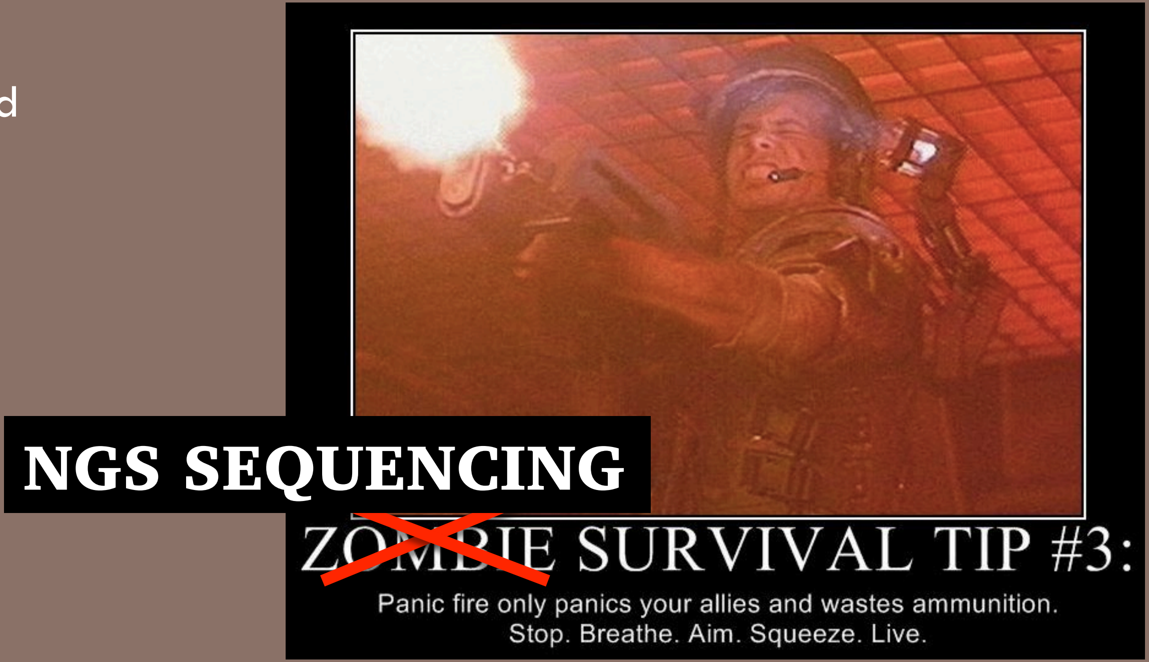
WHAT TO DO IN THE EVENT OF A DATA DELUGE

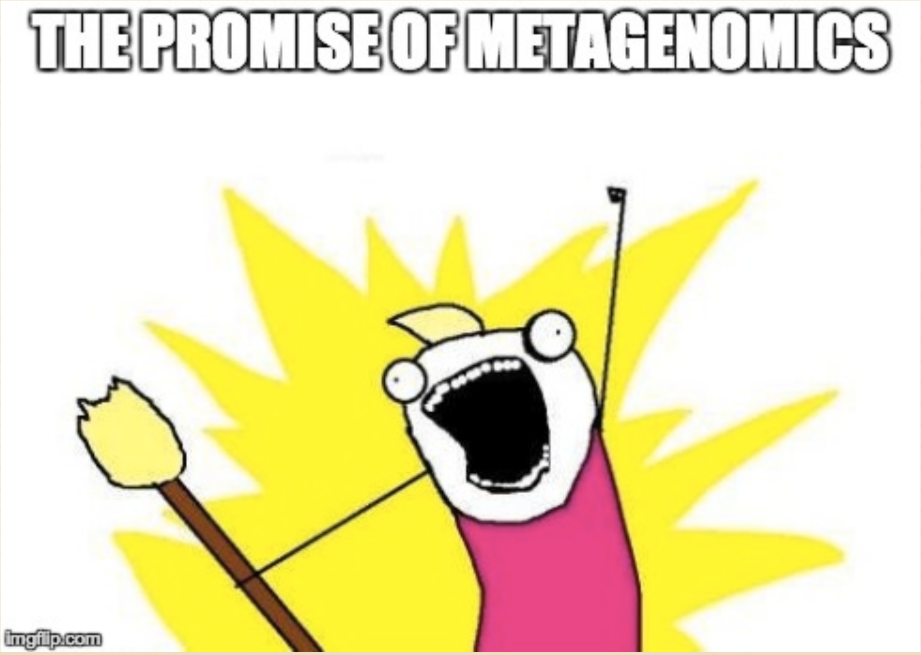
Adina Howe
germslab.org (Genomics and Environmental Research in Microbial Systems)

Iowa State University, Ag & Biosystems Engr (January)

Slides available at www.slideshare.com/adinachuanghowe


NGS SEQUENCING
~~ZOMBIE~~ SURVIVAL TIP #3:
Panic fire only panics your allies and wastes ammunition.
Stop. Breathe. Aim. Squeeze. Live.

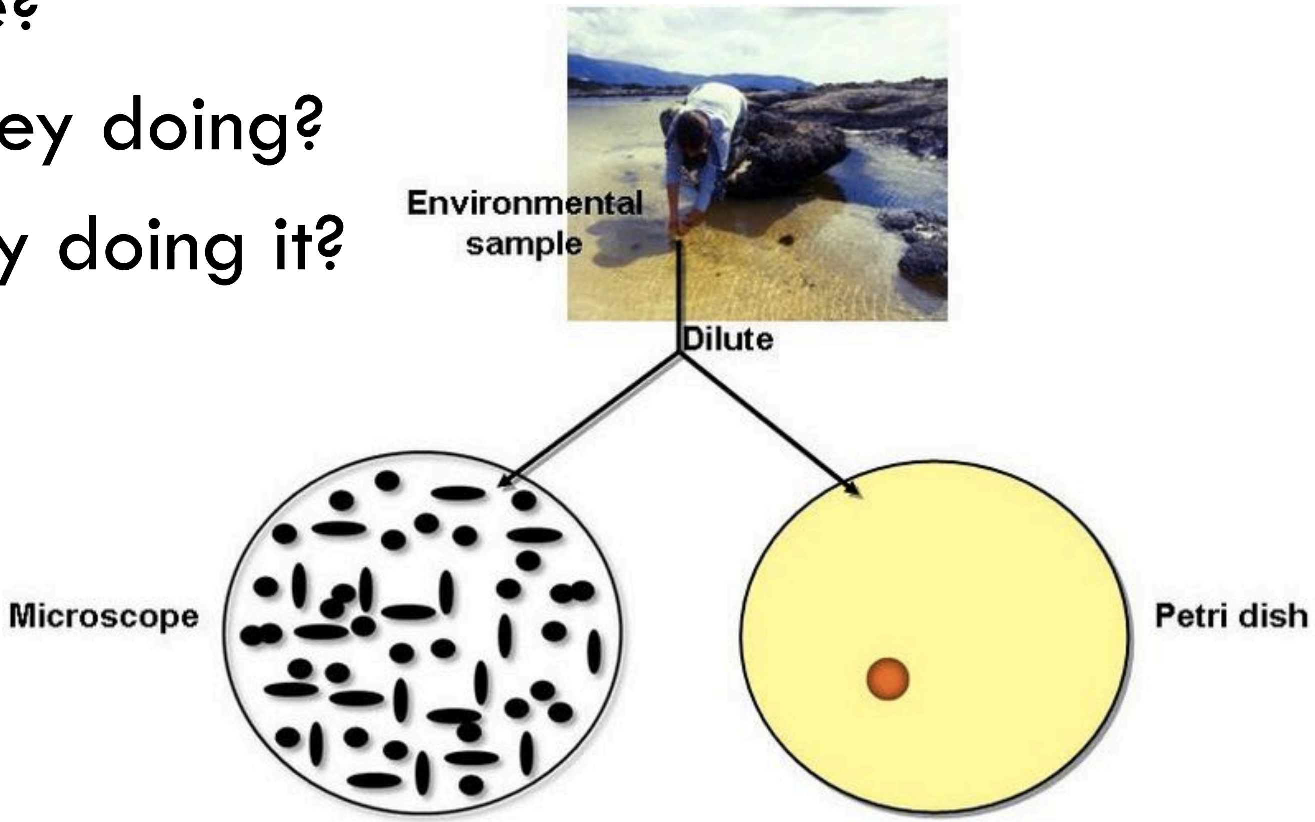
CIBNOR workshop, La Paz, 5/28/2015



HOW DID WE GET HERE

Understanding community dynamics

- □ Who is there?
- □ What are they doing?
- □ How are they doing it?



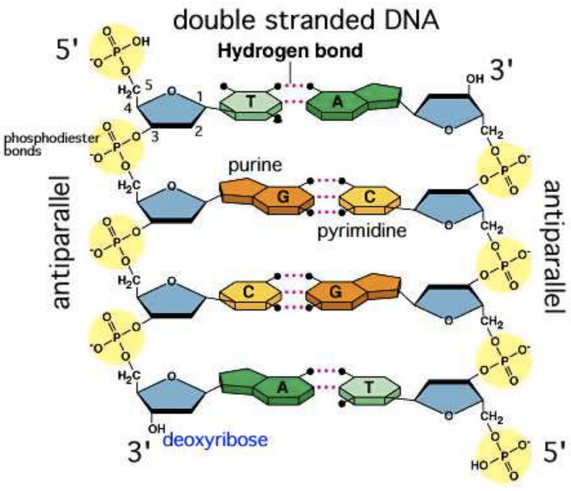
"THE GREAT PLATE COUNT ANOMALY"

~100 times > cells than colonies, 99% unculturable

Kim Lewis, 2010

Gene / Genome Sequencing

- □ Collect samples
- □ Extract DNA
- □ Sequence DNA
- □ “Analyze” DNA to identify its content and origin



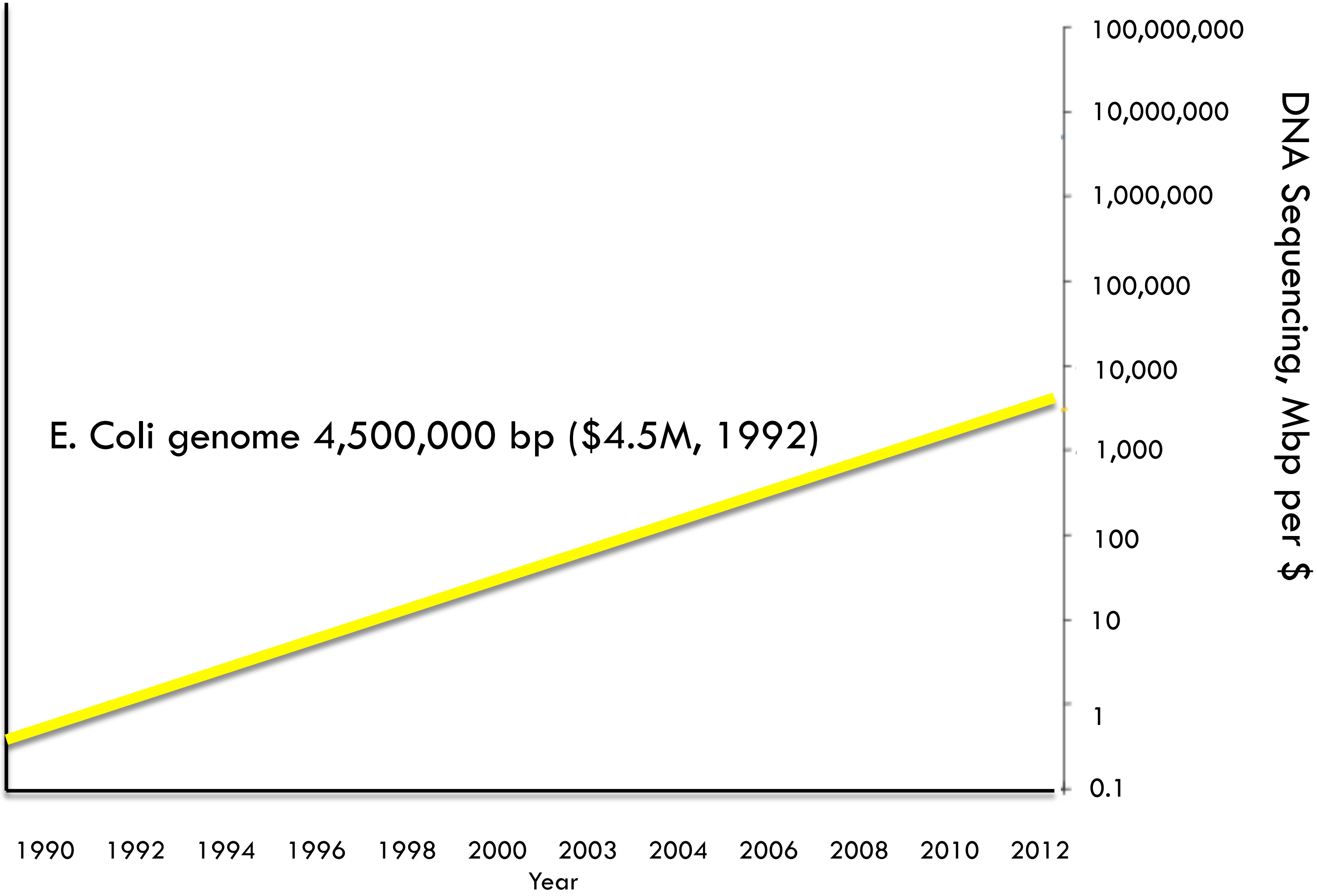
Taxonomy

(e.g., pathogenic E. Coli)

Function

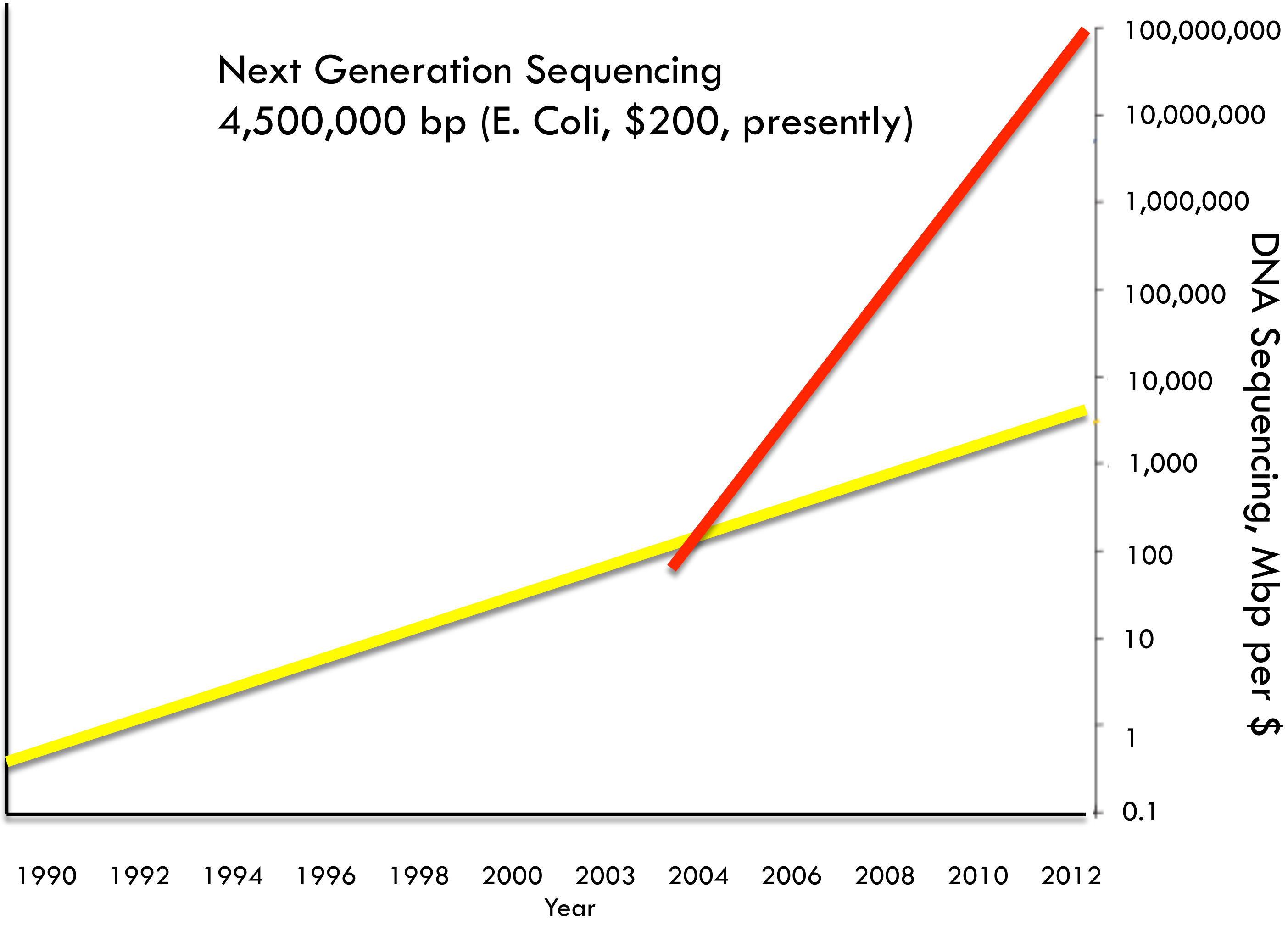
(e.g., degrades cellulose)

Cost of Sequencing


E. Coli genome 4,500,000 bp ($4.5M, 1992)

Stein, Genome Biology, 2010

Rapidly decreasing costs with NGS Sequencing



Stein, Genome Biology, 2010

Effects of low cost sequencing…

First free-living bacterium sequenced for **billions of dollars** and **years** of analysis

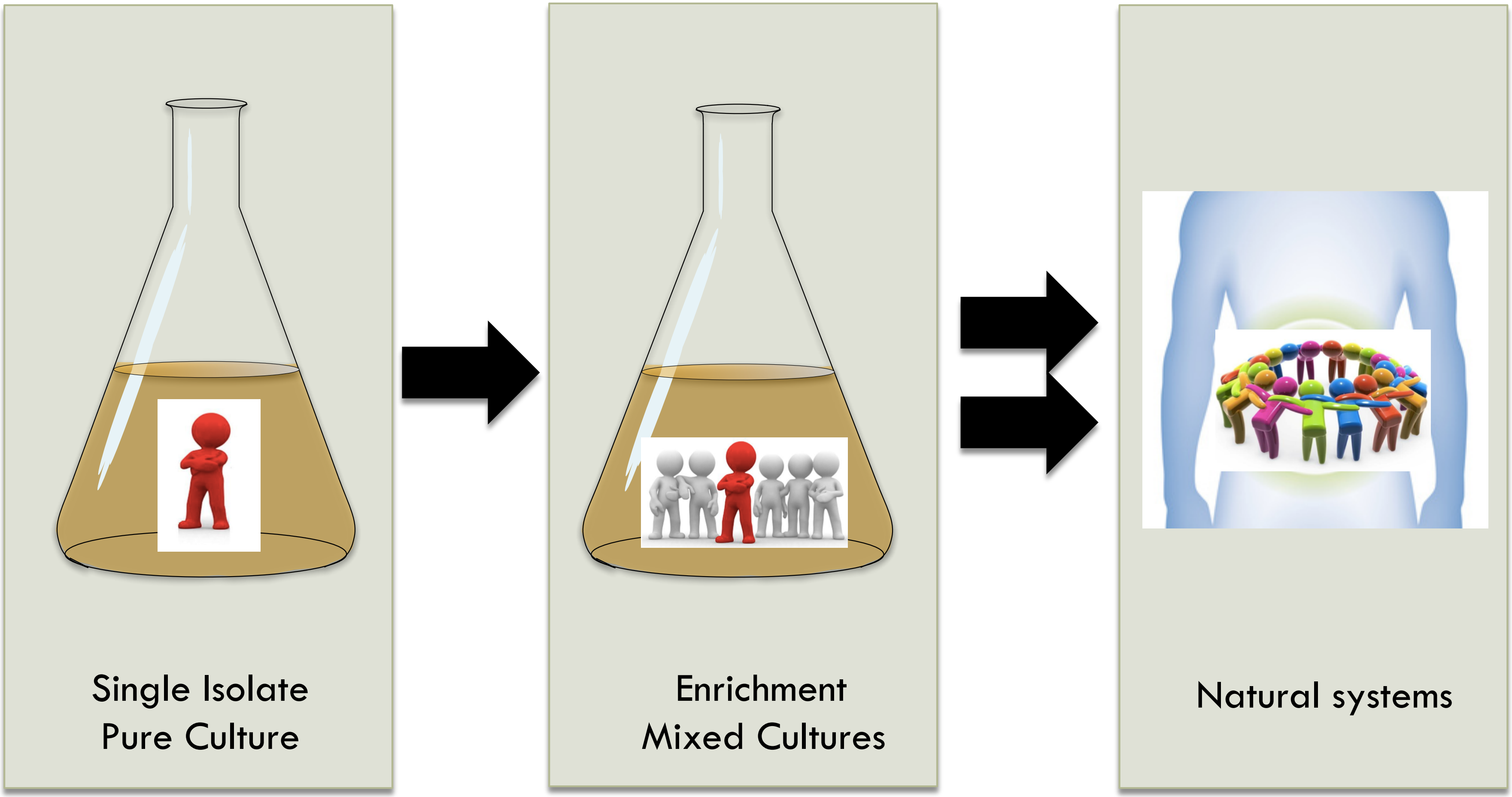
23andMe can help you manage risk and make informed decisions...

Learn valuable health & ancestry information.

$99

Personal genome can be mapped in a **few days** and **hundreds to few thousand dollars**

The experimental continuum

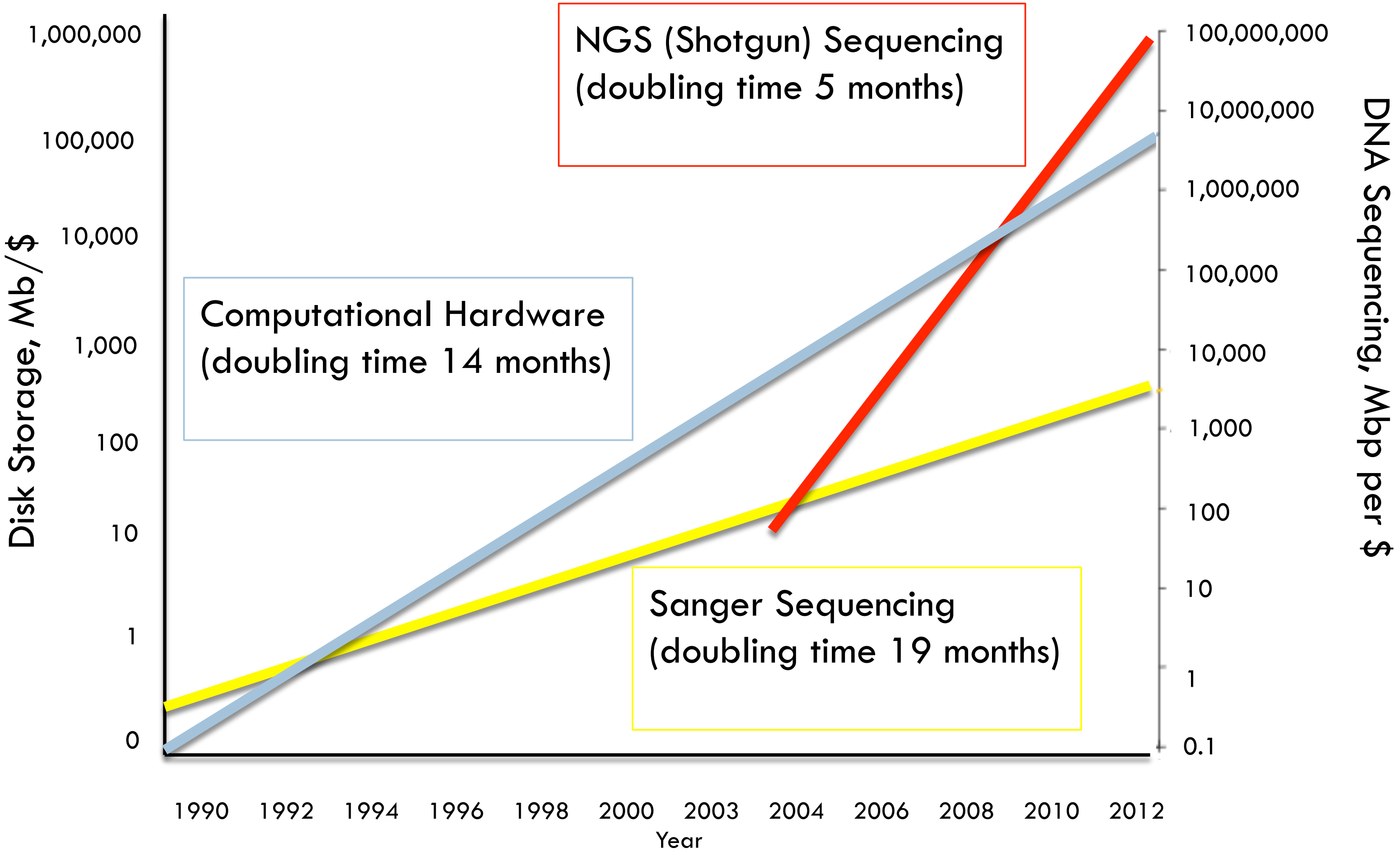


Single Isolate
Pure Culture

Enrichment
Mixed Cultures

Natural systems

The era of big data in biology



Stein, Genome Biology, 2010

Postdoc experience with data

2003-2008 Cumulative sequencing in PhD = 2000 bp

2008-2009 Postdoc Year 1 = 50 Gbp

2009-2010 Postdoc Year 2 = 450 Gbp

2014 = 50 Tbp

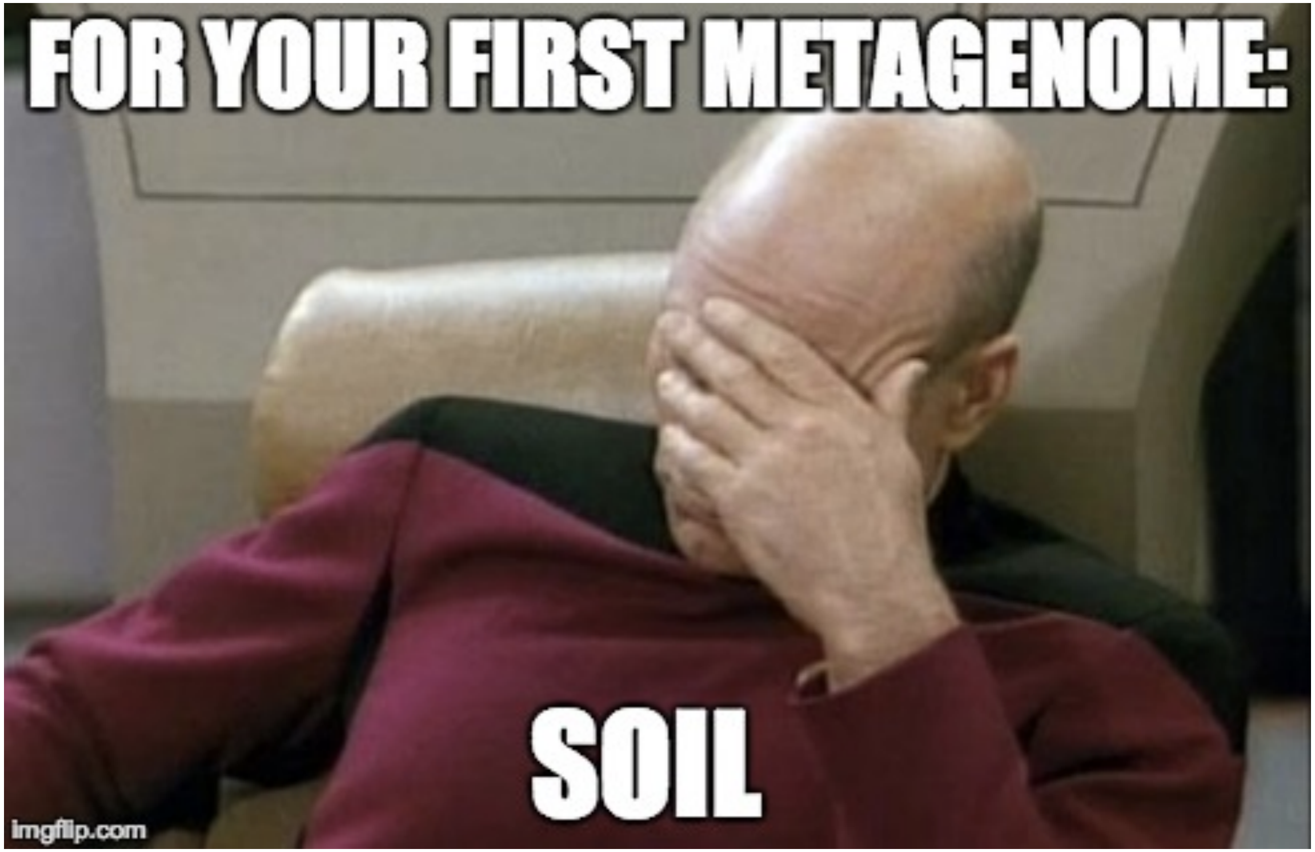
2015 = 500 Tbp budgeted

1 lane Illumina
50 million reads

8 lanes of Illumina
500 million reads

16+ lanes of Illumina
1800 million reads

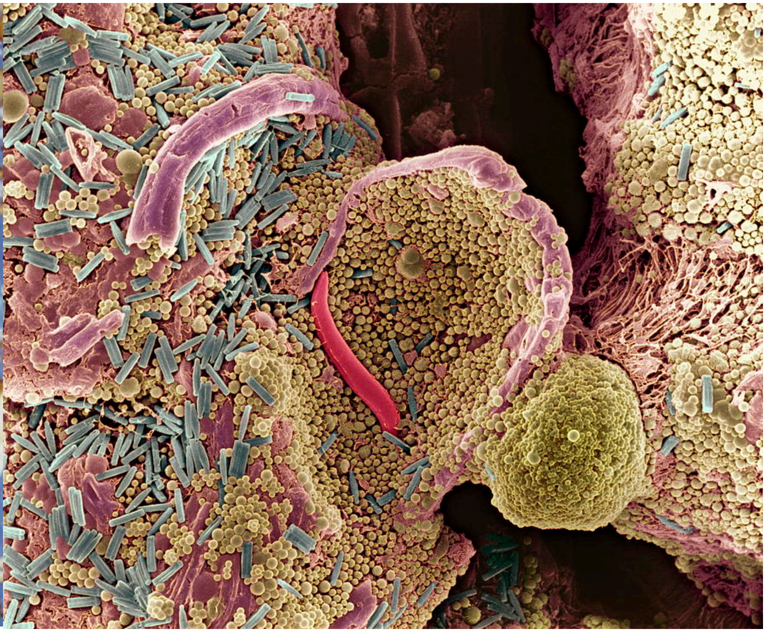
TOTAL 455 Gb of data!
(100,000 genomes)


FOR YOUR FIRST METAGENOME:
SOIL
imgflip.com

THE DIRT ON SOIL

MAGNIFICENT BIODIVERSITY

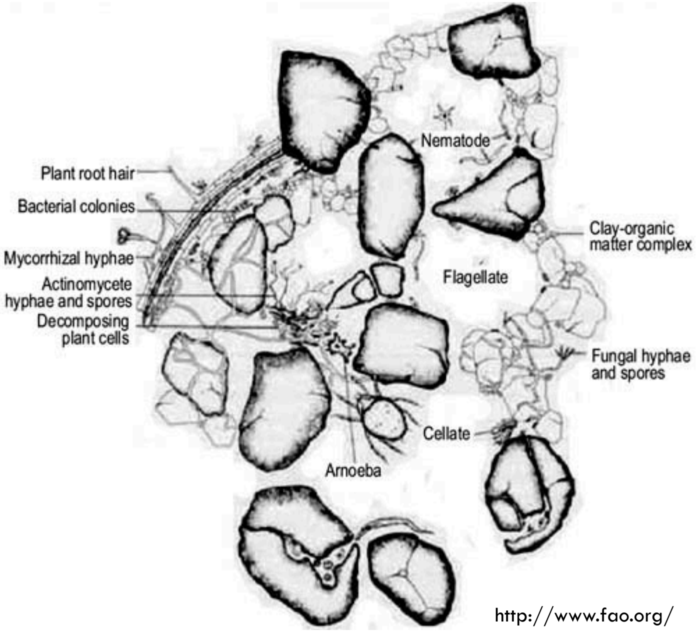
Biodiversity in the dark, Wall et al., Nature Geoscience, 2010



Jeremy Burgress

THE DIRT ON SOIL

SPATIAL HETEROGENEITY



http://www.fao.org/

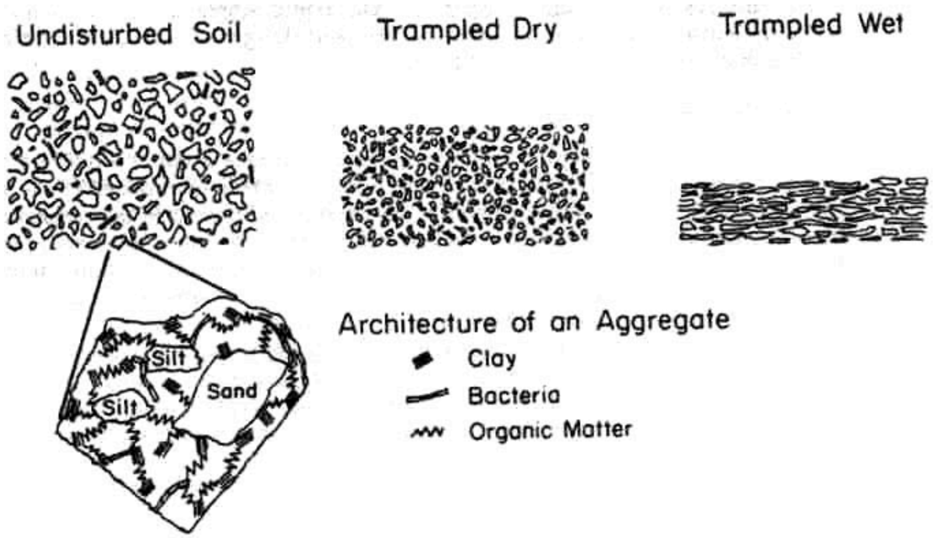
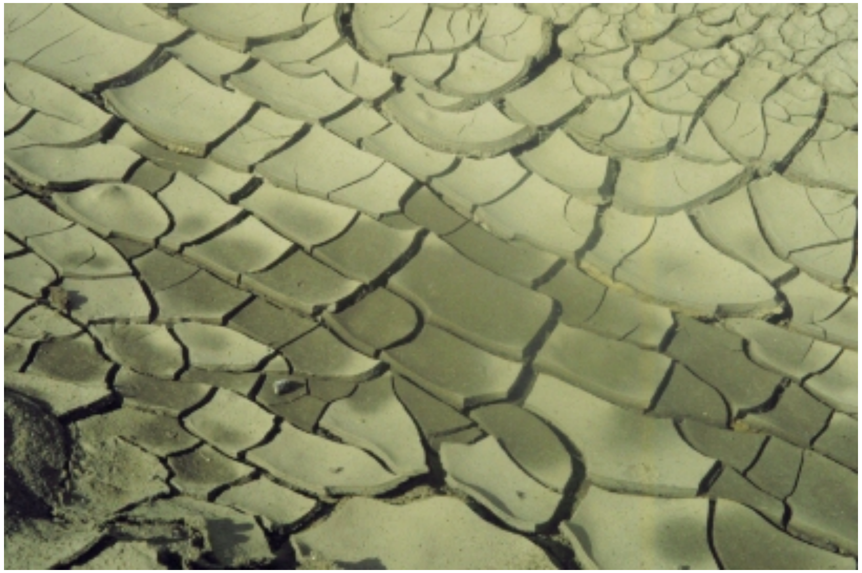


Figure 6.2. **Conceptual architecture of a soil aggregate and the changes in soil aggregate structure caused by trampling under wet and dry conditions.**

www.cnr.uidaho.edu

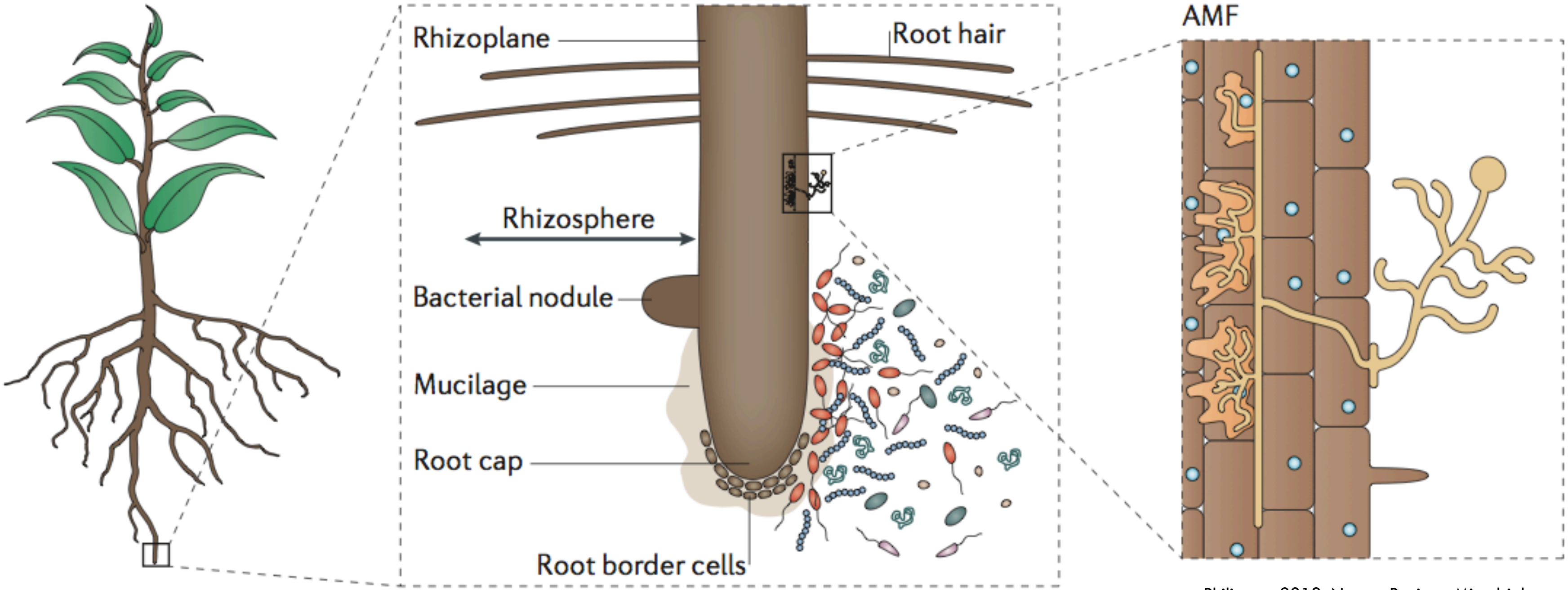
THE DIRT ON SOIL

DYNAMIC



THE DIRT ON SOIL

INTERACTIONS: BIOTIC, ABIOTIC, ABOVE, BELOW, SCALES



Philippot, 2013, Nature Reviews Microbiology

I. Methods to tackle metagenomic datasets
   - Computational
   - Experimental

I. Bottlenecks for microbiologists

Tackling Soil Biodiversity

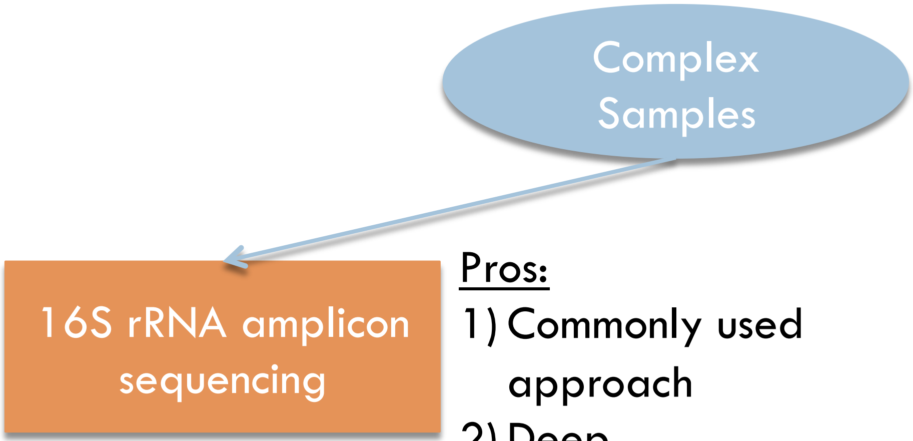
C. Titus Brown, James Tiedje, Qingpeng Zhang, Jason Pell (MSU)
Janet Jansson, Susannah Tringe (JGI)

Source: Chuck Haney

A Slight Digression: Decisions for the new microbial ecologist

Getting the most out of your data

ID, Abundance, Function



Complex Samples

16S rRNA amplicon sequencing

Pros:

1) Commonly used approach
2) Deep characterization

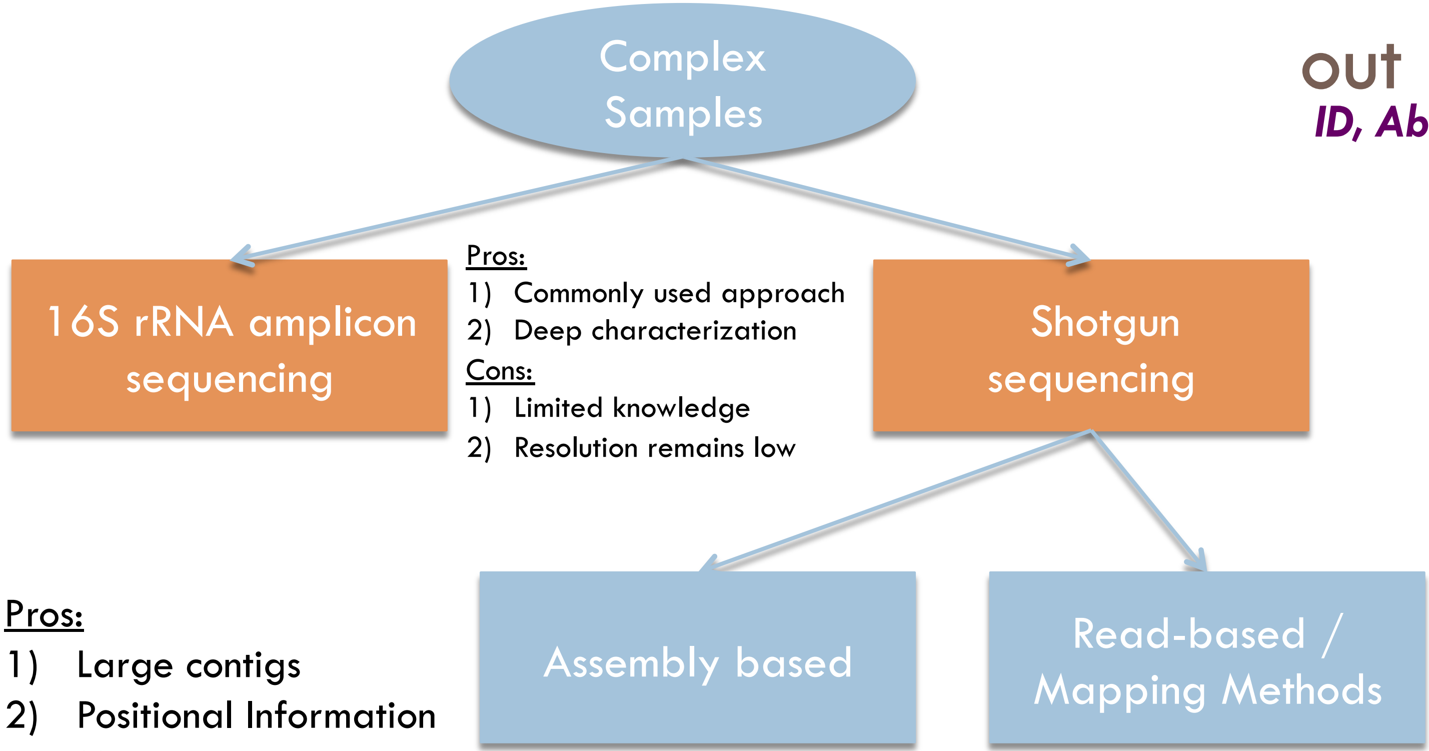
Cons:

1) Limited knowledge
2) Resolution remains low

Patrick Chain

Getting the most out of your data

ID, Abundance, Function



Complex Samples

16S rRNA amplicon sequencing

Shotgun sequencing

Assembly based

Read-based / Mapping Methods

<u>Pros:</u>

1) Commonly used approach
2) Deep characterization

<u>Cons:</u>

1) Limited knowledge
2) Resolution remains low

<u>Pros:</u>

1) Large contigs
2) Positional Information
3) Most direct method to identify novel orgs/genes

<u>Cons:</u>

1) Computational resource intensive
2) Assembling difficulties
   - Sequencing error
   - genomic redundancy
     - chimeras

<u>Pros:</u>

1) Massive data
2) Identity and abundance answered simultaneously
3) Look at all data**

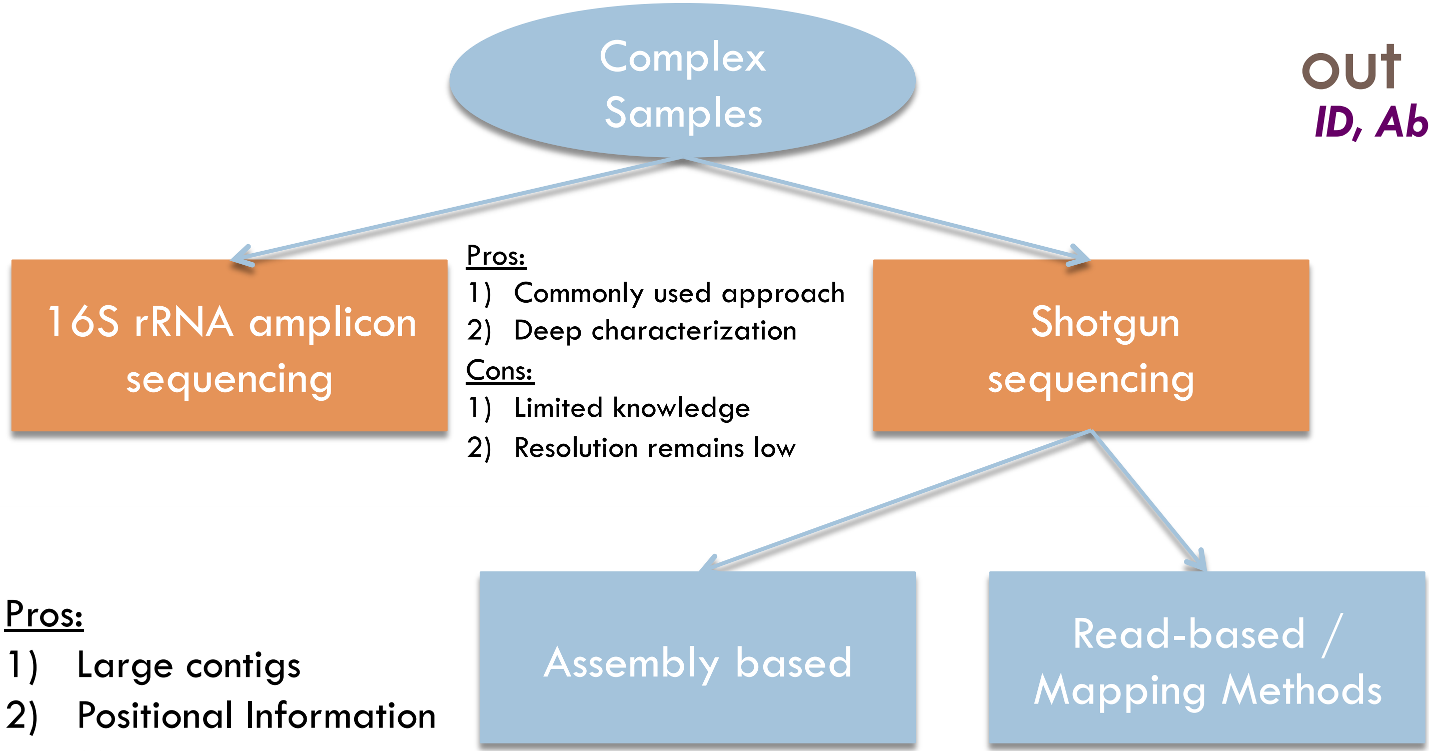
<u>Cons:</u>

1) Massive data (short + with errors)
2) Lack of specificity due to FPs from genomic redundancy
3) Difficult to detect novel genomes – must infer

Patrick Chain

Getting the most out of your data

ID, Abundance, Function



Pros:

1) Commonly used approach
2) Deep characterization

Cons:

1) Limited knowledge
2) Resolution remains low

Pros:

1) Massive data
2) Identity and abundance answered simultaneously
3) Look at all data**

Pros:

1) Large contigs
2) Positional Information
3) Most direct method to identify novel orgs/genes

Cons:

1) Computational resource intensive
2) Assembling difficulties
   - Sequencing error
   - genomic redundancy

     - chimeras

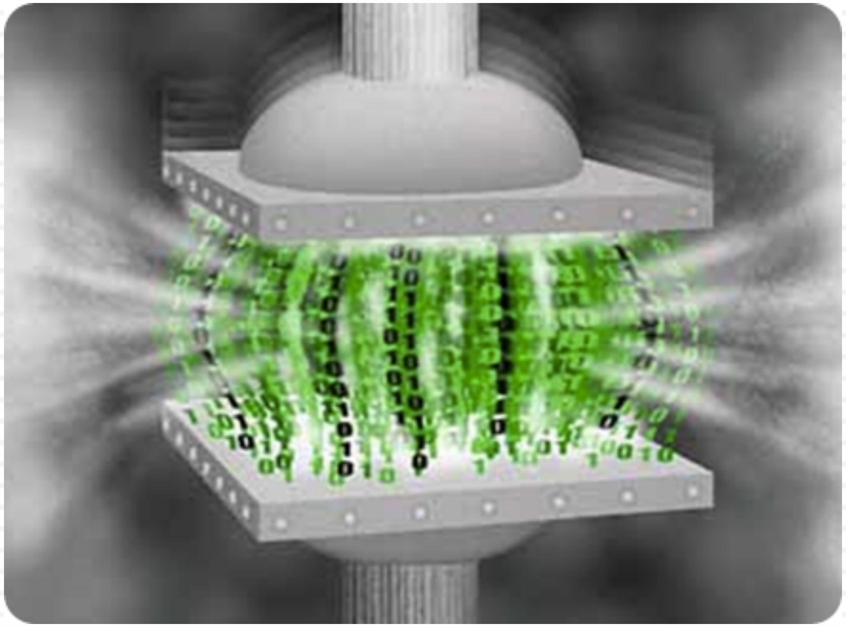
Cons:

1) Massive data (short + with errors)
2) Lack of specificity due to FPs from genomic redundancy
3) Difficult to detect novel genomes – must infer

The answer is always "it depends".

Patrick Chain

Example #1: Data compression



http://siliconangle.com/files/2010/09/image_thumb69.png

de novo assembly

Raw sequencing data (“reads”)

Computational algorithms

The New York Times.

ARMISTICE SIGNED, END OF THE WAR!
BERLIN SEIZED BY REVOLUTIONISTS;
NEW CHANCELLOR BEGS FOR ORDER;
OUSTED KAISER FLEES TO HOLLAND

Informative genes / genomes

- ❖Compresses dataset size significantly
- ❖Improved data quality (longer sequences, gene order)
- ❖Reference not necessary (novelty)

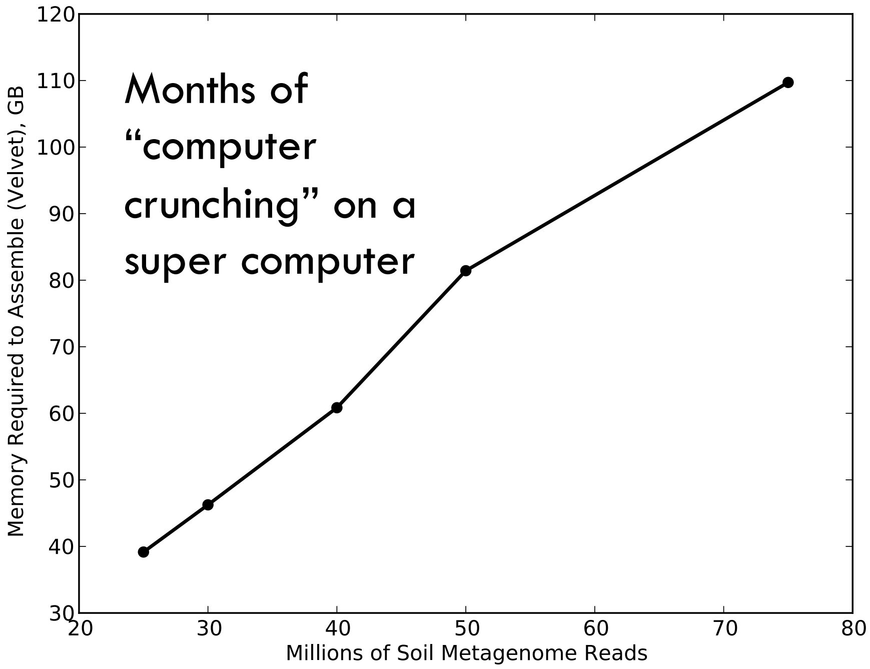
Metagenome assembly…a scaling problem.

Shotgun sequencing and de novo assembly

It was the Gest of times, it was the wor
, it was the worst of timZs, it was the
isdom, it was the age of foolisXness
, it was the worVt of times, it was the
mes, it was Ahe age of wisdom, it was th
It was the best of times, it Gas the wor
mes, it was the age of witdom, it was th
isdom, it was tle age of foolishness

It was the best of times, it was the worst of times, it was the age of
wisdom, it was the age of foolishness

Practical Challenges – Intensive computing



Months of "computer crunching" on a super computer

1 lane Illumina
50 million reads

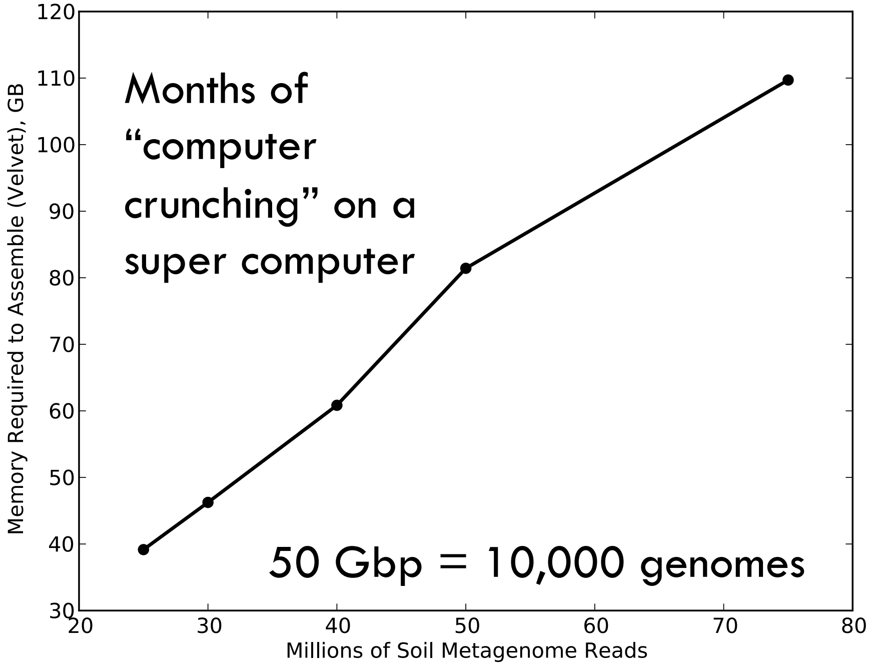
8 lanes of Illumina
500 million reads

16+ lanes of Illumina
1800 million reads

TOTAL 455 Gb of data!
(100,000 genomes)

Howe et al, 2014, PNAS

Practical Challenges – Intensive computing



Months of "computer crunching" on a super computer

50 Gbp = 10,000 genomes

Assembly of 300 Gbp (70,000 genomes worth) can be done with any assembly program in less than 14 GB RAM and less than 24 hours.

16+ lanes of Illumina
1800 million reads

8 lanes of Illumina
500 million reads

TOTAL 455 Gb of data!
(100,000 genomes)

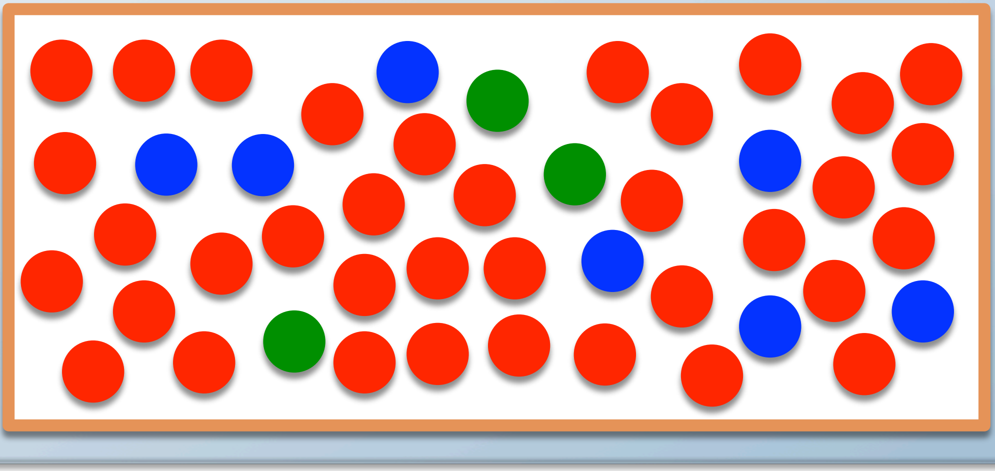
1 lane Illumina
50 million reads

Howe et al, 2014, PNAS

Natural community characteristics

- **Diverse**

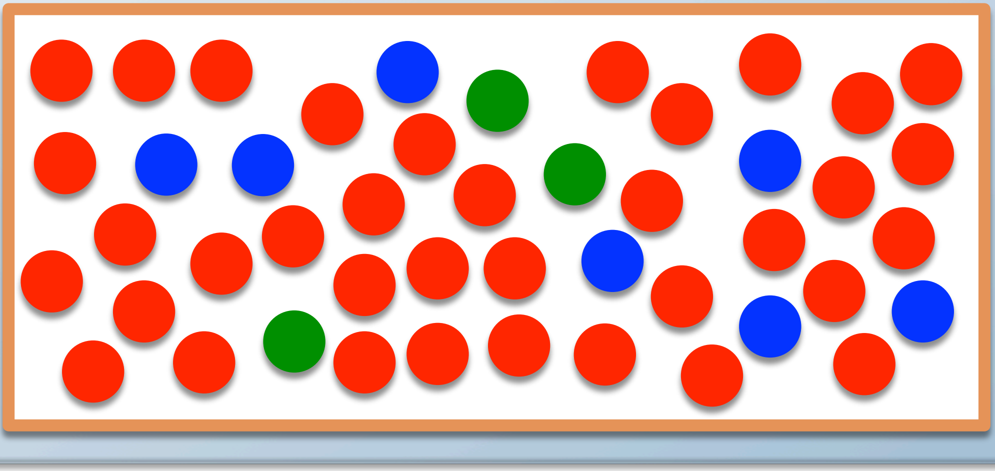
  → Many organisms (genomes)



Natural community characteristics

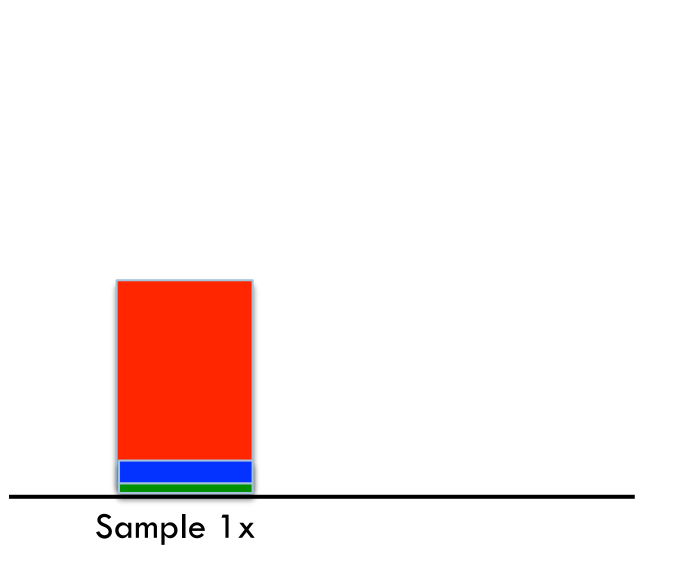
Diverse

- ➔ Many organisms (genomes)



Variable abundance

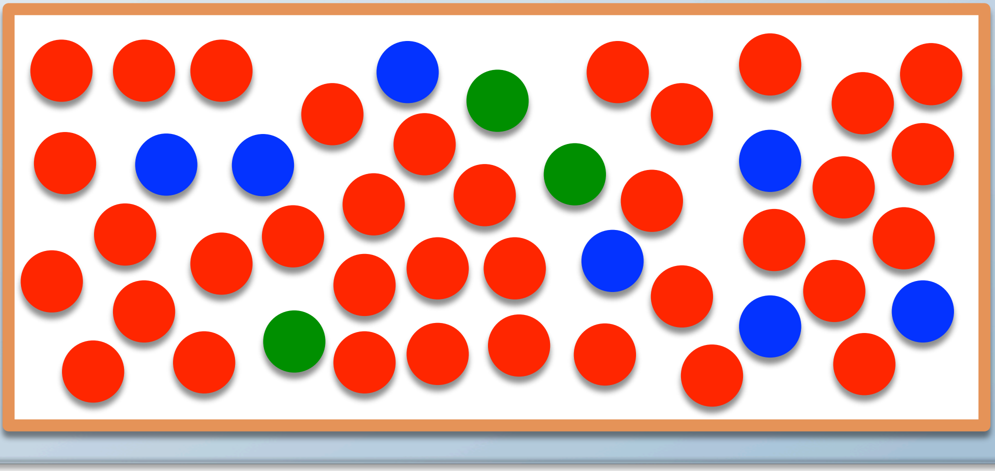
- ➔ Most abundant organisms, sampled more often
- ➔ Assembly requires a minimum amount of sampling
- ➔ More sequencing, more errors



Natural community characteristics

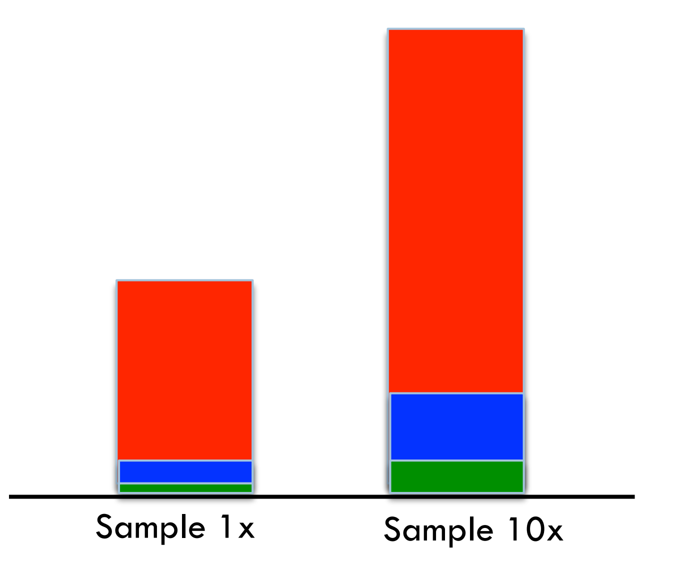
Diverse

- ➔ Many organisms (genomes)



Variable abundance

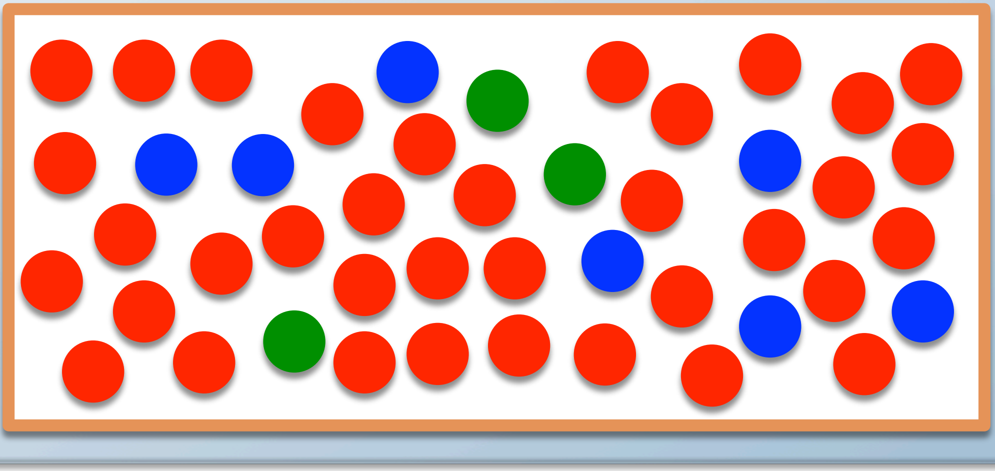
- ➔ Most abundant organisms, sampled more often
- ➔ Assembly requires a minimum amount of sampling
- ➔ More sequencing, more errors



Natural community characteristics

◆ Diverse

➔ Many organisms (genomes)

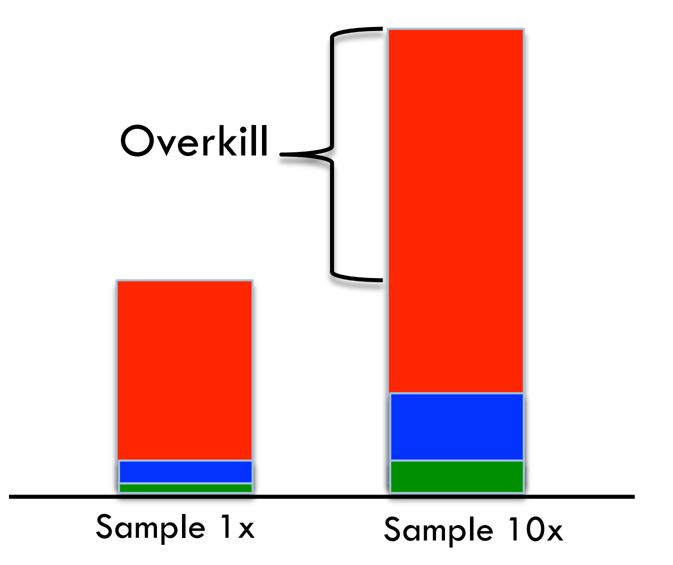


◆ Variable abundance

➔ Most abundant organisms, sampled more often

➔ Assembly requires a minimum amount of sampling

➔ More sequencing, more errors



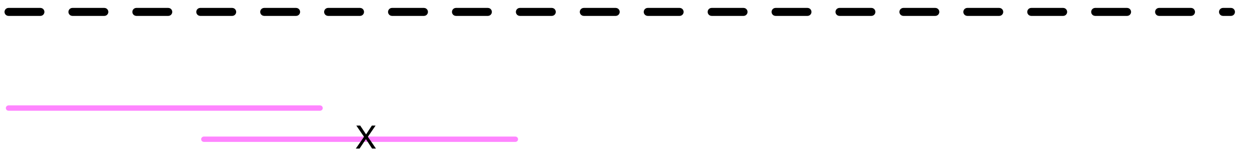
Digital normalization

True sequence (unknown)

Reads
(randomly sequenced)

Brown et al., 2012, arXiv
Howe et al., 2014, PNAS
Zhang et al., 2014, PLOS One

Digital normalization

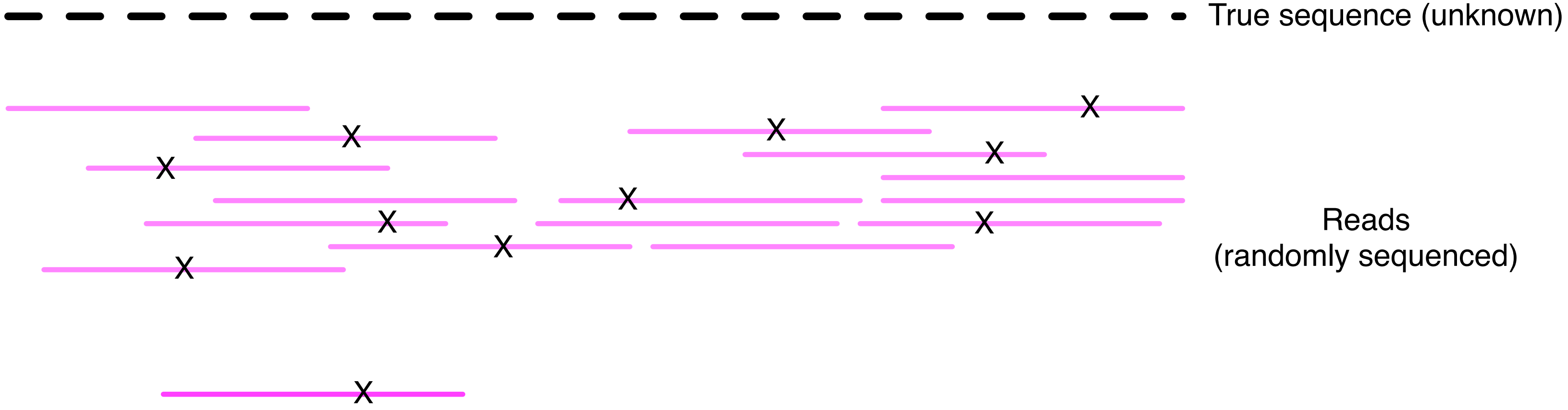


True sequence (unknown)

Reads
(randomly sequenced)

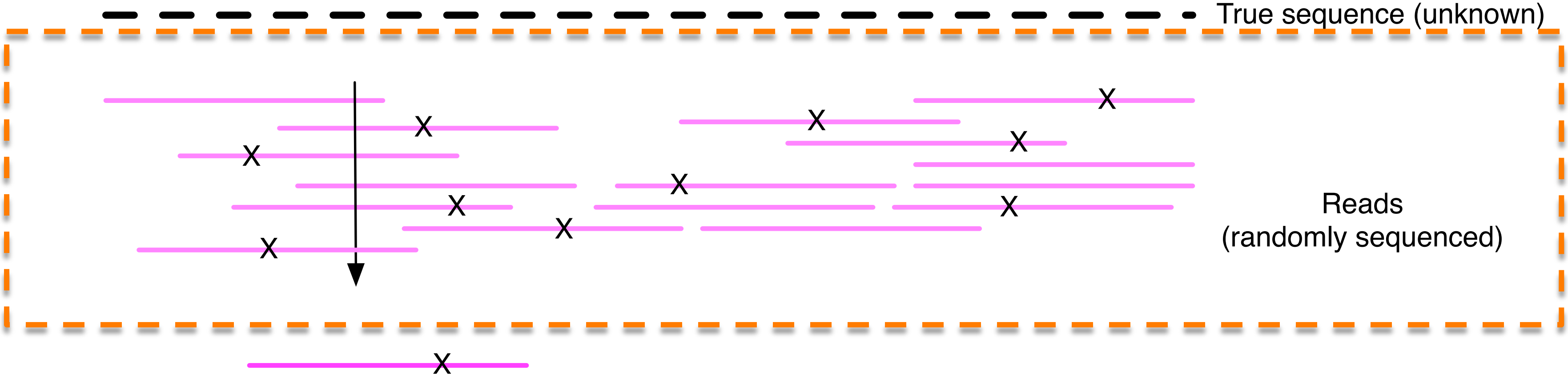
Brown et al., 2012, arXiv
Howe et al., 2014, PNAS
Zhang et al., 2014, PLOS One

Digital normalization



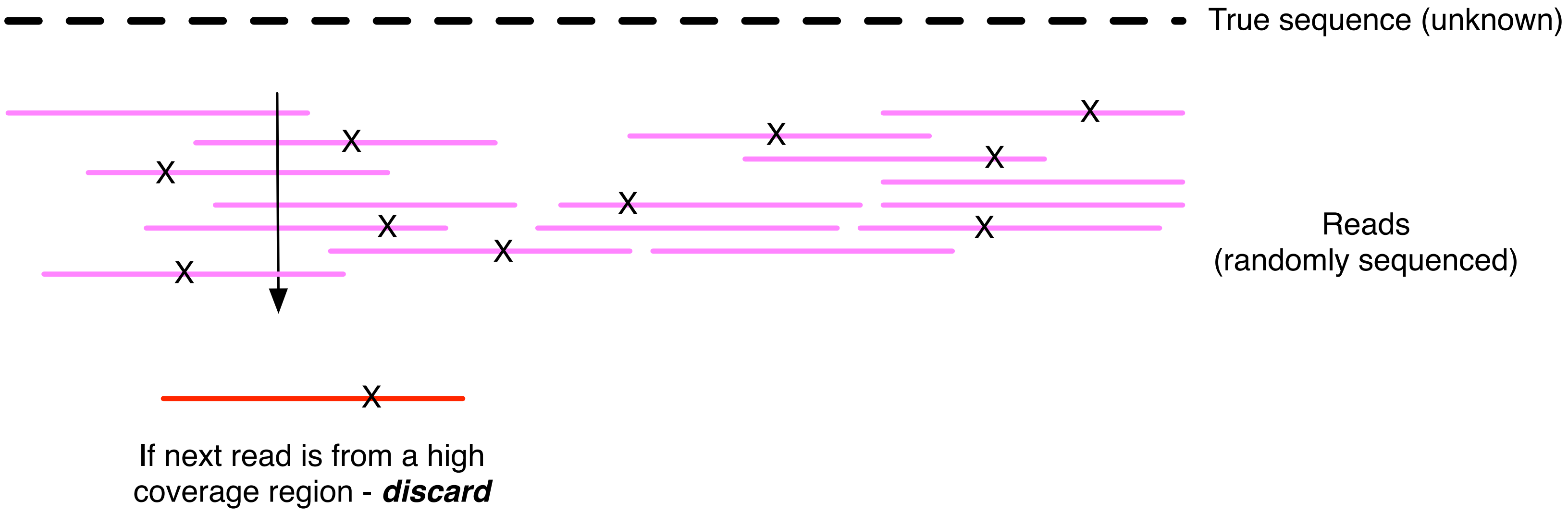
Brown et al., 2012, arXiv
Howe et al., 2014, PNAS
Zhang et al., 2014, PLOS One

Digital normalization



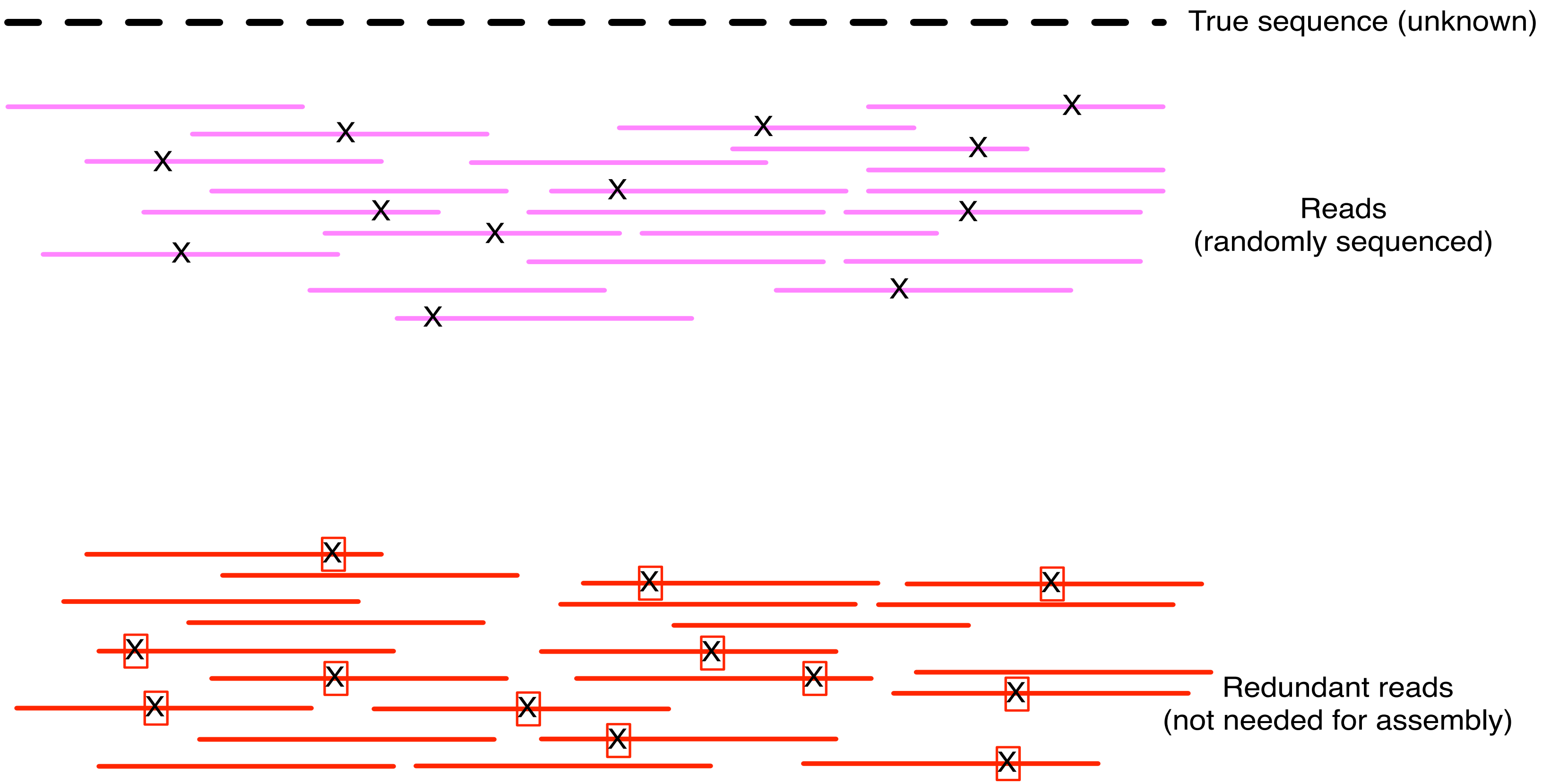
Brown et al., 2012, arXiv
Howe et al., 2014, PNAS
Zhang et al., 2014, PLOS One

Digital normalization



Brown et al., 2012, arXiv
Howe et al., 2014, PNAS
Zhang et al., 2014, PLOS One

Digital normalization



- ❖ Scales datasets for assembly up to 95% - same assembly outputs.
- ❖ Genomes, mRNA-seq, metagenomes (soils, gut, water)

Brown et al., 2012, arXiv
Howe et al., 2014, PNAS
Zhang et al., 2014, PLOS One

Tackling Soil Biodiversity

C. Titus Brown, James Tiedje, Qingpeng Zhang, Jason Pell (MSU)
Janet Jansson, Susannah Tringe (JGI)

Source: Chuck Haney

The reality?

More like…

Howe et. al, 2014, PNAS

Source: Chuck Haney

The Future

- More data, more samples, better references
- Expense will be in sampling – not sequencing or even data analysis
- All biologists will need to know how to use a pipette and write computer programs
- Large-scale, collaborative projects rather than single PI efforts