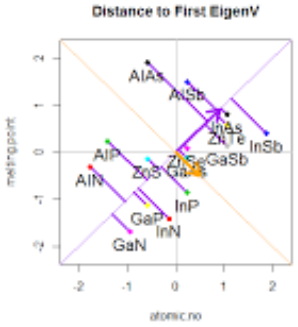
worm





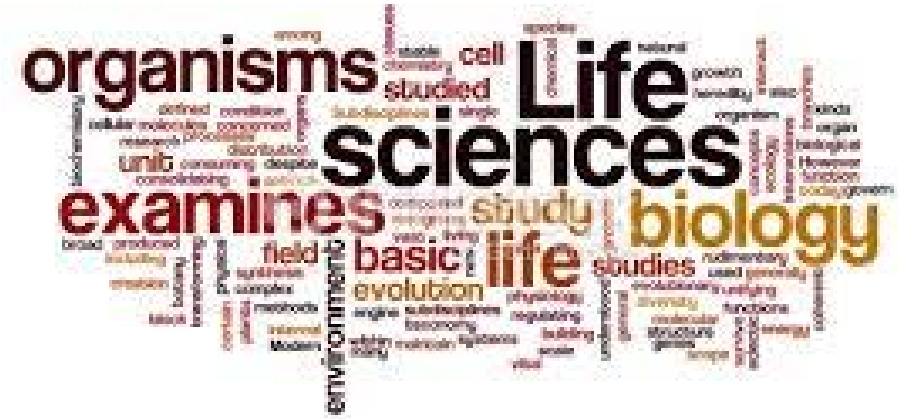
virus

vector



II. Preservation: *Retrieval & key words or search terms*

The "modern" card catalog



CURRENT DATA MANAGEMENT & PRESERVATION STRATEGIES USED BY BIOLOGISTS

Data Management

Data Preservation

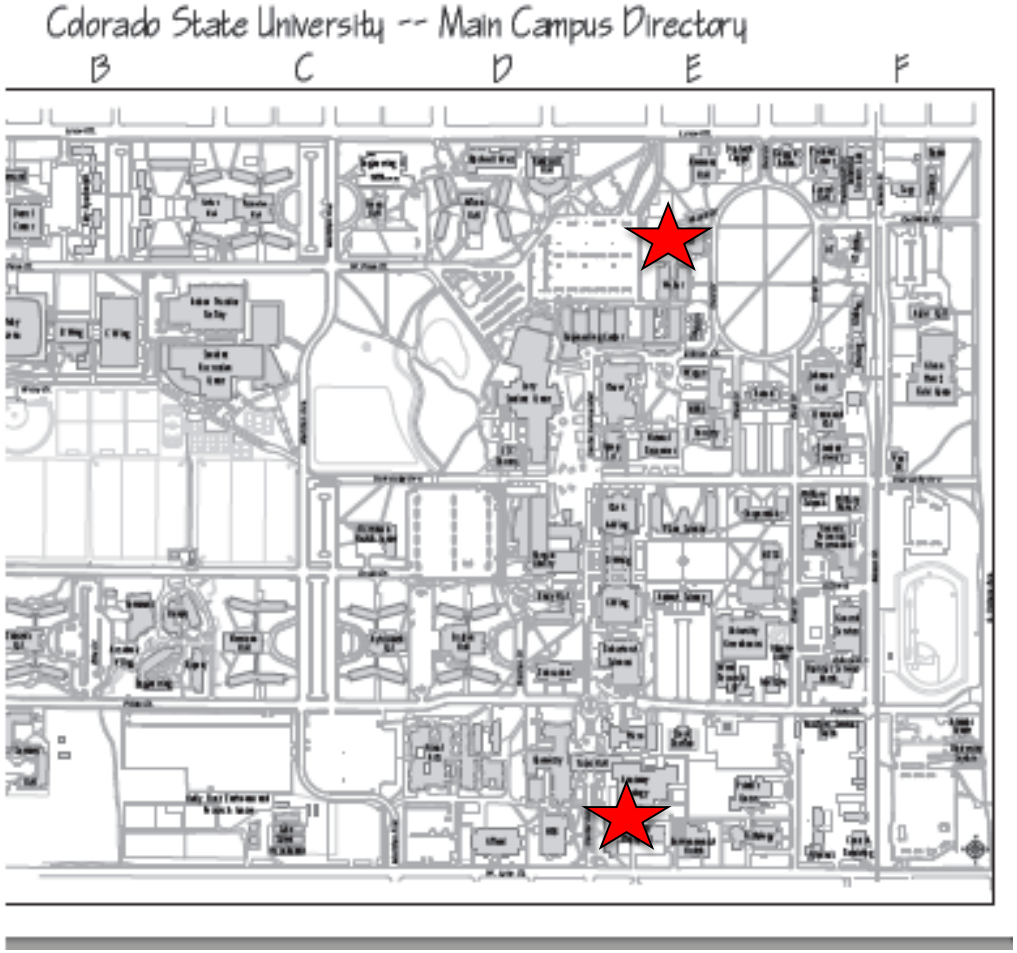
II. Preservation: *Evolution of laboratory data storage*

Current Data Storage systems used by biologists:

- ✓*Individual local computers or servers*
- ✓*Not readily accessible by multi local investigators*
- ✓*Not accessible by outside collaborators*
- ✓*Not routinely backed-up*
- ✓*Deletion of large raw data sets*
- ✓*Data cannot be integrated into multi-investigator programs*

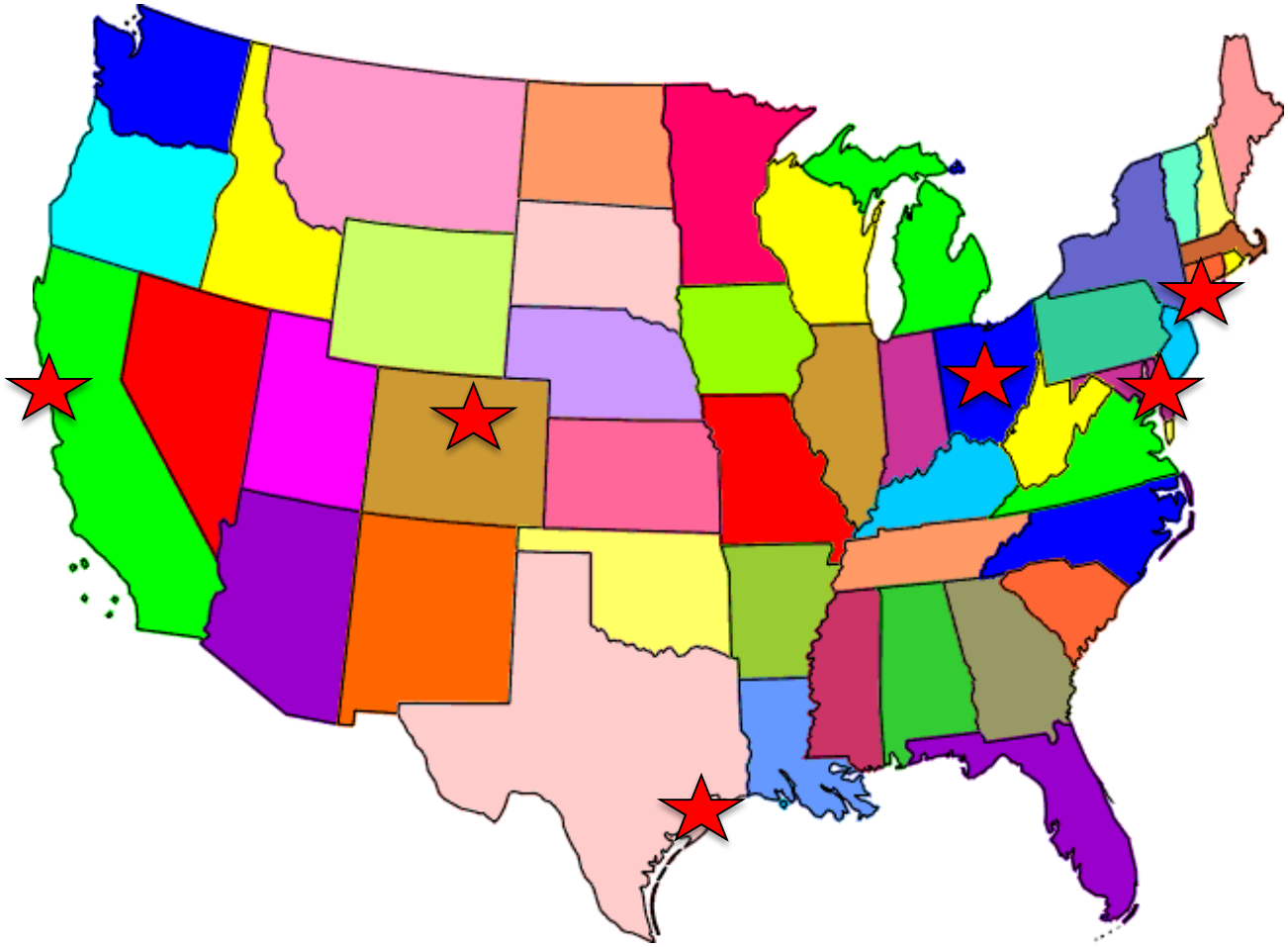
II. Preservation: *Evolution of laboratory data transfer*

Beyond a single laboratory-Data access between experimental sites



II. Preservation: *Evolution of laboratory data transfer*

Beyond a single laboratory-Data access between experimental sites



II. Preservation: *Evolution of laboratory data transfer*

III. Data integrity: *Accidental or intentional*

Many steps involving data manipulation or normalization

1. Information integration- *is the experimental details associated with the data*

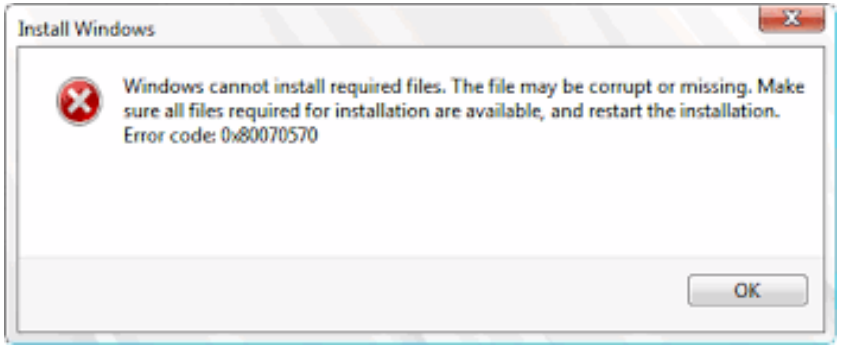
2. Versioning-*how has the data been changed, by who, and for what reason*

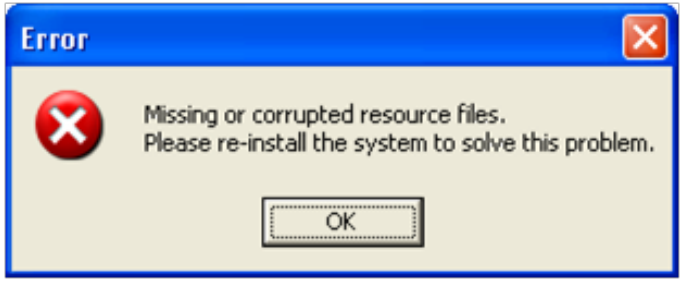
3. Is the data publically available and can the data be audited-*interface with data for manipulation and data analysis & output*

User error > than intentional corruption

III. Data integrity: *Big data troubles*

Where are key features of the data: *What does one do when you need 4,917 files and only receive 4,916 files*





III. Data integrity: *Big data troubles*

Integration of biology and data: *Big data biology and computational analysis is inconsistent in many cases.*

Examples resulting in differences in data outcomes:

- ✓ *Low number of representatives for each group or data set.*
- ✓ *Gravitate to familiarity*
- ✓ *What to do with "missing data"*
- ✓ *Variability in materials, resources, animal species, age, sex & strains*
- ✓ *Level of comfort of researcher(s)*

Approaches: data organization and structure

III. Data integrity: *"deliberately vague"*

"Read me file":

Low information content, that *requires contextual information* for meaning

Lacks or contains very general details or *has limited precision* to description

Applied to data:

Missing details that limits the *extent of the information*

Provided in a format that *cannot be readily integrated* with other information

Impact:

Stalls progress

Provides the opportunity for alternative interpretations based on known uncertainty

Future data management, preservation & integrity needs

Envisioned needs in context of the BIOLOGIST:

1. Data Storage-*where is the data*

2. Maintenance-*has it been changed, if so in what way, and by who, and for what reason*

3. Access to data files-*interface with data for manipulation and data analysis & output*

4. Distribution of data files-*Provide data in "universal" format where state of analysis is embedded and can be integrated with other data*

1. Compatibility of analytical software and future interfaces

2. When is redundancy needed and precision and accuracy?

Future data management, preservation & integrity needs

Envisioned Support Needs provided by COMPUTER SCIENTISTS or DATA PEOPLE:

1. Data Storage-*maintenance, cost, updating hardware, backup, secure, dynamic*
2. Facilitate access to data files-*from remote locations and software-software integration*
3. Movement of data-*without corruption more important than speed*

1. Distribution of data files-*across the US and beyond*

2. Automated work processing-*Dealing with data*

1. "Modern Help Desk"-*move beyond software updates and wireless mouse*

2. Facilitate success & compliance "COLLABORATIVE NOTEBOOK & POLICY"

Questions: