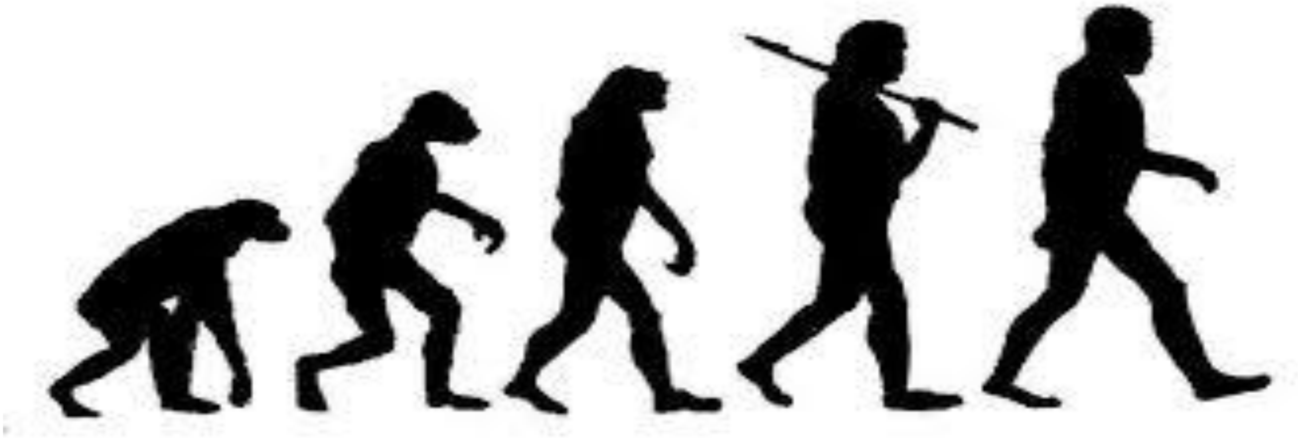
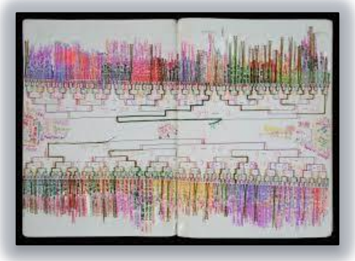
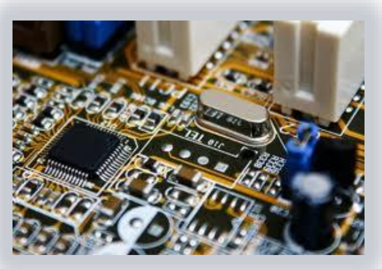
21st Century library: Next Era Biological Data Hurdles for Information Storage, Access, Distribution and Preservation

Richard Slayden, PhD.
Microbiology, Immunology & Pathology
Center for Environmental Medicine

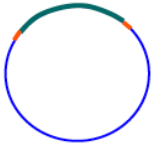
Evolution of the laboratory notebook: *From recording to logging*



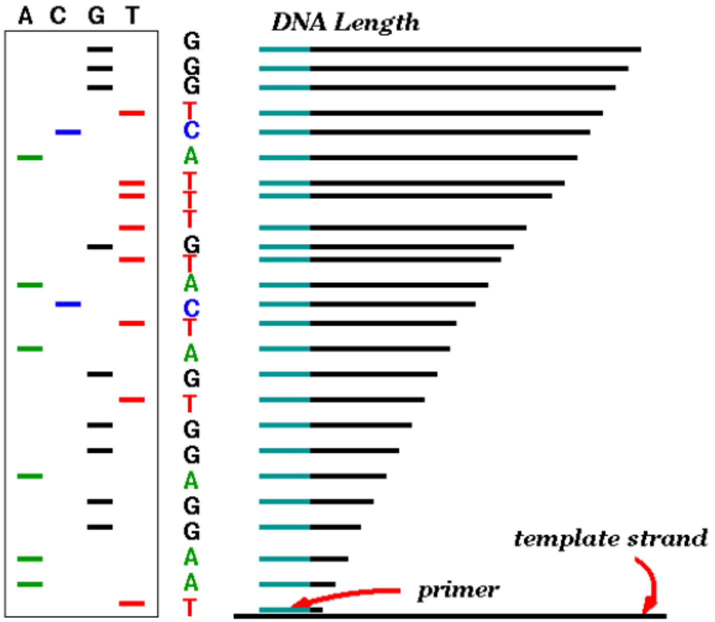




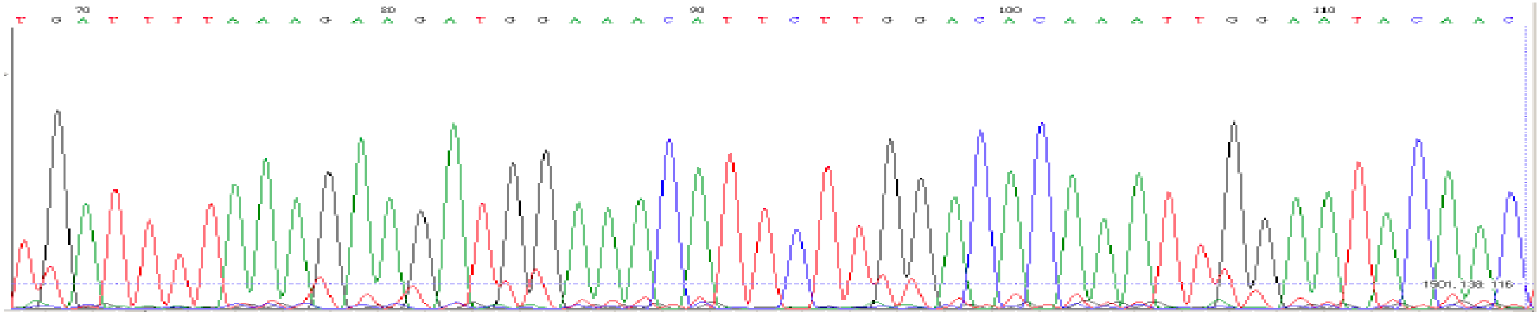
Example of data explosion: *Traditional sequencing*



1. Start at primer
2. Grow DNA chain
3. Include dideoxynucleoside (modified a, c, g, t)
4. Stops reaction at all possible points
5. Separate products with length, using gel electrophoresis



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

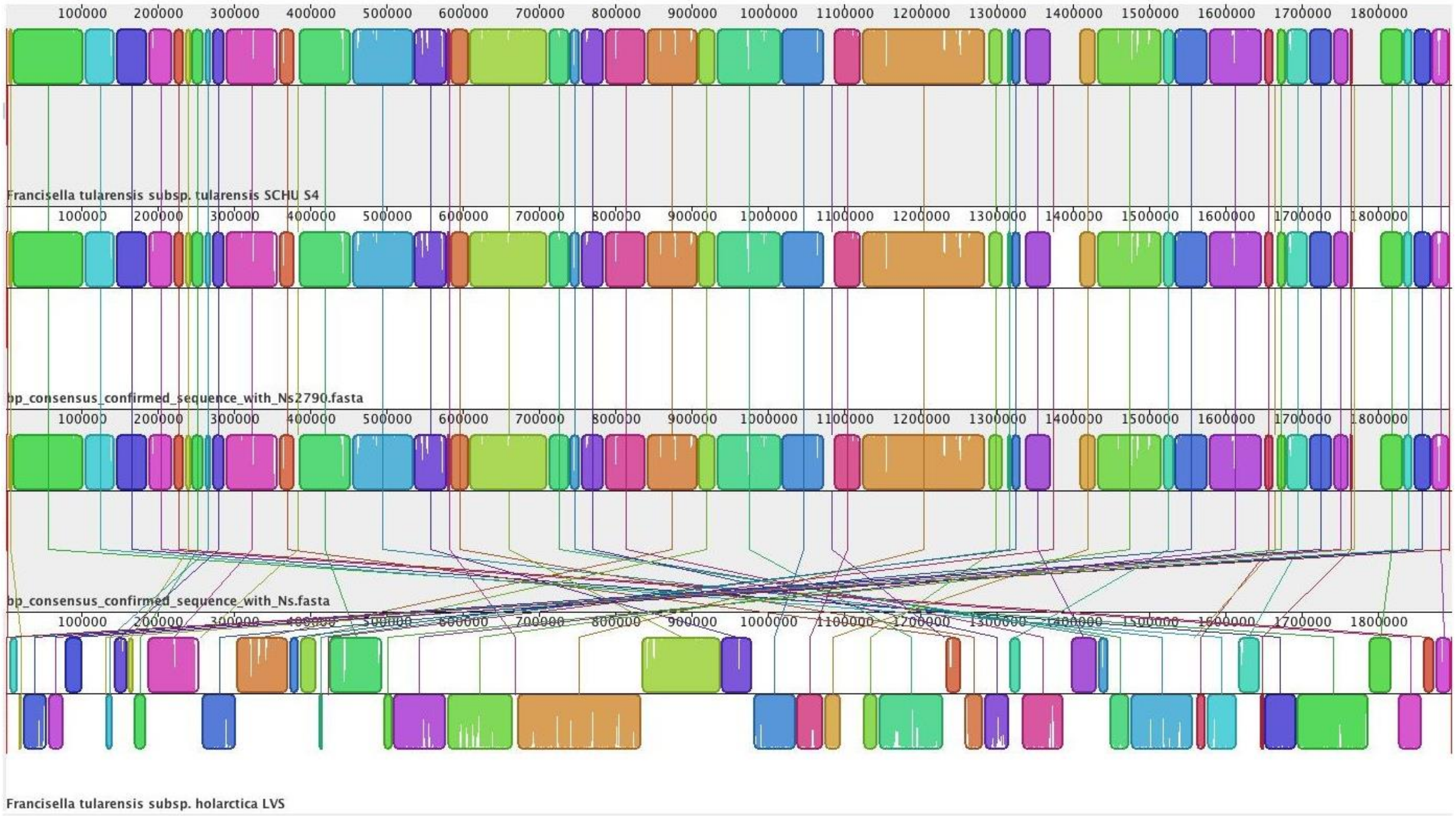


Example of where data is coming from: *Next Generation Sequencing Technology*

Examples of biological data: *Not limited to genome sequencing*

- ✓ Reference or *De Novo* Genome sequence data
- ✓ Resequencing/SNP Analysis
- ✓ Whole Transcriptome/small RNA/microbial RNA/human RNA
- ✓ Epigenetics
- ✓ Gene Essentiality
- ✓ Metagenomic studies

Integration of data: *Genome Analysis-Genome structure and arrangement*


100000 200000 300000 400000 500000 600000 700000 800000 900000 1000000 1100000 1200000 1300000 1400000 1500000 1600000 1700000 1800000
Francisella tularensis subsp. tularensis SCHU S4
bp_consensus_confirmed_sequence_with_Ns2790.fasta
bp_consensus_confirmed_sequence_with_Ns.fasta
Francisella tularensis subsp. holarctica LVS

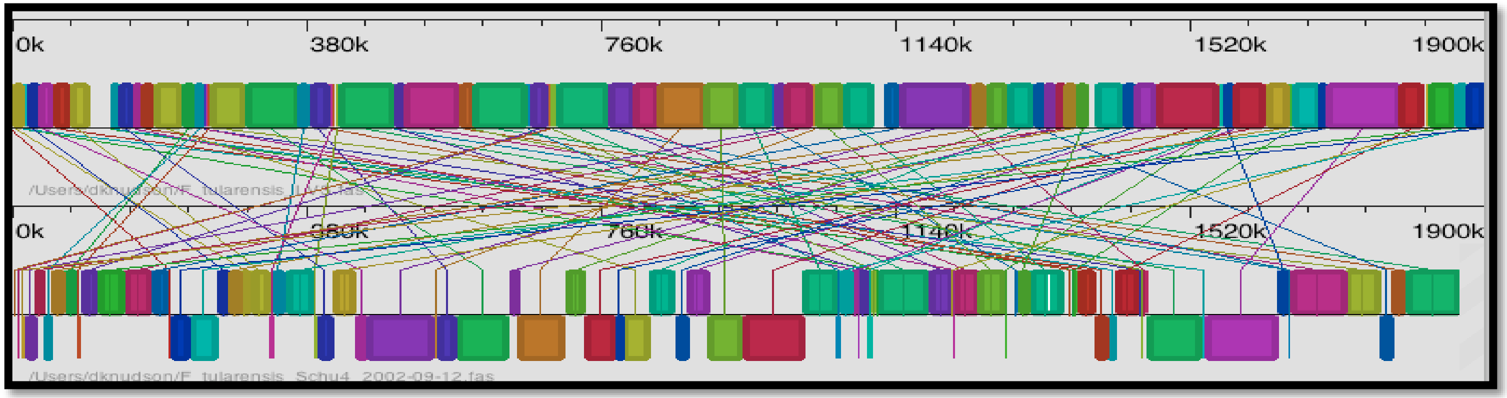
Schu4

Isolate #1

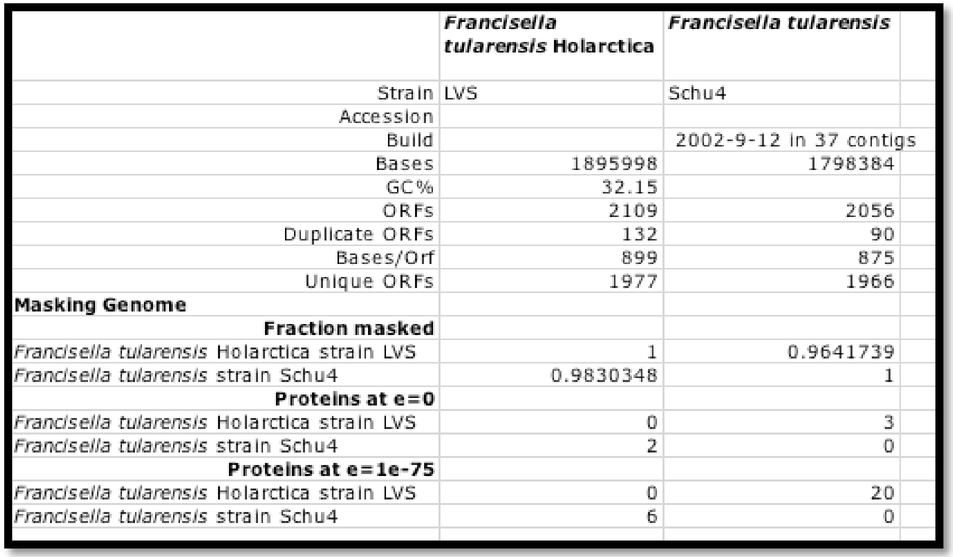
Isolate #2

LVS

Capturing and Updating Biological information and Function

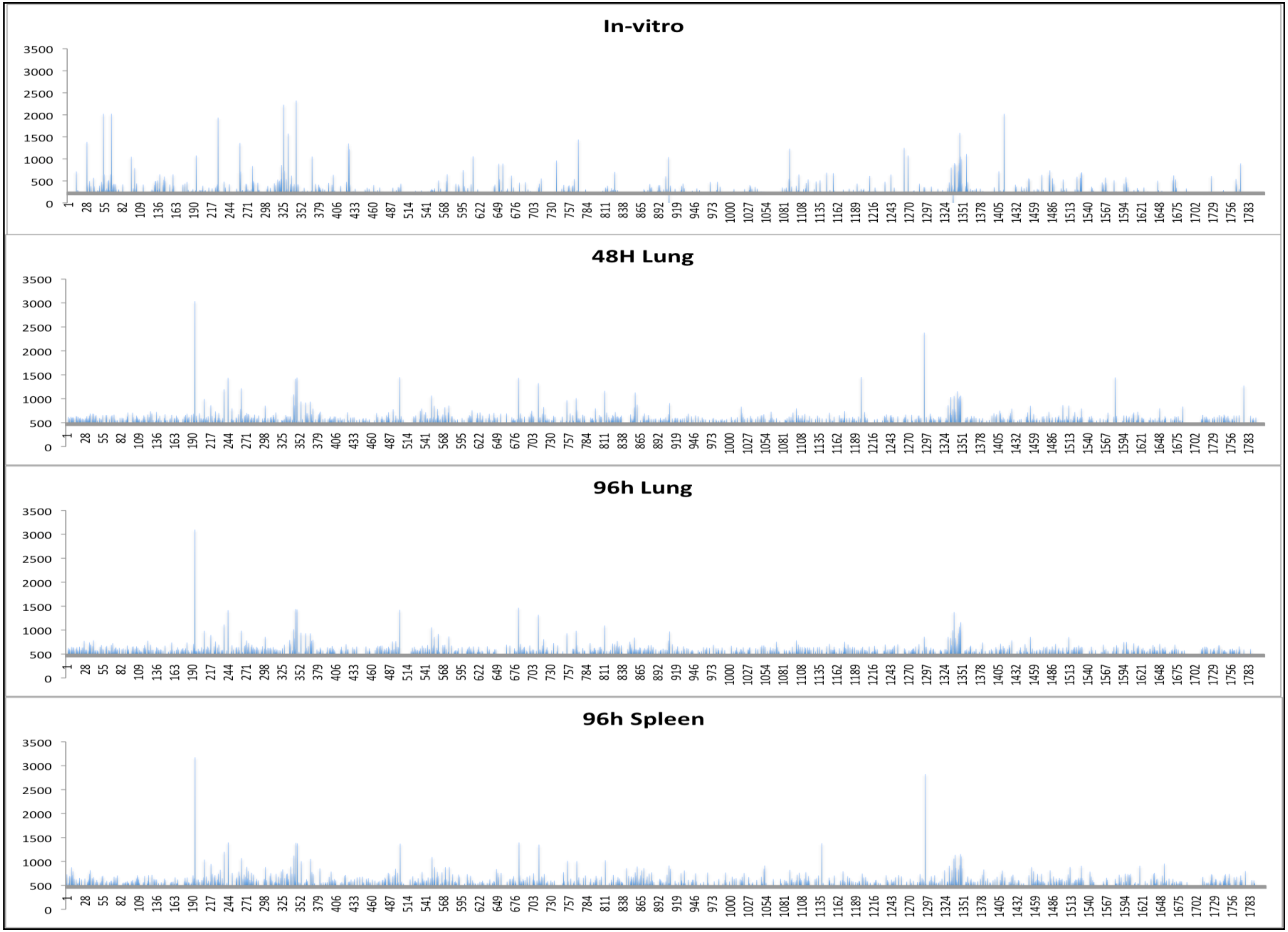


0k
380k
760k
1140k
1520k
1900k
/Users/dknudson/F_tularensis_LVS.fas
0k
380k
760k
1140k
1520k
1900k
/Users/dknudson/F_tularensis_Schu4_2002-09-12.fas

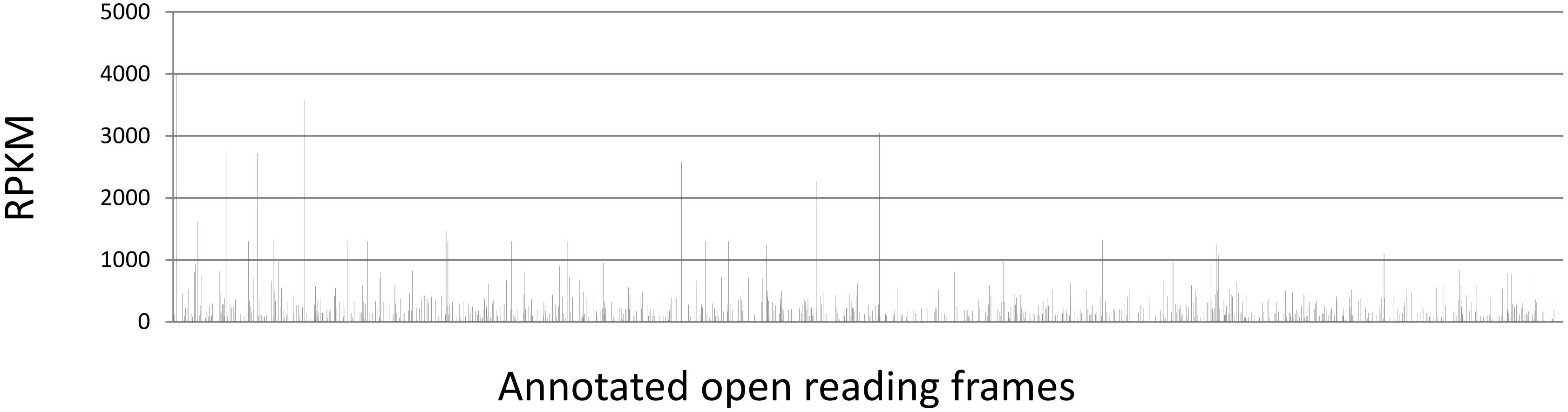


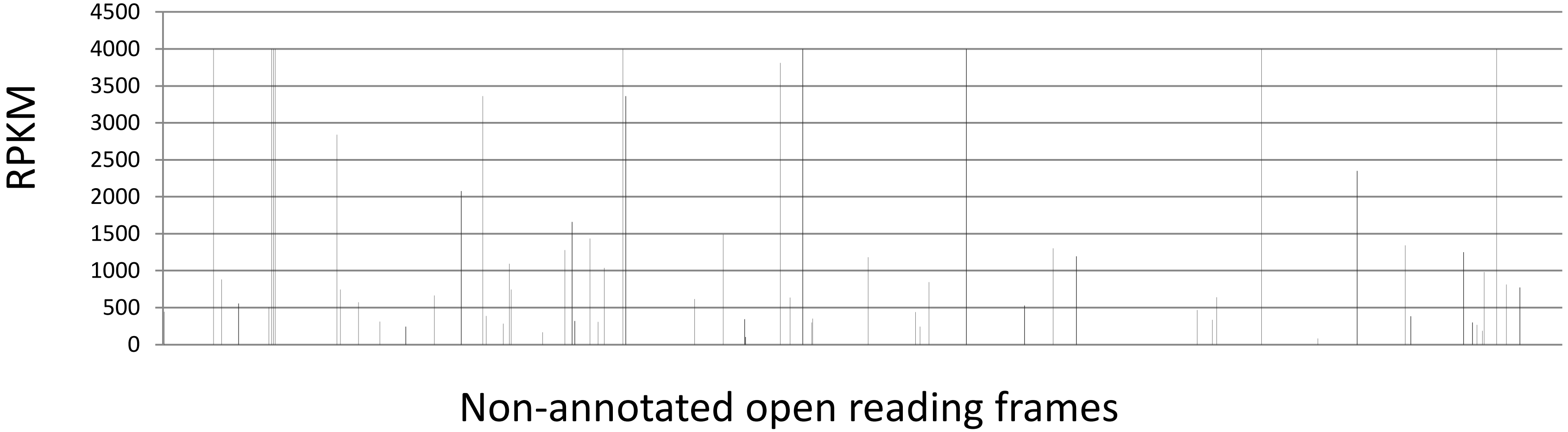
	Francisella tularensis **Holarctica**	***Francisella tularensis***
Strain	LVS	Schu4
Accession		
Build		2002-9-12 in 37 contigs
Bases	1895998	1798384
GC%	32.15	
ORFs	2109	2056
Duplicate ORFs	132	90
Bases/Orf	899	875
Unique ORFs	1977	1966
Masking Genome		
Fraction masked		
Francisella tularensis Holarctica 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

RESOLUTION OF DATA-*UNIQUE DATA FROM A SINGLE INFECTION*

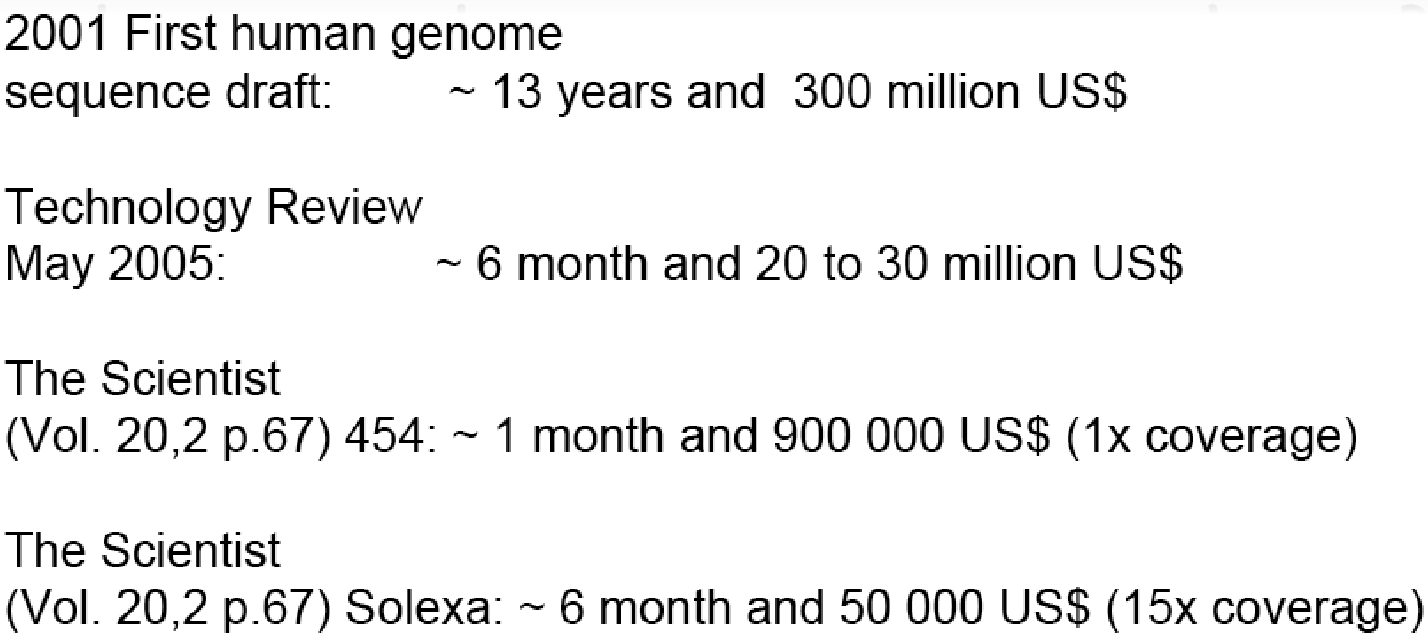


IDENTIFICATION OF NEW GENOMIC INFORMATION: *Assignment of Function*





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)

Example of data explosion: METAGENOMICS 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.

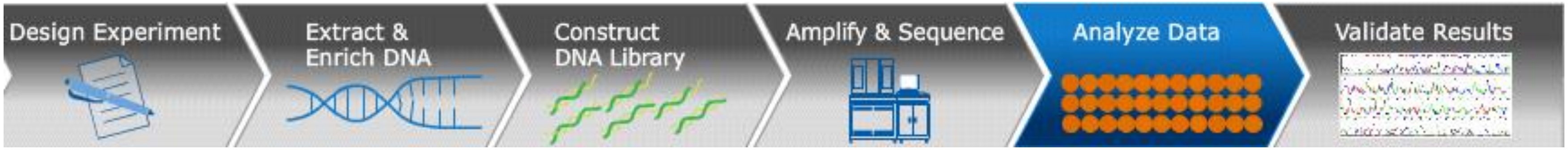
Substantial horsepower, or some data reduction methods, or fairly small highly targeted databases, to make feasible runs.

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 handle metagenomic-sized datasets.

A metagenomics data analysis pipeline is in place for handling NGS sequence data. It’s available to anyone using the CSU sequencers.

Complexity of the data set: *From the Bench to the Data*

Workflow & complexity of the information required



Next Generation Sequencing: *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***

Current Data issues

Current Data Storage:

✓*Individual local computers or servers*

✓*Not readily accessible by multi local investigators*

✓*Not accessible by outside collaborators*

✓*Not routinely backed-up*

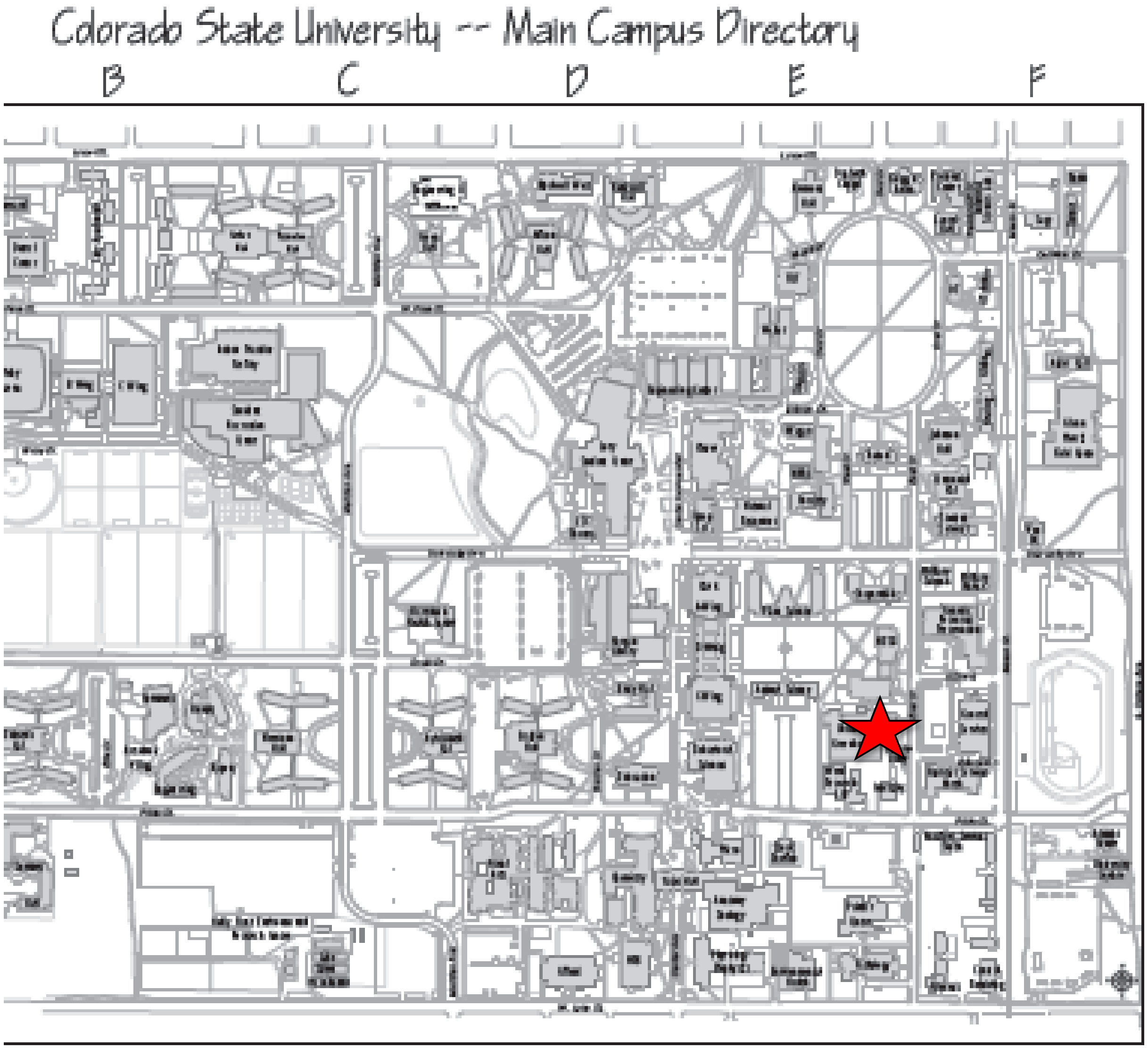
✓*Deletion of large raw data sets*

✓*Data cannot be integrated into multi-investigator programs*

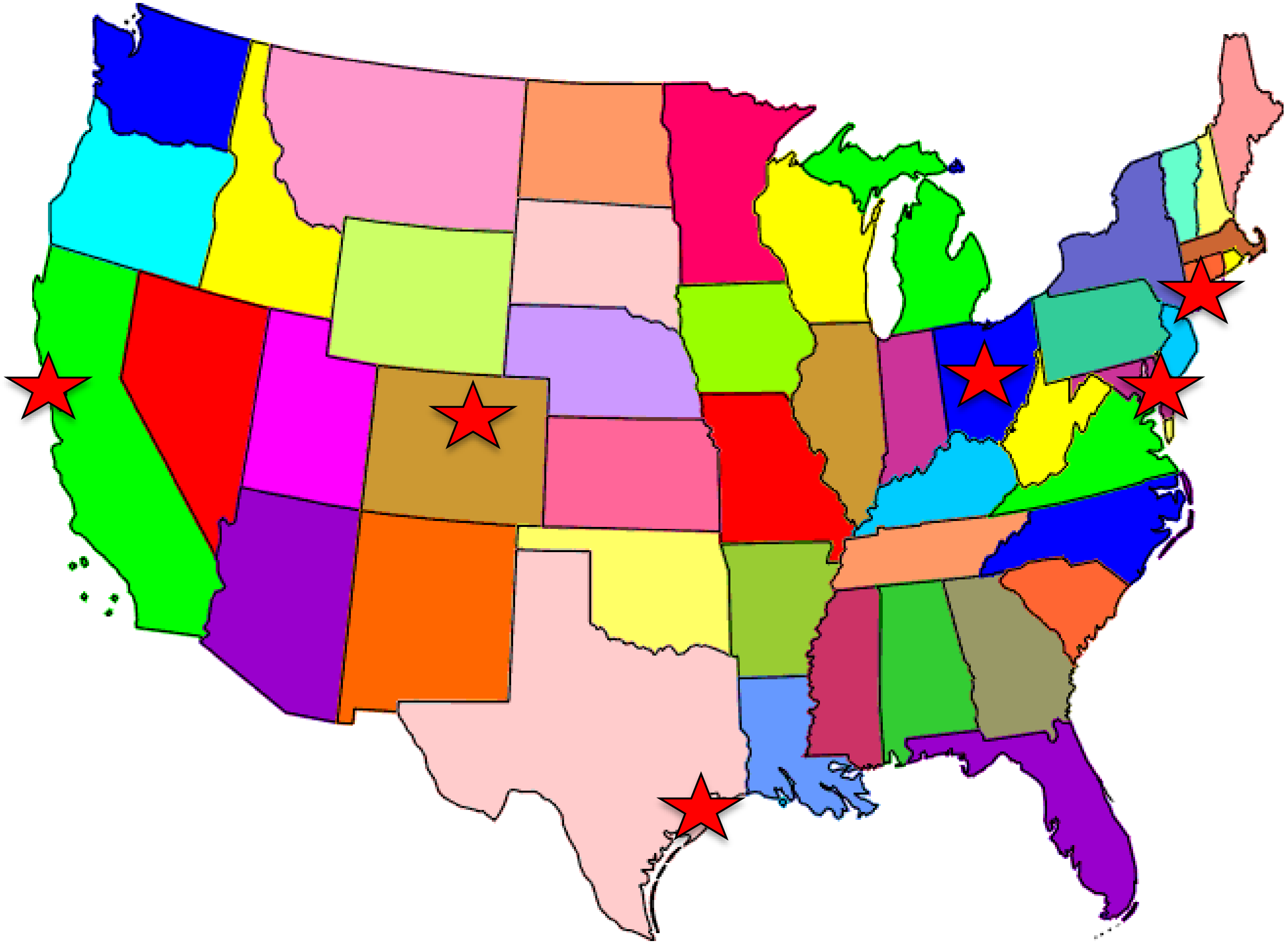
Beyond a single laboratory-Data access between experimental sites

★

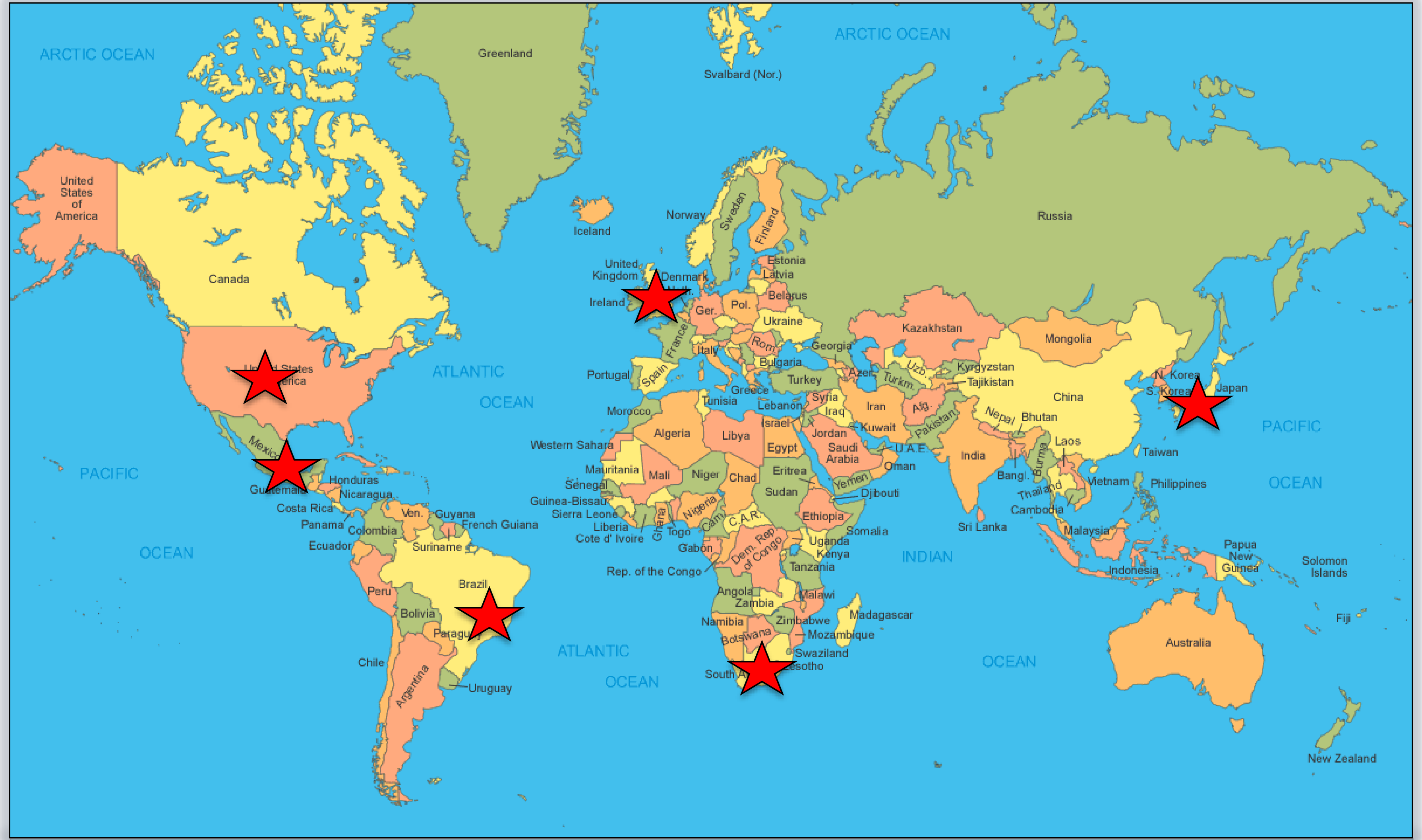
Foothill Campus



Beyond a single laboratory-interaction between experimental sites



Beyond a single laboratory-interaction between experimental sites



CURRENT DATA MANAGEMENT & PRESERVATION STRATEGIES USED BY BIOLOGISTS

Data Management

Data Preservation

What experimental data makes up information?

Study Design
Experiment
Complexity
Data Reduction
Analysis

Information

Current data issues

✓Sequencing: 1-400 genomes (bacterial)

✓Analysis: reference annotation vs re-annotation

✓Source of data: Historical data or newly generated

✓Integration of biological information, data complexity and "version"

Envisioned Support Needs

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-*send data from remote location and analysis*

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

4. Facilitate the development of the COLLABORATIVE LABORATORY NOTEBOOK

Envisioned Needs in context of the BIOLOGIST:

1. Data Storage-*where is the data*

2. Access & 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

The 21st Century Library

Problem solving in 3 phases:

1. Information gathering
   a. *ISTeC*
   b. *ISTeC committee*
   c. *Surveys*
   d. *collaborators*

2. Assessment & validation
   a. *Develop a plan-Initiatives 1-4*
   b. *Provide committee report document-*
   c. *Roll-out to faculty-data management forum & exit survey*

3. Implementation-present to 2020 vision.

The 21st Century Library-current state

“The Initiatives” a biologists perspective

1. Affinity Groups-*misery loves company*
    a. Crowd sourcing strategies & approaches

1. Education-*who, what, where, when & WHY*

1. Physical infrastructure-*Library of the 21st century*

1. Administrative framework-*Facilitation, and sustainable (not regulation)*

Note: bioinformatics/comp-bio is not included

The 21st Century Library-proposed path forward

“Action plan” from my perspective

1. Build on ISTeC Committee report

2. Roll-out to faculty (forum)-*get feed back (exit survey)*

3. Prioritize initiatives based on faculty feed back identify others

4. Organize around initiatives-
   i. *Identify key people willing to participate*
   ii. *Envision the future demand & expectations*
   iii. *Cost*

5. Implementation-*Time frame for achievement [Sparc & Vision 2020]*
   i. *Centralized or decentralized model or hybrid (University/Colleges)*
   ii. *Centralized funding or DC from investigators (fee for service) comb*

thoughts?

Are these the most appropriate initiatives (1-4)?

Other initiatives?

Follow-up opportunity.