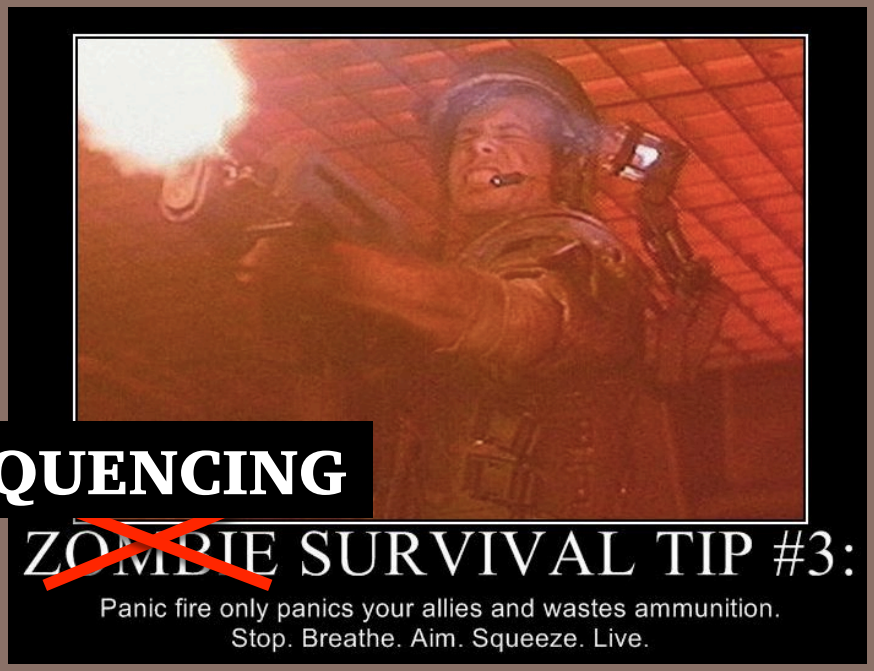
WHAT TO DO IN THE EVENT OF A DATA DELUGE

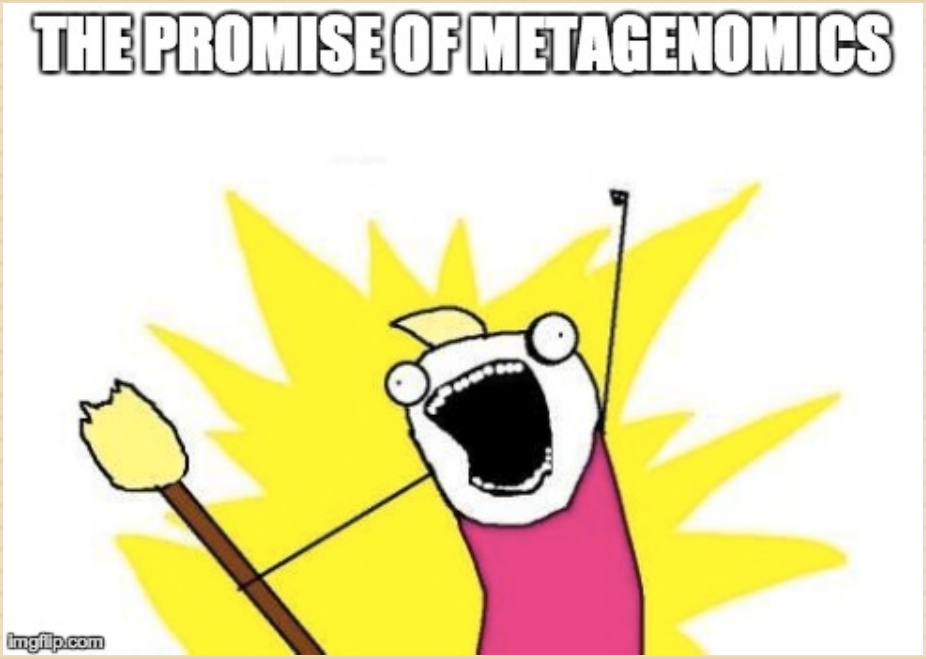
Adina Howe
germslab.org (Genomics and Environmental Research in Microbial Systems)

Iowa State University, Ag & Biosystems Engr (January)

Slides available at www.slideshare.com/adinachuanghowe



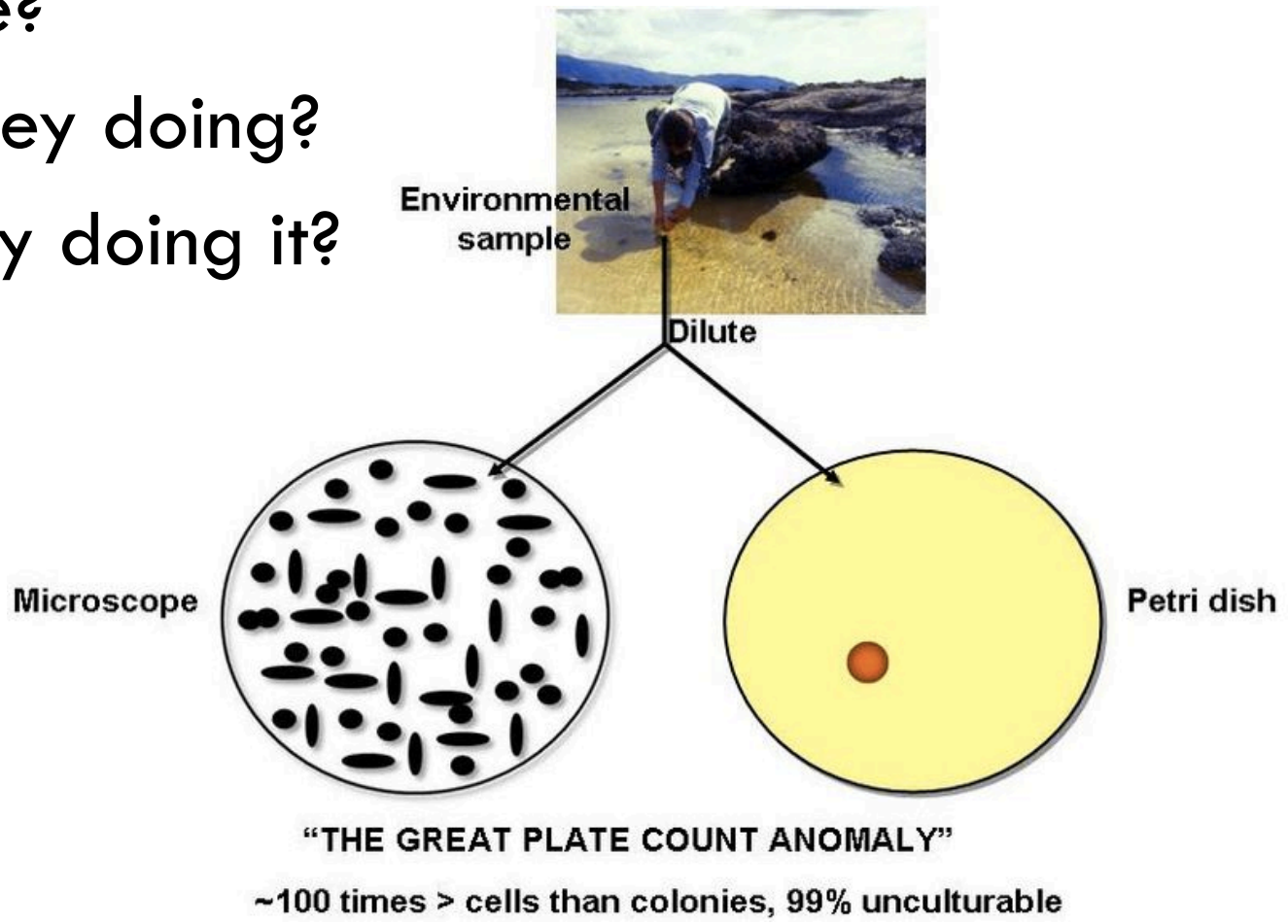
EEMiS Workshop, Öland, Sweden, 12/1/2014



HOW DID WE GET HERE

Understanding community dynamics

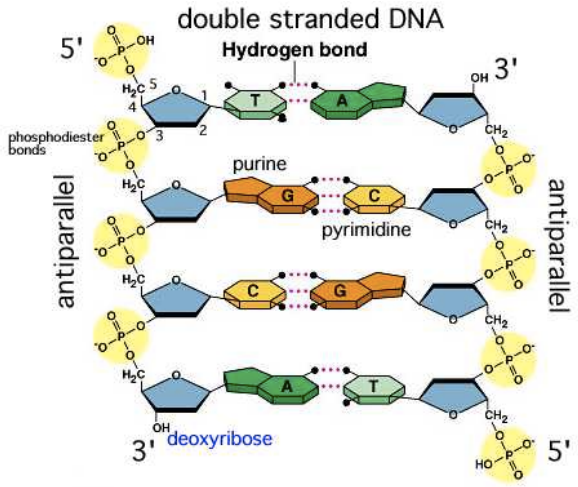
- Who is there?
- What are they doing?
- How are they doing it?



Kim Lewis, 2010

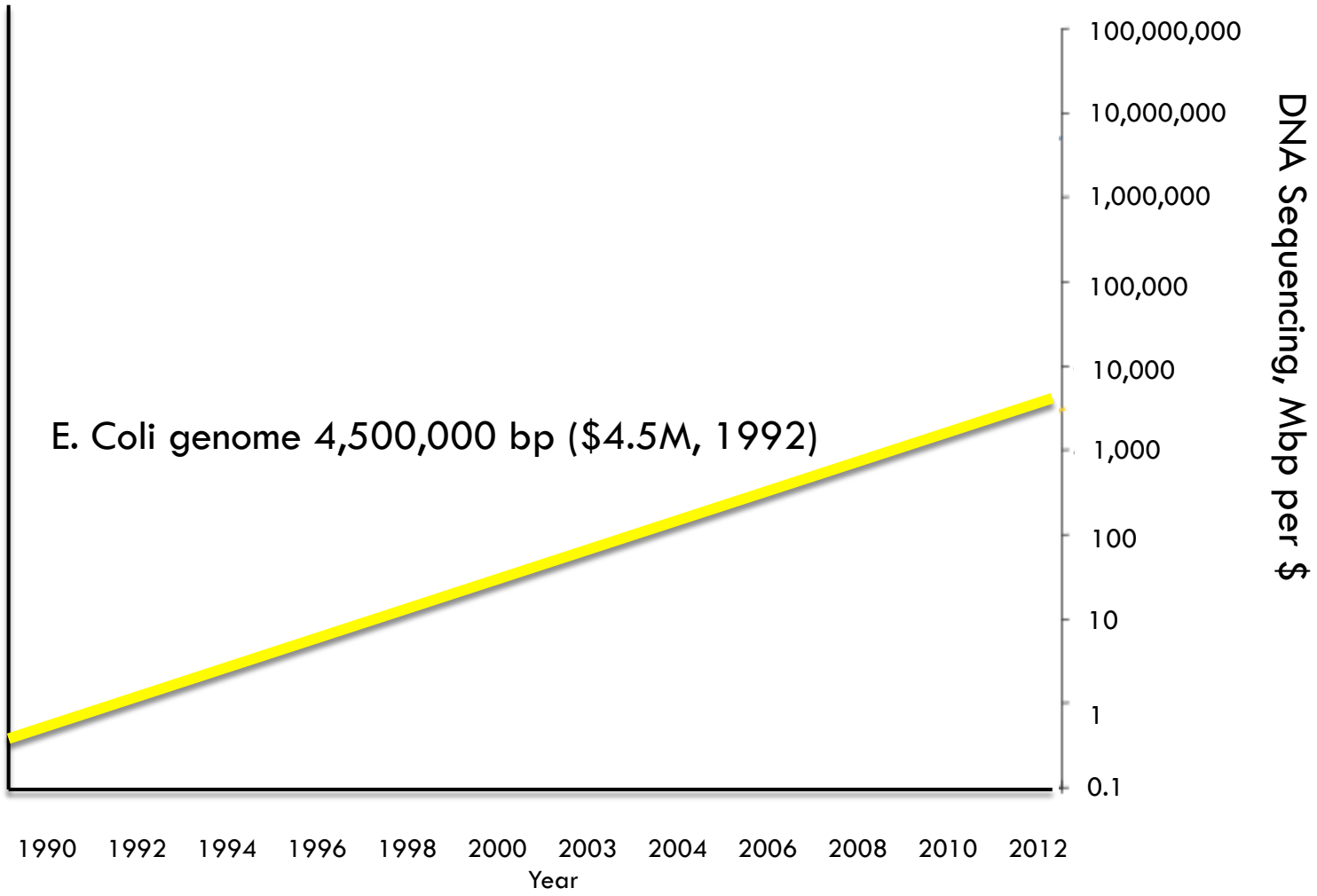
Gene / Genome Sequencing

- Collect samples
- Extract DNA
- Sequence DNA
- "Analyze" DNA to identify its content and origin



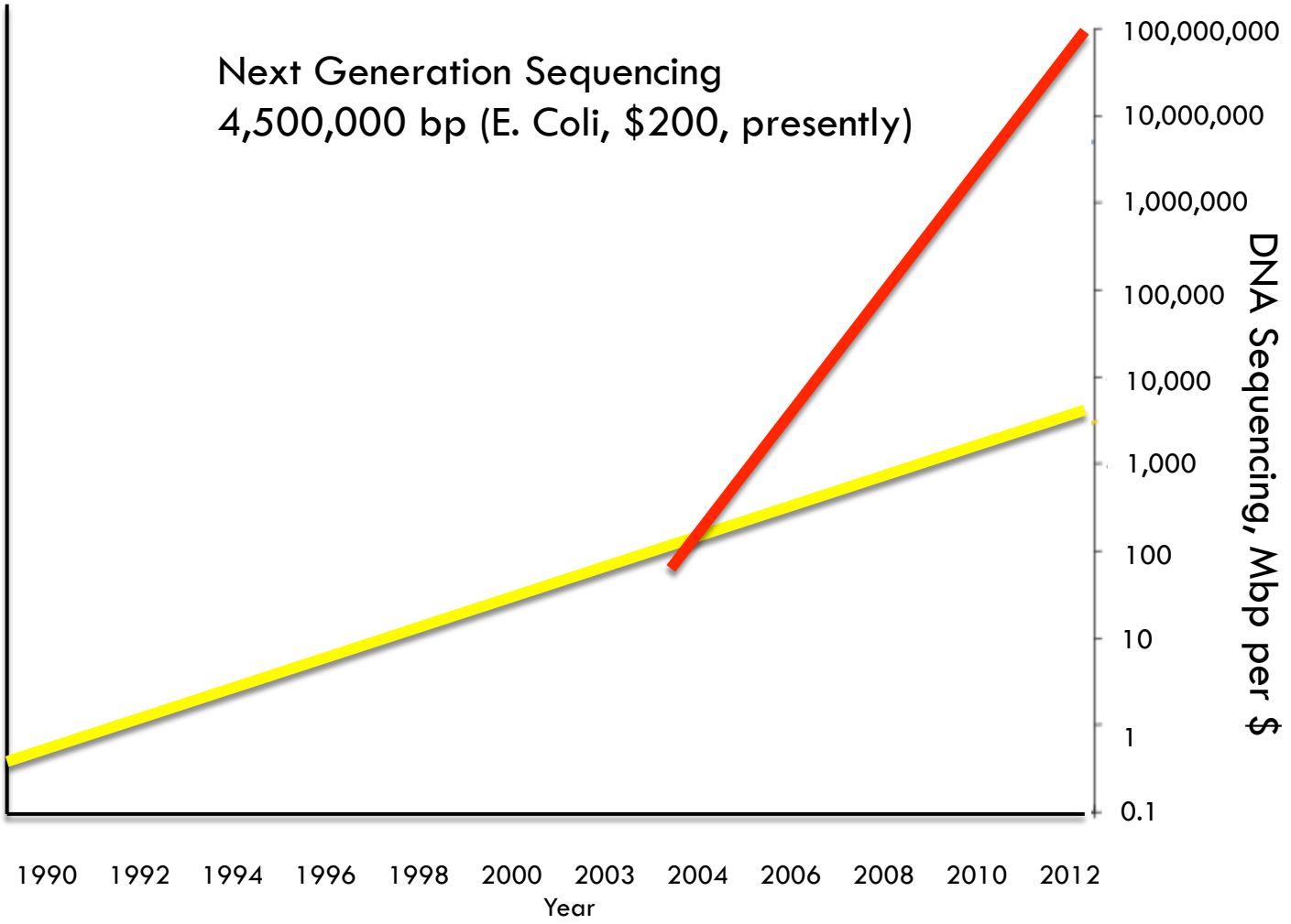
Taxonomy
(e.g., pathogenic E. Coli)
Function
(e.g., degrades cellulose)

Cost of Sequencing



Stein, Genome Biology, 2010

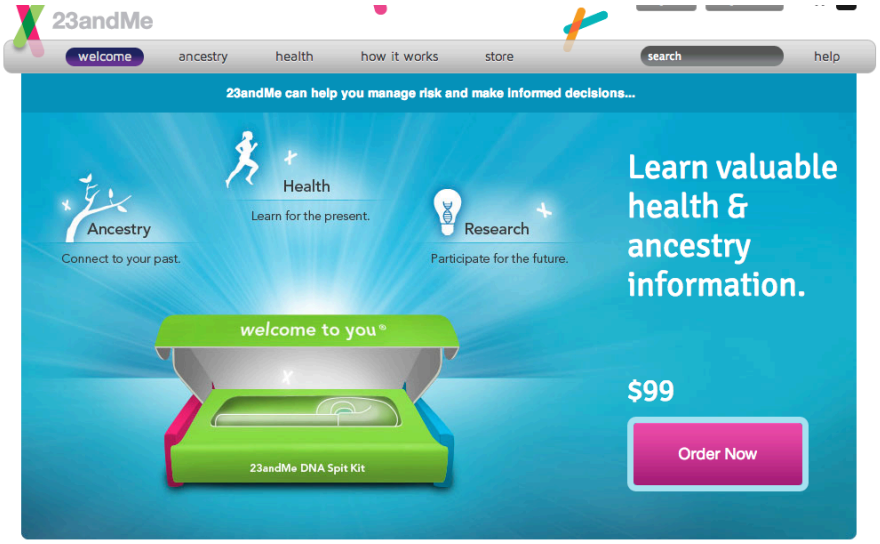
Rapidly decreasing costs with NGS Sequencing



Stein, Genome Biology, 2010

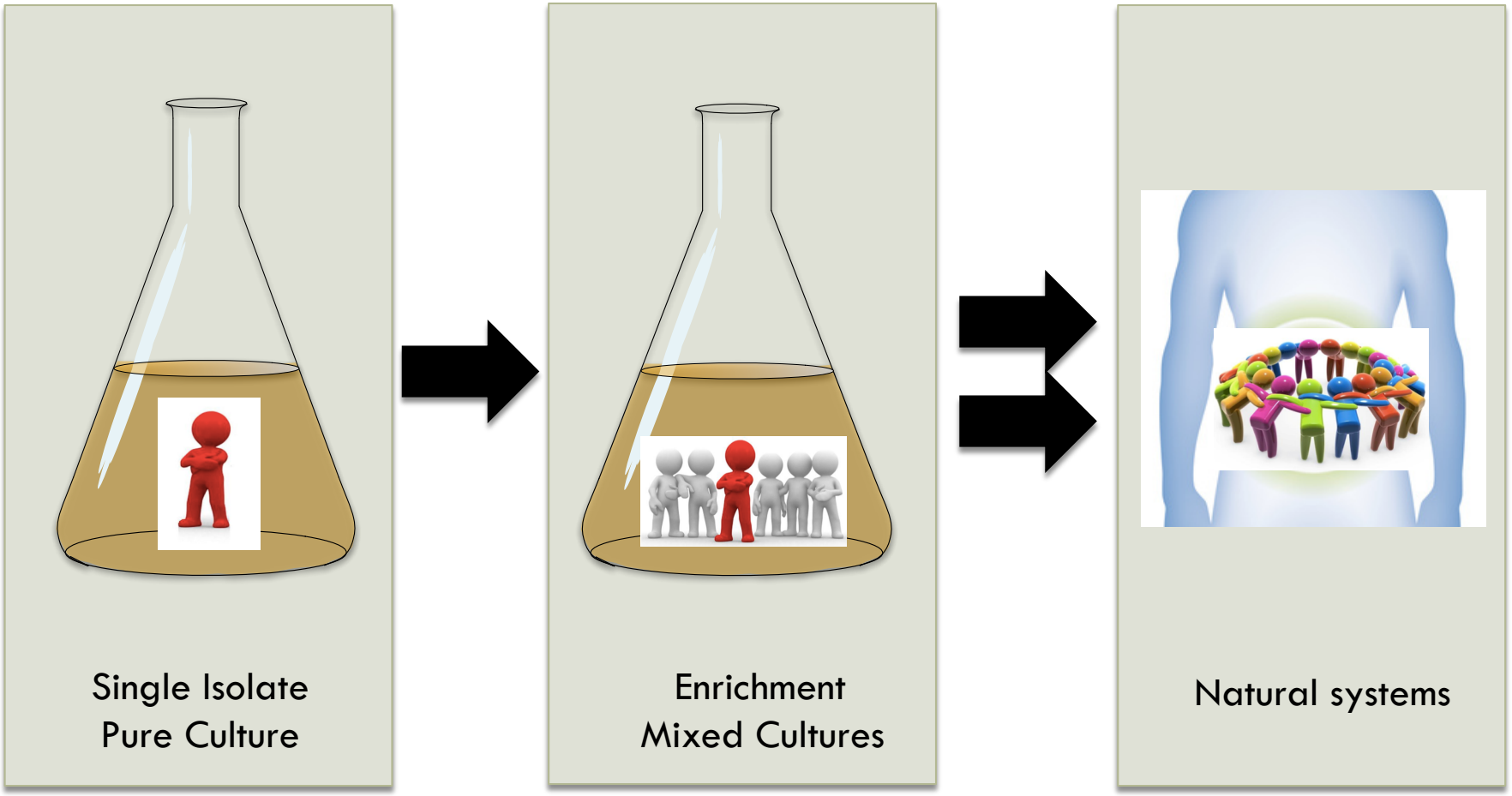
Effects of low cost sequencing…

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

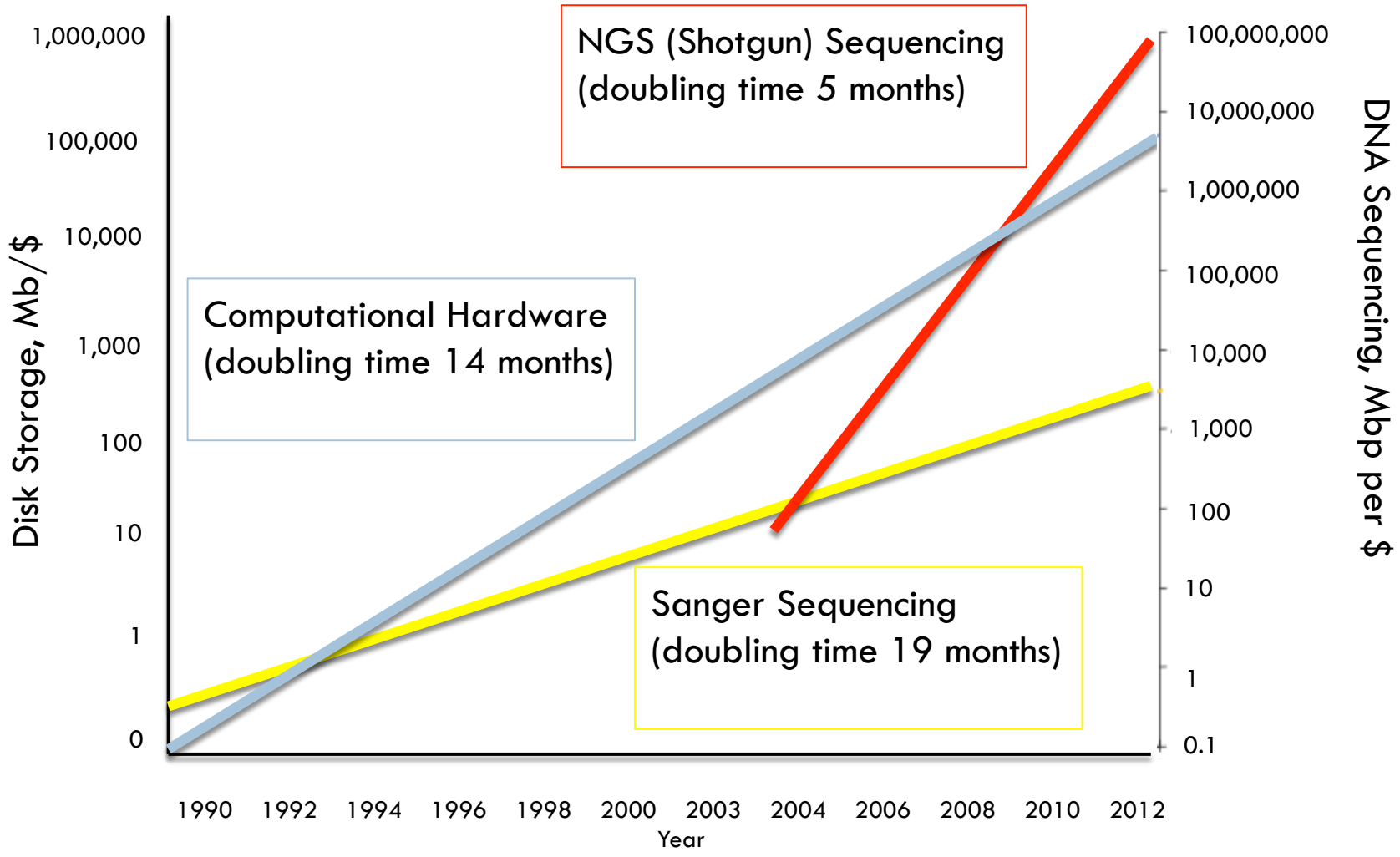


Personal genome can be mapped in a **few days** and **hundreds to few thousand dollars**

The experimental continuum



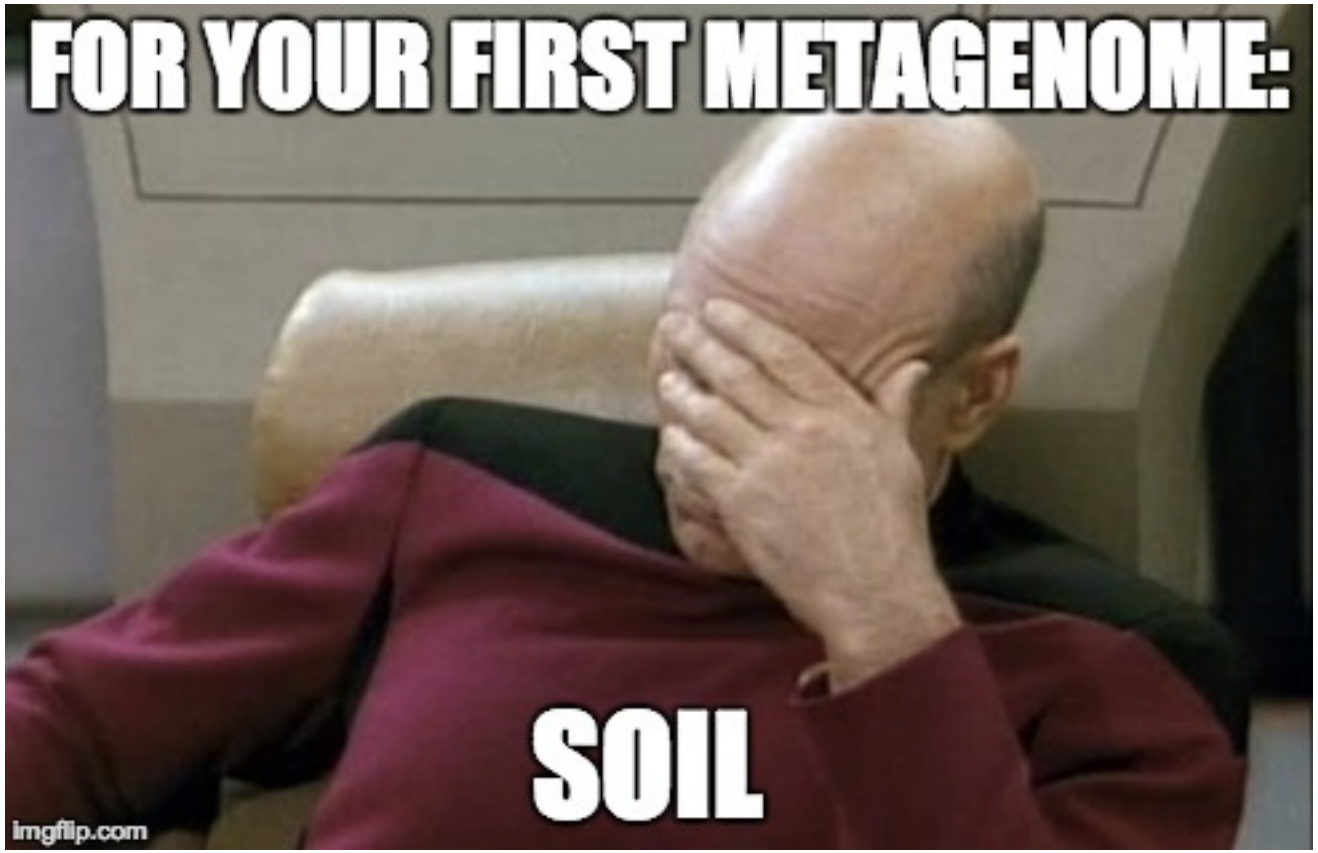
The era of big data in biology



Stein, Genome Biology, 2010

Postdoc experience with data

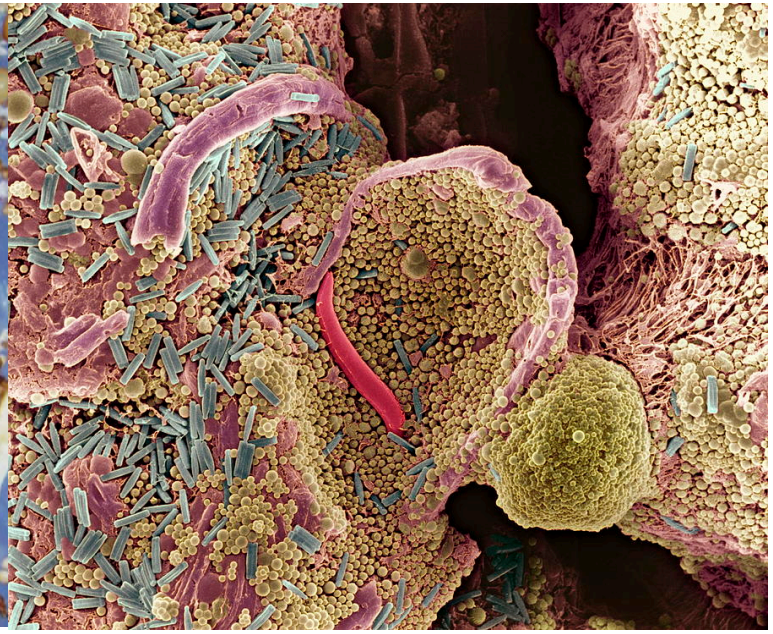
2003-2008 Cumulative sequencing in PhD = 2000 bp
2008-2009 Postdoc Year 1 = 50 Gbp
2009-2010 Postdoc Year 2 = 450 Gbp
2014 = 50 Tbp
2015 = 500 Tbp budgeted


FOR YOUR FIRST METAGENOME:
SOIL
imgflip.com

THE DIRT ON SOIL

MAGNIFICENT BIODIVERSITY

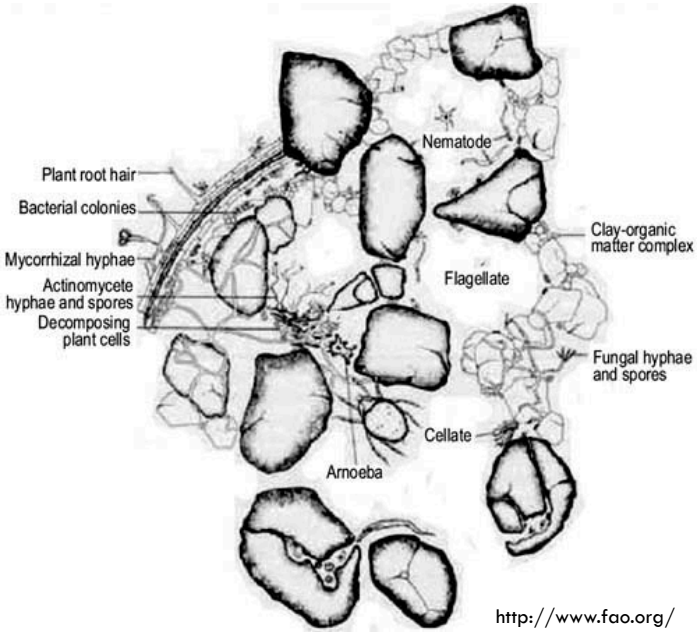
Biodiversity in the dark, Wall et al., Nature Geoscience, 2010



Jeremy Burgress

THE DIRT ON SOIL

SPATIAL HETEROGENEITY



http://www.fao.org/

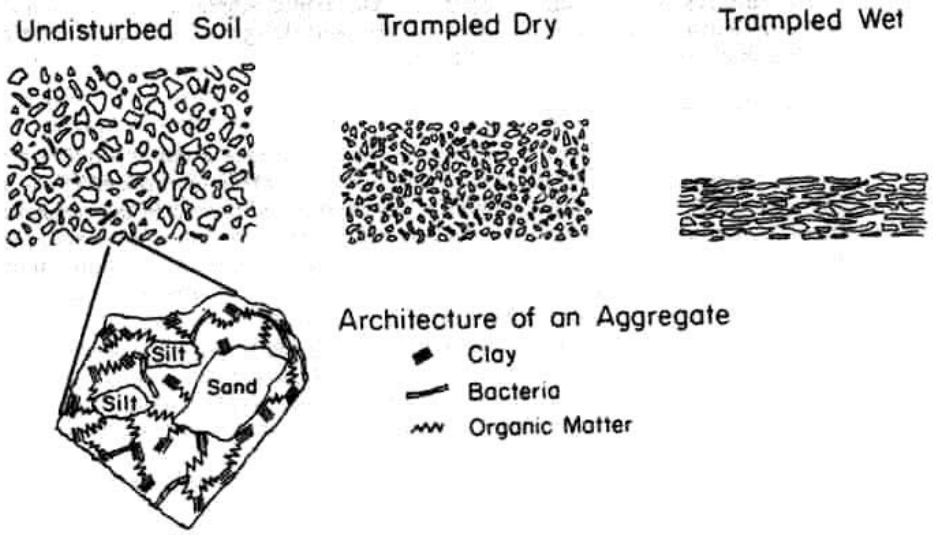
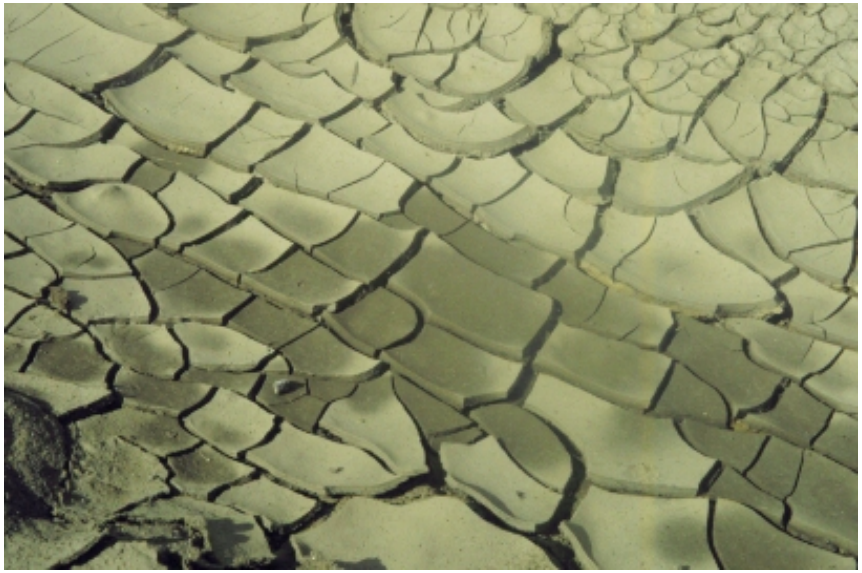


Figure 6.2. Conceptual architecture of a soil aggregate and the changes in soil aggregate structure caused by trampling under wet and dry conditions.

www.cnr.uidaho.edu

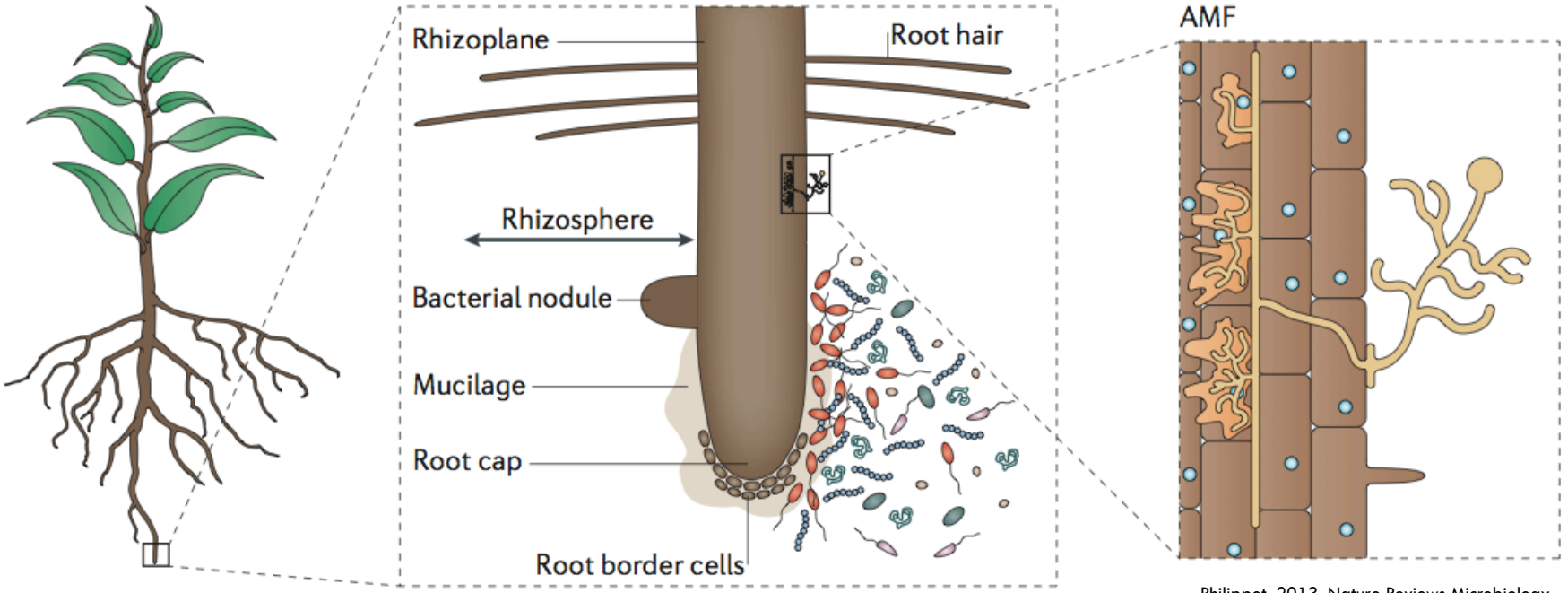
THE DIRT ON SOIL

DYNAMIC



THE DIRT ON SOIL

INTERACTIONS: BIOTIC, ABIOTIC, ABOVE, BELOW, SCALES



Philippot, 2013, Nature Reviews Microbiology

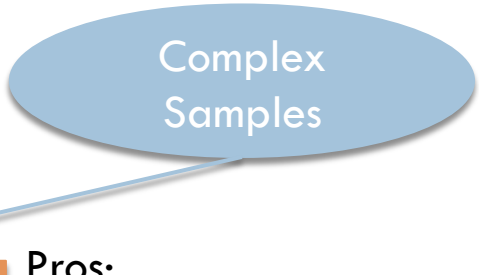
Tackling Soil Biodiversity

Source: Chuck Haney

A Slight Digression: Decisions for the new microbial ecologist

Getting the most out of your data

ID, Abundance, Function



16S rRNA amplicon sequencing

Pros:

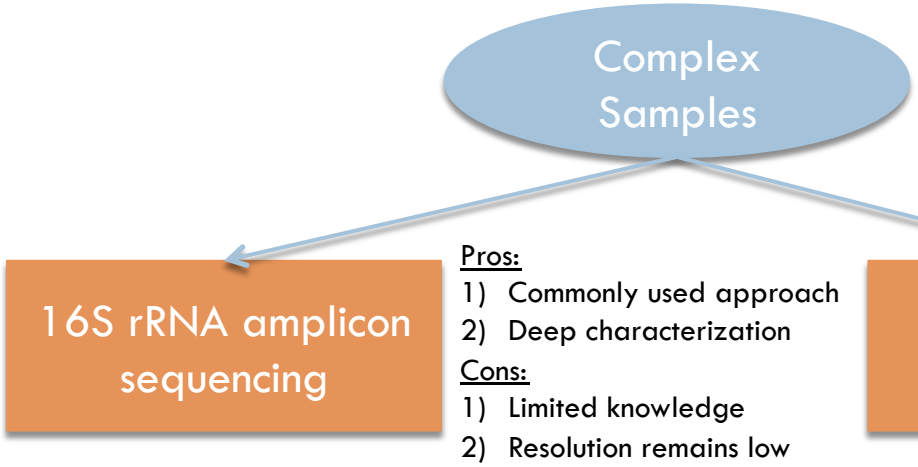
1) Commonly used approach
2) Deep characterization

Cons:

1) Limited knowledge
2) Resolution remains low

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) 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

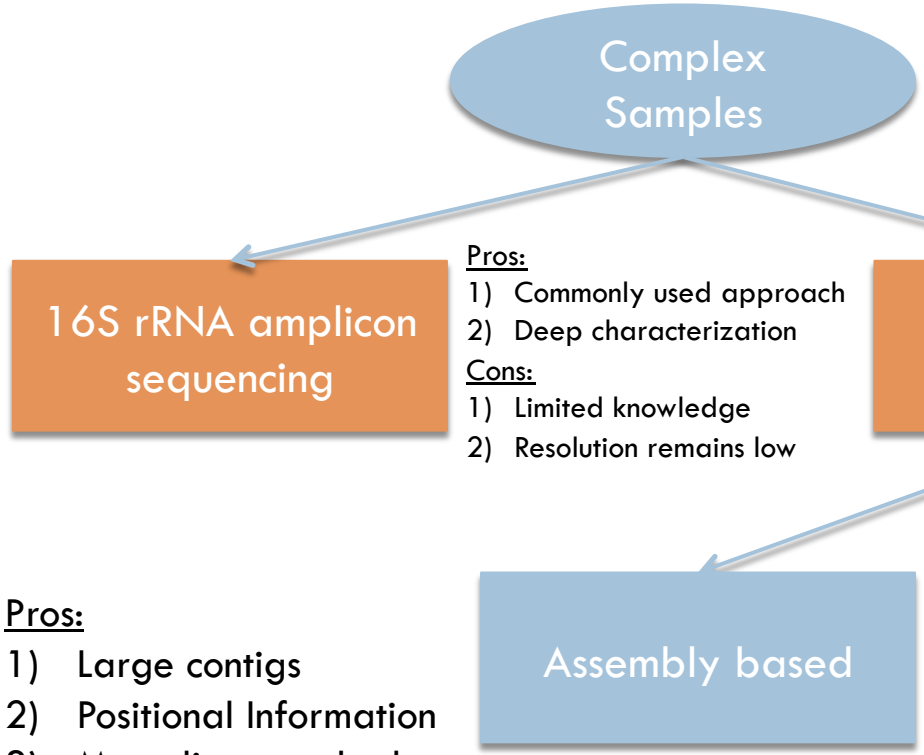
Pros:
1) Massive data
2) Identity and abundance answered simultaneously
3) Look at all data**

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

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) 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

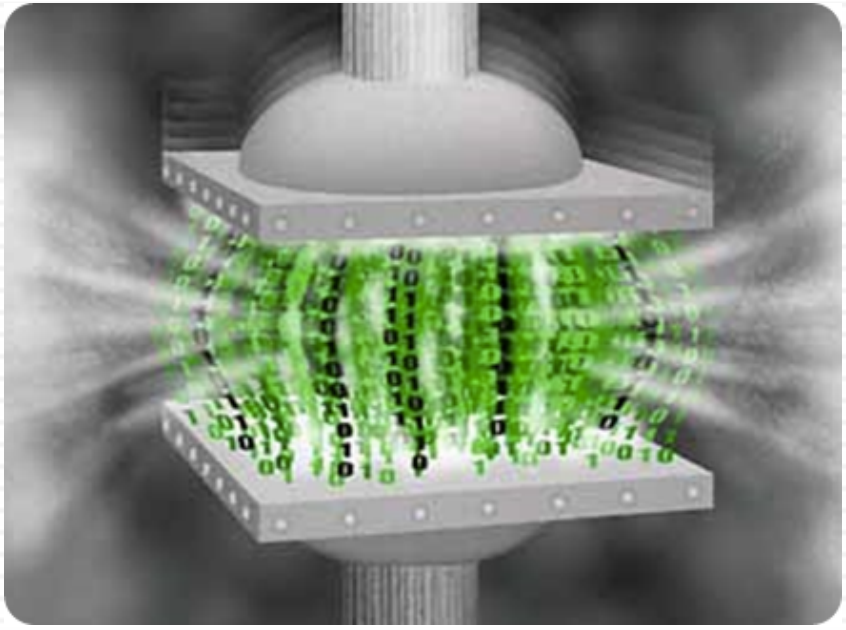
Pros:
1) Massive data
2) Identity and abundance answered simultaneously
3) Look at all data**

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

Informative genes / genomes

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

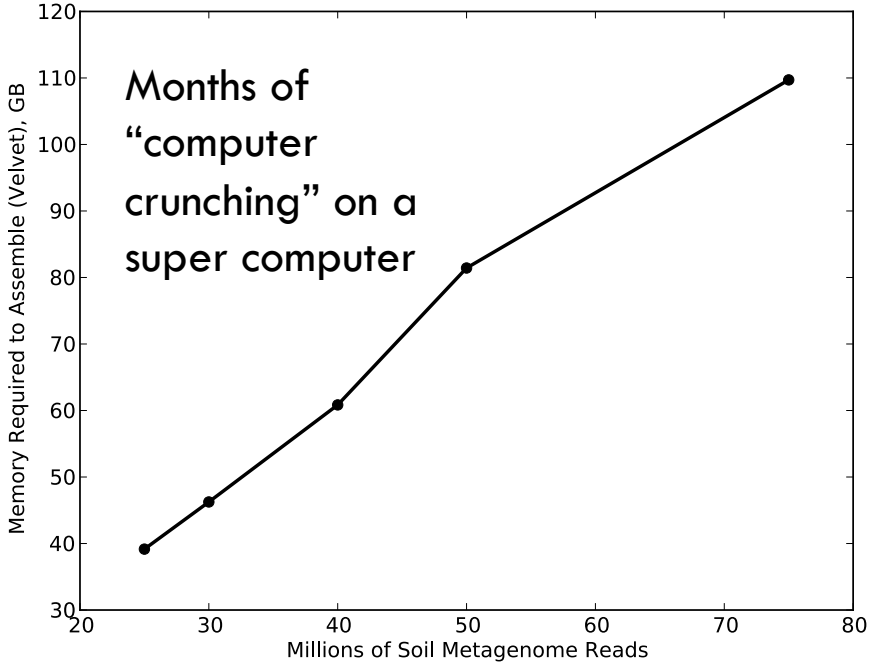
Metagenome assembly…a scaling problem.

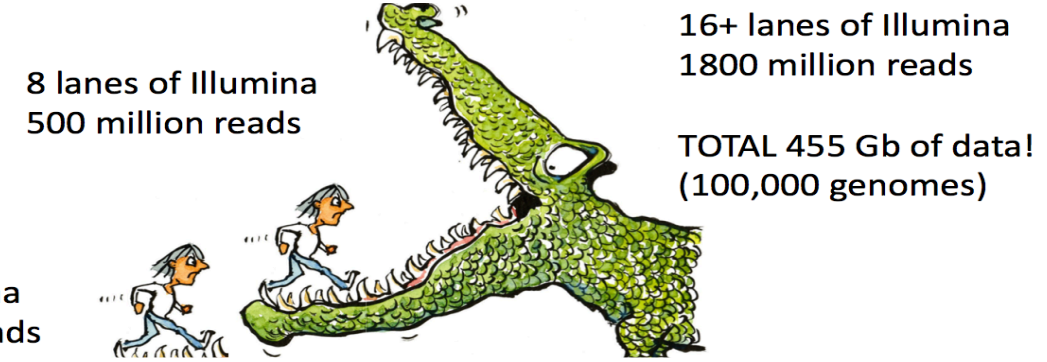
Shotgun sequencing and de novo assembly

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

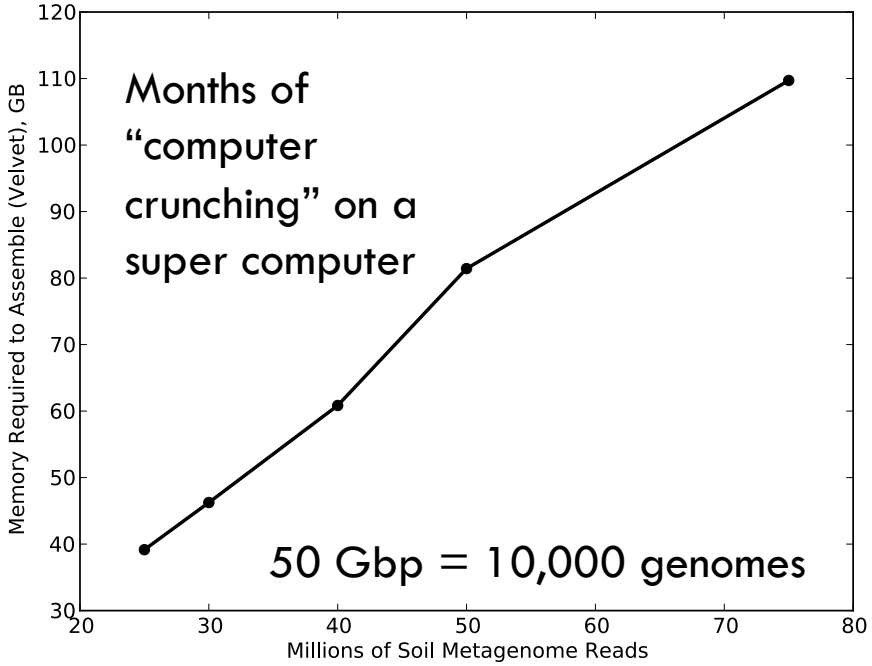
Practical Challenges – Intensive computing



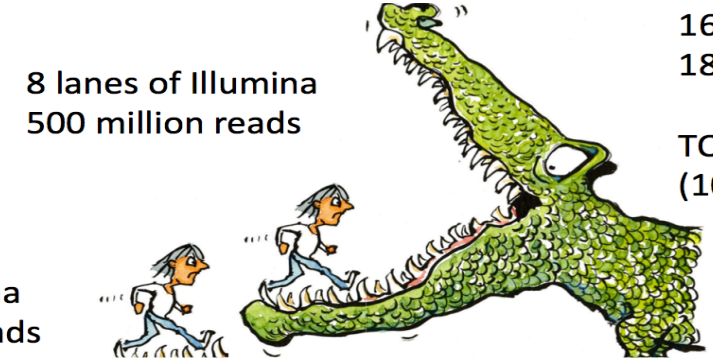


Howe et al, 2014, PNAS

Practical Challenges – Intensive computing



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.



Howe et al, 2014, PNAS

Four main challenges for *de novo* sequencing.

- Repeats.
- Low coverage.
- Errors

These introduce breaks in the construction of contigs.

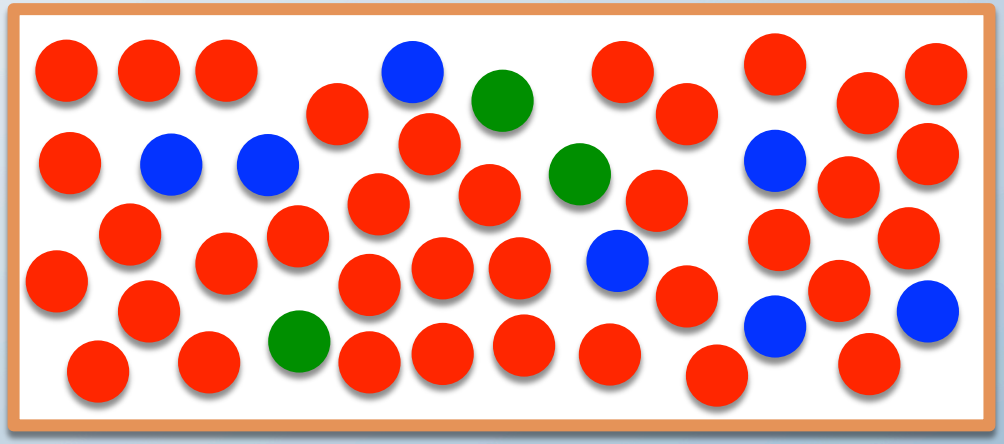
- *Variation* in coverage – transcriptomes and metagenomes, as well as amplified genomic.

This challenges the assembler to distinguish between erroneous connections (e.g. repeats) and real connections.

Natural community characteristics

- **Diverse**

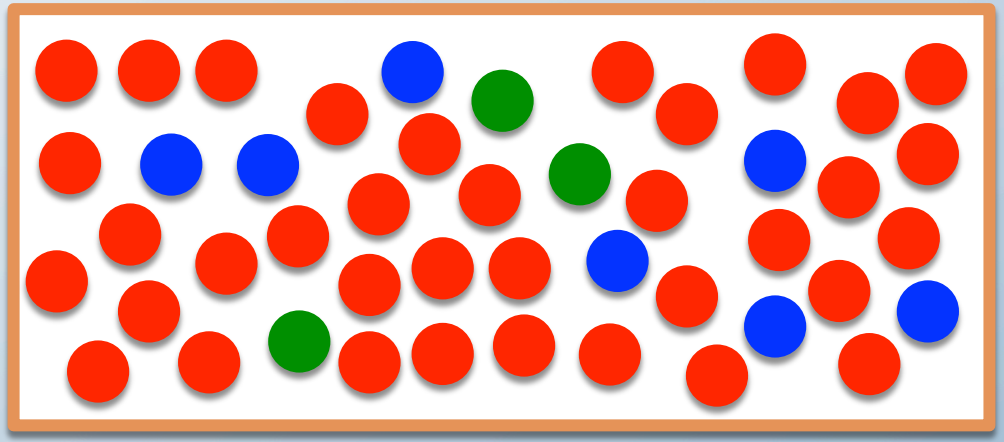
  ➔ Many organisms (genomes)



Natural community characteristics

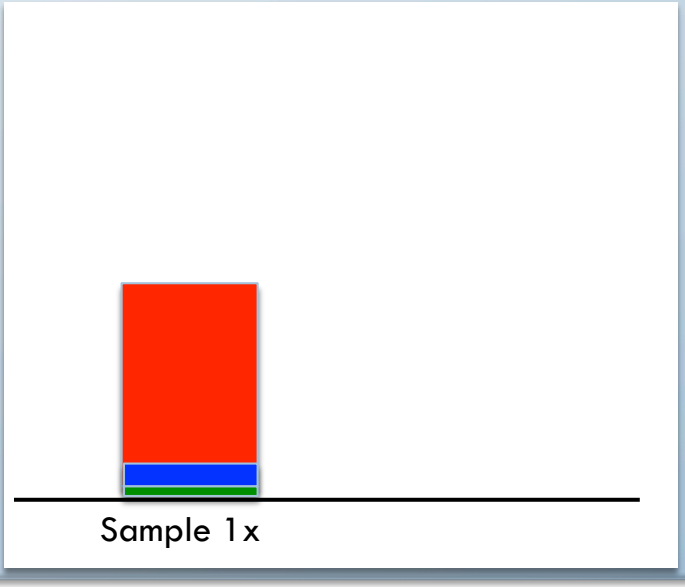
Diverse

➔ Many organisms (genomes)



Variable abundance

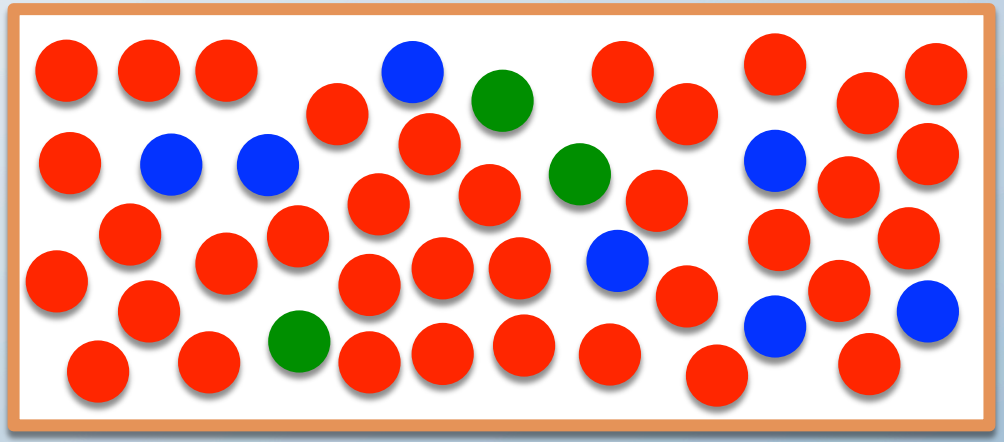
➔ Most abundant organisms, sampled more often
➔ Assembly requires a minimum amount of sampling
➔ More sequencing, more errors



Natural community characteristics

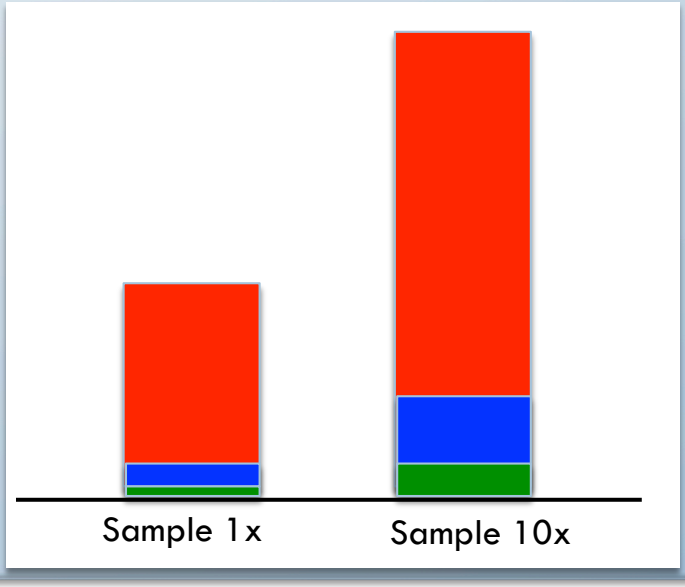
Diverse

➔ Many organisms (genomes)



Variable abundance

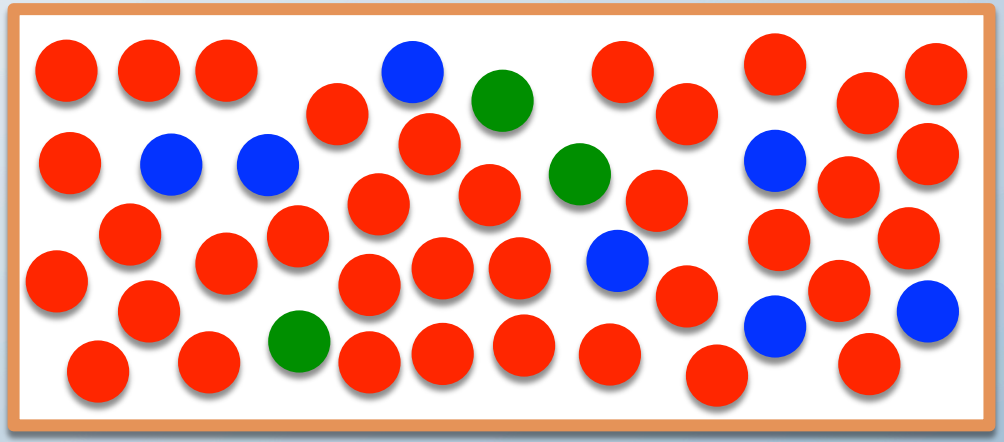
➔ Most abundant organisms, sampled more often
➔ Assembly requires a minimum amount of sampling
➔ More sequencing, more errors



Natural community characteristics

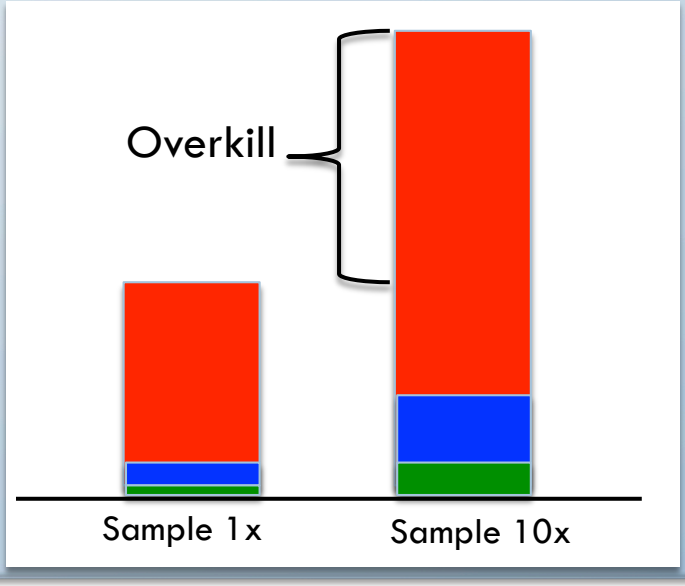
Diverse

➔ Many organisms (genomes)



Variable abundance

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

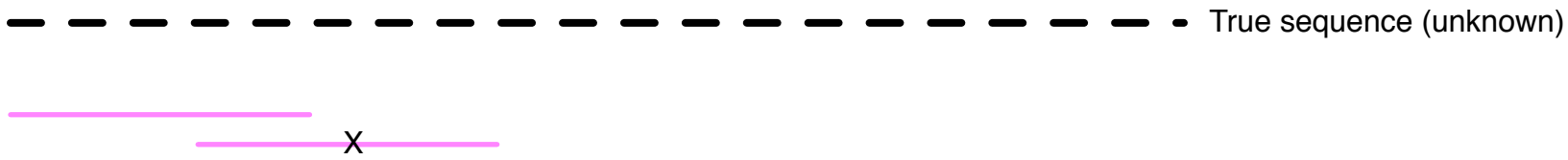


Digital normalization

Reads
(randomly sequenced)

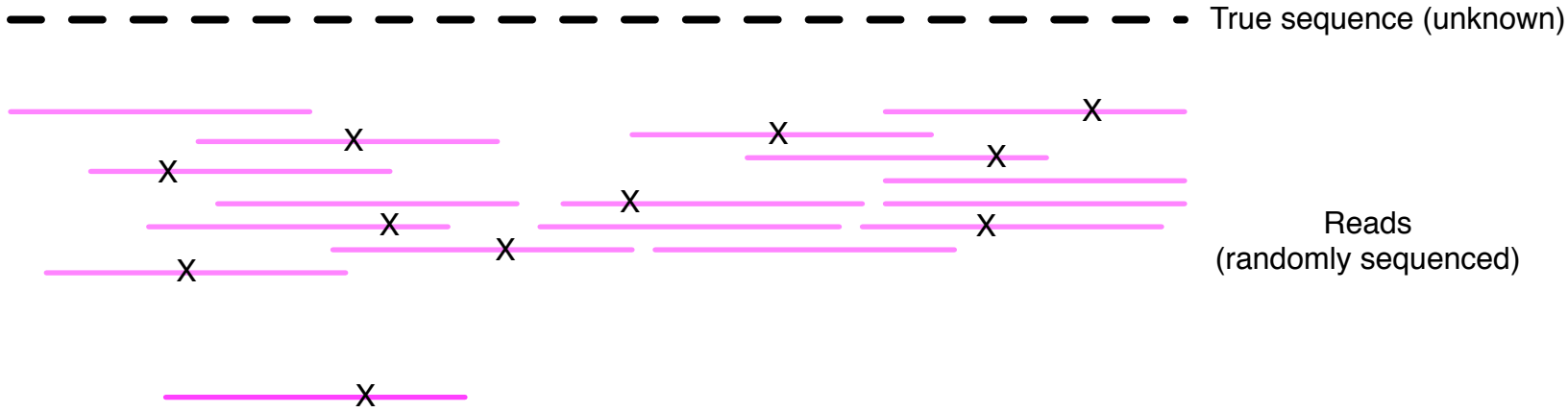
Brown et al., 2012, arXiv
Howe et al., 2014, PNAS
Zhang et al., 2014, PLOS One

Digital normalization



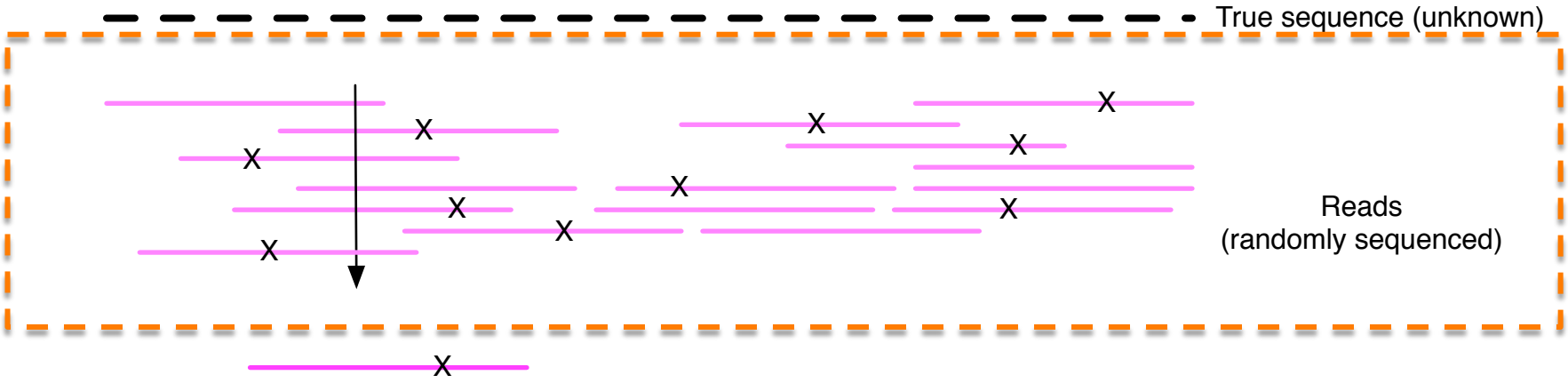
Brown et al., 2012, arXiv
Howe et al., 2014, PNAS
Zhang et al., 2014, PLOS One

Digital normalization



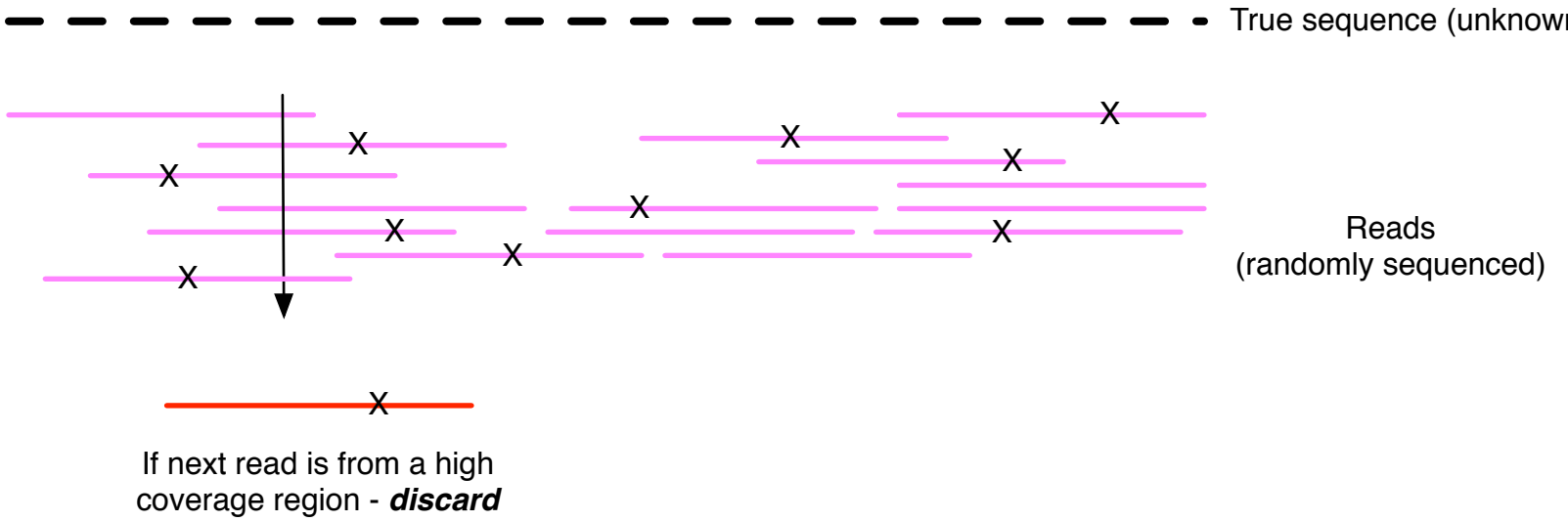
Brown et al., 2012, arXiv
Howe et al., 2014, PNAS
Zhang et al., 2014, PLOS One

Digital normalization



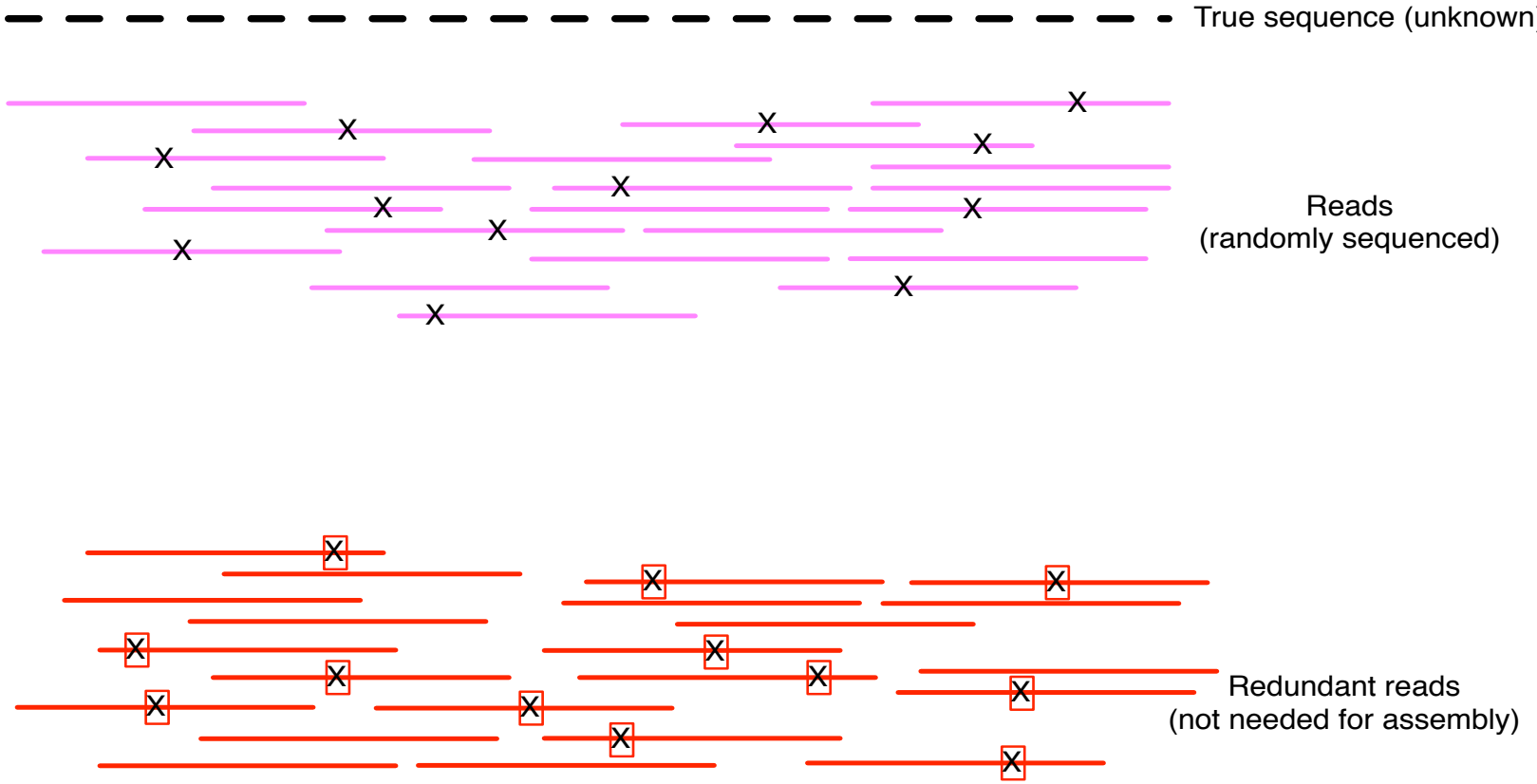
Brown et al., 2012, arXiv
Howe et al., 2014, PNAS
Zhang et al., 2014, PLOS One

Digital normalization



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

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