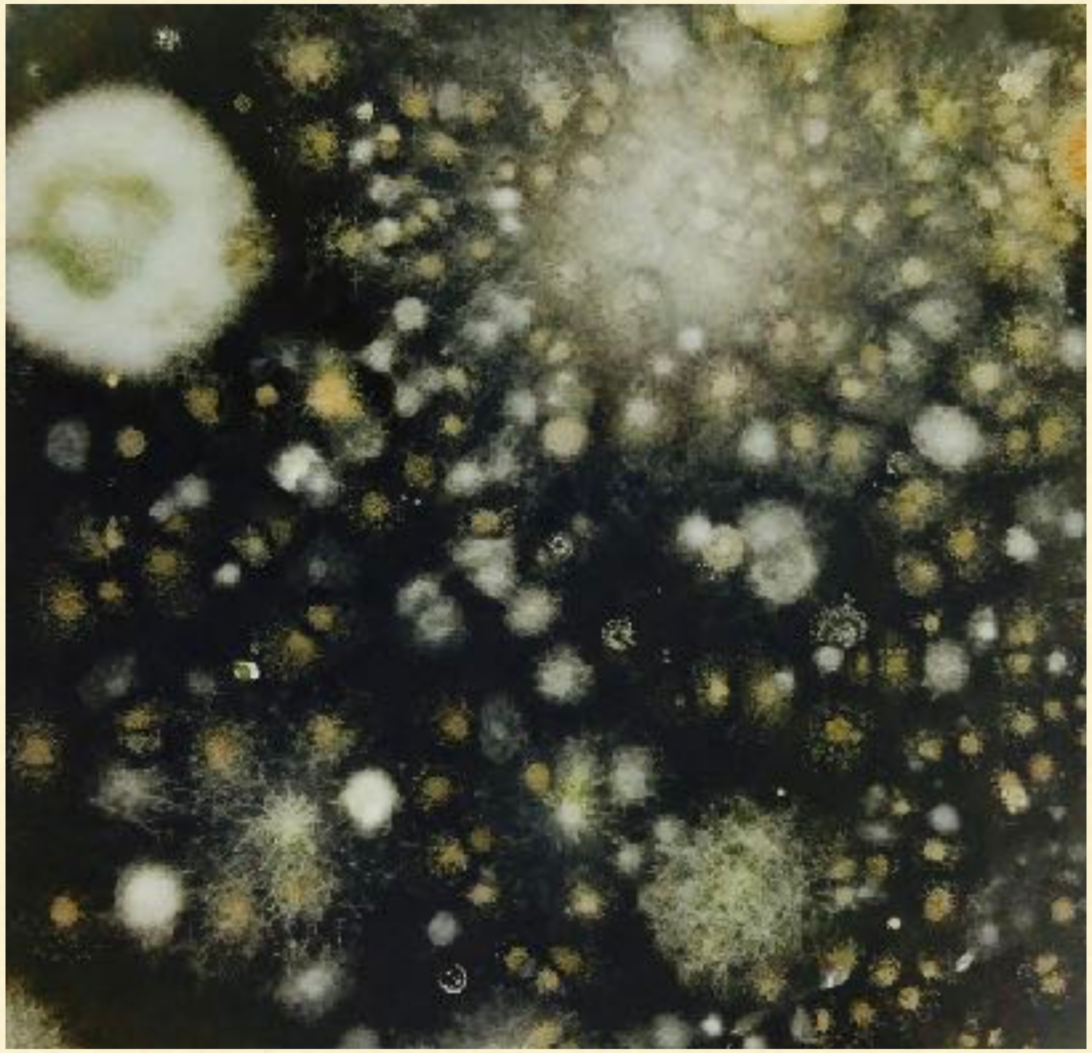


MICROBIOME DATA & ANALYSIS

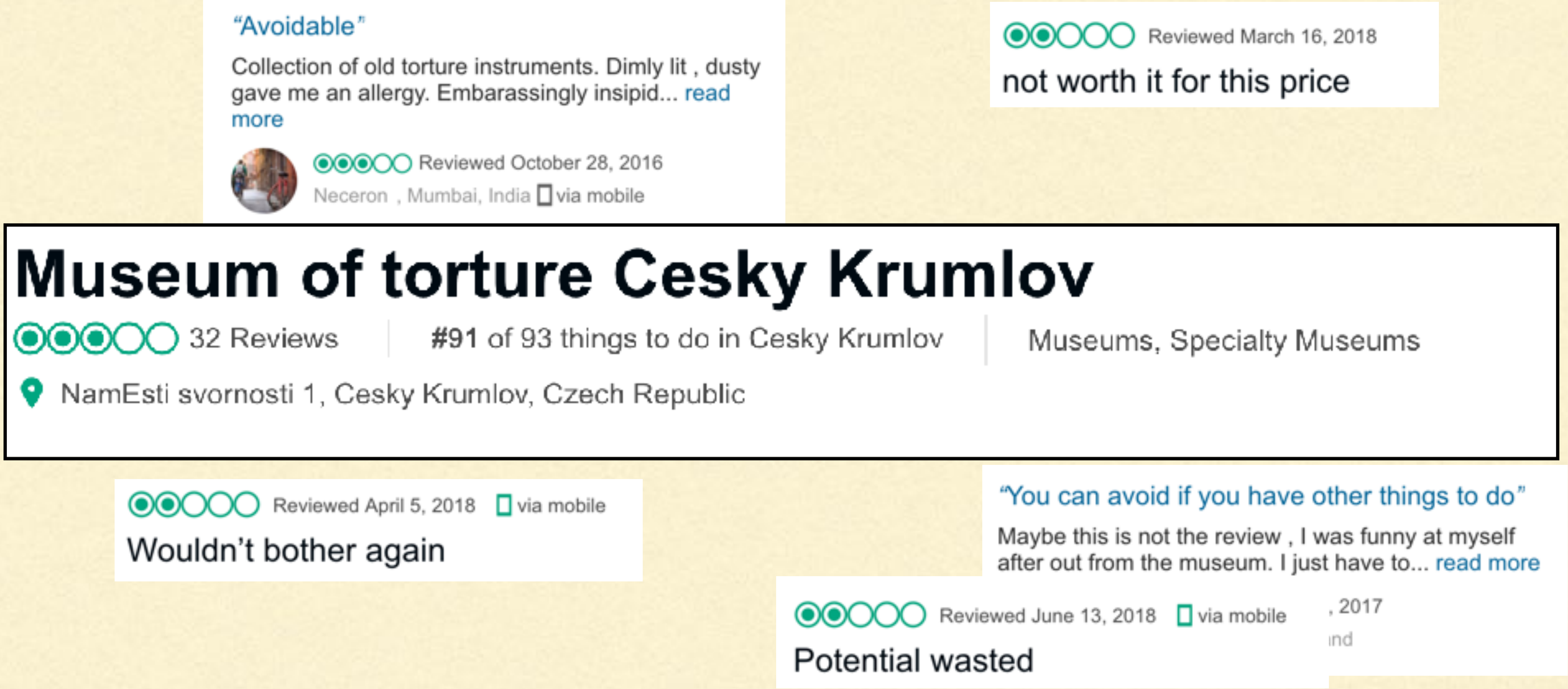
Research Group: Statistical Diversity Lab

PI: Amy D Willis PhD, Assistant Professor, Department of Biostatistics, UW

@AmyDWillis adwillis@uw.edu

Photo credit: T.D. Berry, Whitman lab, UW Madison

THANKS TO OUR SPONSOR



"Avoidable"

Collection of old torture instruments. Dimly lit , dusty gave me an allergy. Embarassingly insipid... read more

●●●○○ Reviewed October 28, 2016

Neceron , Mumbai, India ▯ via mobile

●●○○○ Reviewed March 16, 2018

not worth it for this price

Museum of torture Cesky Krumlov

●●●○○ 32 Reviews | #91 of 93 things to do in Cesky Krumlov | Museums, Specialty Museums

NamEsti svornosti 1, Cesky Krumlov, Czech Republic

●●○○○ Reviewed April 5, 2018 ▯ via mobile

Wouldn't bother again

"You can avoid if you have other things to do"

Maybe this is not the review , I was funny at myself after out from the museum. I just have to... read more

, 2017

ınd

●●○○○ Reviewed June 13, 2018 ▯ via mobile

Potential wasted

OUTLINE

- Why study microbes*?
- How do you study microbes?
- Directions for microbial ecology
  - Opinions and research

* microbes = microscopic organisms; today's focus = bacteria/archaea

HUGE THANKS

- Folks from whom I pilfered material
  - **Sarah Hird** (UConn), **Christian Mueller** (Simons), **Scott Handley** (Wash U)
- My hardworking & brilliant research group, the Statistical Diversity Lab:
  - Bryan Martin (@BryanDMartin_), Pauline Trinh (@paulinetrinh), Kendrick Li (@KendrickLi4), David Clausen, Alex Paynter, Charlie Wolock, Jake Price (@Jake_in_the_Lab)
- Collaborators whose joint work I discuss
  - **Sam Minot** (Fred Hutch), **Alon Shaiber & M Eren** (U Chicago), **Michael McLaren & Ben Callahan** (NC State)
- The **heroic organizers of #evomics2019** and **Daniel McDonald**

HUGE THANKS

YOU!

- For jumping on the 🦠 🚂
- For participating, contributing, correcting me throughout

WHY STUDY MICROBES?

- Microbial:host cells
    - Microbial:host genes
- Impact ecosystem/host health and function
    - Host associated: nutrient absorption, immune system, healing...
    - Environmental: biogeochemical cycling, origins of life...
- Highly localized communities; gene/organism transfer

FUN FACTS

that make studying the microbiome hard

- Hard to culture most microbes
- Microbes can be categorised into groups
  - Strains; taxa; x% similarity on some/all genes
- Every group has some concentration in every environment
  - possibly zero
- Every individual microbe has many genes
- *Microbes of the same strain may not have the same genes*

MICROBIAL QUESTIONS

- What strains are present?
- What genes are present?
- What microbes have what genes?
- How many microbes are there?
- How many different microbes are there?

MICROBIAL POPULATIONS

- Group exercise: (2 minutes)
  - Come up with a microbiome-related question that ***you might want to answer***
    - Preferably one related to your area of interest

HOW DO YOU STUDY MICROBES?

It depends!

TECHNOLOGY

- The technology/technologies that you will use is driven by
    - The scientific question/questions that you have
    - Cost constraints
    - Resource constraints
    - Literature review, opinion of funding agencies, current trends...

TECHNOLOGY

- The technology/technologies that you will use is driven by
  - **The scientific question/questions that you have**
  - Cost constraints
  - Resource constraints
  - Literature review, opinion of funding agencies, current trends...

NEED TO KNOW TECHNOLOGIES

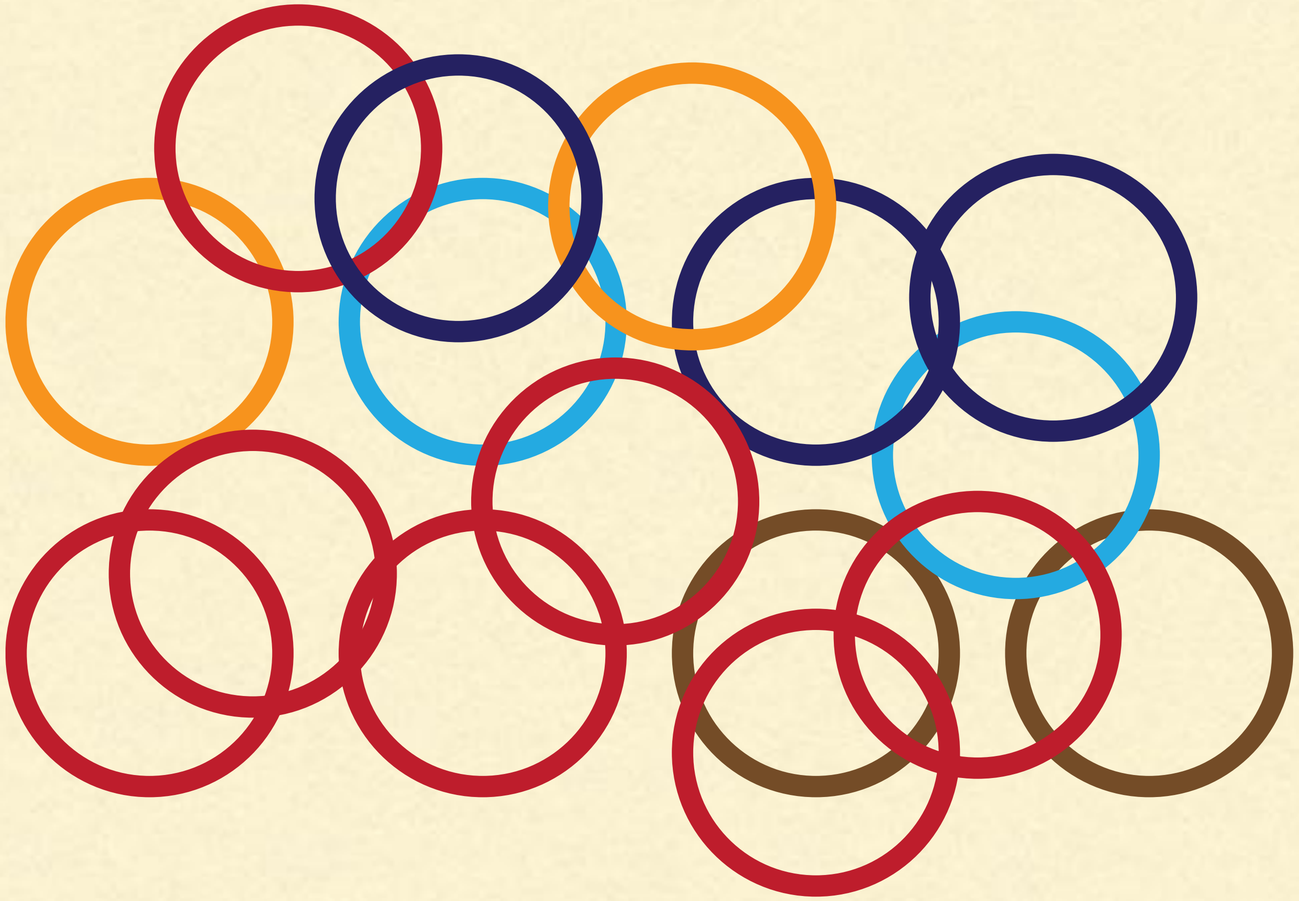
O
THE OPRAH MAGAZINE
OPRAH'S FAVORITE THINGS!
She tasted, she tested, she picked the best—and you have a chance to WIN THEM ALL!

illumina®

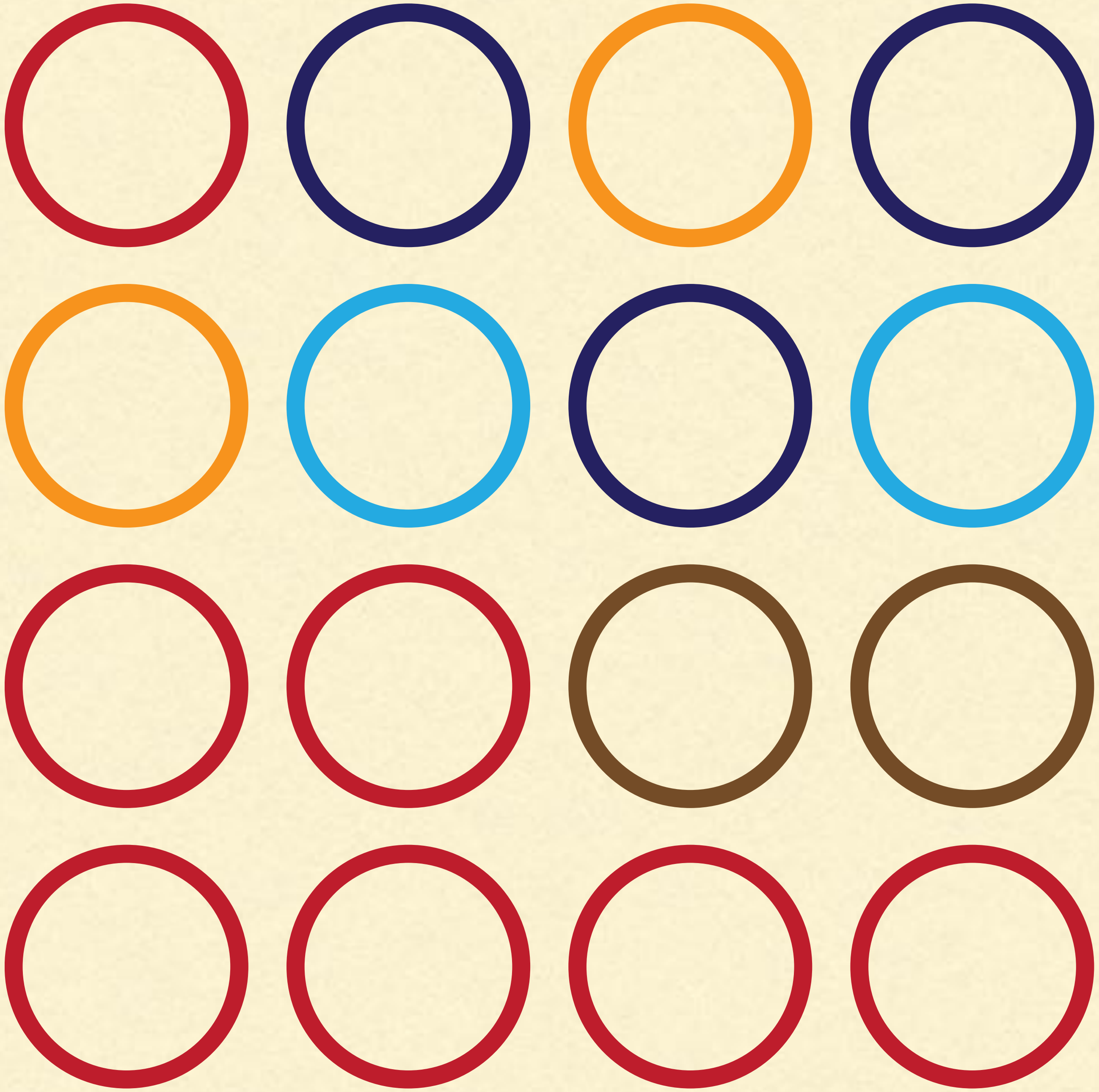
- Amplicon profiling
- Whole genome profiling
- Concentration profiling
- Many others...

AMPLICON PROFILING

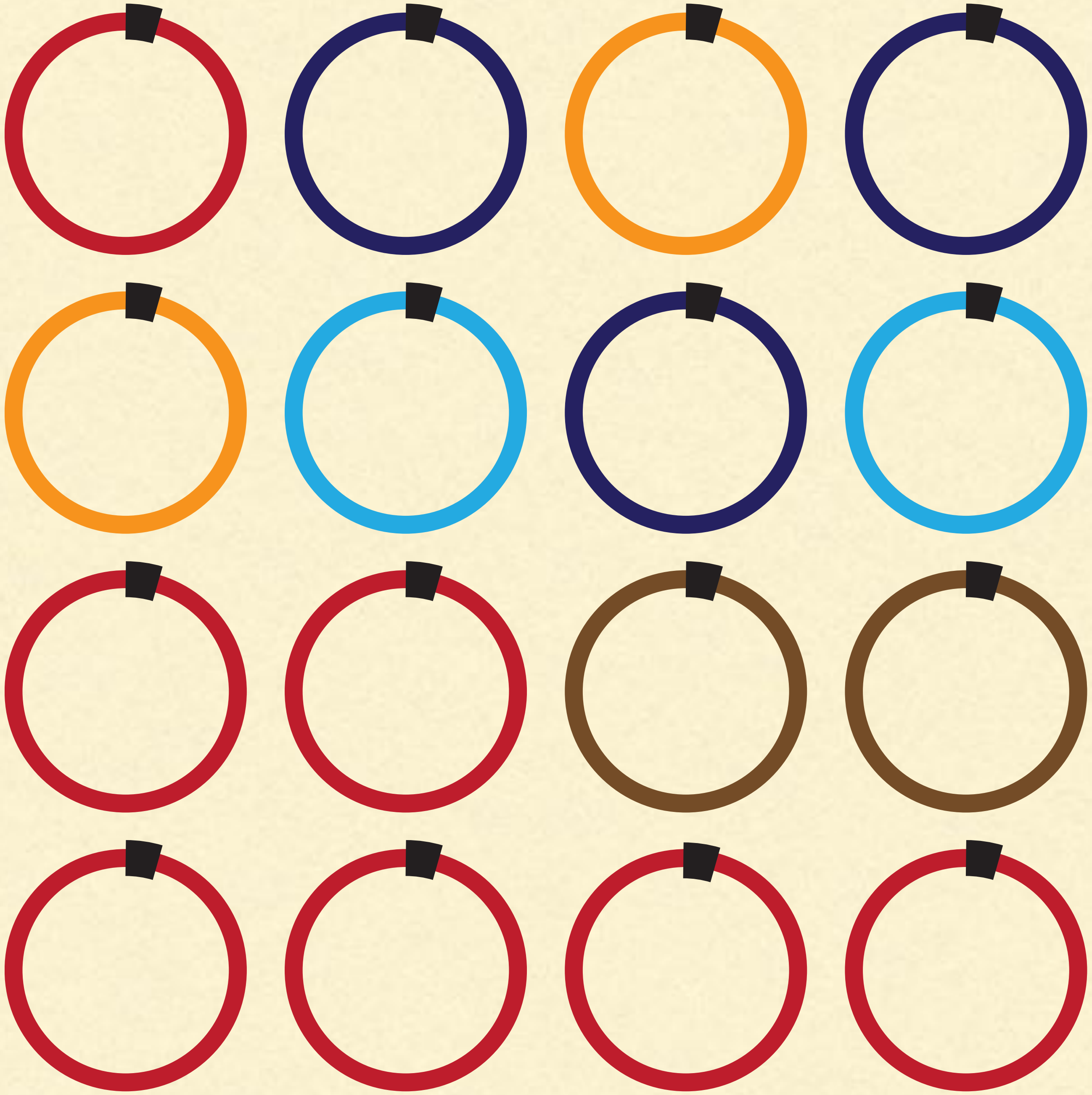
- Amplify (PCR) & sequence a HOMOLOGOUS MARKER (amplicon) shared by all taxa
- e.g.,16S rRNA is bacterial marker gene
- e.g.,18S is marker gene for microbial eukaryotes



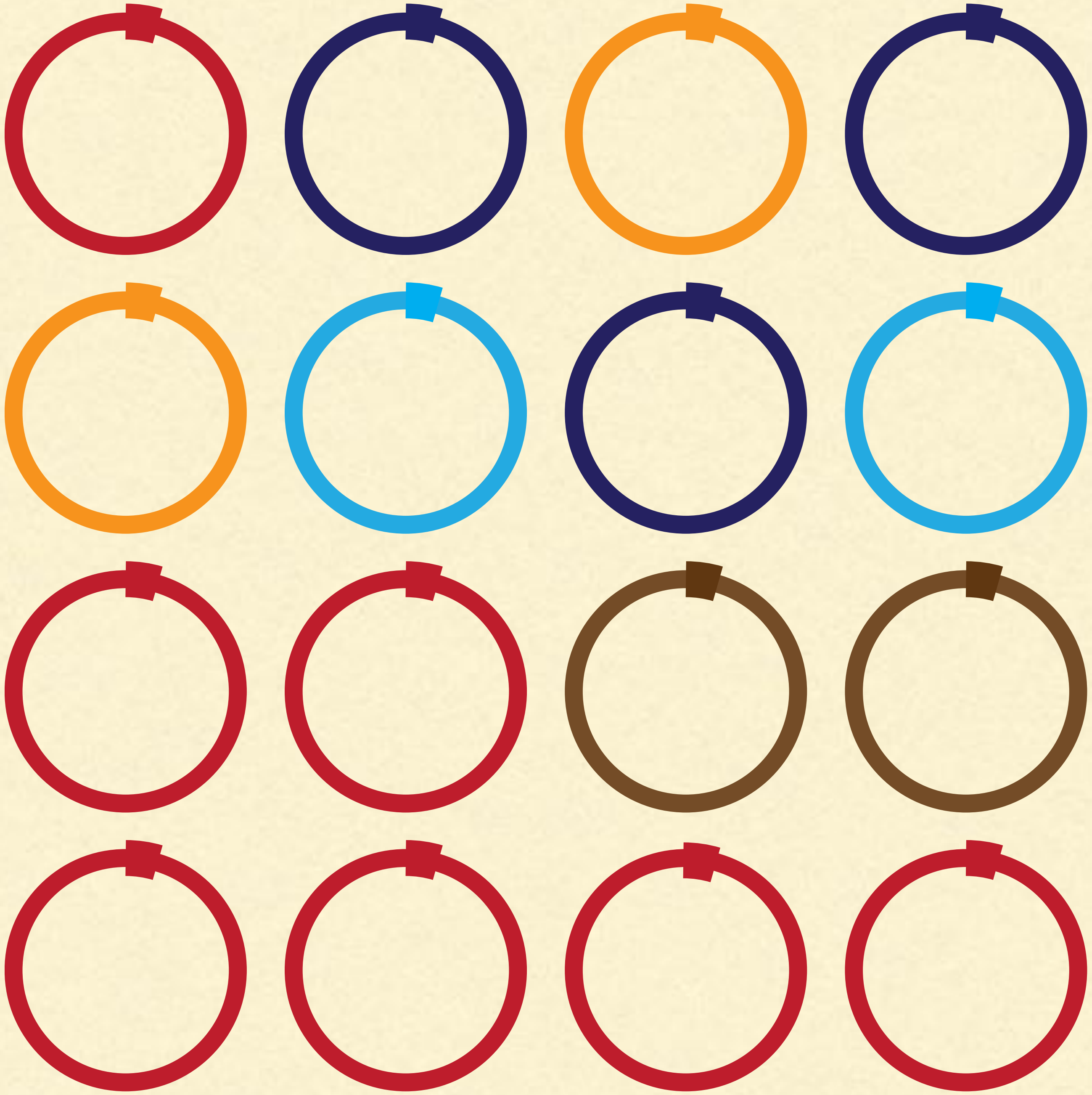
Slide used with permission from Sarah Hird



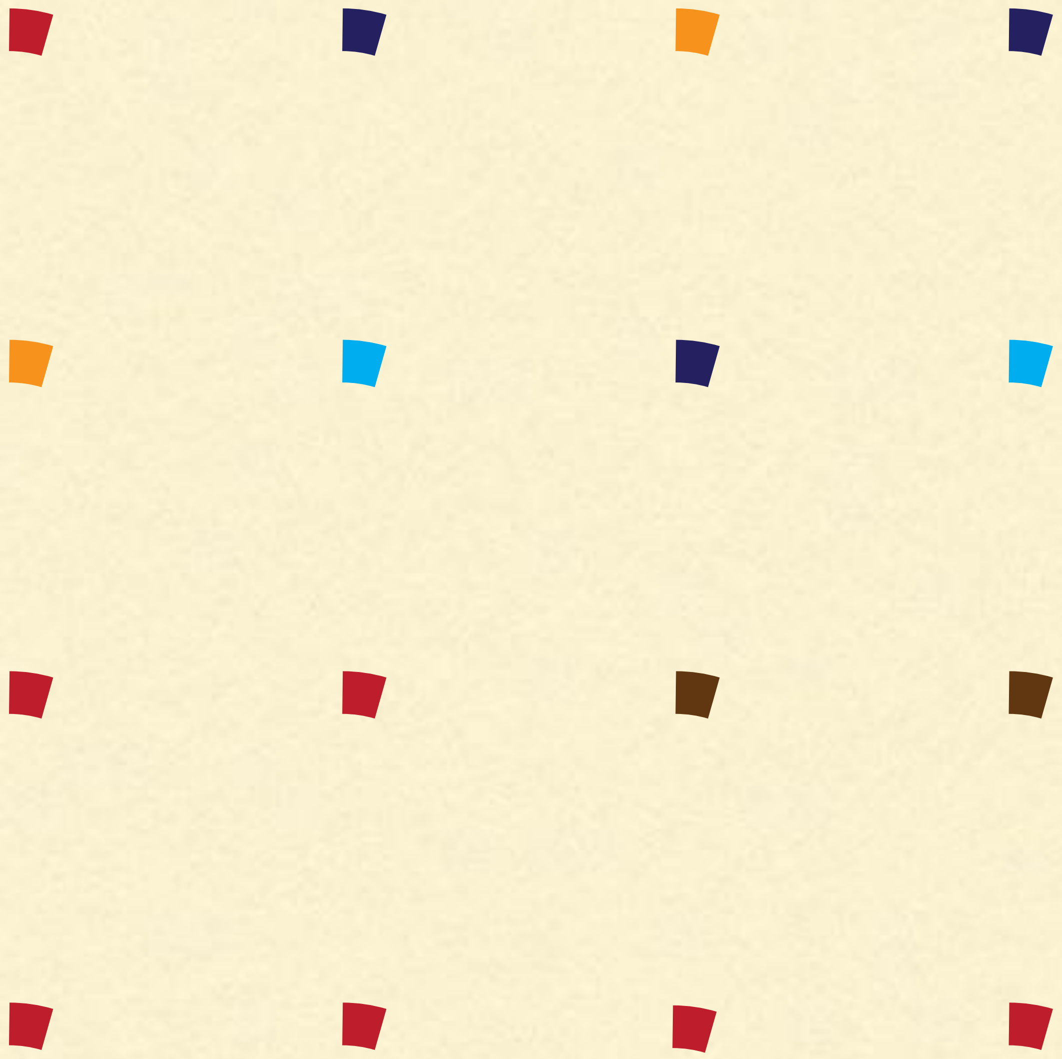
Slide used with permission from Sarah Hird



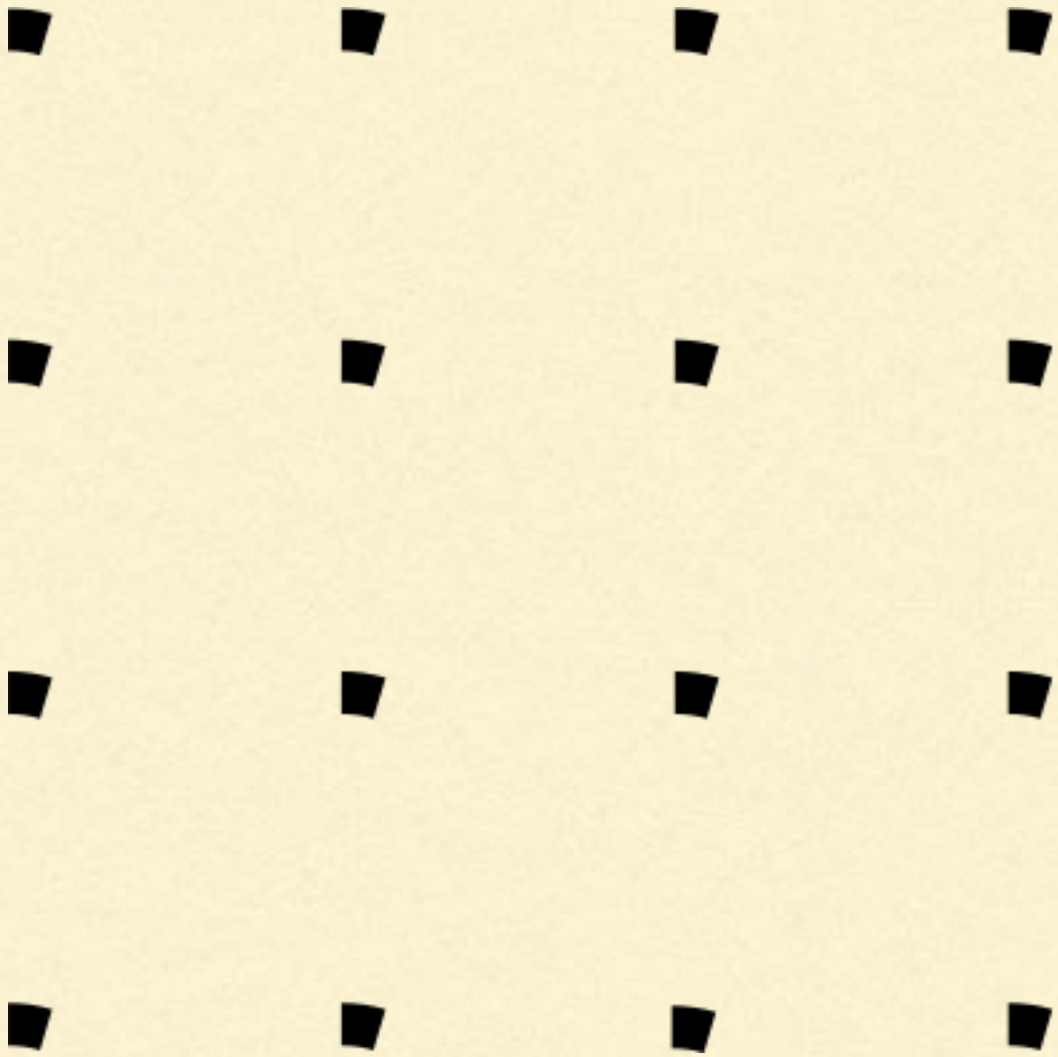
Slide used with permission from Sarah Hird



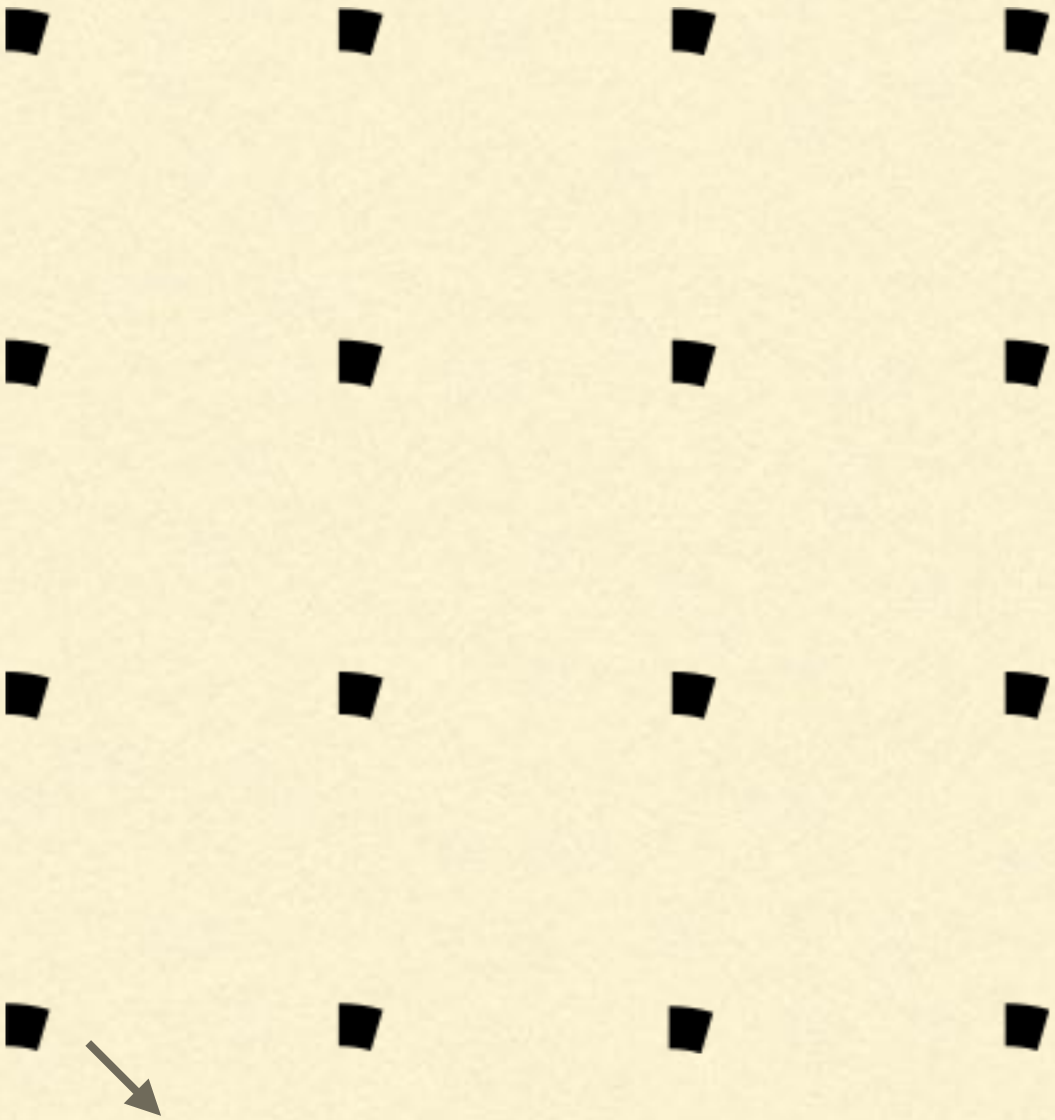
Slide used with permission from Sarah Hird



Slide used with permission from Sarah Hird

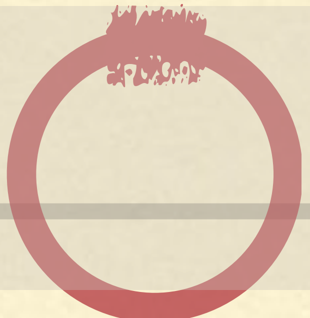


Slide used with permission from Sarah Hird



ACGTGCGTAG…

infer that it [redacted] is from

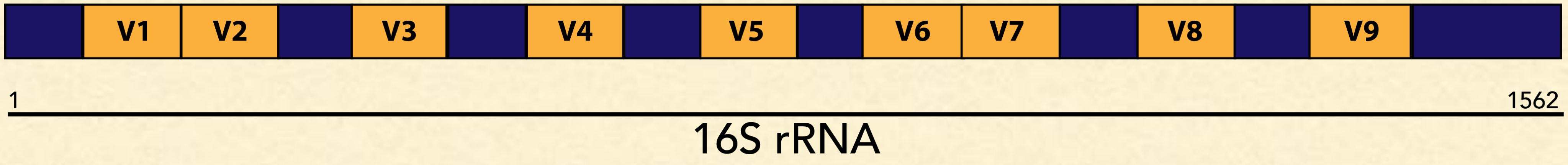


AMPLICON PROFILING

- 16S is a commonly sequenced bacterial marker gene
- Universal: Fancy protein reasons… ask Scott!
- Balance: same in places; different in places
- Not single copy

AMPLICON PROFILING

- 16S is a commonly sequenced bacterial marker gene
- Universal: Fancy protein reasons… ask Scott!
- Balance: same in places; different in places
-

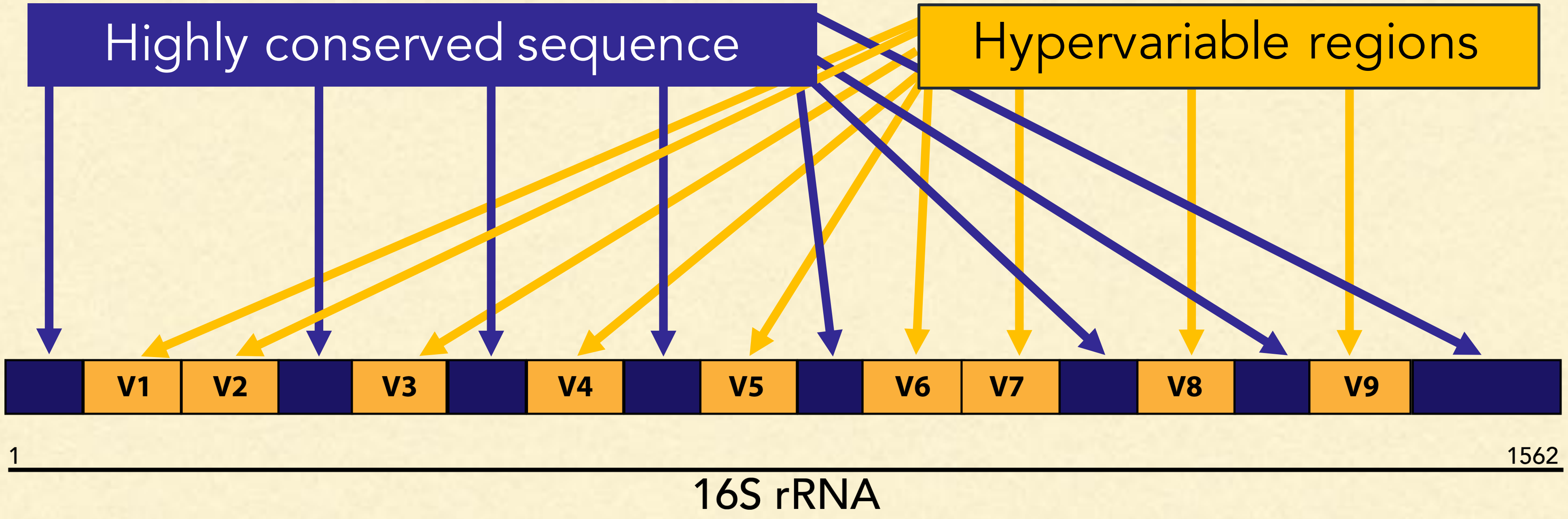


Del Chierico et al. 2015

Slide modified with permission from Sarah Hird

WHY 16S?

- 16S has highly conserved sequences interspacing hypervariable regions

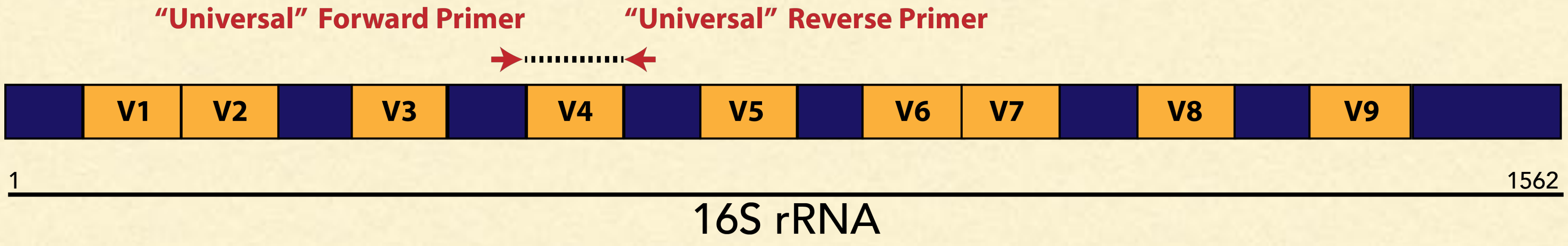


Del Chierico et al. 2015

Slide modified with permission from Sarah Hird

WHY 16S?

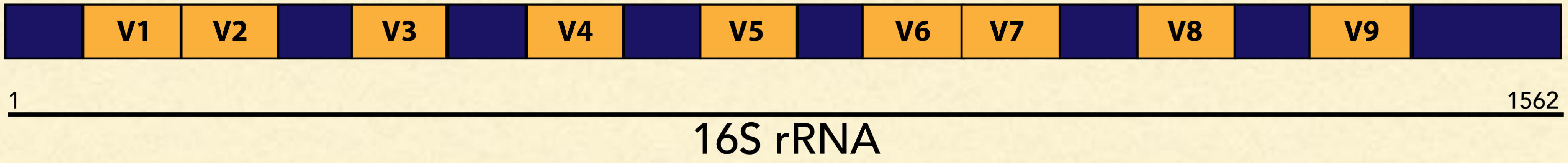
- 16S has highly conserved sequences interspacing hypervariable regions
  - Primers targeting the conserved regions allow us to pull out the hypervariable regions for sequencing



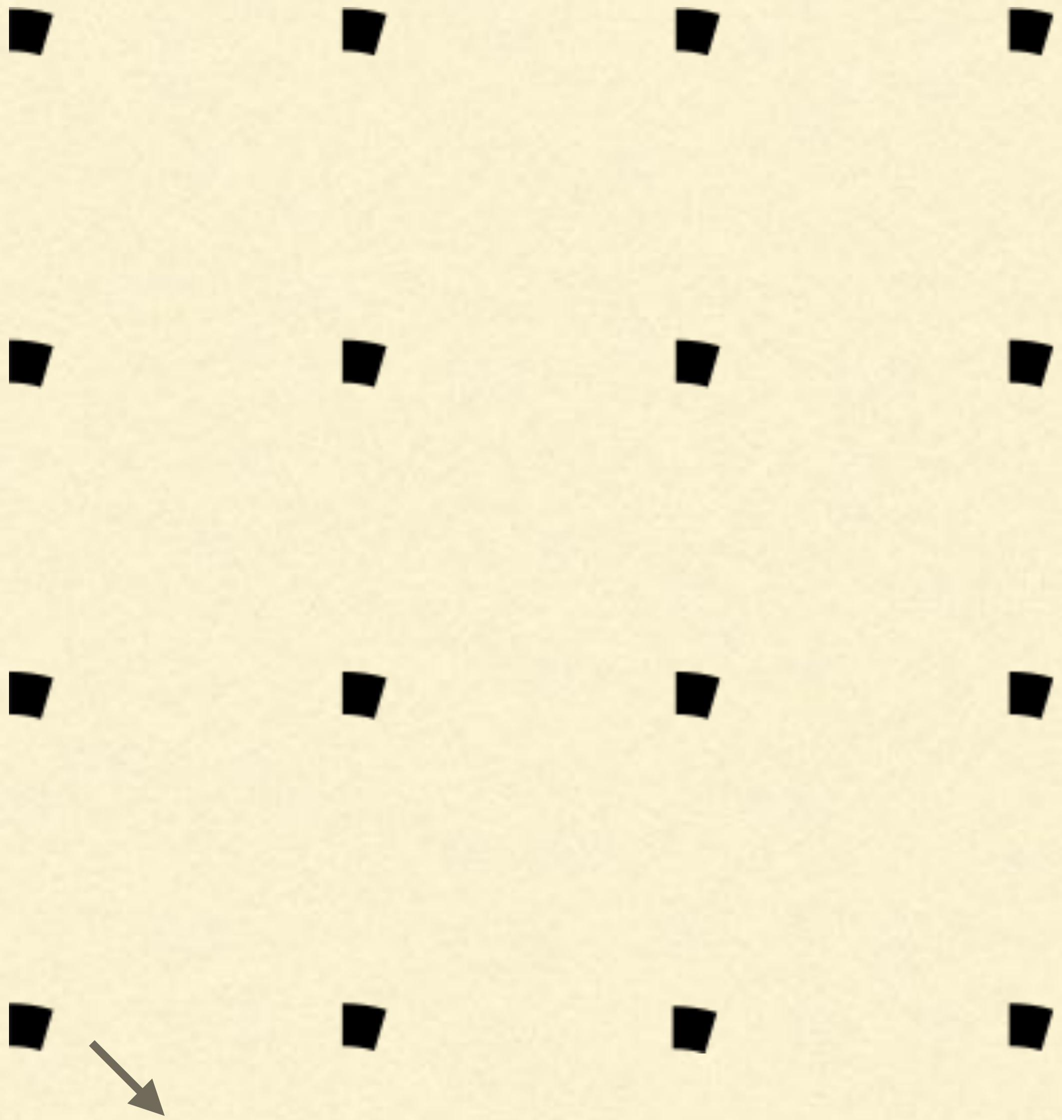
Slide modified with permission from Sarah Hird

WHY 16S?

- 16S has highly conserved sequences interspacing hypervariable regions
  - Primers targeting the conserved regions allow us to pull out the hypervariable regions for sequencing
  - New(ly more common): full length 16S sequencing

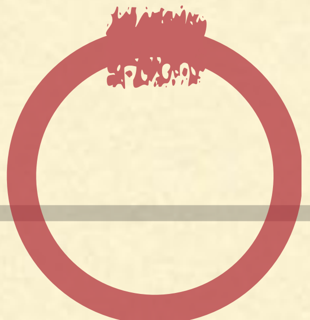


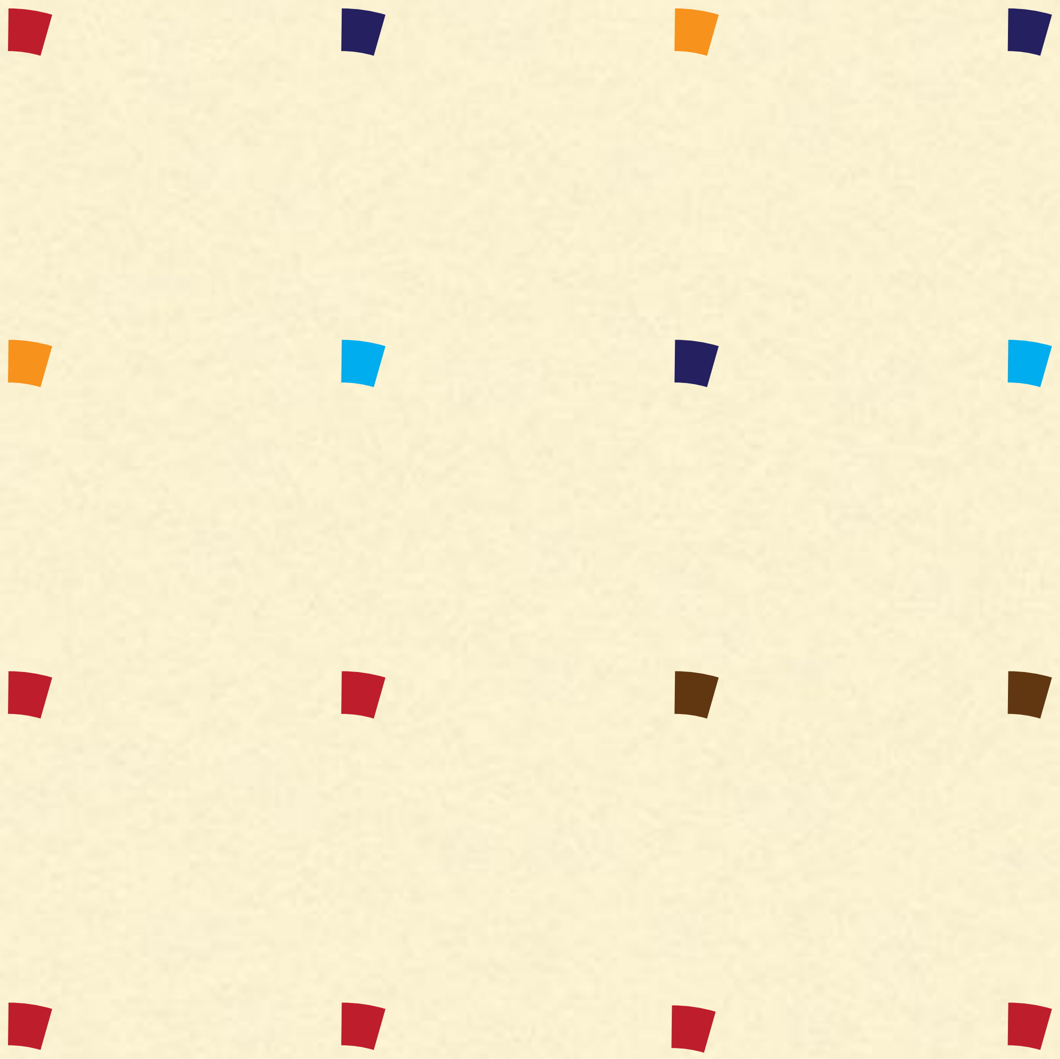
Slide modified with permission from Sarah Hird



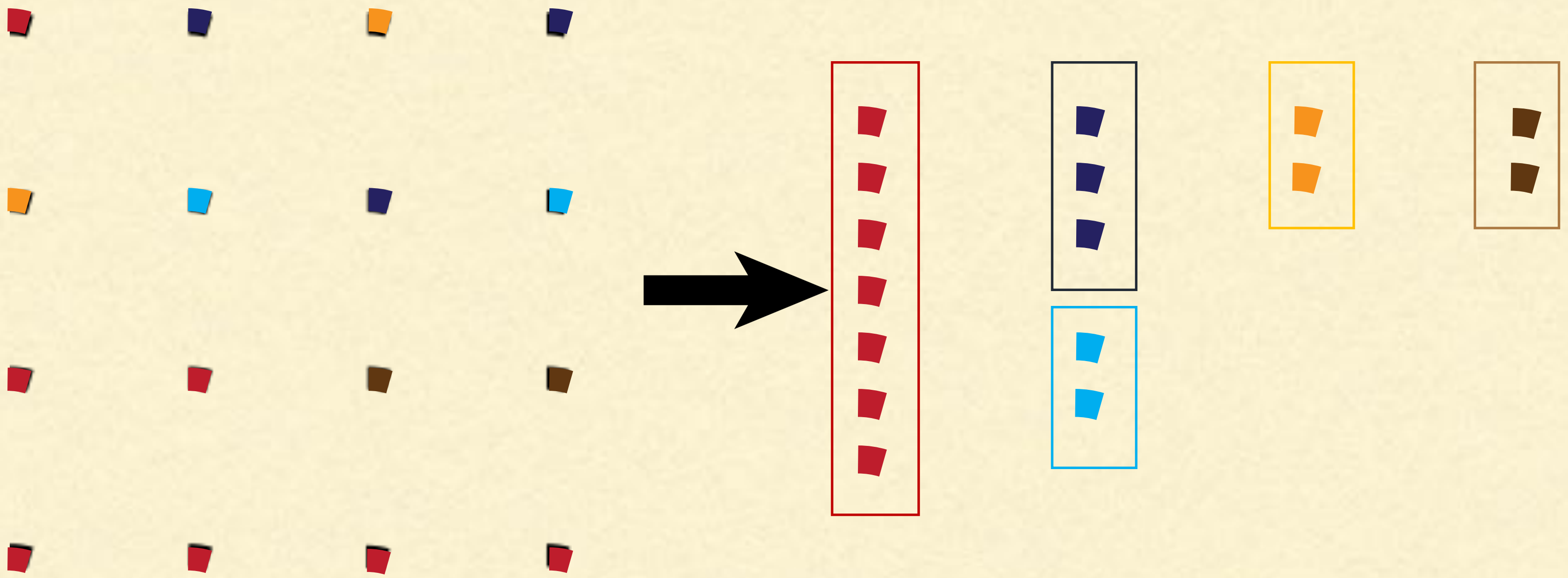
ACGTGCGTAG…

infer that it is from



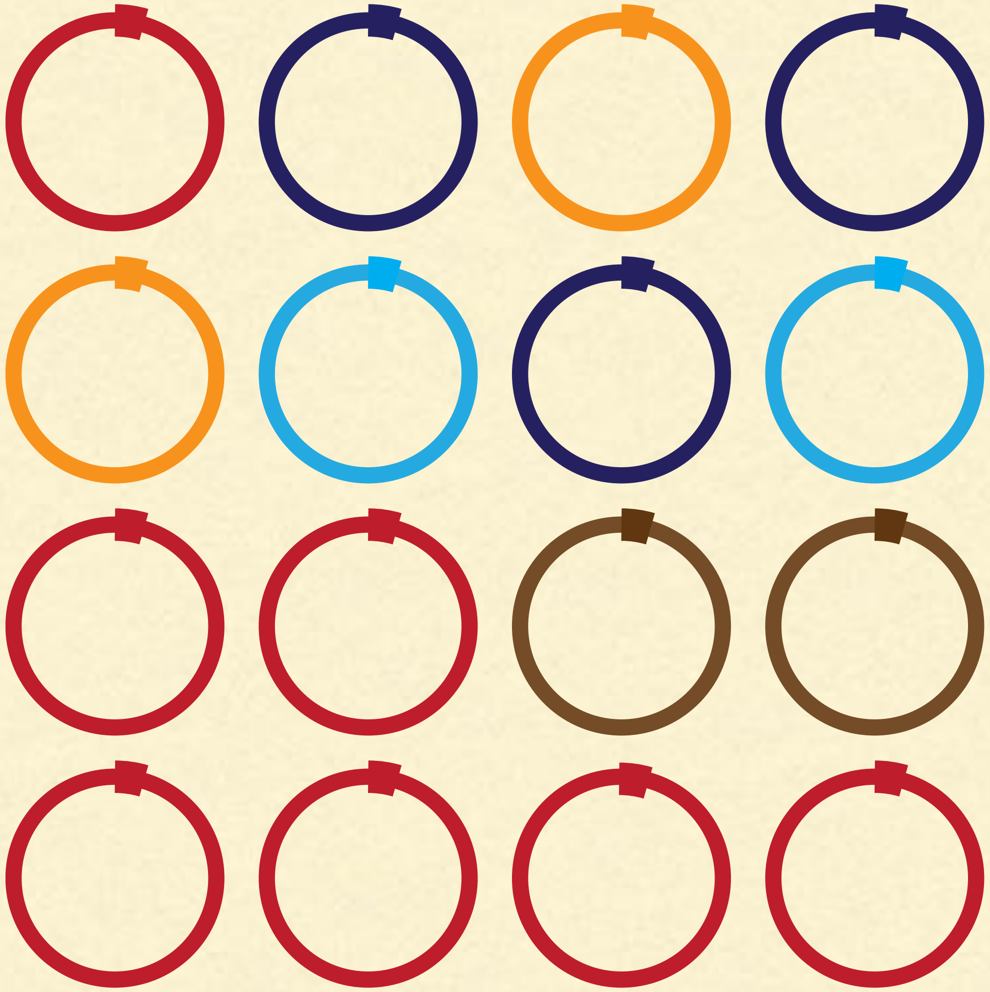


Slide used with permission from Sarah Hird



Slide modified with permission from Sarah Hird

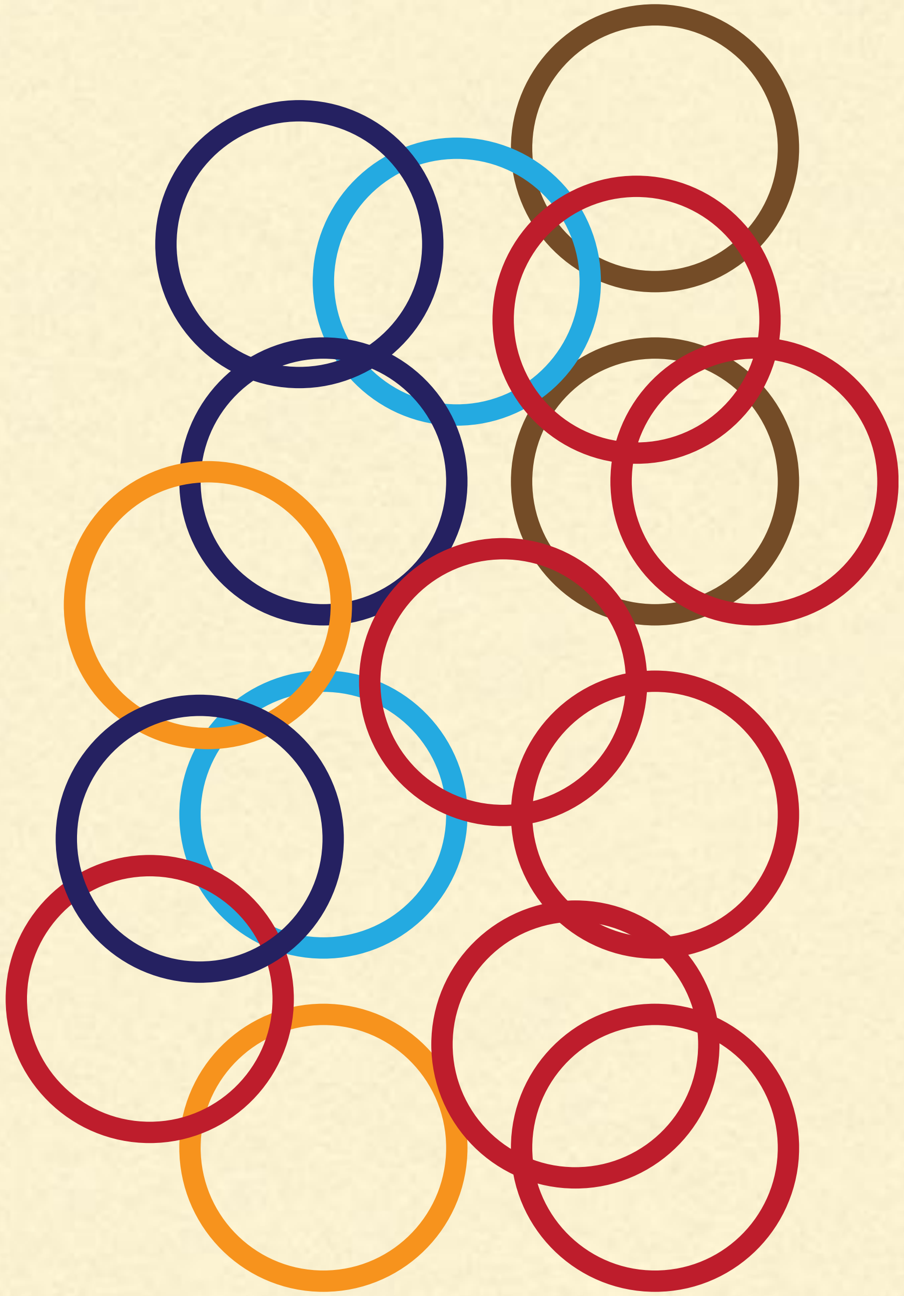
AMPLICON PROFILING

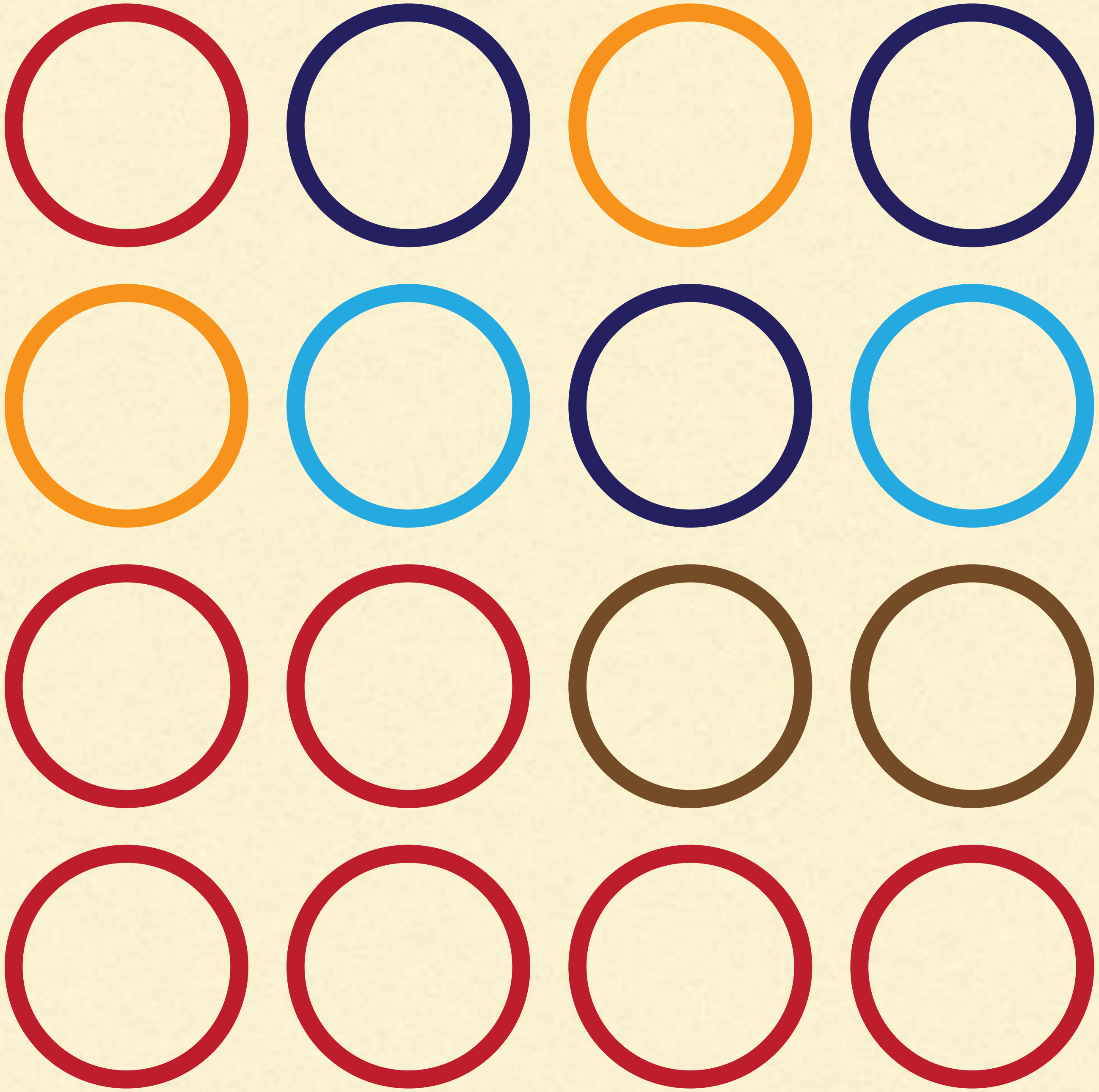


- Cheap, easy, popular… historical reasons
- Most (but not all) taxa amenable
- Severe distortions (PCR, primers, index hopping)
  - Discussed later

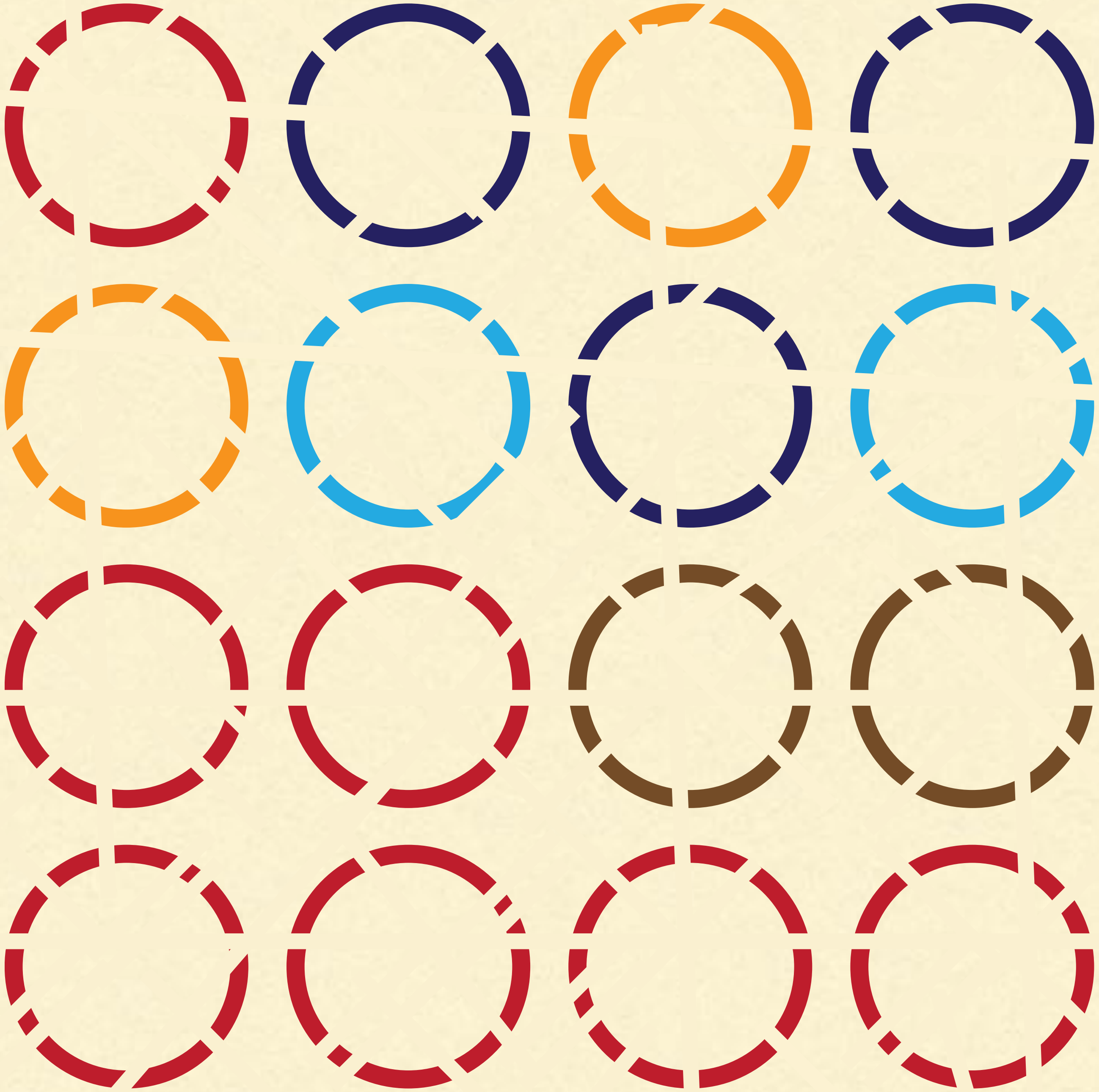
WHOLE GENOME PROFILING

- Whole genome sequencing (WGS)
  - Shear all DNA and sequence fragments
  - *Functional potential*
- Commonly called "metagenomics"
  - metagenome = all the genomes





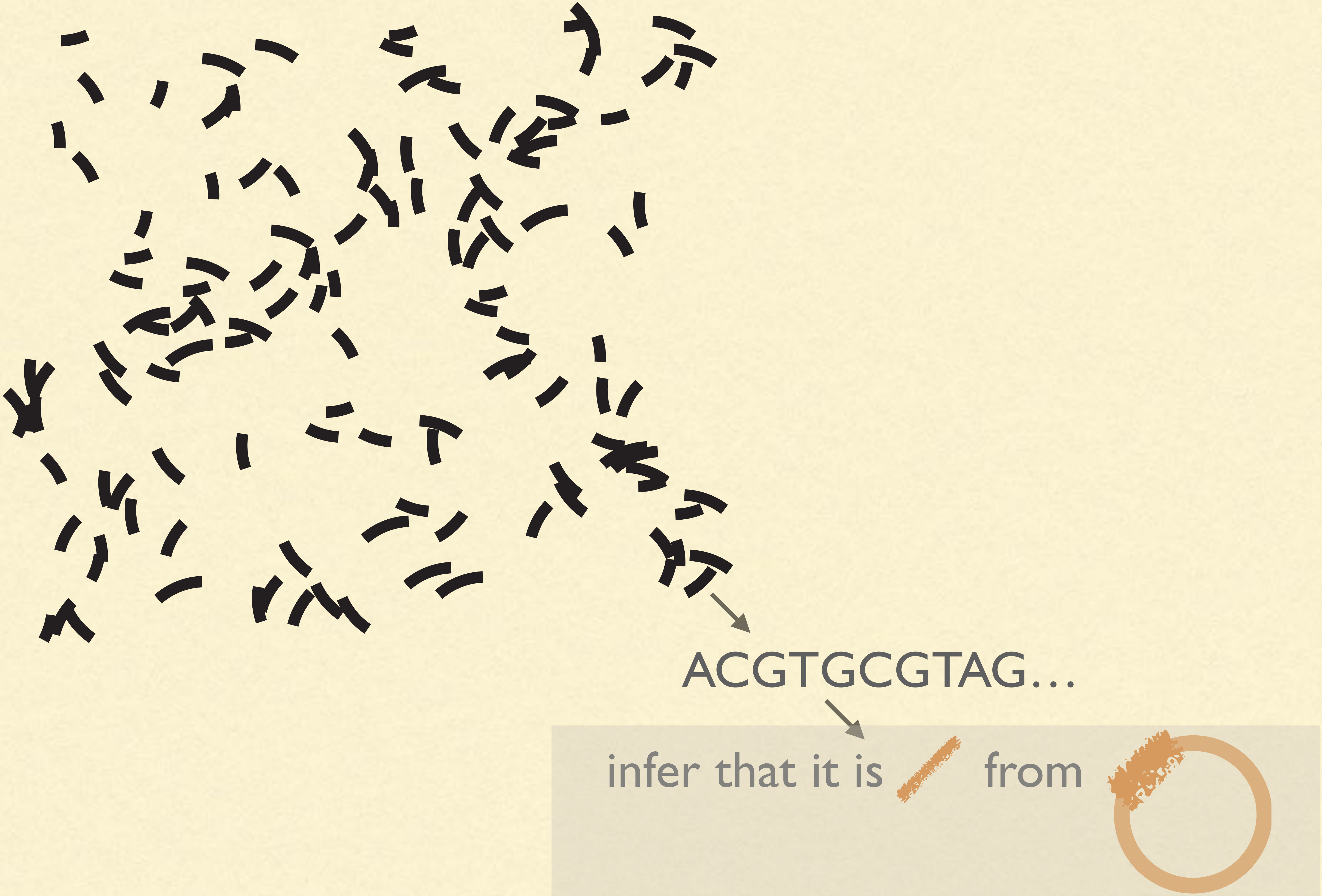
Slide modified with permission from Sarah Hird



Slide modified with permission from Sarah Hird

Slide modified with permission from Sarah Hird

Slide modified with permission from Sarah Hird



ACGTGCGTAG…

infer that it is from

Slide modified with permission from Sarah Hird

WHOLE GENOME PROFILING

- Multilocus
- Gene content! Not just markers
- More expensive (getting cheaper)
- Sequence non-microbial genes
- Widely thought to be less distortion

CONCENTRATION PROFILING

CONCENTRATION PROFILING

- Just believe me that there are more bacteria in some places than others, ok?

CONCENTRATION PROFILING

- Develop primers to target region
  - Region determines *what* concentration
- Amplify (qPCR) and count (calibrate) to see how many instances of that region there are

TECHNOLOGY

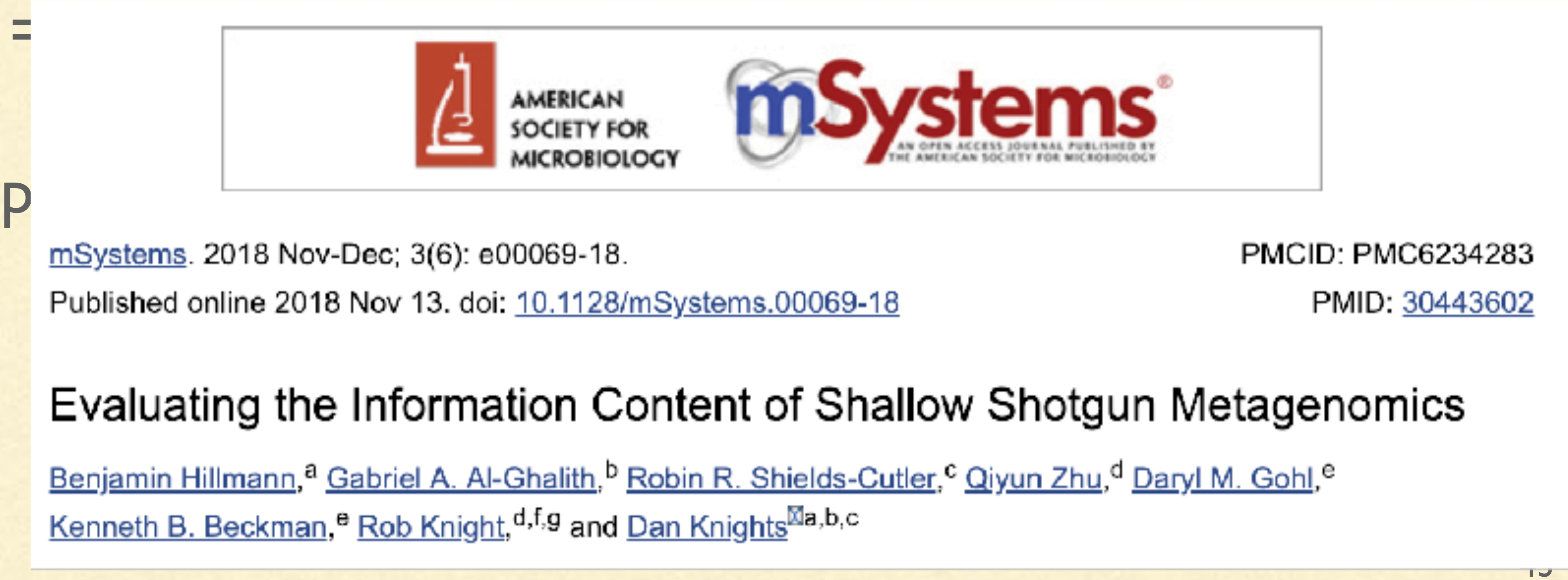
- The technology/technologies that you will use is driven by
  - The scientific question/questions that you have
  - **Cost constraints**
  - **Resource constraints**
  - Literature review, opinion of funding agencies, current trends...

$$ COMPARISON

- Costs *can vary wildly*… here are some recent ballparks:
    - 16S = $17/sample
    - WGS = $100-200 per sample
- 250 samples: 16S = $5k, WGS = $25k-$50k

$$ COMPARISON

- Costs *can vary wildly*… here are some recent ballparks:
  - 16S = $17/sample
  - WGS =
- 250 samp



AMERICAN SOCIETY FOR MICROBIOLOGY

mSystems

AN OPEN ACCESS JOURNAL PUBLISHED BY THE AMERICAN SOCIETY FOR MICROBIOLOGY

mSystems. 2018 Nov-Dec; 3(6): e00069-18.
Published online 2018 Nov 13. doi: 10.1128/mSystems.00069-18

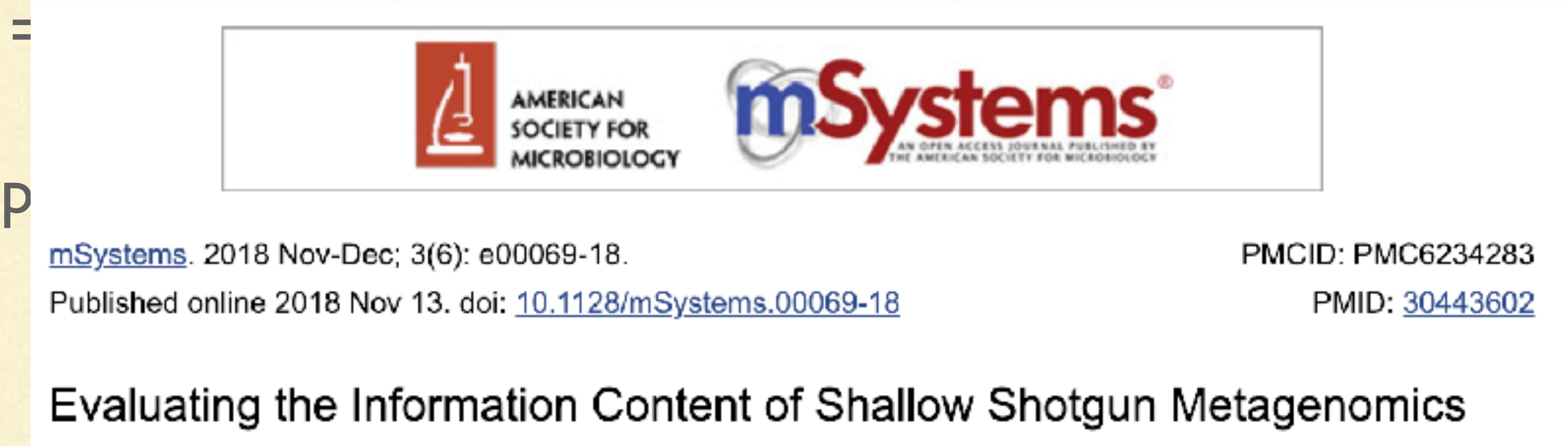
PMCID: PMC6234283
PMID: 30443602

Evaluating the Information Content of Shallow Shotgun Metagenomics

Benjamin Hillmann,[a] Gabriel A. Al-Ghalith,[b] Robin R. Shields-Cutler,[c] Qiyun Zhu,[d] Daryl M. Gohl,[e] Kenneth B. Beckman,[e] Rob Knight,[d,f,g] and Dan Knights[a,b,c]

$$ COMPARISON

- Costs *can vary wildly*… here are some recent ballparks:
  - 16S = $17/sample
  - WGS =
- 250 samp



AMERICAN SOCIETY FOR MICROBIOLOGY

mSystems

AN OPEN ACCESS JOURNAL PUBLISHED BY THE AMERICAN SOCIETY FOR MICROBIOLOGY

mSystems. 2018 Nov-Dec; 3(6): e00069-18. PMCID: PMC6234283

Published online 2018 Nov 13. doi: 10.1128/mSystems.00069-18 PMID: 30443602

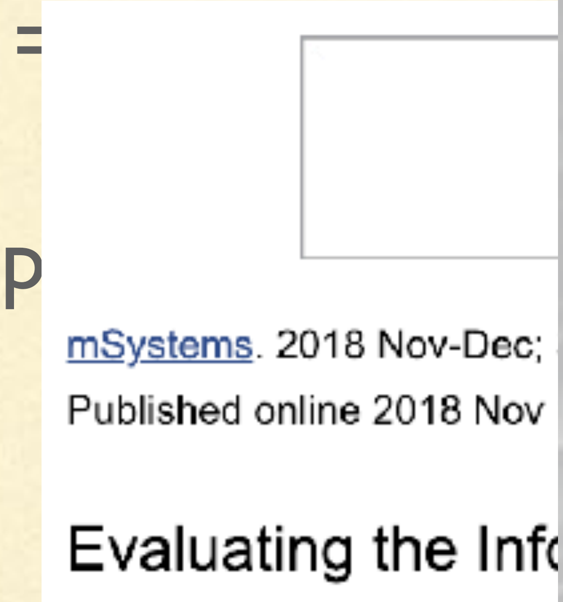
Evaluating the Information Content of Shallow Shotgun Metagenomics

- Other considerations: non-microbial contamination, storage, analysis…

Joke stolen without permission from Scott Handley

$$ COMPARISON

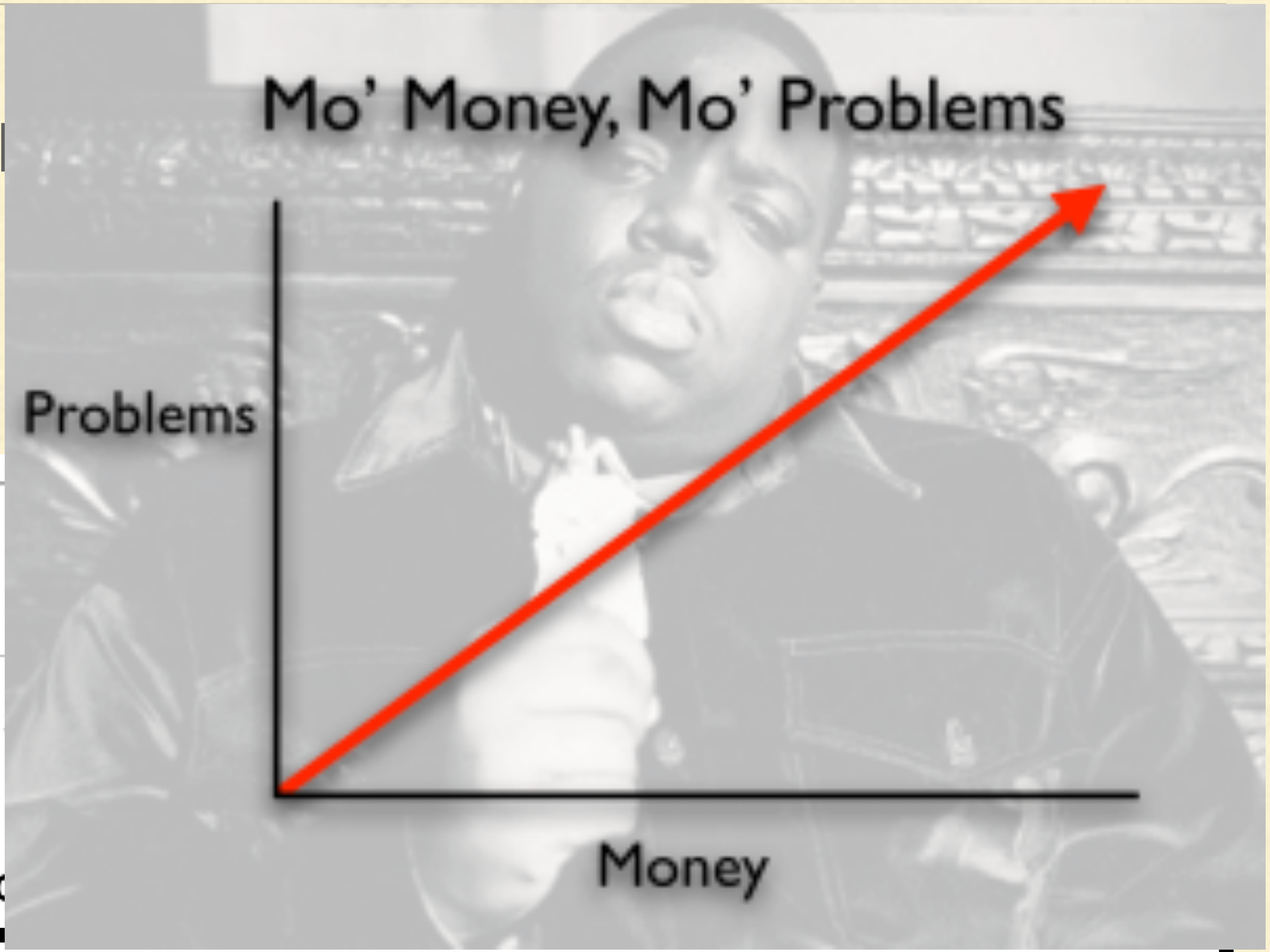
- Costs *can vary wildly…*
  - 16S = $17/sample
  - WGS =
- 250 samp



mSystems. 2018 Nov-Dec;

Published online 2018 Nov

Evaluating the Inf



- Other considerations: non-microbial contamination, storage, analysis…

MICROBIAL POPULATIONS

- Group exercise: (2 minutes)
  - ~~Come up with a microbiome-related question that ***you might want to answer***~~
  - Come up with a microbiome-related question that ***you could answer***
    - *How does sequencing technology influence what you can study?*

ONCE YOU HAVE YOUR DATA...

ONCE YOU HAVE YOUR DATA...

ONCE YOU HAVE YOUR DATA...

- preprocessing
- cleaning
- (iterating)
- analysis

CLEANING AND PREPROCESSING

- Preprocessing
  - sometimes the same as cleaning
  - more often: processing the data into biological units
- Cleaning
  - Basic checks: determine whether sequencing went entirely/a little/~~not at all~~ wrong

BIOLOGICAL UNITS

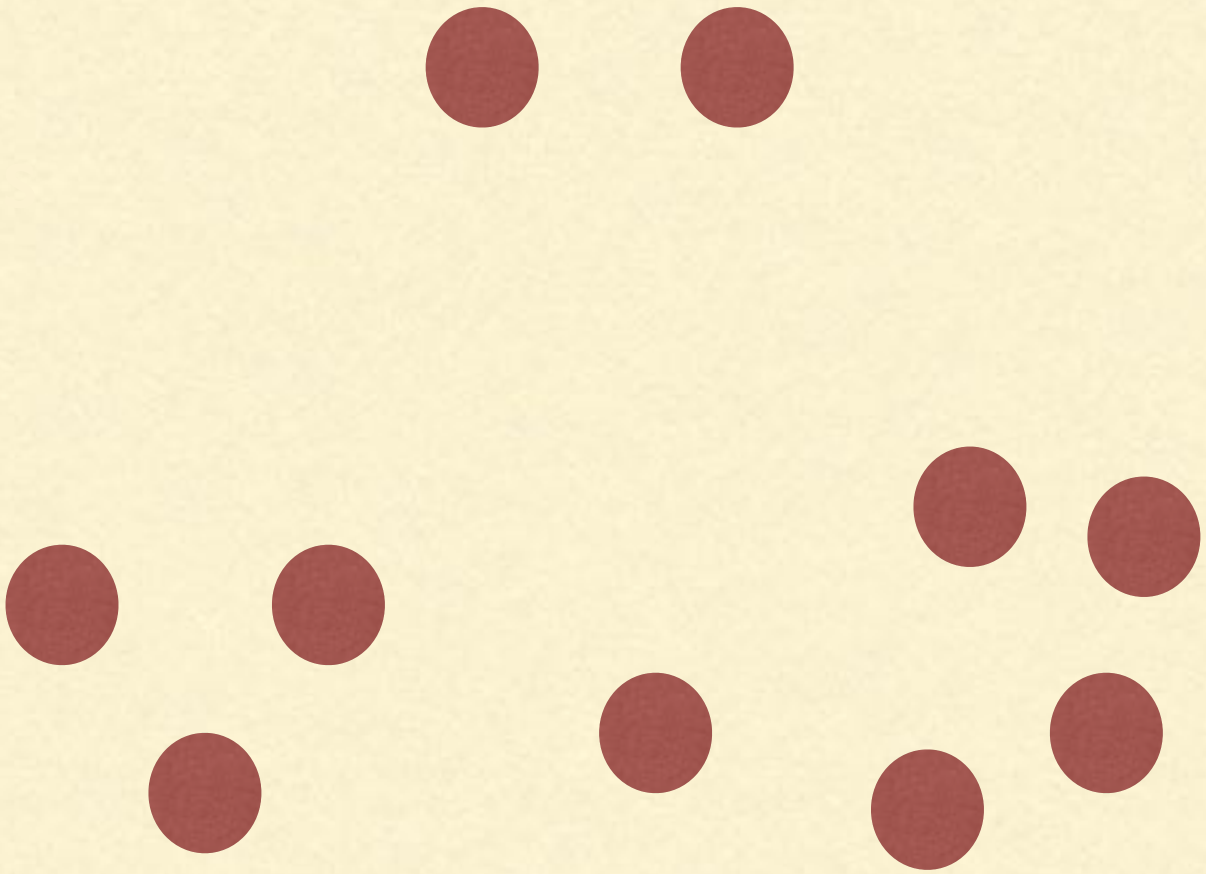
- The units that come off your sequence are usually not immediately useful:
    - … AAACTCTATCTATCTACTXTCGCGCGTACGCGTCAT…
    - …AAACTCTAGCTATCTACTTTCGCXGGTACGCCTCAT…
    - …AACCCCTCGCACGACCAGCACAACACAACTACCA…
    - …AACTCCGTAAAACTACAACTACTACTACCATACACG…
- Idea: group data into units that *simplify analysis* and are *biologically meaningful*

BIOLOGICAL UNITS: TAXONOMIC PROFILING

- If two sequences are the same, should be grouped together
  - Very unlikely that two sequences are the same
- If two sequences are the *same enough*, should be grouped together
  - Idea: clustering!

BIOLOGICAL UNITS: 16S

● = observed 16S sequence

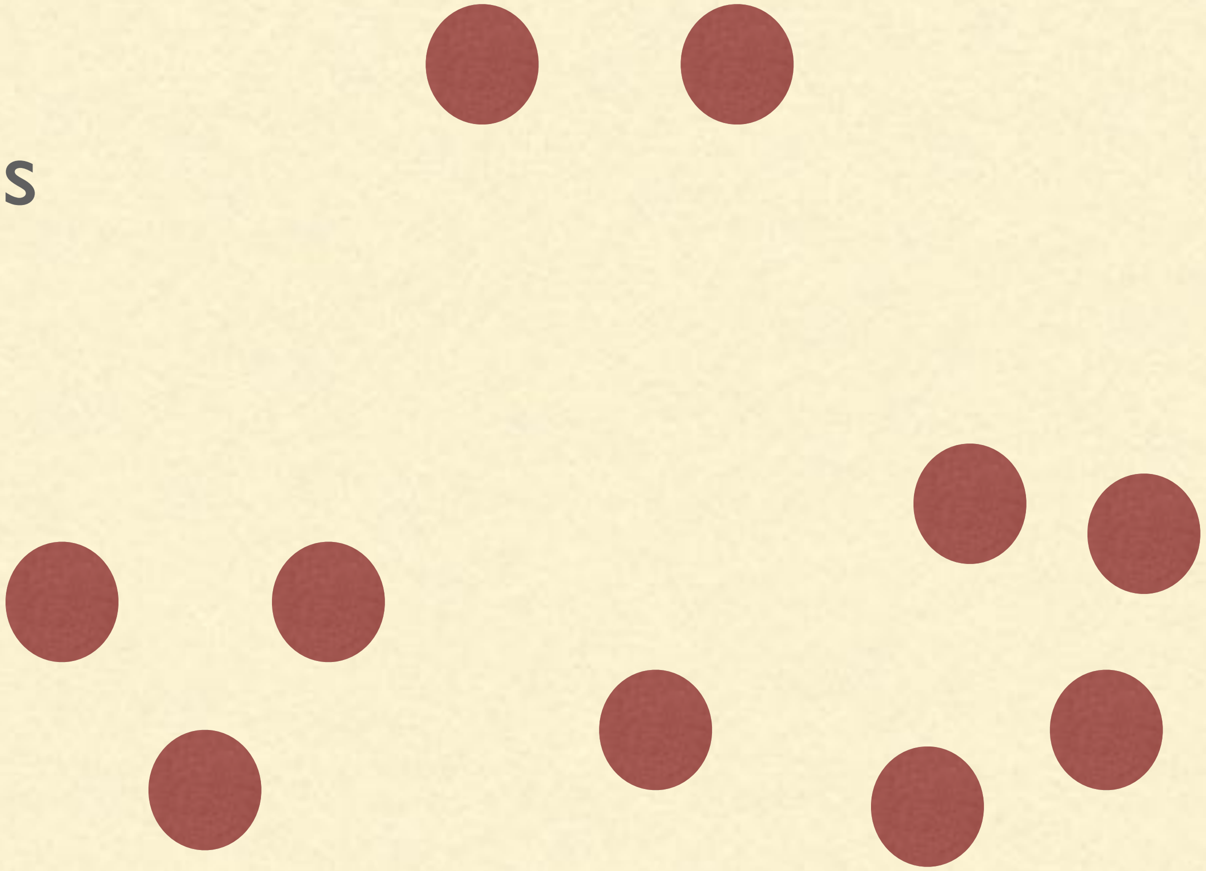


Slide modified with permission from Scott Handley

● = observed 16S sequence

●● = similar sequences

● ● = less similar sequences

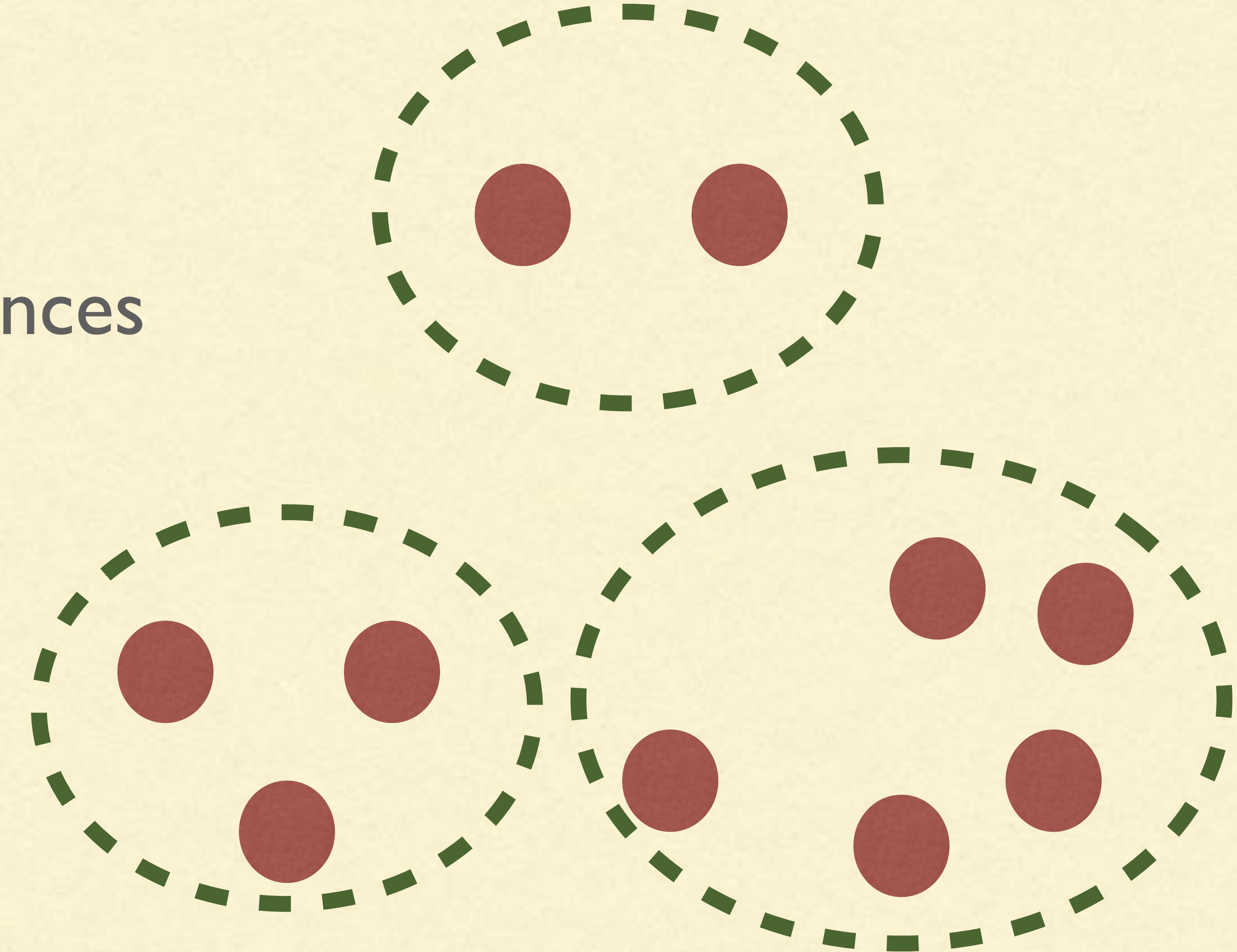


Slide modified with permission from Scott Handley

● = observed 16S sequence

●● = similar sequences

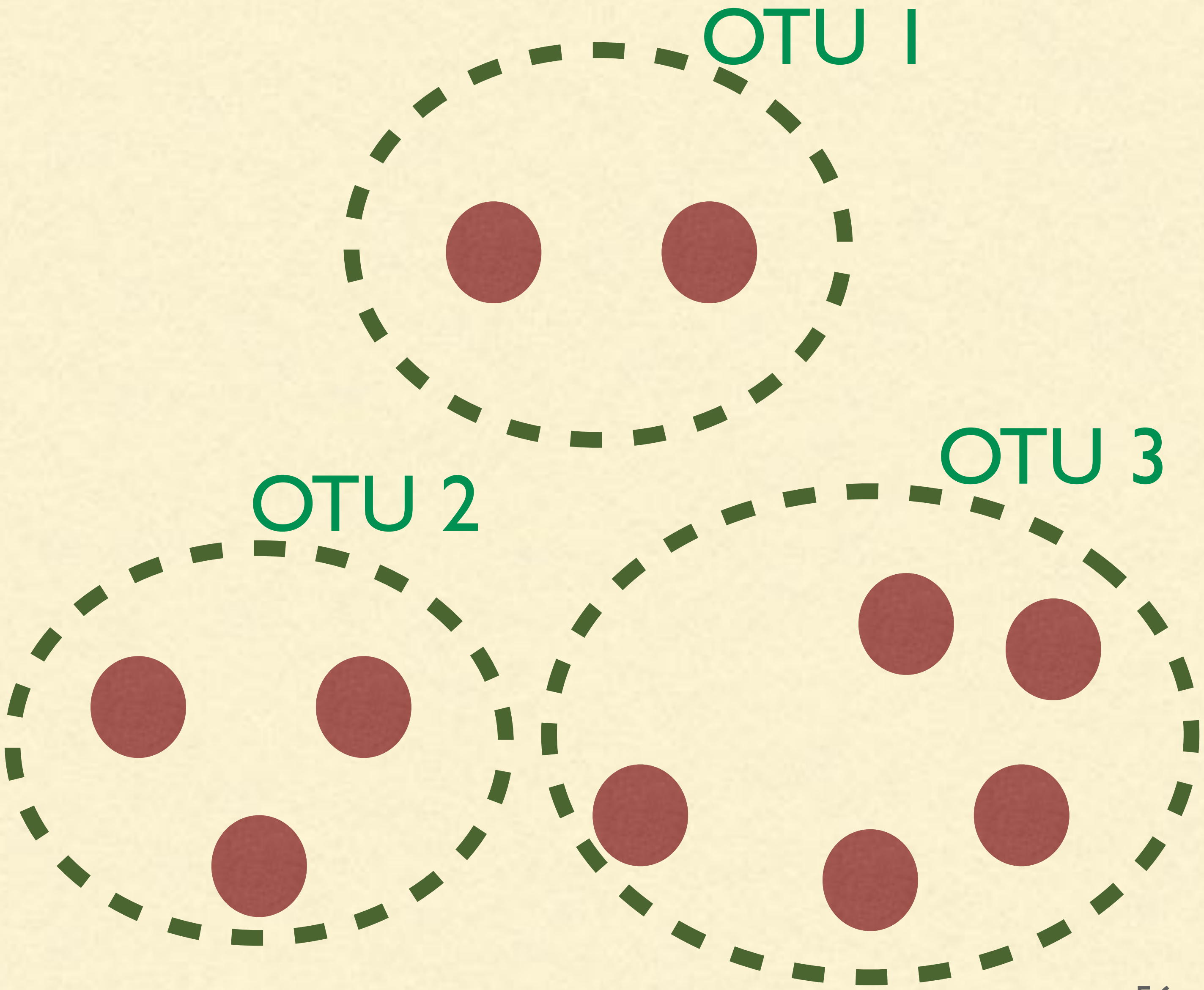
● ● = less similar sequences



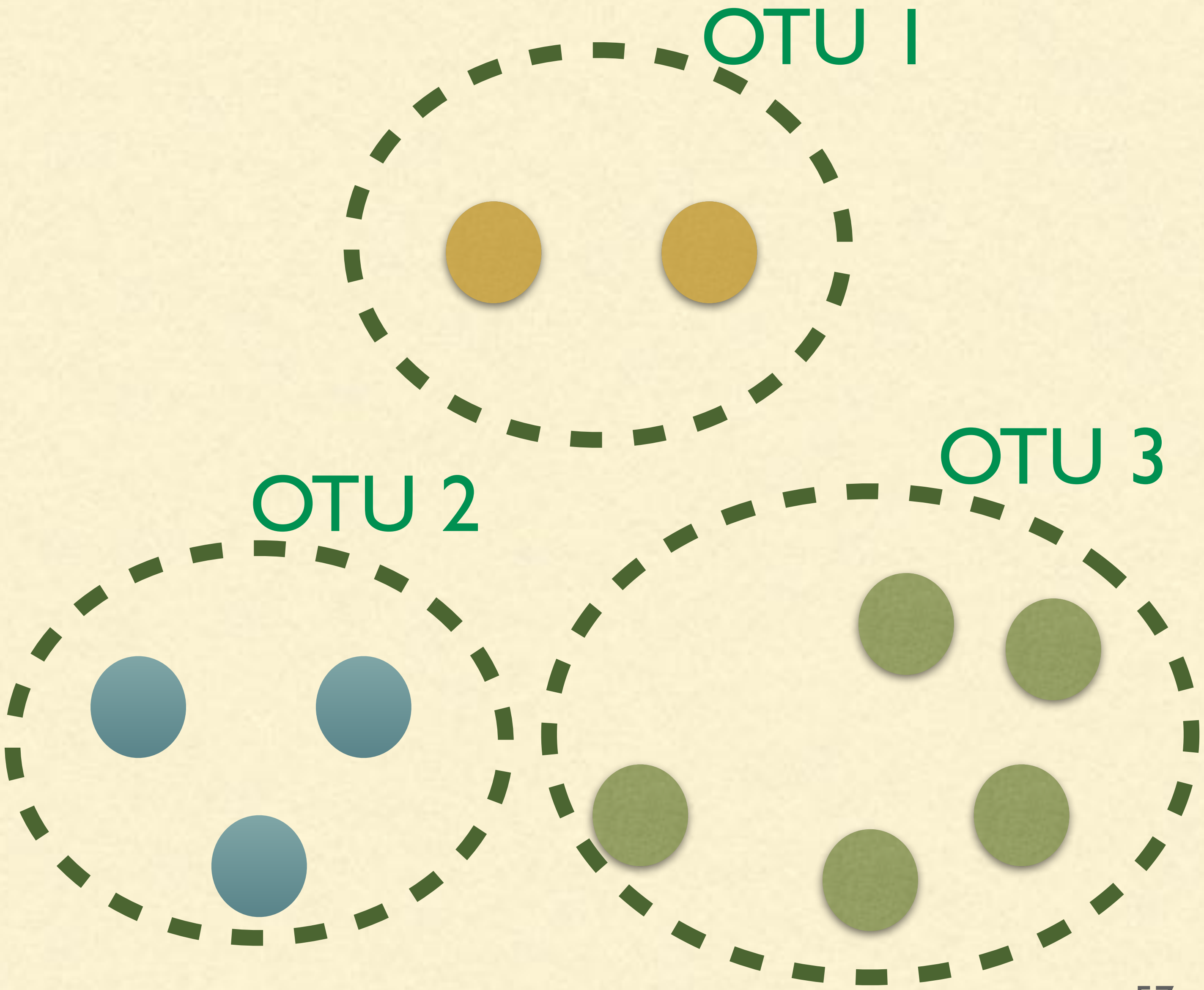
Slide modified with permission from Scott Handley

OTU clustering: Make *operational taxonomic units* by clustering at x% similarity

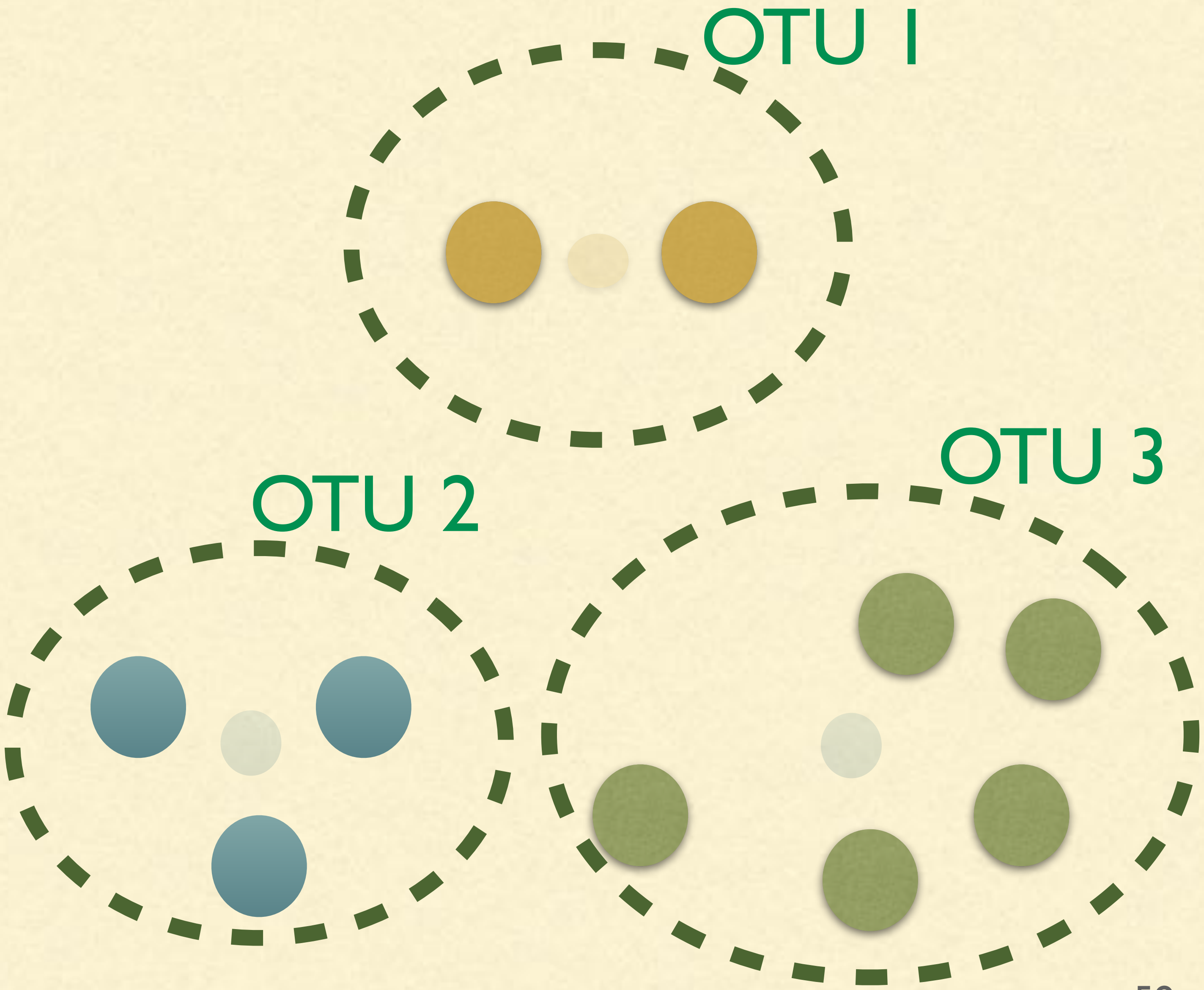
- 97% is common for 16S, but a little arbitrary



Slide modified with permission from Scott Handley



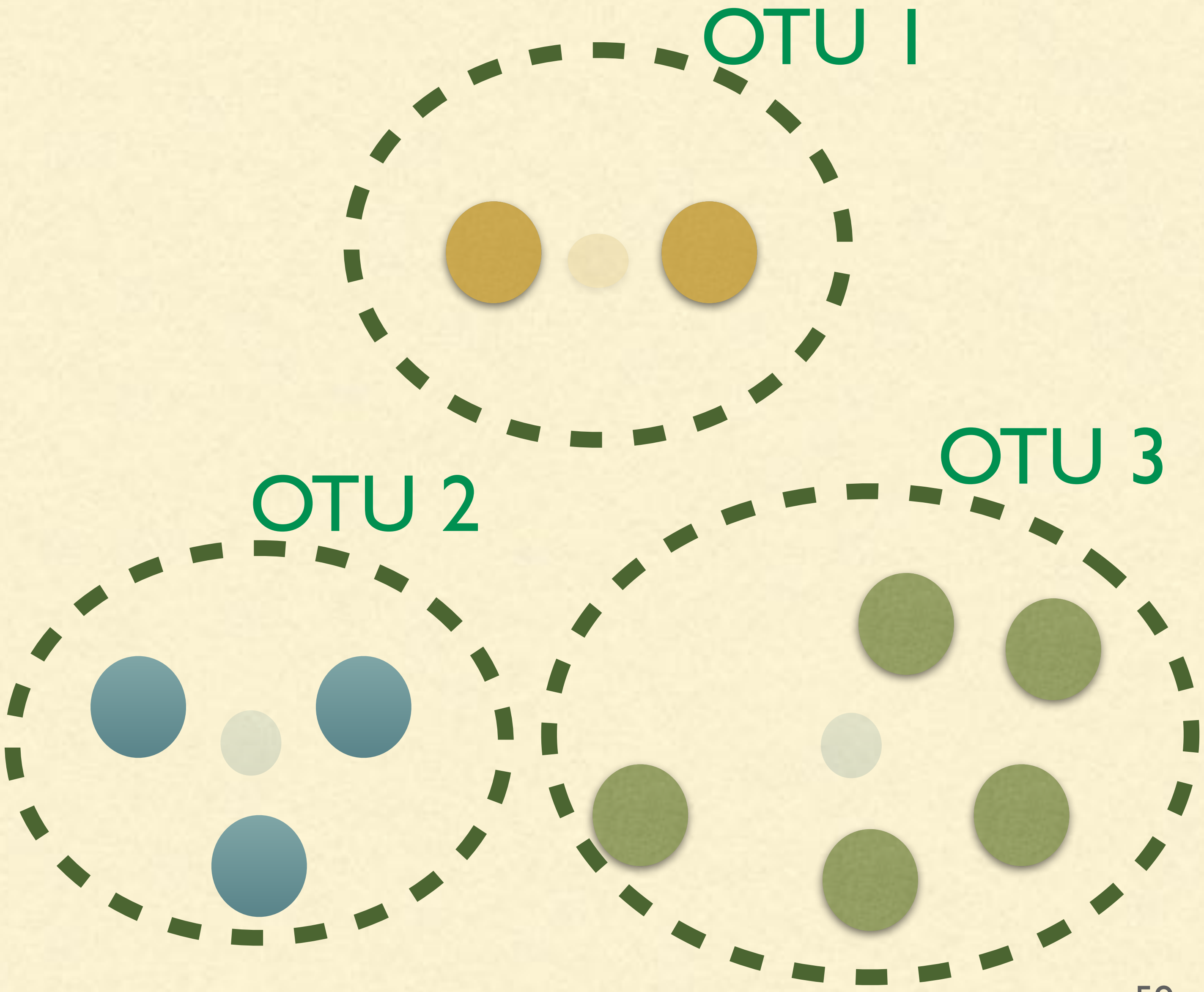
Slide modified with permission from Scott Handley


OTU 1
OTU 2
OTU 3

BIOLOGICAL UNITS: OTUs

OTU clustering: Make *operational taxonomic units* by clustering at x% similarity

- Assign OTU the taxonomy of "most central" sequence

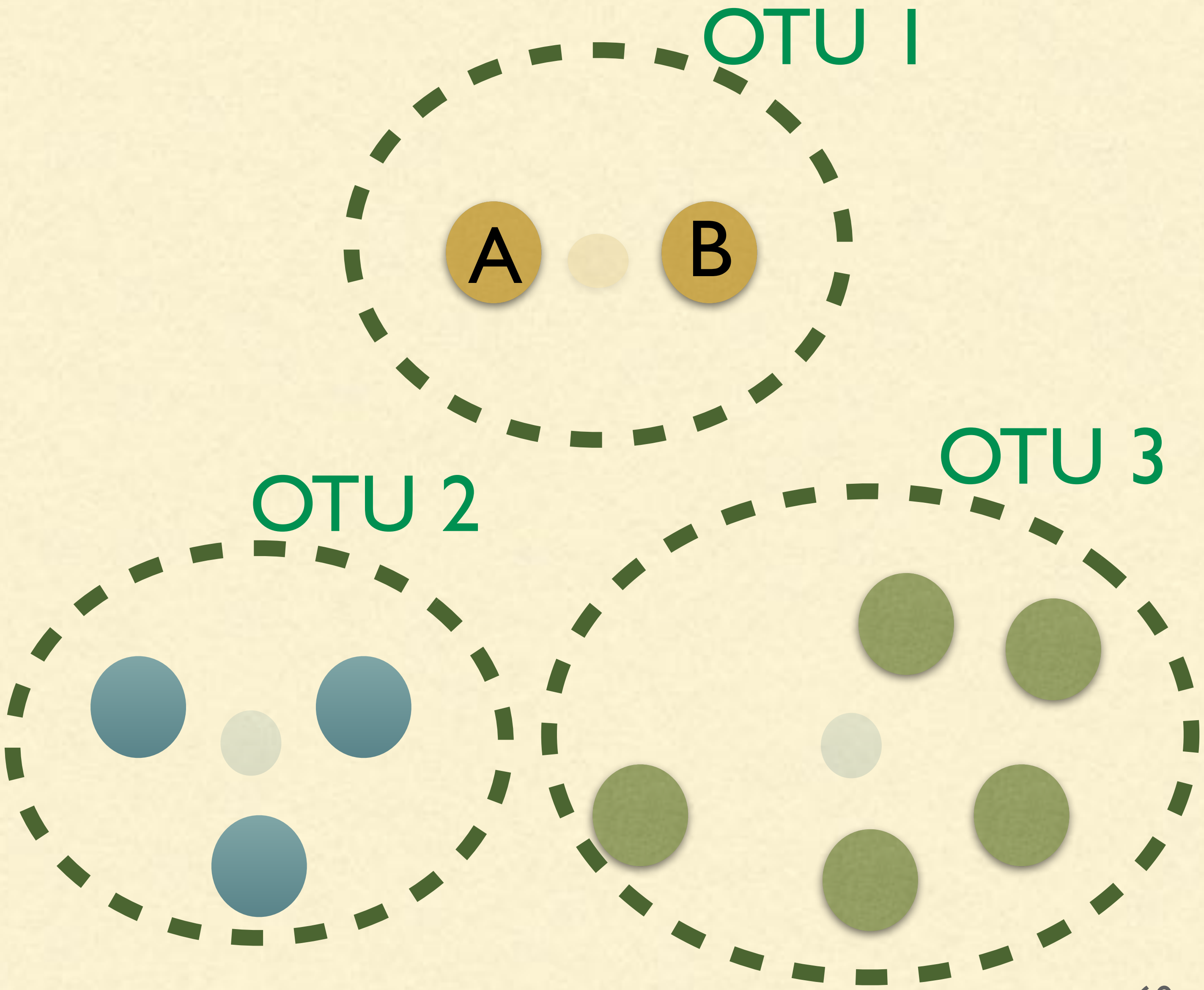


BIOLOGICAL UNITS: 16S DATA

Why are A and B different?

Option 1: sequencing errors

Option 2: actually different



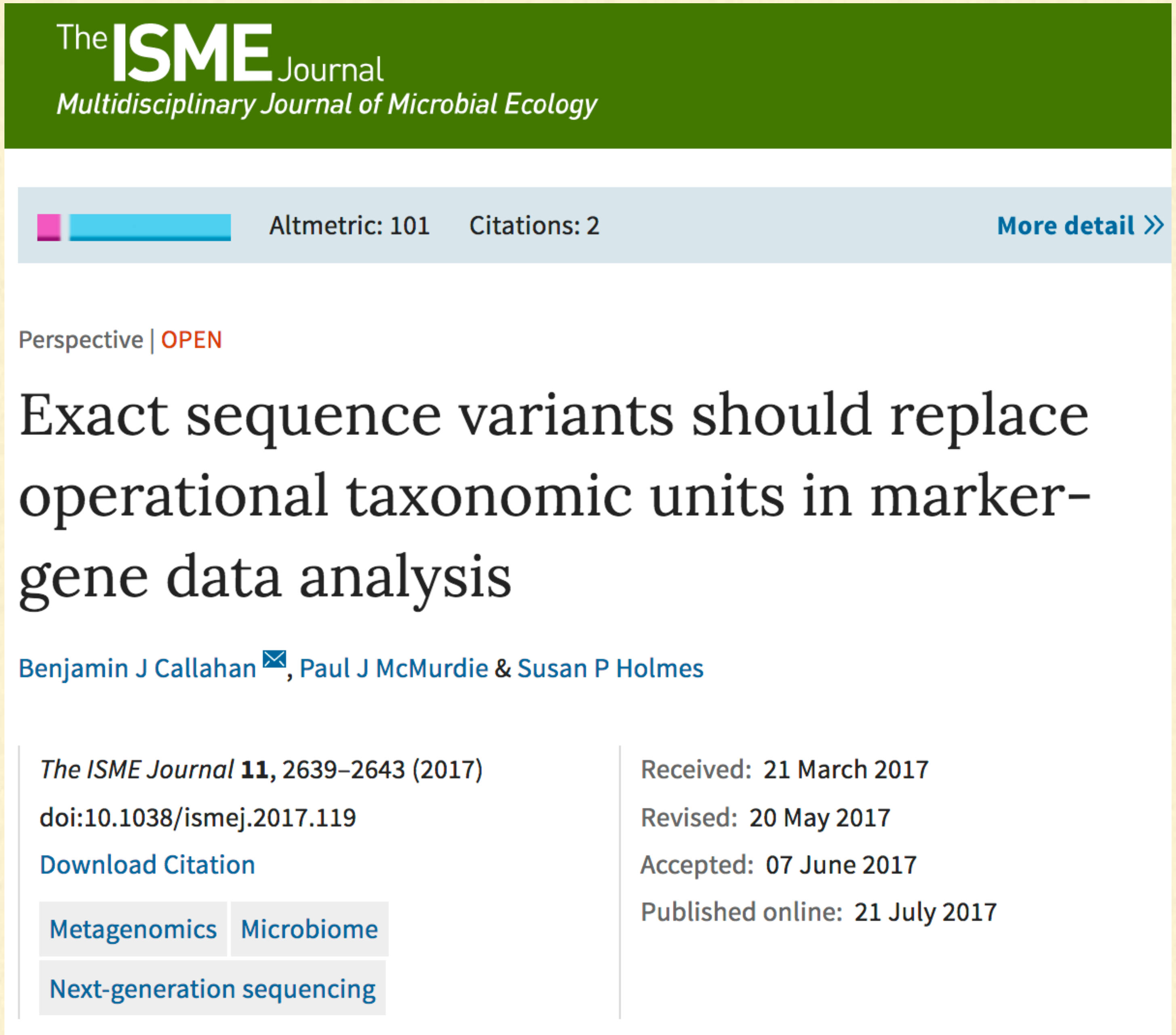
BIOLOGICAL UNITS: 16S DATA

Why are A and B different?

Option 1: sequencing errors

Option 2: actually different

These options should be distinguished!



The ISME Journal

Multidisciplinary Journal of Microbial Ecology

Altmetric: 101 Citations: 2 More detail »

Perspective | OPEN

Exact sequence variants should replace operational taxonomic units in marker-gene data analysis

Benjamin J Callahan, Paul J McMurdie & Susan P Holmes

The ISME Journal **11,** 2639–2643 (2017)

doi:10.1038/ismej.2017.119

Download Citation

Metagenomics Microbiome

Next-generation sequencing

Received: 21 March 2017

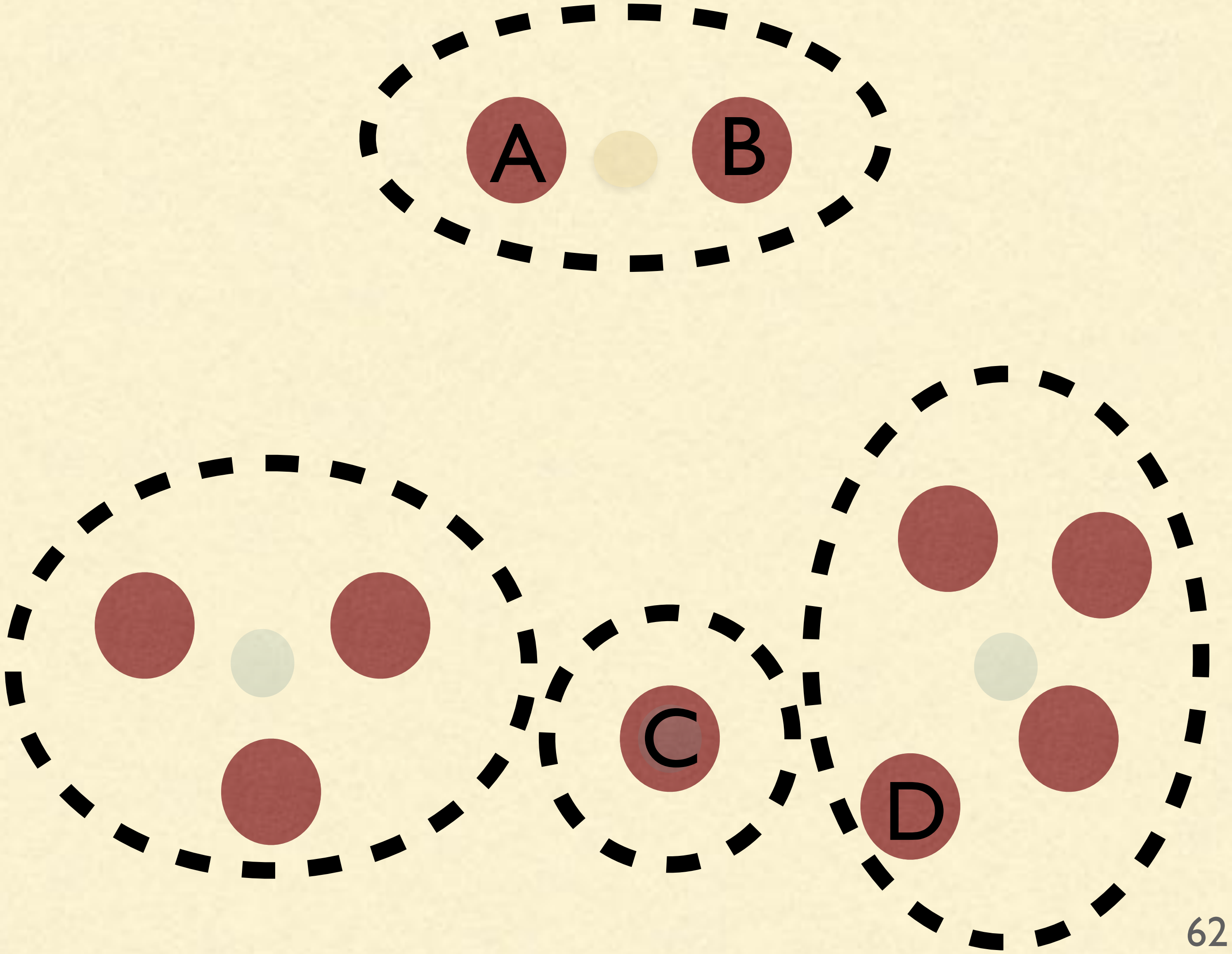
Revised: 20 May 2017

Accepted: 07 June 2017

Published online: 21 July 2017

BIOLOGICAL UNITS: ASVs

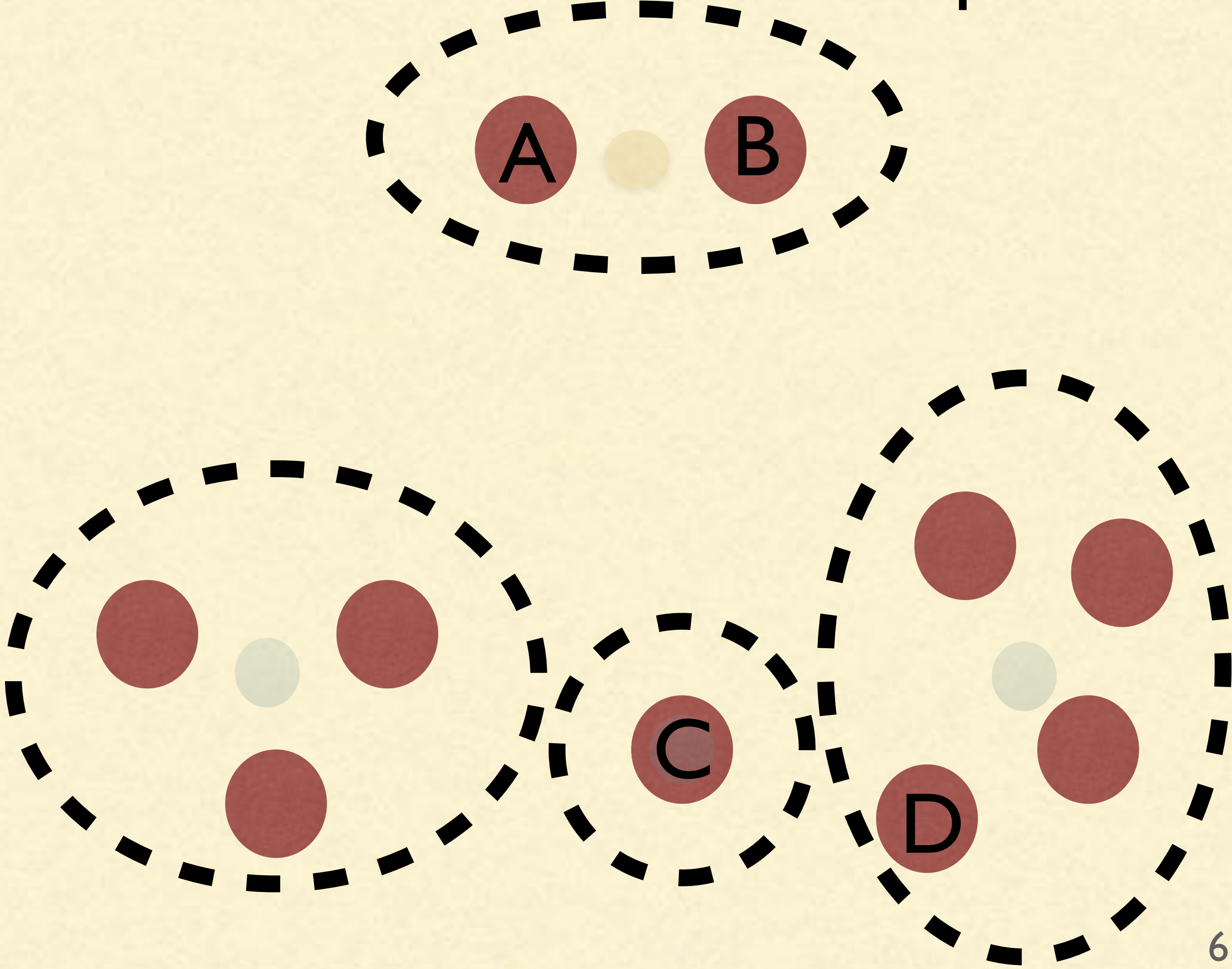
- We can estimate sequencing error rates
    - So can estimate how much observed sequences should vary around "true" sequence



BIOLOGICAL UNITS: ASVs

- We can estimate sequencing error rates
    - So can estimate how much observed sequences should vary around "true" sequence

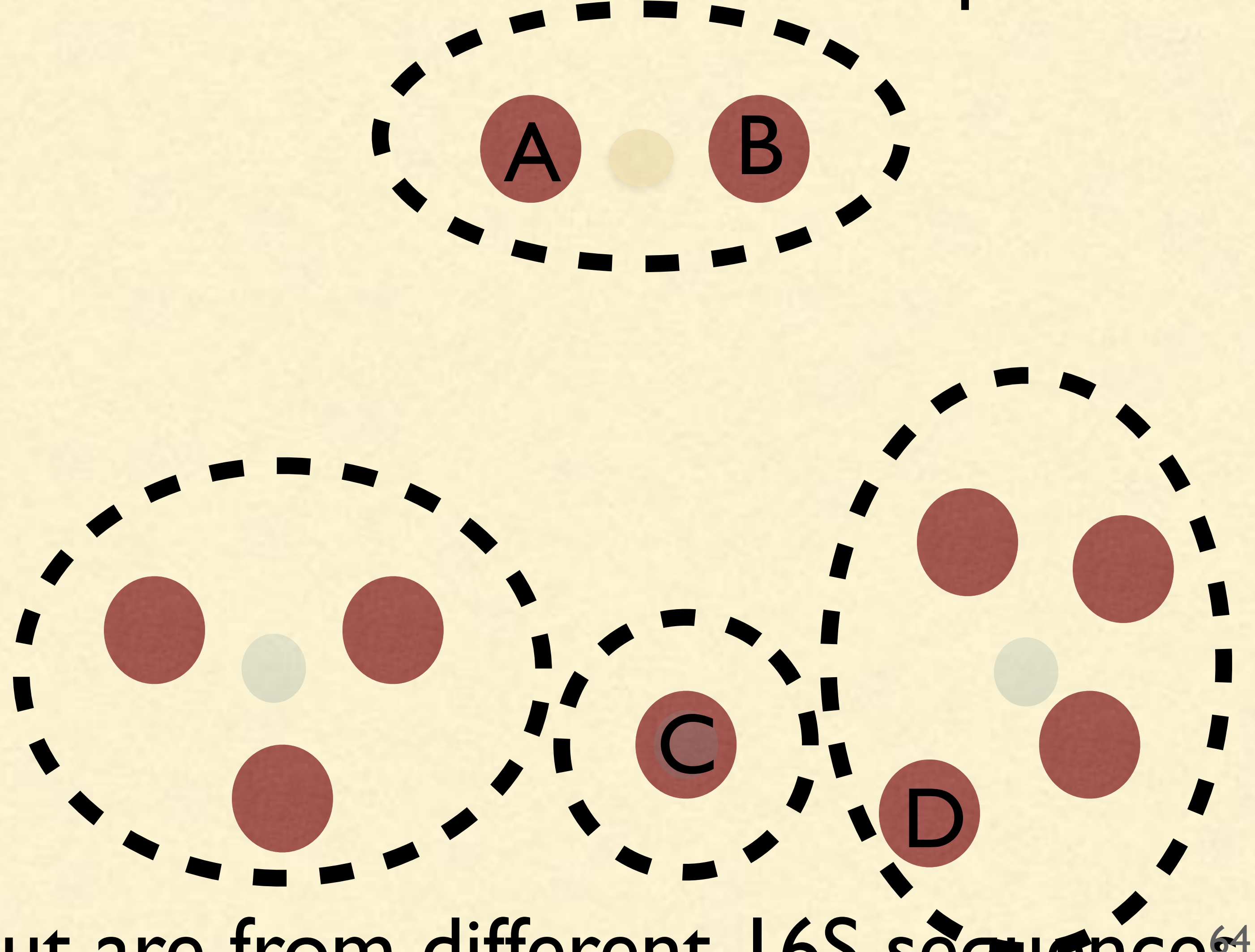
A & B are from the same 16S sequence



BIOLOGICAL UNITS: ASVs

- We can estimate sequencing error rates
  - So can estimate how much observed sequences should vary around "true" sequence

A & B are from the same 16S sequence



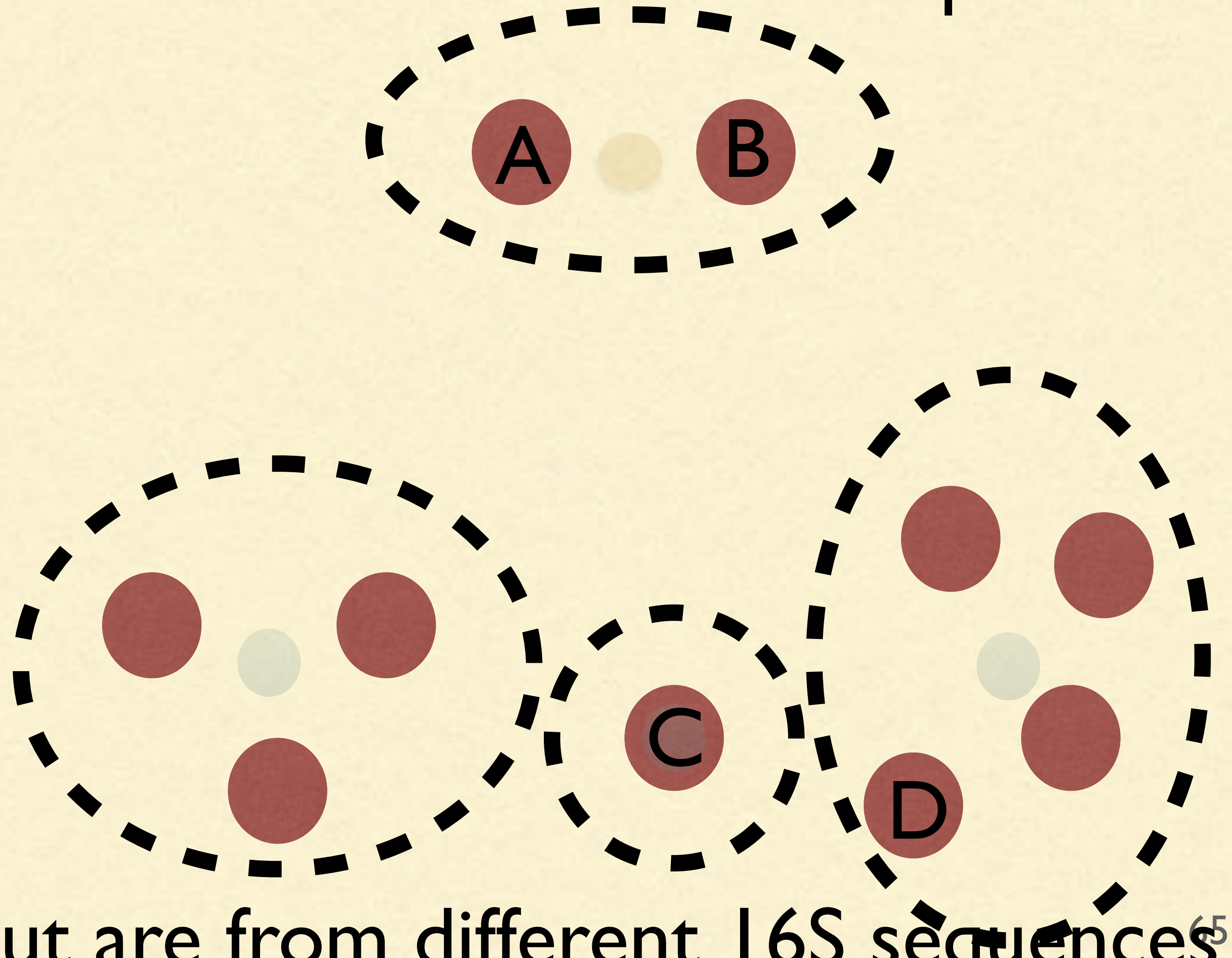
C & D are similar but are from different 16S sequences

(observed difference more than explainable by error rate)

BIOLOGICAL UNITS: ASVs

- Source sequences are called Amplicon Sequence Variants (ASVs)
- DADA2
  - ASV construction
  - Less spurious diversity

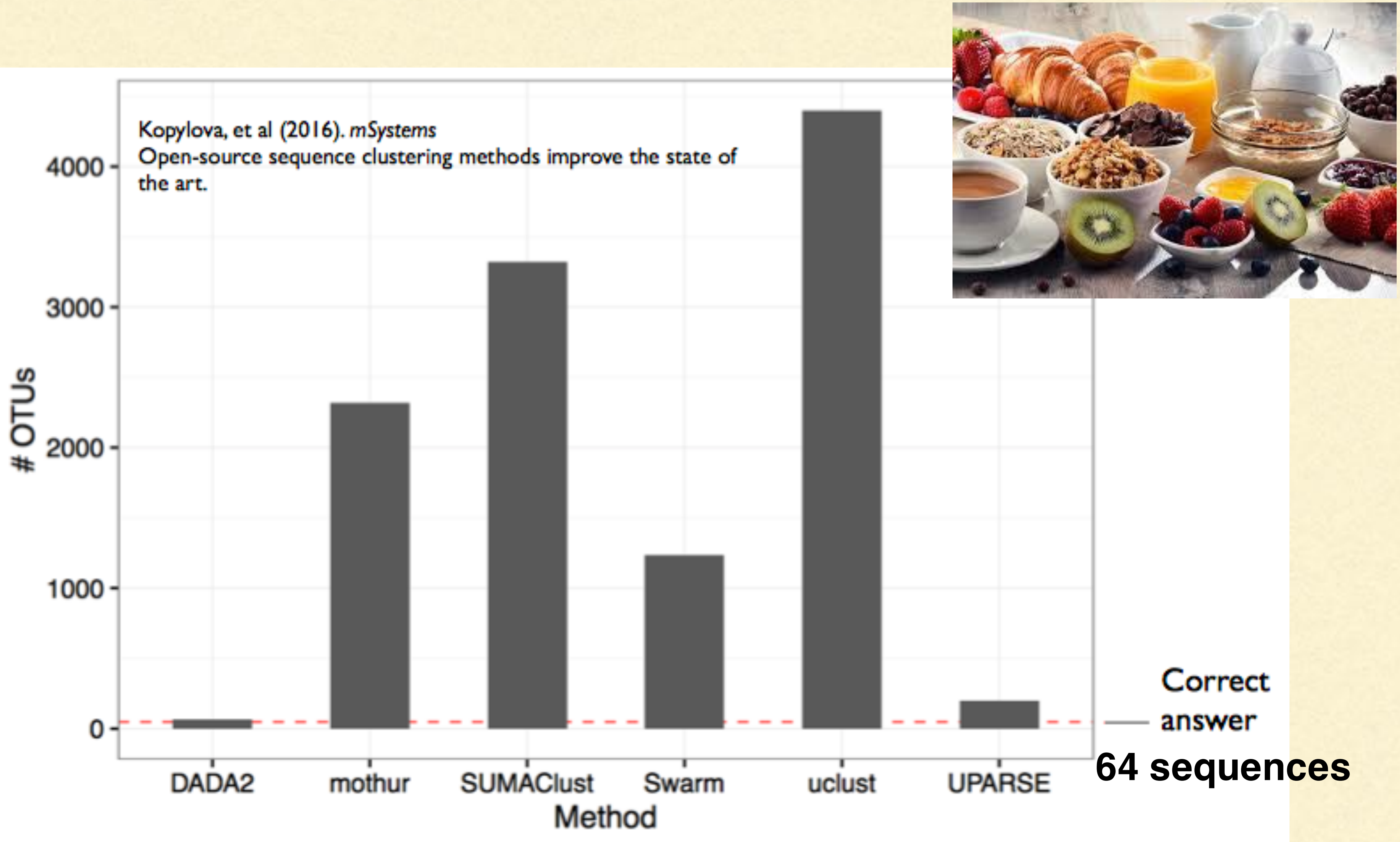
A & B are from the same 16S sequence



C & D are similar but are from different 16S sequences

(observed difference more than explainable by error rate)

DADA2: ASV ALGORITHM



Kopylova, et al (2016). *mSystems*
Open-source sequence clustering methods improve the state of the art.

http://benjjneb.github.io/dada2/R/SotA.html

Slide modified with permission from Scott Handley

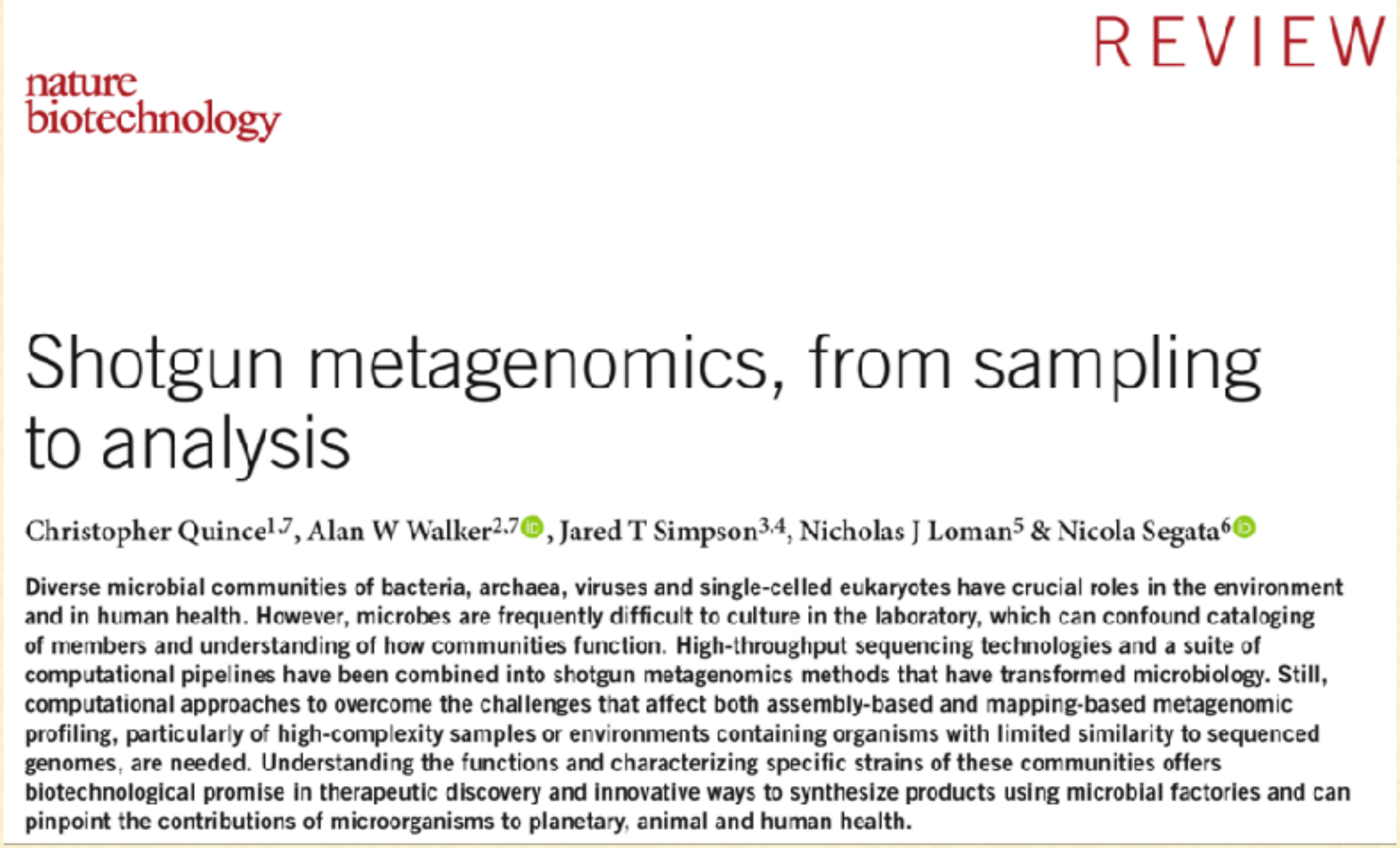
BIOLOGICAL UNITS: 16S

- Biological unit of 16S is 16S sequence
  - i.e. 16S amplicon sequence variants
- 16S sequences need to be clustered…
  - Old: into *operational clusters*
  - Modern: into *sequence variants*

BIOLOGICAL UNITS: WGS

- Many options
  - Genomes
  - Genes
  - Co-abundant genes
  - Others

GENES/GENOMES FROM WGS DATA



REVIEW

nature biotechnology

Shotgun metagenomics, from sampling to analysis

Christopher Quince[1,7], Alan W Walker[2,7], Jared T Simpson[3,4], Nicholas J Loman[5] & Nicola Segata[6]

Diverse microbial communities of bacteria, archaea, viruses and single-celled eukaryotes have crucial roles in the environment and in human health. However, microbes are frequently difficult to culture in the laboratory, which can confound cataloging of members and understanding of how communities function. High-throughput sequencing technologies and a suite of computational pipelines have been combined into shotgun metagenomics methods that have transformed microbiology. Still, computational approaches to overcome the challenges that affect both assembly-based and mapping-based metagenomic profiling, particularly of high-complexity samples or environments containing organisms with limited similarity to sequenced genomes, are needed. Understanding the functions and characterizing specific strains of these communities offers biotechnological promise in therapeutic discovery and innovative ways to synthesize products using microbial factories and can pinpoint the contributions of microorganisms to planetary, animal and human health.

ASSEMBLY

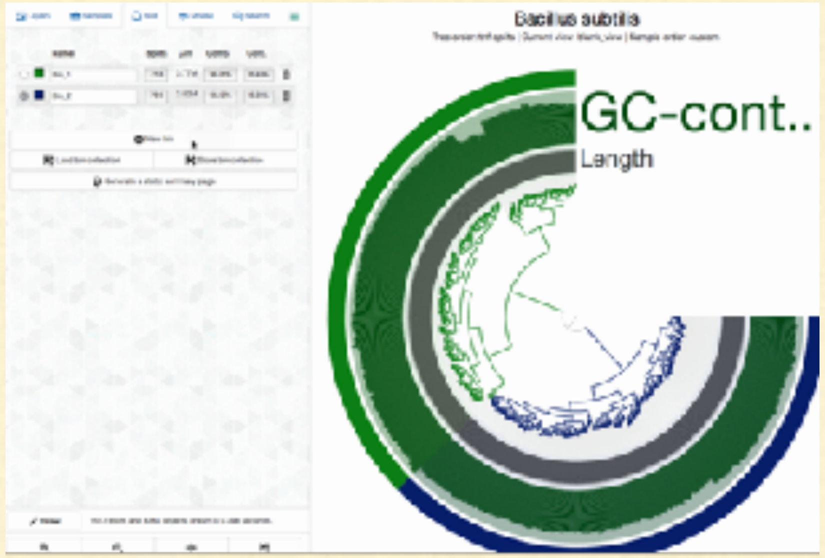
- Every genome is a puzzle, break into pieces, put pieces back together
  - Different microbes contain same genes
  - Microbes of same strain can have very similar genomes
    - e.g., SNVs, same genome but missing gene/operon
  - Can’t assume equal coverage across genomes/samples
    - low coverage => can’t piece puzzle together
    - high coverage => expensive

ASSEMBLY

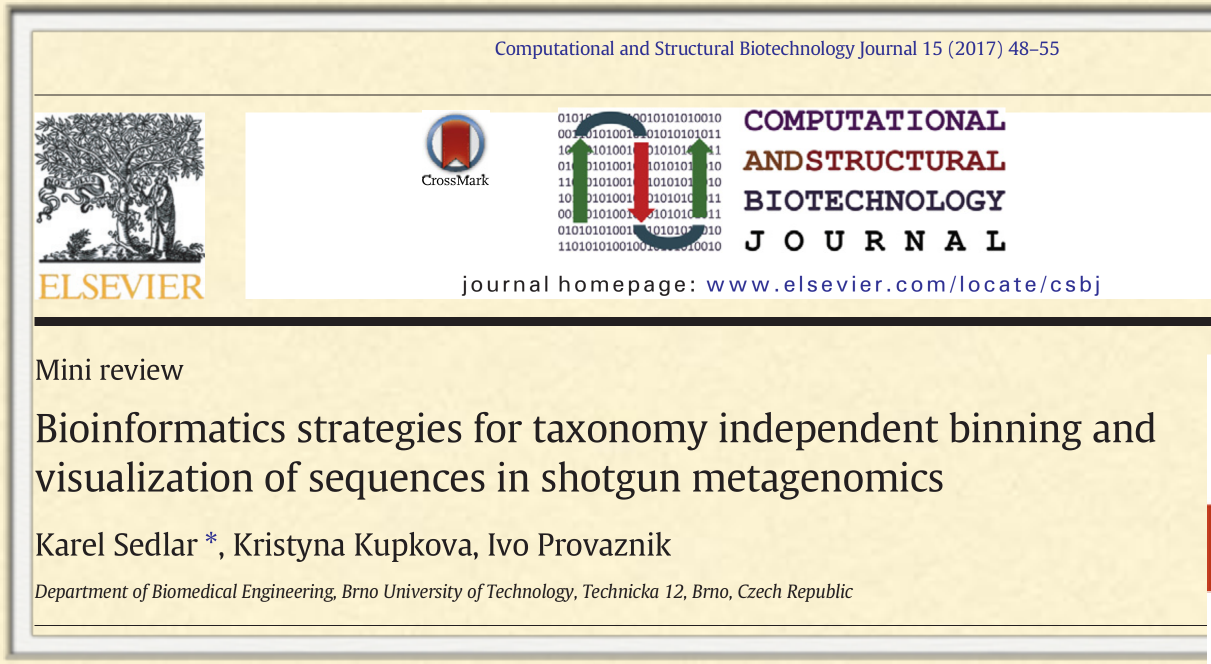
- Assemblers turn reads into (~10^4 - 10^6) **contigs**
- No single assembler "best"
  - Many use de Bruijn graphs: break reads into k-mers; find path
  - Inconsistent coverage is huge challenge
- Options: MEGAHIT, MetaSPAdes, others
  - MEGAHIT: "more genes that can be annotated in complex communities"
- Review article: "Use more than one!"

BINNING

- Contigs come from what genomes? How many genomes?
- Binning groups **contigs** into **genomes**
- Supervised & unsupervised
  - Choice dictated by reliability/availability of reference genomes
- Balance between automation and refinement
  - Anvi’o: helps with manual refinement
    - (More later)



UNSUPERVISED BINNING



Computational and Structural Biotechnology Journal 15 (2017) 48–55

ELSEVIER

CrossMark

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

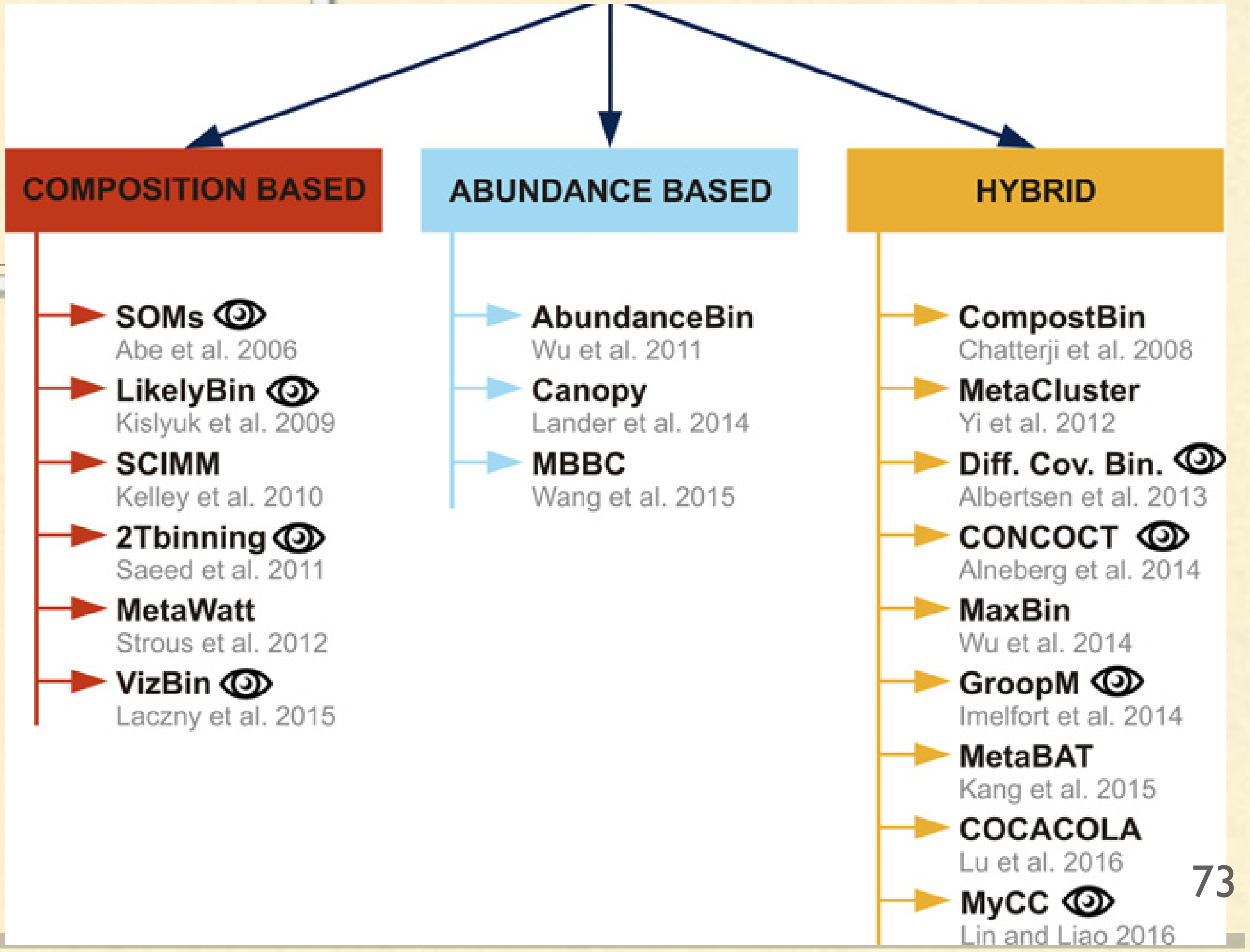
journal homepage: www.elsevier.com/locate/csbj

Mini review

Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics

Karel Sedlar *, Kristyna Kupkova, Ivo Provaznik

Department of Biomedical Engineering, Brno University of Technology, Technicka 12, Brno, Czech Republic



COMPOSITION BASED

- SOMs Abe et al. 2006
- LikelyBin Kislyuk et al. 2009
- SCIMM Kelley et al. 2010
- 2Tbinning Saeed et al. 2011
- MetaWatt Strous et al. 2012
- VizBin Laczny et al. 2015

ABUNDANCE BASED

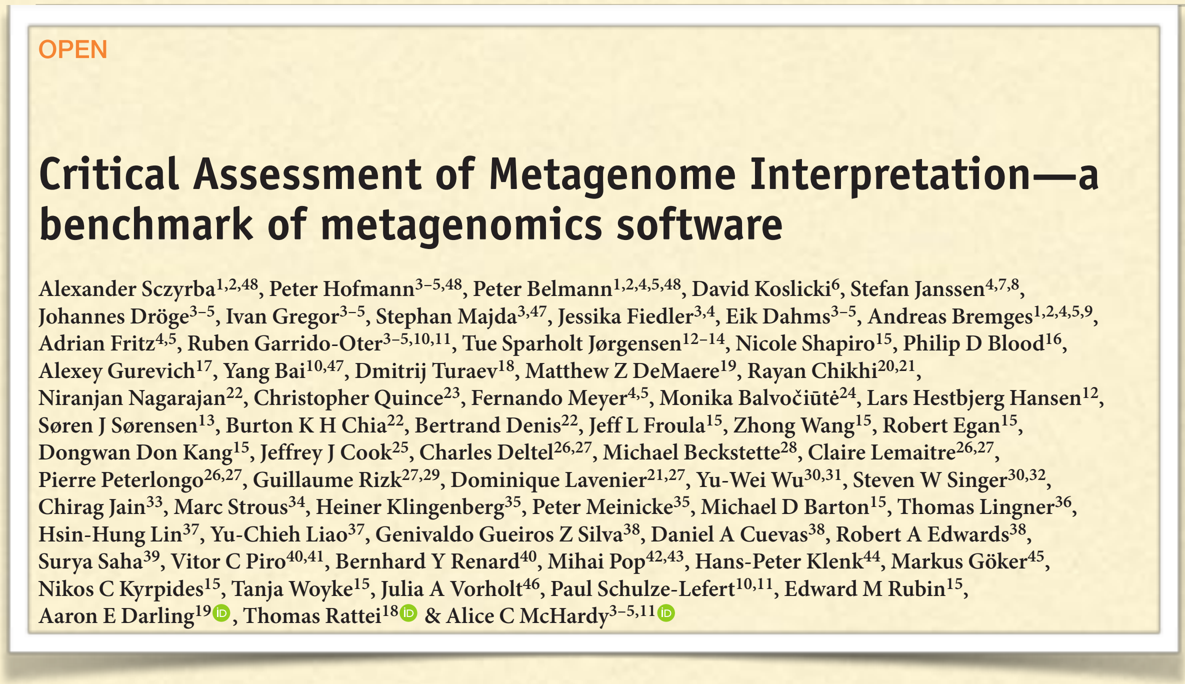
- AbundanceBin Wu et al. 2011
- Canopy Lander et al. 2014
- MBBC Wang et al. 2015

HYBRID

- CompostBin Chatterji et al. 2008
- MetaCluster Yi et al. 2012
- Diff. Cov. Bin. Albertsen et al. 2013
- CONCOCT Alneberg et al. 2014
- MaxBin Wu et al. 2014
- GroopM Imelfort et al. 2014
- MetaBAT Kang et al. 2015
- COCACOLA Lu et al. 2016
- MyCC Lin and Liao 2016

Slide modified with permission from Christian Mueller

UNSUPERVISED BINNING



OPEN

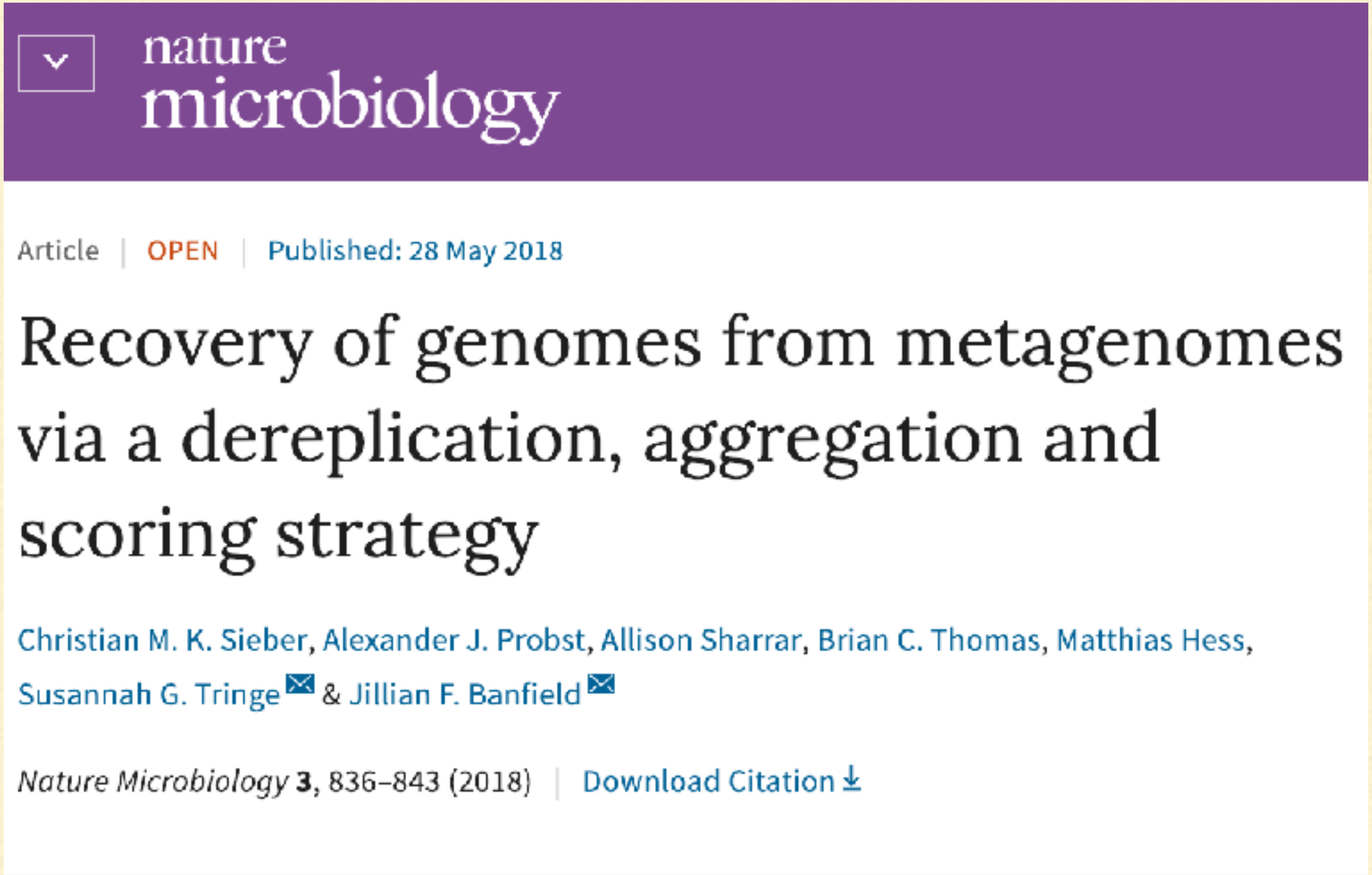
Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software

Alexander Sczyrba[1,2,48], Peter Hofmann[3–5,48], Peter Belmann[1,2,4,5,48], David Koslicki[6], Stefan Janssen[4,7,8], Johannes Dröge[3–5], Ivan Gregor[3–5], Stephan Majda[3,47], Jessika Fiedler[3,4], Eik Dahms[3–5], Andreas Bremges[1,2,4,5,9], Adrian Fritz[4,5], Ruben Garrido-Oter[3–5,10,11], Tue Sparholt Jørgensen[12–14], Nicole Shapiro[15], Philip D Blood[16], Alexey Gurevich[17], Yang Bai[10,47], Dmitrij Turaev[18], Matthew Z DeMaere[19], Rayan Chikhi[20,21], Niranjan Nagarajan[22], Christopher Quince[23], Fernando Meyer[4,5], Monika Balvočiūtė[24], Lars Hestbjerg Hansen[12], Søren J Sørensen[13], Burton K H Chia[22], Bertrand Denis[22], Jeff L Froula[15], Zhong Wang[15], Robert Egan[15], Dongwan Don Kang[15], Jeffrey J Cook[25], Charles Deltel[26,27], Michael Beckstette[28], Claire Lemaitre[26,27], Pierre Peterlongo[26,27], Guillaume Rizk[27,29], Dominique Lavenier[21,27], Yu-Wei Wu[30,31], Steven W Singer[30,32], Chirag Jain[33], Marc Strous[34], Heiner Klingenberg[35], Peter Meinicke[35], Michael D Barton[15], Thomas Lingner[36], Hsin-Hung Lin[37], Yu-Chieh Liao[37], Genivaldo Gueiros Z Silva[38], Daniel A Cuevas[38], Robert A Edwards[38], Surya Saha[39], Vitor C Piro[40,41], Bernhard Y Renard[40], Mihai Pop[42,43], Hans-Peter Klenk[44], Markus Göker[45], Nikos C Kyrpides[15], Tanja Woyke[15], Julia A Vorholt[46], Paul Schulze-Lefert[10,11], Edward M Rubin[15], Aaron E Darling[19], Thomas Rattei[18] & Alice C McHardy[3–5,11]

Genome binner (% contamination)			Recovered genomes (% completeness)		
			>50%	>70%	>90%
Gold standard	—		753	753	753
CONCOCT	—	<10% <5%	275 267	272 265	262 256
MetaWatt 3.5	—	<10% <5%	500 476	475 452	405 393
MetaBAT	—	<10% <5%	247 234	228 216	195 186
MyCC	—	<10% <5%	250 220	240 211	197 173
MaxBin 2.0	—	<10% <5%	390 356	385 352	343 316

Slide modified with permission from Christian Mueller

UNSUPERVISED BINNING



nature microbiology

Article | OPEN | Published: 28 May 2018

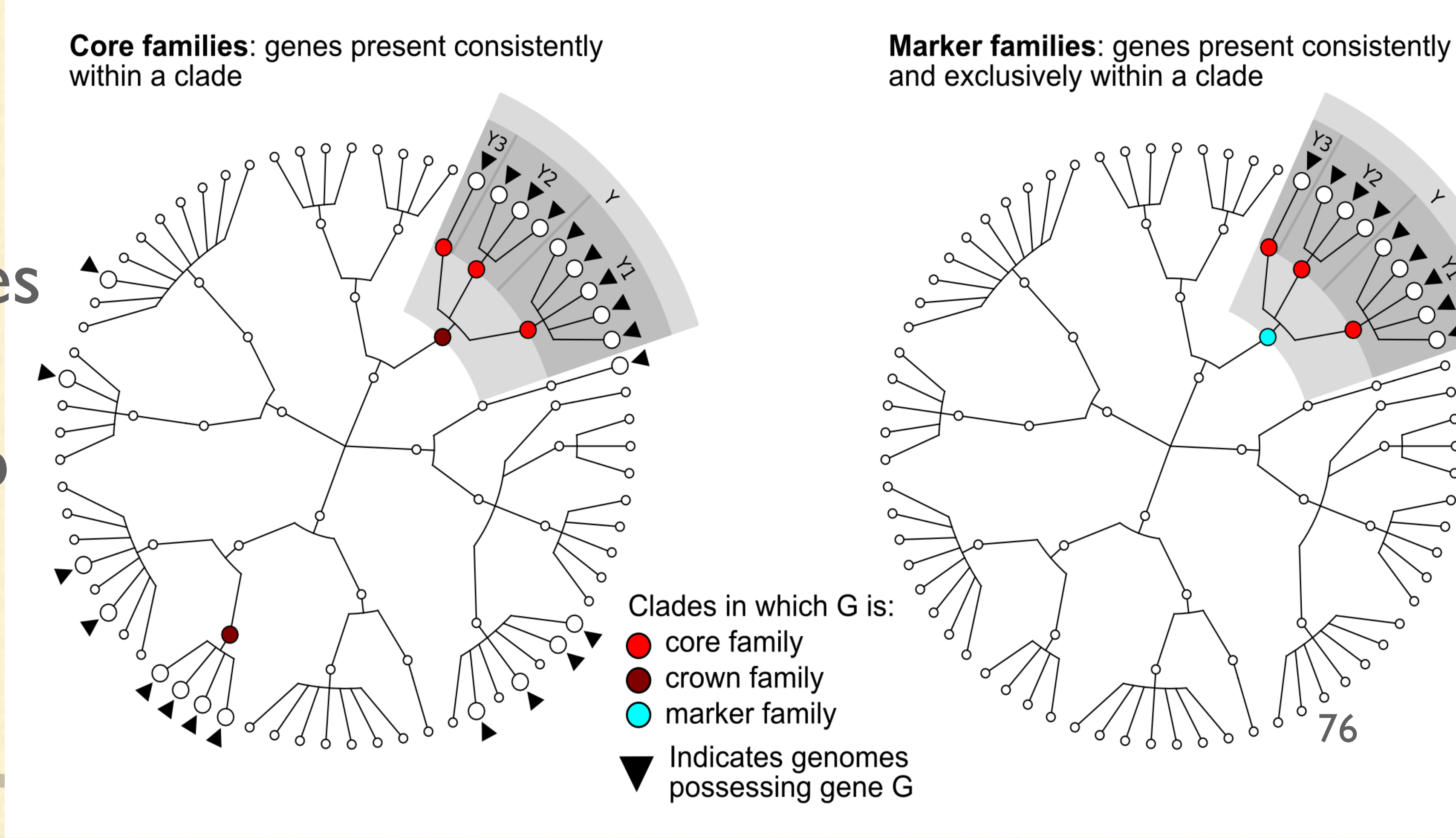
Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy

Christian M. K. Sieber, Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe ✉ & Jillian F. Banfield ✉

Nature Microbiology **3**, 836–843 (2018) | Download Citation ⤓

ASSEMBLY-FREE WGS

- Can map reads to genomes (often not faster than assembly; high FP)
- Better idea: use specific genes (not all genes)
- **MetaPhlAn**
  - Core & marker genes
  - Great for human mb


Core families: genes present consistently within a clade
Marker families: genes present consistently and exclusively within a clade
Y3
Y2
Y
Y1
Clades in which G is:
● core family
● crown family
● marker family
▼ Indicates genomes possessing gene G

BIOLOGICAL UNITS: GENES

- From genomes… genes!
- Adapt tools from single-genome world
  - Challenge: microbial genes mostly uncharacterized

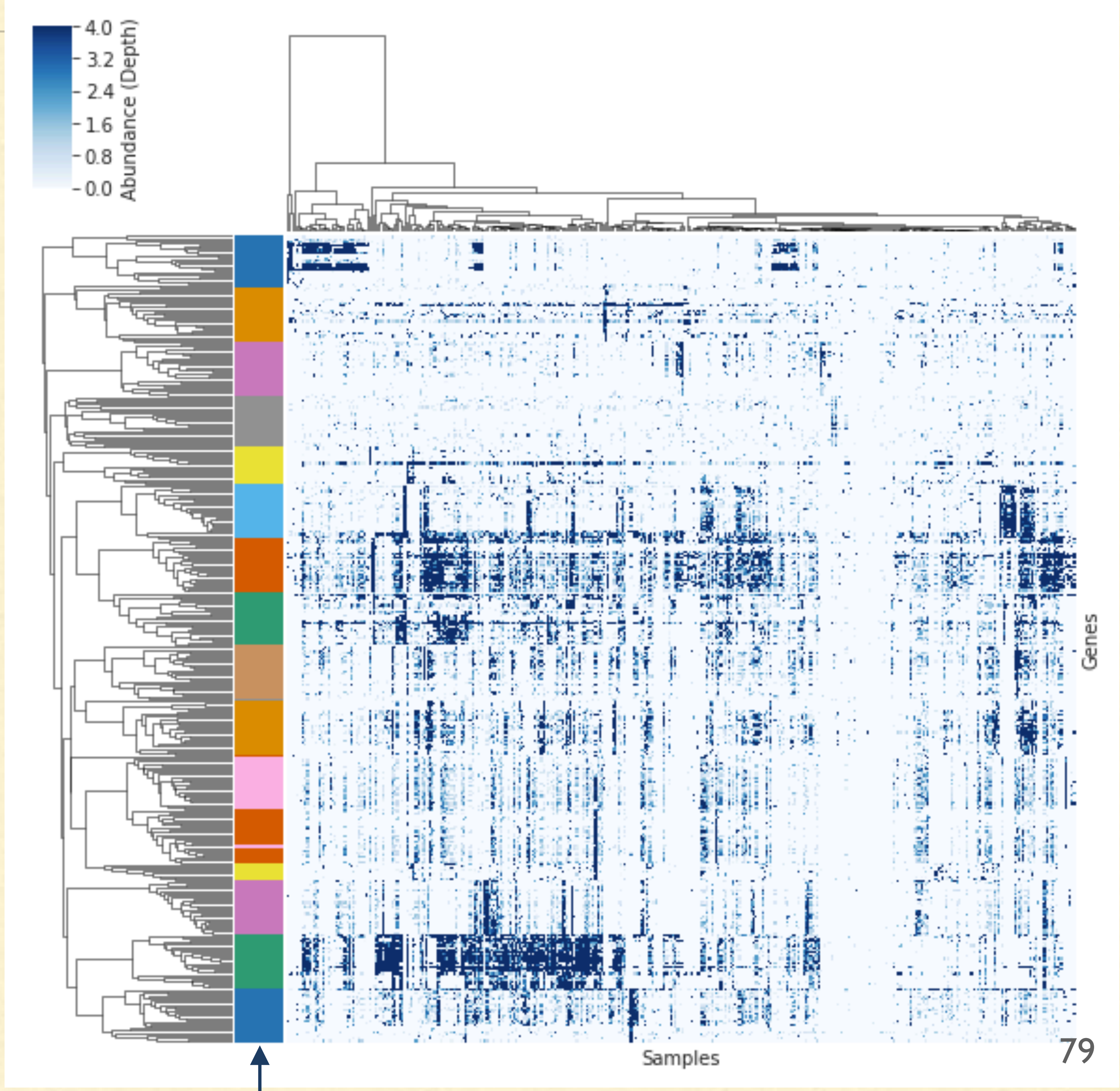
BIOLOGICAL UNITS

- Challenges with WGS include
  - lots of genes
  - Choice of database has enormous impact
- Advantage: lots of redundancy = genes that occur together
  - Genes that consistently occur together arguably biological unit
    - CAGs = co-abundant genes; grouping of genes that are consistently present/absent together across samples

CAGs AS BIOLOGICAL UNITS

Work lead by Sam Minot (Fred Hutch)

- Co-abundant gene (CAG) construction algorithm
- No databases
- Reproducibly associated with disease



CAG grouping

BIOLOGICAL UNITS TO CLEANING

- Once you have your data sorted into biological units, you may need to do some cleaning
- Often cleaning = filtering
  - e.g., low yield
  - e.g., low quality score data
  - e.g., likely sequencing errors (sometimes low abundance 🚩)
  - e.g., contaminants (e.g., with *decontam*)

SUMMARY: FIRST HALF

- Microbes, their relevance, questions
- Technology to study microbes
- Processing data into meaningful units

- Next up: analysis; open problems

BREAK

☕☕☕☕☕☕☕☕🍪☕

☕☕☕☕☕☕☕☕☕☕

☕☕☕☕☕☕☕☕☕☕

ANALYSIS

- The type of data that you have affect how you will analyse
  - e.g., compositional/relative/absolute
- The questions that you have affect how you analyse
  - e.g., exploratory/confirmatory

SCENARIO

PROPORTION DATA

ABSOLUTE ABUNDANCE	MICROBE A	MICROBE B	MICROBE C
ENVIRO 1	5	5	20
ENVIRO 2	10	10	40

↓ observe

# OBSERVED	MICROBE A	MICROBE B	MICROBE C	TOTAL
ENVIRO 1	1.01 / 6	1/6	3.99 / 6	1
ENVIRO 2	0.99 / 6	0.99/6	4.02 / 6	1

Can compare across rows & columns

SCENARIO

ABSOLUTE DATA

ABSOLUTE ABUNDANCE	MICROBE A	MICROBE B	MICROBE C
ENVIRO 1	5	5	20
ENVIRO 2	10	10	40

↓ observe

# OBSERVED	MICROBE A	MICROBE B	MICROBE C	TOTAL
ENVIRO 1	4	5	18	27
ENVIRO 2	9	9	37	55

Can compare across rows & columns

SCENARIO

COMPOSITIONAL/RELATIVE DATA

ABSOLUTE ABUNDANCE	MICROBE A	MICROBE B	MICROBE C
ENVIRO 1	5	5	20
ENVIRO 2	10	10	40

↓ observe

# OBSERVED	MICROBE A	MICROBE B	MICROBE C	TOTAL
ENVIRO 1	499	500	2001	3000
ENVIRO 2	250	251	1010	1511

Can compare across rows only

DATA

- 16S and WGS data are compositional/relative
  - Can compare observed values within samples
  - Common (users/software): convert to proportions
    - ADW: Disagree, this loses information about precision
    - ADW: Good statistical methods model precision
  - Implications for analysis

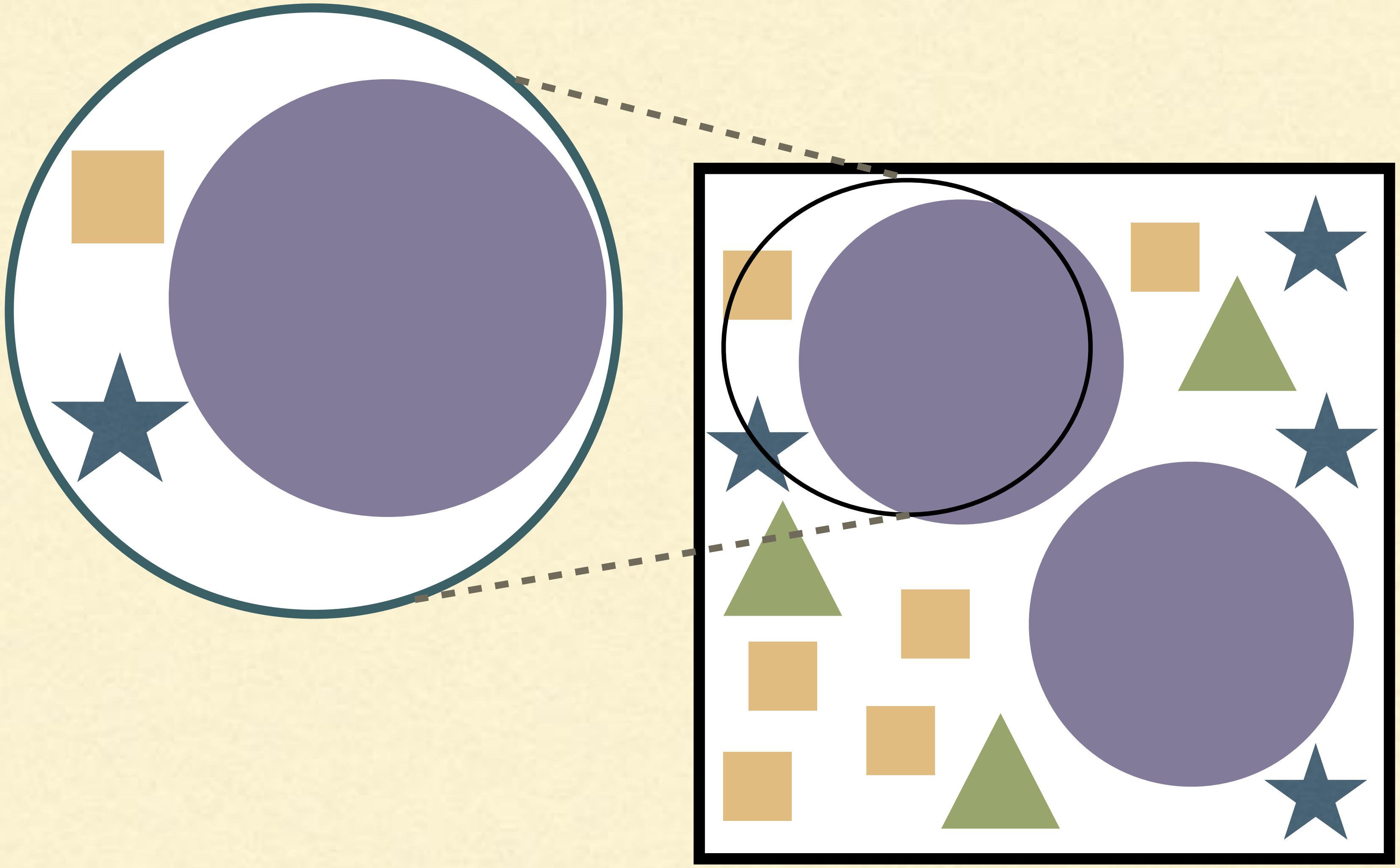
PARAMETERS

- Estimation: using information about the sample to estimate something about the population

★ = 1/3

● = 1/3

■ = 1/3



★ = 4/15

● = 2/15

▲ = 1/5

■ = 2/5

PARAMETERS

- something about the population = “parameter”
  - Genus-level relative abundance of Streptococcus in your saliva right now
  - Proportion of Krumlovians with MRSA infections
  - Mean phylum-level diversity on the hands of #evomics19 faculty

PARAMETERS FOR COMPOSITIONAL DATA

- Diversity parameters: α, β
  - sometimes called diversity "indices"
    - ADW: this terminology reflects a lack of understanding of statistical concept of parameters
- Relative abundance of taxon/gene
- Relative abundance within an environment ("enrichment")

DIVERSITY

- Low dimensional summaries of entire communities
  - α-diversity: one community
    - e.g., species richness, Shannon diversity
  - β-diversity: multiple communities
    - e.g., UniFrac, Bray-Curtis
- **Diversity is relevant in lots of contexts... not just the microbiome!**

DIVERSITY & PARAMETERS

- There are multiple choices to make when talking about diversity
  - Which taxonomic level? (strain/species/genus...)
  - Which diversity parameter?
  - Which estimate of the diversity parameter?

DIVERSITY & PARAMETERS

- There are multiple choices to make when talking about diversity
  - Which taxonomic level? (strain/species/genus...)
  - **Which diversity parameter?**
  - Which estimate of the diversity parameter?

ALPHA DIVERSITY

- Suppose we have C groups in our environment in proportions p_1, p_2, …, p_C
- Any function of
  - p_1, p_2, …, p_C OR
  - p_1, p_2, …, p_C and phylogeny ~~some info about relationships amongst groups~~

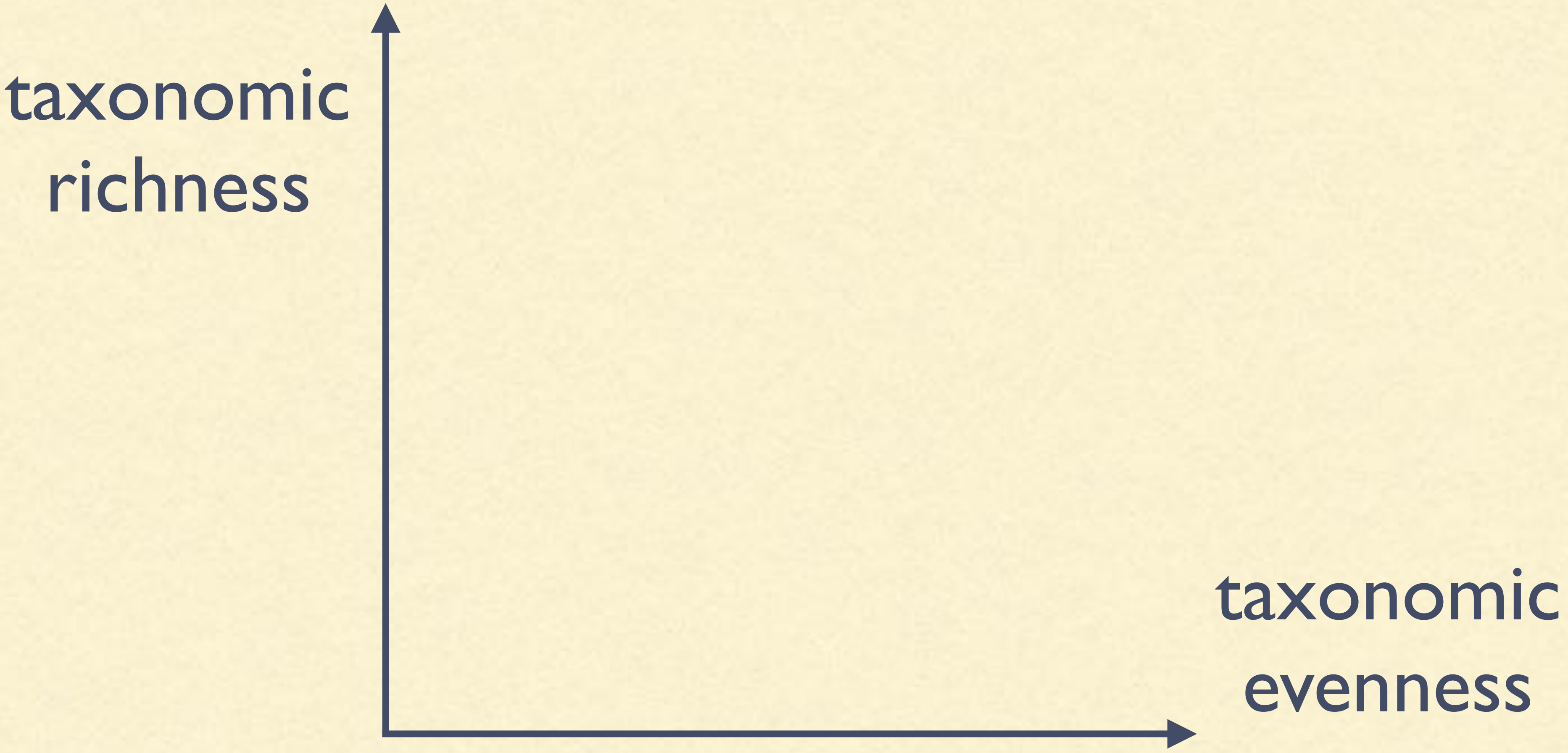
is a valid α-diversity parameter

ALPHA DIVERSITY

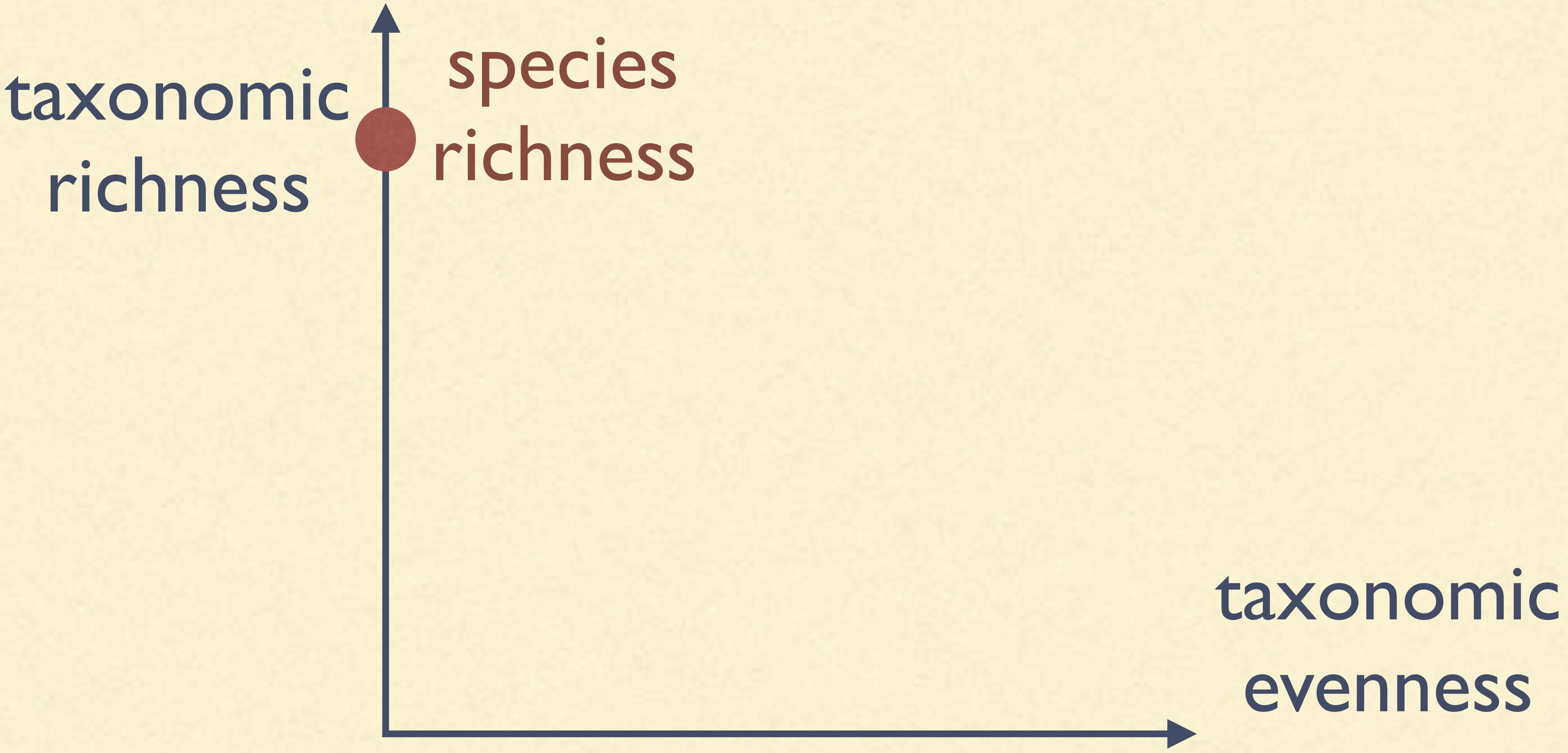
- Some examples of α-diversity measures include
  - Species richness: C
  - Simpson’s index: $\displaystyle\sum_{i=1}^{C} p_i^2$
  - Shannon diversity: $\displaystyle-\sum_{i=1}^{C} p_i \ln p_i$
  - Shannon's E: $\dfrac{-\sum_{i=1}^{C} p_i \ln p_i}{\ln C}$

YOUR CHOICE

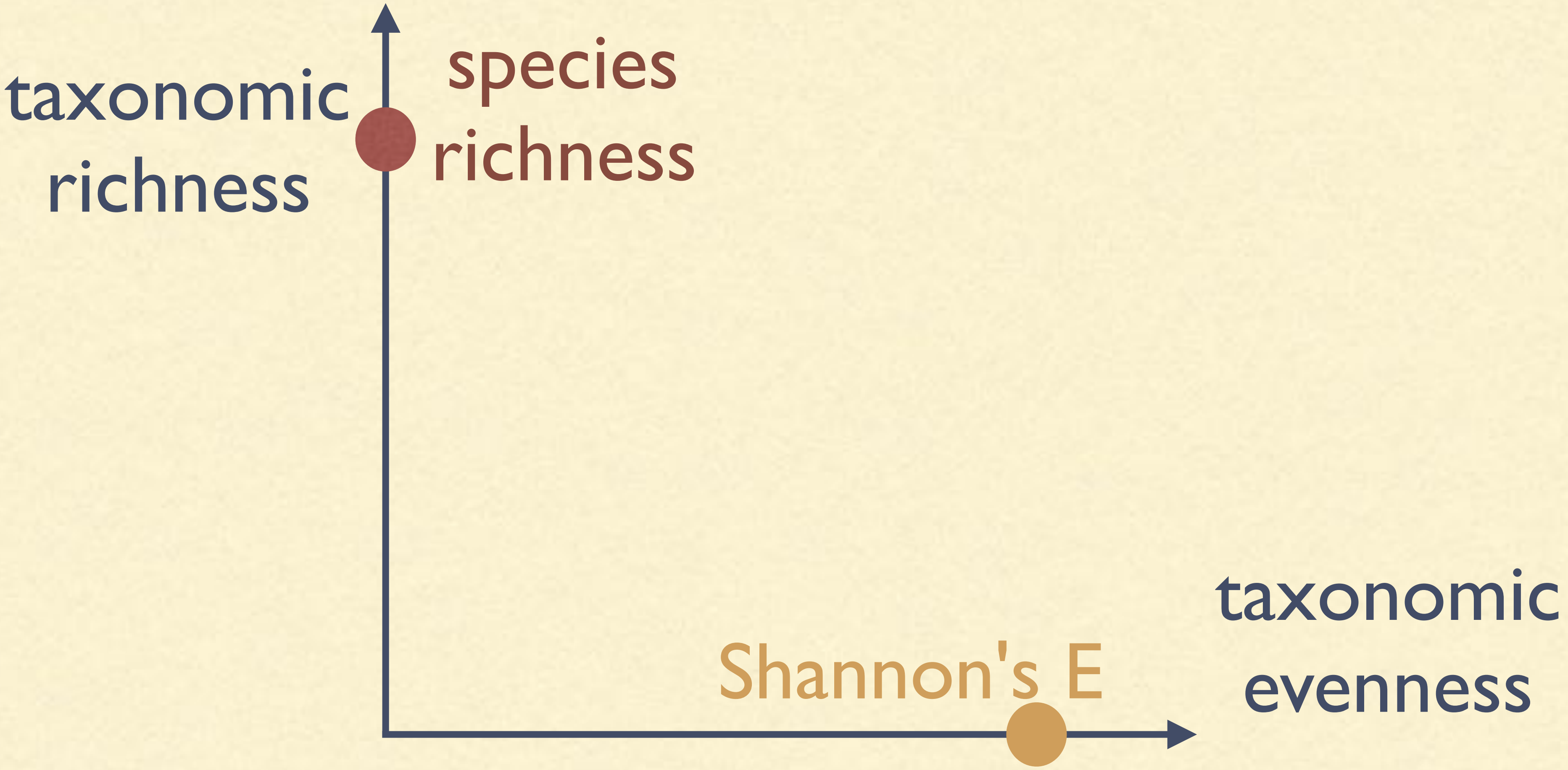
- Think: What difference do you want to highlight?



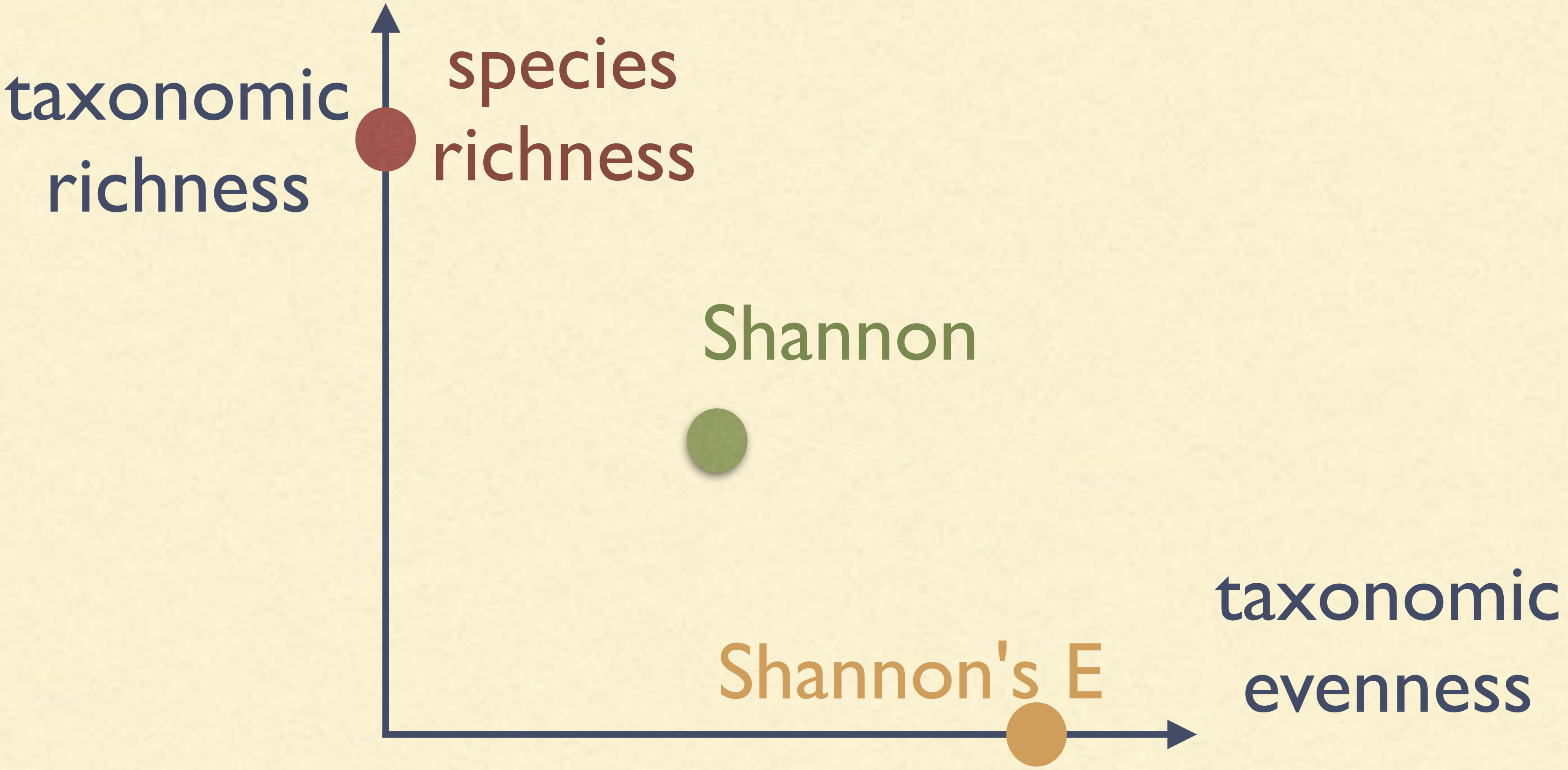
YOUR CHOICE



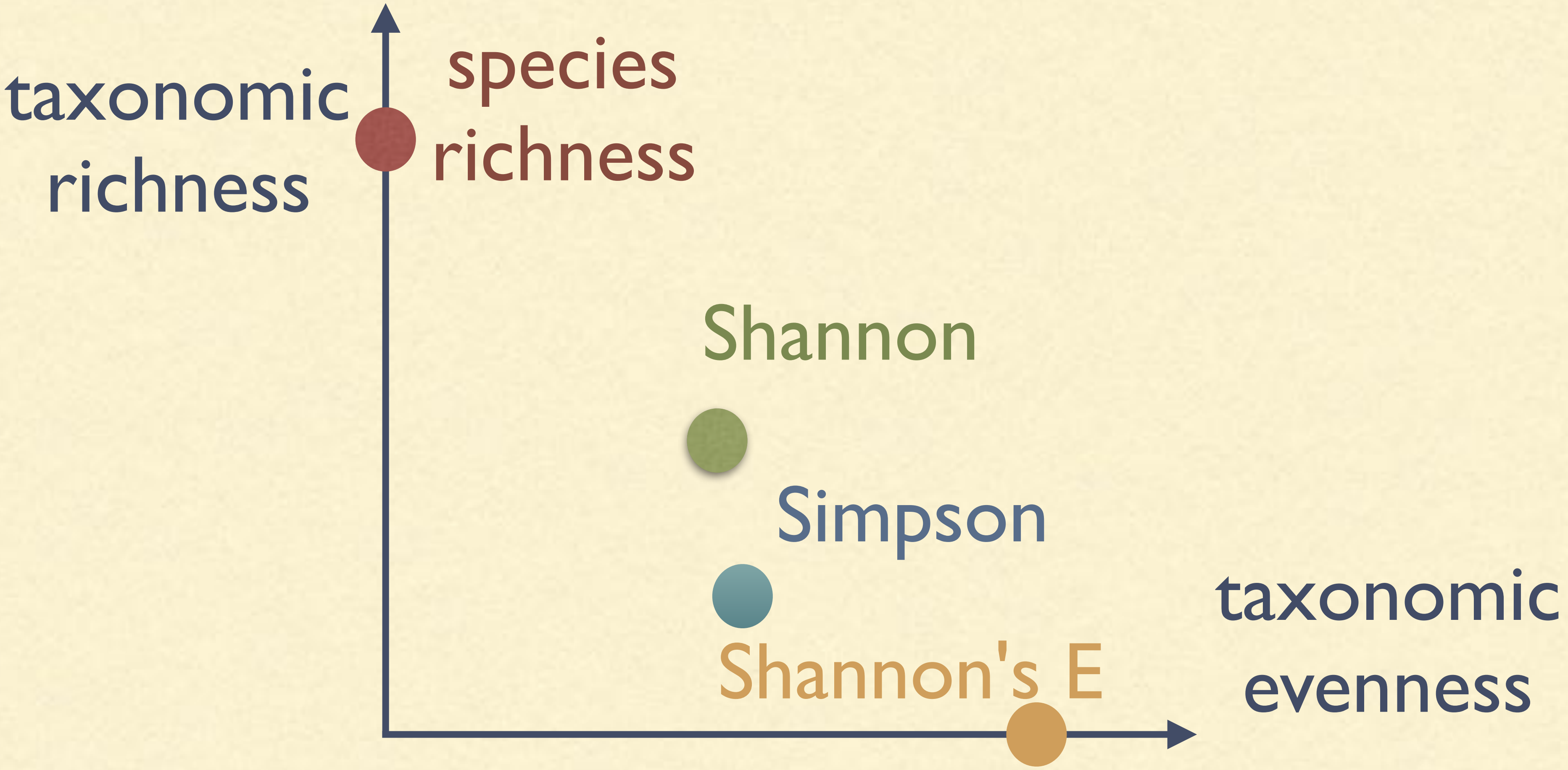
YOUR CHOICE


taxonomic richness
species richness
Shannon's E
taxonomic evenness

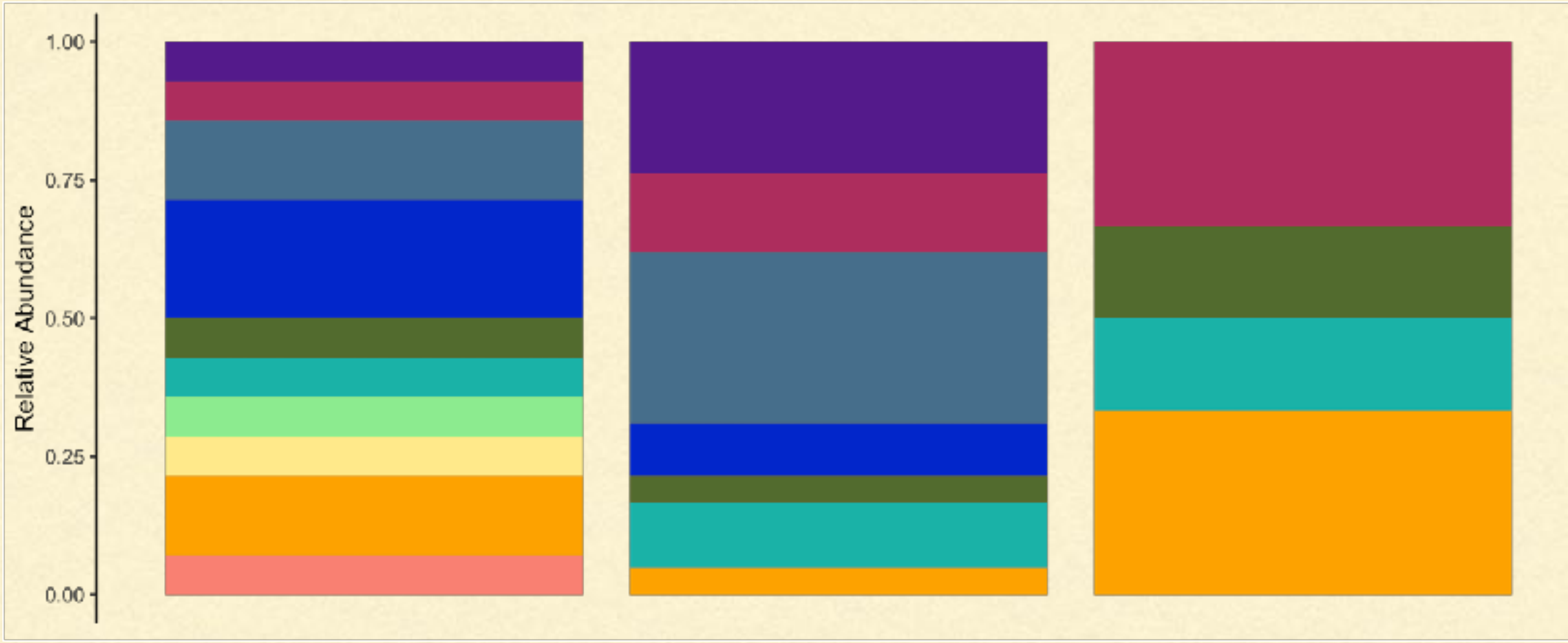
YOUR CHOICE


taxonomic richness
species richness
Shannon
Shannon's E
taxonomic evenness

YOUR CHOICE



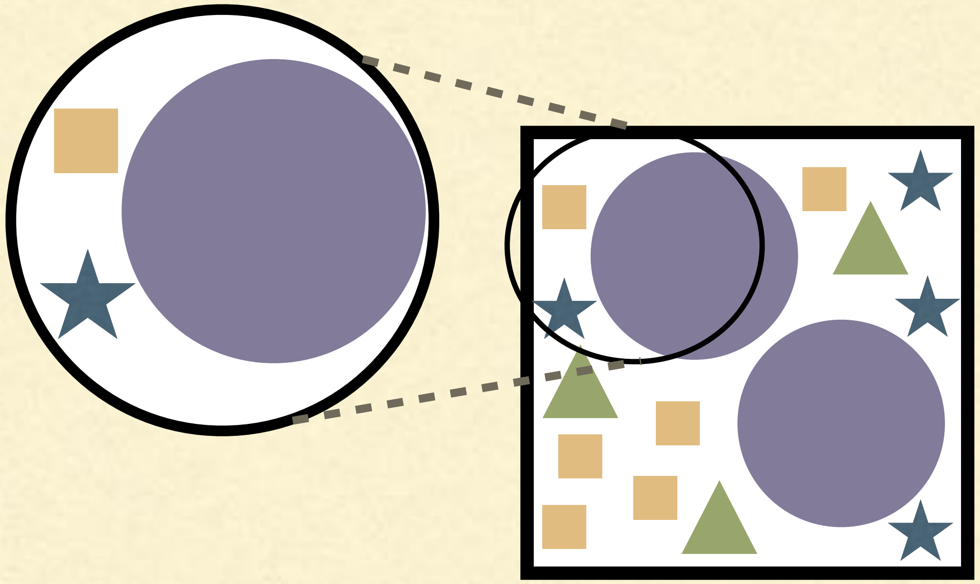
This is a question of *parameter choice:*
Which parameter highlights the differences I care about?



Richness	10	7	4
Shannon	2.21	1.75	1.33
Evenness	0.96	0.90	0.96
Simpson's	0.88	0.80	0.72
Inverse Simpson's	8.17	4.98	3.60

THE PROBLEM

- In practice, we don't observe the entire community, just a sample from it
  - we don't know C or $p_1, p_2, \ldots, p_C$
- **We need to *estimate* them using the data we collected**



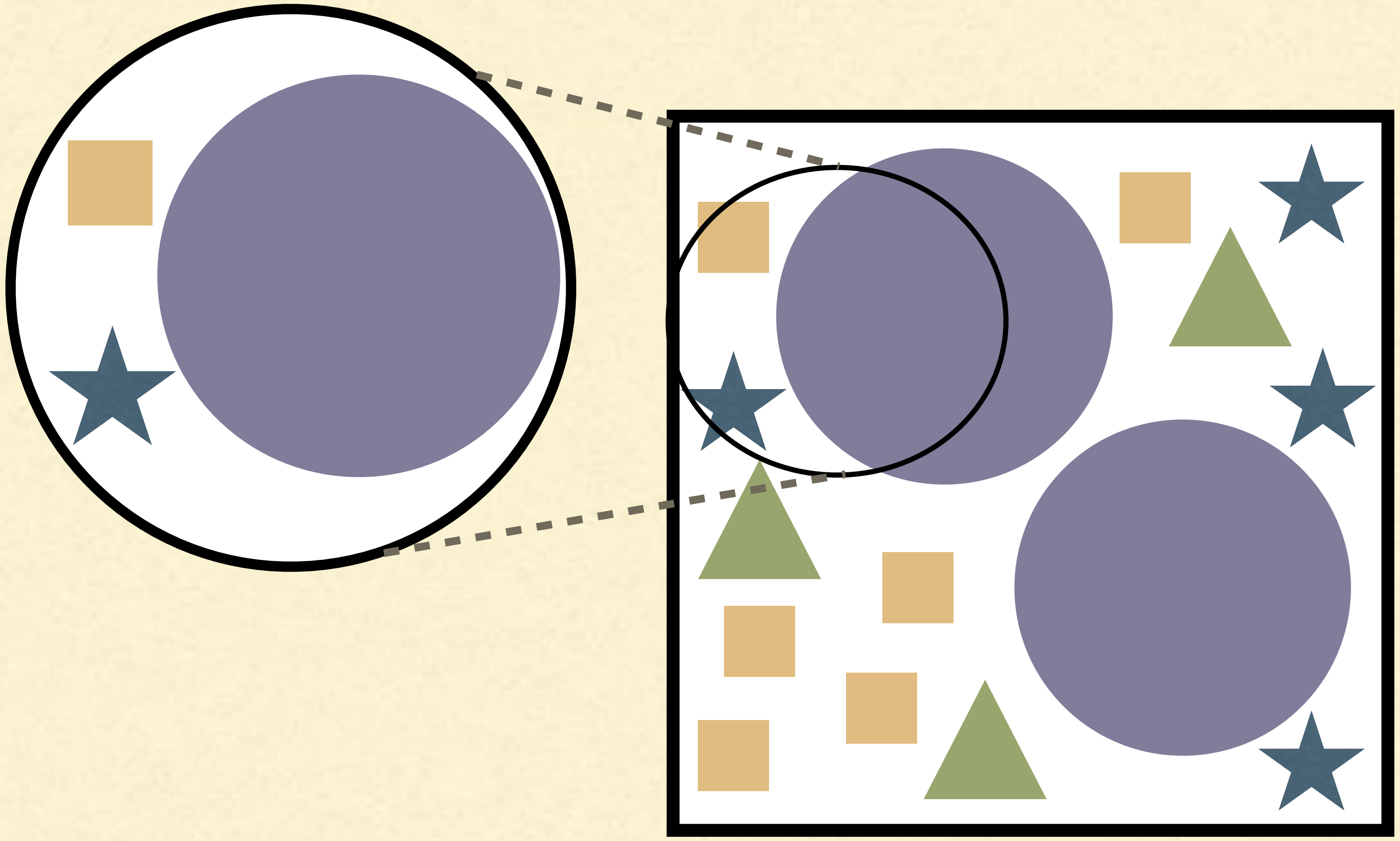
  - *Research interest of ADW: how to estimate diversity* naive

THE ~~"CLASSICAL"~~ APPROACH

- Substitute the observed abundances $\hat{p}_1, \ldots, \hat{p}_c$ for the unknown, true abundances p_1, p_2, …, p_c and pretend nothing happened
  - e.g. Estimate the richness with: $c = \#\{i : \hat{p}_i \neq 0\}$
  - e.g. Estimate the Simpsons index: $\sum_{i=1}^{c} \hat{p}_i^2$

ONE PROBLEM (OF MANY)

- Species richness: plug-in estimate *underestimates*
- Simpson: estimate *overestimates*
- ~~Need new indices~~
- Need new estimators



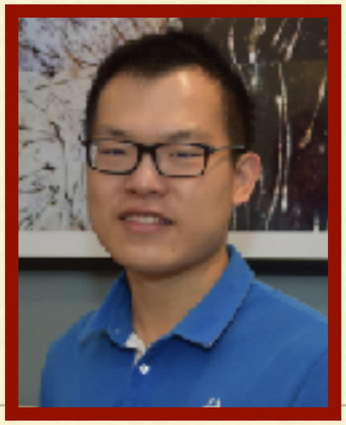
HOW TO FIX

- 2 things are wrong here:
  - The bias (under/overestimation)
  - The variance (how big are the error bars — you'll never be exactly right)

SPECIES RICHNESS

- The "species problem": how many species were missing from the sample
- Idea
  - If many rare species in sample, likely there are many missing species
  - If few rare species in sample, likely there are few missing species
  - Use data on rare species to predict # missing species

SPECIES RICHNESS



Kendrick Li

Alex Paynter

CATCHALL
species richness estimation in R

BREAKAWAY

- CatchAll: mixed Poisson models
  - stable, restrictive, hard to use
- breakaway: non-mixed Poisson models
  - Higher variance, flexible models, in R

SPECIES RICHNESS ESTIMATION

- Good options
  - breakaway::breakaway(); QIIME2 breakaway plug-in
  - breakaway::chao_bunge()
  - breakaway:: objective_bayes_*()
  - CatchAll
- Bad options
  - QIIME2: chao1; scikitbio…
  - R:vegan::…

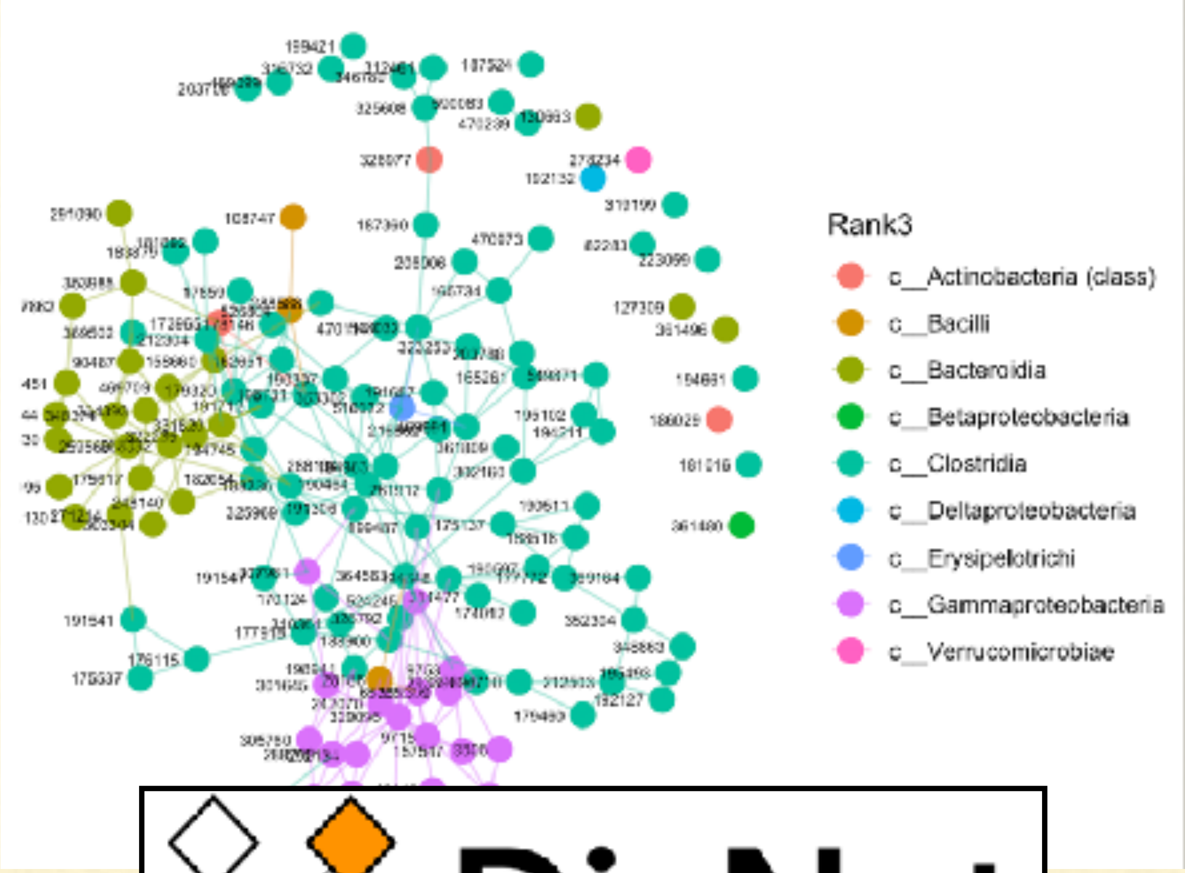
BREAKAWAY

qiime2

Pauline Trinh
(Q2 wizard)

ALPHA DIVERSITY: SHANNON DIVERSITY, SIMPSON, ETC.

- Slightly different approach:
  - Share strength across multiple samples to estimate C and $p_1, p_2, \ldots, p_C$, then use network models to get variance



DivNet

DIVNET

Bryan Martin

Pauline Trinh

- This idea works for estimating any diversity index (**α or β**) that is a function of relative abundances
- It can also be used to estimate any diversity index that is a function of the tree

github.com/adw96/DivNet

DivNet

Coming soon...

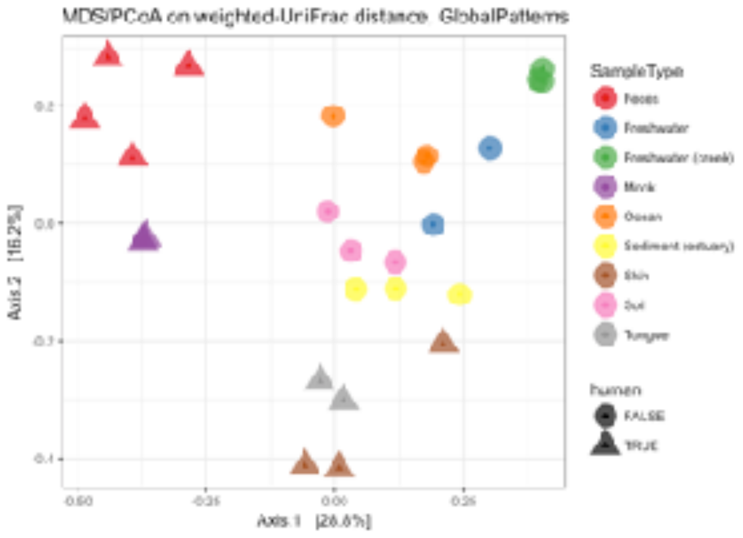
PhyloDivNet

BETA DIVERSITY

- Community 1: $p_1^{(1)}, p_2^{(1)}, \ldots, p_C^{(1)}$; Community 2: $p_1^{(2)}, p_2^{(2)}, \ldots, p_C^{(2)}$
  - β-diversity parameters are usually distances between compositional vectors
- Bray-Curtis: $\beta_{BC} = 1 - \sum_{i=1}^{C} \min(p_i^{(1)}, p_i^{(2)})$
- Jaccard: $\beta_J =$ % taxa not shared
- UniFrac: Weights phylogeny

DIVERSITY: HYPOTHESIS TESTING

- Sometimes diversity is analysed as an exploratory tool
  - e.g., ordination



- Other times you want to do inference
  - e.g., H_0: two communities have zero dissimilarity
  - e.g., H_0: communities A & B have same dissimilarity as communities A & C

HYPOTHESIS TESTING FOR DIVERSITY

- Common approach: PERMANOVA
- Critical issue: adjust for different resolution
  - Good solution = use error bars
    - breakaway::betta(); DivNet::testDiversity
  - (Bad solution = rarefy)

VARIANCE AND HYPOTHESIS TESTS

- Why is estimating variance important?
- Hypothesis testing
- Most hypothesis tests take the form

$$\frac{\text{estimate}}{\text{standard error}} \sim N(0,1)$$

VARIANCE AND HYPOTHESIS TESTS

- If your estimate was 1, and the (true) standard deviation is 1…

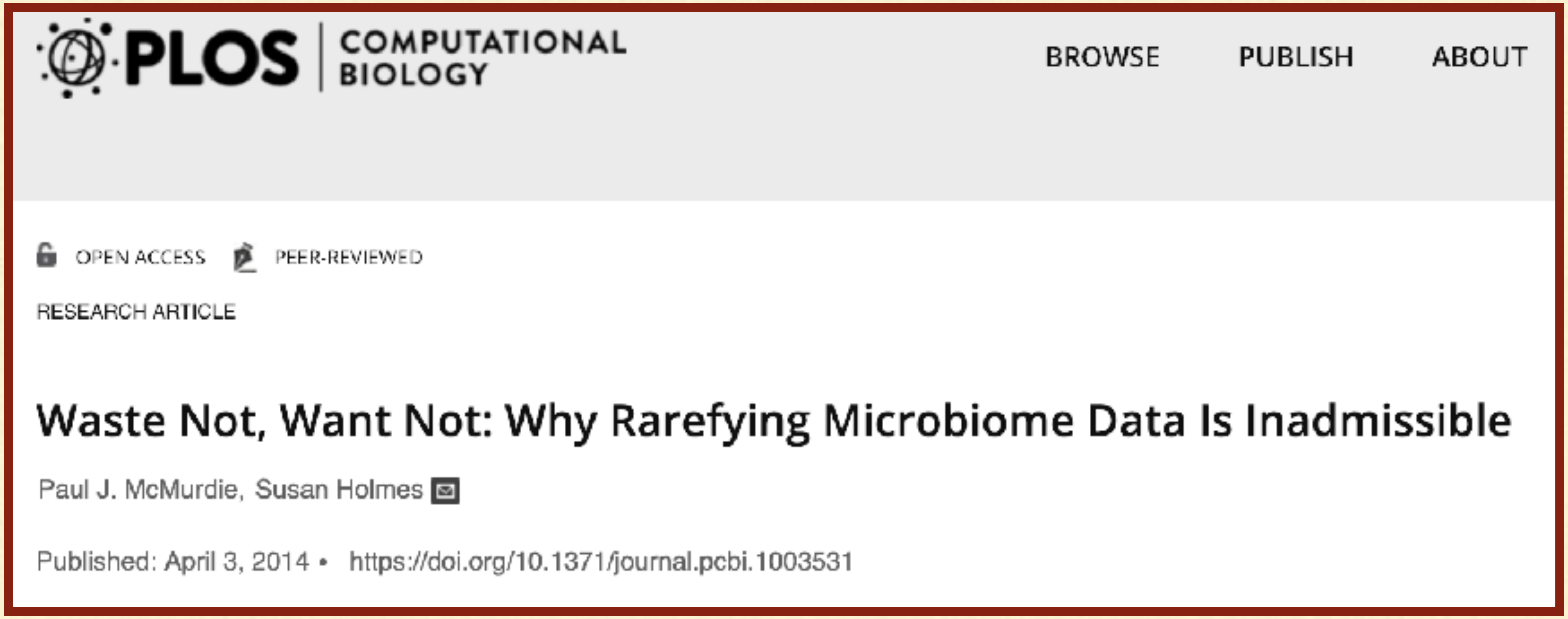
STANDARD ERROR	1	0.5	0.33	0.25
P-VALUE	0.318	0.046	0.002	<0.001

BIAS AND DIVERSITY

- Alternative approach that I loathe: rarefaction
- Idea:
  - Discover more diversity with more sequencing
  - Can’t directly compare samples with different depths
  - Randomly throw away reads until all samples have same depth
- Better idea:
  - Statistical estimation accounts for different sequencing depths!

BIAS AND DIVERSITY

- Alternative approach that I loathe: rarefaction



RESEARCH ARTICLE

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

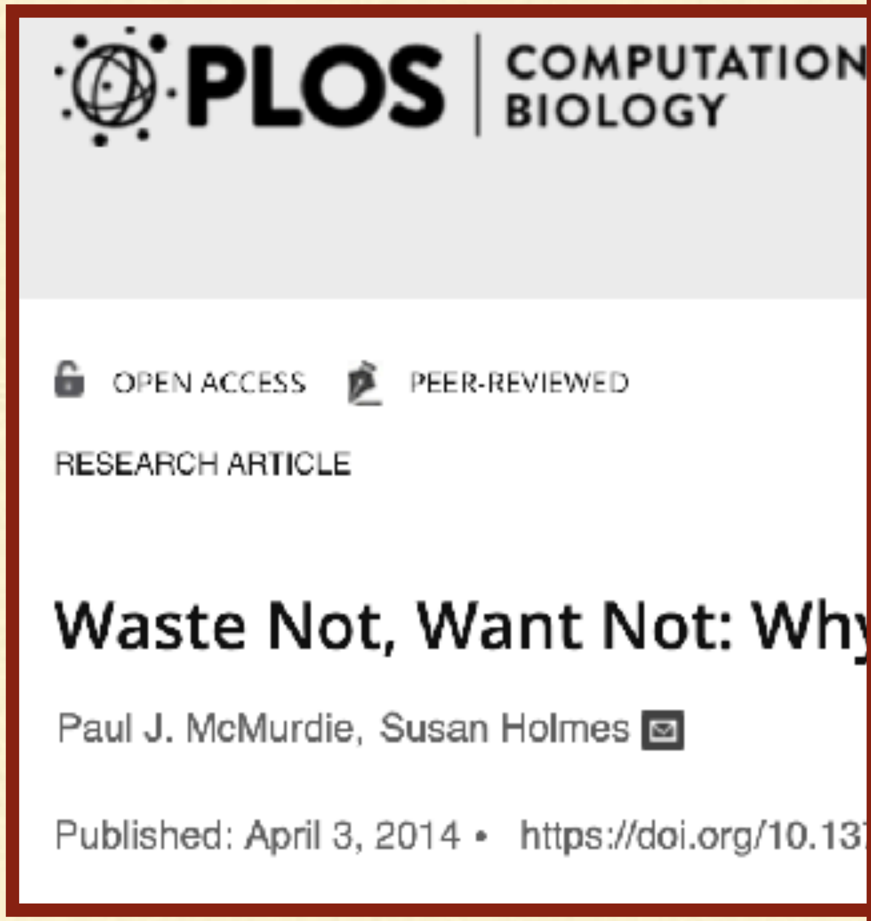
Paul J. McMurdie, Susan Holmes

Published: April 3, 2014 • https://doi.org/10.1371/journal.pcbi.1003531

- Statistical estimation accounts for different sequencing depths!

BIAS AND DIVERSITY

- Alternative approach



PLOS | COMPUTATIONAL BIOLOGY

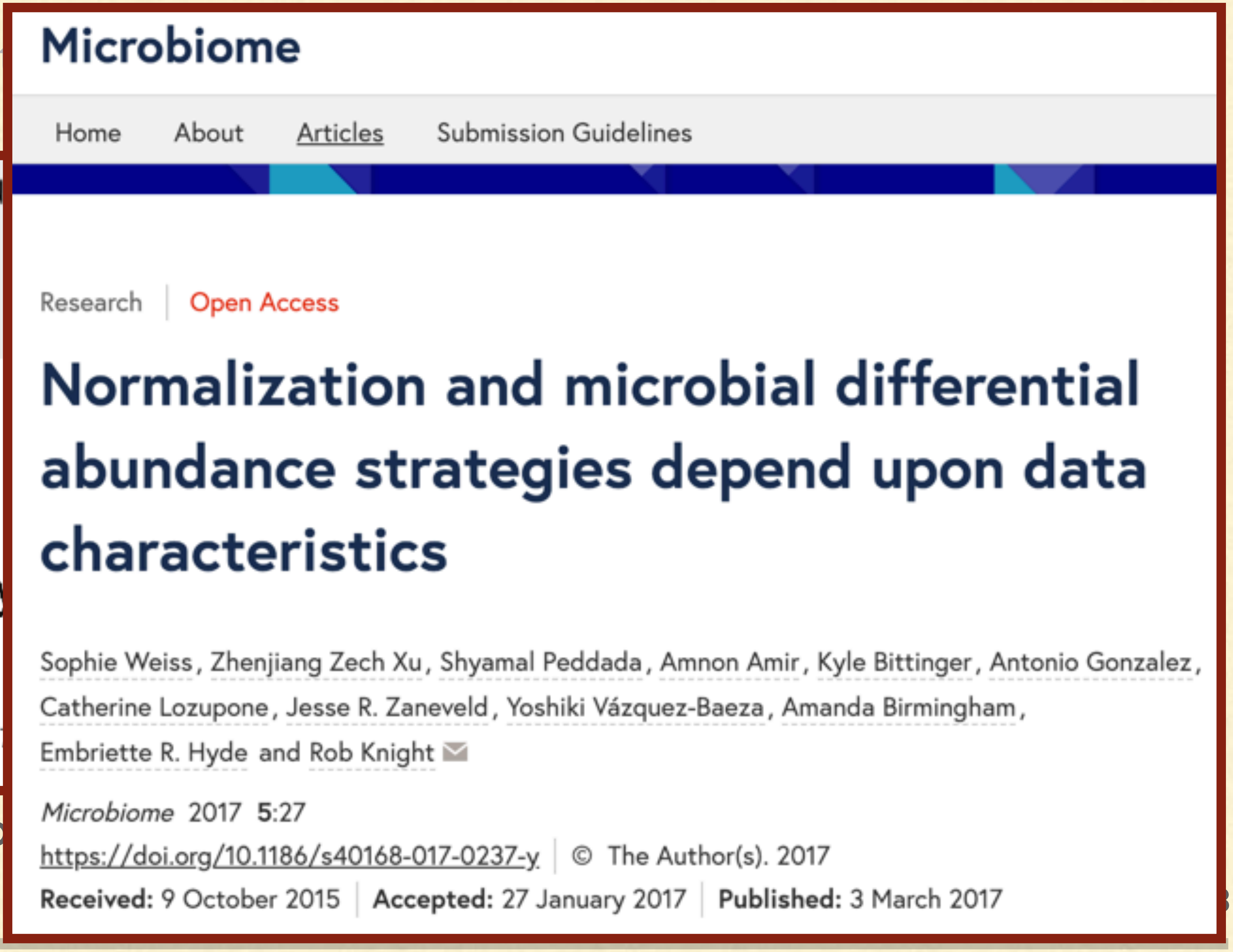
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Waste Not, Want Not: Why

Paul J. McMurdie, Susan Holmes

Published: April 3, 2014 • https://doi.org/10.13



Microbiome

Home About Articles Submission Guidelines

Research | Open Access

Normalization and microbial differential abundance strategies depend upon data characteristics

Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde and Rob Knight

Microbiome 2017 5:27

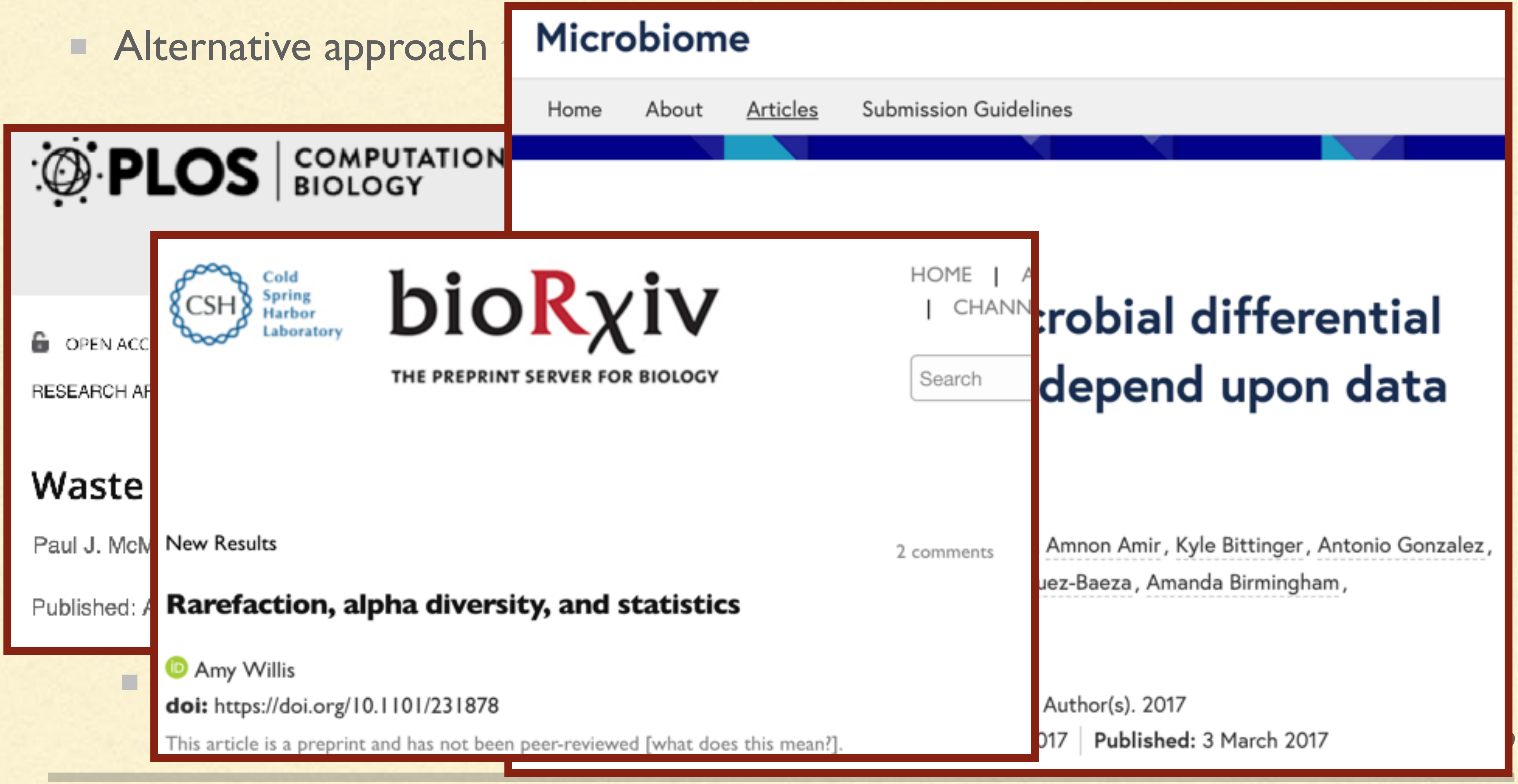
https://doi.org/10.1186/s40168-017-0237-y | © The Author(s). 2017

Received: 9 October 2015 | **Accepted:** 27 January 2017 | **Published:** 3 March 2017

- Statistical estimatio

BIAS AND DIVERSITY

- Alternative approach



PLOS | COMPUTATION BIOLOGY

OPEN ACC

RESEARCH AR

Waste

Paul J. McM

Published: A

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | A | CHANN

Search

New Results

2 comments

Rarefaction, alpha diversity, and statistics

Amy Willis

doi: https://doi.org/10.1101/231878

This article is a preprint and has not been peer-reviewed [what does this mean?].

Microbiome

Home About Articles Submission Guidelines

robial differential depend upon data

Amnon Amir, Kyle Bittinger, Antonio Gonzalez, uez-Baeza, Amanda Birmingham,

Author(s). 2017

017 | **Published:** 3 March 2017

DIVERSITY

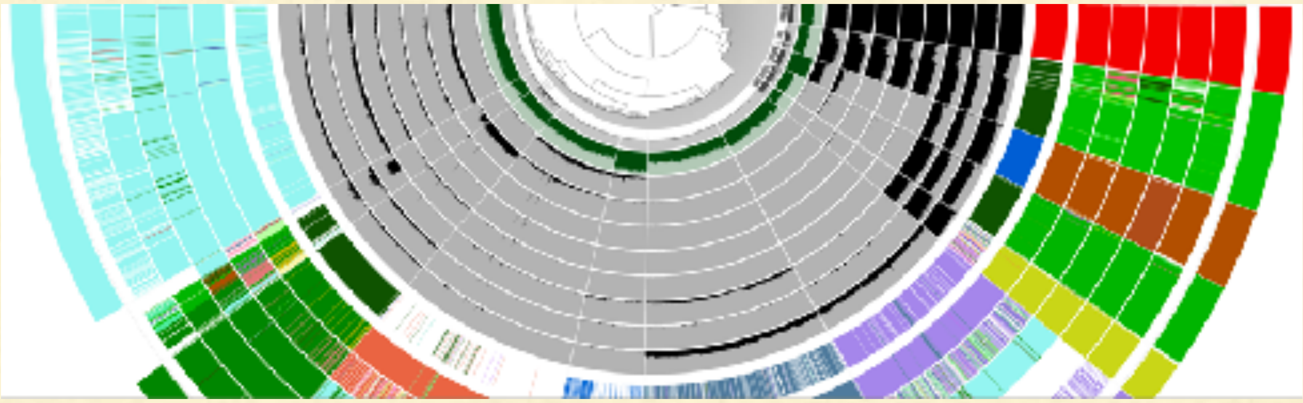
- Very useful summary of (high-dimensional) compositional data… in many settings!
- Diversity is a useful *first question*
- Drawback: Changes in diversity don't indicate what composition(s) are changing….

ABUNDANCE

- How do we talk about changes in the amount of something?
  - Fraction of environments with a characteristic
  - Relative abundance: proportions only
  - Relative abundance: count data
  - Absolute abundance (same tools for DE analysis e.g., DESeq2, edgeR)

ABUNDANCE: ENRICHMENT

enrichment of genes/functions/pathways: higher presence in one group vs another group



- Need to know: anvi'o
  - Amazingly powerful tools for lots of things, including WGS
  - Fantastic workflows and tutorials for all things WGS

merenlab.org/

Image credit: *anvi'o* development team

ENRICHMENT

- If samples the genomes came from were observed independently, the ***enrichment analysis*** in anvi’o gives a hypothesis test for enrichment

- Key points: adjusts for different numbers of genomes in each group; hypothesis testing & false discovery control

RELATIVE ABUNDANCE: COUNTS

- Observe $\mathbf{W_i}$ counts out of $\mathbf{M_i}$ total counts for samples i=1…**n**
- For each sample have $\mathbf{X_i}$, information about treatment/disease/source environment
- Goal: Hypothesis test for changes in mean relative abundance with $\mathbf{X_i}$
    - Options: CORNCOB, LEFse, ANCOM, MaAsLin, gneiss, DESeq2, ALDEx2, many others

CORNCOB

COmpositional RegressioN for Correlated Observations with the Beta-binomial

Daniela Witten, UW

- Latent variable model & hypothesis testing for **relative abundance**
- Adjusts for different depths
- Flexible model: individual microbes correlated
- Bonus: Mean *and variance* ("dysbiosis") testing

Bryan Martin

CORNCOB AND DESEQ2

corncob

- Designed for marker gene (compositional) data
- Models relative abundance, overdispersion, and correlation parameters
- Different structure for different taxa
- Uses within-taxon correlation to model zeros

DESeq2

- Designed for RNAseq (different data structure)
- Tests changes in abundance
- Constrained dispersion
- Individual microbes are assumed independent

OTHER ANALYSIS APPROACHES

- Networks
  - Can be very interesting… if your data is very good
- Source tracking
  - Can be very interesting… if your data is very good
- Many, many others

COMMON IDEA

- "I didn't collect the data that I really wanted, so I will use what I have to try to reconstruct the data that I really wanted"
  - e.g., microbial concentration (16S qPCR x 16S rel abundance)
  - e.g., functional information (PiCRUST)
  - **Very very serious caveats! Use with extreme caution!**

CONSIDERATIONS FOR MICROBIOME SCIENCE

- Too many microbiome papers list significant associations
    - Taxon A, B C; genes X, Y, Z are significantly higher/lower abundance in [folks with disease D]
    - Observations are interesting, often unhelpful
        - Does the microbiome cause the disease, or the other way around?
        - Studies involving (intelligent) interventions can help
            - e.g., paired data/longitudinal sampling

EXPERIMENTAL DESIGN

The population that you want to study may not be the population that you get to study

- Before undertaking a microbiome study, think carefully about
    - the question you want to answer,
    - the data you have access to, and
    - the questions you can answer with the data that you have access to

WHAT CAN WE DO?

- Replicate, replicate, replicate
    - Independently repeating the experiment is the gold standard for confirming a result is “real”
- Think critically
- Use plots, not p-values

WHAT ELSE CAN WE DO?

- Be honest
  - Keep all analyses that you ran, not just the final one
- Write down all of the hypotheses that you care about
  - Before doing the experiment
  - Before doing the analysis
- Your university might house a statistician; try to involve them...
  - ...in the entire process, not just calculating p-values

SUMMARY

- Technology: Taxonomic, functional, concentration profiling

- Data cleaning & preprocessing: organising data into biological units (16S = ASVs; WGS = genomes/genes/CAGs)

- Statistical estimation & hypothesis testing

    - Diversity analysis: α, β

    - Abundance analysis: enrichment, proportions (count/proportion), abundance

- and *many other things* that couldn't be fit into this lecture

RESOURCES:
HOW DO YOU STUDY MICROBES?

- Your university probably has a microbiology department
- Your university probably has a statistical consulting service
- STAMPS: Strategies and Techniques for Analyzing Microbial Population Structures at the MBL (Marine Biological Laboratory)
    - Apply by April 19
- The Statistical Diversity Lab @ UW

statisticaldiversitylab.com

DIRECTIONS FOR MICROBIAL ECOLOGY

- Research
  - Reproducibility
  - Calibrating sequencing results with reality
  - Lab goals & wrap up

REPRODUCIBILITY

- Microbiome Quality Control Project
  - Sent same set of samples to 10+ sequencing labs, 8 bioinformatics labs
  - Compared results
  - Question for you: what is the best case scenario?

REPRODUCIBILITY

David Clausen

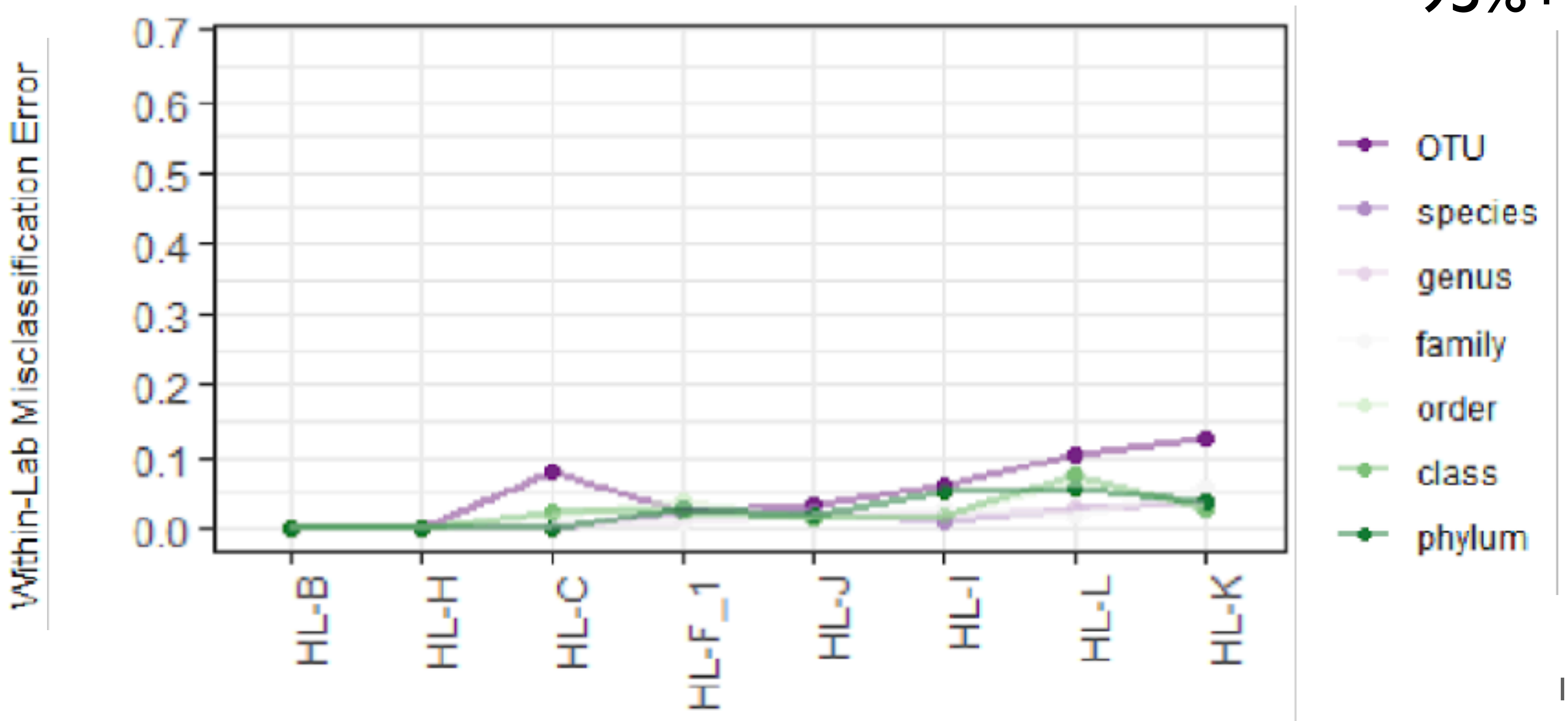
- Reproducibility evaluation
  - Ideal: every lab gets identical results
  - Good enough: Not identical, but consistent ability to discriminate
  - Our qtn: Are technical replicates of Sample A more similar to each other than technical replicates from Sample B?
    - Within lab? Across lab?
    - How likely are results obtained from one lab to be reproduced in another lab?

REPRODUCIBILITY: WITHIN-LAB

David Clausen

What percentage of the time can we determine sample type based on within-lab replicates?

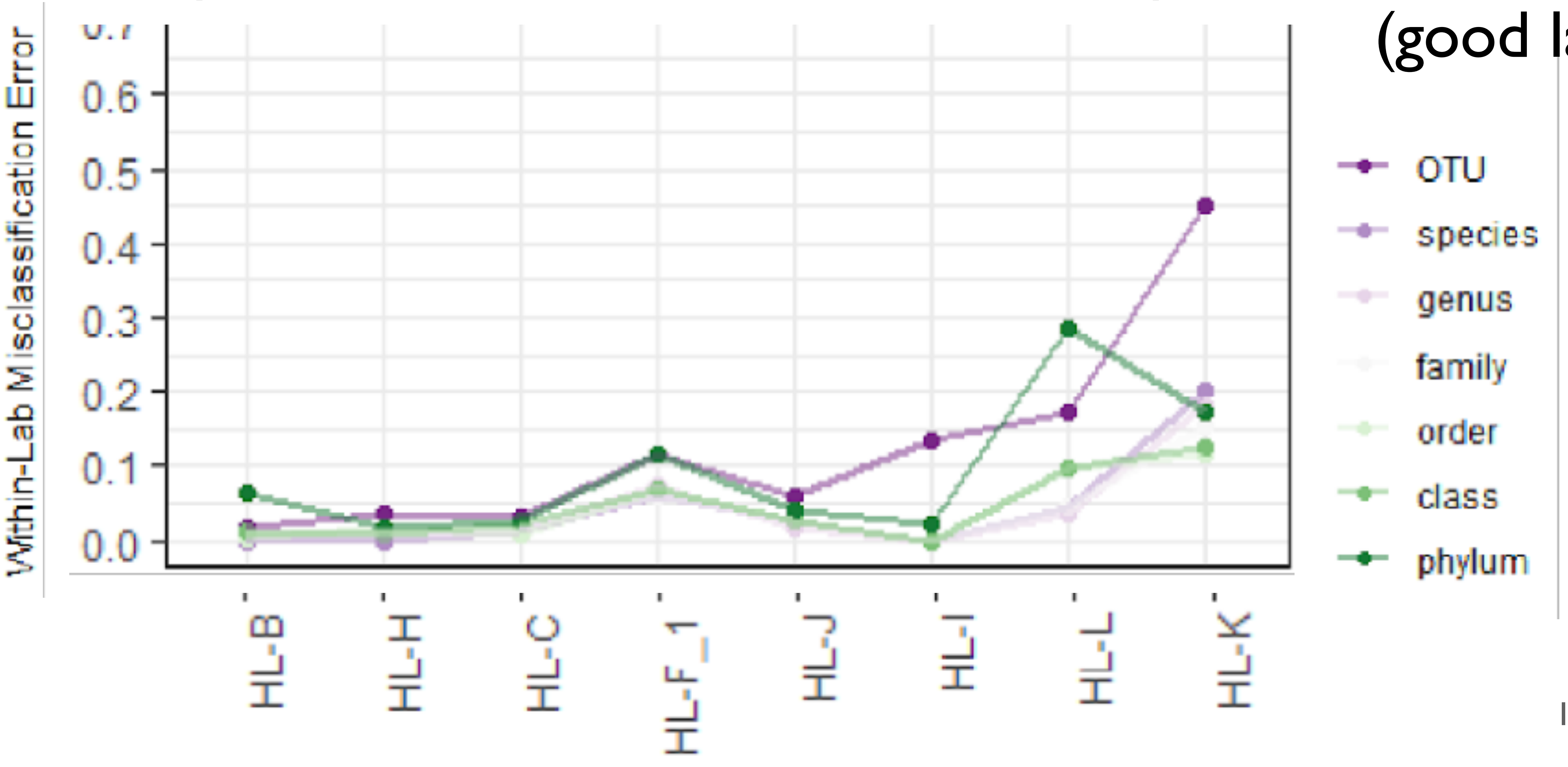
~95%+



REPRODUCIBILITY: WITHIN-LAB

What percentage of the time can we determine sample identifier based on within-lab replicates?

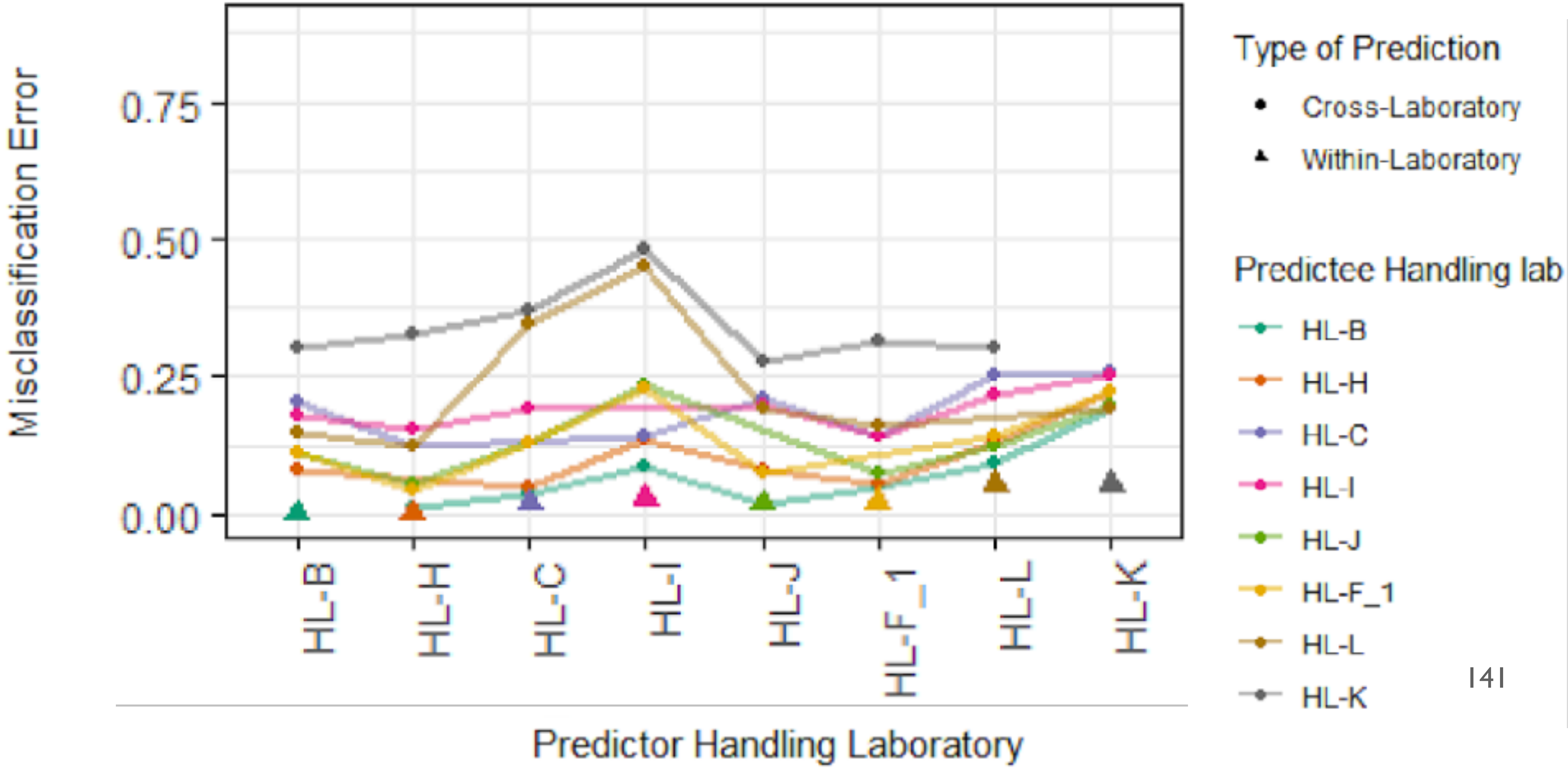
~90%
(good labs)



REPRODUCIBILITY: ACROSS LABS

What percentage of the time can we determine sample type based on another lab's results?

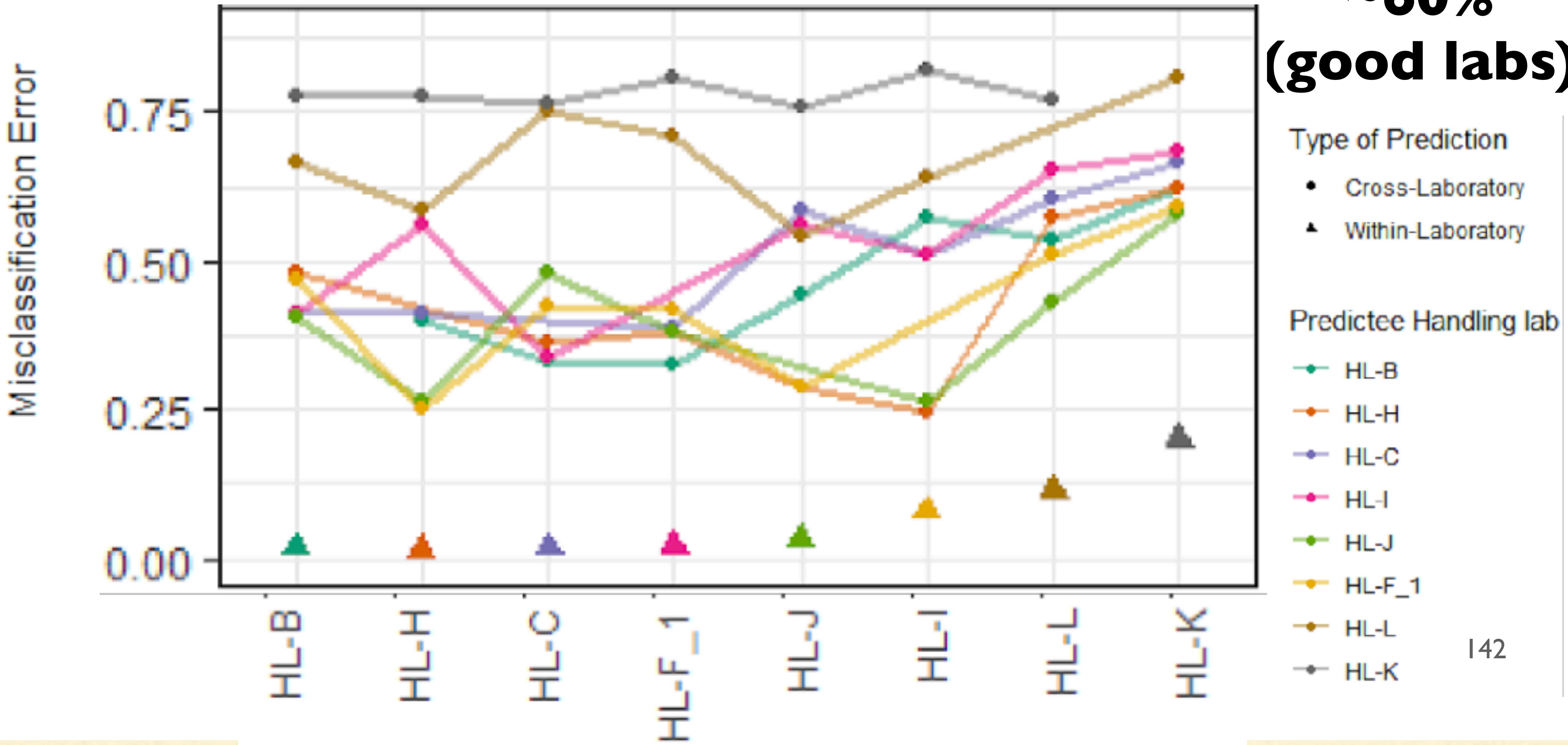
~80-90%



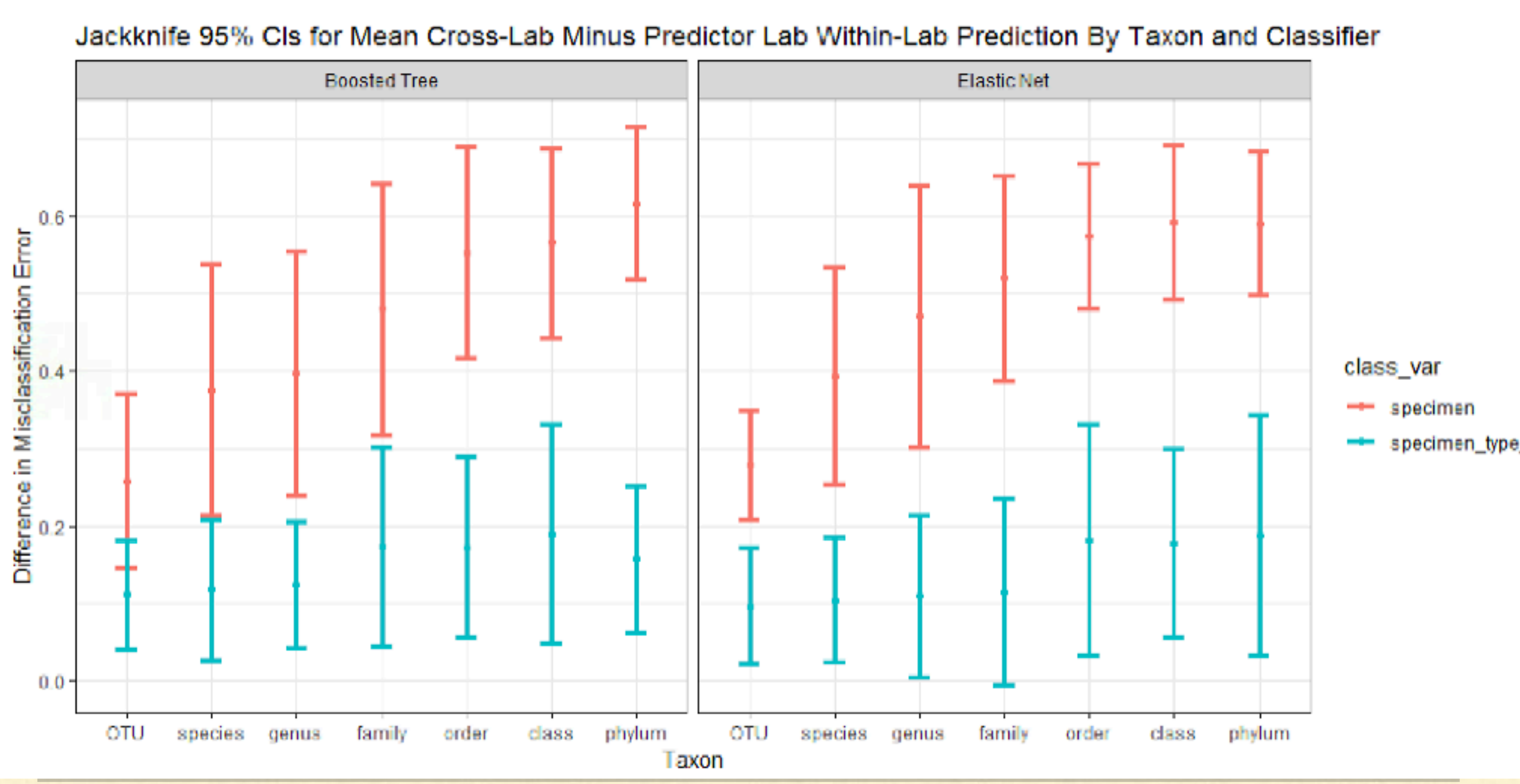
REPRODUCIBILITY: ACROSS LABS

What percentage of the time can we determine sample identifier based on another lab's results?

~60% (good labs)



HOW MUCH WORSE IS REPRODUCIBILITY ACROSS VS WITHIN LABS?



MODELING EFFICIENCY

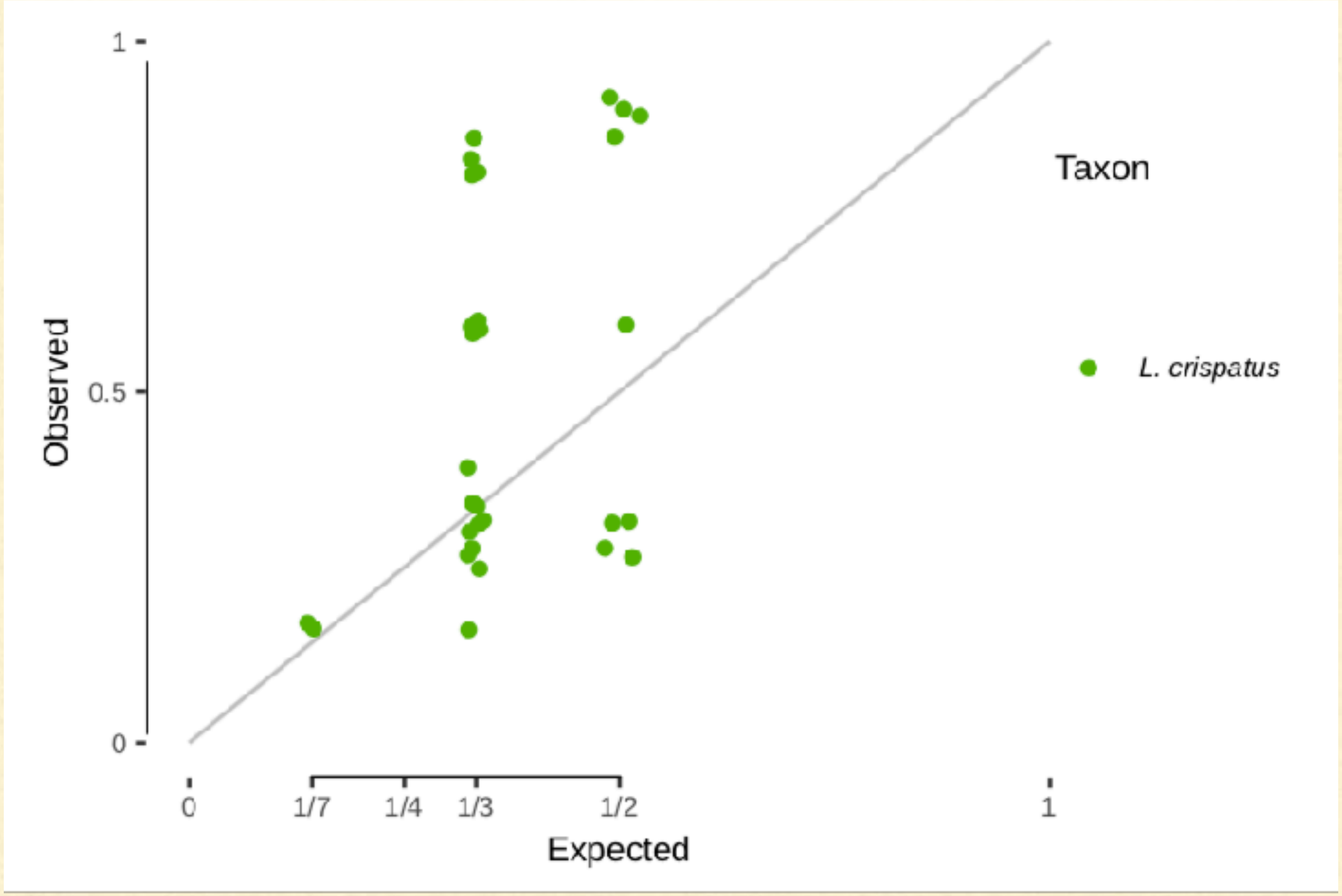
Michael McLaren (NCSU)

Ben Callahan (NCSU)

David Clausen

- Big picture goal: correct cross-lab differences
- Current step: understand how taxon abundances are distorted by sequencing process
- Approach: mock communities!

MODELING EFFICIENCY



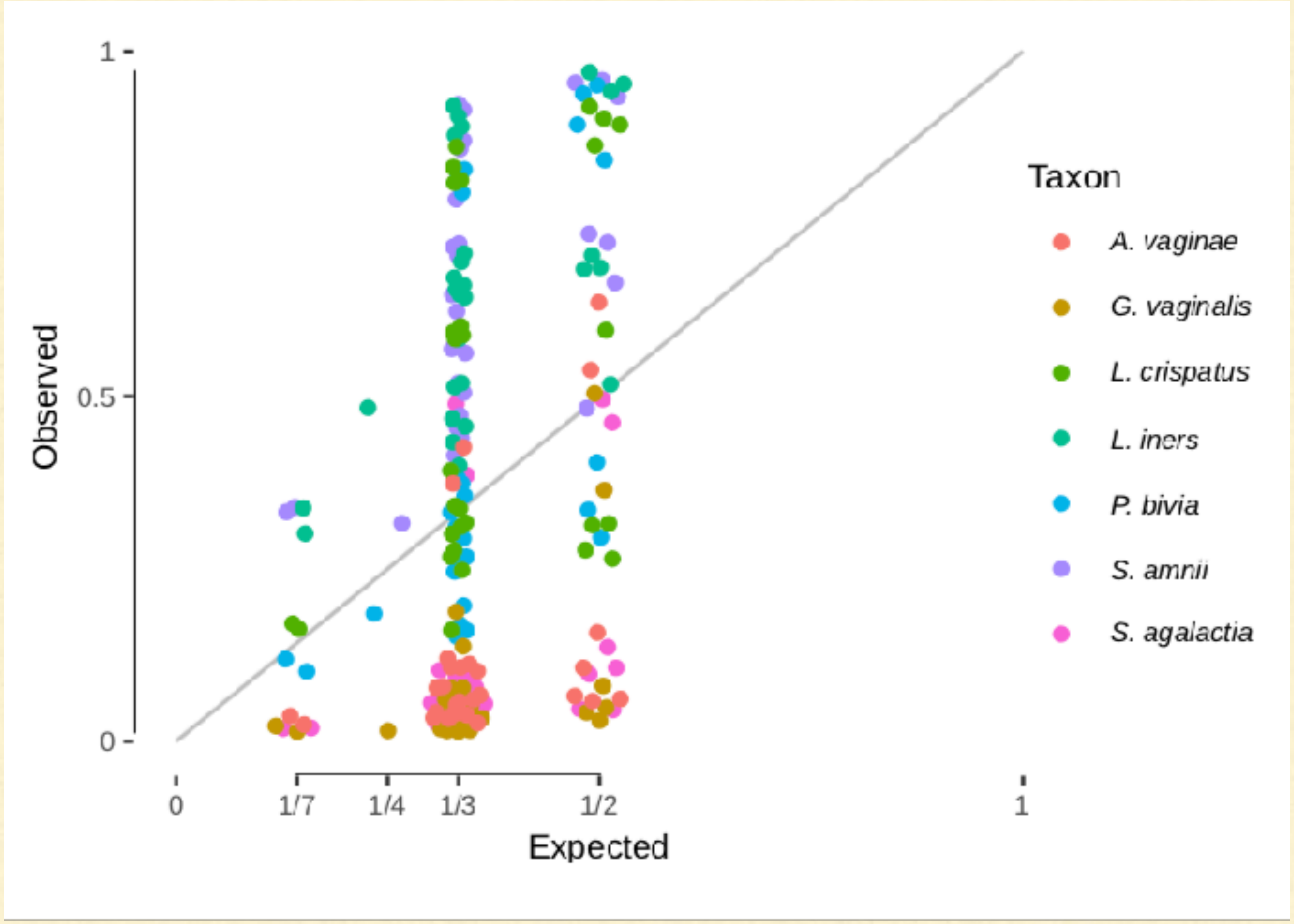
Michael McLaren (NCSU)

Ben Callahan (NCSU)

David Clausen

MODELING EFFICIENCY

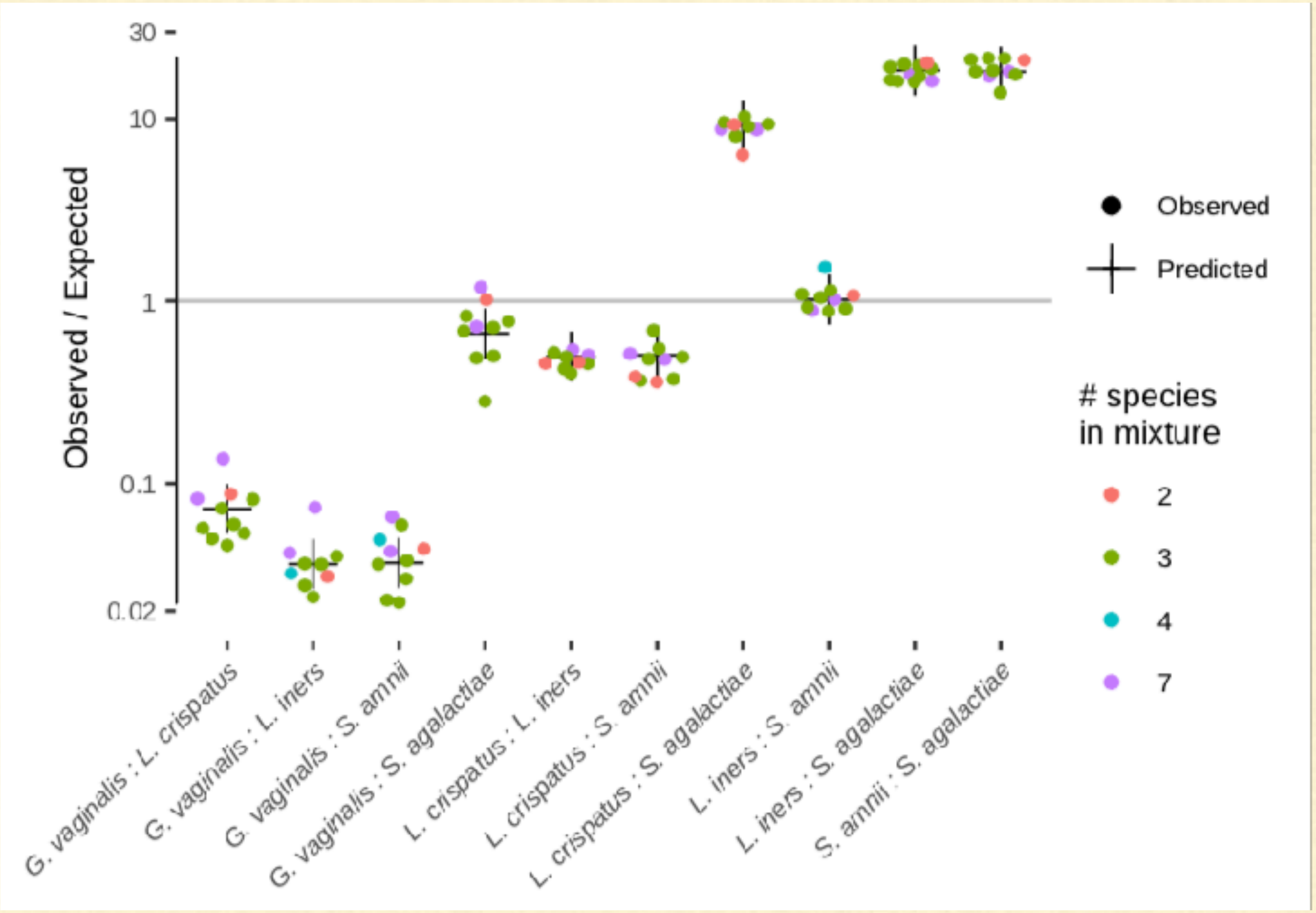
Michael McLaren (NCSU)



Ben Callahan (NCSU)

David Clausen

MODELING EFFICIENCY

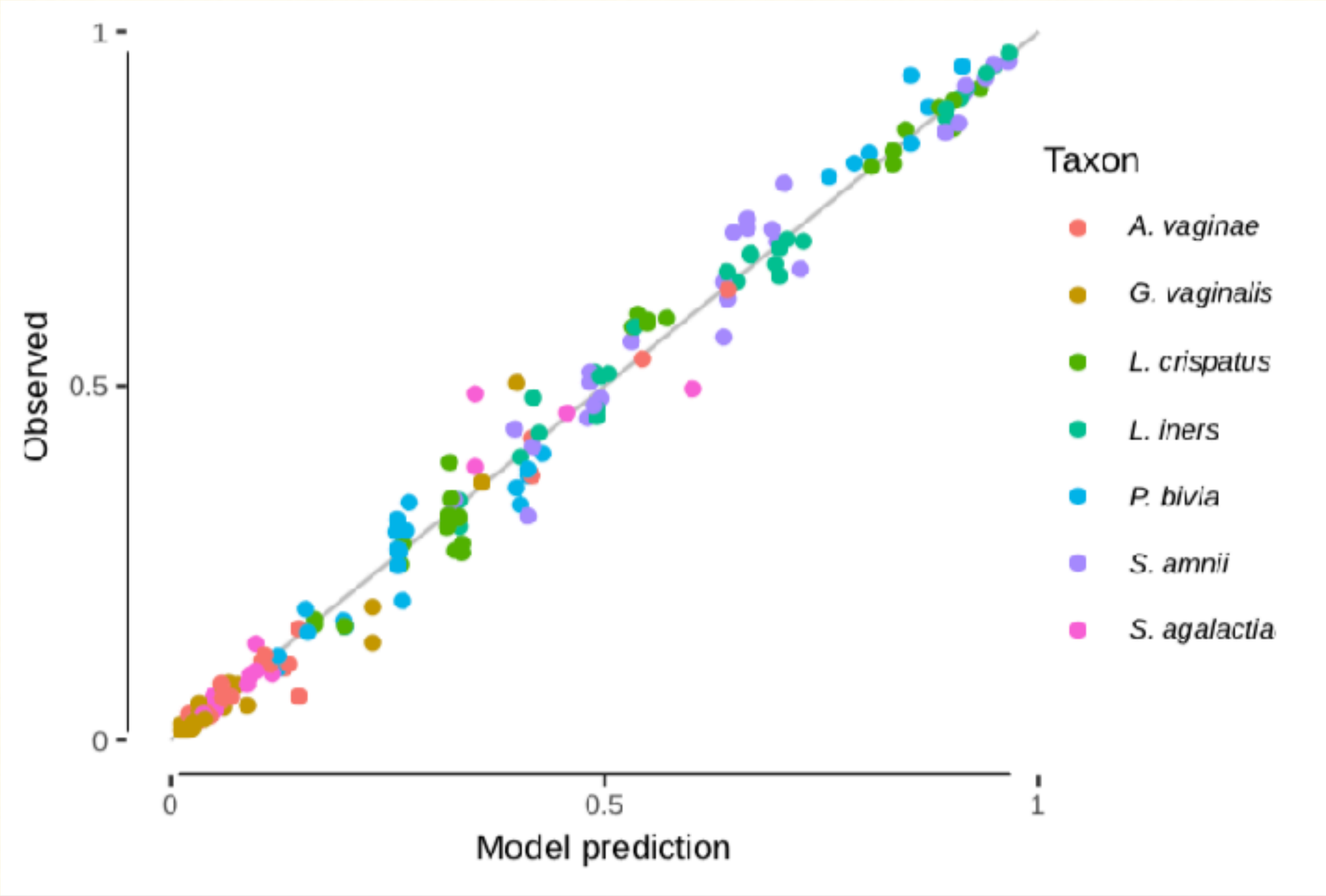


Michael McLaren (NCSU)

Ben Callahan (NCSU)

David Clausen

MODELING EFFICIENCY



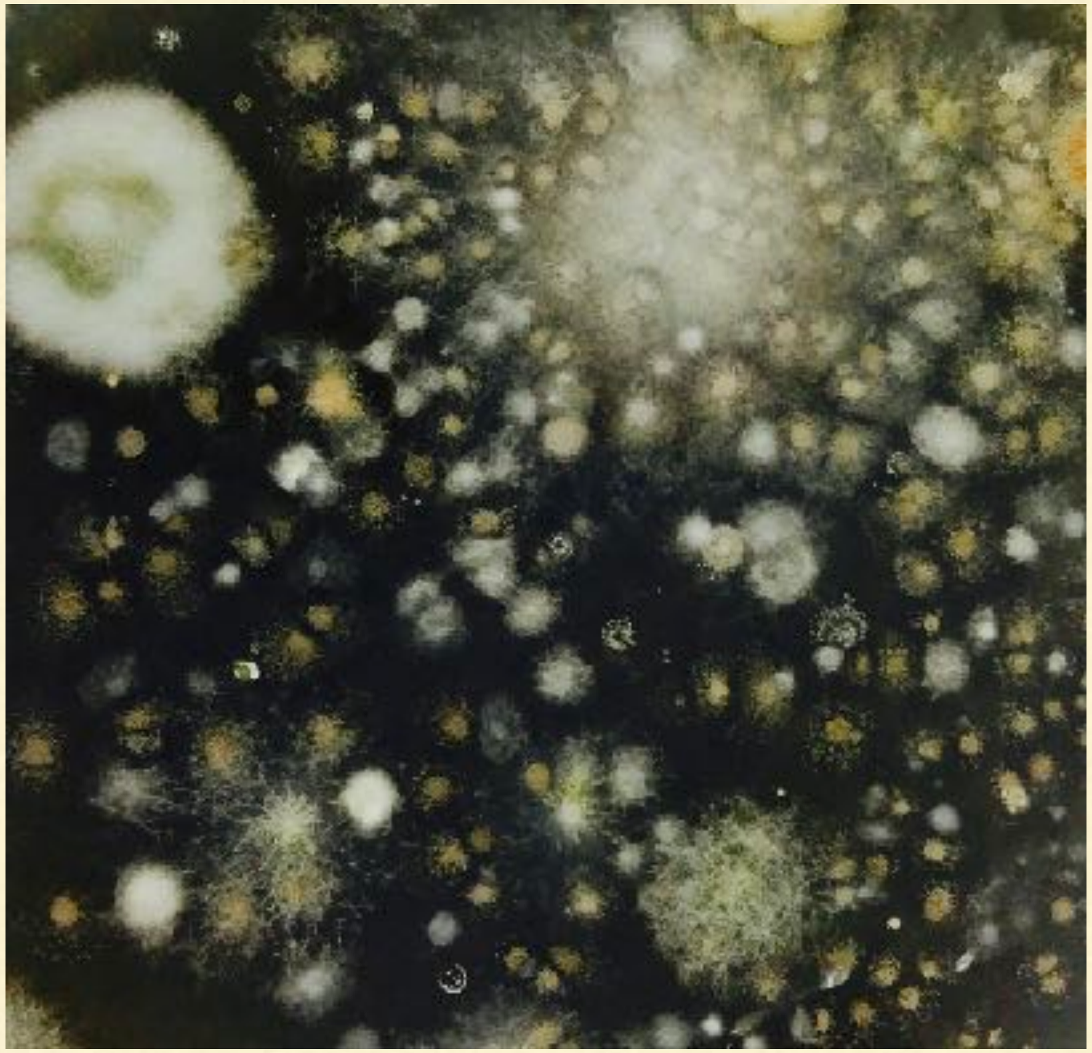
Michael McLaren (NCSU)

Ben Callahan (NCSU)

David Clausen

STATISTICAL DIVERSITY LAB GOALS

- Develop statistical and computational tools for reproducible microbiome science
  - Address model misspecification
  - Make use of existing data (yours and others')
  - Model sequencing process and errors
  - Outreach: why statistical estimation and good statistical practice matters



MICROBIOME DATA & ANALYSIS

Research Group: Statistical Diversity Lab

PI: Amy D Willis PhD, Assistant Professor, Department of Biostatistics, UW

@AmyDWillis adwillis@uw.edu

Photo credit: T.D. Berry, Whitman lab, UW Madison