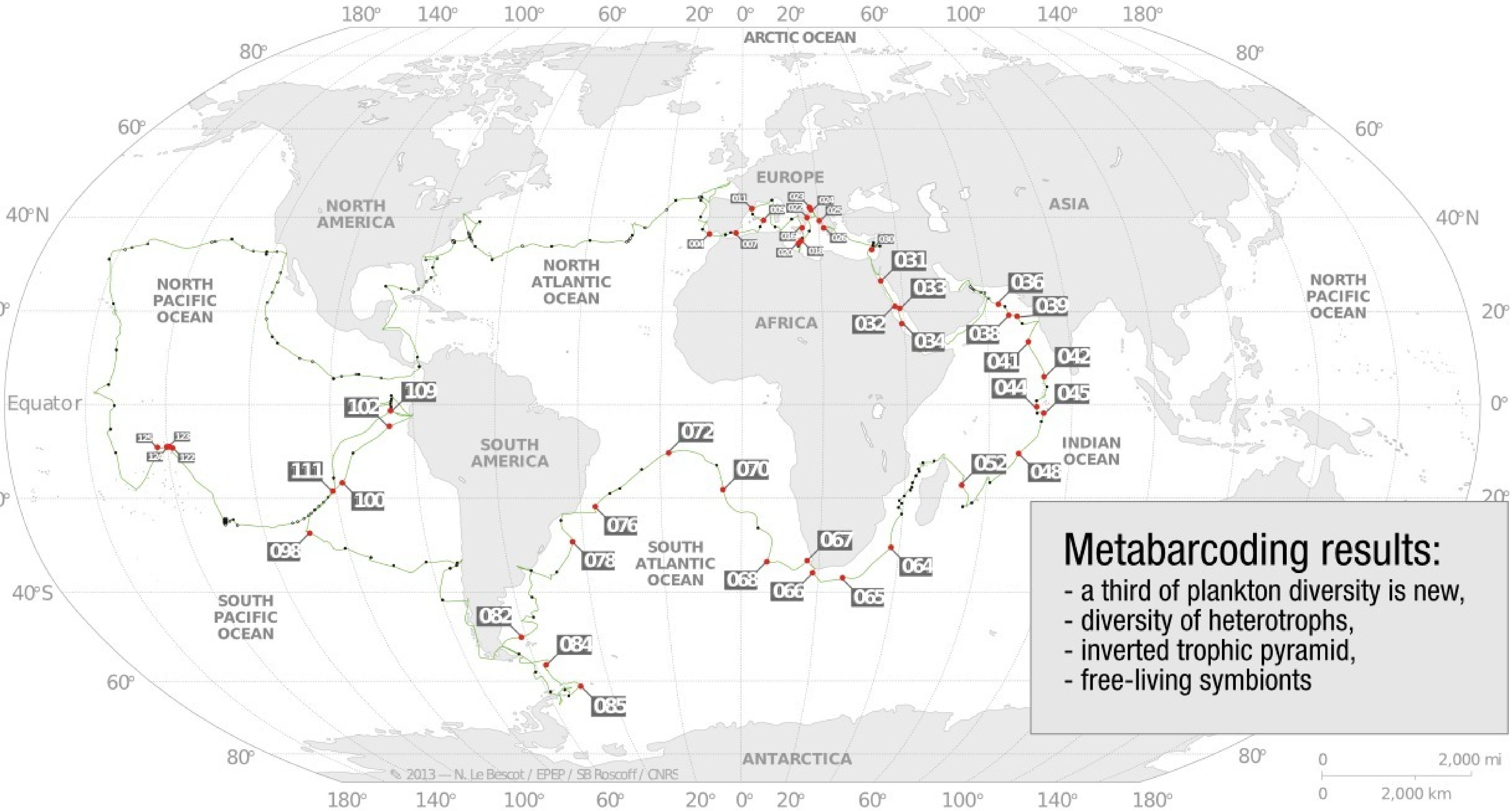
Metabarcoding

bioinformatics progress & challenges

Frédéric Mahé
March 8, 2016
CBGP

Small organisms play a major role but are hard to inventory



© 2013 — N. Le Bescot / EPEP / SB Roscoff / CNRS

Metabarcoding results:

- a third of plankton diversity is new,
- diversity of heterotrophs,
- inverted trophic pyramid,
- free-living symbionts

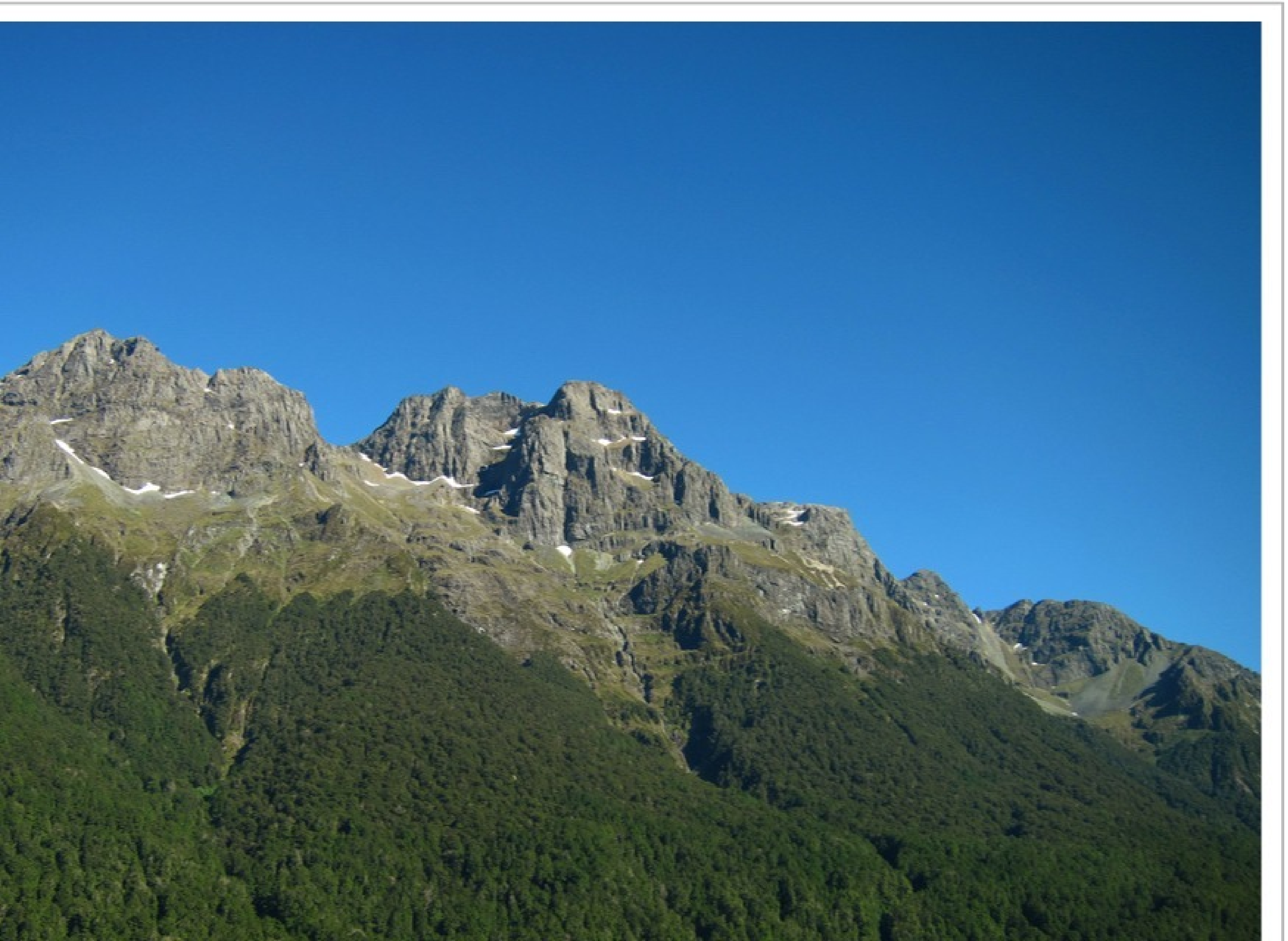
First trip of the TARA OCEANS project

(1.3 billion reads, the 47 sampling stations published so far are in red)

de Vargas et al., 2015 Science

Microbial diversity at the tree line level

David Wardle & Jordan Mayor, Swedish University of Agricultural Sciences



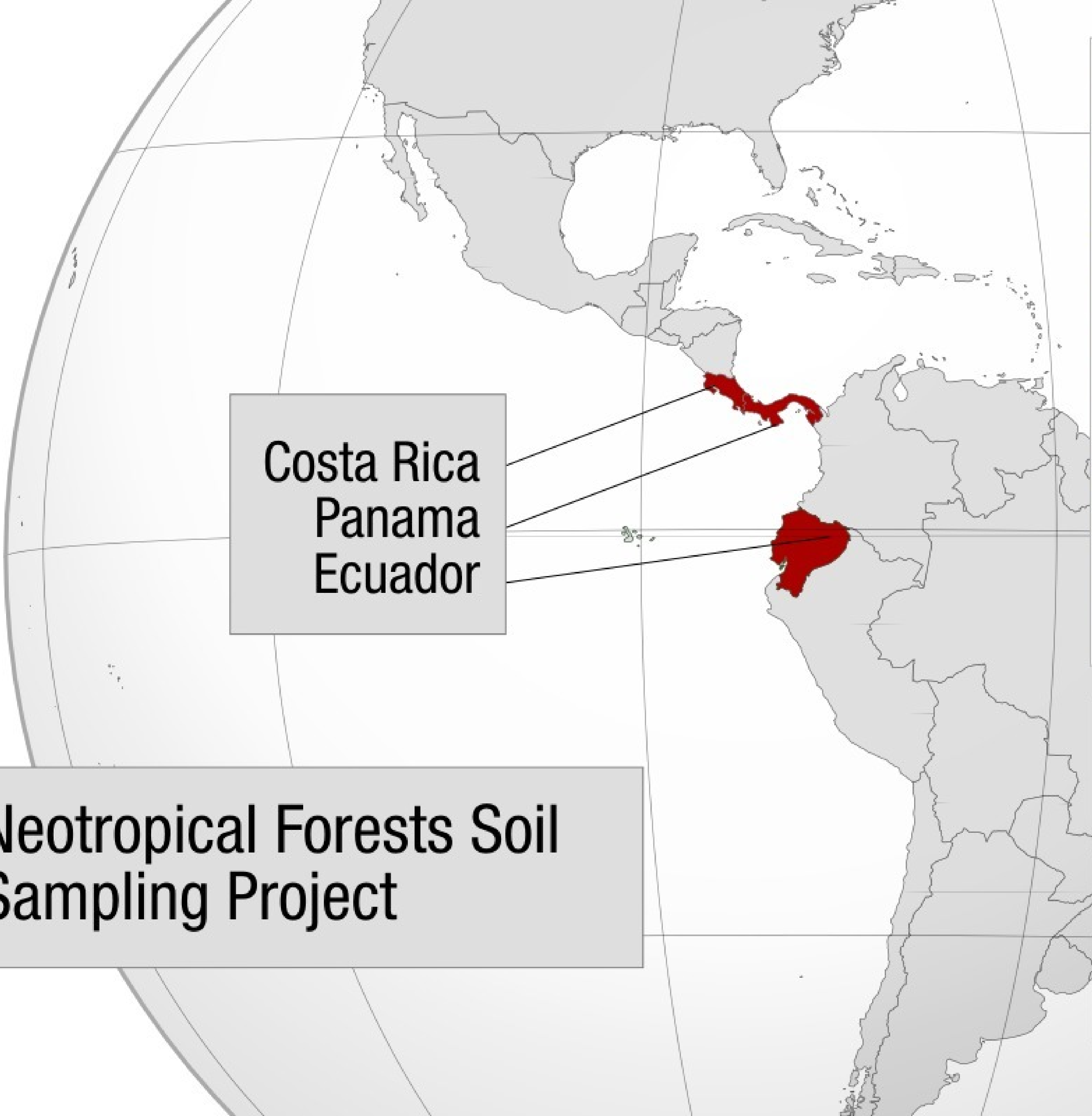
Sampling

- 7 countries (Australia, Austria, Canada, Chile, Japan, New Zealand, USA),
- 5 transects per country,
- 8 replicates per transect,
- 5 soil samples per replicate,
- soil chemistry

Early results

- few unknowns,
- dominance of fungi,
- weak endemism,
- microbial communities?
- geographical analysis?

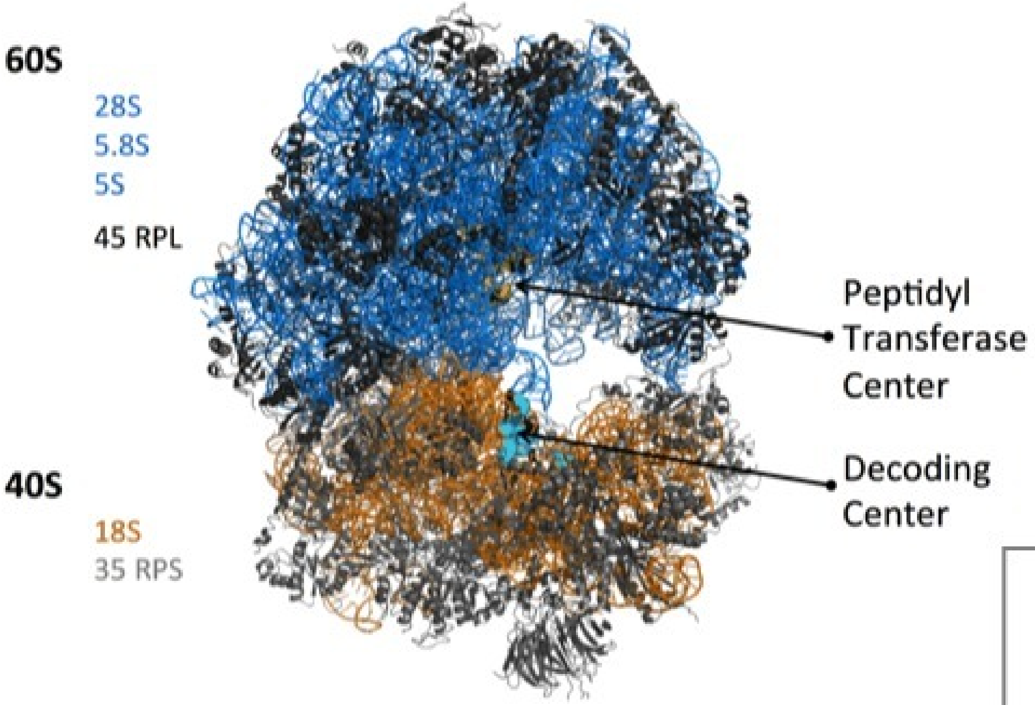
Neotropical Forests Soil Sampling Project

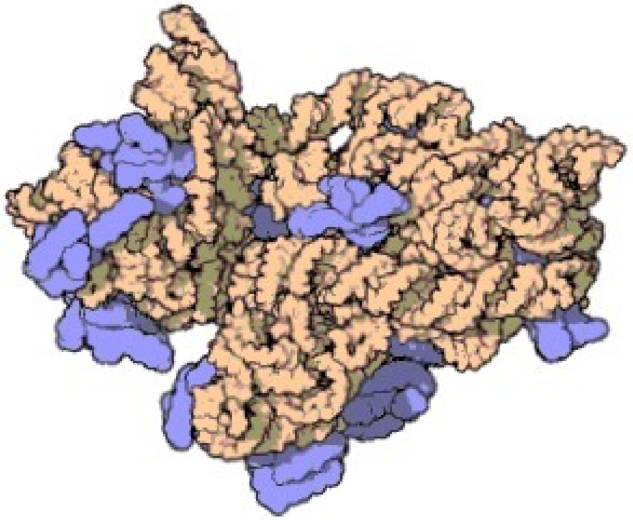


Early results

- half of unknowns,
- not-so-many fungi,
- dominance of parasits,
- notable endemism,
- hyperdominant taxa

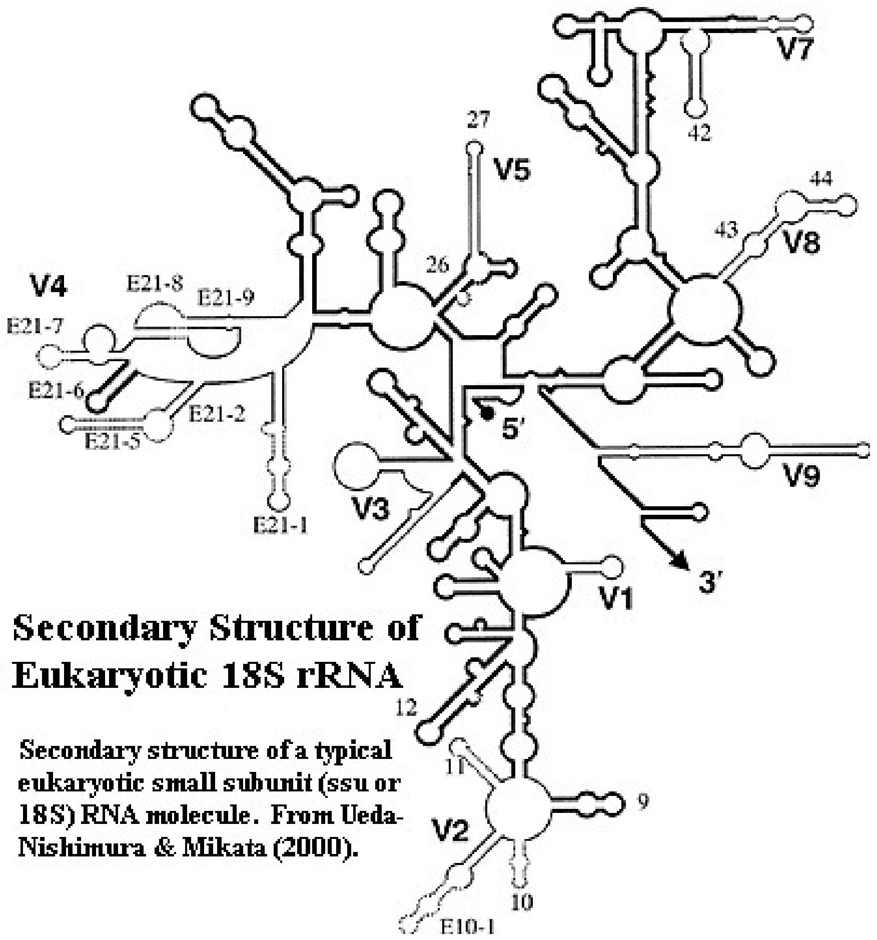
A universal gene: ribosomal RNA





Small Sub-Unit (SSU)

16S	Bacteria Archaea Mitochondria Chloroplasts
18S	Eukaryota



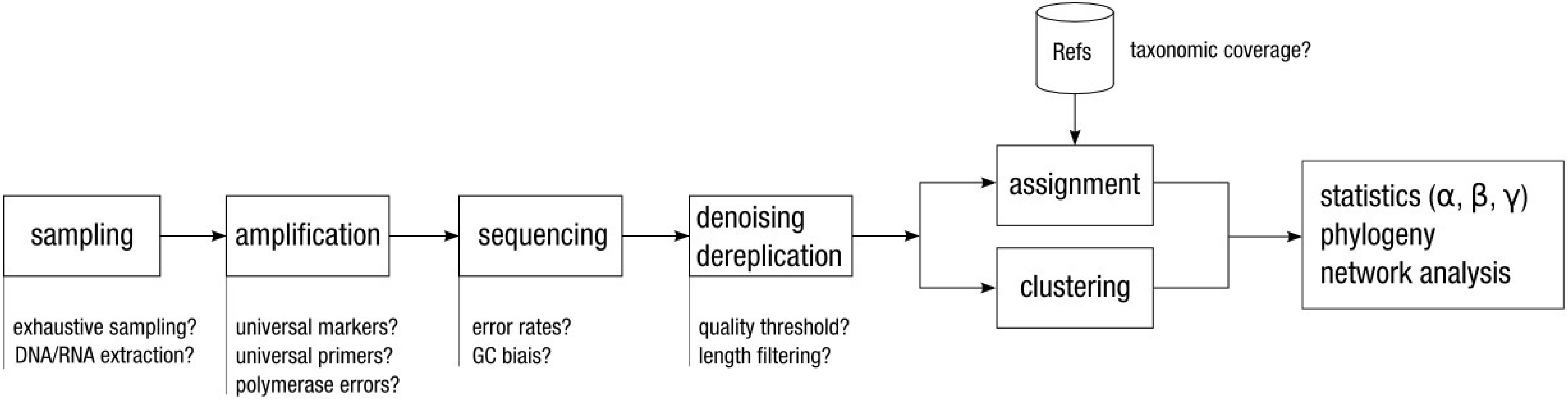
Secondary Structure of Eukaryotic 18S rRNA

Secondary structure of a typical eukaryotic small subunit (ssu or 18S) RNA molecule. From Ueda-Nishimura & Mikata (2000).

other markers can be used (e.g., ITS). Requirements are: conserved distal regions for primers, variable internal regions, and available sets of reference sequences.

Environmental Metagenomics

targeted amplification



sampling

exhaustive sampling?
DNA/RNA extraction?

amplification

universal markers?
universal primers?
polymerase errors?

sequencing

error rates?
GC biais?

denoising
dereplication

quality threshold?
length filtering?

Refs

taxonomic coverage?

assignment

clustering

statistics (α, β, γ)
phylogeny
network analysis

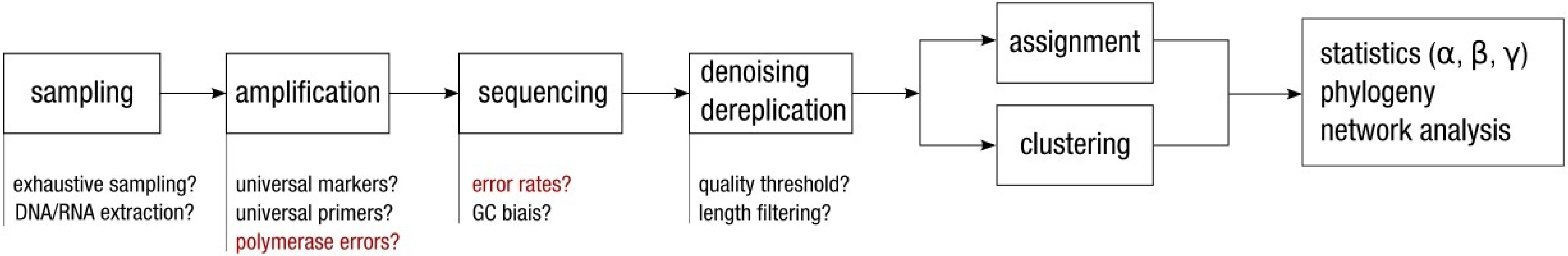
chemistry, physics and randomness

bioinformatics
(preparation)

biostatistics
(actual analyses)

Noise is the real challenge

amplification and sequencing are imperfect processes



sampling → amplification → sequencing → denoising dereplication → assignment / clustering → statistics (α, β, γ) phylogeny network analysis

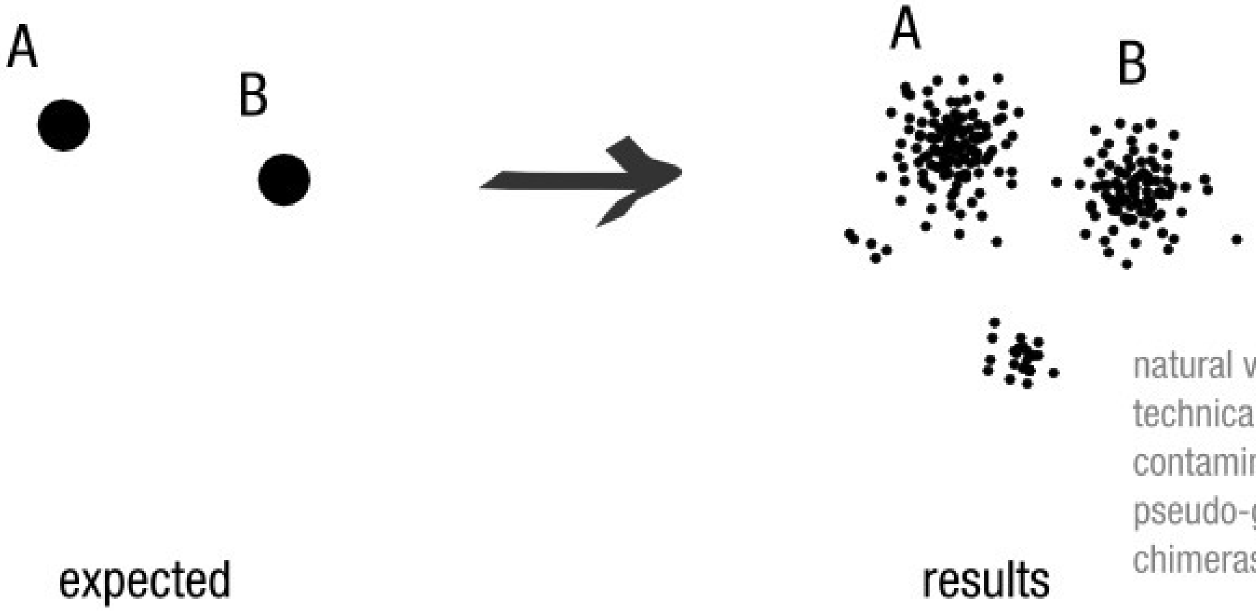
sampling: exhaustive sampling? DNA/RNA extraction?

amplification: universal markers? universal primers? polymerase errors?

sequencing: error rates? GC biais?

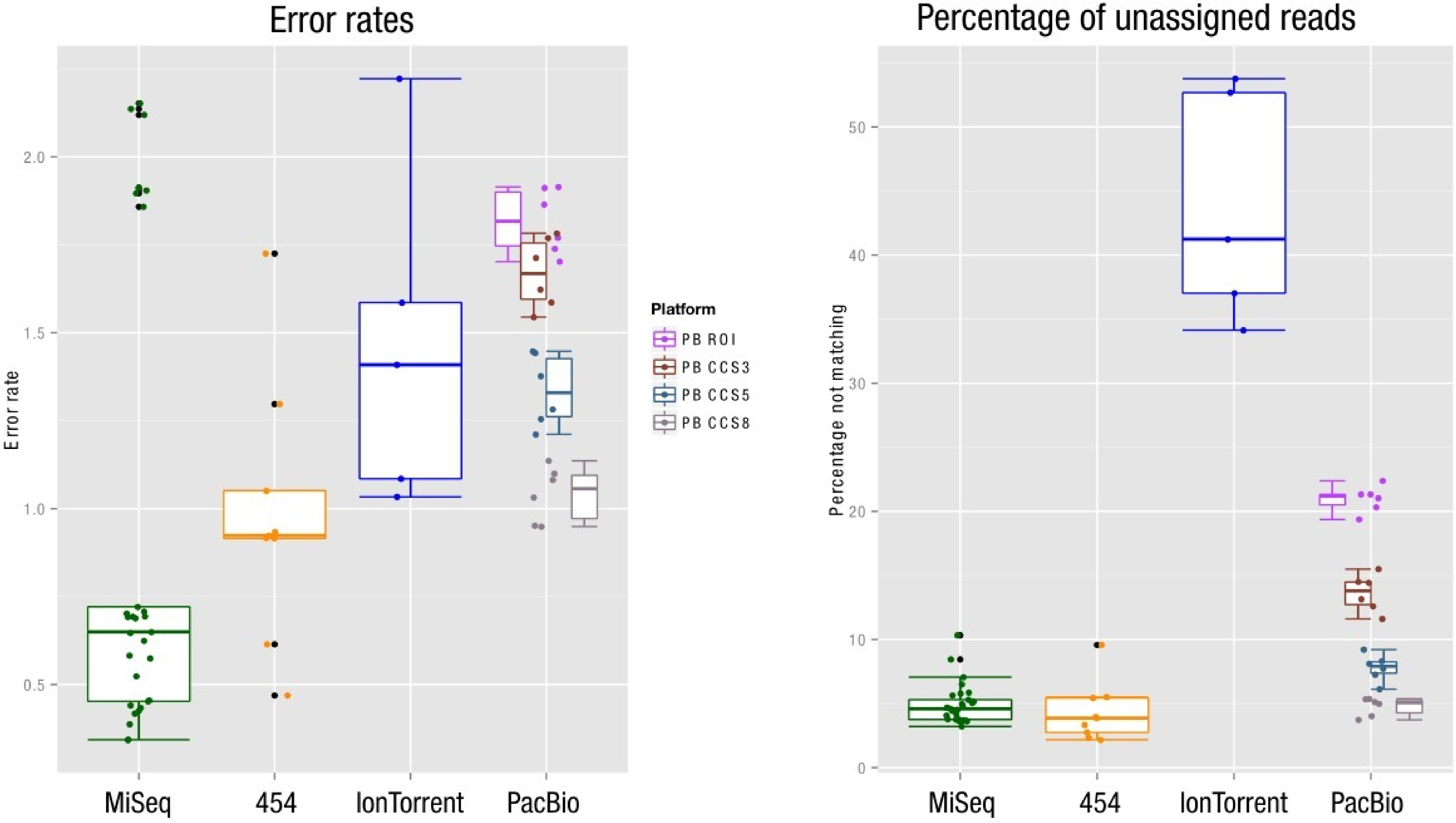
denoising dereplication: quality threshold? length filtering?

chemistry, physics and randomness



natural variability?
technical noise?
contaminant?
pseudo-gene?
chimeras?

Error rate per sequencing platform

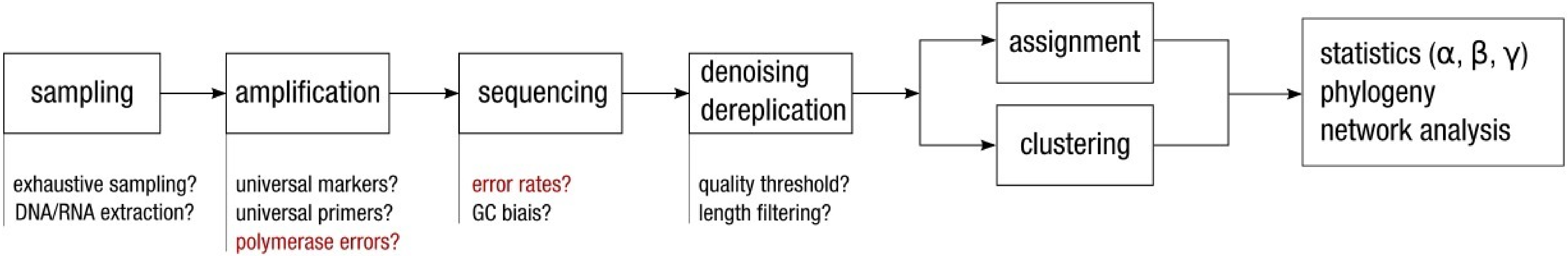

Error rates

Percentage of unassigned reads

D'Amore et al. (2016) A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics.

Noise is the real challenge

amplification and sequencing are imperfect processes



sampling → amplification → sequencing → denoising dereplication → assignment / clustering → statistics (α, β, γ) phylogeny network analysis

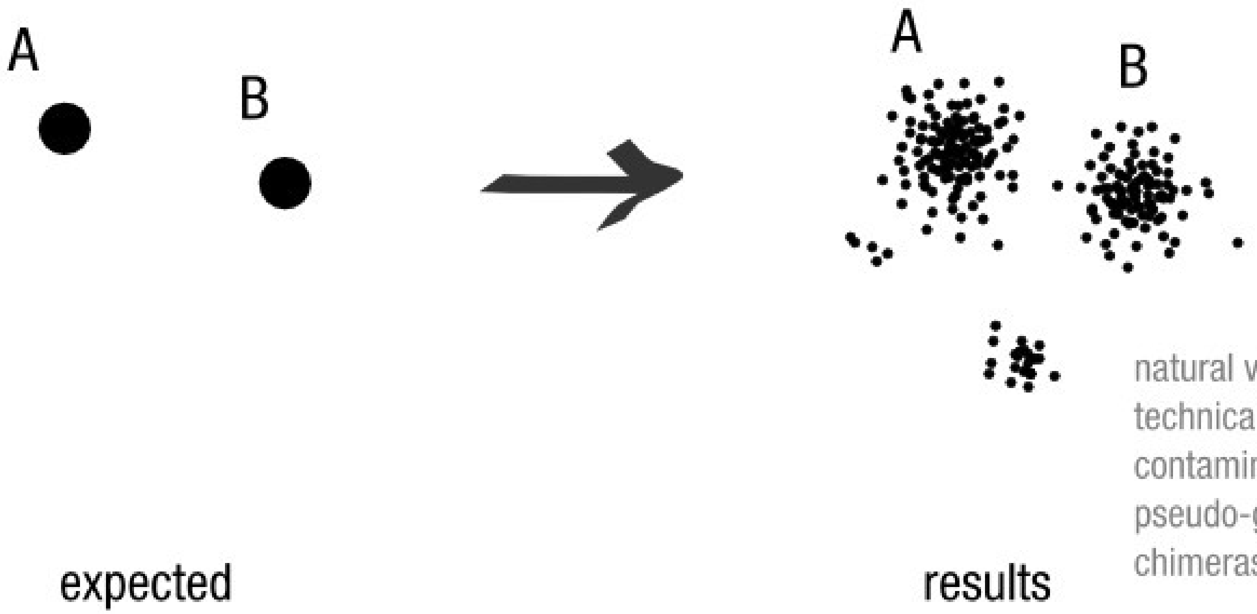
sampling: exhaustive sampling? DNA/RNA extraction?

amplification: universal markers? universal primers? polymerase errors?

sequencing: error rates? GC biais?

denoising dereplication: quality threshold? length filtering?

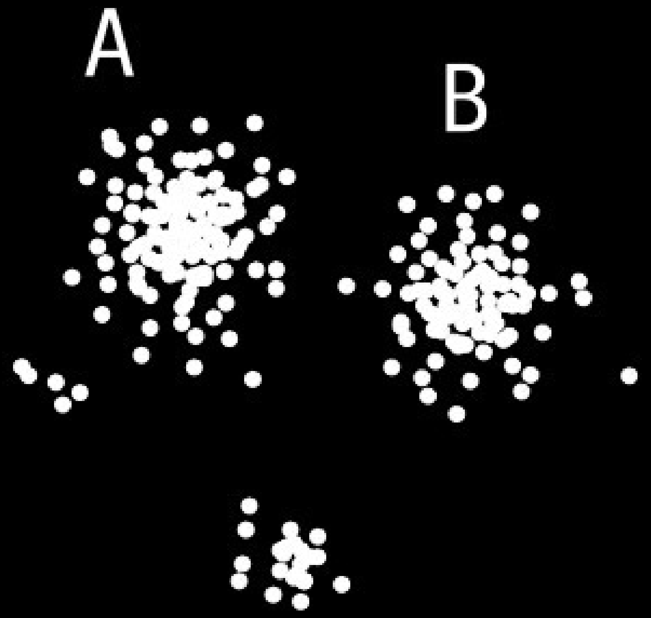
chemistry, physics and randomness



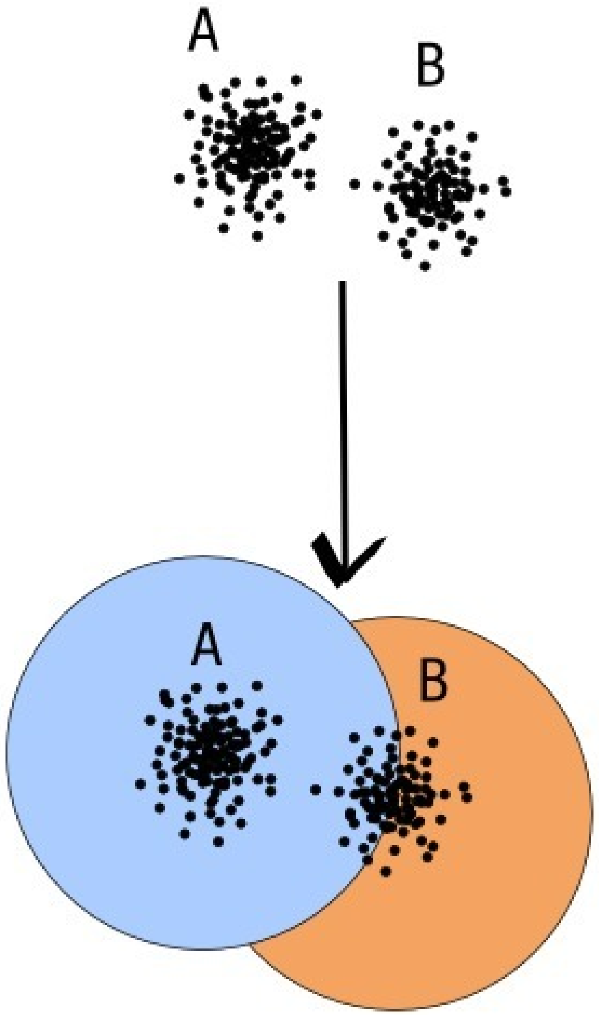
natural variability?
technical noise?
contaminant?
pseudo-gene?
chimeras?

dereplication & clustering

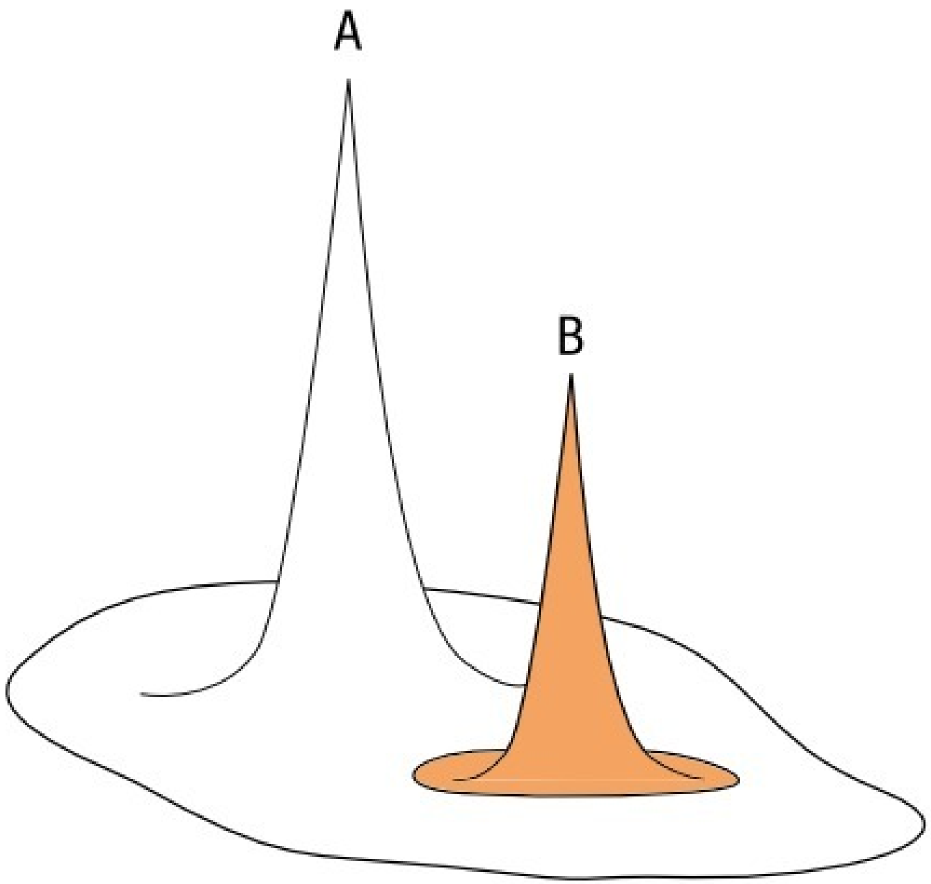
use errors at your own advantage



Swarm: fast, exact and high-resolution clustering



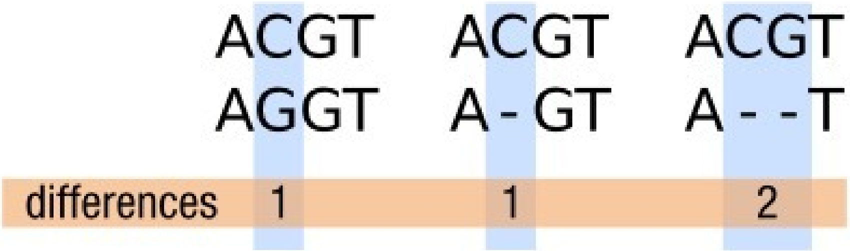
clustering threshold (often 97%) is most of the time unadapted and can mask diversity.



swarm uses abundance values and a new clustering strategy to delineate natural high-quality OTUs.

Swarm clustering method
growth phase

take advantage of PRC and sequencing errors



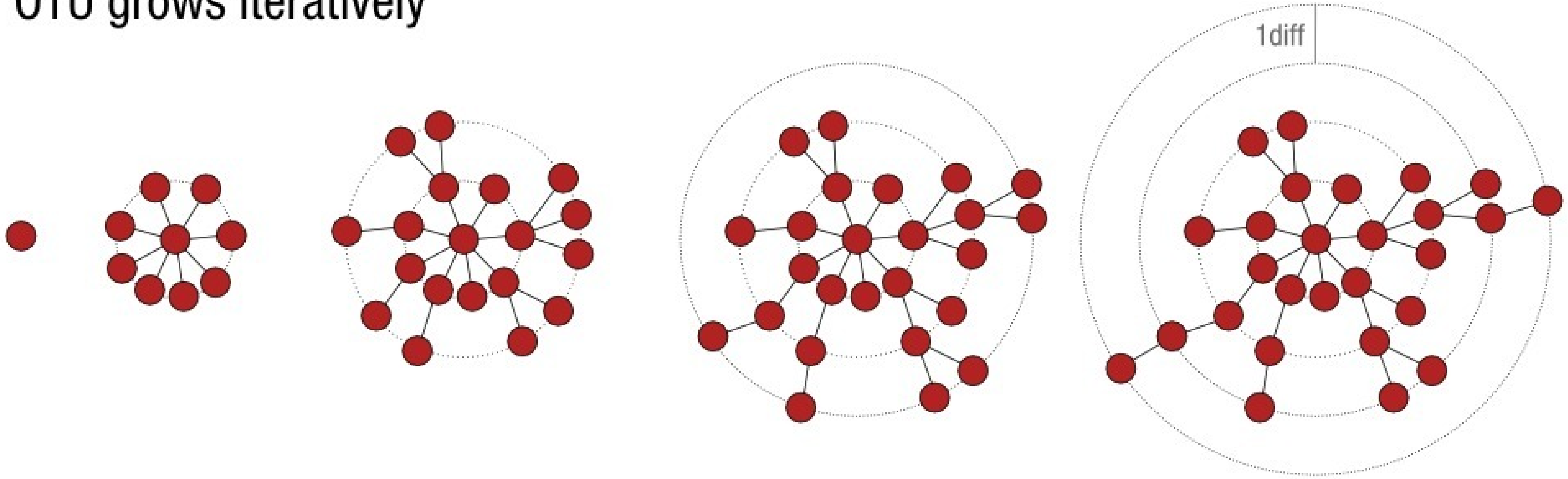
ACGT ACGT ACGT
AGGT A - GT A - - T

differences 1 1 2

Avoid & speed-up comparisons

- composition-based prefiltering
- memoization
- fast Needleman-Wunsch

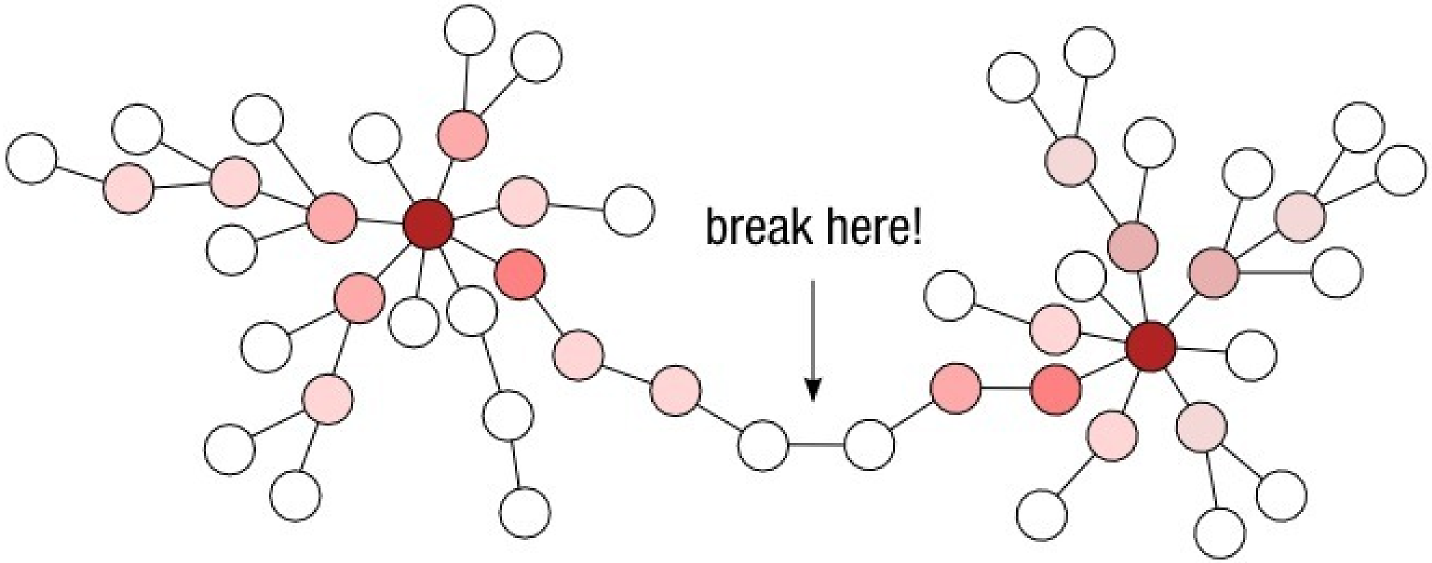
OTU grows iteratively



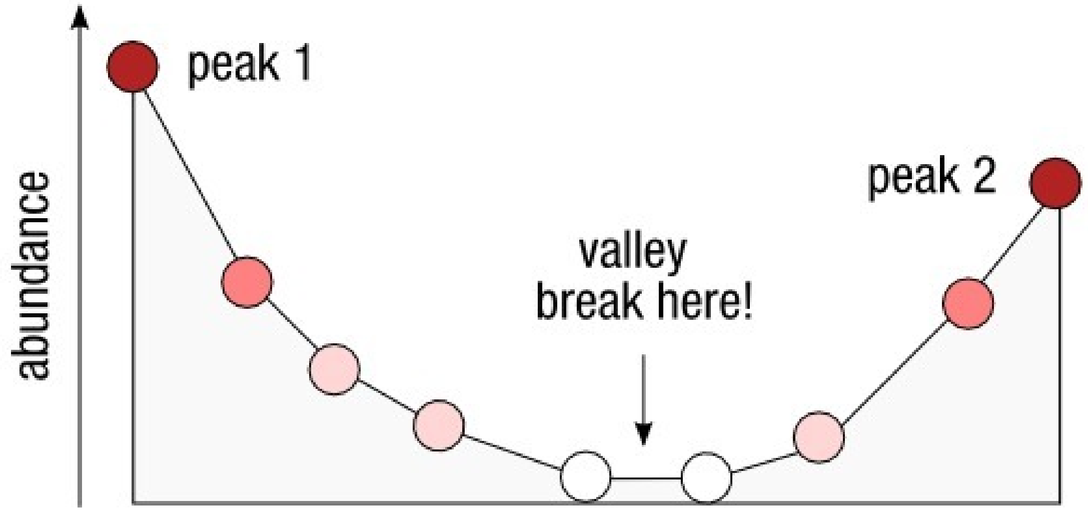
initial seed (randomly picked
from amplicon dataset)

no more closely related amplicons,
the process stops

Swarm clustering method
breaking phase



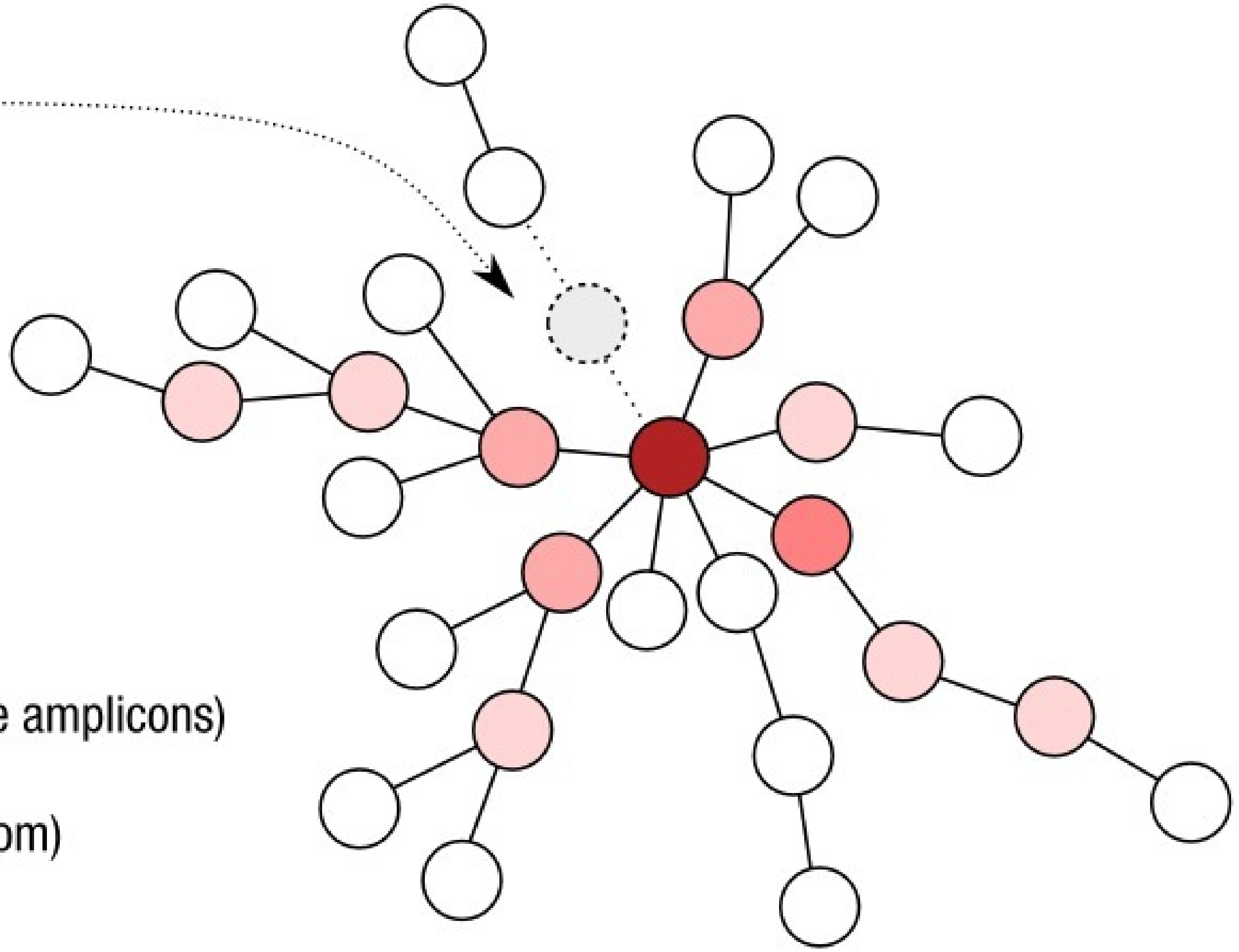
Take into account the abundance of amplicons to produce higher-resolution clusters.



Assuming that original sequences are more abundant than erroneous copies.

Swarm clustering method
grafting phase

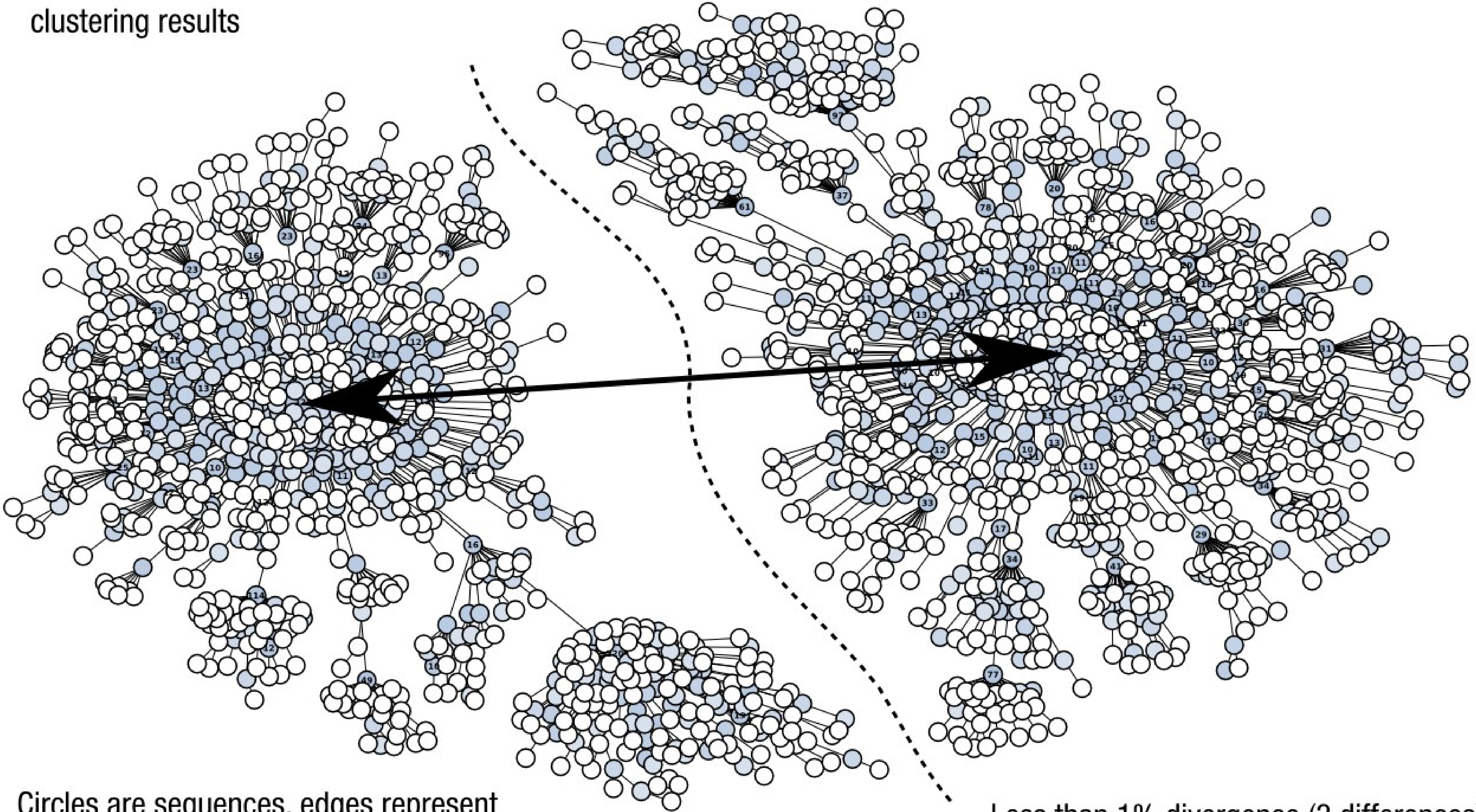
Postulate the existence of an intermediate amplicon to be able to graft a small OTU onto a bigger one.



○—○ small OTU (made of 2 rare amplicons)

◌ virtual amplicon (or phantom)

Swarm produces high-resolution clustering results



Circles are sequences, edges represent one difference (substitution or indel)

Less than 1% divergence (3 differences) between the two peaks of abundance

Swarm 2.0 is a highly scalable denoising-clustering method

earth microbiome project

Suomi (Finland)
Island (Iceland)
Sverige (Sweden)
Norge (Norway)
Polska (Poland)
Україна (Ukraine)
Türkiye (Turkey)
(China)
(South Korea)
(Japan)
North Pacific Ocean
North Atlantic Ocean
México (Mexico)
(Algeria)
(Libya)
(Egypt)
(Iraq)
(Iran)
(Pakistan)
(Saudi Arabia)
India
(Thailand)
Mali
Niger
Tchad (Chad)
(Sudan)
Nigeria
Venezuela
Colombia
DR Congo
Tanzania
Indonesia
Papua New Guinea
Peru (Peru)
Brasil (Brazil)
Bolivia
Angola
Namibia
Botswana
Madagasikara (Madagascar)
Indian Ocean
Australia
South Pacific Ocean
South Atlantic Ocean
Chile
South Africa
Argentina
erreur cartographiq

28,275 samples
2.3 billion reads

swarm: 5 hours
usearch: >150 days

chimera detection

```
A: GTCGCTACTACCGATTGAACGTTTTAGTGAGGTCCTCGGACTGTGAGCCTGGCGGGTTG
                      ||||||||||||||
B: TACTACCAAACTGAGTTAGCGTTTTAGTGAGGTAAGACGACCAAACTGTAGCGTTTAG

C: GTCGCTACTACCGATTGAACGTTTTAGTGAGGTAAGACGACCAAACTGTAGCGTTTAG
```

vsearch: open-source alternative for usearch

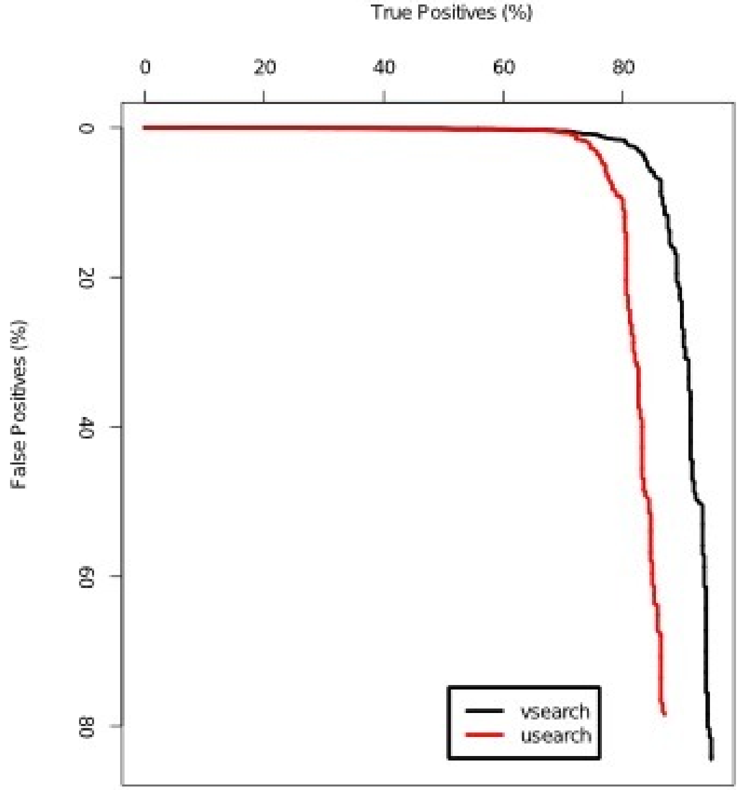
clustering, chimera detection, dereplication, searching, sorting, masking and shuffling

usearch (Rob Edgar):
- very important for metagenomics,
- 1,000 citations,
- fundation for QIIME,
- closed-source & costly

qiime Quantitative Insights Into Microbial Ecology

growing success:
- many happy users,
- faster and improved,
- fundation for QIIME 2.0

vsearch:
- free and open-source,
- fast,
- documented,
- revive the research field



Torbjørn Rognes
Oslo University

vsearch: open-source alternative for usearch

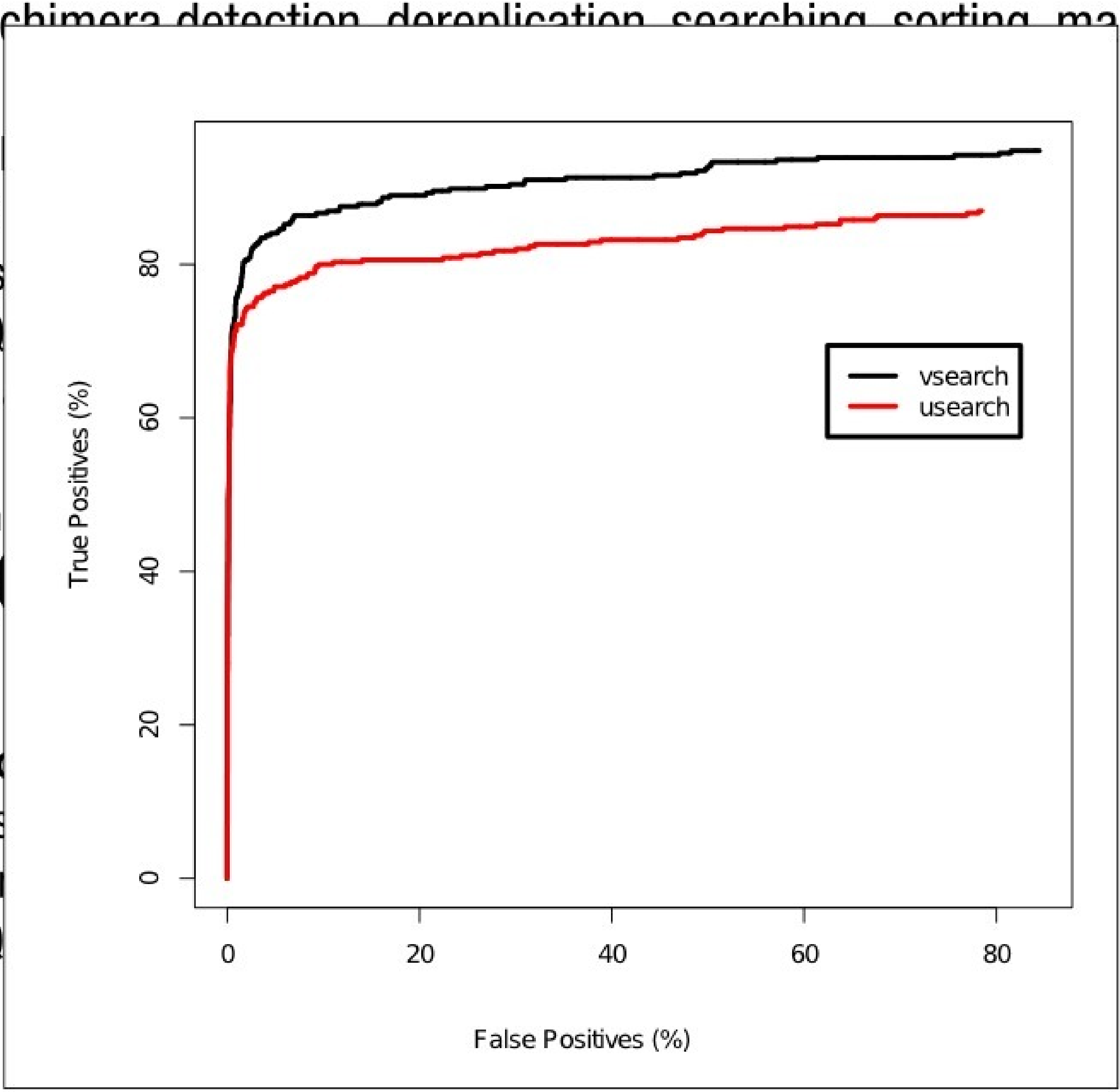
clustering, chimera detection, dereplication, searching, sorting, masking and shuffling

usearch (Rob I

- very important
- 1,000 citations
- fundation for Q
- closed-source

growing succ

- many happy us
- faster and impr
- fundation for Q



Torbjørn Rognes
Oslo University

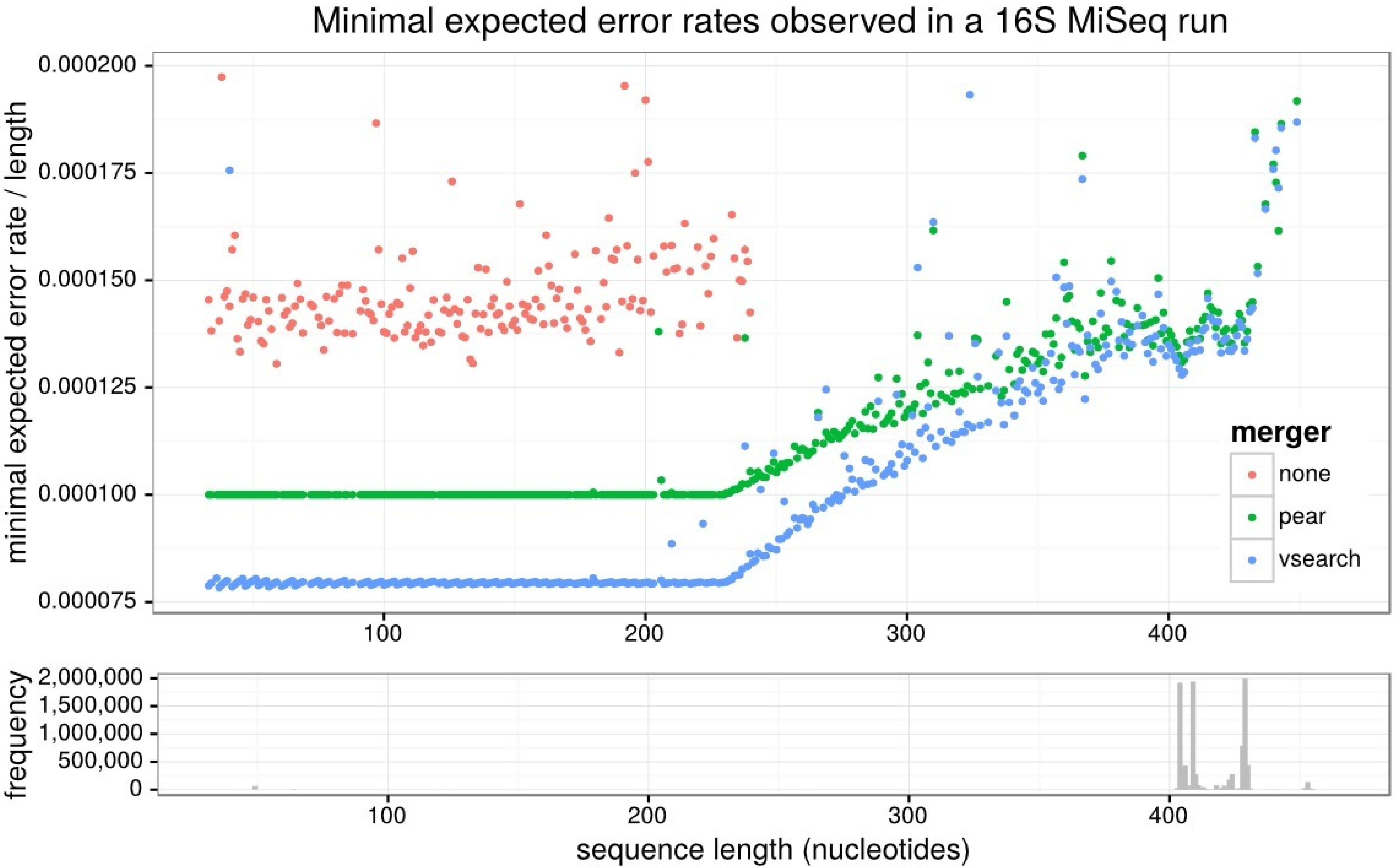
quality filtering

```
@M00185:171:000000000-AJ75U:1
TACGGGAGGCAGCAGTGGGGAATCTTGCGCAATGCGCGAAAGCGTGACGCAGCAACGCCG
+
BCCCC@BBCCCCGGGGGGGGGGHHHHHHGGGGGHFEGGGGGGHGGGGEGGGGGHHHFGGG

@M00185:171:000000000-AJ75U:2
CGGCGTTGCTGCGTCACGCTTTCGCGCATTGCGCAAGATTCCCCACTGCTGCCTCCCGTA
+
HEGCG-BCFGGGGGGGGGGFFFFFFGGGGAGAAADFFFFFFFFFFFFFFFFFFFFEDA;F
```

expected error rate = Σ Qv

ee = 1.0 (50% chance to have zero error)


Minimal expected error rates observed in a 16S MiSeq run
minimal expected error rate / length
0.000200
0.000175
0.000150
0.000125
0.000100
0.000075
100
200
300
400
merger
none
pear
vsearch
frequency
2,000,000
1,500,000
1,000,000
500,000
0
100
200
300
400
sequence length (nucleotides)

Perspectives

faster and more efficient filters (denoising and clustering)

better understanding of PCR/sequencer noise

stronger mathematical background (sequence-space)

robust α, β statistics able to deal with noise

repeated experiments (technical/biological replicates)