ibest

Initiative for Bioinformatics & Evolutionary Studies

Combining Bacterial Fingerprints

James A. Foster
Biological Sciences
Bioinformatics & Computational Biology
Institute for Bioinformatics & Evolutionary STudies (IBEST)

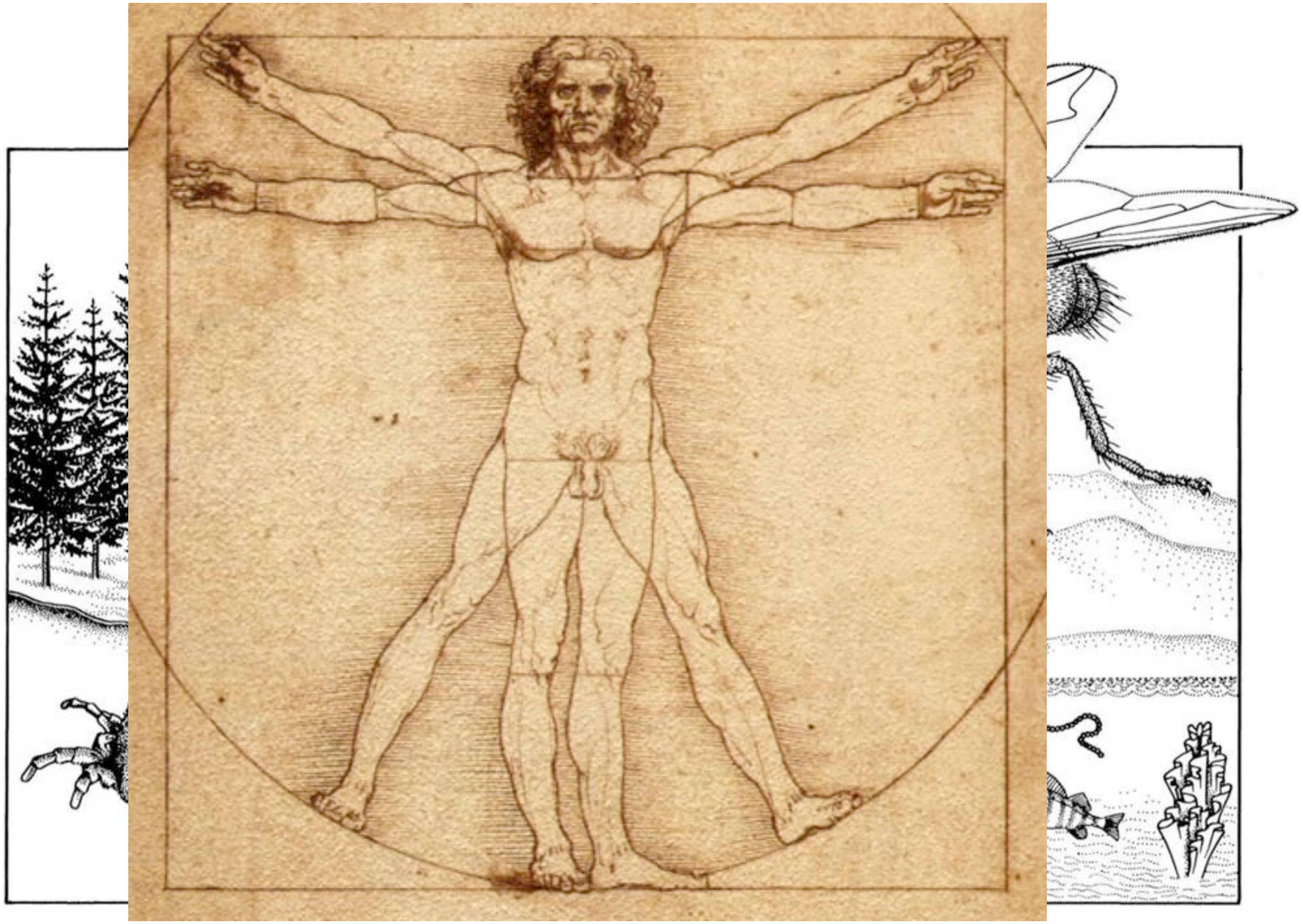
5 May 2014

University of Idaho

Who's world is this?

ibest
Initiative for Bioinformatics & Evolutionary Studies

We are one out of 2 million named species (5-100m est.)



Ignores 1 Billion species Bacteria!

Initiative for Bioinformatics & Evolutionary Studies

Million years on earth

3600 Bacteria

500 Fish

130 Flowers

190 Mammals

0.2 Humans

Bacterial inventions

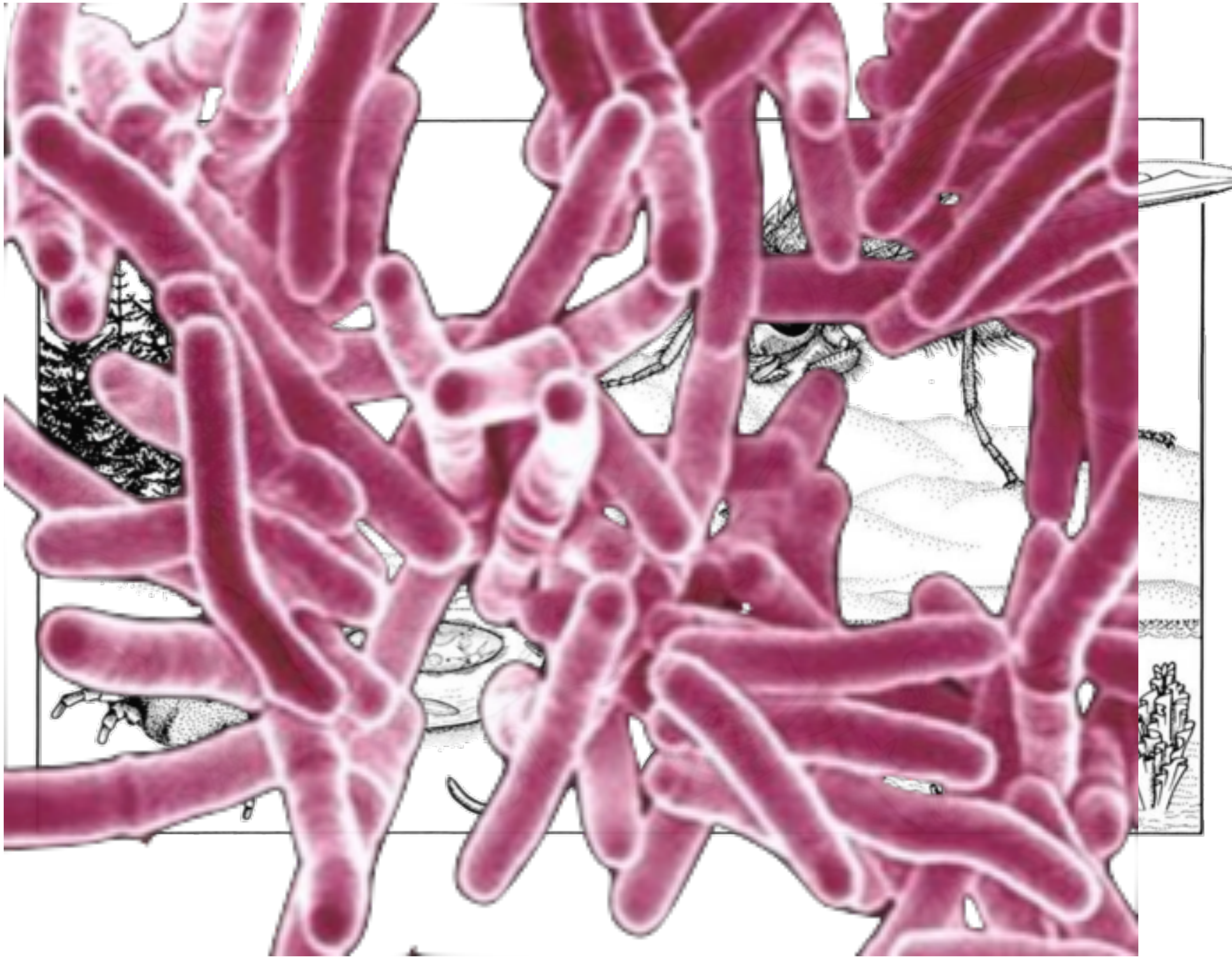
Oxygen

Photosynthesis

Nitrogenation

the Nucleus

Mitochondria



University of Idaho

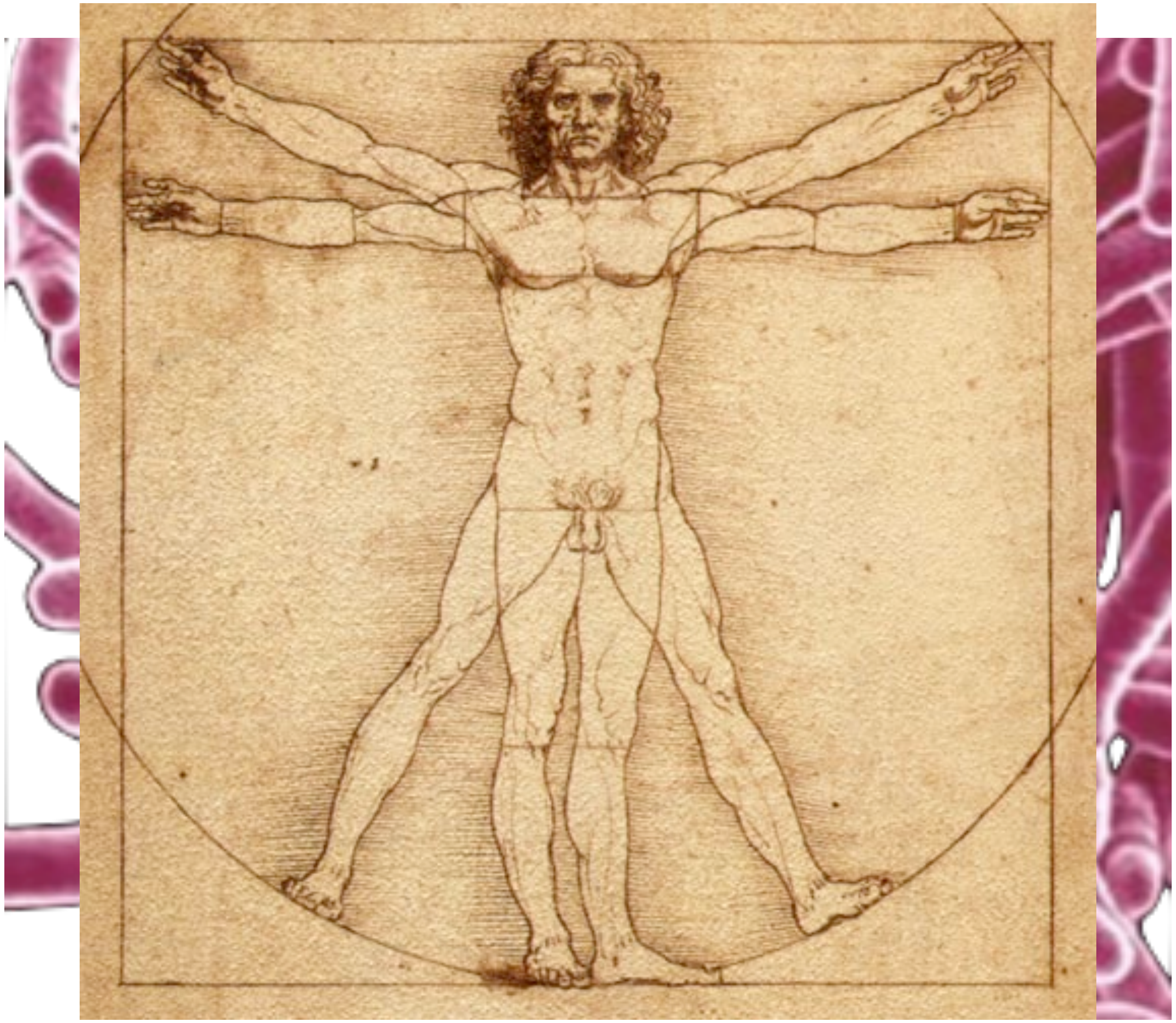
Whose world are you?

Initiative for Bioinformatics & Evolutionary Studies

90% of the cells in your body are bacterial

≥ 99.99% of the gene products in your body are bacterial

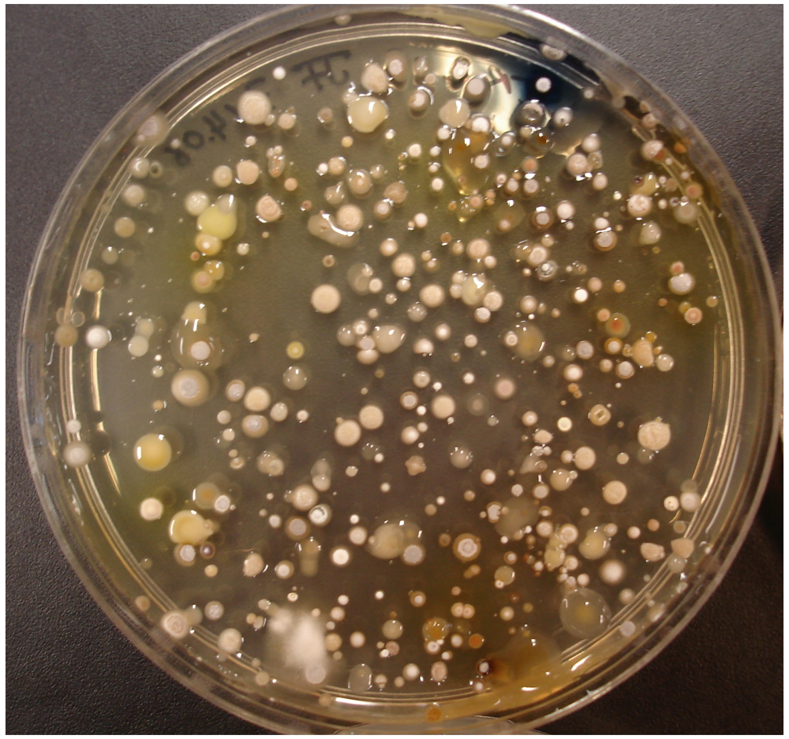
Biologically, you are mostly a bacterial ecosystem



The dark microbiome

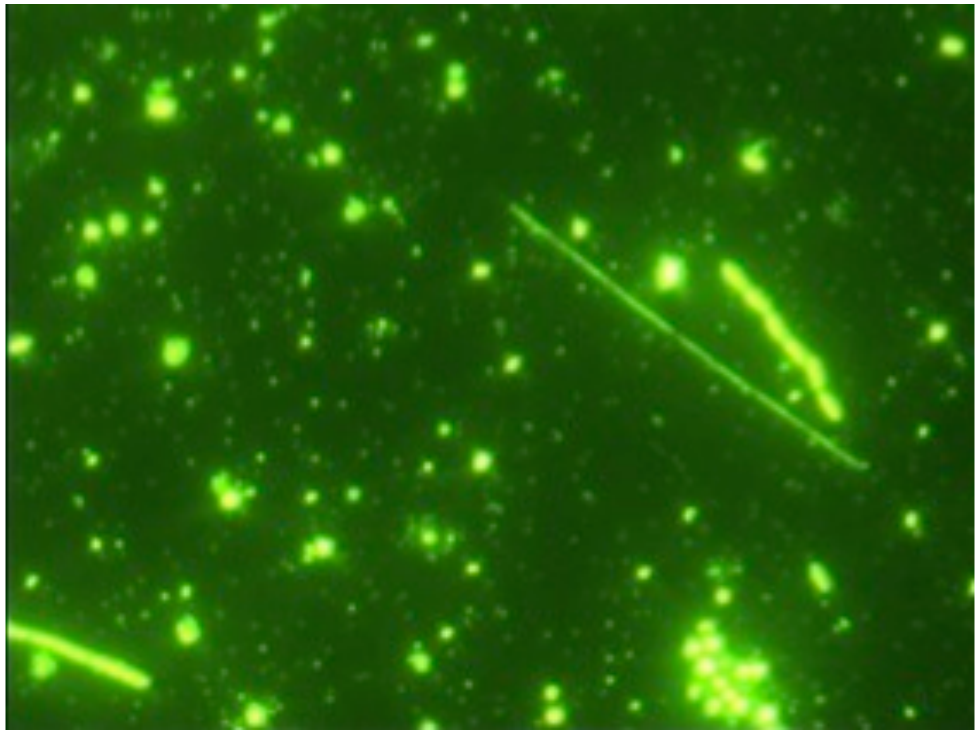
Initiative for Bioinformatics & Evolutionary Studies

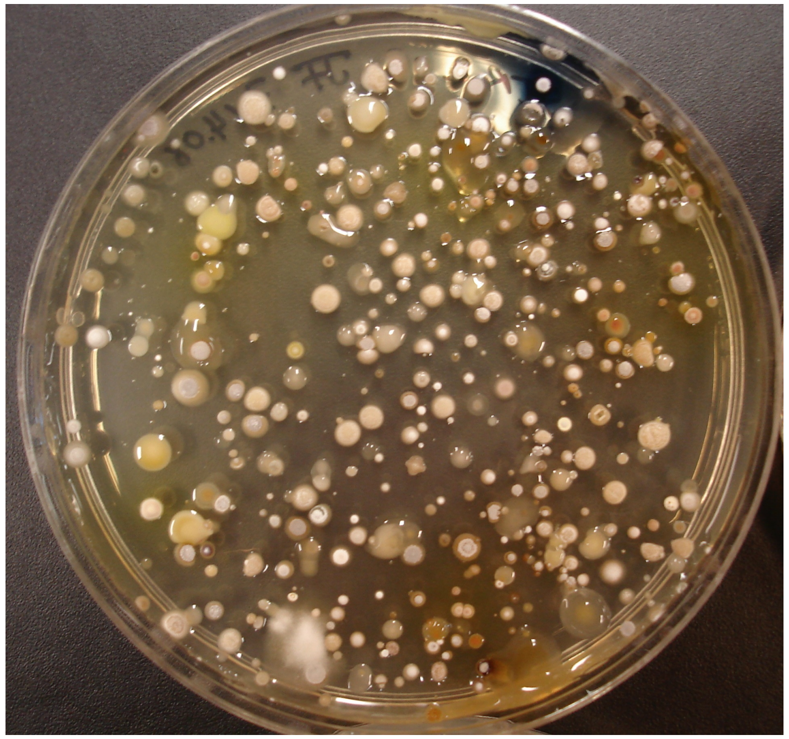
- Up to 1 Billion species, about 5,000 known
  - 10K species in a gram of soil
  - 1cm^2 intestine: bacteria > all humans, ever
  - Half of all Phyla undiscovered (human vs sponge)
- Great plate count anomaly: approx. 97% of bacteria cannot be grown



The dark microbiome

- Great plate count anomaly: approx. 97% of bacteria cannot be grown
- What we do know is *highly biased*

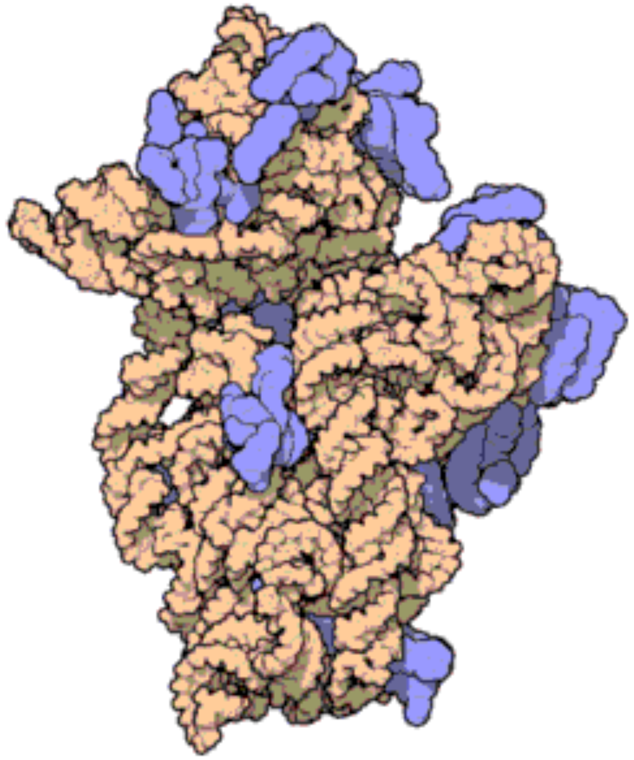


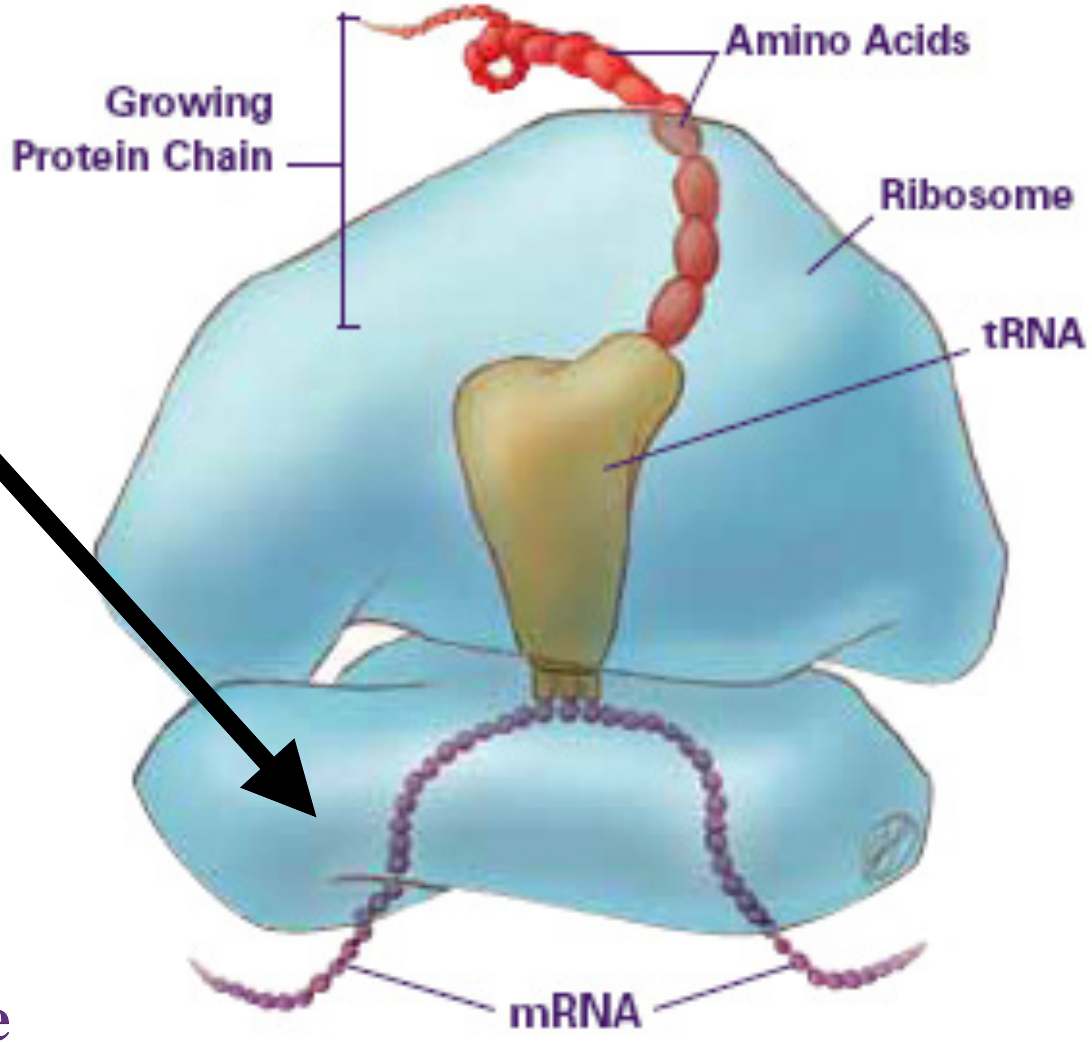


How to count the invisible

All bacteria:

- Translate genome to proteins
- Using *ribosome*: RNA + proteins
- Small subunit attaches first
- Coded by 16S gene





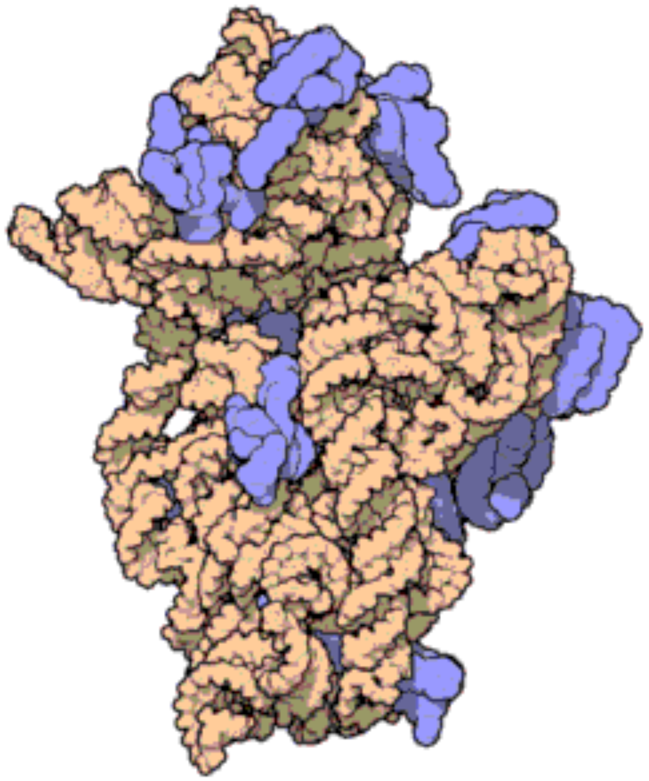
Genome

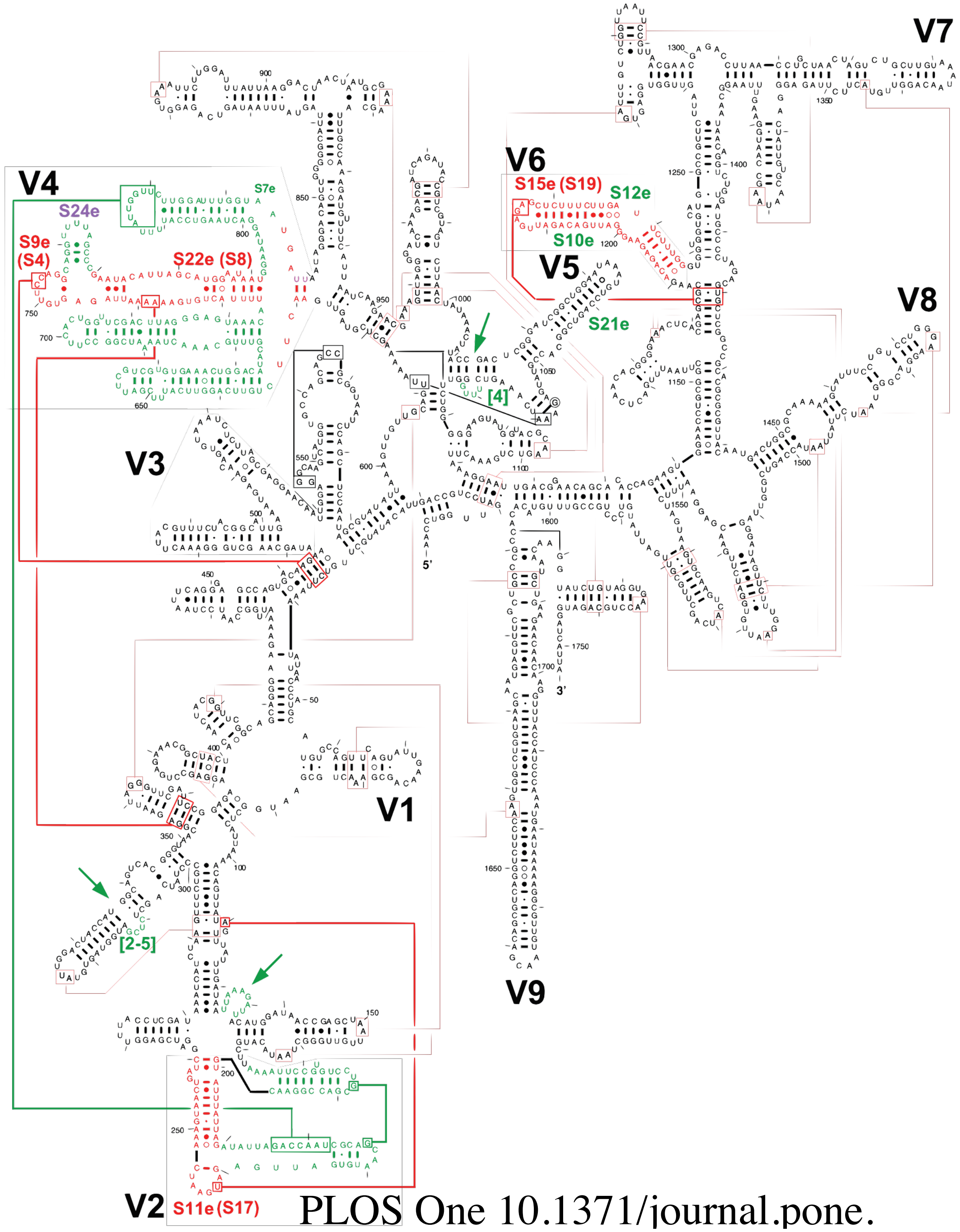
16S small subunit fingerprints

Secondary structure

Variable/conserved regions

V1 – V9: Fingerprints





PLOS One 10.1371/journal.pone.

High throughput fingerprinting

1. Get “*every*” DNA molecule in a sample: break cells up, wash, filter
2. Isolate fingerprint regions from *all* bacteria
3. Sequence them *all*

↓

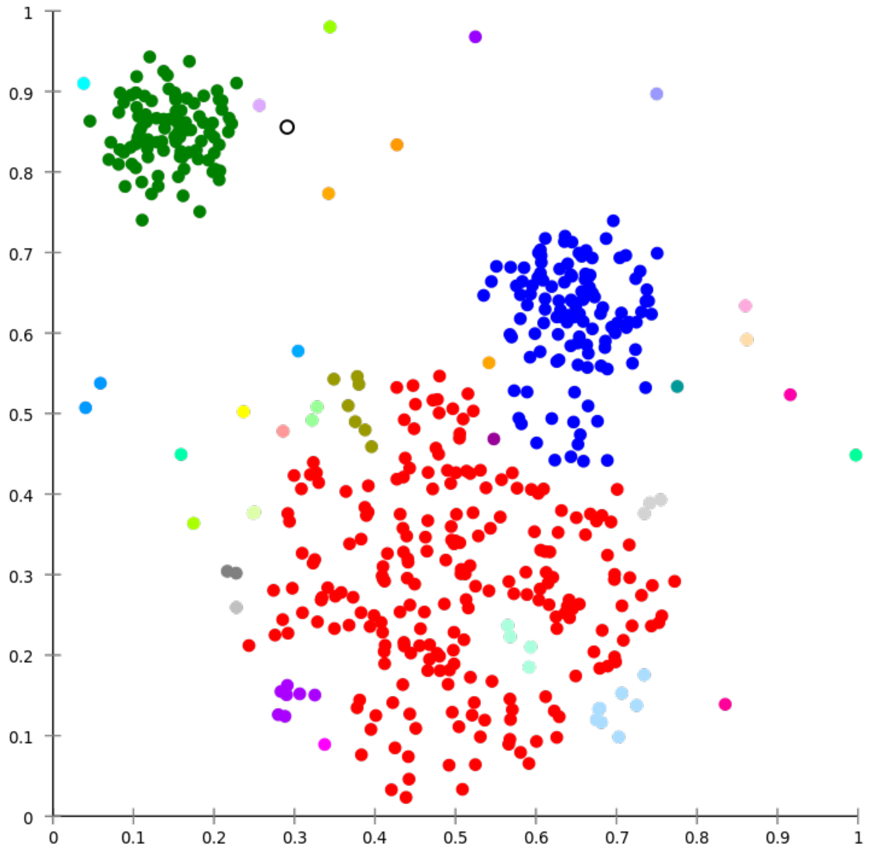
10–20 million fingerprint sequences

↓

Infer how many of which species were there

Interpreting fingerprint data

Initiative for Bioinformatics & Evolutionary Studies



1. Compute similarity (distance) between fingerprint sequences
2. Cluster, call a cluster a “species”
3. Number of clusters is species richness
4. Size of clusters is species abundance

Problem – Solution

?

- Fingerprints evolve at different rates: which to use?
- Varies with species: need to know who is there to choose best fingerprint!

! Use multiple fingerprints!

?

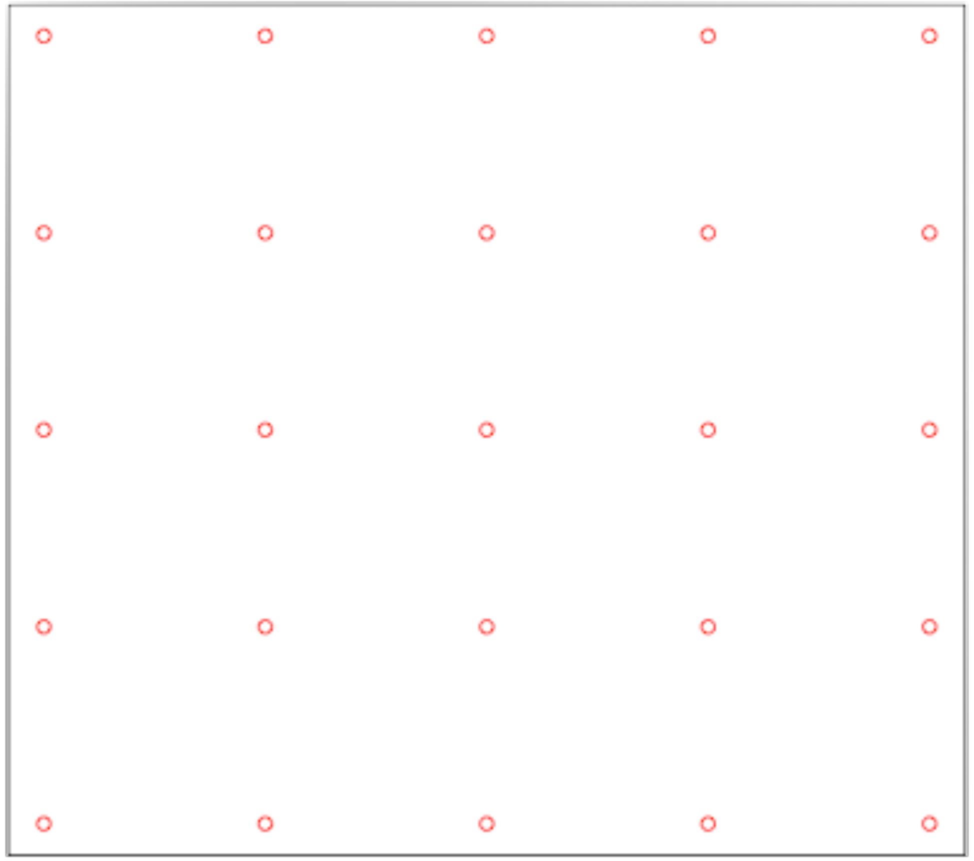
- How to compare distances between sequences from unknown species?
- With multiple fingerprints with unknown biases?

Solution: Find registration marks

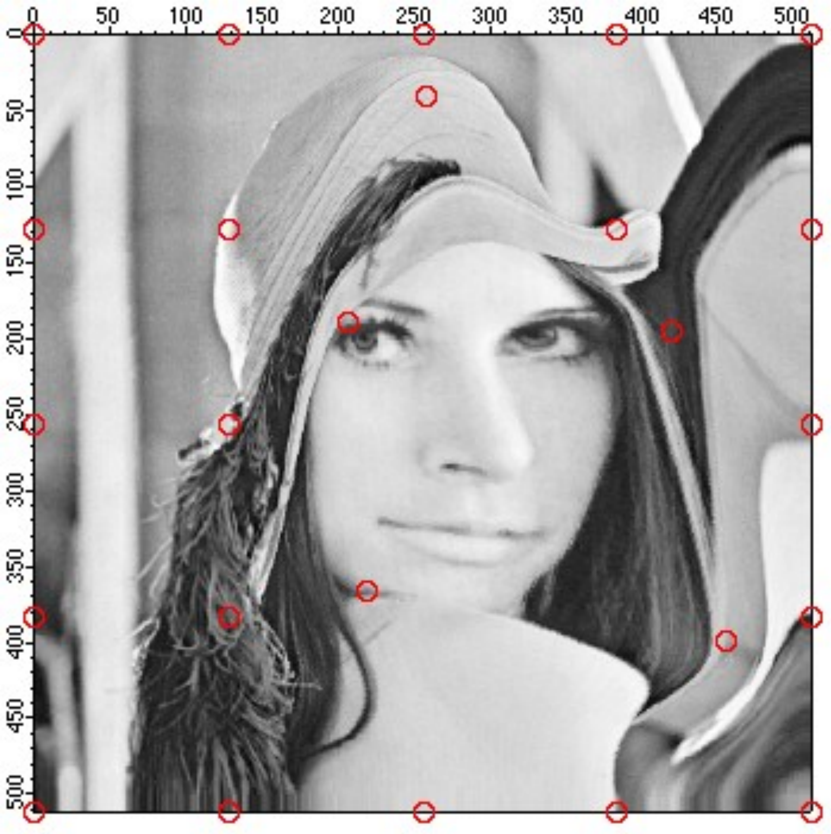
!

Translate distances using reference points from known fingerprints

Registration marks



Observed image

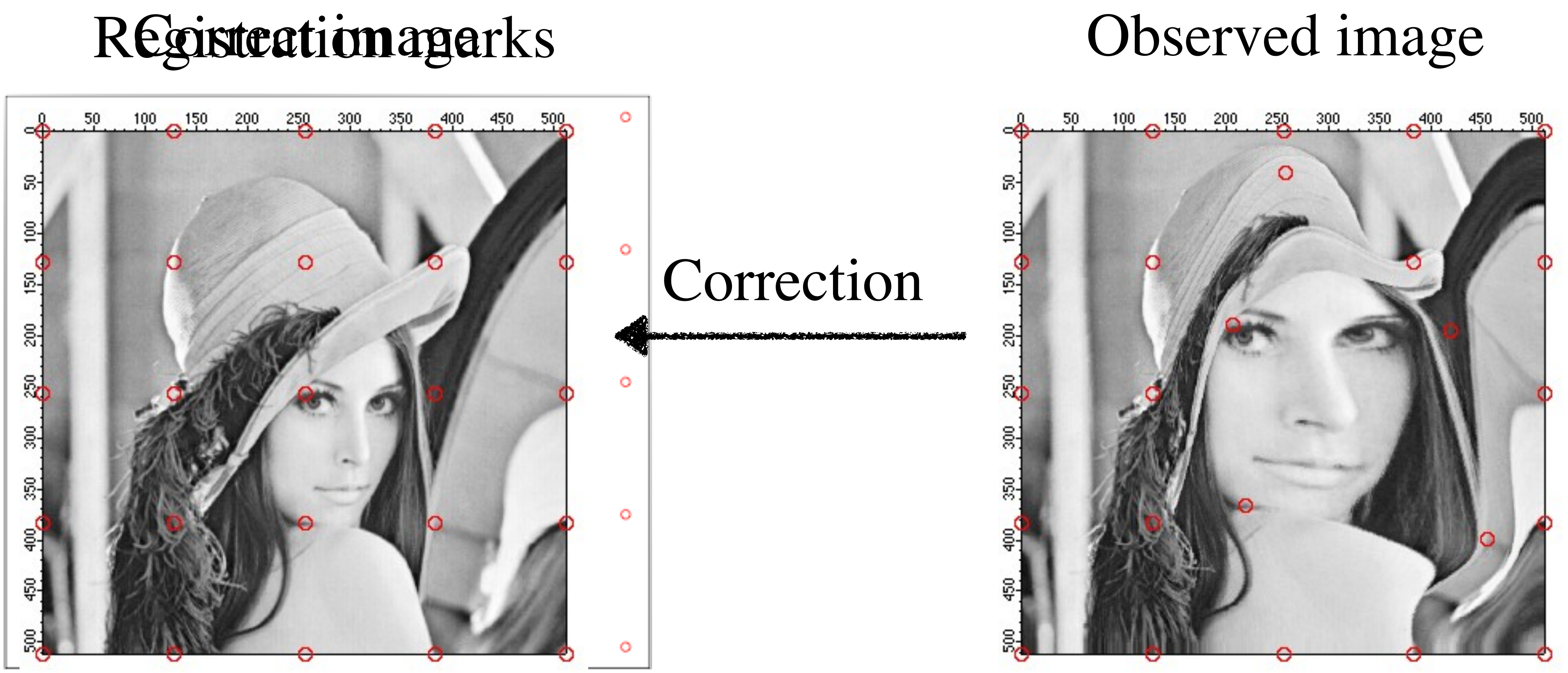


Solution: Use image registration

ibest
Initiative for Bioinformatics & Evolutionary Studies

!

Translate distances using reference points from known fingerprints



Solution: Find registration marks

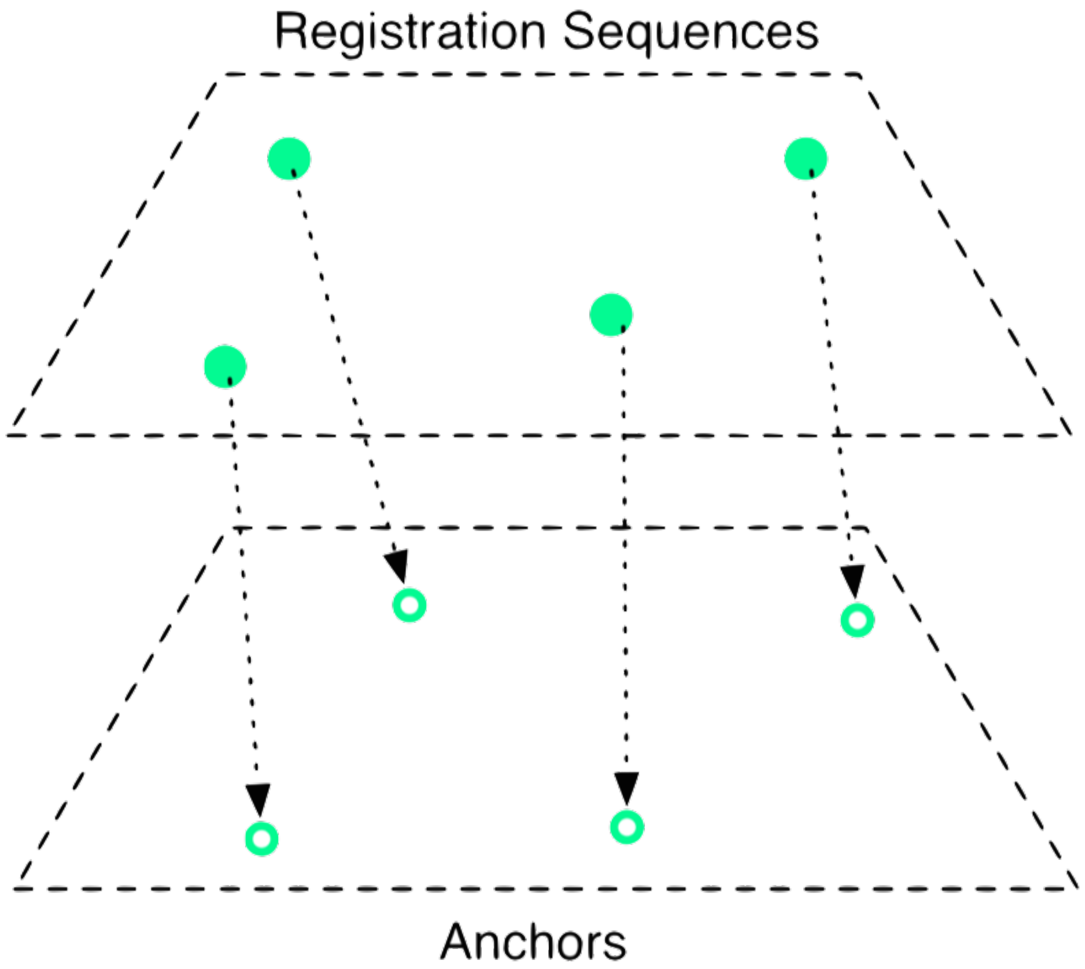
Determine distances for known full 16S sequences

Extract a fingerprint

Determine distances for reference fingerprints (anchors)

● Known full 16S sequences

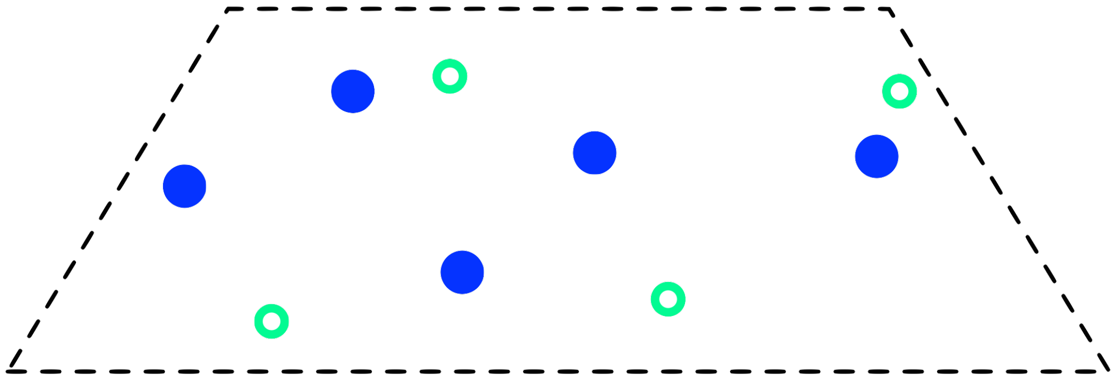
○ Fingerprint subsequences



Solution: add empirical reads

Combine DNA sequences with (known) anchors

Compute distances from reads to *all* sequences



Empirical Reads

● Actual reads

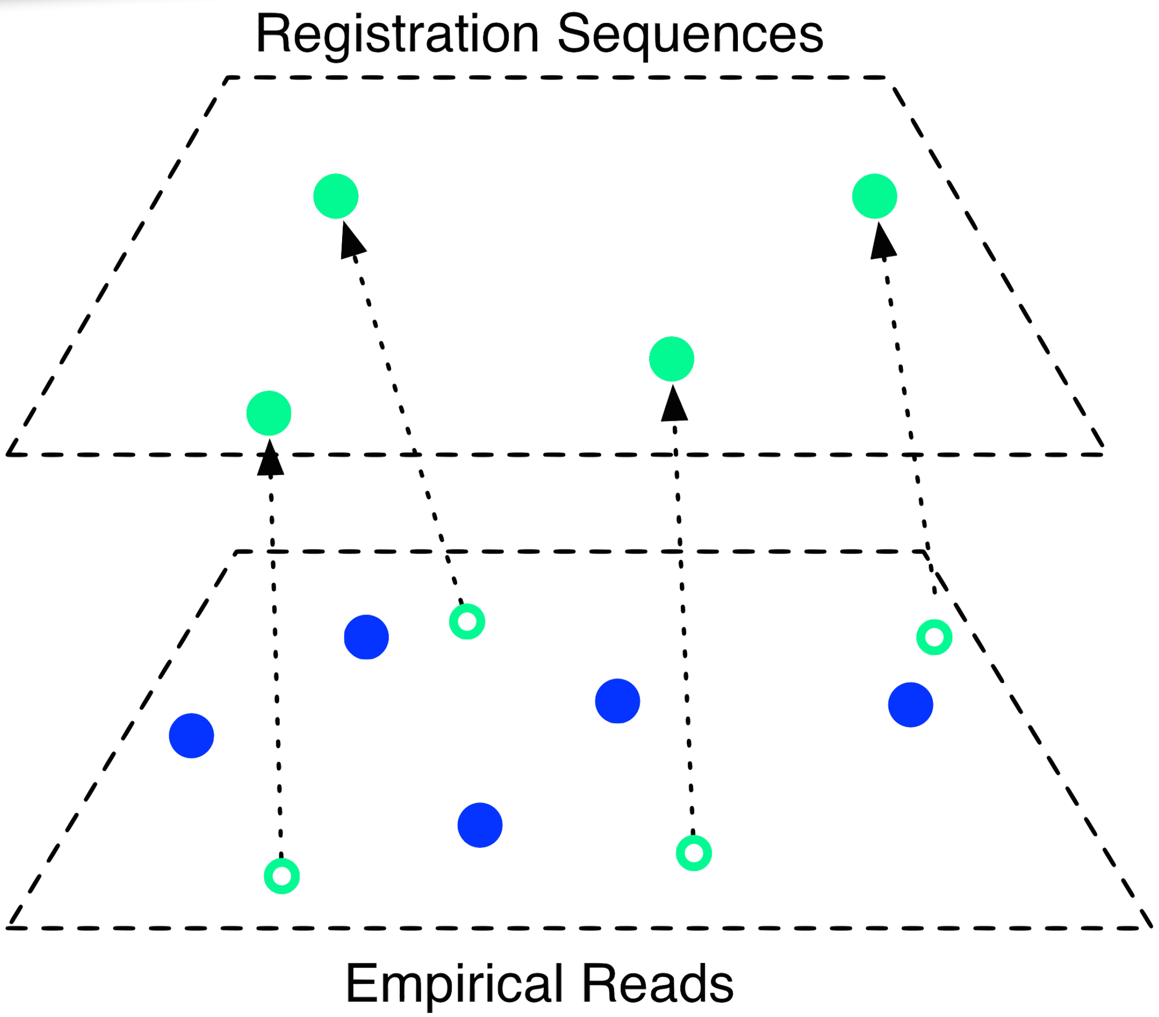
○ Fingerprint sequence

Observed image

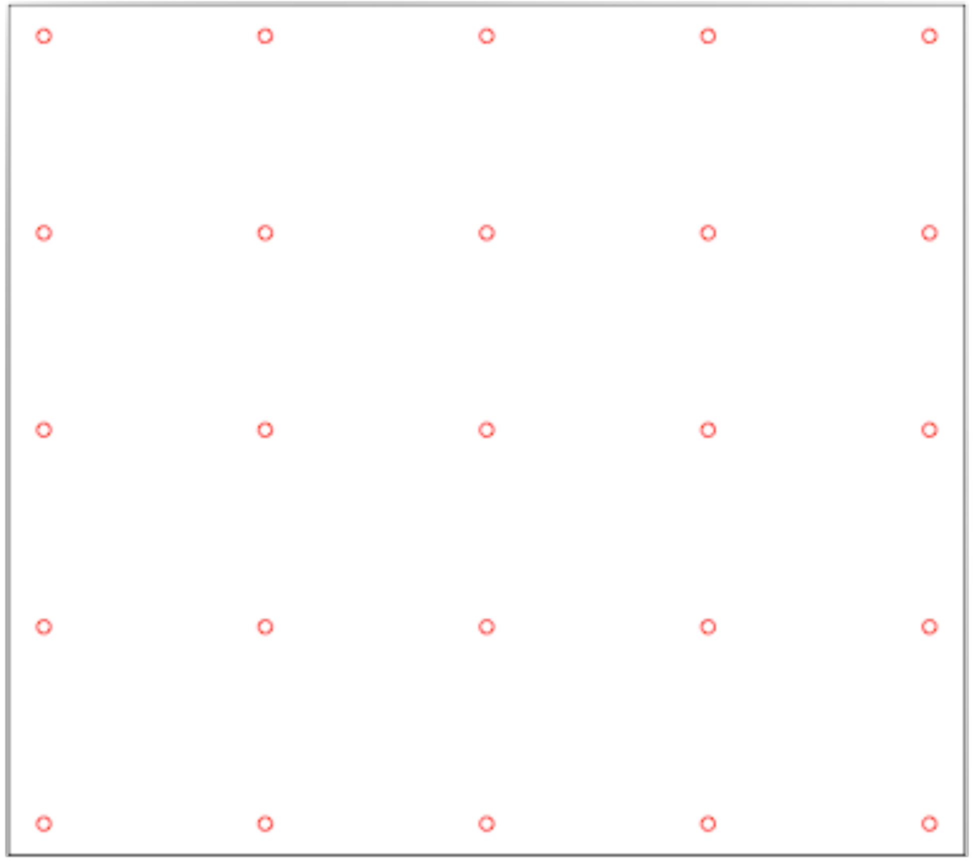
Solution: match registration

Initiative for Bioinformatics & Evolutionary Studies

Map anchors back to known registration marks



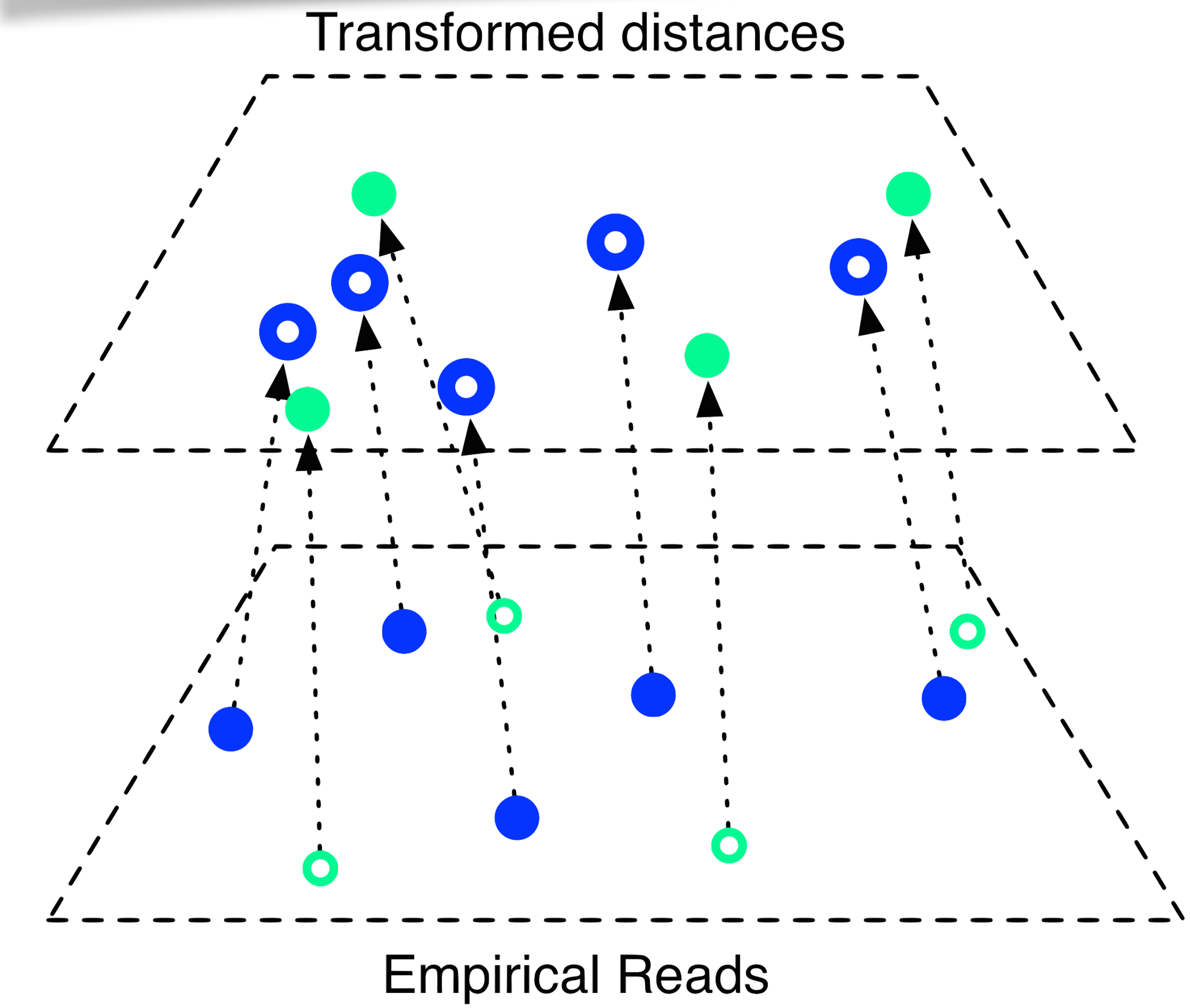
Registration marks



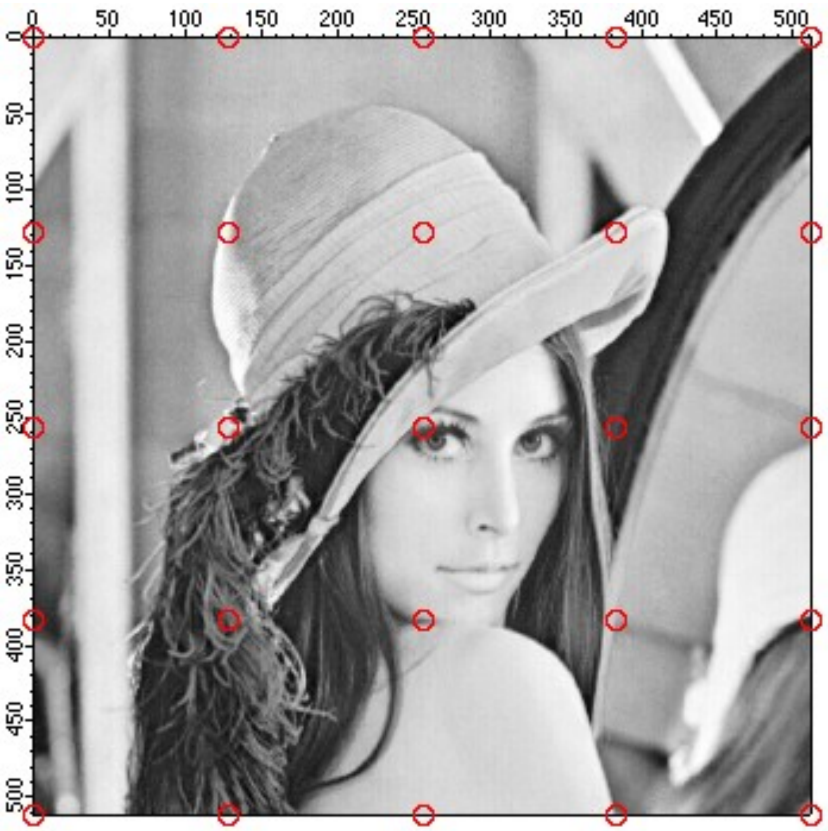
University of Idaho

Solution: Move empirical data

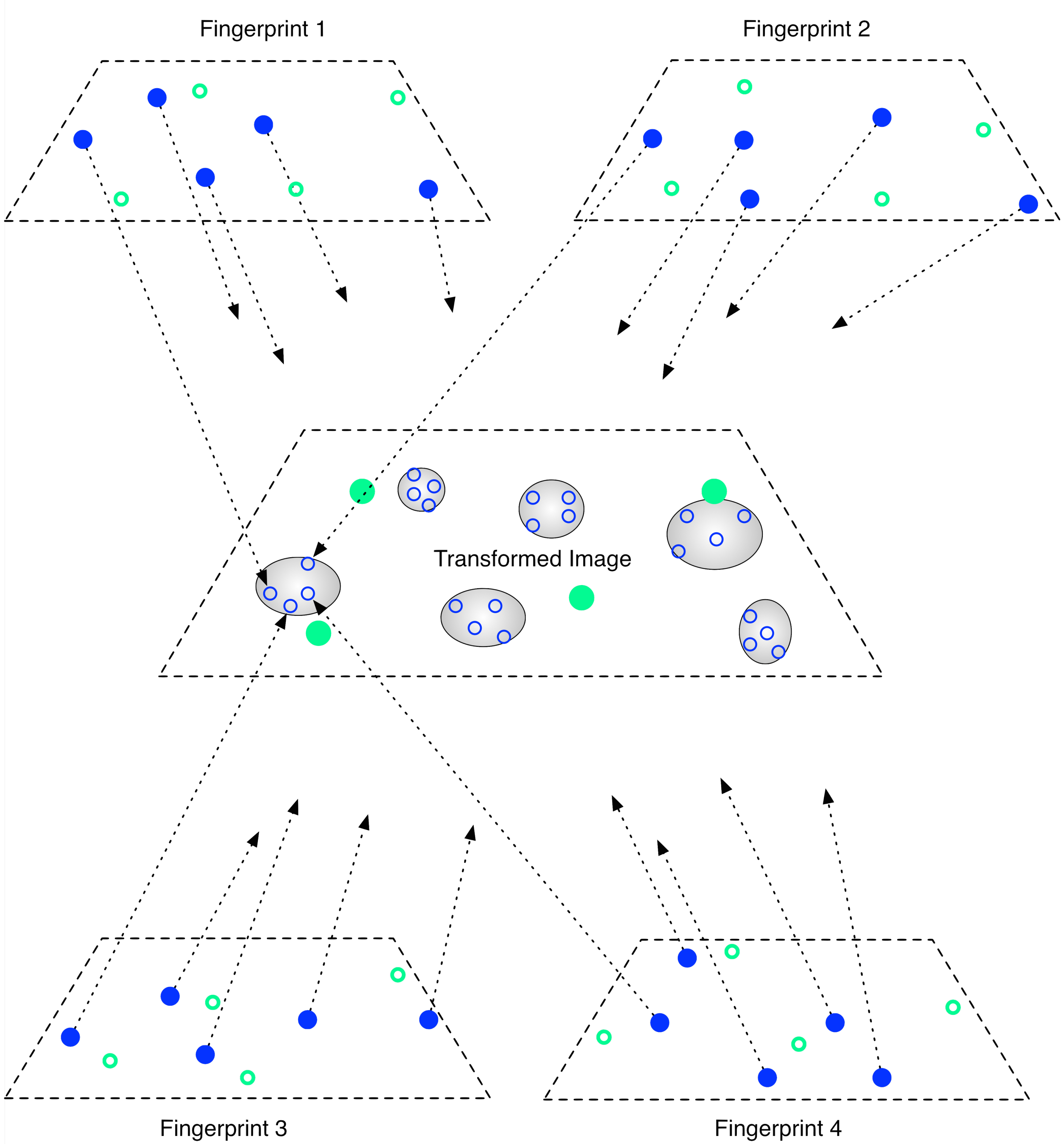
Carry empirical reads along



Correct image



Solution: combine fingerprints



Repeat for multiple fingerprints

Remove outliers

Cluster results

Current activity

- Find efficient 2D mapping: nonmetric multidimensional scaling (NMDS)
- Develop fast distance computation algorithms: pre clustering plus hashing
- Develop accuracy statistics: perturbation analysis
- Determine how many "registration sequences" are best for 20 million empirical points
- Precompute registration libraries for different sample types (soil, human microbiome, ocean, etc.)
- Determine accuracy with simulation and known sequences

Future work

- ✦Parallelize: use reference triangles
- ✦What to do with outliers?
- ✦What do cluster shapes/density say about fingerprints, species, ecology?
  - Which fingerprints are good for which species?
  - Which fingerprints are most misleading in given environments?
- ✦How are clusters and evolution related?
- ✦Application to empirical data: milk project
- ✦Many more!

Acknowledgements

Initiative for Bioinformatics & Evolutionary Studies

Funding

- NSF DBI0939454 BEACON Evolution in Action
- NSF IAA 041485301 “What is normal milk?”
- NIH P20GM16448 COBRE center for research in processes of evolution

Institute for Bioinformatics and Evolutionary Studies (IBEST)

- IBEST Computational Resources Core

Students

- Ilya Zhbnanikov, BCB PhD Candidate
- Janet Williams, BCB PhD student
- Daniel Beck, BCB PhD