

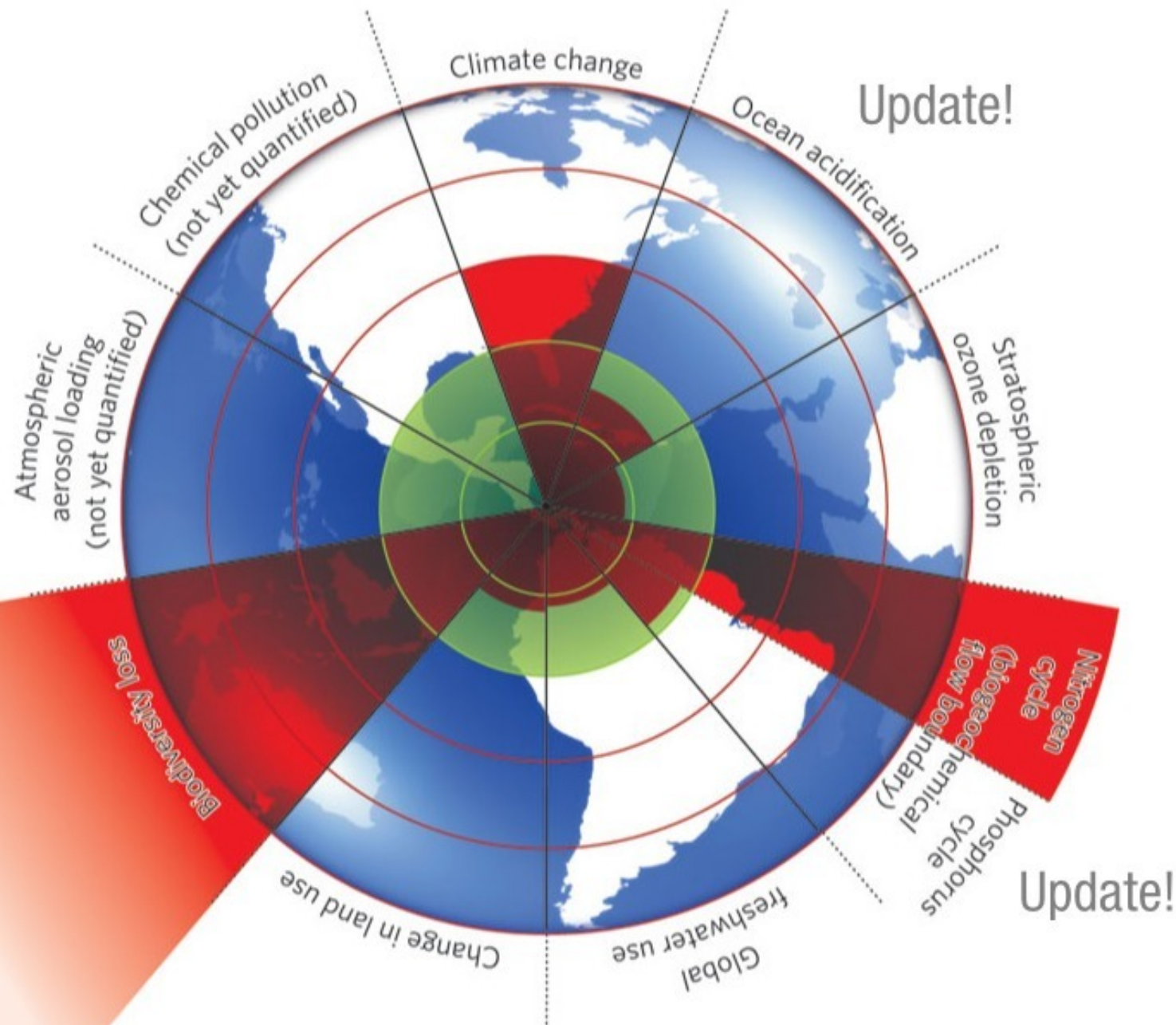
Sequencing technologies & metagenomics

| bioinformatics progresses & challenges

Frédéric Mahé
Nov. 25, 2016
Marseille

Planetary boundaries and human development

Rockström et al. (2009) Nature 461 and Steffen et al., (2015) Science 347:6223



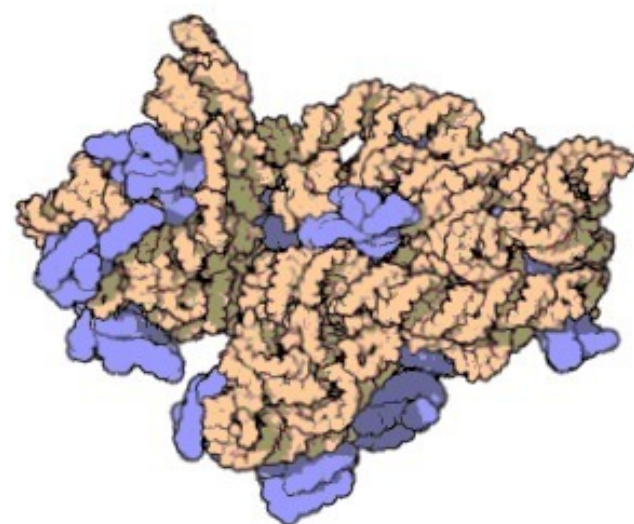
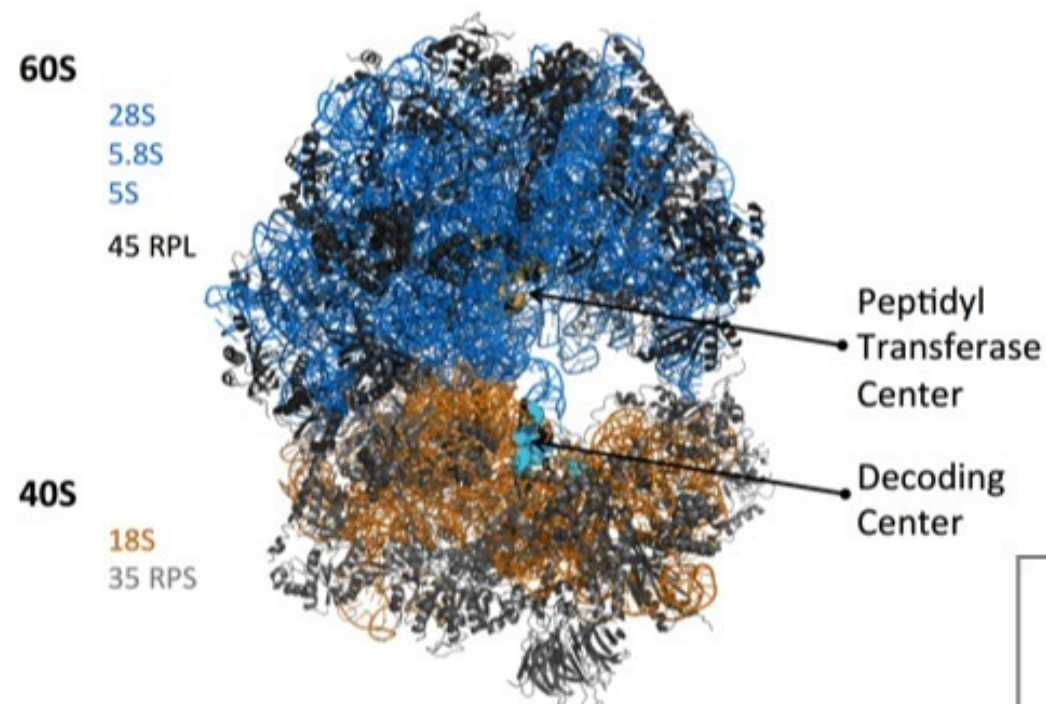
Biodiversity loss (genetic and functional) is a major short-term threat

How can we measure biodiversity?



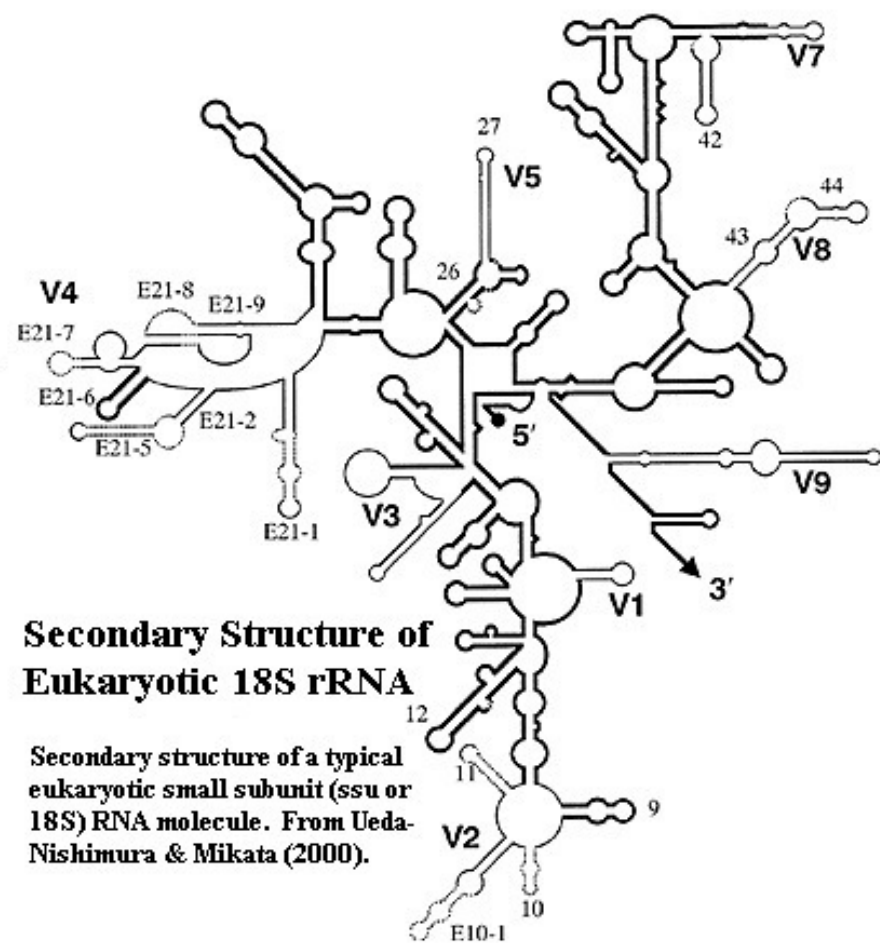
Small organisms play a major role but are hard to inventory

A universal gene: ribosomal RNA



Small Sub-Unit (SSU)

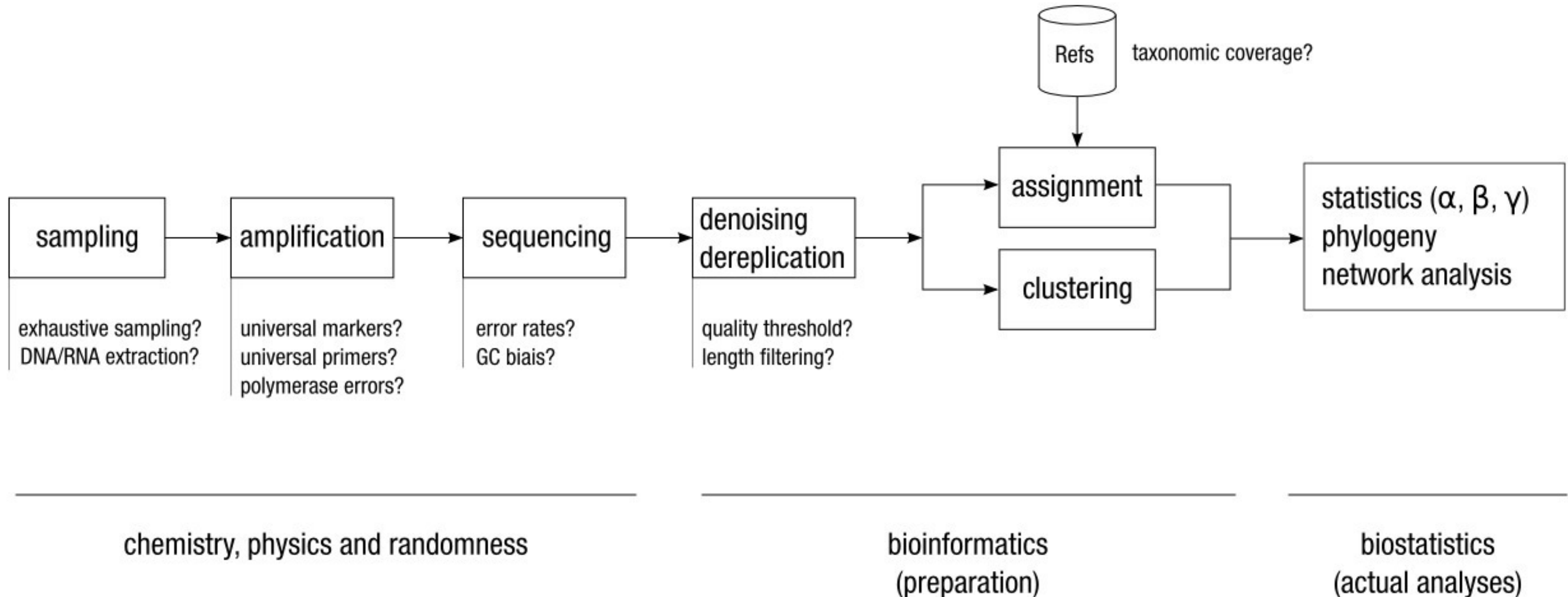
16S	Bacteria Archaea Mitochondria Chloroplasts
18S	Eukaryota



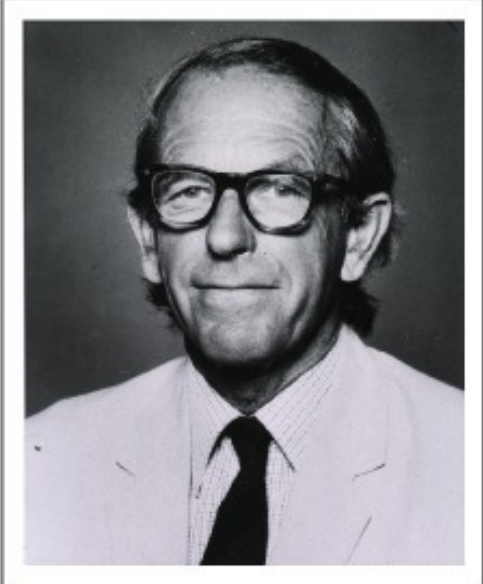
other markers can be used (e.g., ITS). Requirements are: conserved distal regions for primers, variable internal regions, and available sets of reference sequences.

Environmental Metagenomics

targeted amplification

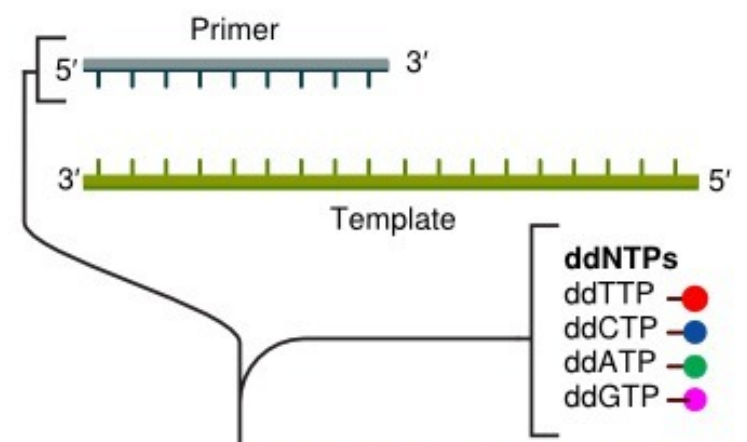


Sanger sequencing technique (1977)

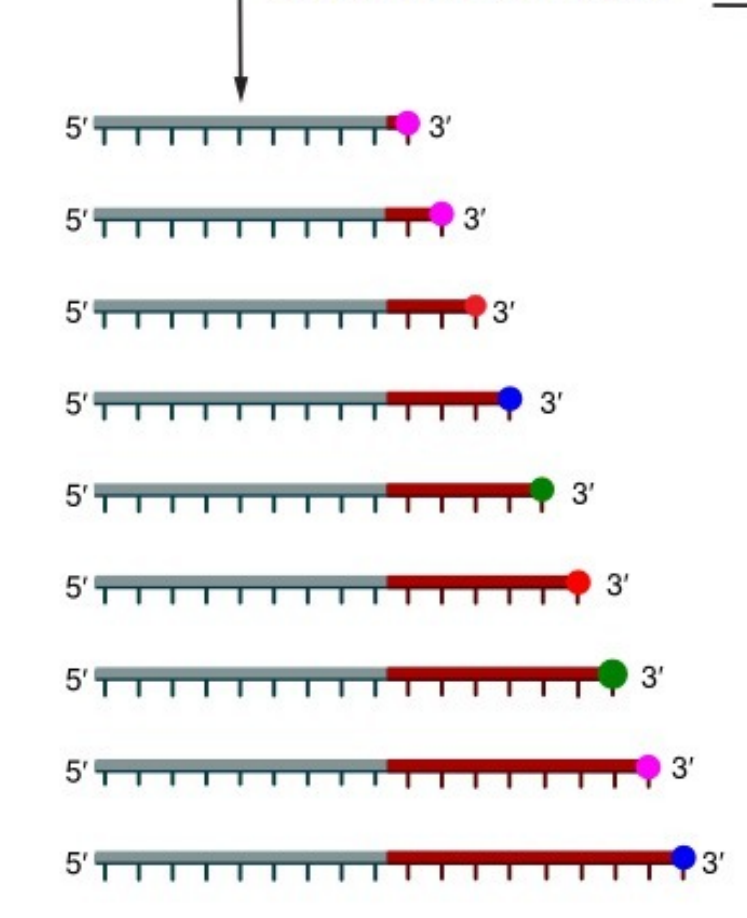


① Reaction mixture

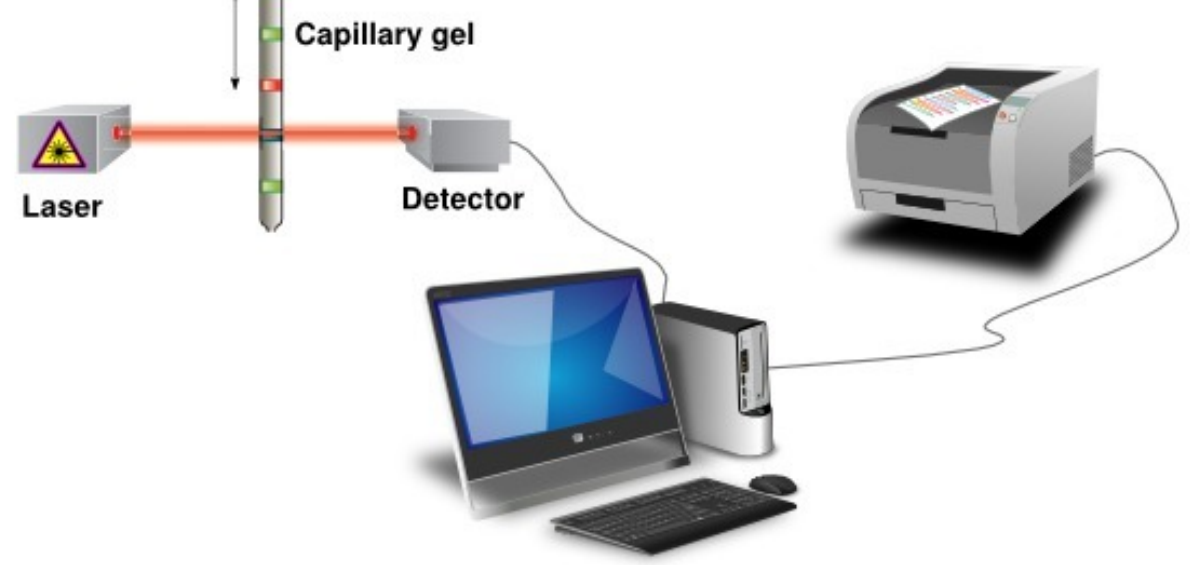
- ▶ Primer and DNA template
- ▶ DNA polymerase
- ▶ ddNTPs with fluorochromes
- ▶ dNTPs (dATP, dCTP, dGTP, and dTTP)



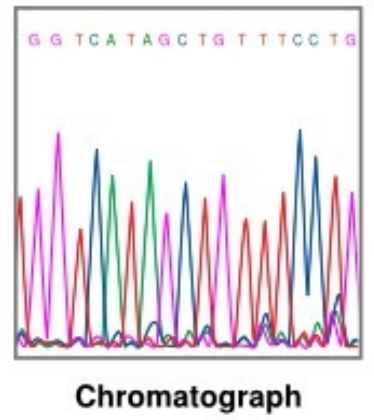
② Primer elongation and chain termination



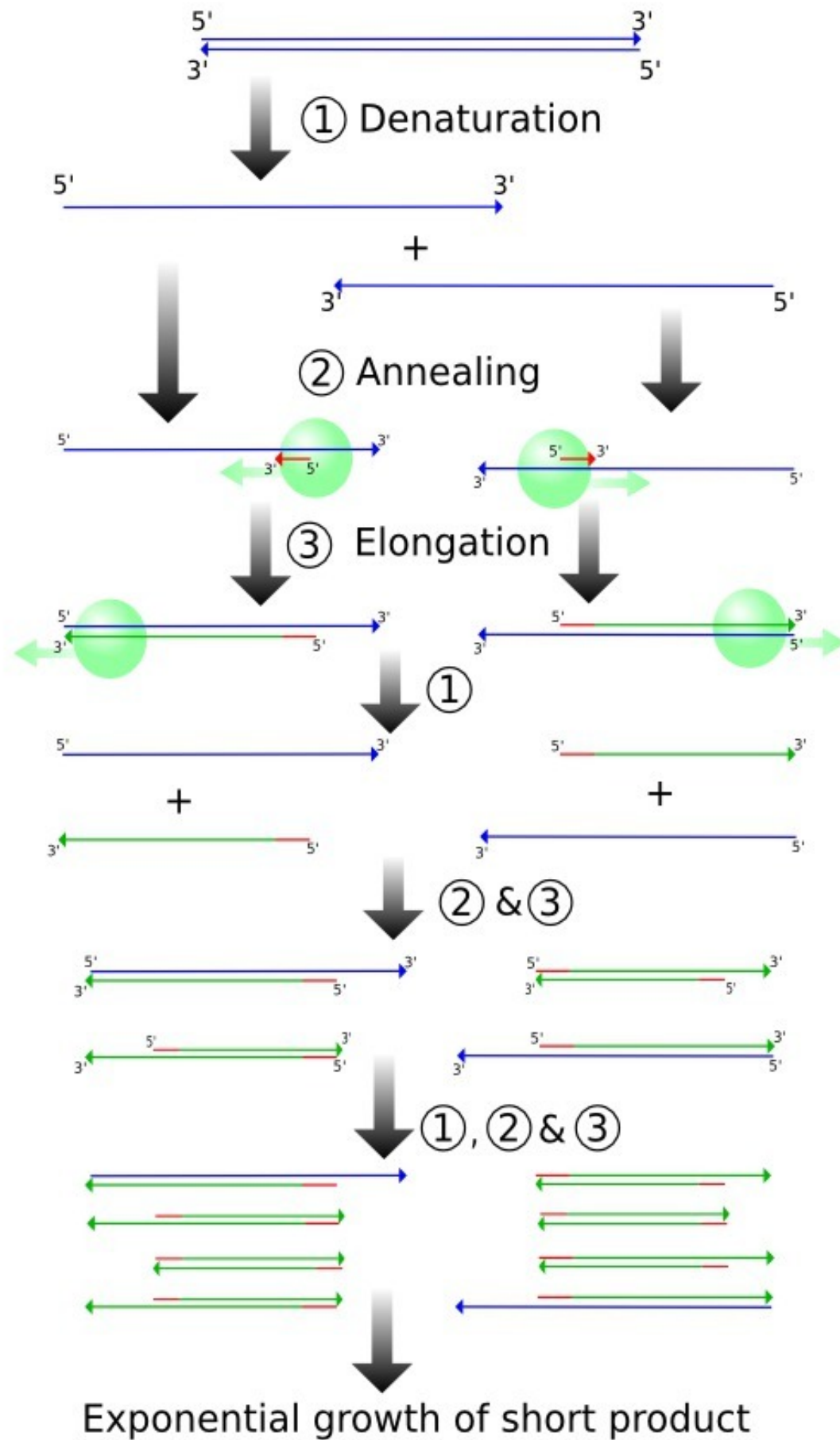
③ Capillary gel electrophoresis separation of DNA fragments



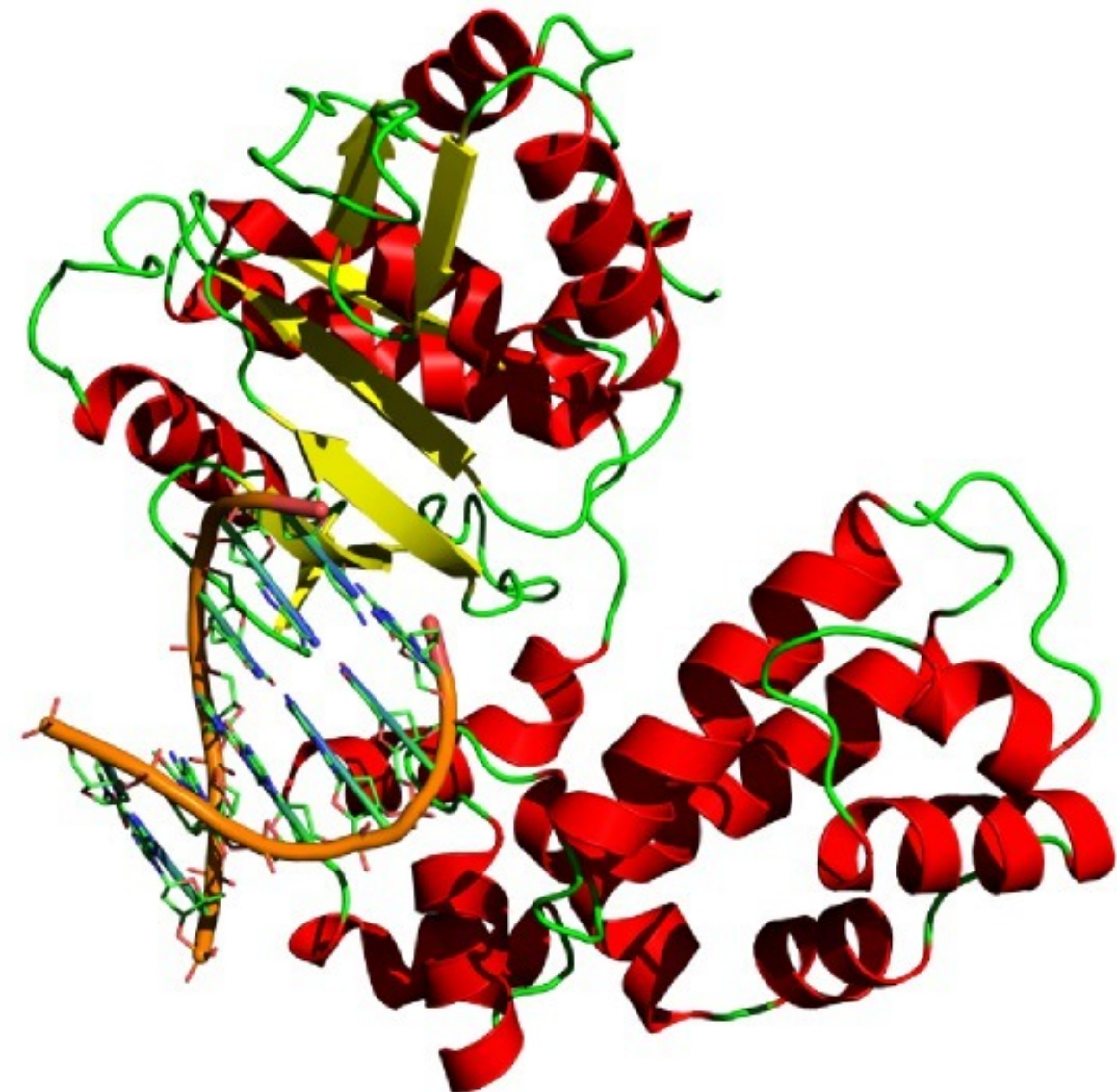
④ Laser detection of fluorochromes and computational sequence analysis



Polymerase chain reaction (1983)

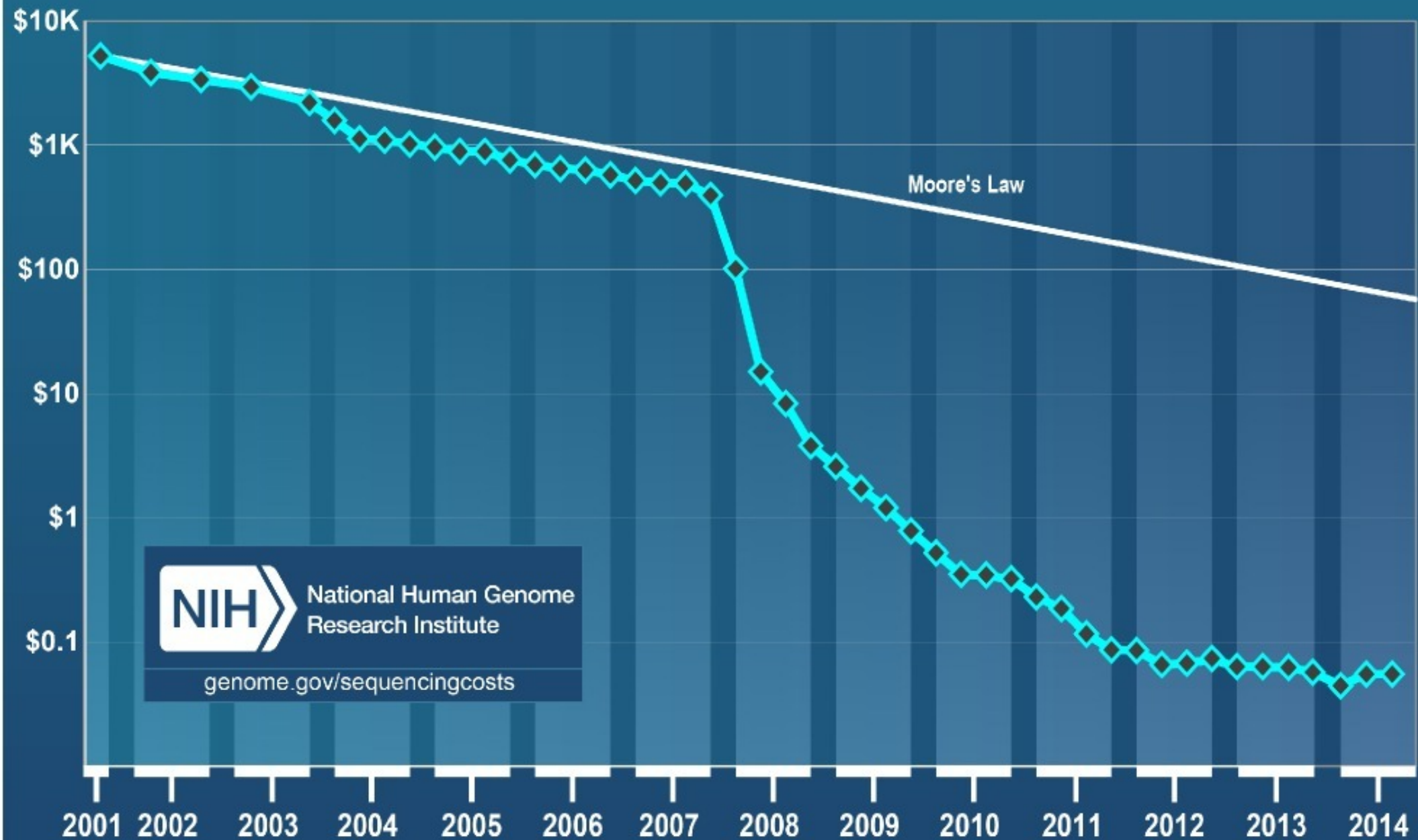


Kary Mullis (Nobel Prize 1993)



Yikrazuul CC 3.0 By-SA

Cost per Raw Megabase of DNA Sequence



IonTorrent

Ion PGM, Ion Proton

Illumina

HiSeq 3000/4000, MiSeq 3000

allows lots of samples and deep sequencing

Pacific Biosciences

Oxford Nanopore Technologies

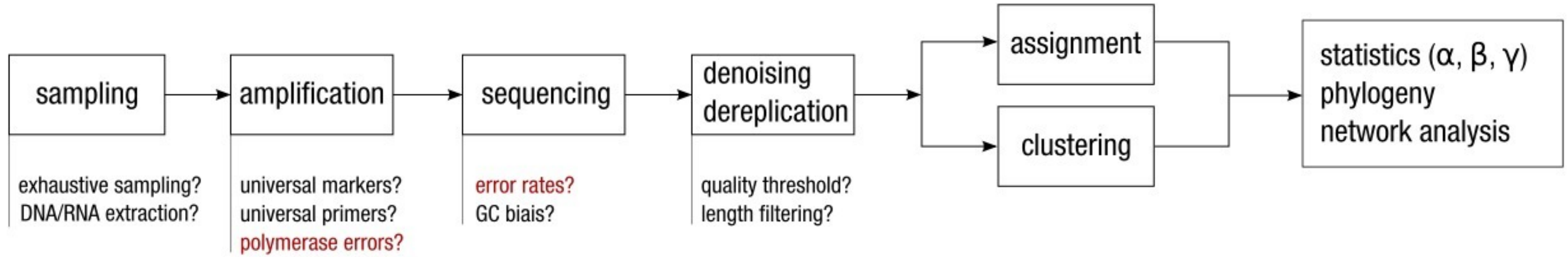


IBM
Blue Gene/P

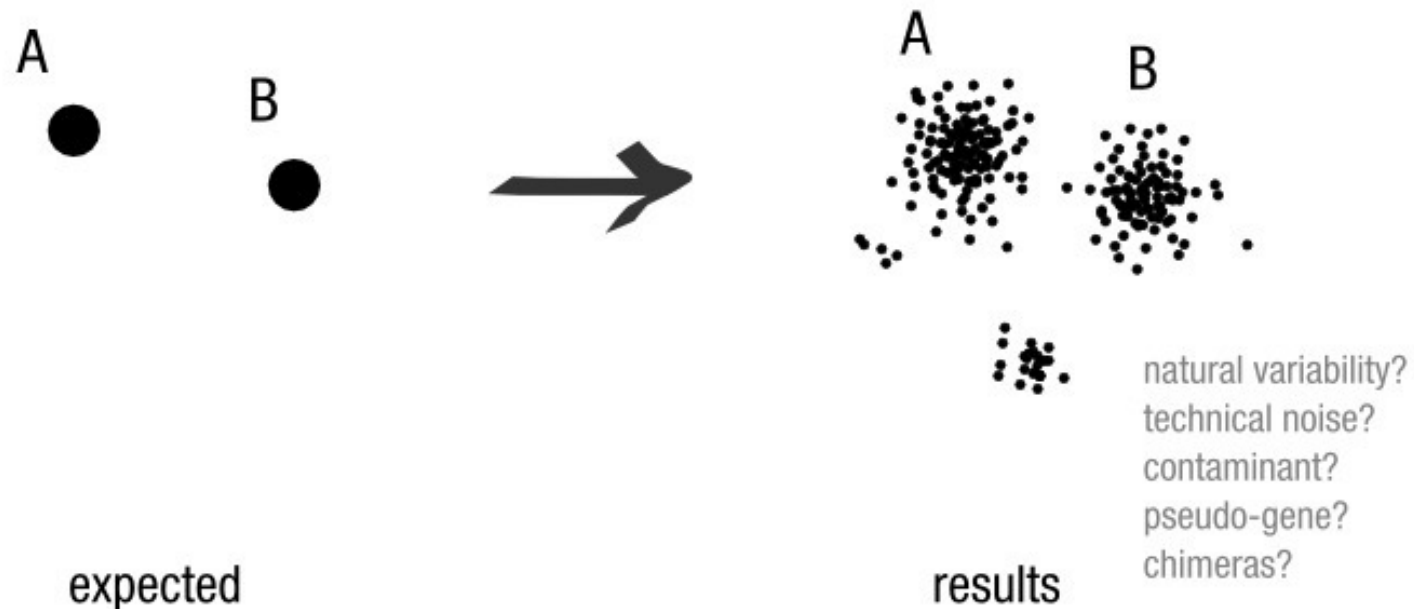
Yes, we need more computers and storage capacity to keep up, but...

Noise is the real challenge

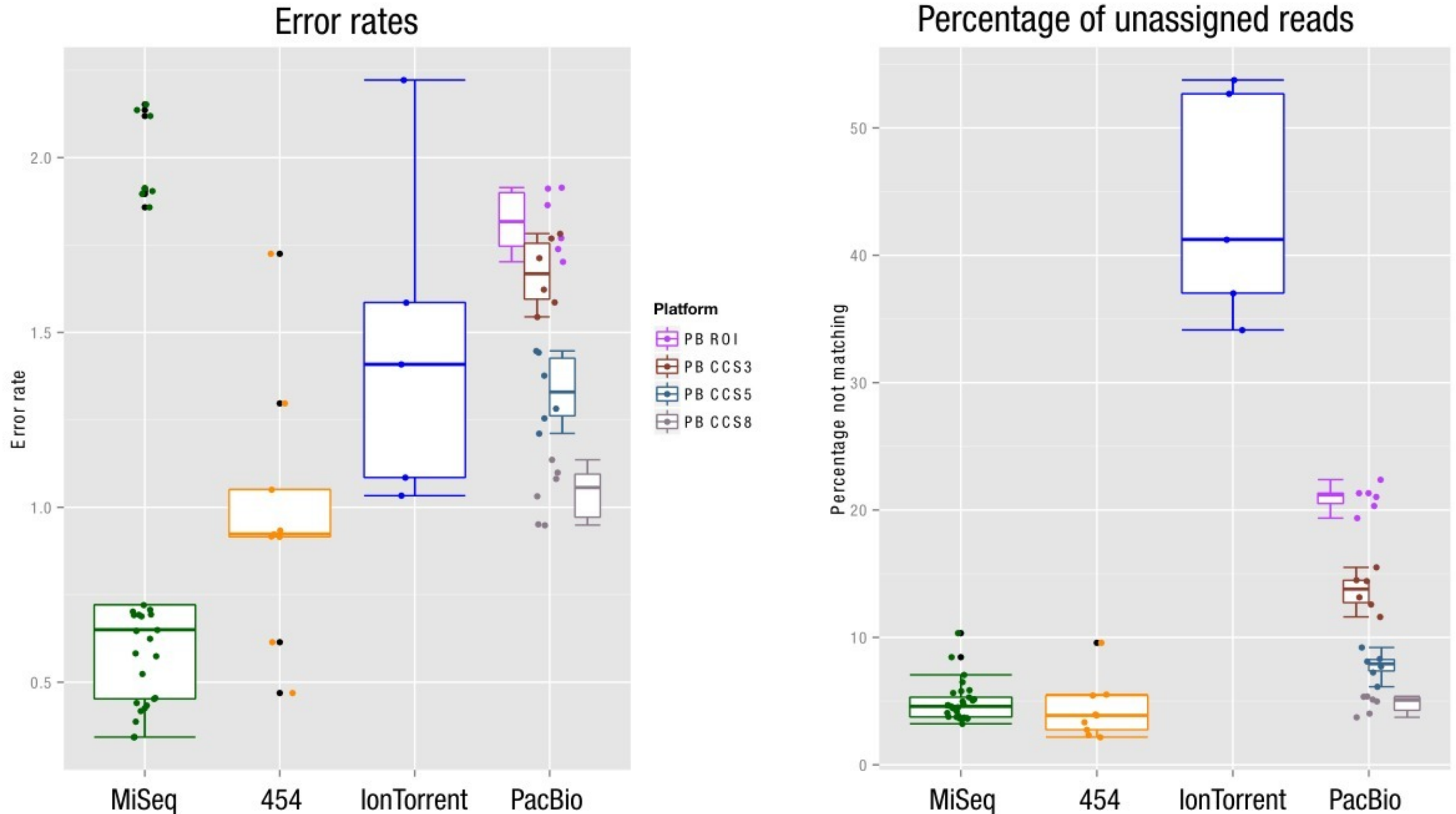
amplification and sequencing are imperfect processes



chemistry, physics and randomness



Error rate per sequencing platform



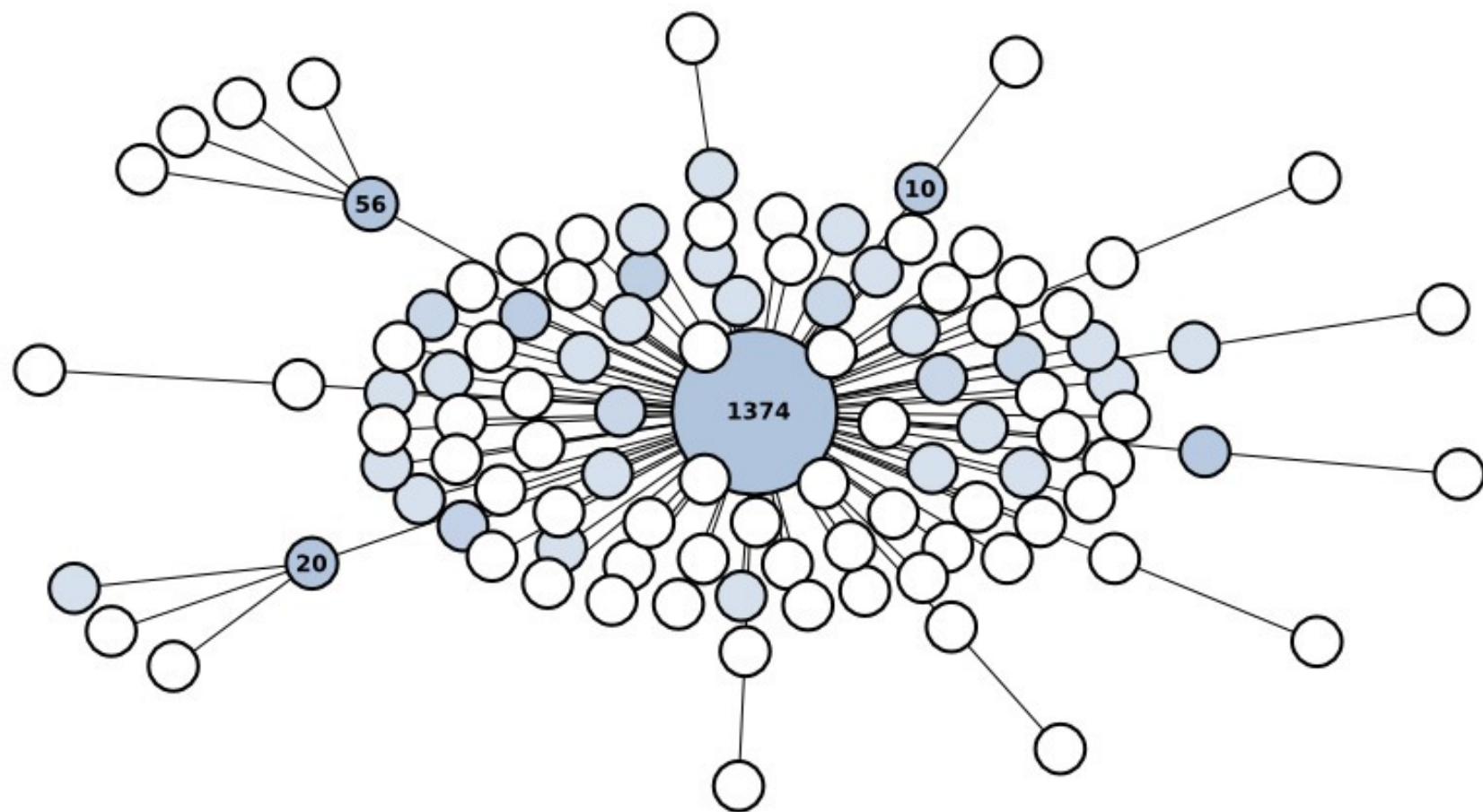
What is the nature of these errors?

2
ACGT
AAGT
AGGT
ATGT
A-GT
AACGT
...

At each position in a sequence:

- 3 substitutions (C -> A, G or T)
- 1 deletion,
- 4 insertions

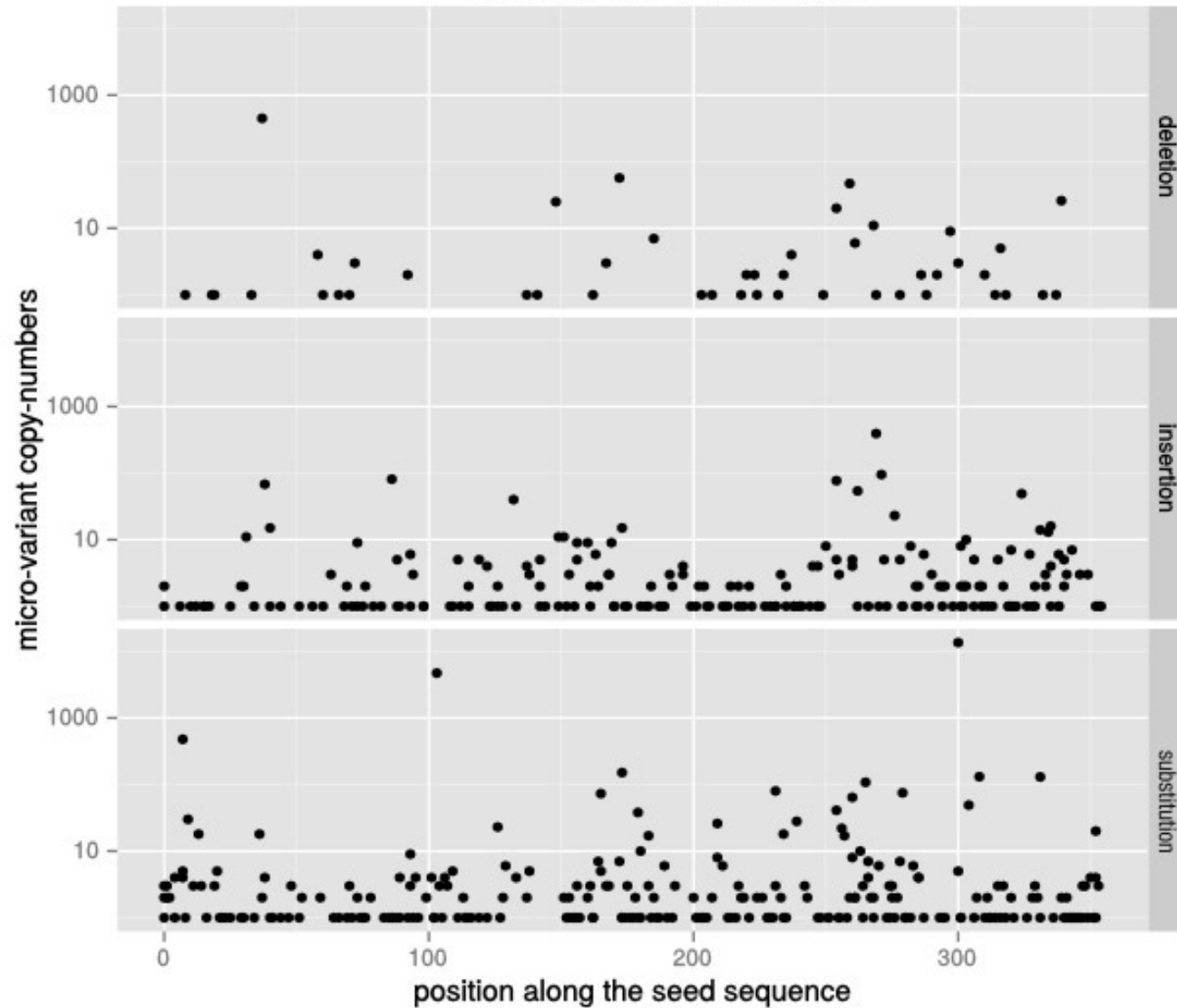
A 380 nucleotide seed (V4), can have up to 2,664 first degree unique micro-variants.



A 130 nucleotide seed (V9), can have up to 914 first degree unique micro-variants.

How are errors distributed?

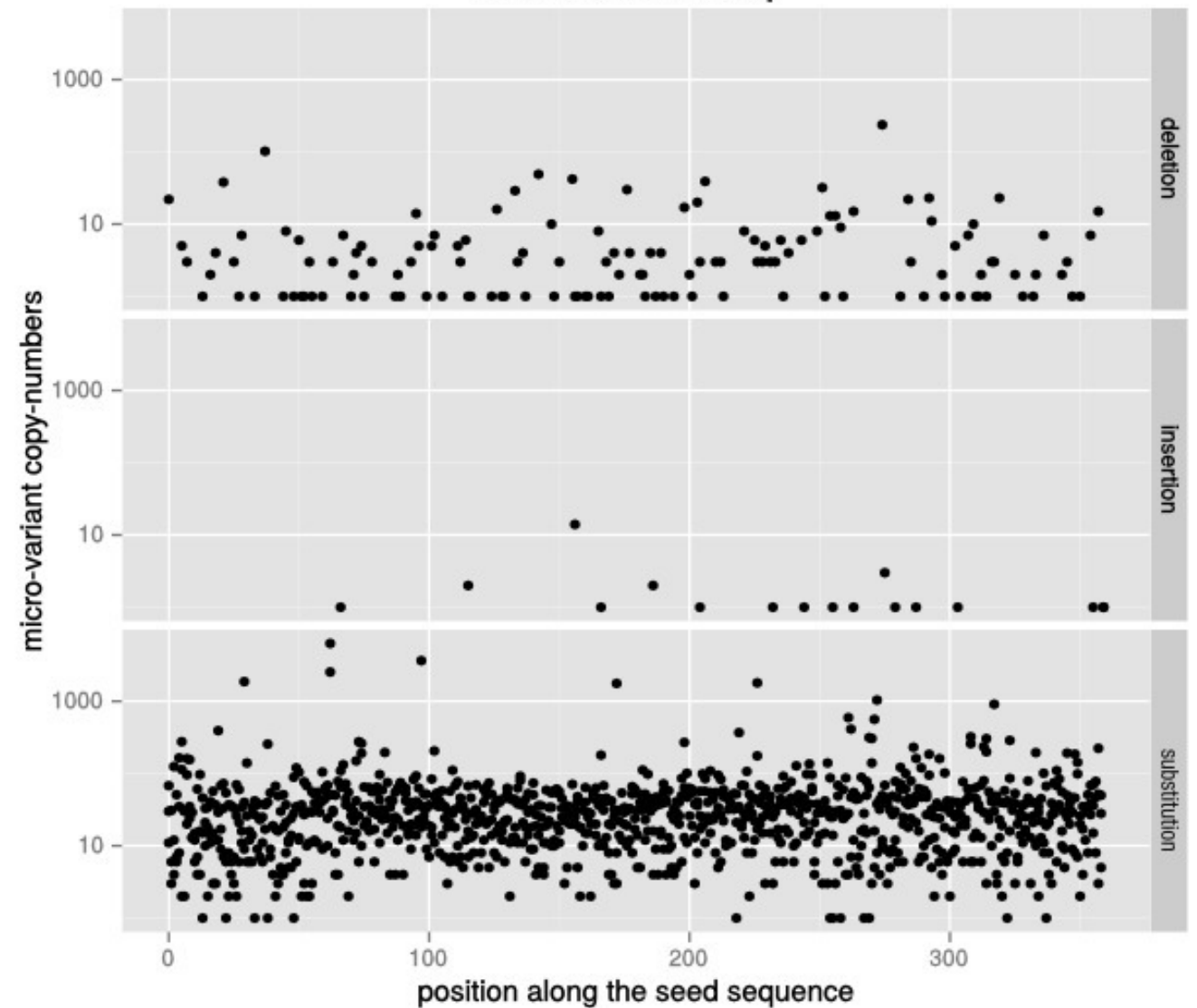
Roche 454 Titanium



Roche 454 Titanium (for 100 bp):

- 0.4011 insertions/deletions,
- 0.0543 substitutions.

Illumina MiSeq

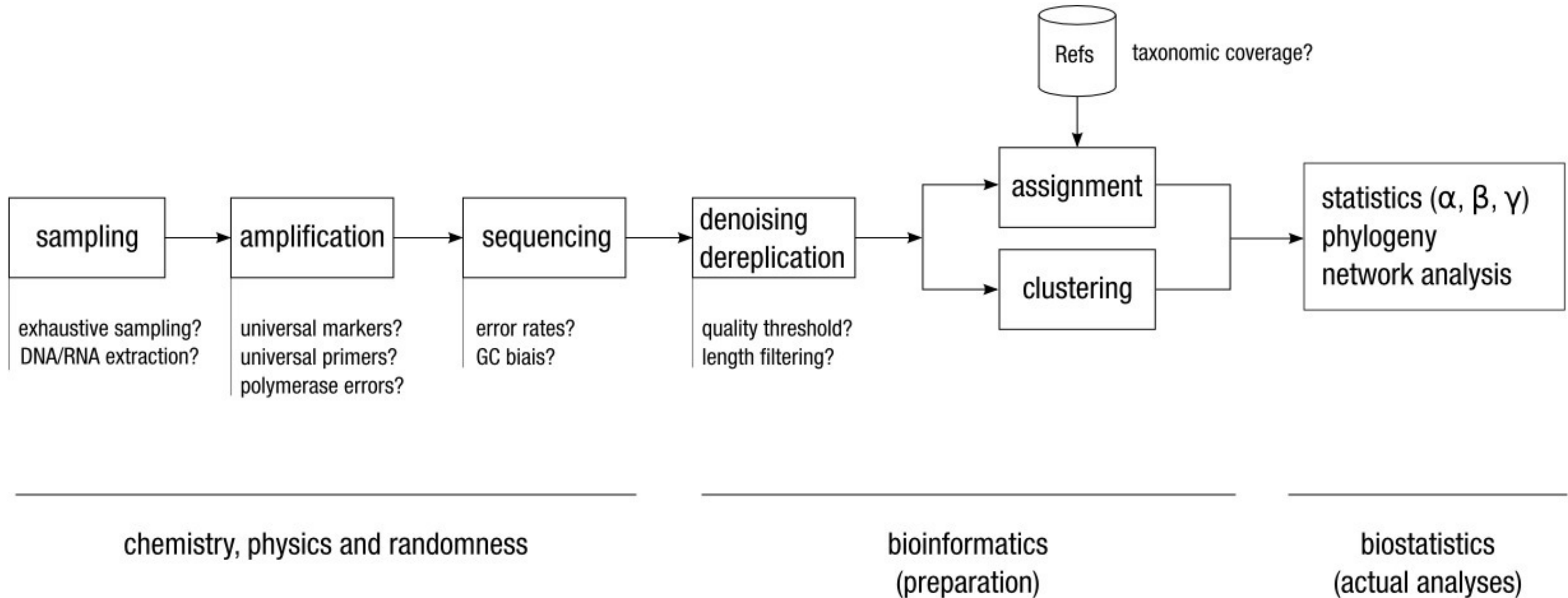


Illumina MiSeq v2 (for 100 bp):

- 0.0009 insertions/deletions,
- 0.0940 substitutions.

Environmental Metagenomics

targeted amplification



Major metagenomics pipelines

targeted amplification



Mothur (2009)
Patrick Schloss
open-source
single piece
most cited
stats

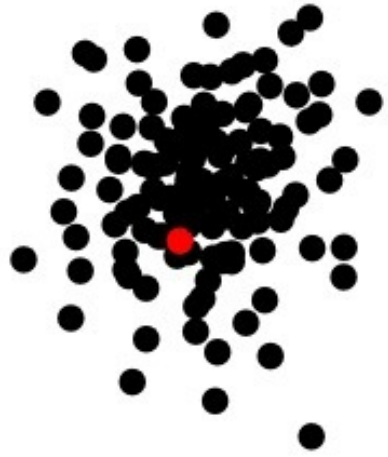


Qiime (2010)
Gregory Caporaso
open-source
python wrapper
most used(?)
stats

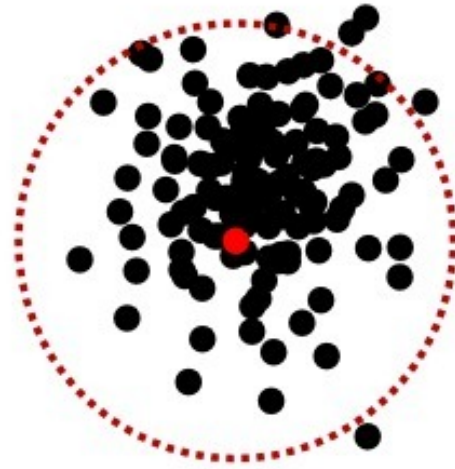
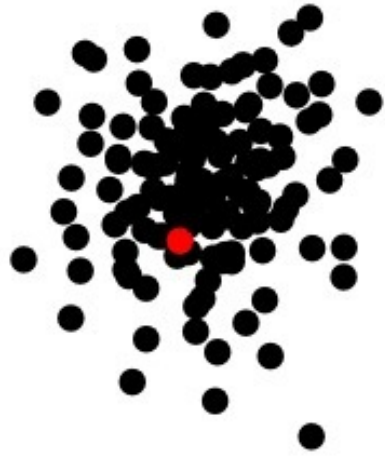
Uparse

Uparse (2013)
Robert Edgar
closed-source
usearch commands
popular
no stats

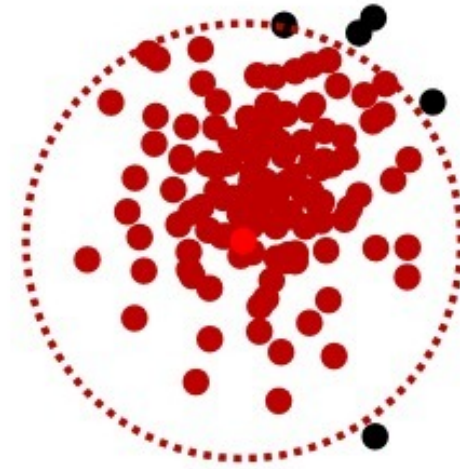
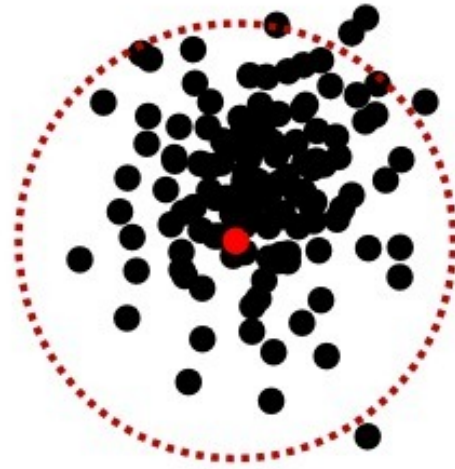
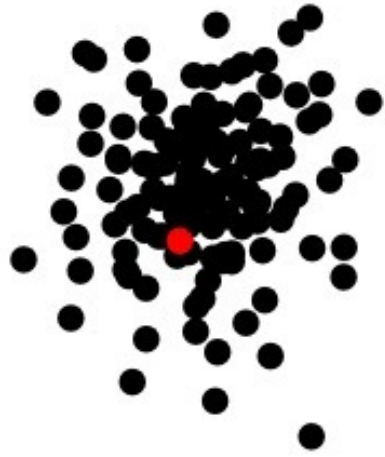
How traditional clustering works?



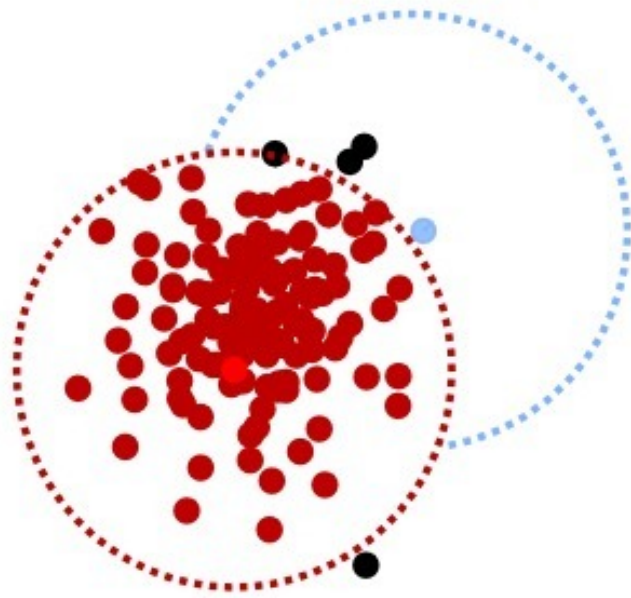
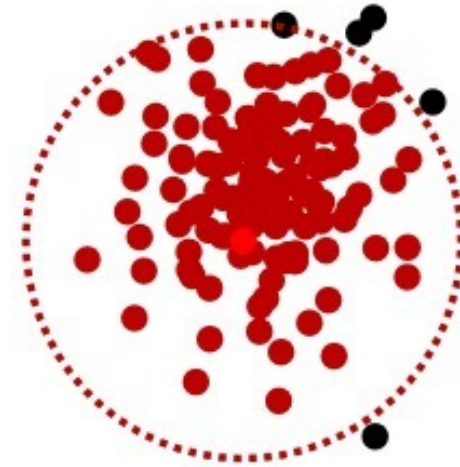
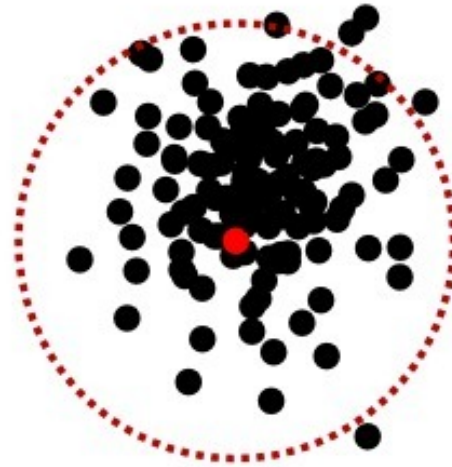
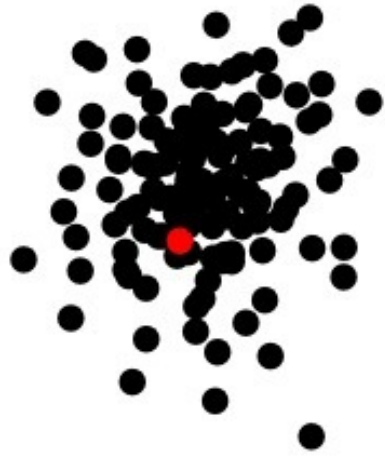
How traditional clustering works?



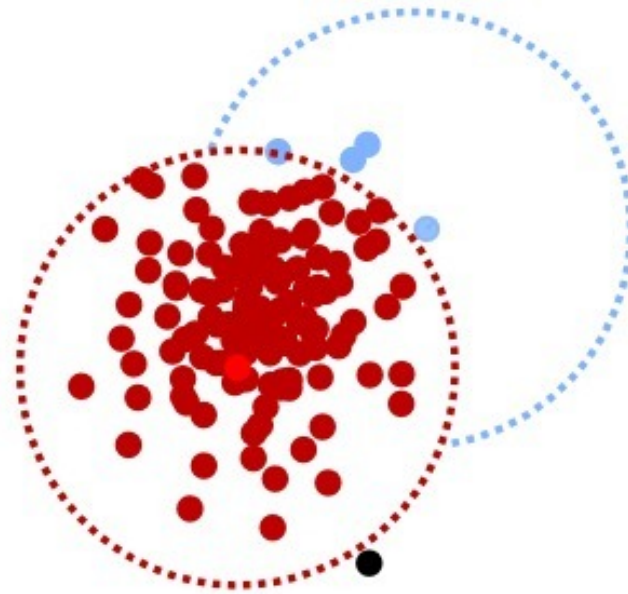
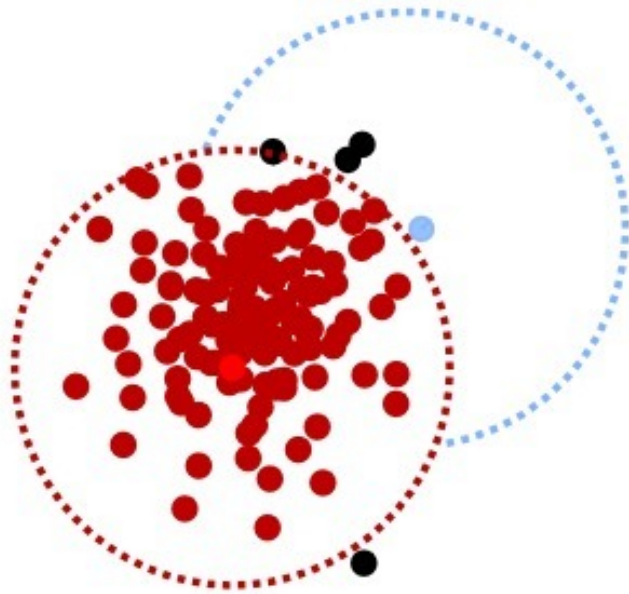
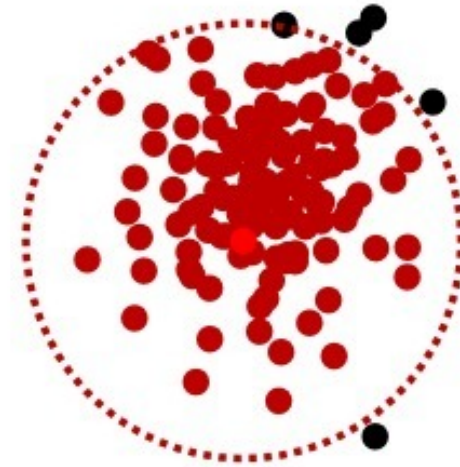
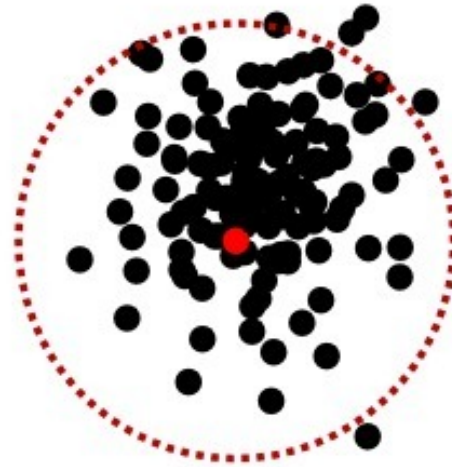
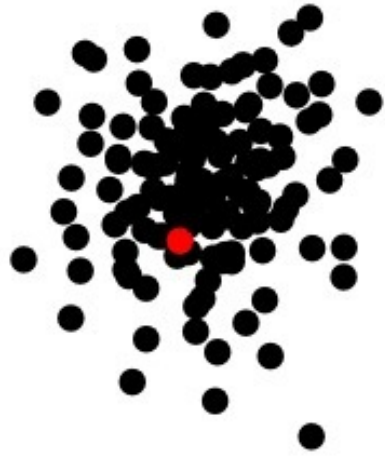
How traditional clustering works?



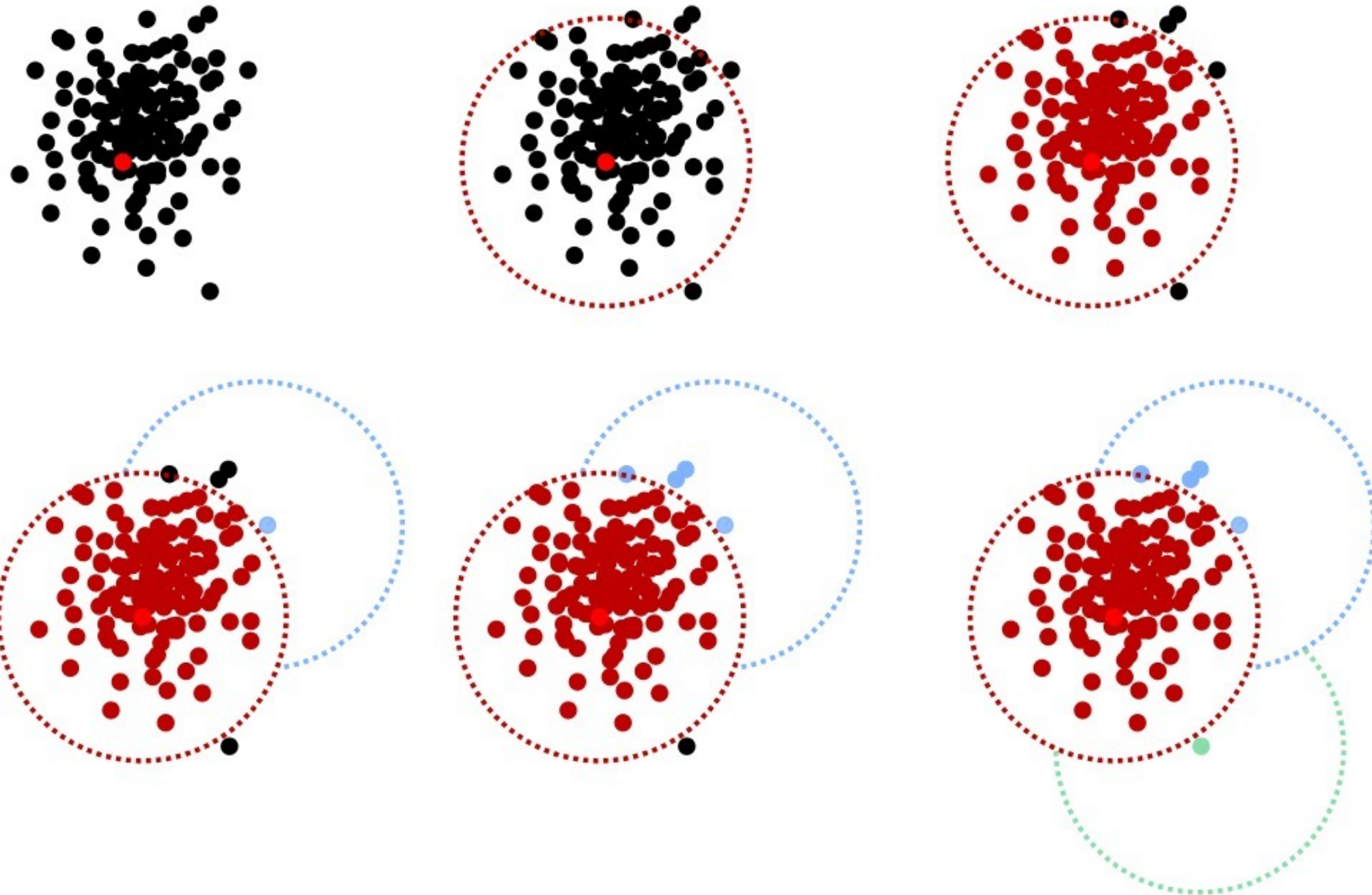
How traditional clustering works?



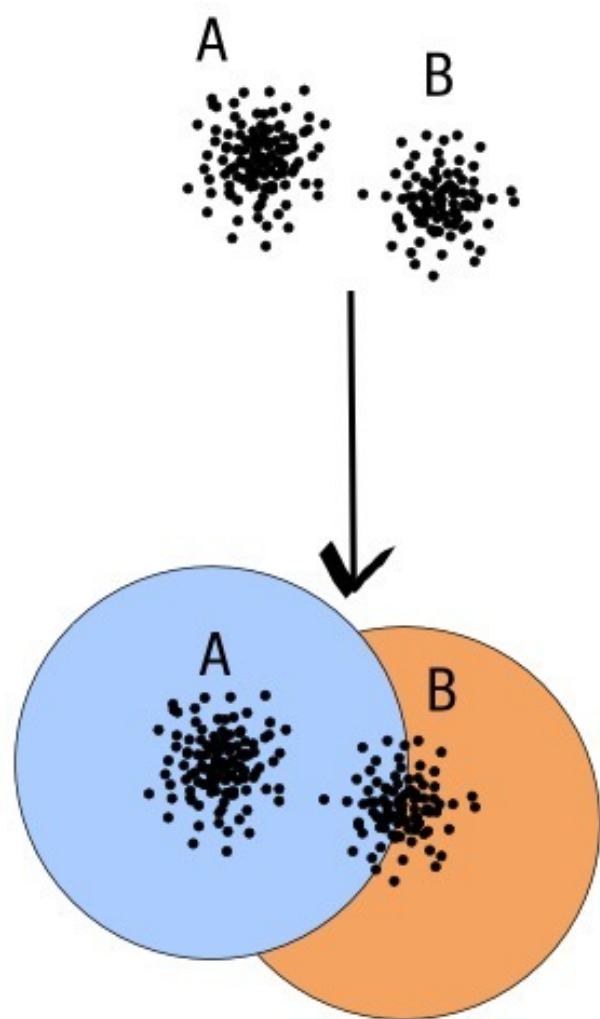
How traditional clustering works?



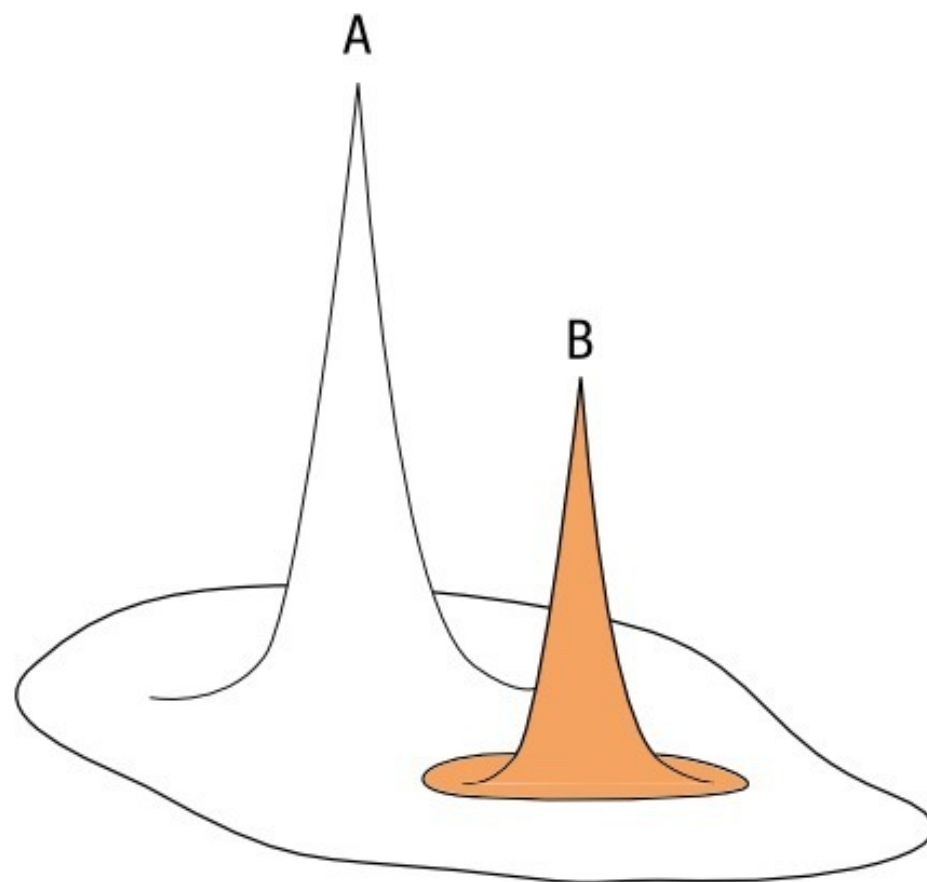
How traditional clustering works?



Swarm: fast, exact and high-resolution clustering



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



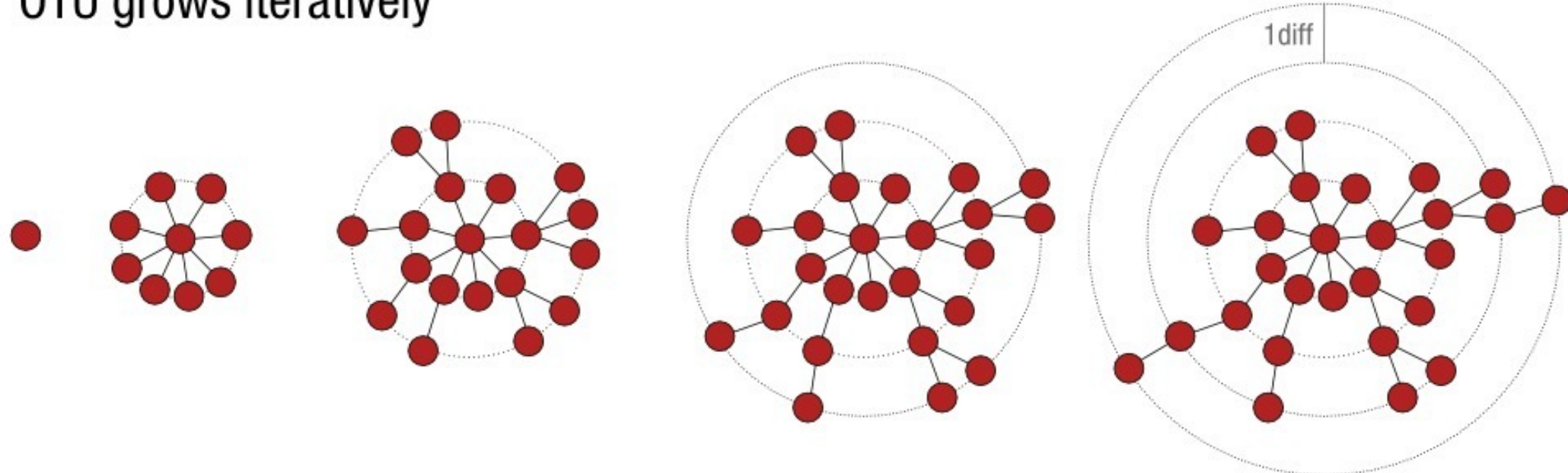
swarm uses abundance values and a new
clustering strategy to delineate natural
high-quality OTUs.

Swarm clustering method

growth phase



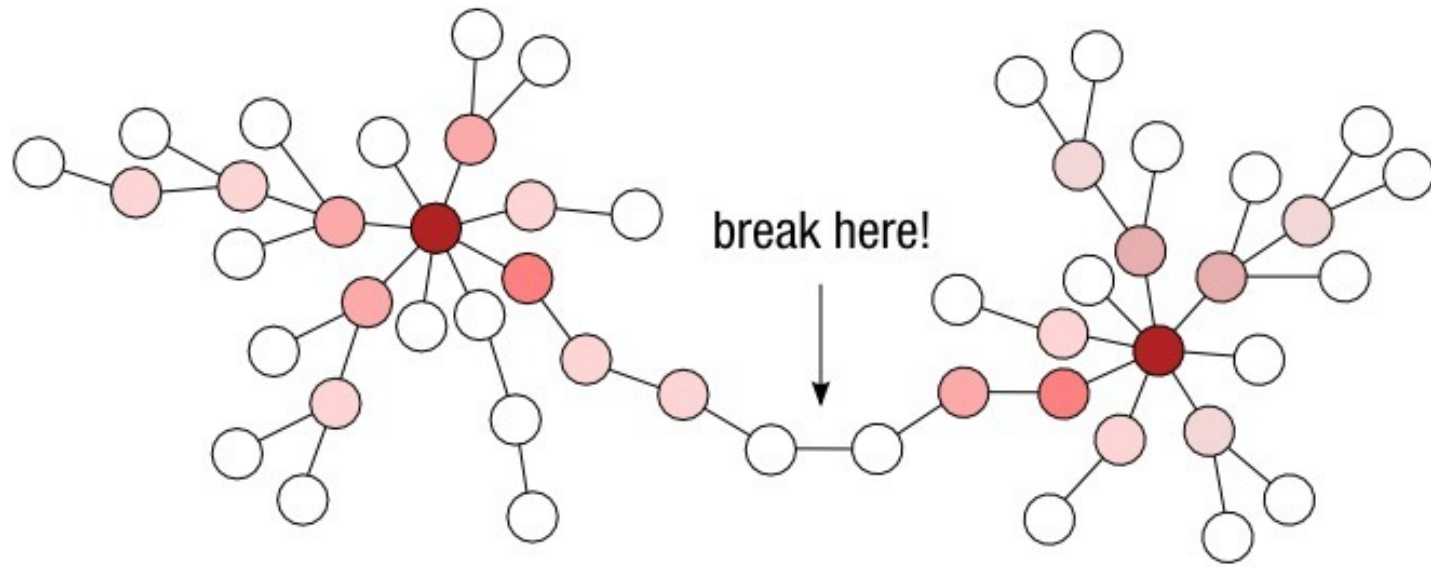
OTU grows iteratively



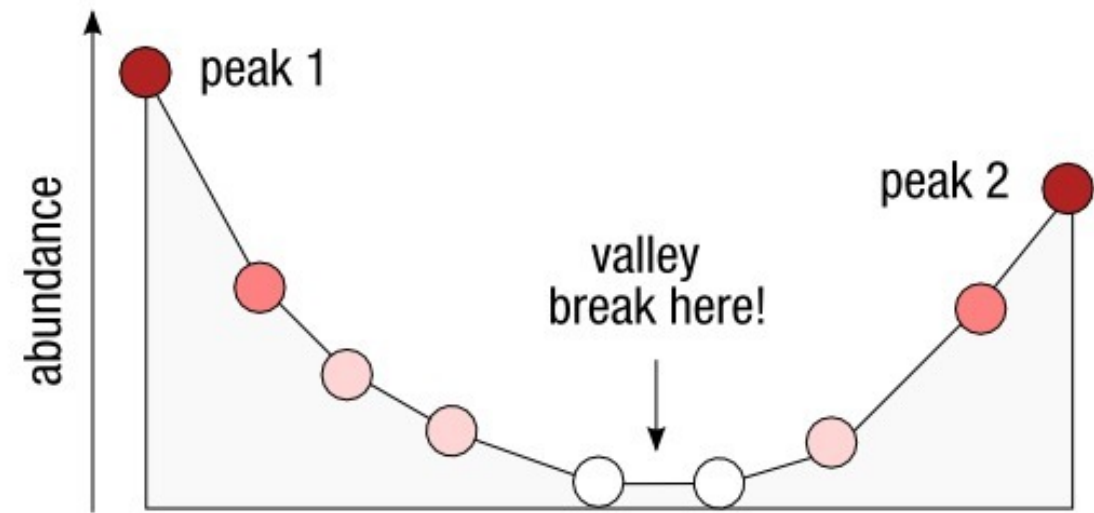
initial seed (randomly picked from amplicon dataset)

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.

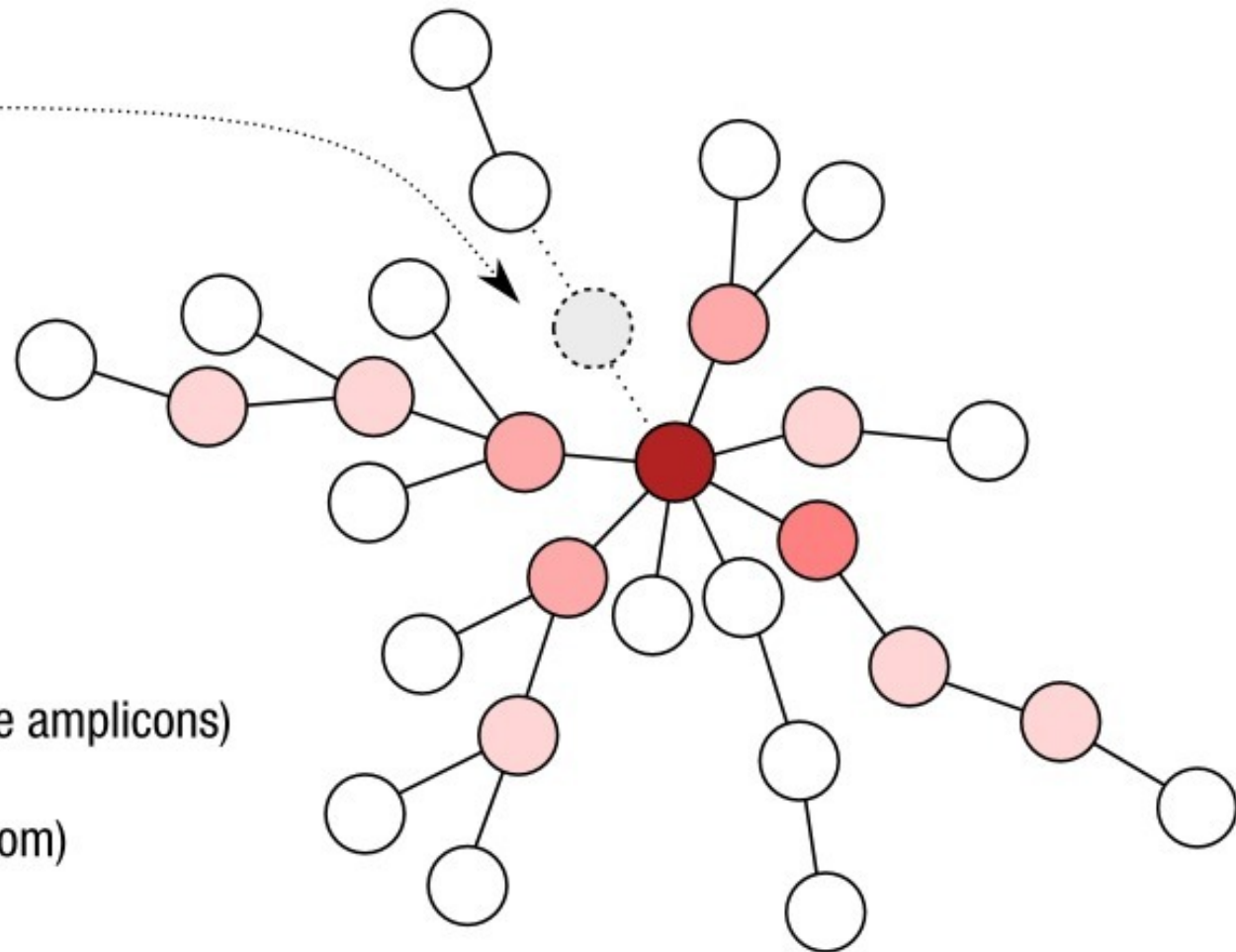
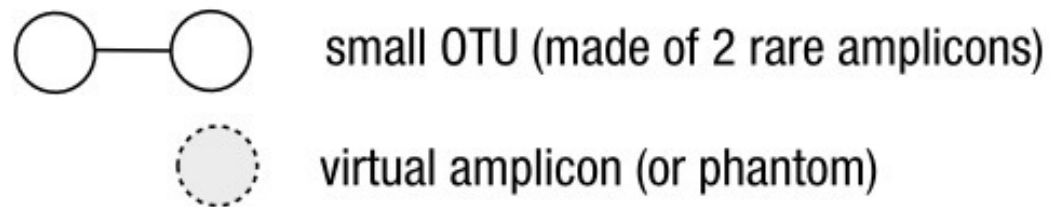


Assuming that original sequences are more abundant than erroneous copies.

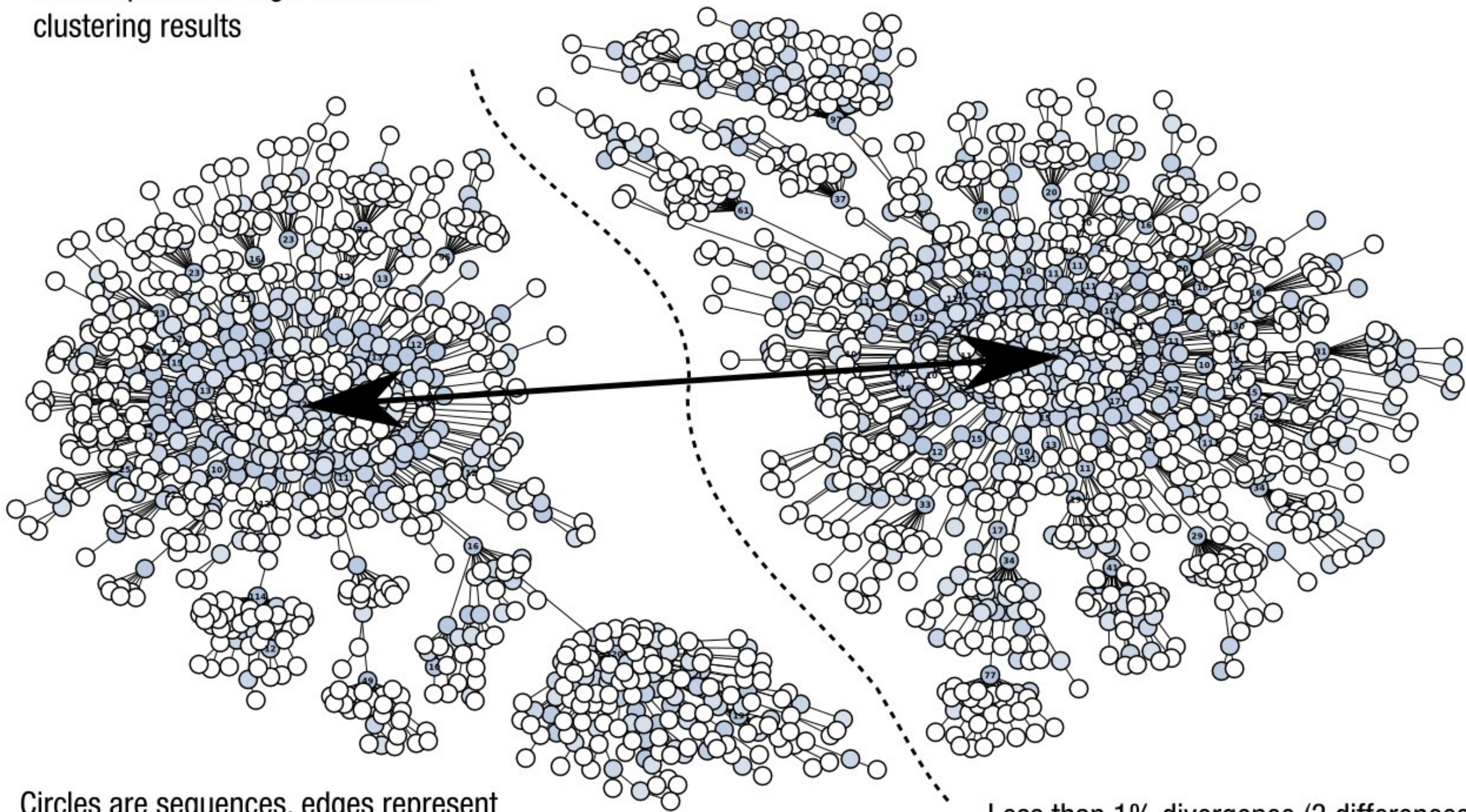
Swarm clustering method

grafting phase

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



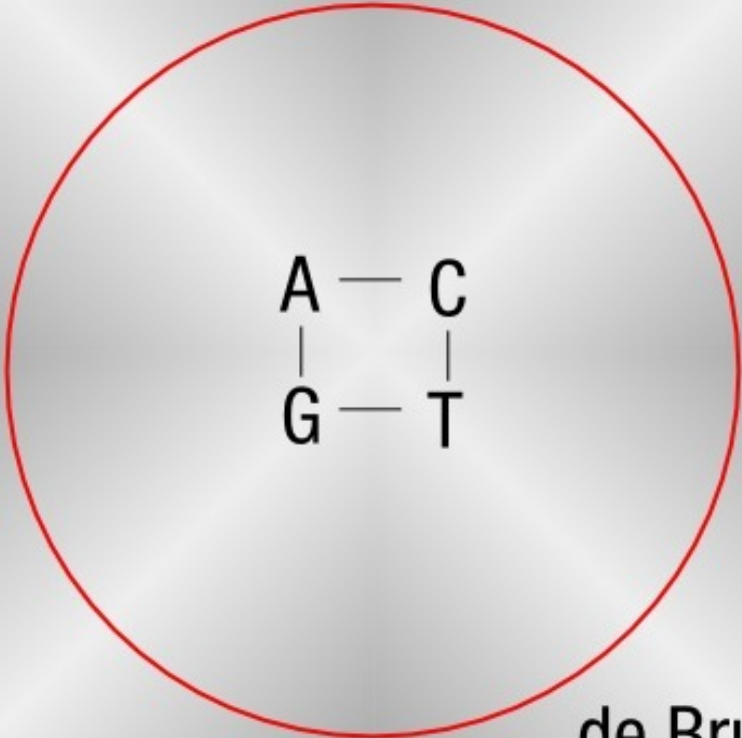
Swarm produces high-resolution clustering results



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

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

sequence-space



A — C
| |
G — T

swarm derives from basic observations of the sequence-space

de Bruijn graph in genome assembly

Swarm 2.0 is a highly scalable denoising-clustering method

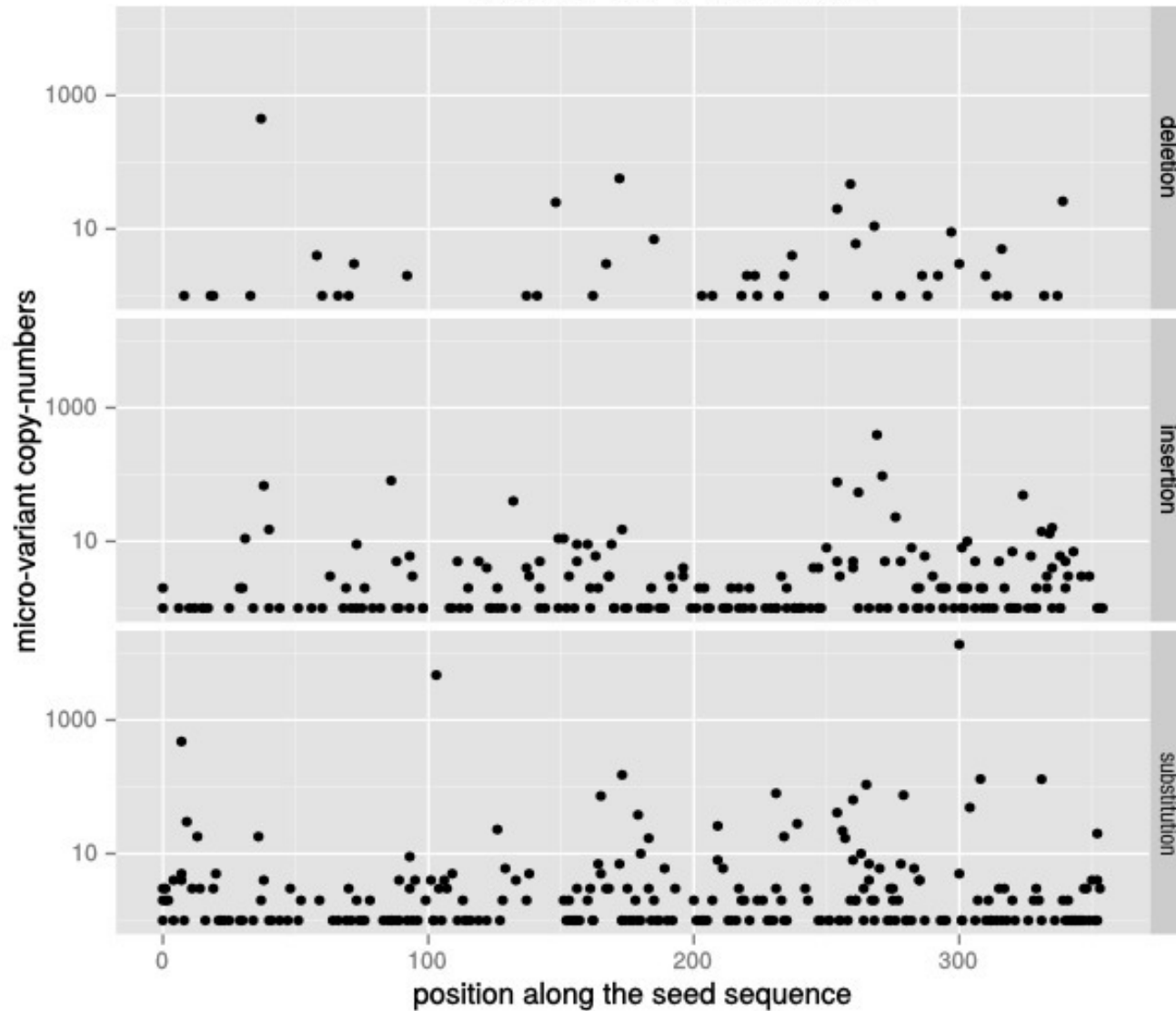


28,275 samples
2.3 billion reads

swarm: 5 hours
usearch: >150 days

How are errors distributed?

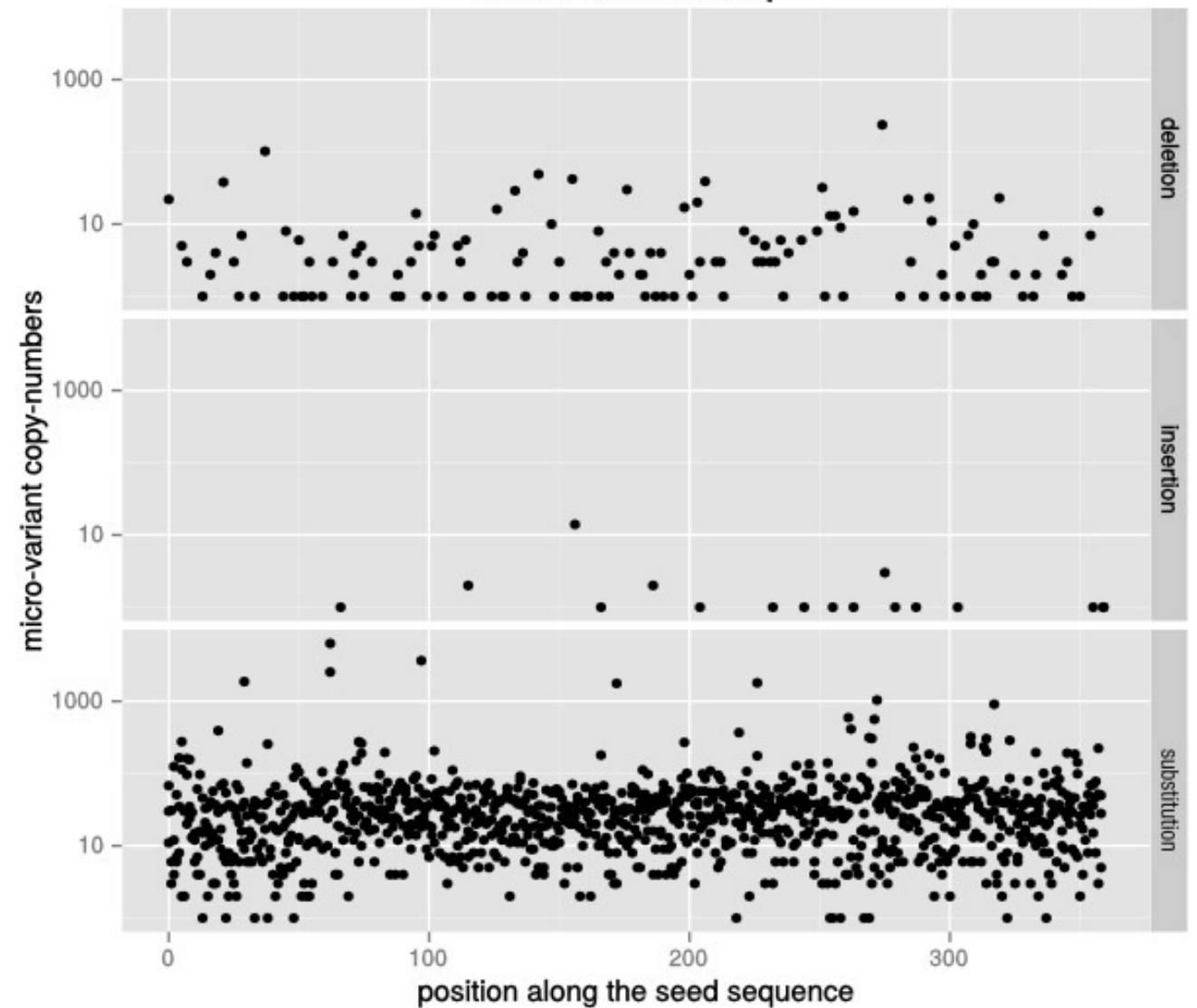
Roche 454 Titanium



Roche 454 Titanium (for 100 bp):

- 0.4011 insertions/deletions,
- 0.0543 substitutions.

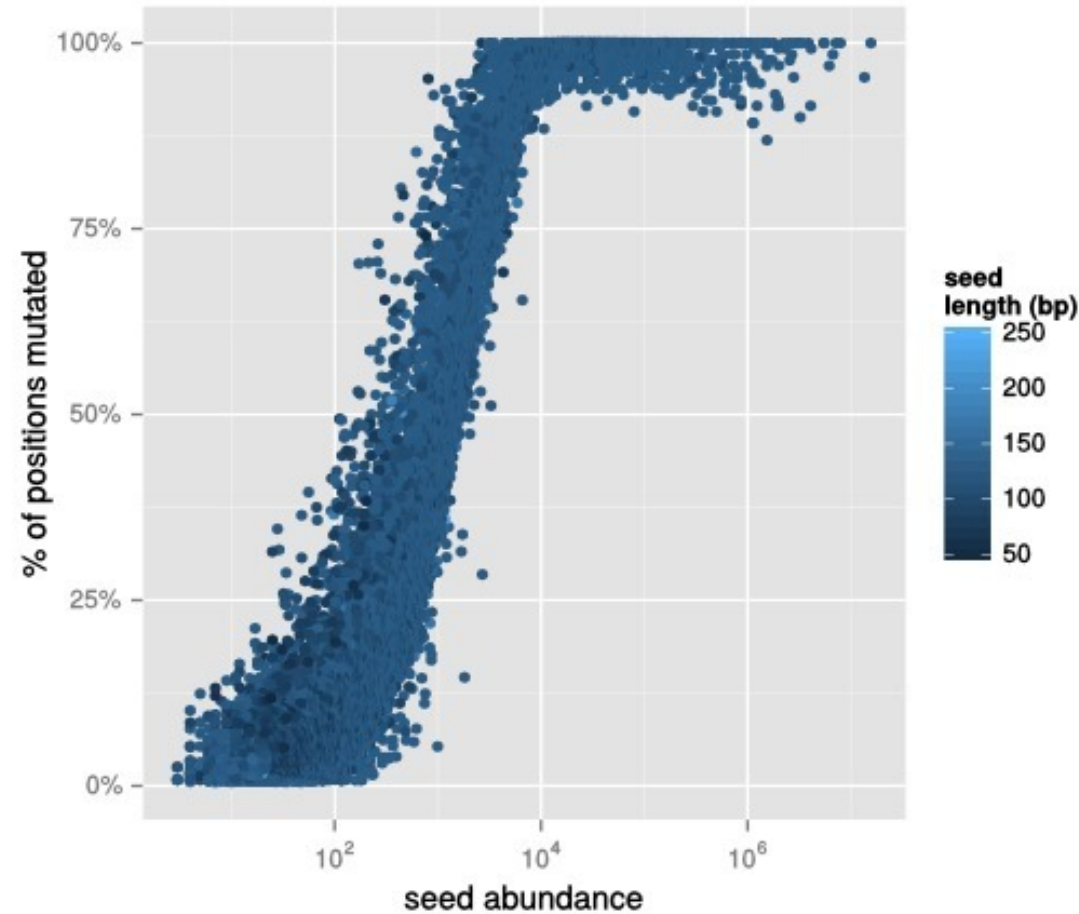
Illumina MiSeq



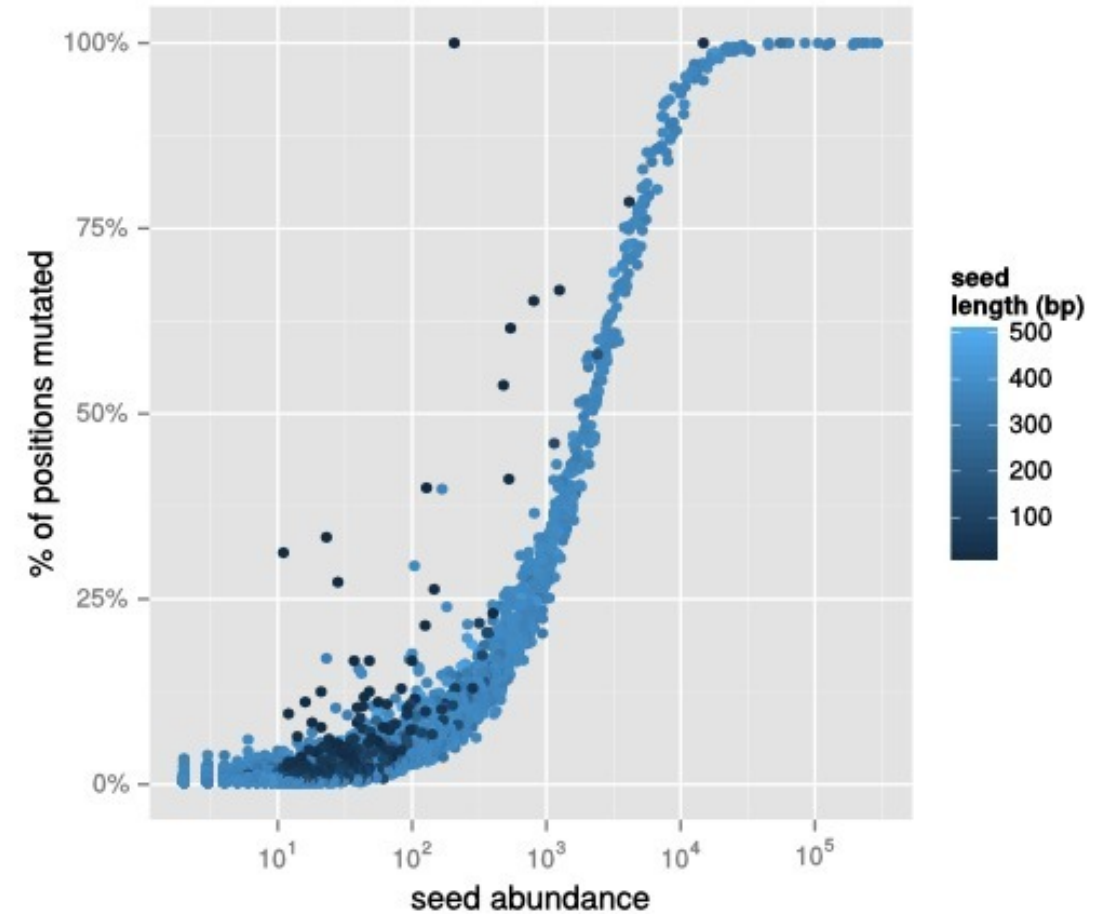
Illumina MiSeq v2 (for 100 bp):

- 0.0009 insertions/deletions,
- 0.0940 substitutions.

How fast errors accumulate with sequencing depth?



TARA OCEANS
V9

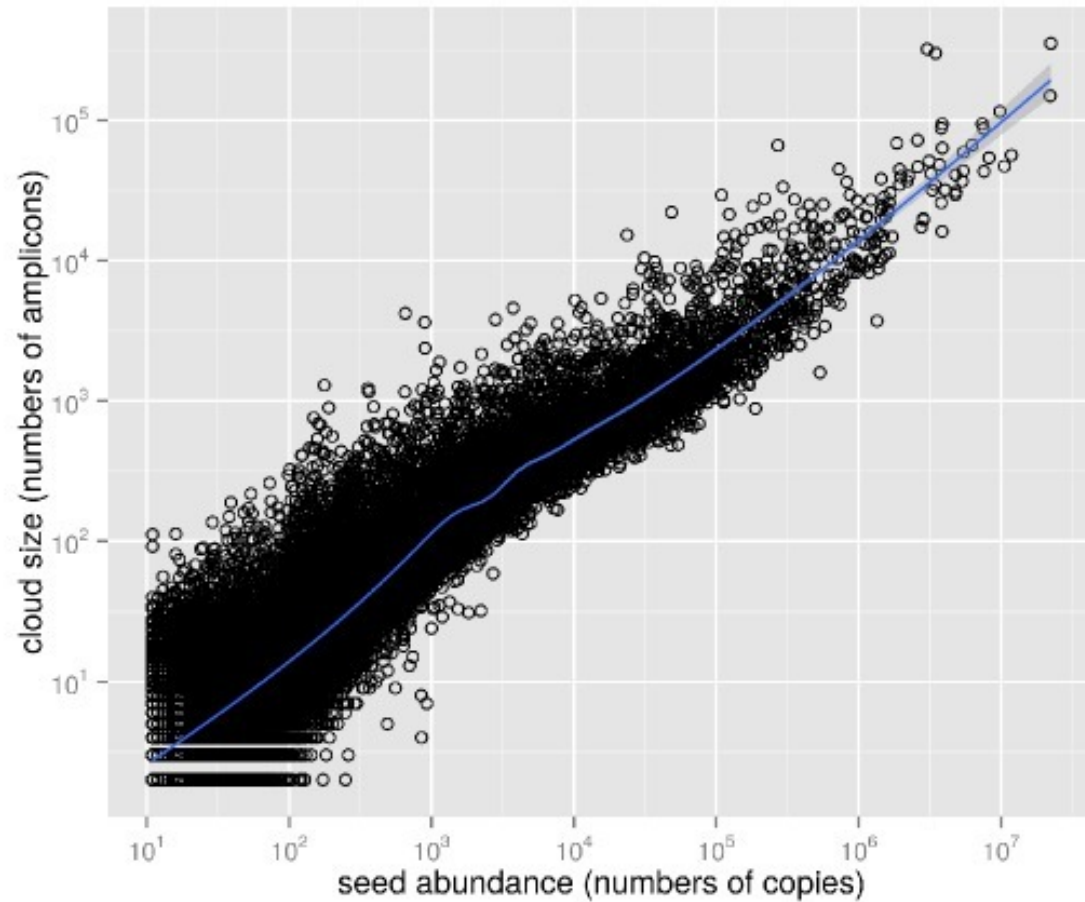


Neotropical Forests Soils
V4

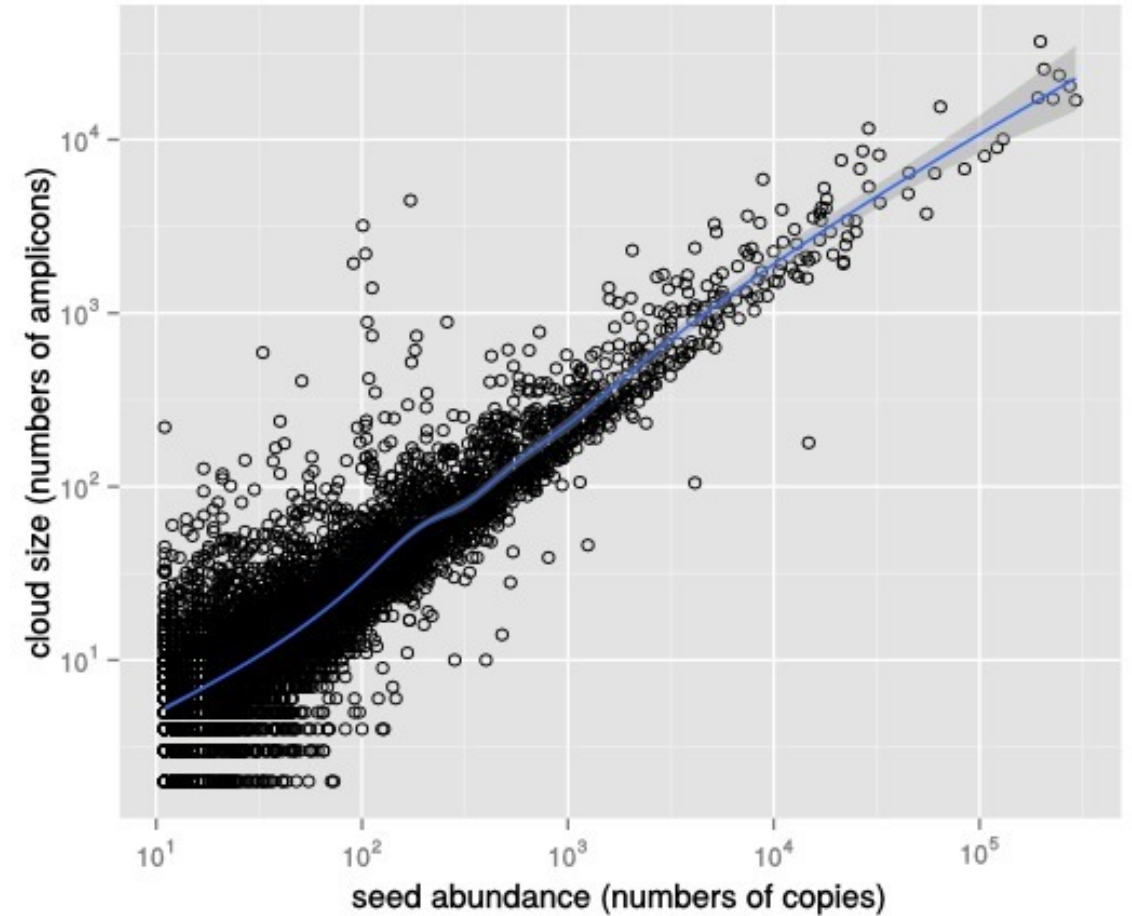
Saturation curve depends on the average seed length and sequencing platform, not on sequencing depth.

Cloud size (and diameter) will vary with the taxa abundance. There is no optimal per-taxa clustering threshold.

Seed versus Cloud



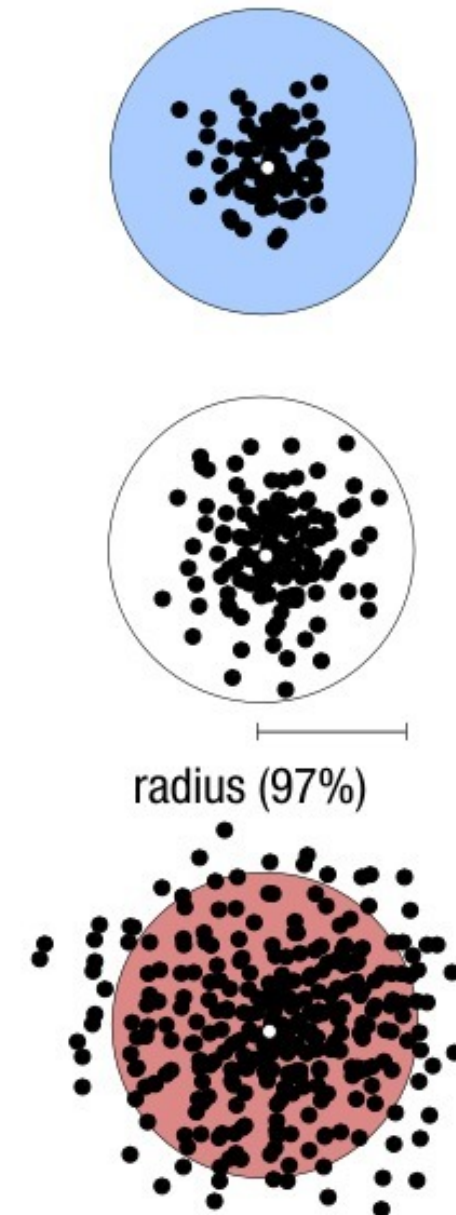
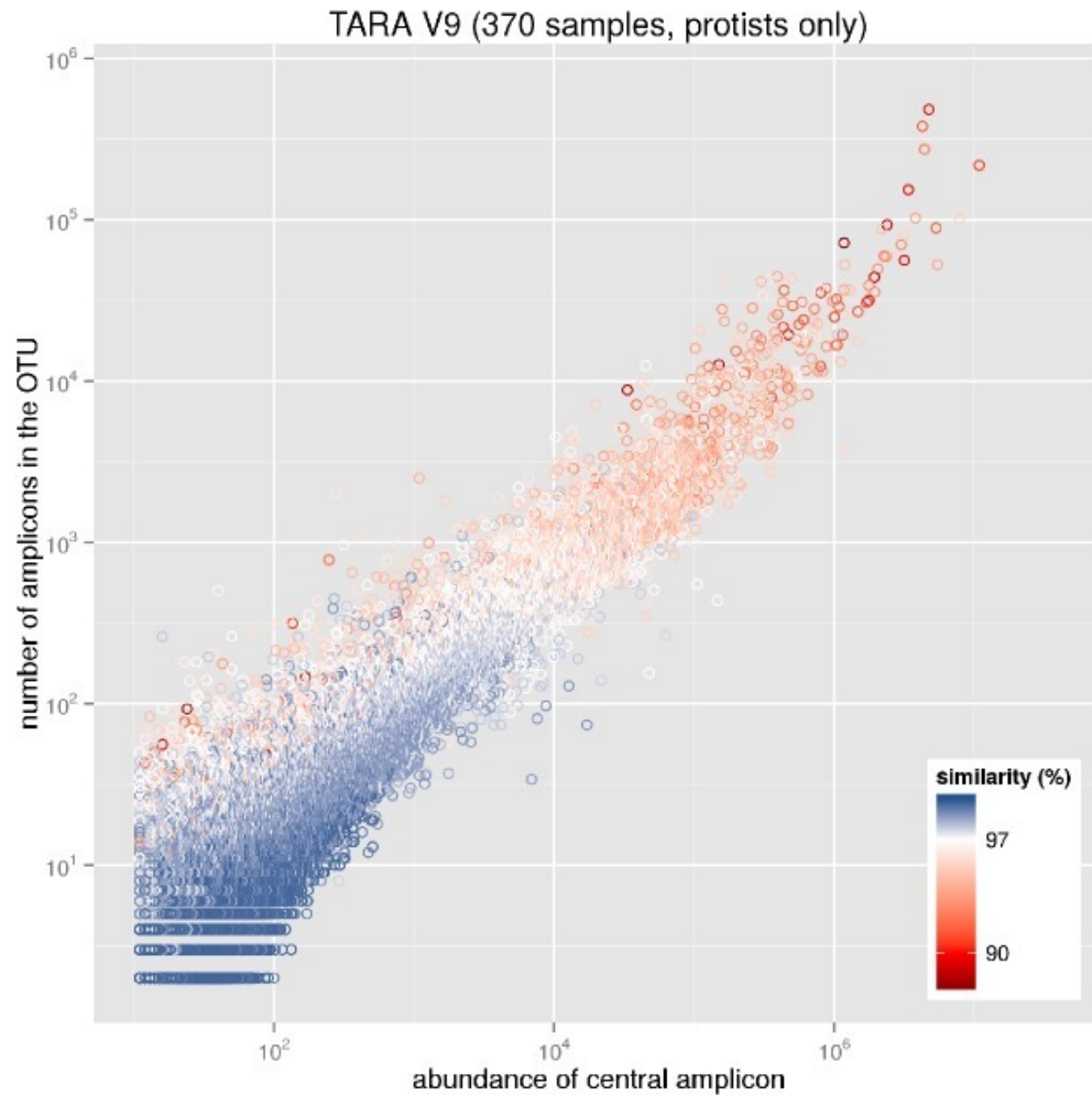
TARA OCEANS
V9



Neotropical Forests Soils
V4

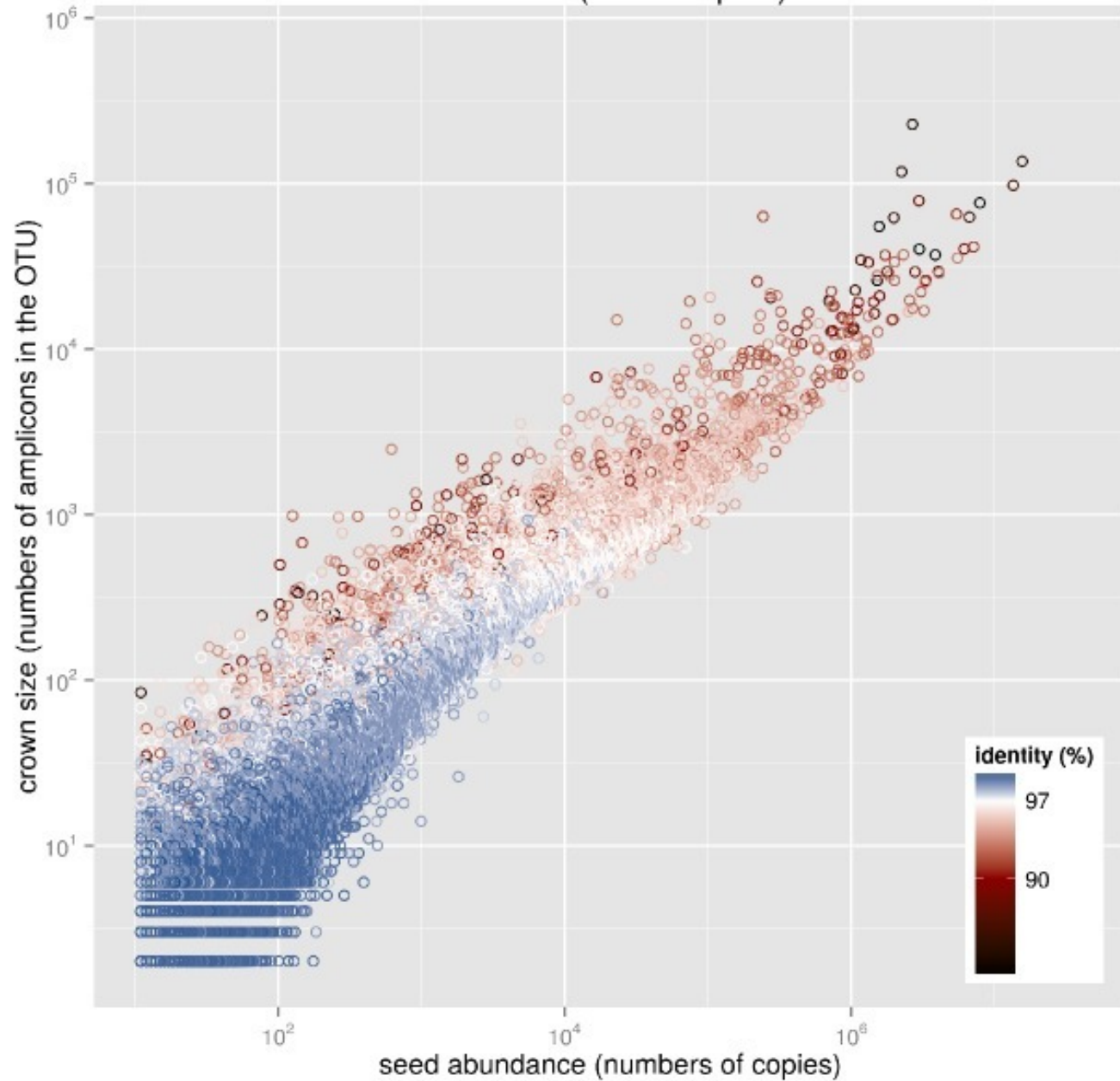
Strong correlation between abundance and the number of micro-variants. Abundant taxa are not known to have more diversified V9 sequences than less abundant taxa. What we see are technical errors.

What if we'd used a fix 97%-clustering threshold?

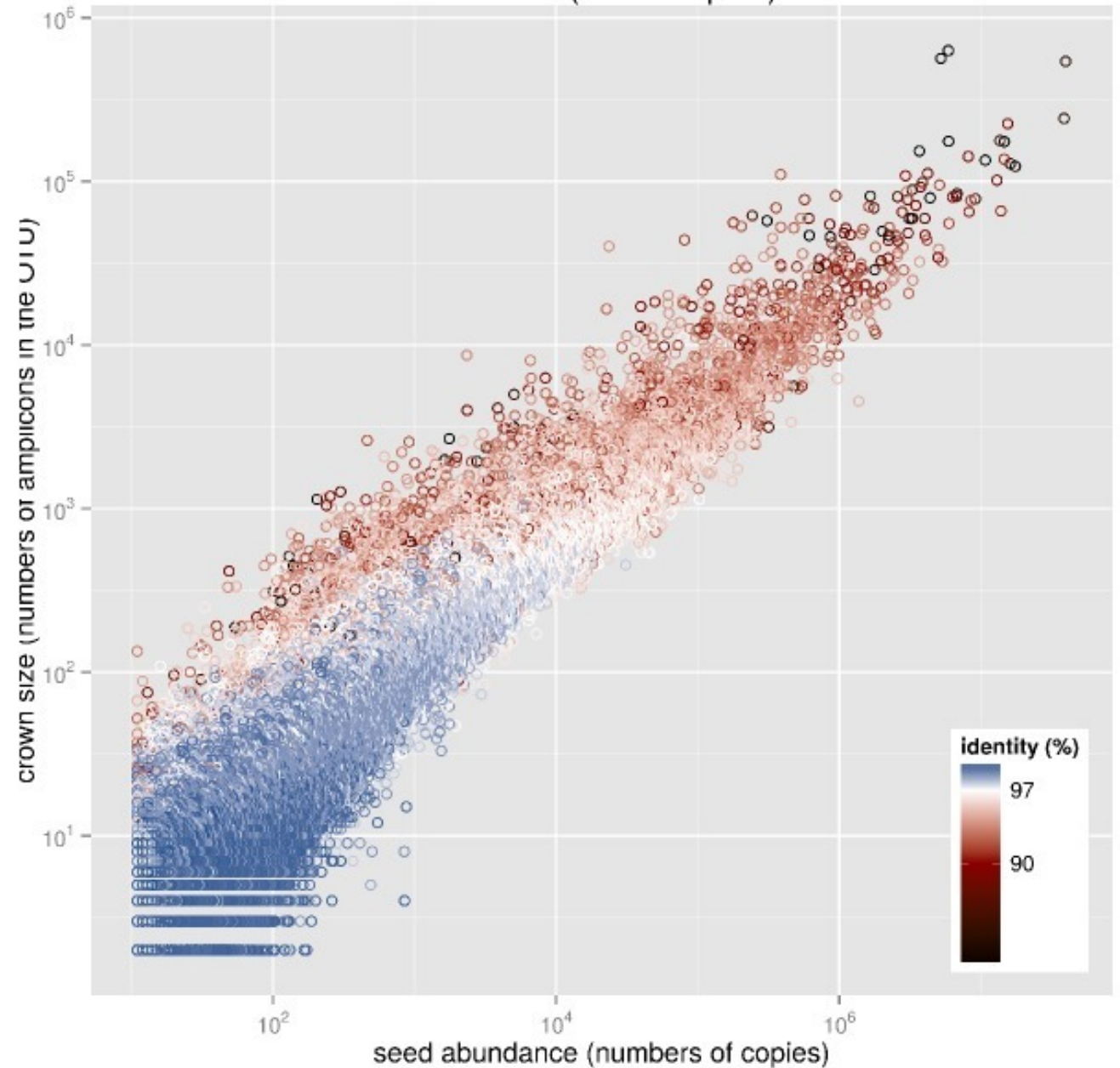


Seed abundance vs cloud vs cluster radius shows 97%-threshold inadequacy

TARA V9 (264 samples)

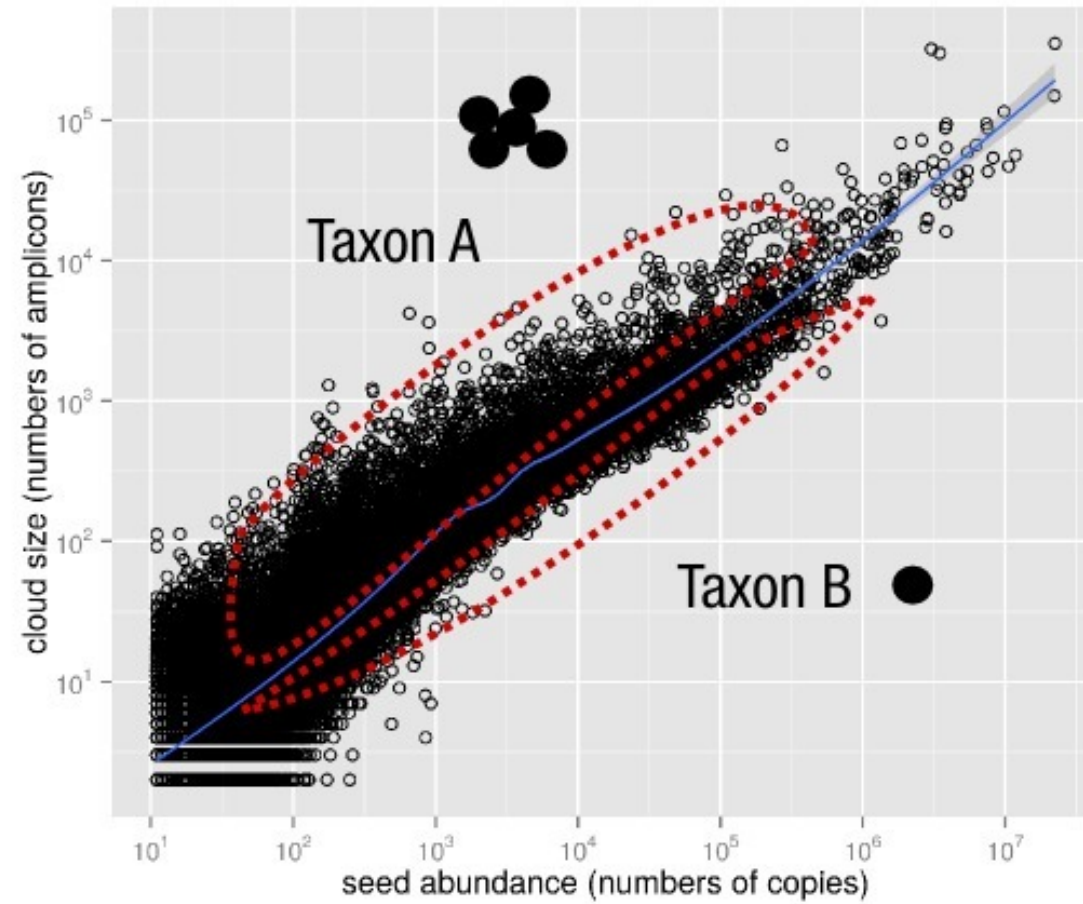


TARA V9 (908 samples)



clusters produced with swarm using $d = 1$

Natural variability?



TARA OCEANS
V9

New algorithms can deal with big data.
Now, how can we reduce noise further?

uchime

UCHIME improves sensitivity and speed of chimera detection

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R.

Bioinformatics. 2011 Aug 15;27(16):2194-200. doi: 10.1093/bioinformatics/btr381

R.C.Edgar et al.

```
A      81 CCTTGGTAGGCCGtTGCCCTGCCAACTAGCTAATCAGACGCgggtCCATCtcaCACCaaccggAgtTTTtcTCaCTgTacc 160
Q      81 CCTTGGTAGGCCGCTGCCCTGCCAACTAGCTAATCAGACGCATCCCCATCCATCACCGATAAAATCTTTAATCTCTTTCAG 160
B      81 TCTTGGTgGGCCGtTaCCcCGCCAACaAGCTAATCAGACGCATCCCCATCCATCACCGATAAAATCTTTAAaCTCTTTCAG 160
Diffs  A      A      p A      A      A      BBBB      BBB      BBBB BB      BBa B      B BBB
Votes  +      +      0 +      +      +      +++++      +++      +++++ ++      ++! +      + +++
Model  AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAxxxxxxxxxxxxxxxxBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

Fig. 3. Chimeric alignment showing diffs and votes. This figure shows a region from an alignment generated by UCHIME. Diffs and votes are annotated. The 'Model' row indicates the three segments of the alignment which are closer to A, the crossover (X) and closer to B, respectively. Diffs are 'A' = diff with Q closer to A in the A segment, 'a' = diff with Q closer to A in the B segment, and similarly for 'B' and 'b'. A 'p' diff indicates that the parents agree but are different from Q. Votes are '+' (yes), '!' (no) and '0' (abstain), indicating whether the corresponding diff supports or contradicts the model.

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,
- foundation for QIIME,
- closed-source & costly

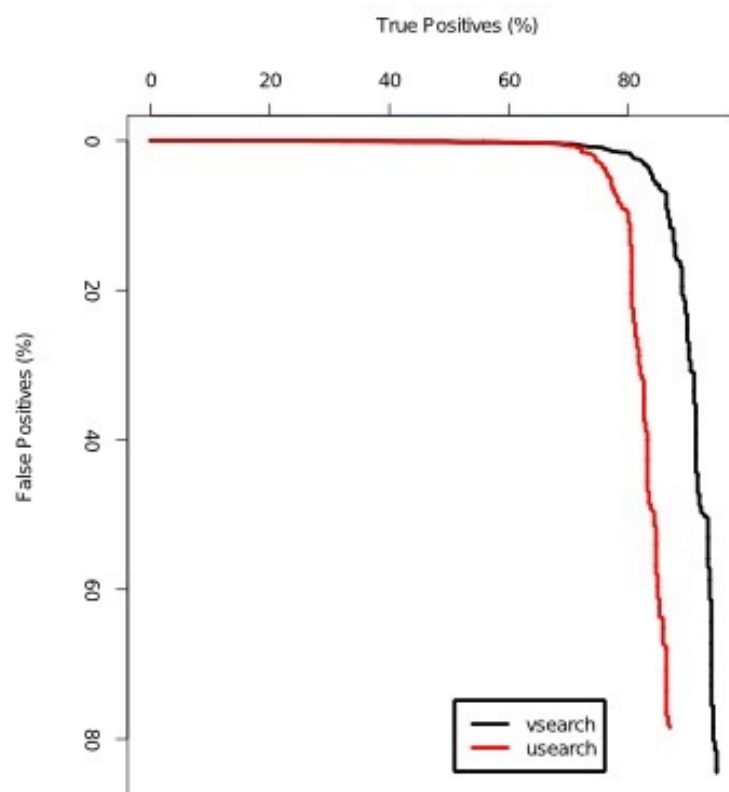


growing success:

- many happy users,
- faster and improved,
- foundation for QIIME 2.0

vsearch:

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



Torbjørn Rognes
Oslo University



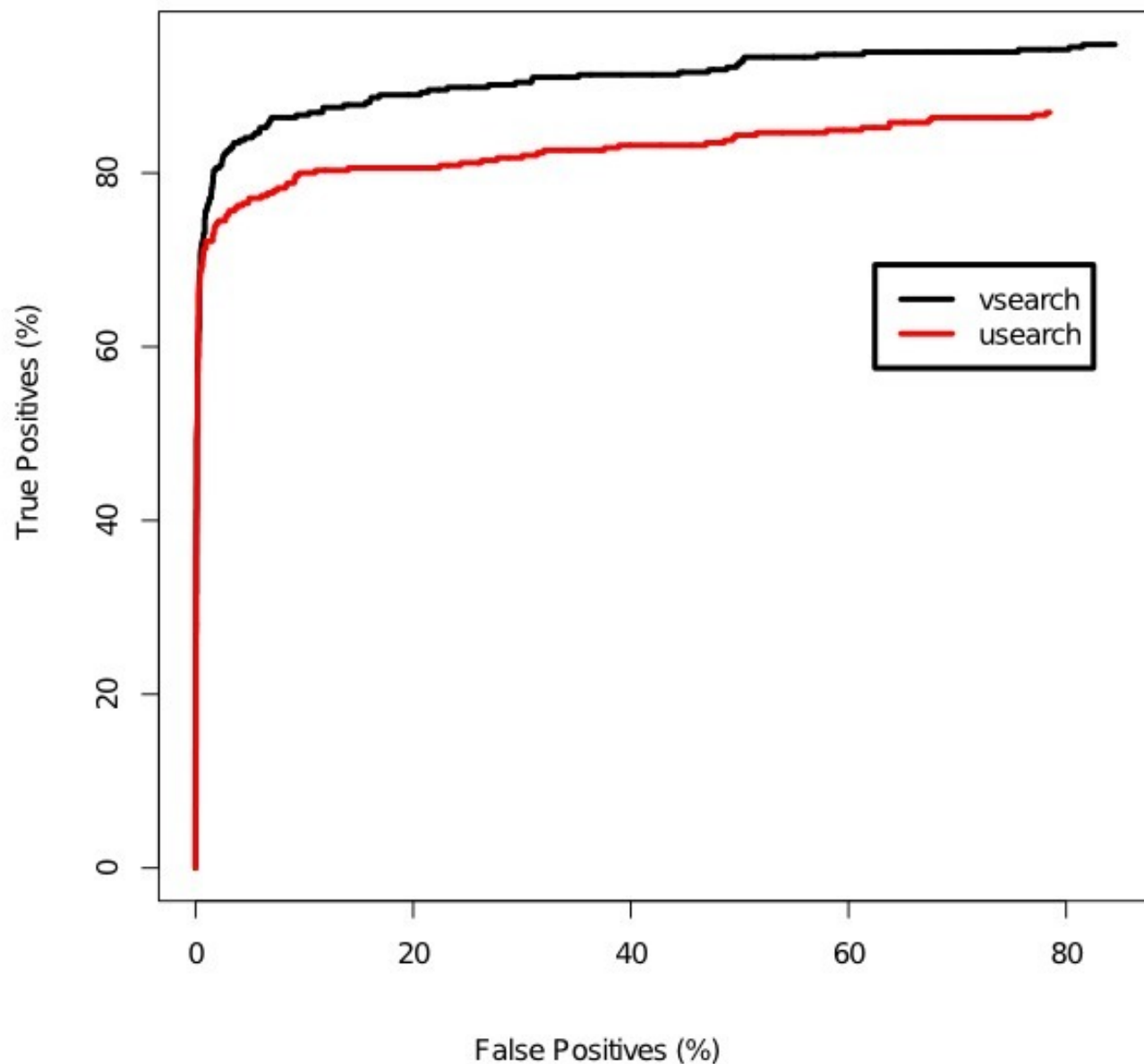
vsearch: open-source alternative for usearch

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

- usearch** (Rob)
- very important
 - 1,000 citations
 - foundation for Q
 - closed-source



- growing success**
- many happy users
 - faster and improved
 - foundation for Q



Torbjørn Rognes
Oslo University



quality filtering

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

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

expected error rate = $\sum Qv$

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

Unique sequences	ΣQ	length
------------------	------------	--------

ea51412efcd0b3cc8510436791667c67f186d983	0.0037	48
--	--------	----

57e9fa79a7d578ee3fa5dbd4960f8d283b4020b1	0.0042	54
--	--------	----

bc42f3dc4a98988d39c23fa5bbf2f3f20c372b66	0.0044	57
--	--------	----

8e6a5352d2ba02eb5291ce5a2c4fccc95e5279c6	0.0048	61
--	--------	----

282382d34ca8346e1b7968b2084f75249840f819	0.0048	62
--	--------	----

092bf0fdd2cf634aba6e3cc5b6ae323db37bbf62	0.0057	73
--	--------	----

41873e83fb3666536a7bf0f702cd24805cd11e17	0.0057	73
--	--------	----

448b14d7a8657f19f7340da3a31d25652f40b67e	0.0057	73
--	--------	----

61c4bb94bf70f1f86e62cad5d7951941a67b8479	0.0057	73
--	--------	----

6fc98ad4798e85e9b8b6643786d786701e1eb0d7	0.0057	73
--	--------	----

ad967f92c57493caf41f8b86f5bf3397e6bd4a80	0.0057	73
--	--------	----

b6b4e3fd88e23a68773ea00a530dcf320826322b	0.0687	73
--	--------	----

2a46cc936ef8a263360257420a72921a769440c0	0.0063	80
--	--------	----

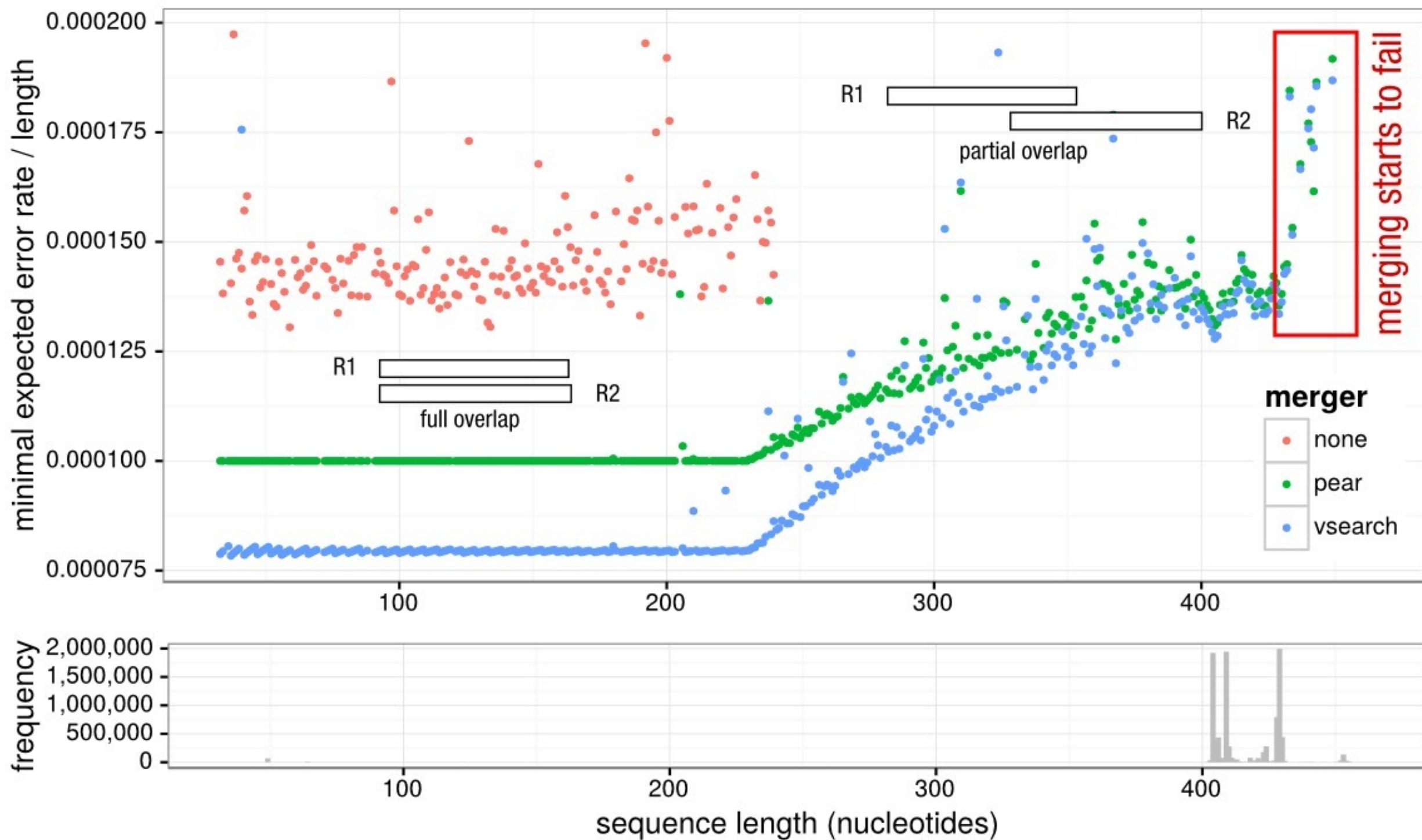
03dcb427634a4fe6c204a276a9237a3a20c3fe5e	0.0072	92
--	--------	----

10dbb1a0a9555c64ad2c0b4f91aec6055175a02b	0.0072	92
--	--------	----

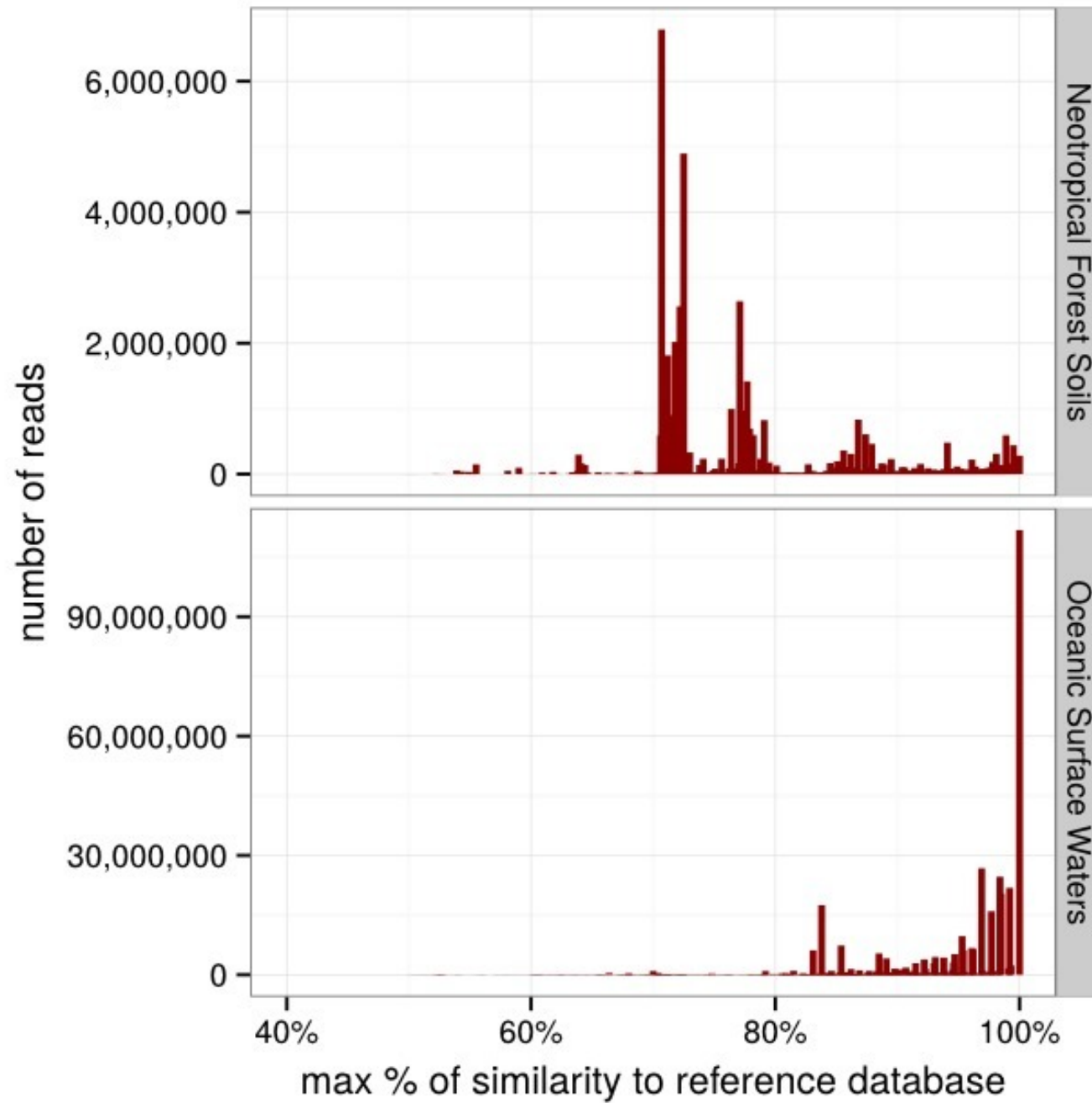
121c9a9ac6f8a6fdf1dcf3ca5ecb7c484d8e7fb0	0.0072	92
--	--------	----

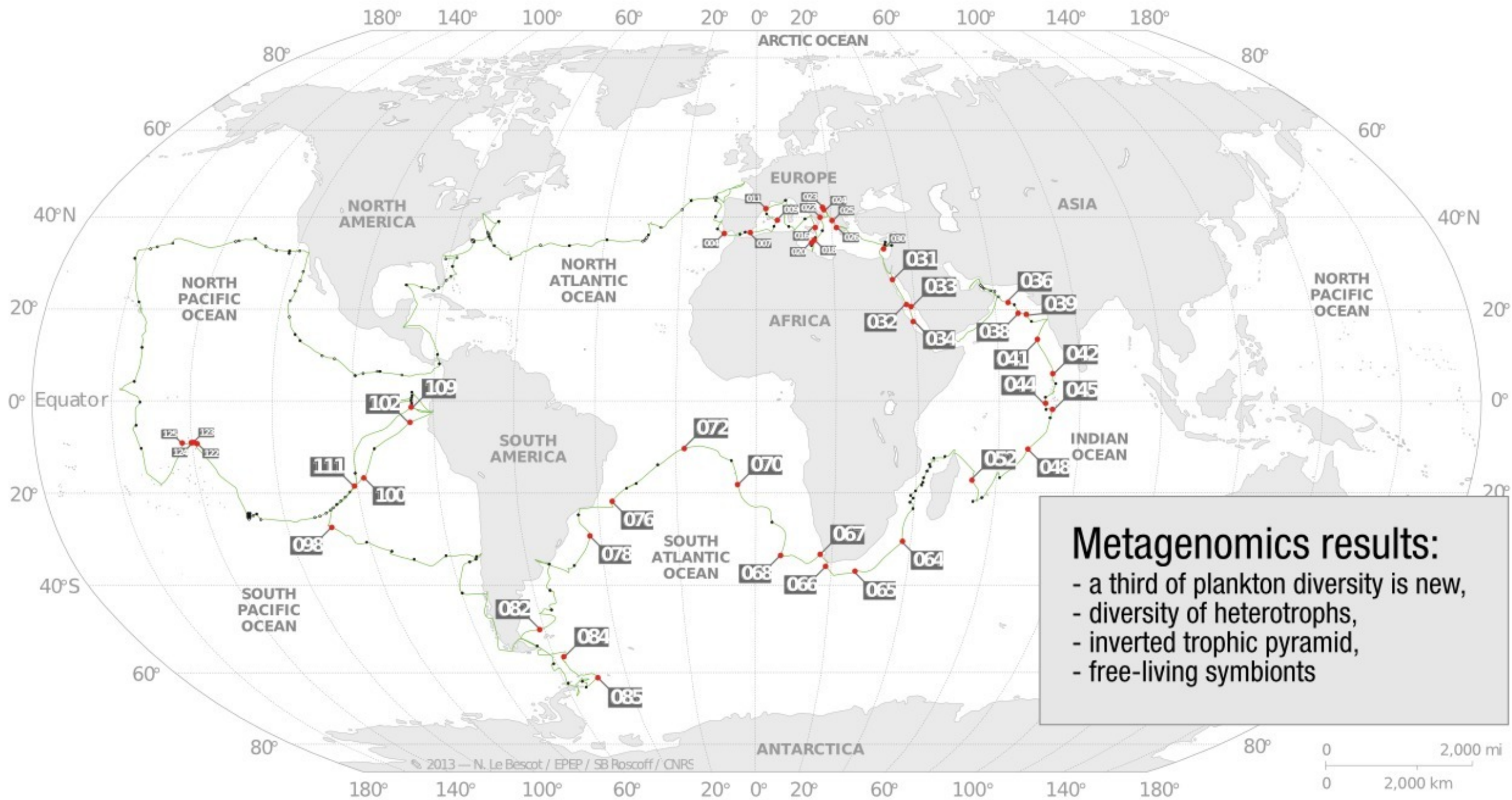
1222c57cba5c99b5b6013716ec42954f872c27e4	0.0072	92
--	--------	----

Lowest expected error observed per sequence length in a 16S V3-V4 MiSeq 2x250 bp dataset



Filtering by taxonomic assignment





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

A map of the Neotropical region, including Central and South America, with a grid of latitude and longitude lines. Three countries are highlighted in red: Costa Rica, Panama, and Ecuador. Lines connect these red areas to a central text box. A larger text box is at the bottom left, and a list of early results is in a box at the bottom right. An inset photograph of a person in a forest is in the top right.

Neotropical Forests Soil Sampling Project

Costa Rica
Panama
Ecuador

Early results

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

Microbial diversity at the tree line level

David Wardle & Jordan Mayor, Swedish University of Agricultural Sciences



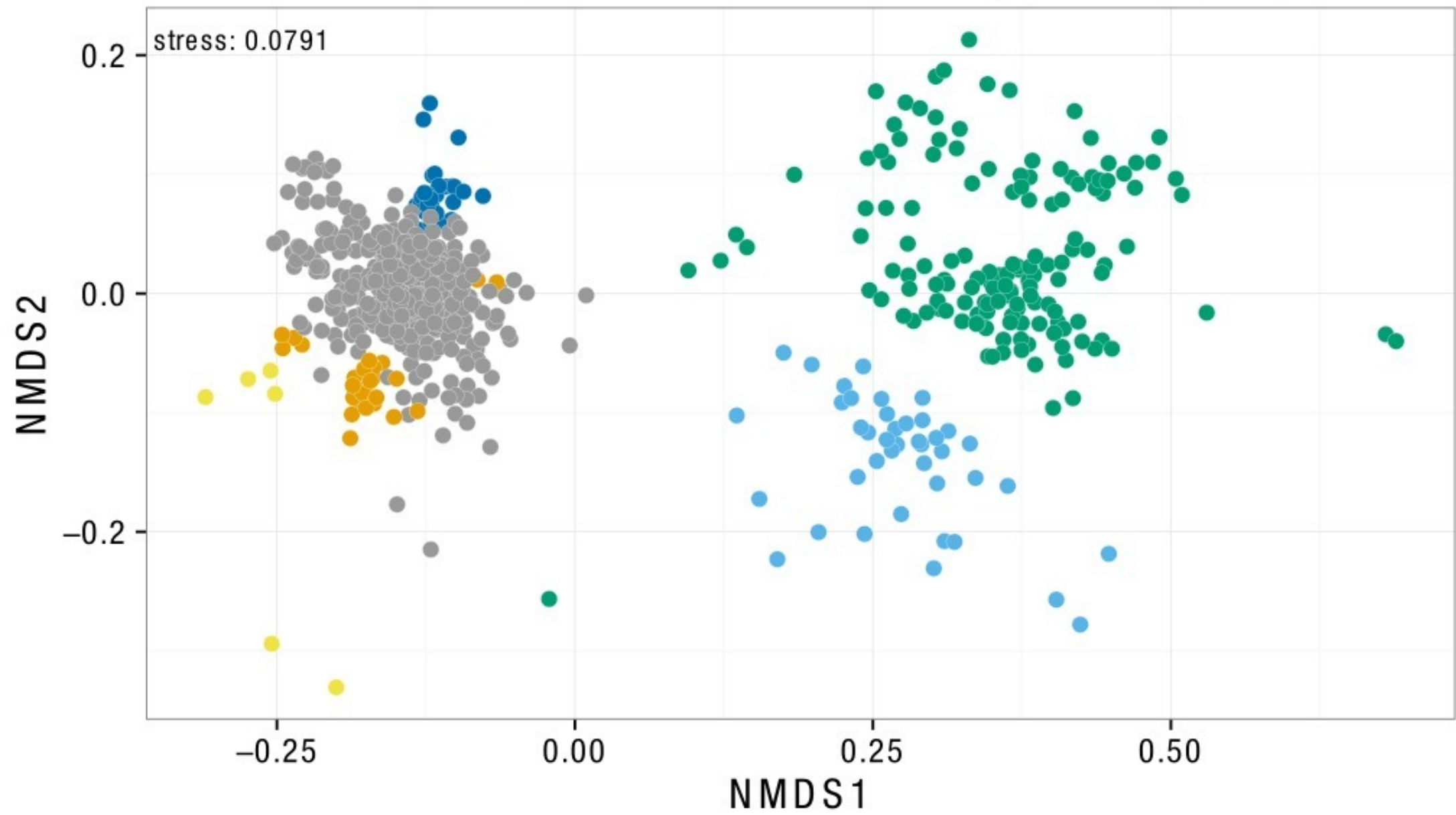
Sampling

- 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?

Meta-analysis of 18S rRNA V4 (672 samples)



- | | |
|---|---|
| ● bathypelagic | ● seafloor |
| ● fresh water | ● temperate mountain forest soil |
| ● open ocean | ● tropical lowland forest soil |

What do we need to deal with noise?

- 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)

Technical noise and robust stats
are our main challenges

Thanks!

<https://github.com/torognes/swarm>

<https://github.com/torognes/vsearch>

