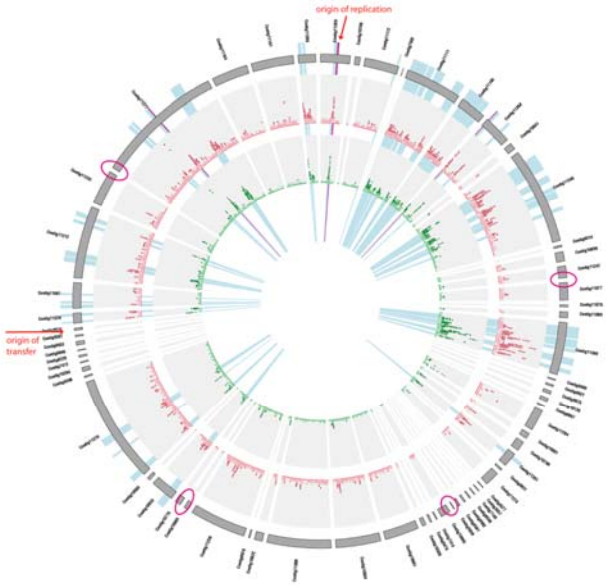
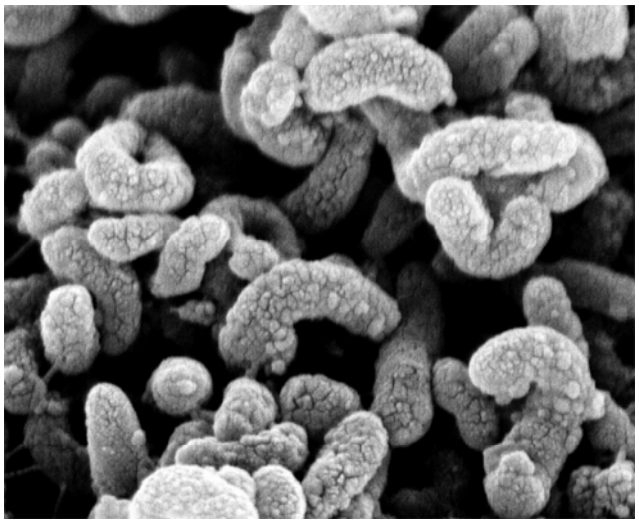
Diversity in microbial populations in the wild

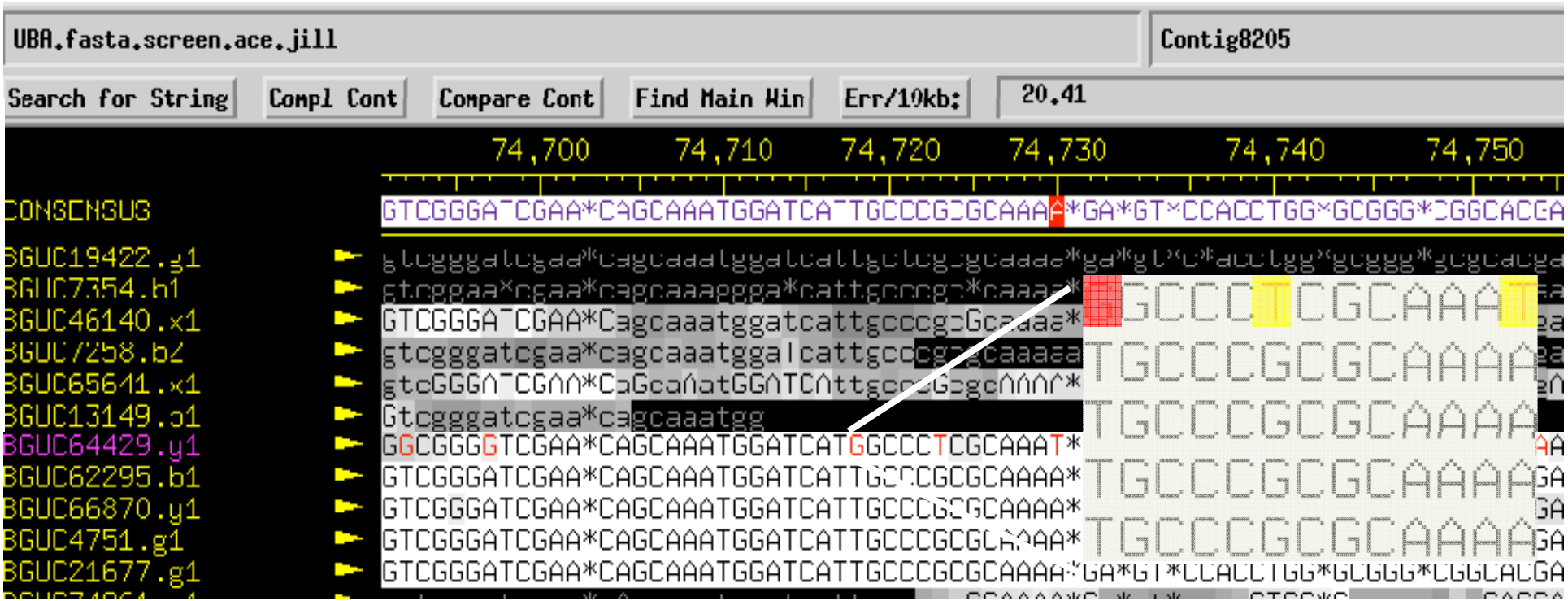


Sheri Simmons
Assistant Scientist, Bay Paul Center
Marine Biological Laboratory
Woods Hole, MA

Sequencing: a window into microbial life

Cells of the iron oxidizing bacterium *Leptosprillum* group II (Clara Chan)

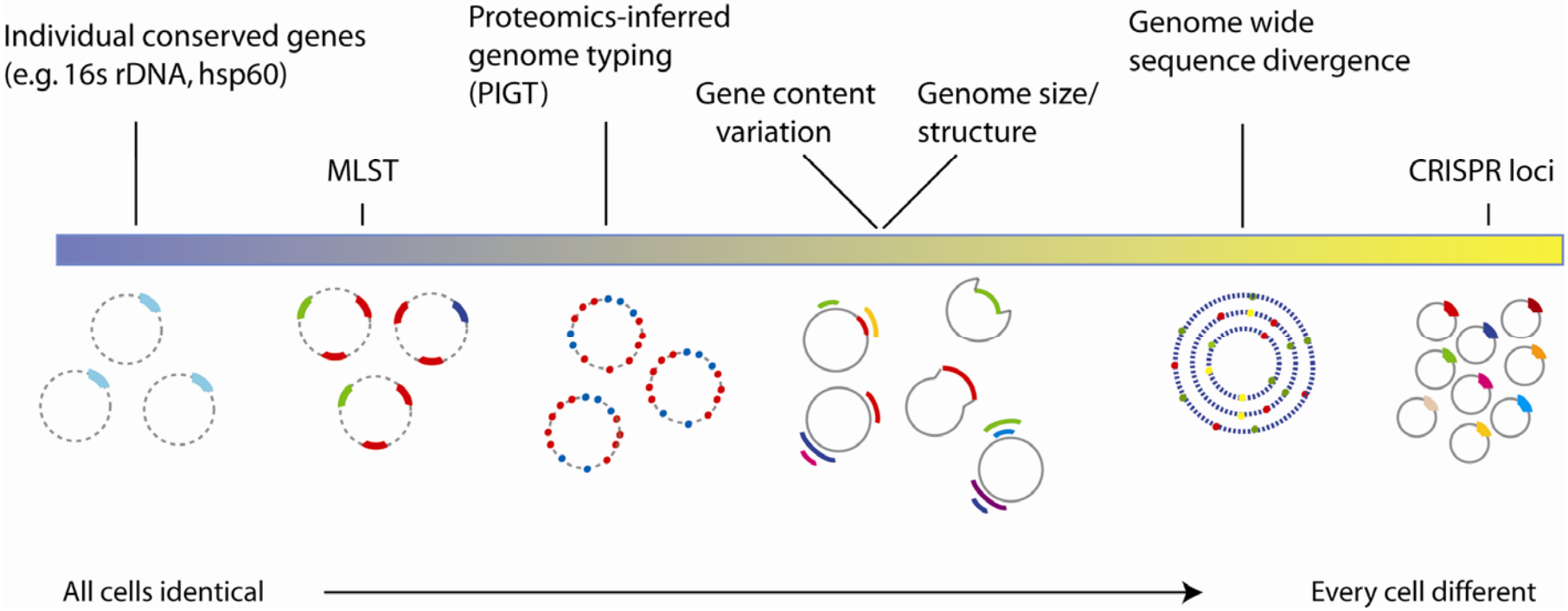




Sequencing allows us to access genetic variation within microbial communities

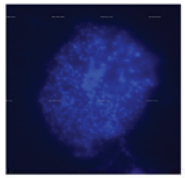
- What processes regulate its distribution and dynamics?
  - Key evolutionary rates largely unknown in most communities (HGT, recombination)
  - Importance of neutral processes relative to selection
- How do ecological parameters affect rates of fundamental evolutionary processes in communities?

Spectrum of variation accessible through sequencing



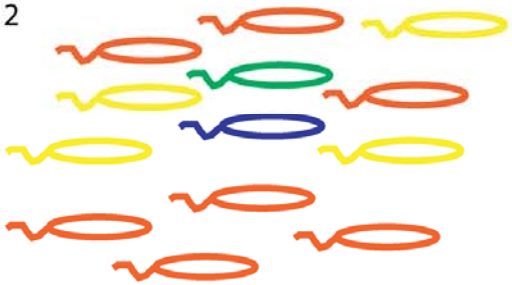
Wilmes, Simmons et al. (2009) FEMS Microbiol Rev 33:109-32.

1



Natural microbial community

2

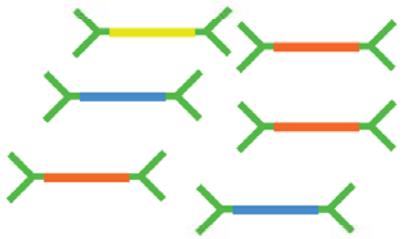


Multiple species of bacteria coexist

3

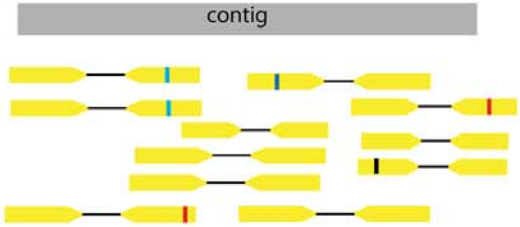
Isolate and fragment DNA

4

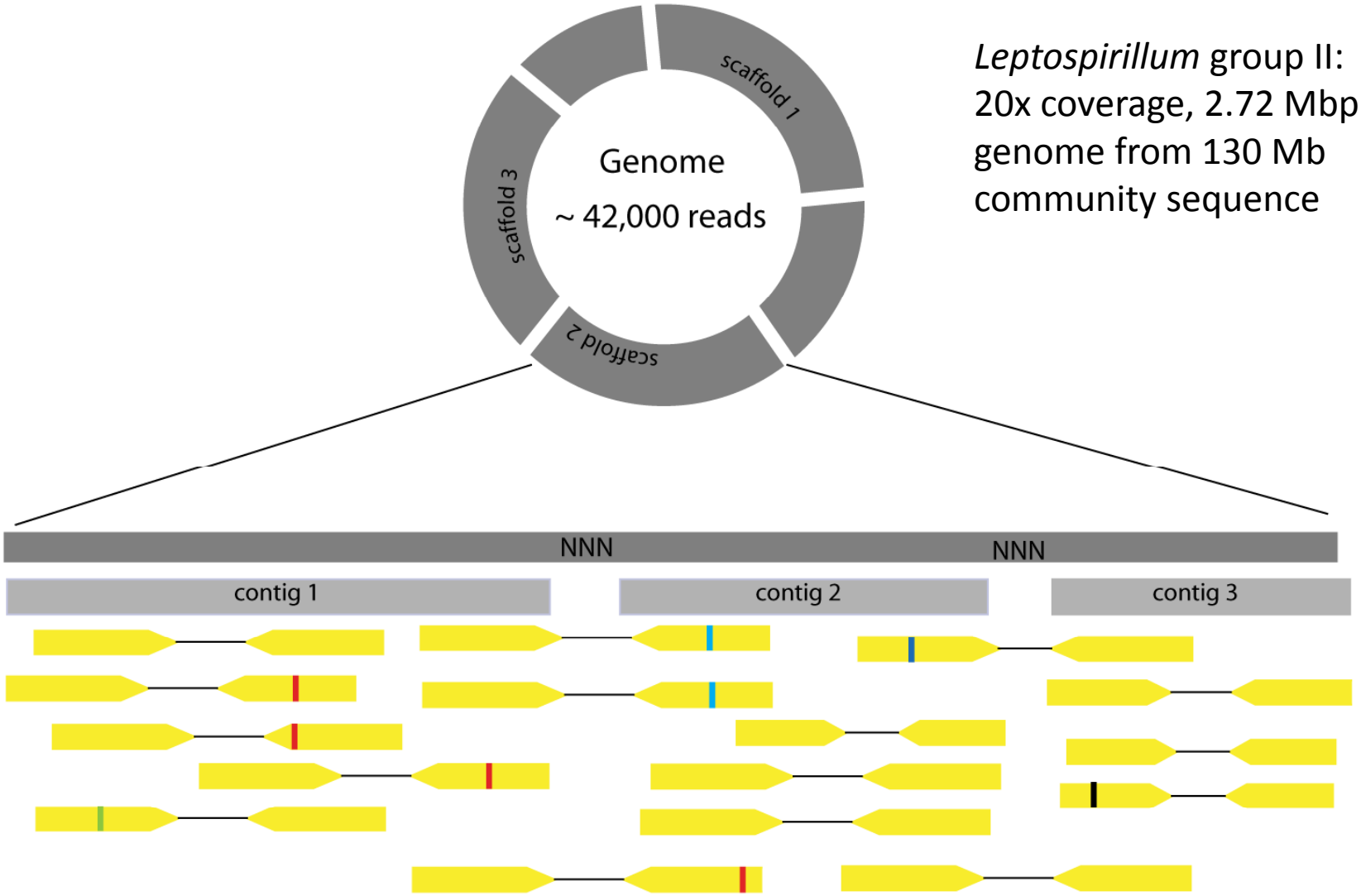


Build and sequence a molecular library

Annotate reads and bin by functional category;
Compare the distribution of functions between samples



Assemble partial "population genomes":
Analysis of genetic variation and evolutionary processes



Leptospirillum group II: 20x coverage, 2.72 Mbp genome from 130 Mb community sequence

Variation is visible within the population genome

New technology greatly increases power of sequencing for metagenomics

Old- ABI capillary sequencer

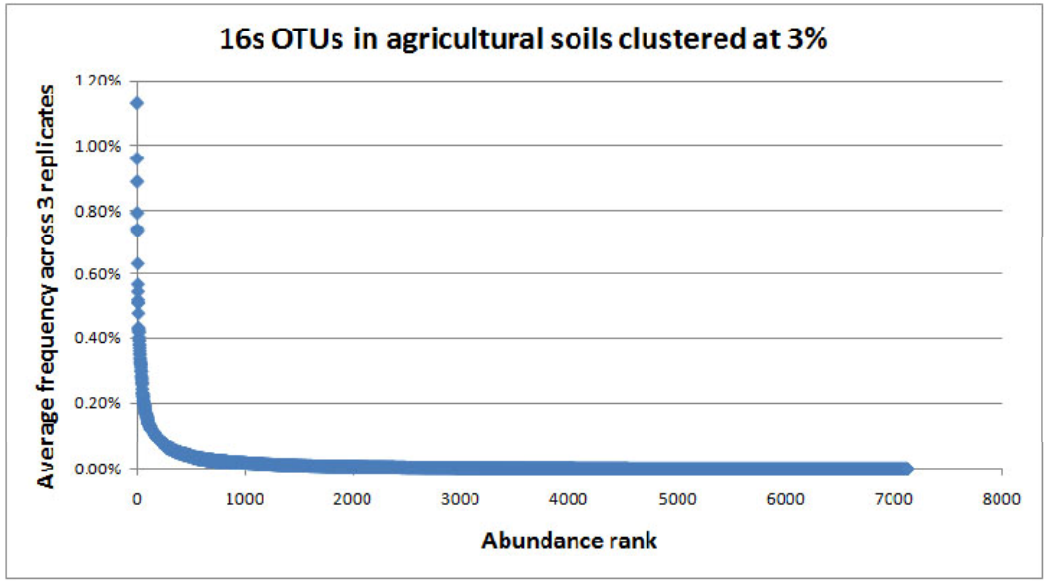
- 1,000 reads/day =~ 20% of a microbial genome
- ~700-1000 bp reads
- First microbial metagenome: 100 Mbp, published 2004

New- Illumina HiSeq 2000

- 100 million reads/day =~ 10,000 microbial genomes
- Reads are 35-150 bp
- 200 Gigabases sequence output from a single run
- First metagenome: 572 Gbp, published 2010
- Not yet widely used for metagenomics, but tremendous potential

So why don't we sequence everything?

- Most microbial communities in nature have highly skewed abundance distributions
- We undersample nearly all species in the community using shotgun metagenomics



Cardon, Simmons (unpub)

Illumina HiSeq 2000	Max yield 200 Gbp per run
Reads per lane	125 million
Gbp per lane	12.5
Yield of a genome present at 1% of the community	125 Mbp
Coverage of a 5 Mbp genome at 1%	25x
Coverage of a 5 Mbp genome at 0.1%	2.5x

Why don't we sequence everything (2)

- Cost of 1 lane of paired end sequencing on a HiSeq 2000 (max yield = 125 million reads) is ~$2500
- But…
  - Cost of constructing 1 genomic library = ~$500
  - Cost of constructing one meta-transcriptomic library = ~$900
  - To multiplex 10 genomic samples your cost is ~$5,000
- The biggest cost is computational!
  - 2 Tb of data produced per run of the HiSeq
- Assembly, mapping, and annotation are computationally intensive
- Pipelines are not standardized

Central questions in microbial ecology

- What is the relative importance of selective and neutral processes in microbial populations?
  - Crucial for determining the functional significance of diversity
  - Tells us how communities will recover from perturbation
- How repeatable is community assembly over space and time?
- What is the rate of genetic exchange in the wild?
- Are there common dynamic processes in microbial communities?
- How do patterns of microbial diversity depend on spatial scale?

Outline

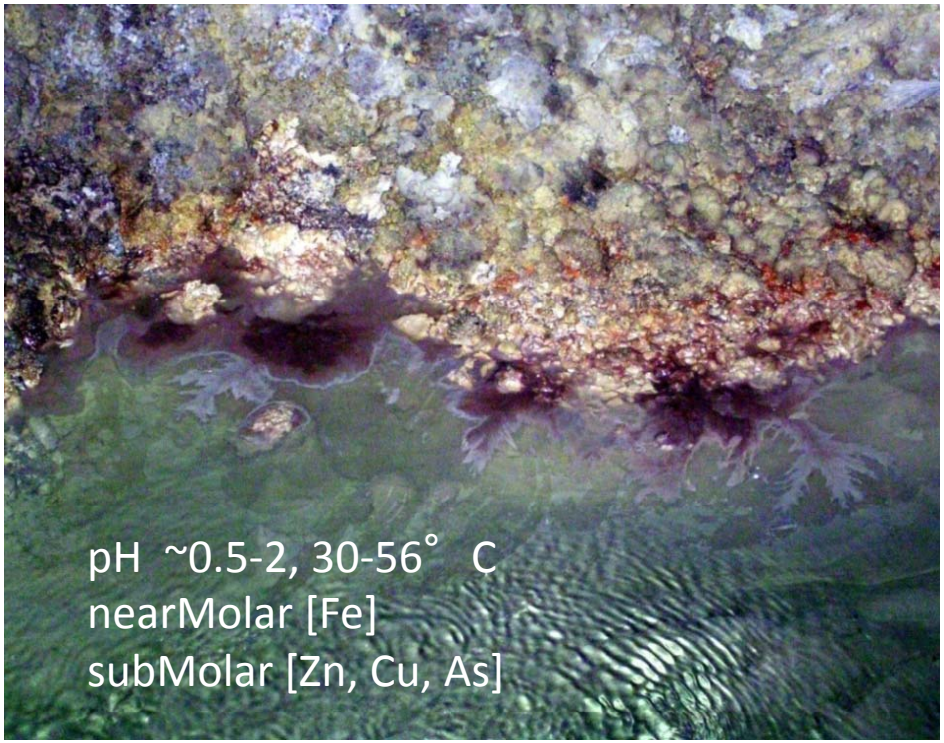
Microevolutionary variation in microbial communities in acid mine drainage (work with Jill Banfield)

At the MBL:

1. Model microbial systems:
   - **Dynamics of microbes on the surface of plant leaves**
   - *Horizontal gene transfer in natural communities*
2. Open microbial systems:
   - Haplotype tracking in mixed communities
   - *Taxonomic, genetic, and functional diversity in agricultural soils*

Microevolutionary variation in acid mine drainage microbial communities

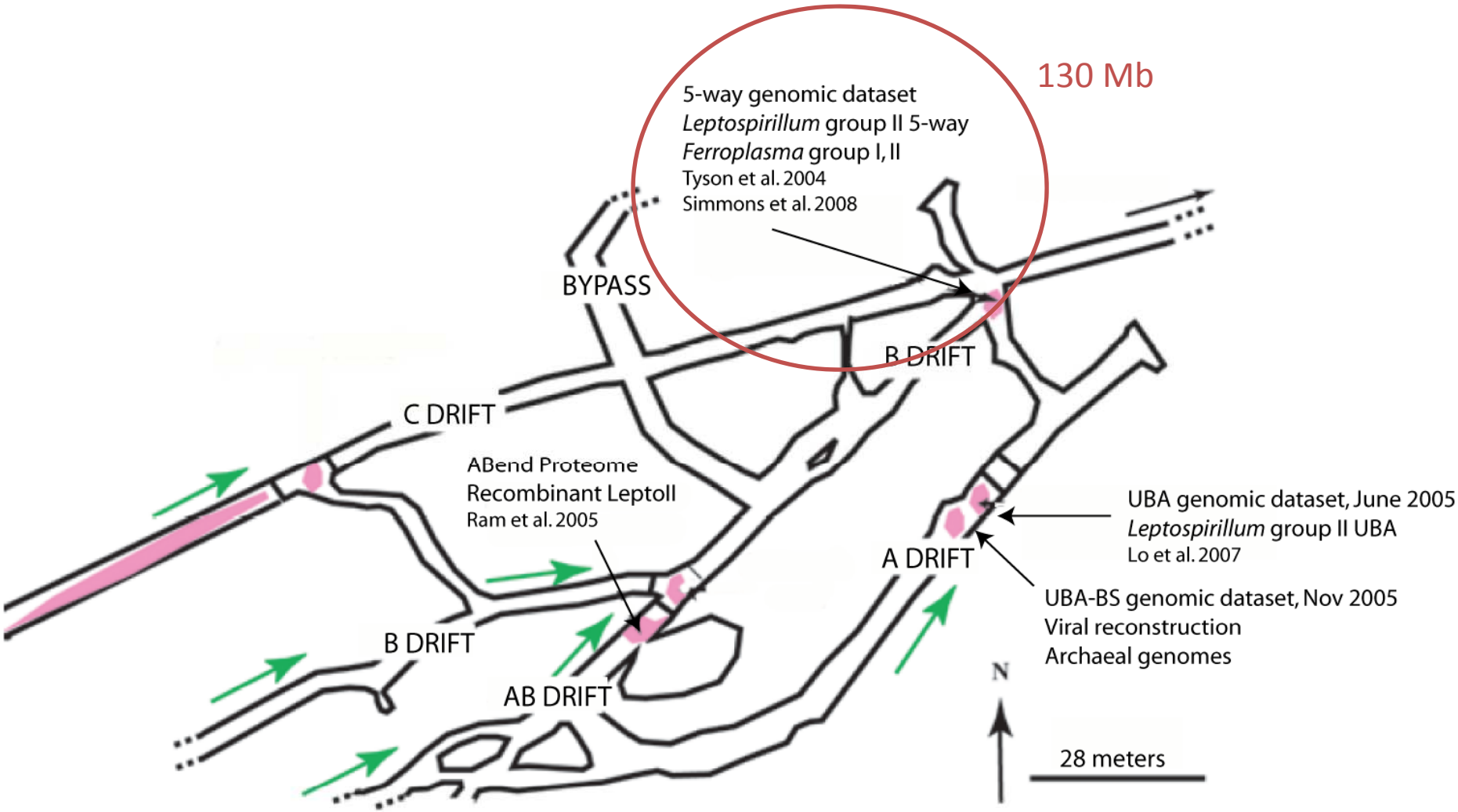
- Sequencing of microbial communities in acid mine drainage
  - Does selection maintain variation within a population?



Postdoctoral work with Jill Banfield, UC Berkeley

DOE GENOMICS:GTL
SYSTEMS BIOLOGY FOR ENERGY AND ENVIRONMENT
OFFICE OF SCIENCE
U.S. DEPARTMENT OF ENERGY

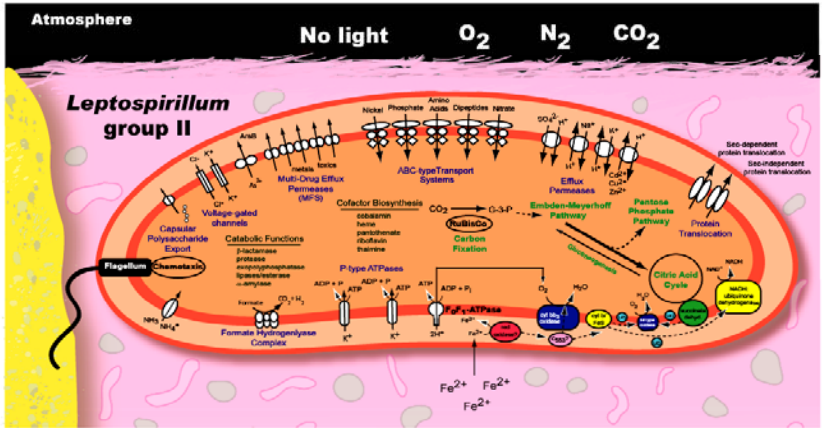
NSF



Sampling location of all community genomic datasets within the Richmond Mine Iron Mountain, California

A snapshot of the 5-way biofilm community

- Dominated by *Leptospirillum* group II: ~20x coverage, 2.72 Mbp complete genome
- *Ferroplasma acidarmanus,* other archaea, *Leptospirillum* group III present at low abundance
- Iron oxidation is the predominant metabolic process



pH ~ 0.7, 30-56° C
nearMolar [Fe]
subMolar [Zn, Cu, As]

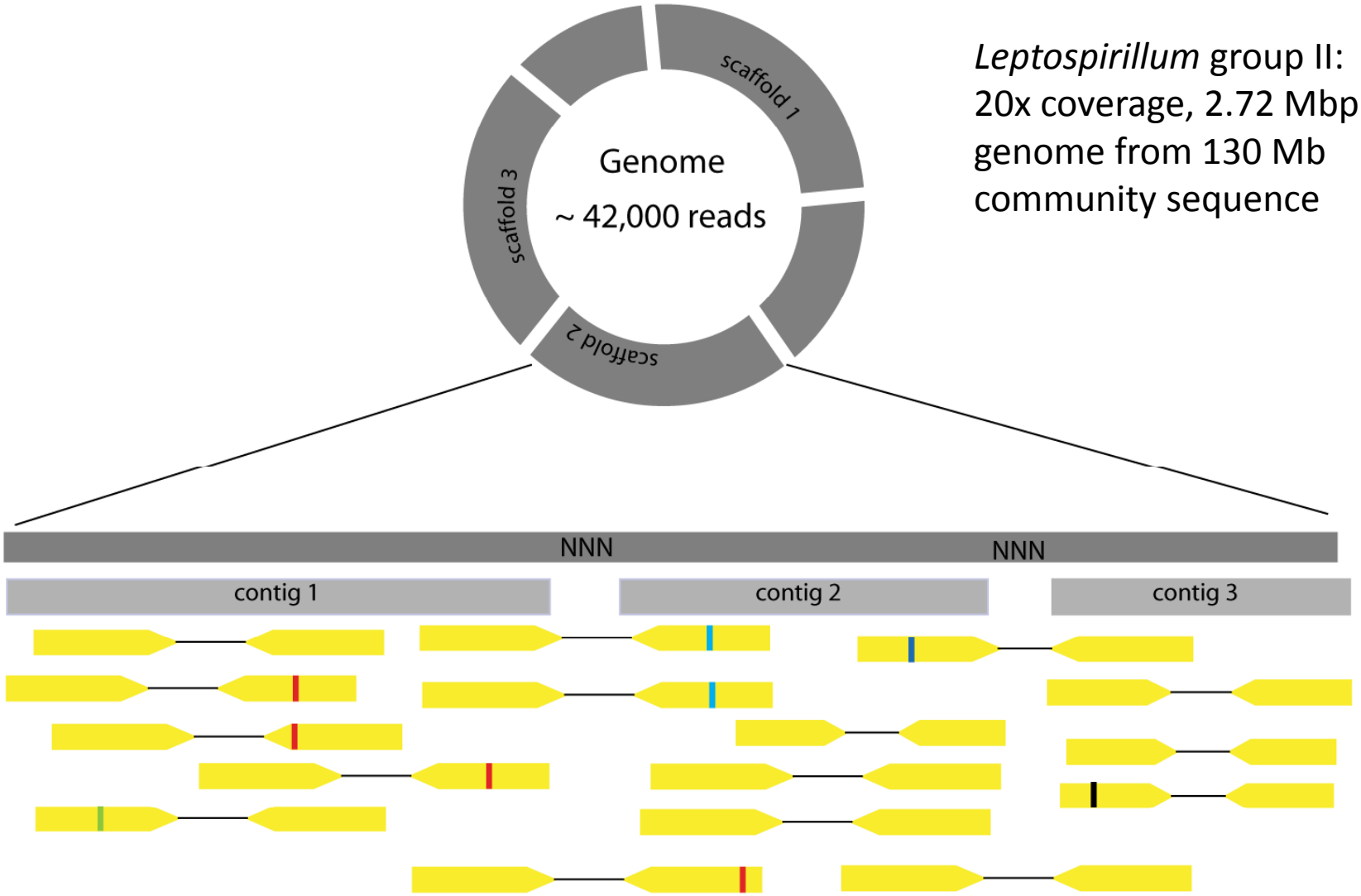
Net increase in soluble metal

Acidification

$$FeS_2 + 14Fe^{3+} + 8H_2O \rightarrow 15Fe^{2+} + 2SO_4^{2-} + 16H^+$$

Microbes

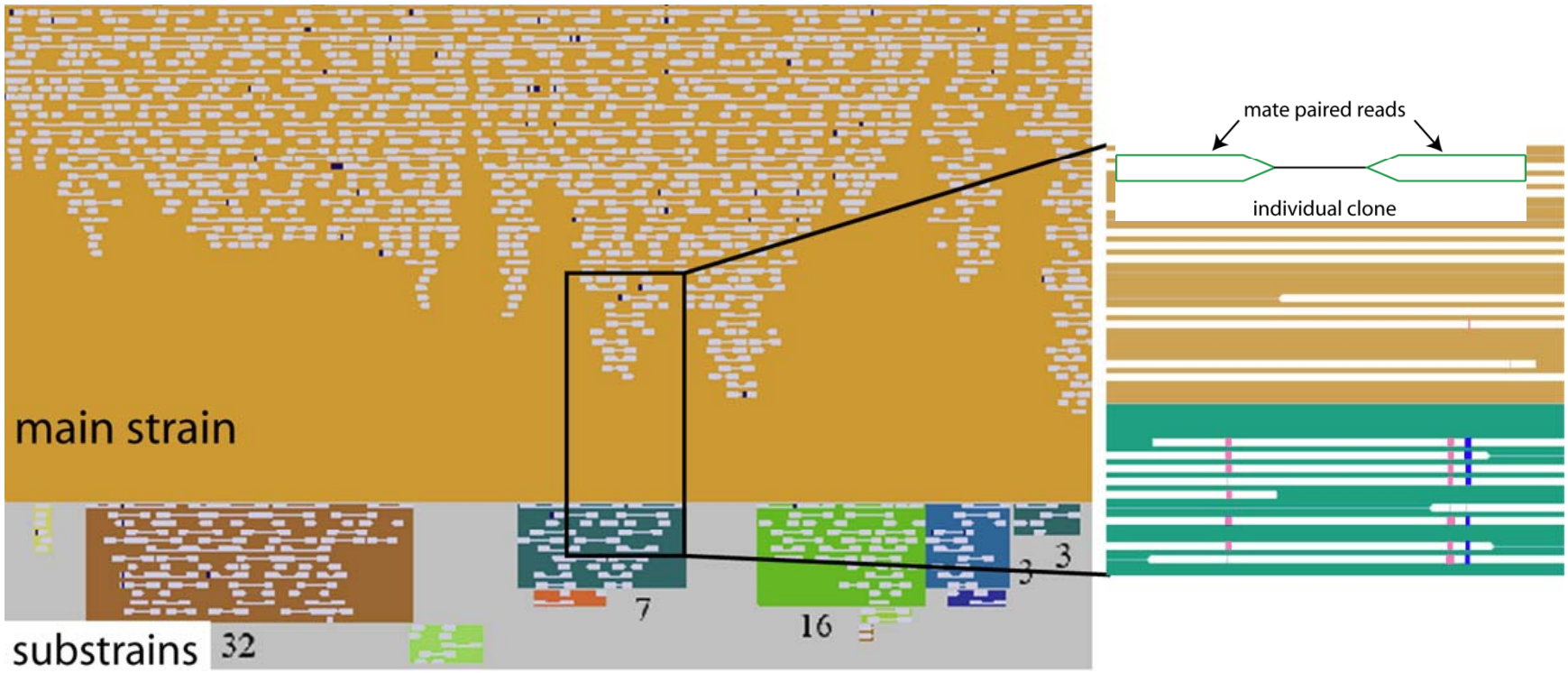
Tyson et al. *Nature* 2004



Leptospirillum group II: 20x coverage, 2.72 Mbp genome from 130 Mb community sequence

Variation is visible within the population genome

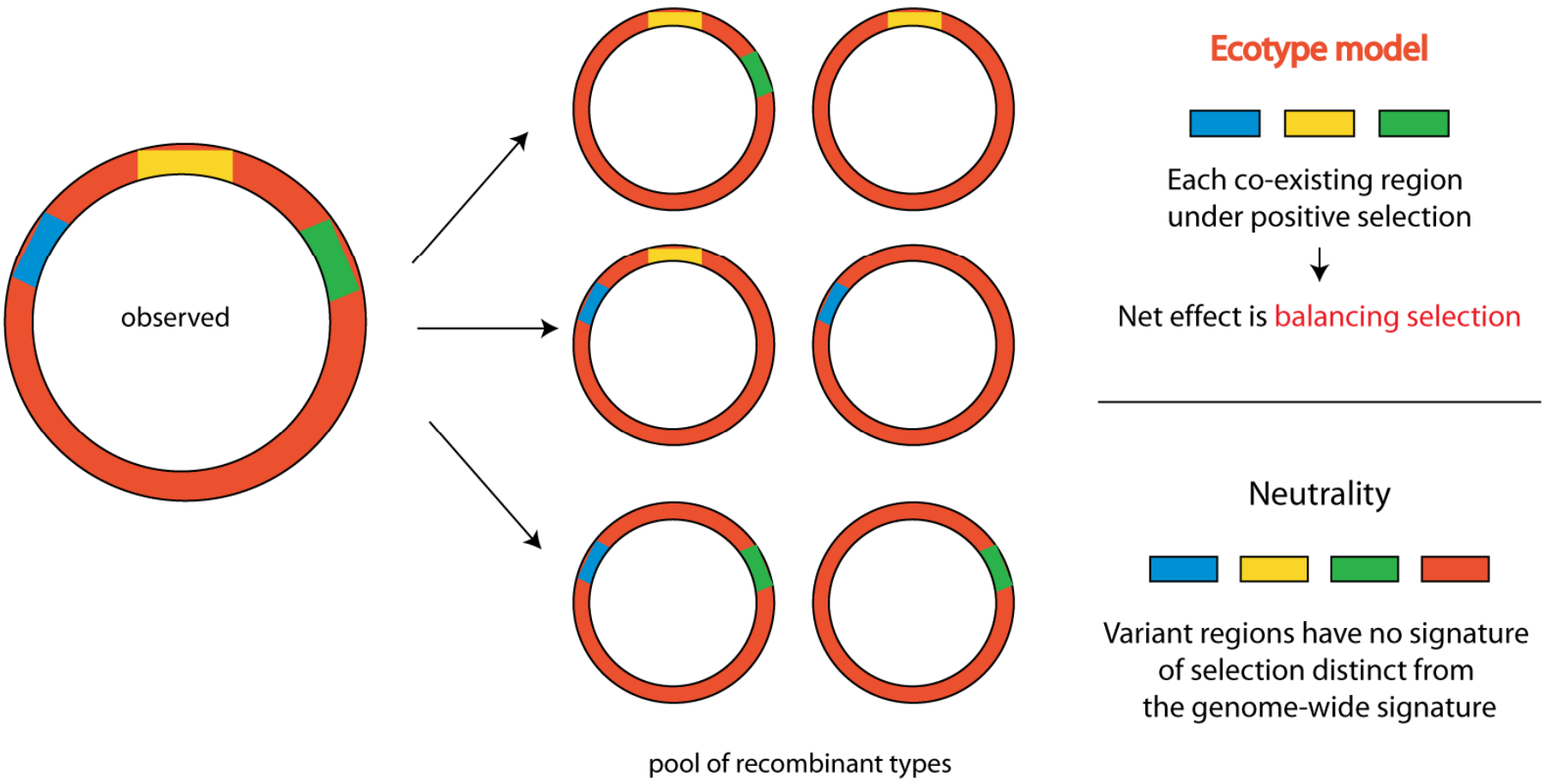
Distinct strains are present within the population



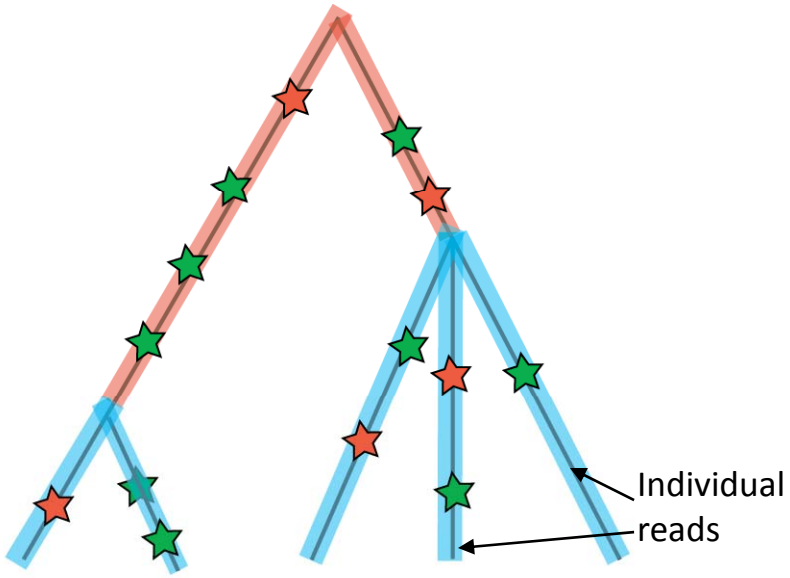
We use synonymous and nonsynonymous polymorphisms to test potential adaptive value of strains

Simmons et al. 2008, PLoS Biol. 6(7): e177

Hypotheses explaining strain variation

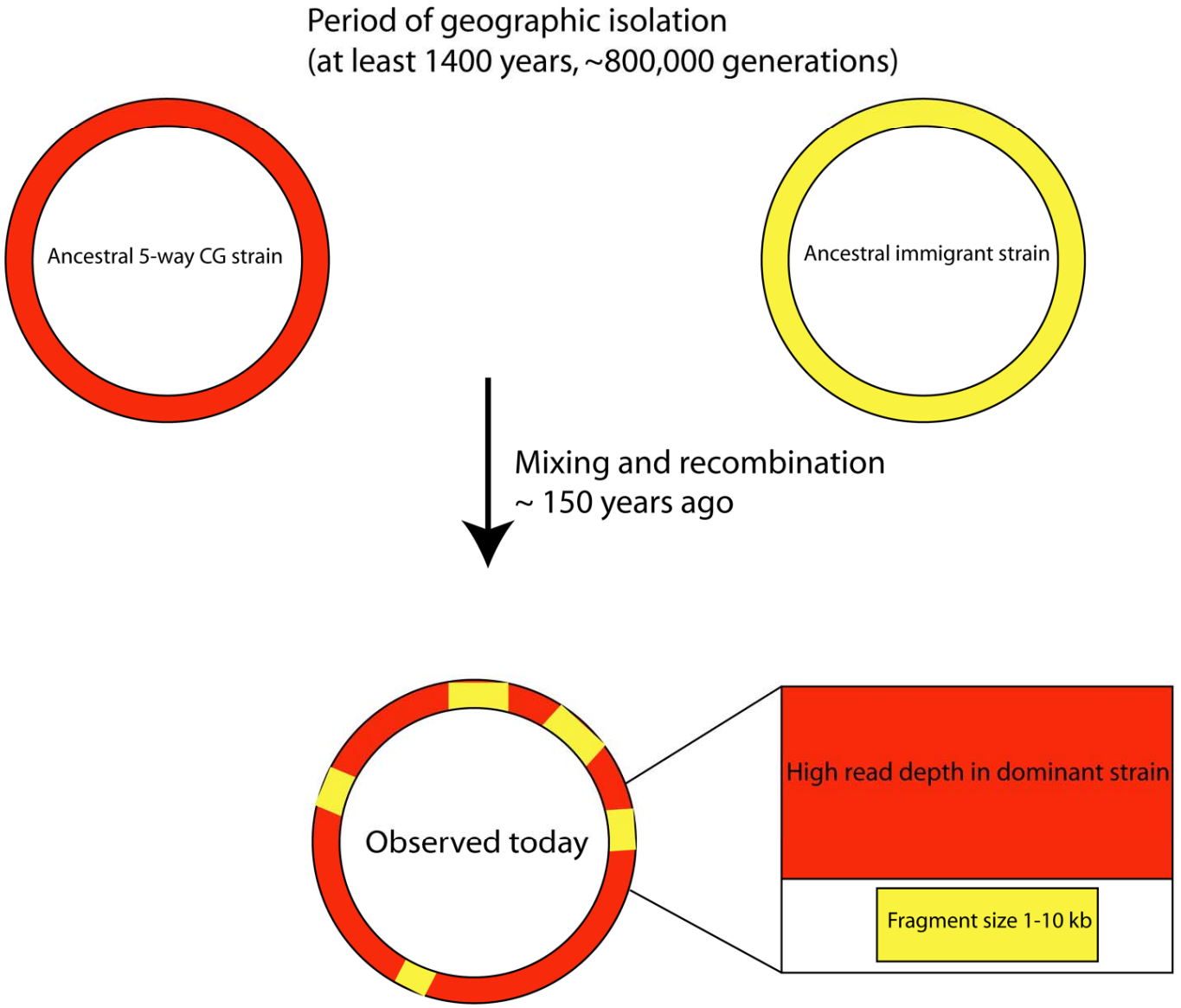


McDonald-Kreitman test



Between species		Within species	
# of NS subs between species / # of syn subs between species	=	# of NS subs within species / # of syn subs within species	Neutral evolution
# of NS subs between species / # of syn subs between species	>	# of NS subs within species / # of syn subs within species	Positive selection
# of NS subs between species / # of syn subs between species	<	# of NS subs within species / # of syn subs within species	Negative selection

Lack of selection on strains supports isolation and drift model



Model versus open microbial systems

- Why do we need a model system for microbial ecology?
  - Most communities of environmental importance are very hard to manipulate experimentally: e.g. soils and human gut
  - Open systems do not allow accurate estimates of key parameters in the neutral model: dispersal rate and source pool composition
  - Biological replication is crucial to establish patterns
  - Controlled time series sampling is needed to determine the repeatability of assembly processes

The tomato phyllosphere: a new model system

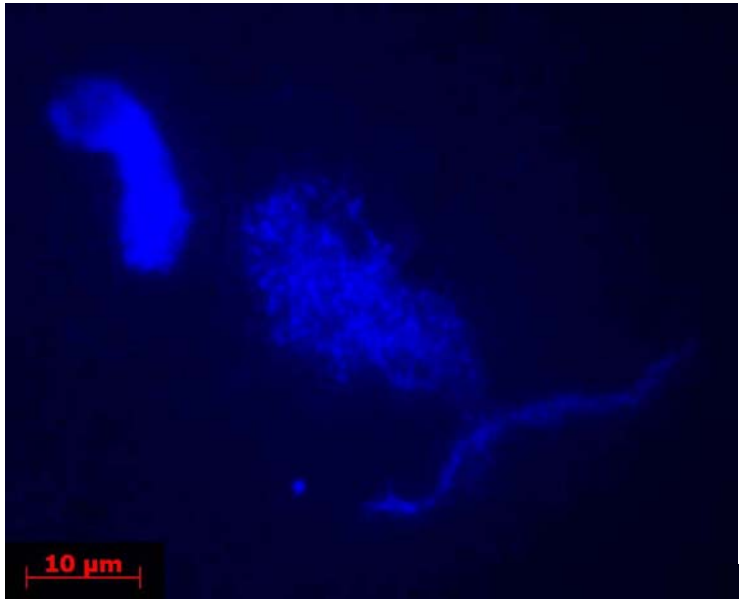
- Rapid growth, maintained under controlled conditions in the MBL greenhouse
- Large number of biological replicates
- Continual emergence of new leaf substrates
- Direct measurements of dispersal rates and composition of colonizing pool
- Complex, yet controlled

17-day old tomato seedlings growing in the MBL Greenhouse

Seedlings germinated in sterile phyto-agar in lab incubator

11-day seedling

Compound leaf of *Solanum lycopersicum* "Green Zebra"

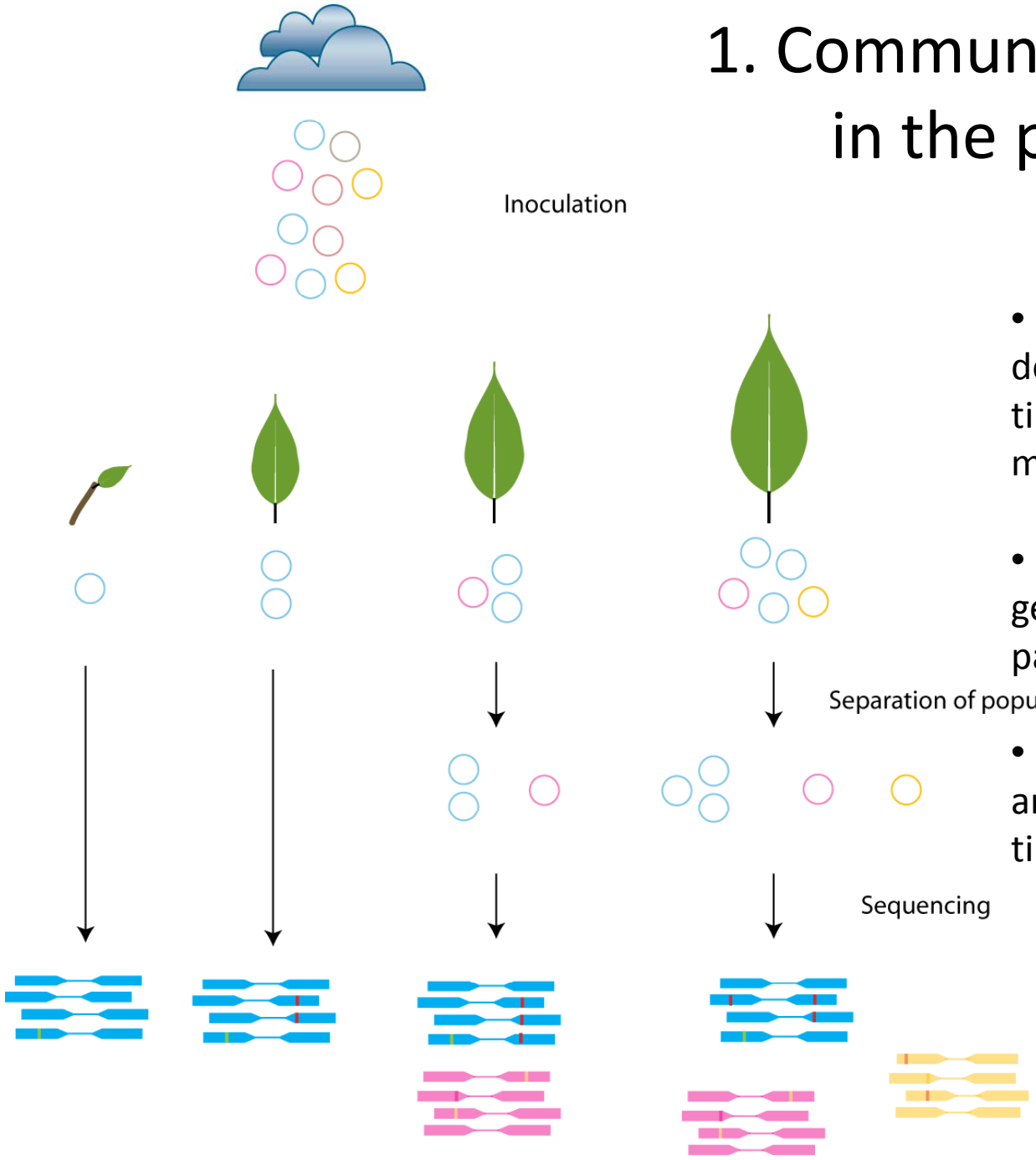


Microbial biofilm detached from the leaf surface

Key questions

1. Do microbial communities develop in parallel over time?
2. Is the development of genomic variation over time repeatable?
3. What is the contribution of niche versus neutral processes to community assembly?

1. Community development in the phyllosphere



- Do parallel communities develop in the same way over time given the same starting material?
- Do taxonomic and functional genes show similar dynamical patterns?
- Large-scale metagenomic and amplicon sequencing of replicate time series samples

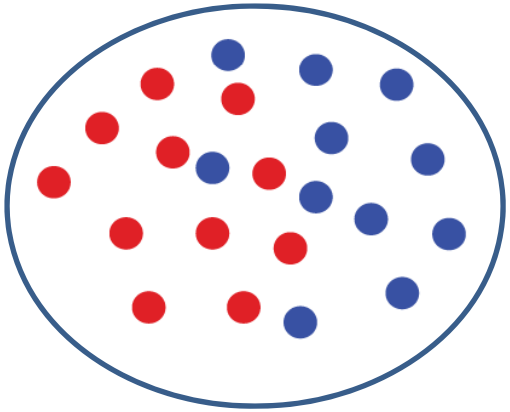
2. Testing neutral and niche based models of community assembly

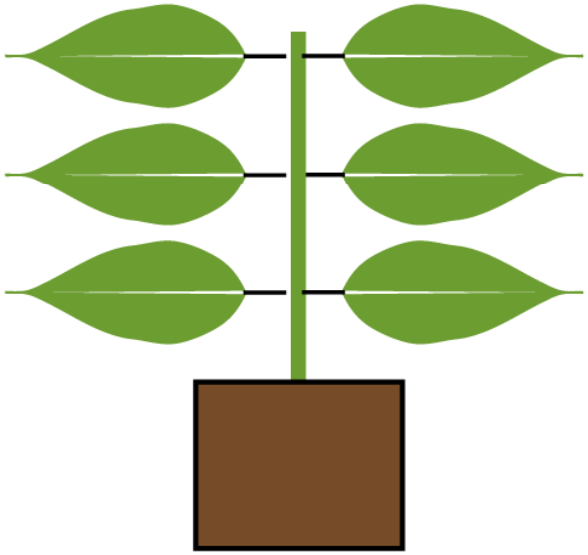
- Prediction of the neutral model: the average abundance of a taxon across multiple local communities = its abundance in the source pool

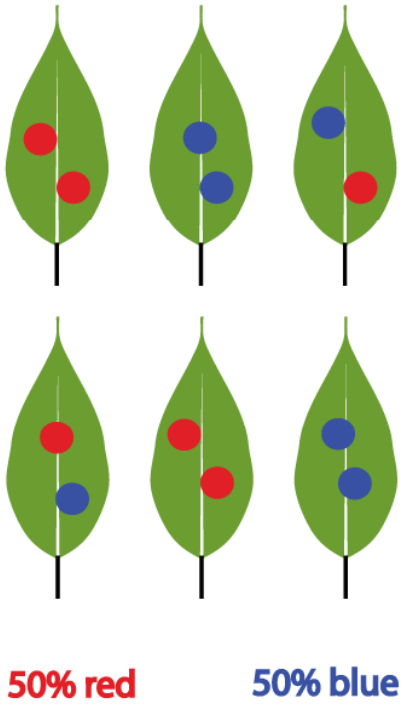
- We will measure both parameters directly for taxonomic and functional genes

Prediction of the neutral model

Capture airborne microbes





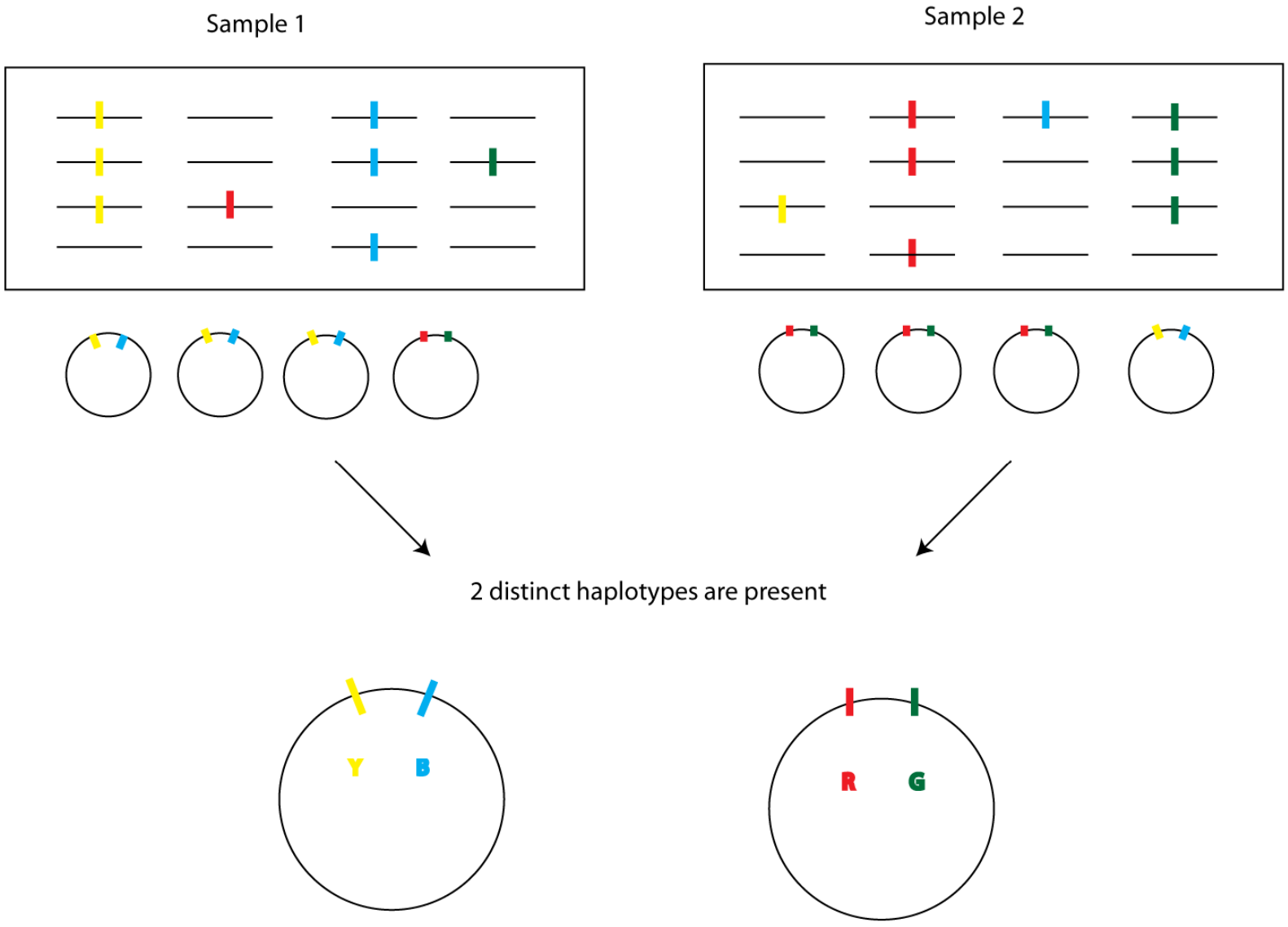


Wash from leaf surface

Experimental design

- Do taxonomic and functional gene abundances fit the simple neutral model, or are they correlated with abiotic characteristics?
- We will quantify:
  - Size and composition of the airborne colonizing pool
  - The migration rate onto leaf surfaces
  - The taxonomic and functional composition of individual leaf communities using metagenomics
  - Abiotic conditions, including plant nutrient status

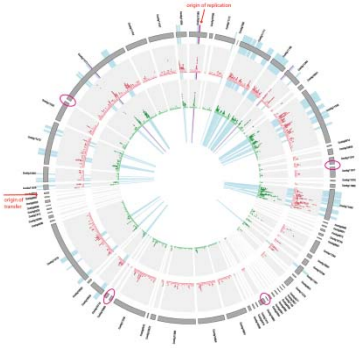
Determining haplotype frequencies in natural populations



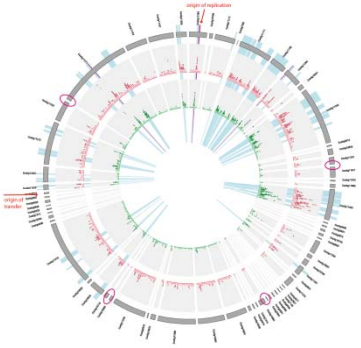
Approach

- Paired end Illumina sequencing of time series samples
- *De novo* assembly and alignment to reference genomes
- Custom pipeline for extracting allele frequencies
- Pilot study: Illumina sequencing of strain mixes in defined proportions


Applications Places System
34 °F Tue Feb 16, 3:31 PM
sheri
IGV
File View Tracks Help
Ferroplasma_acidarma...
Ferroplasma_acidarmanus_fer1
Go
6,286 bp
26,000 bp
27,000 bp
28,000 bp
29,000 bp
30,000 bp
31,000 bp
Coverage
Mix1_F_q25_sorted.bam
3 tracks loaded
168M of 275M
[sheri@ubuntu: ~]
sheri@ubuntu: ~
samtools.1 - Mozilla Fi...
[sheri@ubuntu: ~]
Tag sorting statistics (...
IGV



Summary



- Next-generation sequencing gives us tremendous power to investigate microbial evolution and ecology
- We can use it to ask large-scale questions about the fundamental processes structuring microbial communities in nature, in a way not previously possible
- A complementary approach using model and open systems allows targeting of ecological/evolutionary mechanisms of community assembly

Acknowledgements

MBL:

Meghan Chafee

Emelia DeForce

Zoe Cardon

Funding:

G. Unger Vetleson

foundation, anonymous donor

Banfield lab, UC Berkeley:

Vincent Denef

Paul Wilmes

Brett Baker

Funding: NSF, DOE