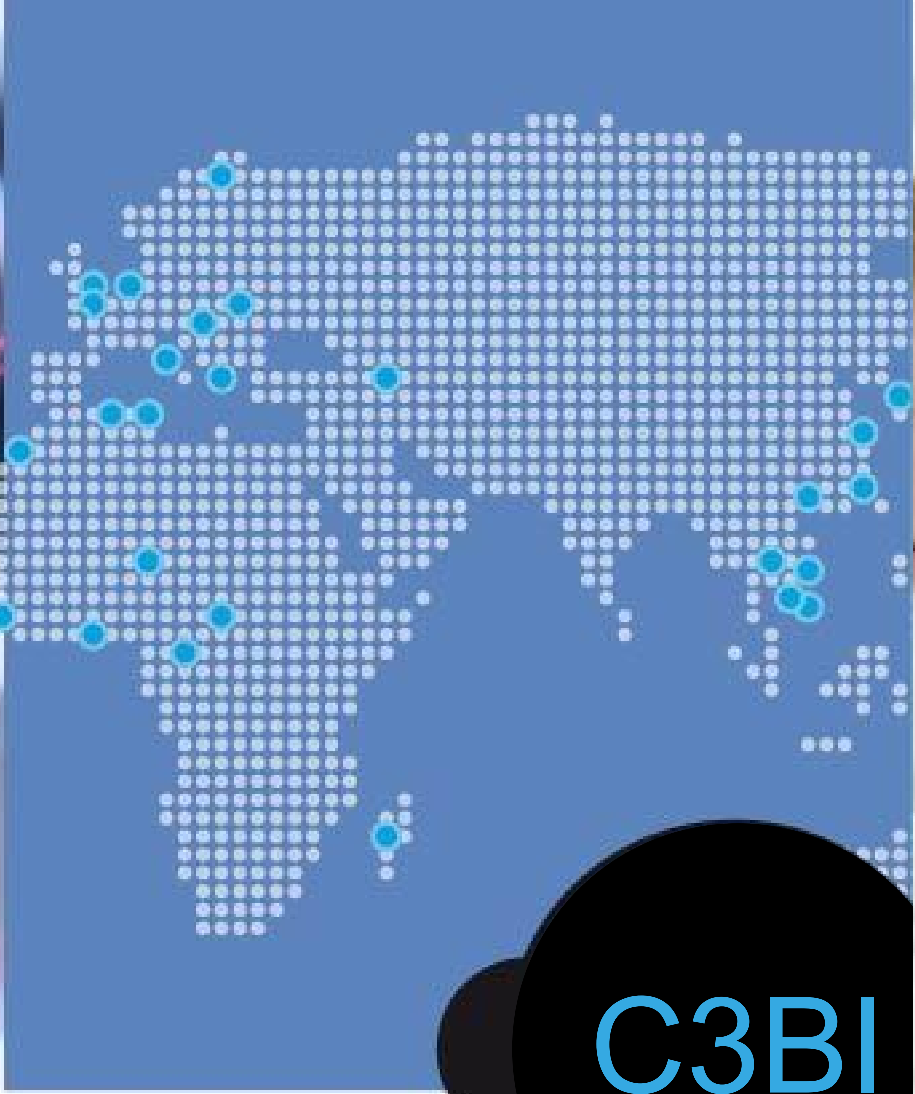
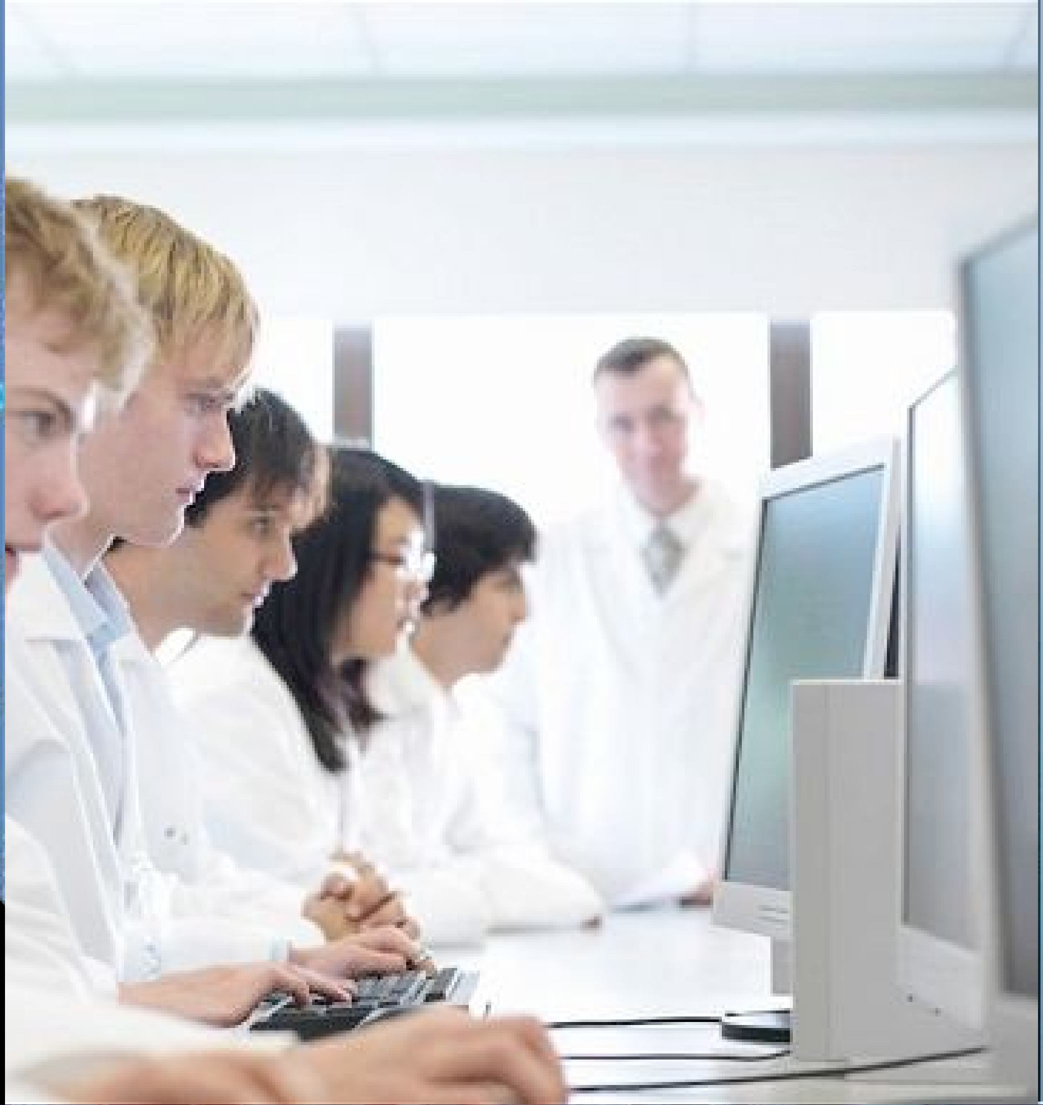




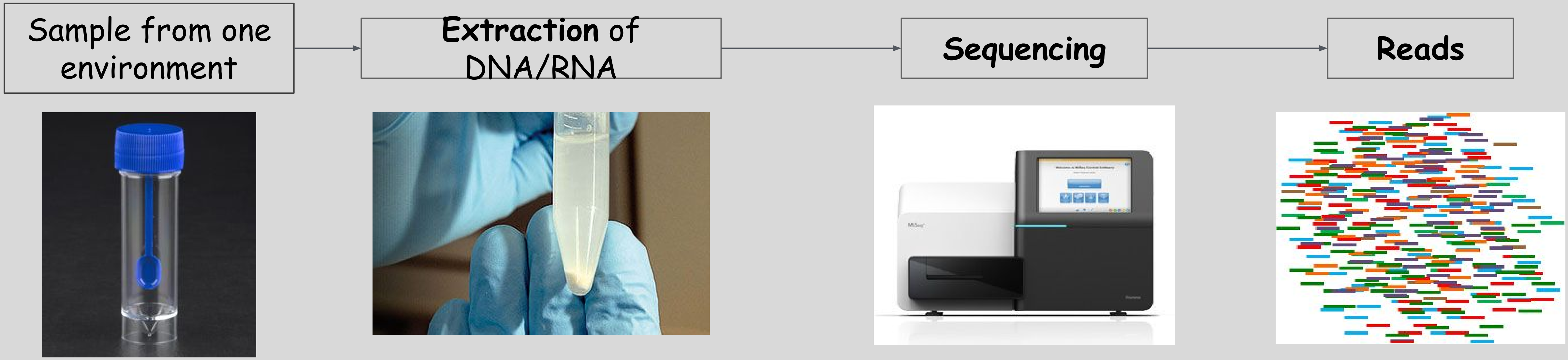
C3BI

# Targeted metagenomics

Amine Ghozlane

June 2016

# High-throughput sequencing as a tool for exploring the microbiome



Sample from one environment → **Extraction** of DNA/RNA → **Sequencing** → **Reads**

"Metagenomics is like a disaster in a jigsaw shop" Iddo Friedberg

## Who is there ?

- **Taxonomical annotation**
- Co-Abundance Gene groups (CAG)
- Binning
- Assembly

## What are they able to do ?

- Gene/protein prediction
- Functional annotation
- Metabolic network reconstruction

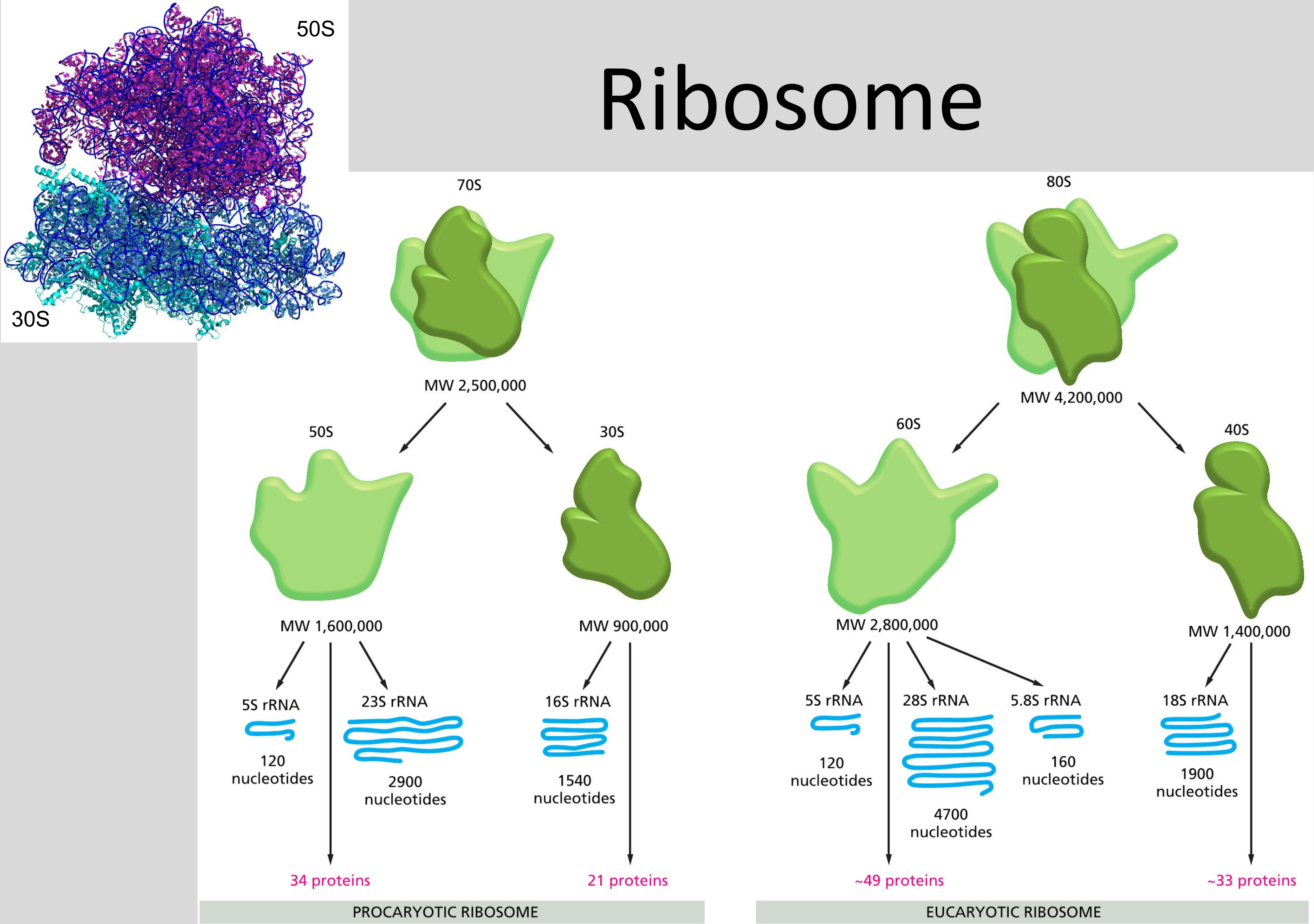
## What are they doing ?

- RNA/Protein quantification

## What is the difference between these environments ?

- Comparative metagenomics
- Quantitative metagenomics

# Ribosome



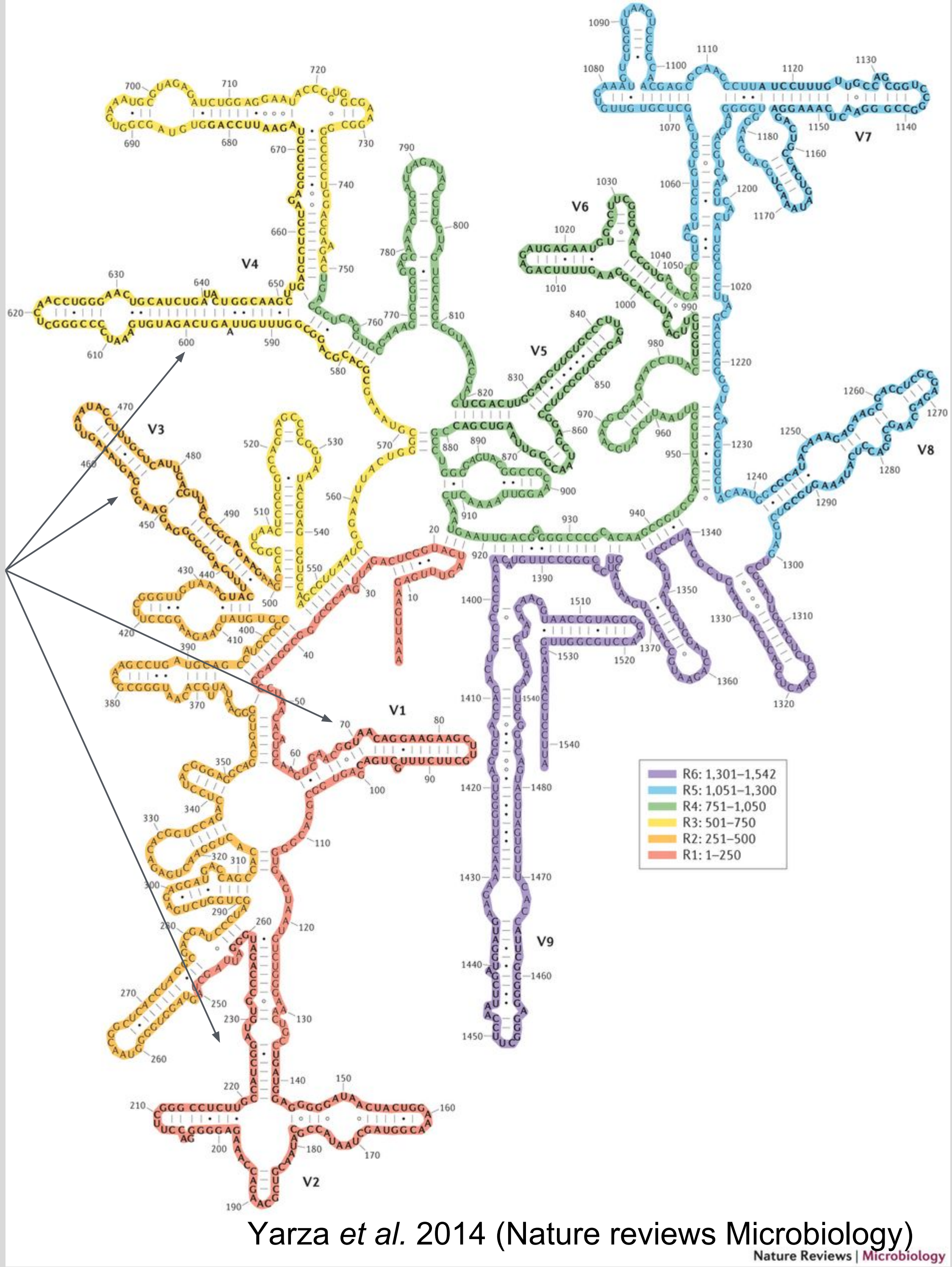
ITS : located between 18S and 5.8S rRNA genes

Image : Alberts Molecular Biology of the Cell 5th

# 16S rRNA

- Weakly affected by horizontal gene transfer*
- 9 variable regions surrounded by conserved regions
- Universal primers**, 25 PCR cycle***
- Most well represented gene in Genbank
- Sequencing kits : V1-V3, V3-V4, V3-V5, V5-V6...

Variable regions



Yarza *et al.* 2014 (Nature reviews Microbiology)

Nature Reviews | Microbiology

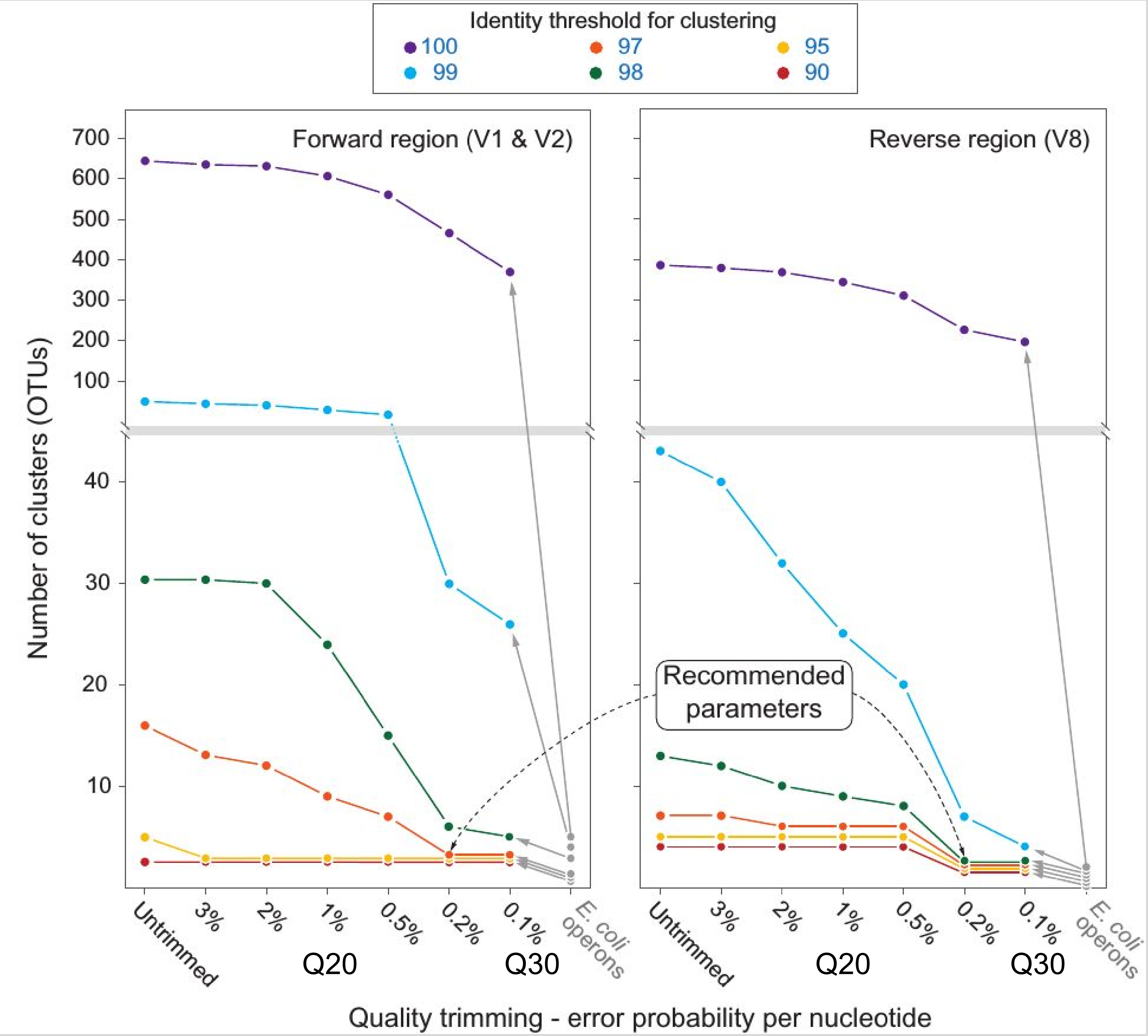
*Daubin et al. 2003 (Science) **Weisburg *et al.* 1991 (J Bacteriol.) *** Illumina protocol

# Operational Taxonomic Unit (OTU)

*Définition : "Group of DNA sequences that share a defined level of similarity"**

Kunin et al. 2010

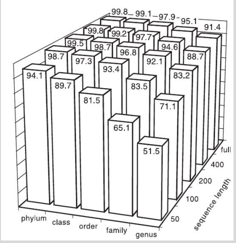
- 454 sequencing of V1-V2 & V8 E. coli MG1655
- Theoretical number
  - 5 phylotypes for V1-V2 at 100% id
  - 1 phylotype for V8
- Results
  - **0.1-0.2% error probability**
  - **97% similarity threshold**



*Vetrovsky and Baldrian 2013 (Plos One)

# Targeted metagenomics strategies

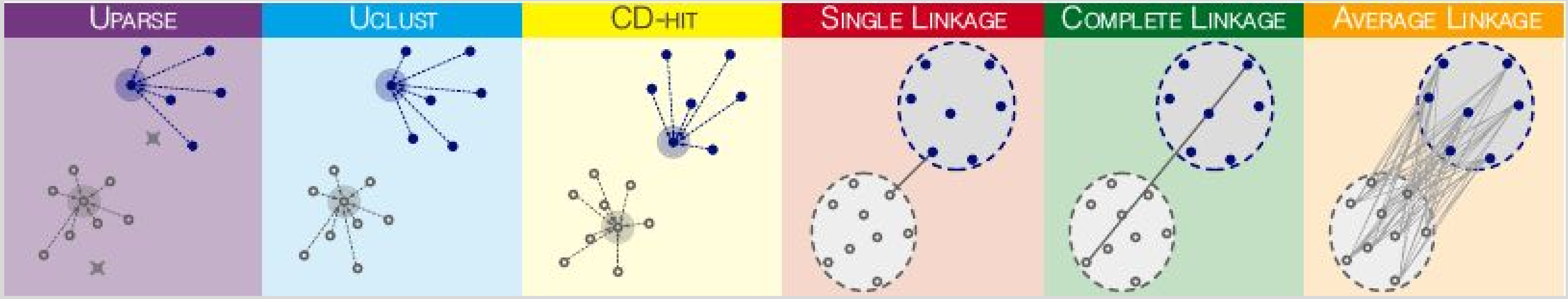
- ❖ **CLOSED REFERENCE CLUSTERING**
  - ➢ **Clustering in a OTU the sequence that are similar to a reference**
  - ➢ **Classification**



- ❖ **DE NOVO CLUSTERING**
  - ➢ **Distance between the sequence is used to cluster sequence into OTUs**

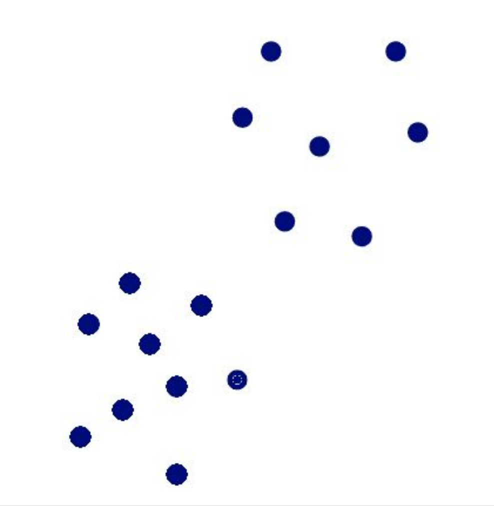
- ❖ **OPEN REFERENCE CLUSTERING**
  - ➢ **Closed-reference clustering followed by de novo clustering for sequence that are not similar to the reference**

**CLUSTERING ALGORITHMS**



*Westcott, Schloss, 2016 PeerJ; Rideout 2014; Schmidt et al. 2015

# Hierarchical clustering

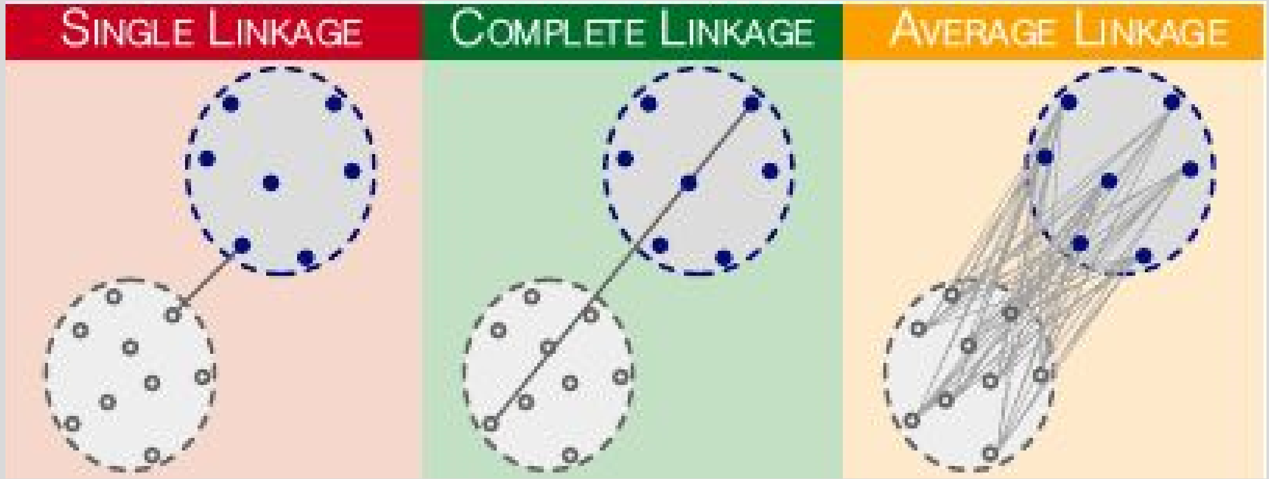


**Algorithm:**

- **Initial n groups**
- **Each step:**
  - **Merge of two group considering linkage**

http://www.saedsayad.com/clustering_hierarchical.htm

# Hierarchical clustering



**Algorithm:**

- **Initial n groups**
- **Each step:**
  - **Merge of two group considering linkage**

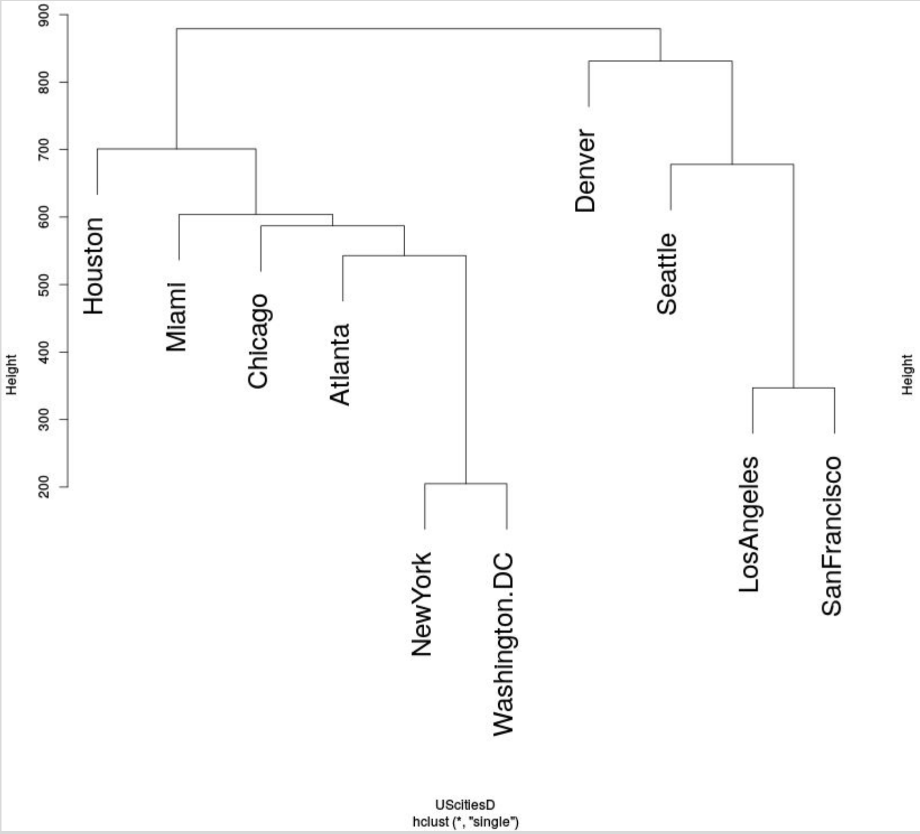
**Distances:**

- **Single linkage**
  - the distance between two clusters is defined as the ***shortest*** distance between two points in each cluster
- **Complete linkage**
  - the distance between two clusters is defined as the ***longest*** distance between two points in each cluster
- **Average linkage**
  - the distance between two clusters is defined as the ***average*** distance to every point in the other cluster
- …

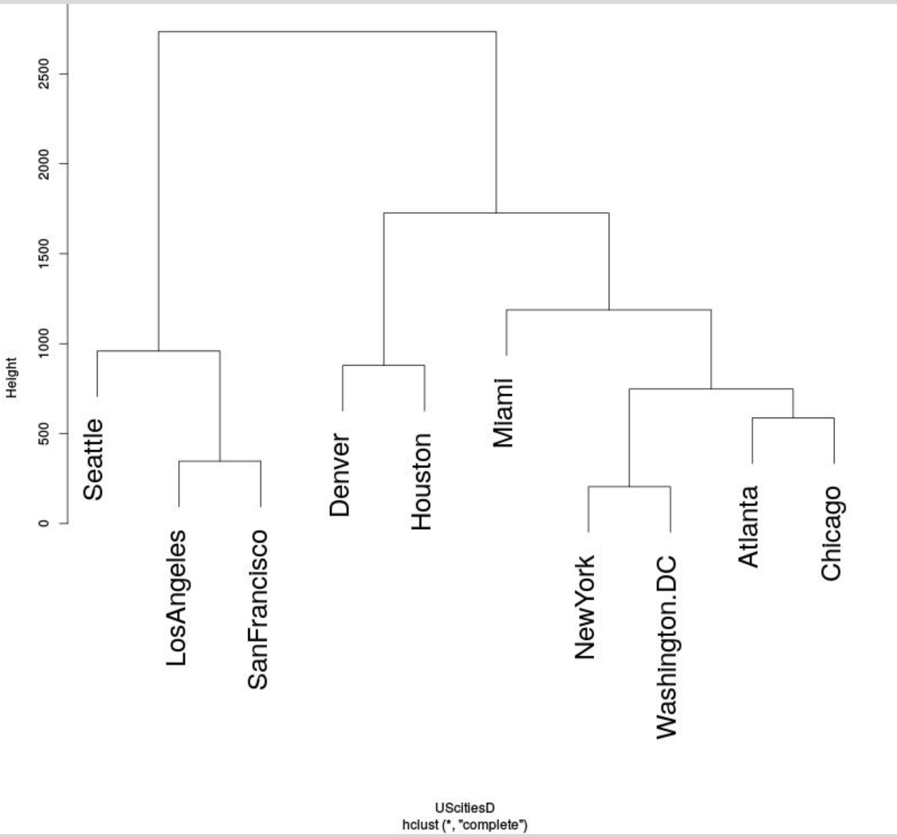
http://www.saedsayad.com/clustering_hierarchical.htm

# Hierarchical clustering

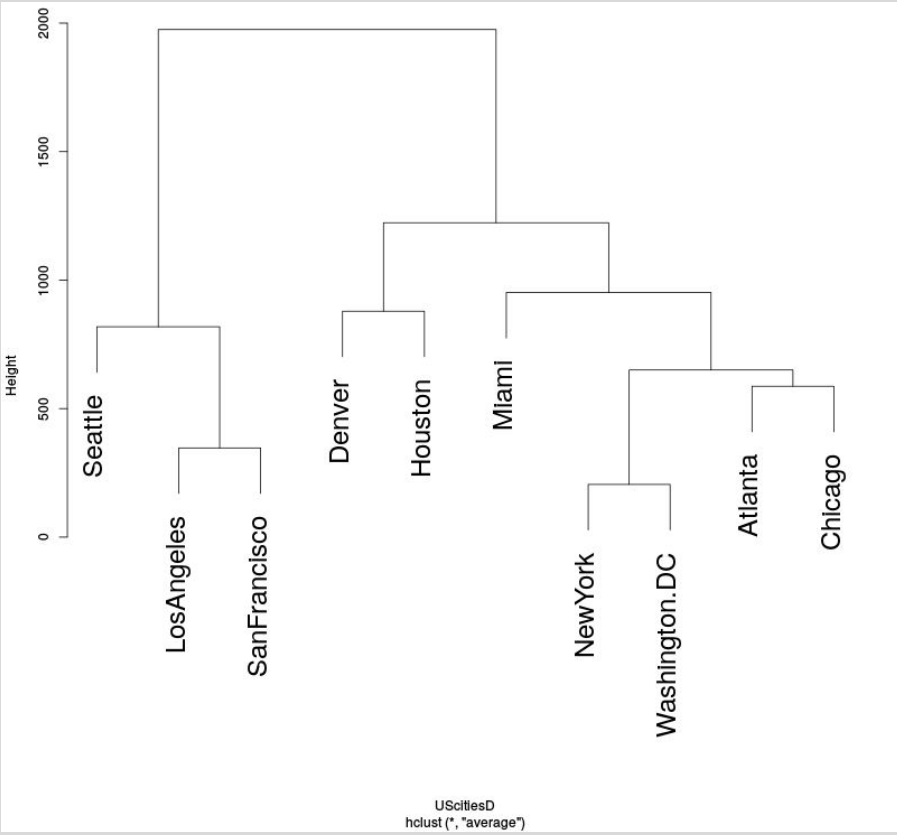
### SINGLE linkage



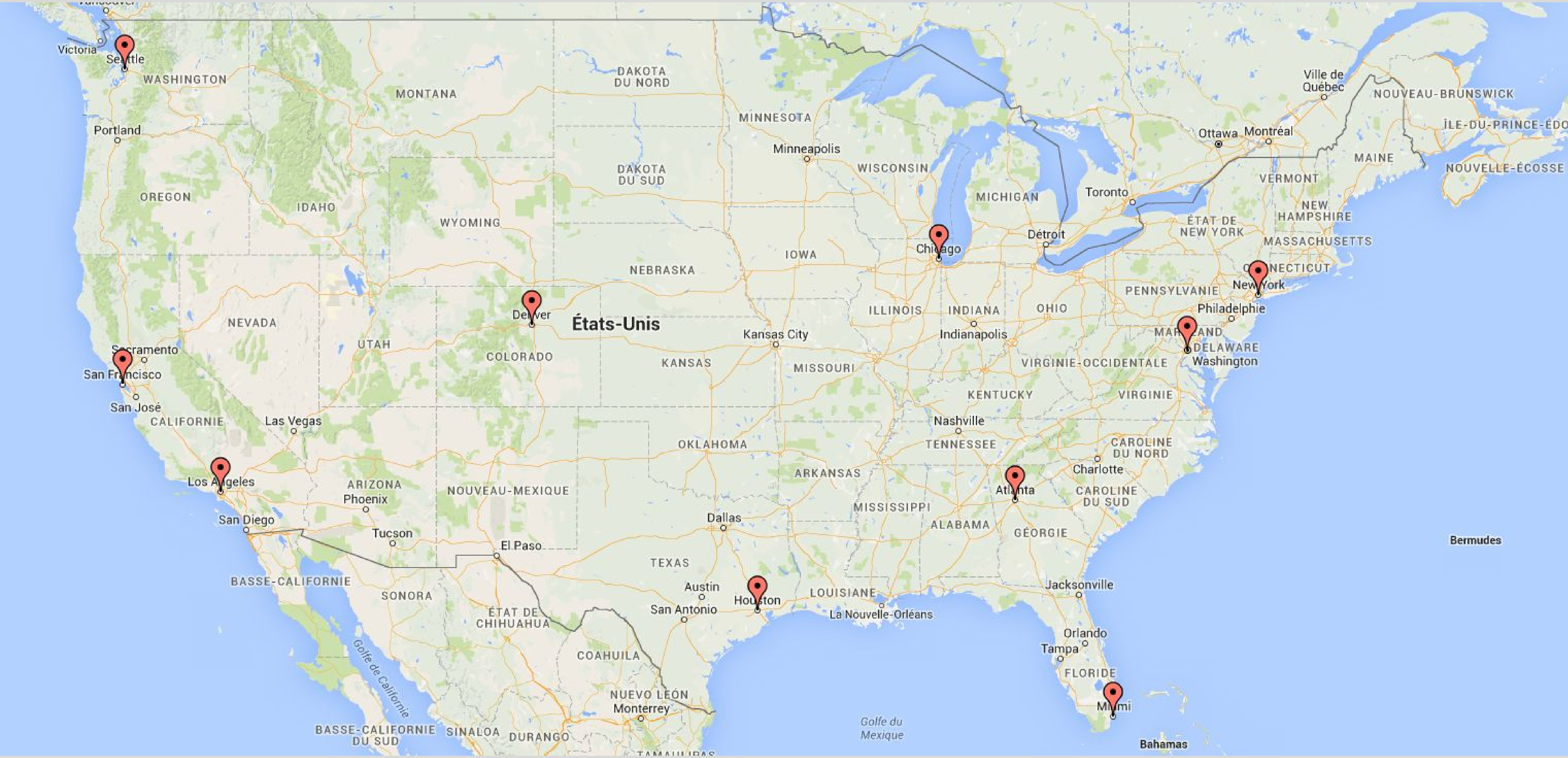
### COMPLETE linkage



### AVERAGE linkage



# Let’s play at clustering

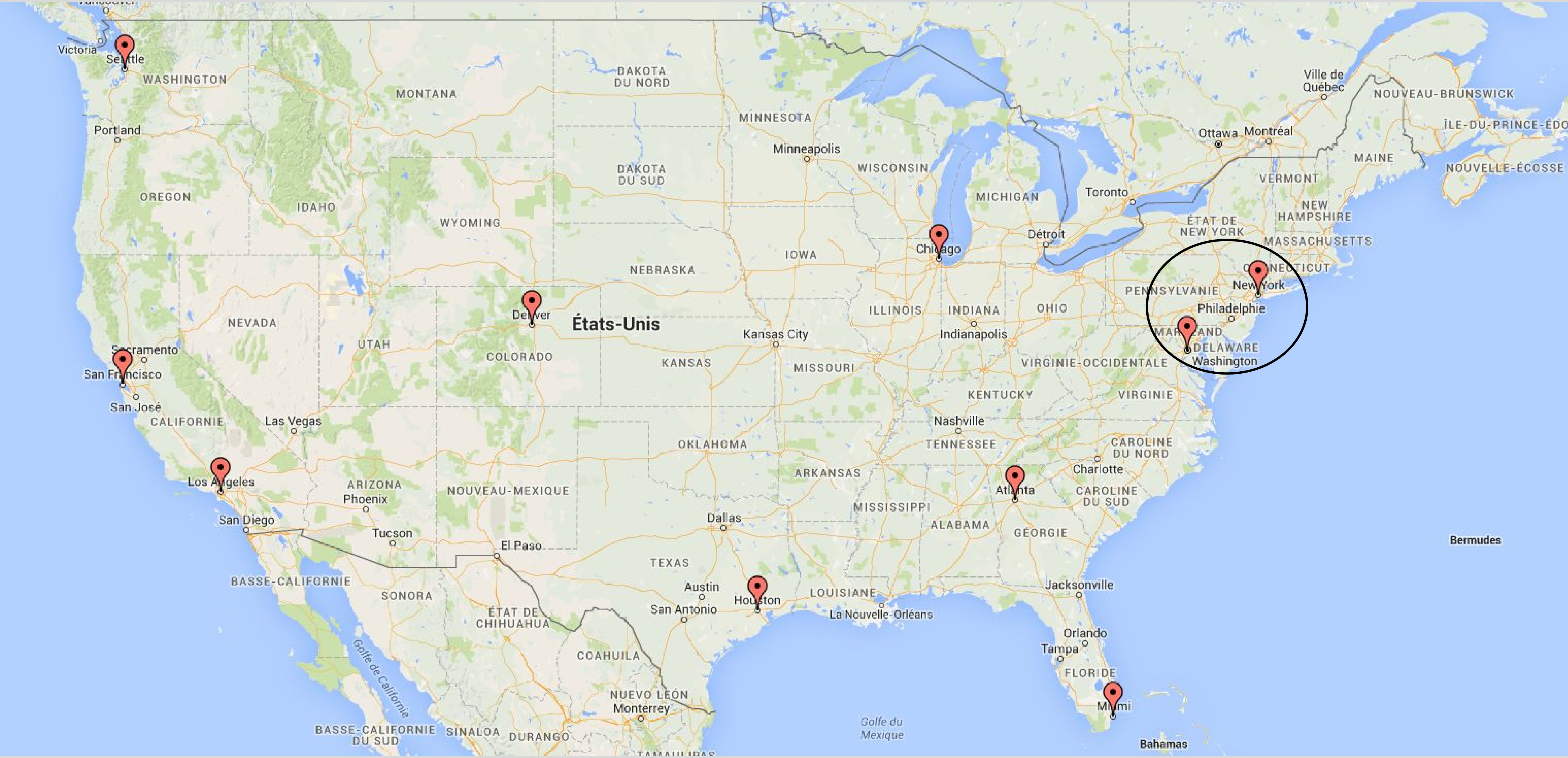


<u>Criteria :</u> “A group is composed of cities with less than 800 km of distance”

# Distance

| Distance between US cities | Atlanta | Chicago | Denver | Houston | Los Angeles | Miami | New York | San Francisco | Seattle |
|---|---|---|---|---|---|---|---|---|---|
| Atlanta | | | | | | | | | |
| Chicago | 587 | | | | | | | | |
| Denver | 1212 | 920 | | | | | | | |
| Houston | 701 | 940 | 879 | | | | | | |
| Los Angeles | 1936 | 1745 | 831 | 1374 | | | | | |
| Miami | 604 | 1188 | 1726 | 968 | 2339 | | | | |
| New York | 748 | 713 | 1631 | 1420 | 2451 | 1092 | | | |
| San Francisco | 2139 | 1858 | 949 | 1645 | 347 | 2594 | 2571 | | |
| Seattle | 2182 | 1737 | 1021 | 1891 | 959 | 2734 | 2408 | 678 | |
| Washington DC | 543 | 597 | 1494 | 1220 | 2300 | 923 | **205** | 2442 | 2329 |

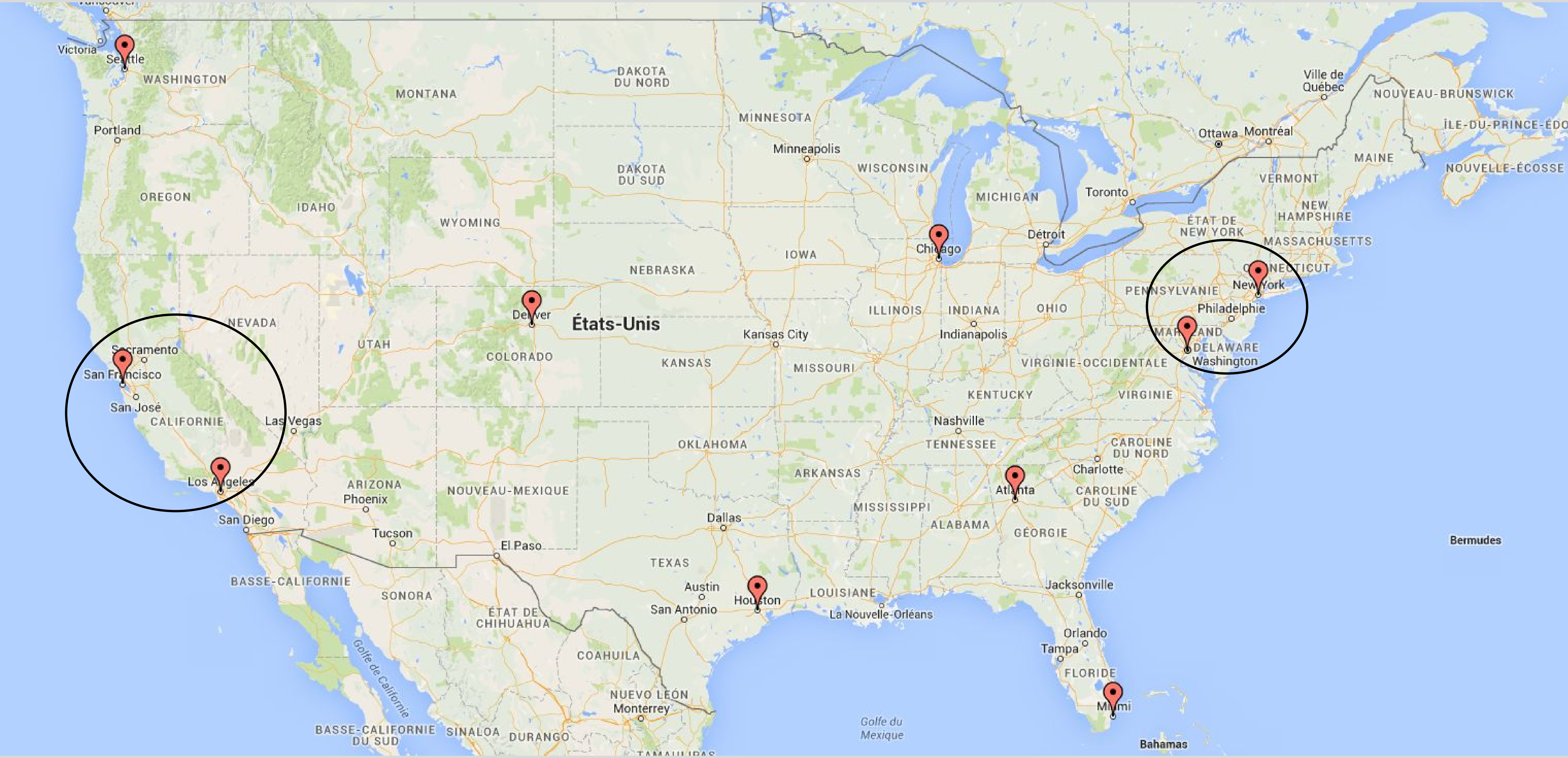
# Hierarchical clustering



# Distance

| Distance between US cities | Atlanta | Chicago | Denver | Houston | Los Angeles | Miami | New York | San Francisco | Seattle |
|---|---|---|---|---|---|---|---|---|---|
| Atlanta | | | | | | | | | |
| Chicago | 587 | | | | | | | | |
| Denver | 1212 | 920 | | | | | | | |
| Houston | 701 | 940 | 879 | | | | | | |
| Los Angeles | 1936 | 1745 | 831 | 1374 | | | | | |
| Miami | 604 | 1188 | 1726 | 968 | 2339 | | | | |
| New York | 748 | 713 | 1631 | 1420 | 2451 | 1092 | | | |
| San Francisco | 2139 | 1858 | 949 | 1645 | **347** | 2594 | 2571 | | |
| Seattle | 2182 | 1737 | 1021 | 1891 | 959 | 2734 | 2408 | 678 | |
| Washington DC | 543 | 597 | 1494 | 1220 | 2300 | 923 | **205** | 2442 | 2329 |

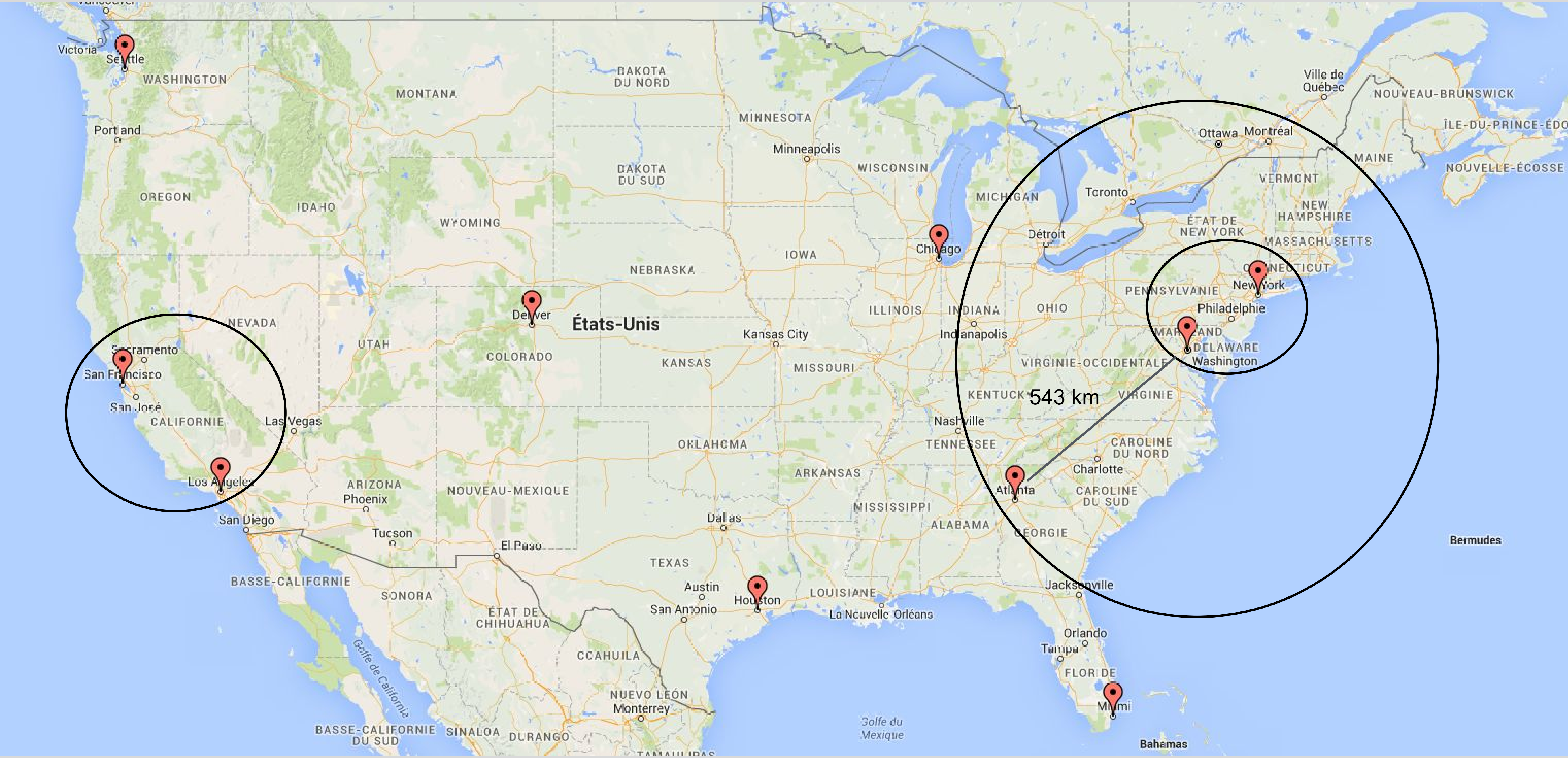
# Hierarchical clustering - single linkage



# Next in single-linkage ?

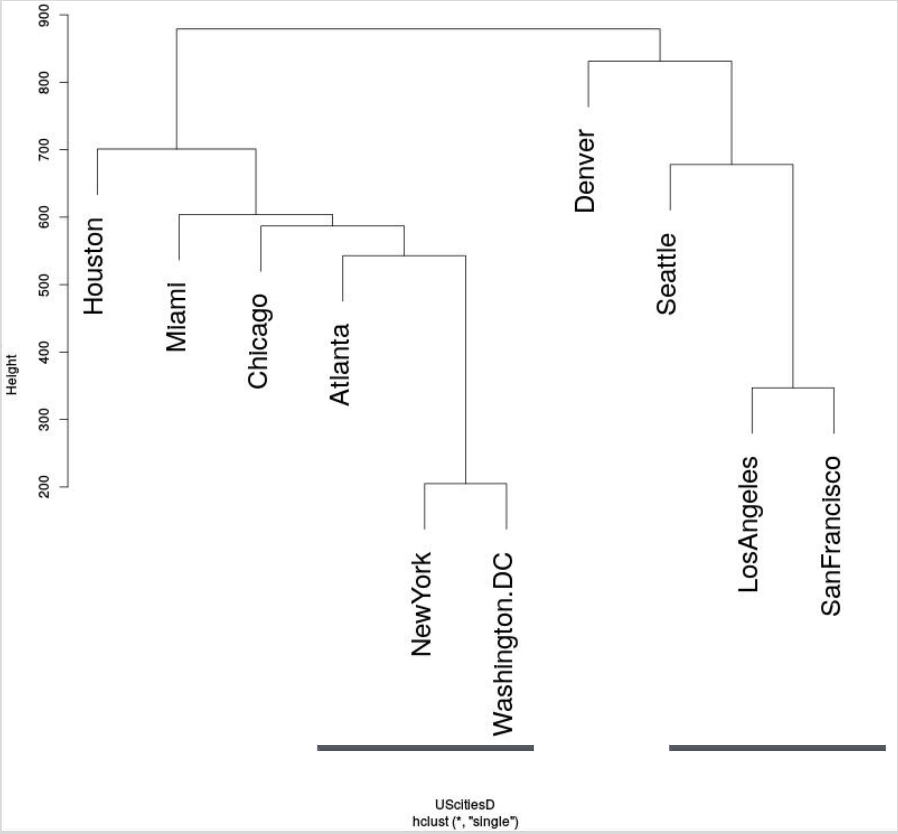
| Distance between US cities | Atlanta | Chicago | Denver | Houston | Los Angeles | Miami | New York | San Francisco | Seattle |
|---|---|---|---|---|---|---|---|---|---|
| Atlanta | | | | | | | | | |
| Chicago | 587 | | | | | | | | |
| Denver | 1212 | 920 | | | | | | | |
| Houston | 701 | 940 | 879 | | | | | | |
| Los Angeles | 1936 | 1745 | 831 | 1374 | | | | | |
| Miami | 604 | 1188 | 1726 | 968 | 2339 | | | | |
| New York | 748 | 713 | 1631 | 1420 | 2451 | 1092 | | | |
| San Francisco | 2139 | 1858 | 949 | 1645 | **347** | 2594 | 2571 | | |
| Seattle | 2182 | 1737 | 1021 | 1891 | 959 | 2734 | 2408 | 678 | |
| Washington DC | **543** | 597 | 1494 | 1220 | 2300 | 923 | **205** | 2442 | 2329 |

# Hierarchical clustering - single linkage



# Hierarchical clustering

**SINGLE linkage**

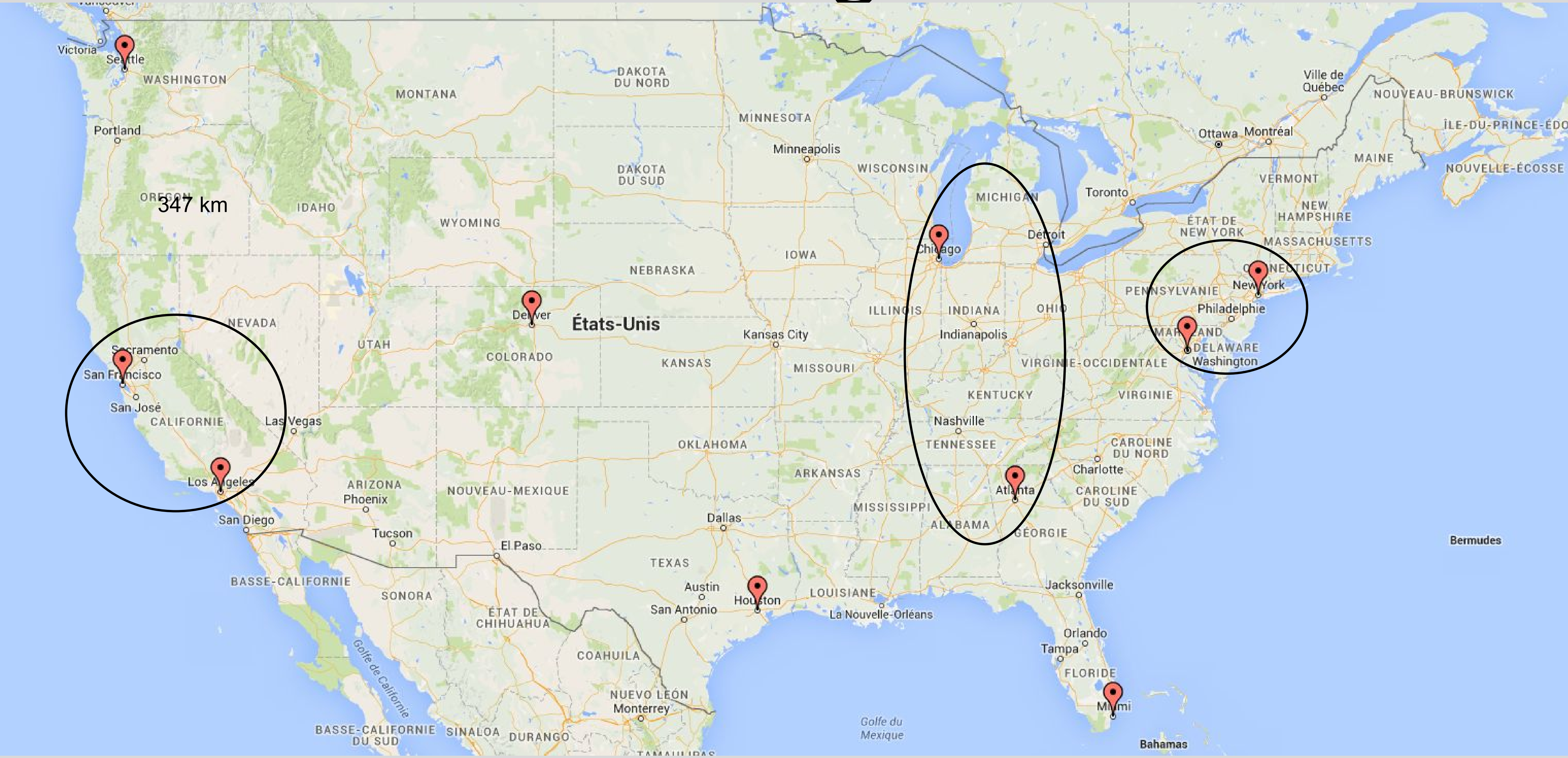


# Now in average-linkage

| Distance between US cities | Atlanta | Chicago | Denver | Houston | Los Angeles | Miami | New York | San Francisco | Seattle |
|---|---|---|---|---|---|---|---|---|---|
| Atlanta | | | | | | | | | |
| Chicago | **587** | | | | | | | | |
| Denver | 1212 | 920 | | | | | | | |
| Houston | 701 | 940 | 879 | | | | | | |
| Los Angeles | 1936 | 1745 | 831 | 1374 | | | | | |
| Miami | 604 | 1188 | 1726 | 968 | 2339 | | | | |
| New York | 748 | 713 | 1631 | 1420 | 2451 | 1092 | | | |
| San Francisco | 2139 | 1858 | 949 | 1645 | **347** | 2594 | 2571 | | |
| Seattle | 2182 | 1737 | 1021 | 1891 | 959 | 2734 | 2408 | 678 | |
| Washington DC | 543 | 597 | 1494 | 1220 | 2300 | 923 | **205** | 2442 | 2329 |

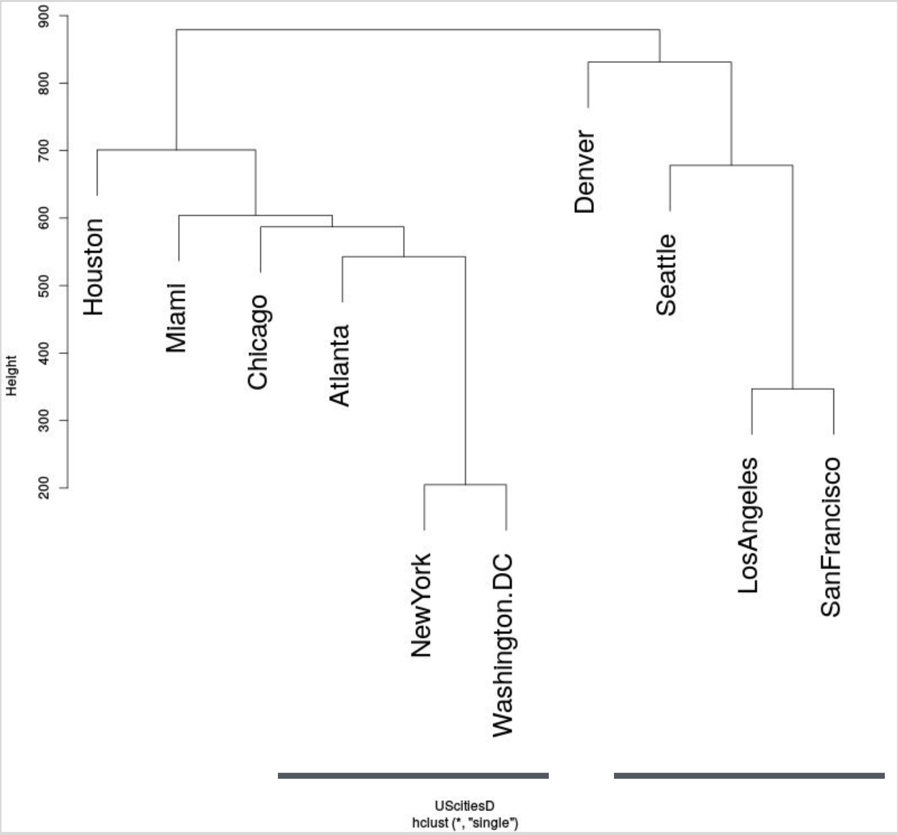
Atlanta - (Washington - New York ) = (543 + 748) / 2 = **645.2**

# Hierarchical clustering - average linkage

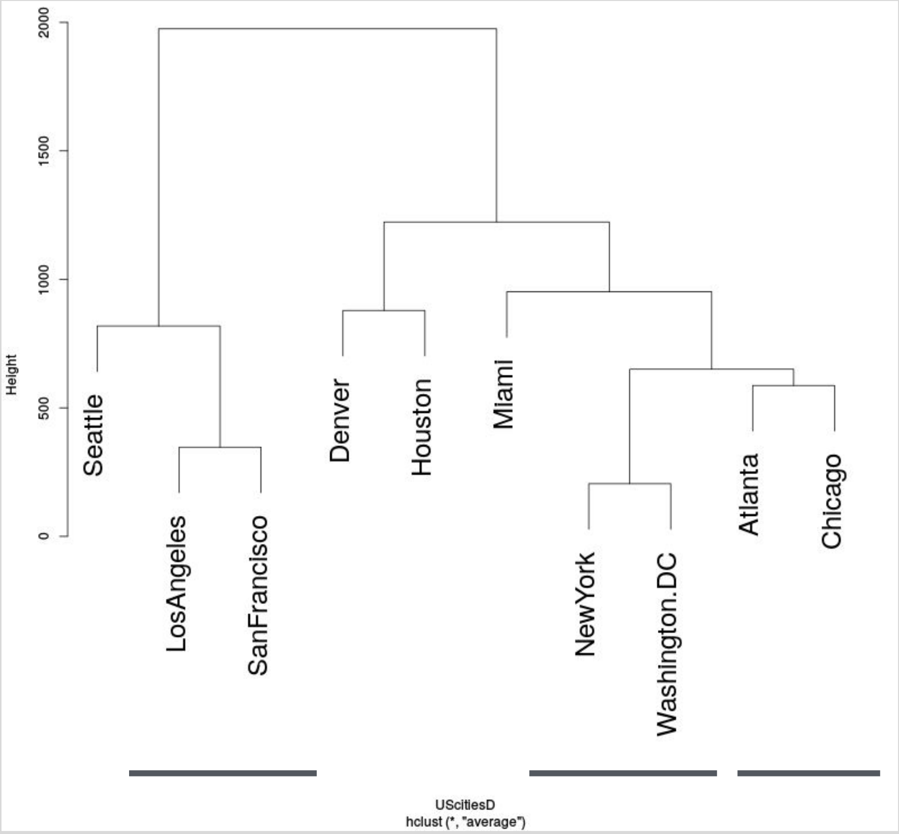


# Hierarchical clustering

## SINGLE linkage



## AVERAGE linkage



# Hierarchical clustering

**Outcome:**

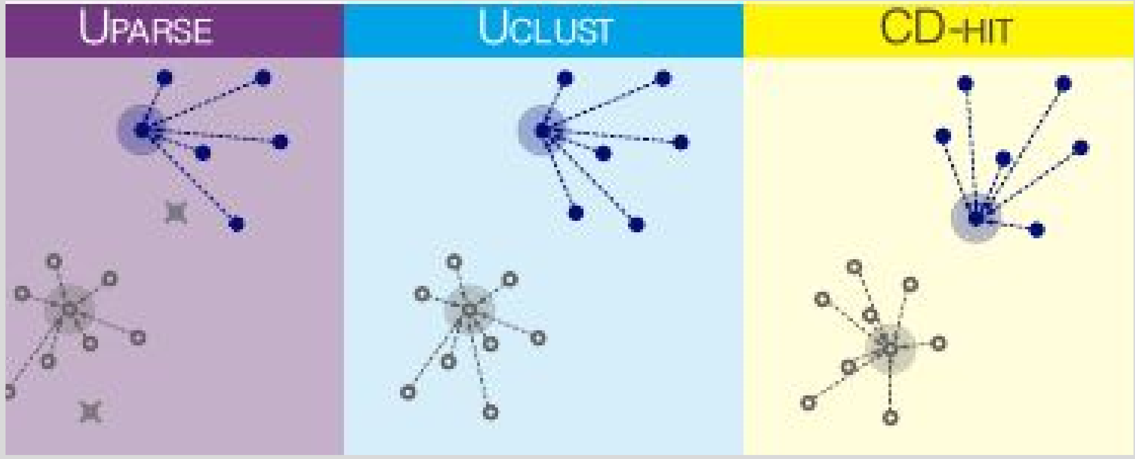
- **Hierarchical clustering depends on the linkage policy**
- **All distances need to be known**
- **Hierarchical clustering is expensive, in general agglomerative strategies cost O($n^3$)…**

**Example :**

**n=200 sequences**

**Cost agglomerative = 8e+06 operations**

# Greedy clustering



**Algorithm:**

- **Initial n groups <u>ordered</u> in particular way**
- **Each step:**
  - **Pick a group and compare to the reference**
  - **If close to the reference:**
    - **Add in reference cluster**
  - **Otherwise:**
    - **Add it as a reference**

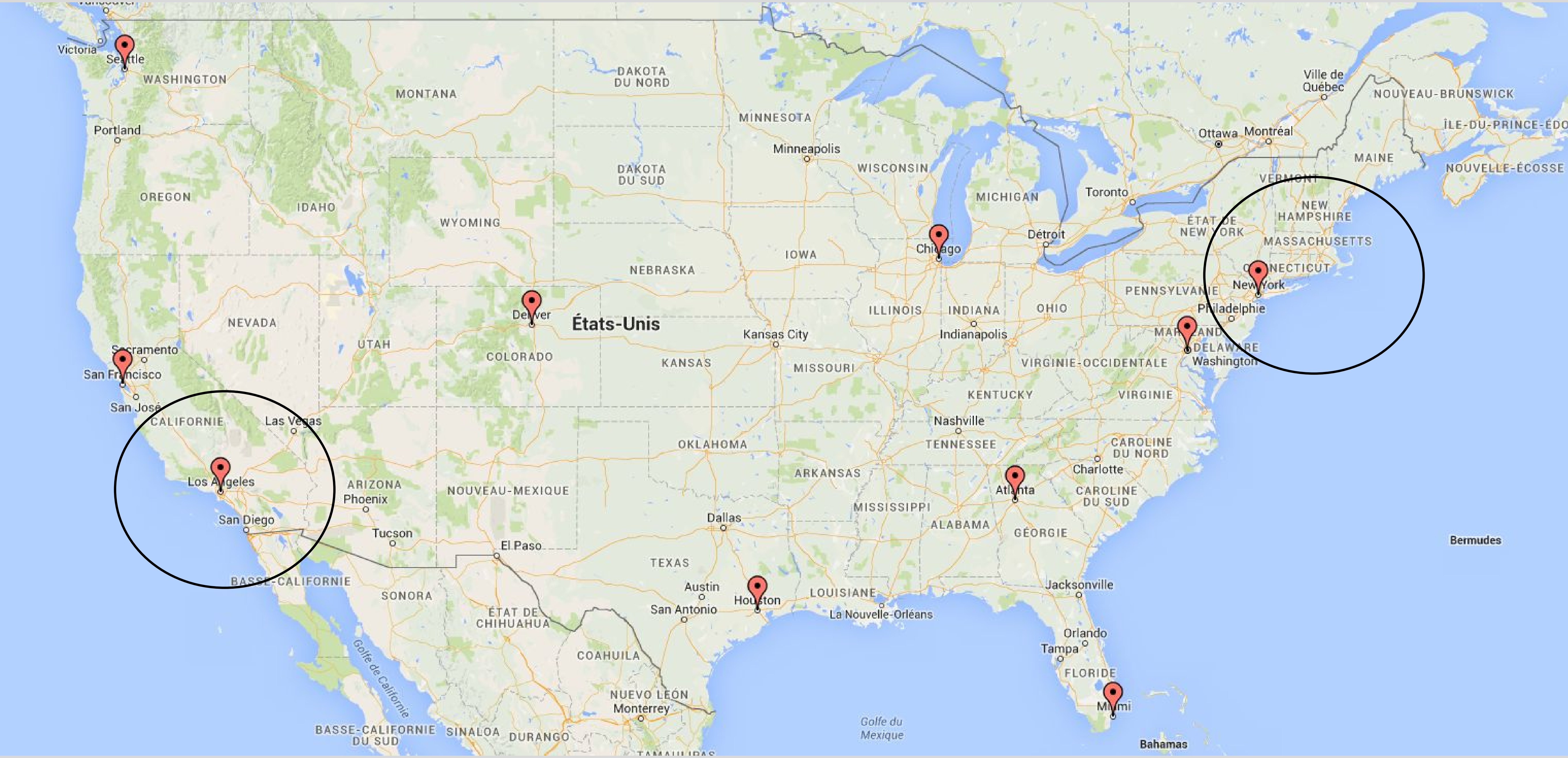
- **Ordering:**
  - **Length-based Greedy Clustering (CD-HIT, Uclust)**
  - **Abundance-based Greedy Clustering (AGC) : “Most-Abundant-centroid”**

# Abundance-based Greedy Clustering methods

| City | Population |
|---|---|
| New York | 8550405 |
| Los Angeles | 3958125 |
| Chicago | 2722389 |
| Houston | 2099451 |
| San Francisco | 852469 |
| Washington DC | 646449 |
| Seattle | 634535 |
| Denver | 634265 |
| Atlanta | 443775 |
| Miami | 430332 |

New York - Los Angeles = 2451 km > 800 km

# Abundance-based Greedy Clustering methods



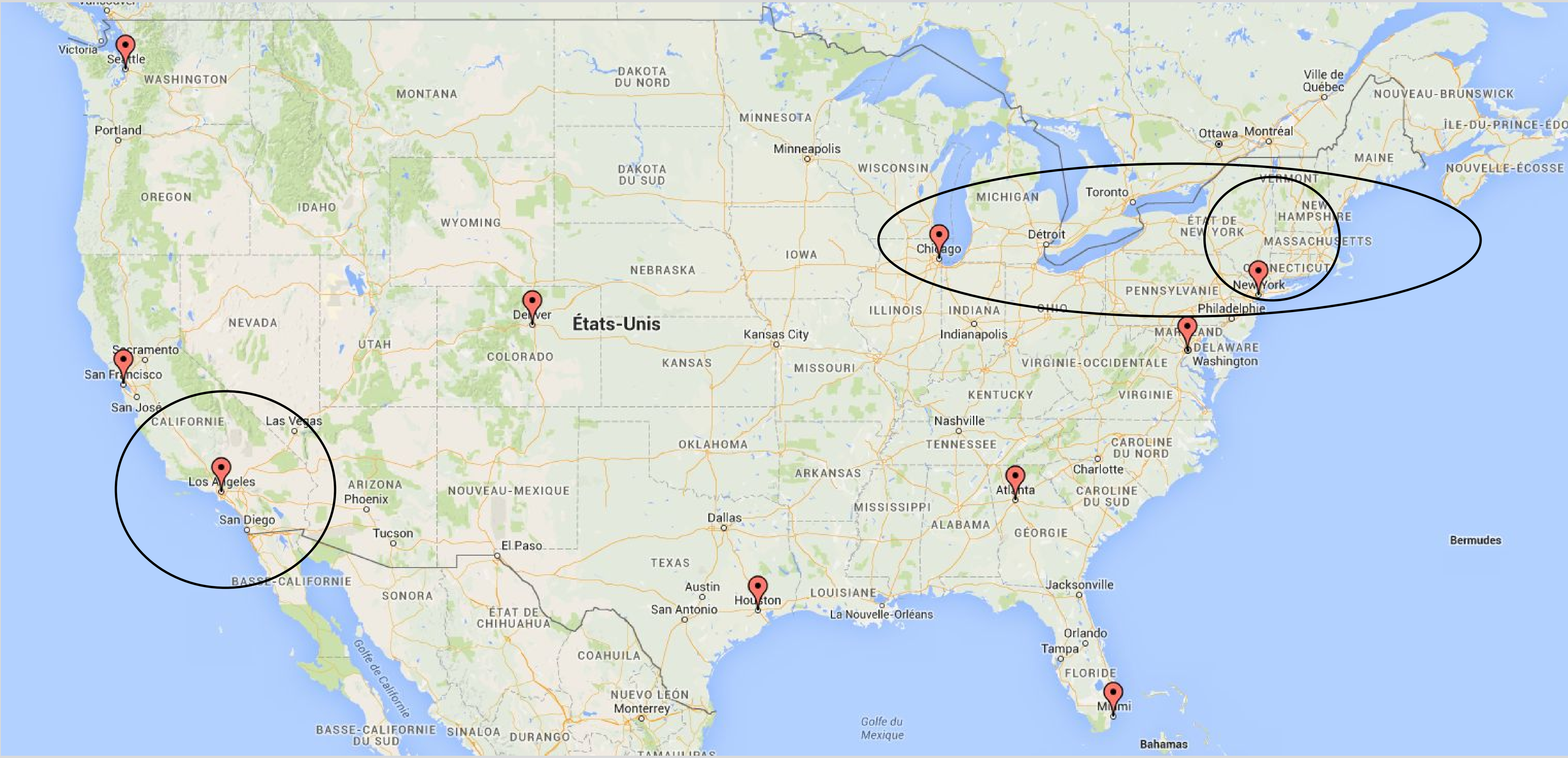
# Abundance-based Clustering methods

| City | Population |
|---|---|
| New York | 8550405 |
| Los Angeles | 3958125 |
| Chicago | 2722389 |
| Houston | 2099451 |
| San Francisco | 852469 |
| Washington DC | 646449 |
| Seattle | 634535 |
| Denver | 634265 |
| Atlanta | 443775 |
| Miami | 430332 |

New york - Los Angeles = 2451 km > 700 km

New york - Chicago = 713 km

# Abundance-based Greedy Clustering methods



# Abundance greedy clustering

**Outcome:**

- ❖ **AGC depend on the sorting strategy (length, abundance…)**
- ❖ **The distance to the reference is guarantee…**
- ❖ **…not the distance between sequences in the OTU**
- ❖ **AGC cost is in the worst case O($n^2$)…**

**Example :**

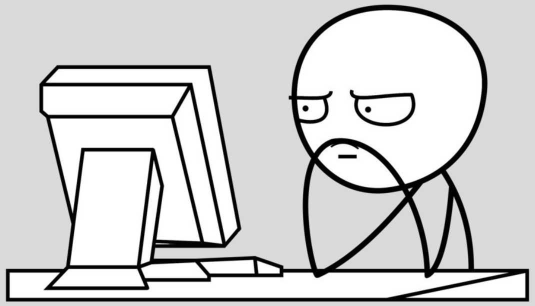
**n=200 sequences**

**Cost = 40000 operations <<< 8e+06 operations in hierarchical clustering**

# What is the best approach ?

**Not a simple question, how to evaluate the different approach ?**

- ❖ **Number of OTU ?**
- ❖ **Stability of OTU ?**
- ❖ **Quality of OTU ?**
- ❖ **Diversity ?**
- ❖ **Computational time ?**

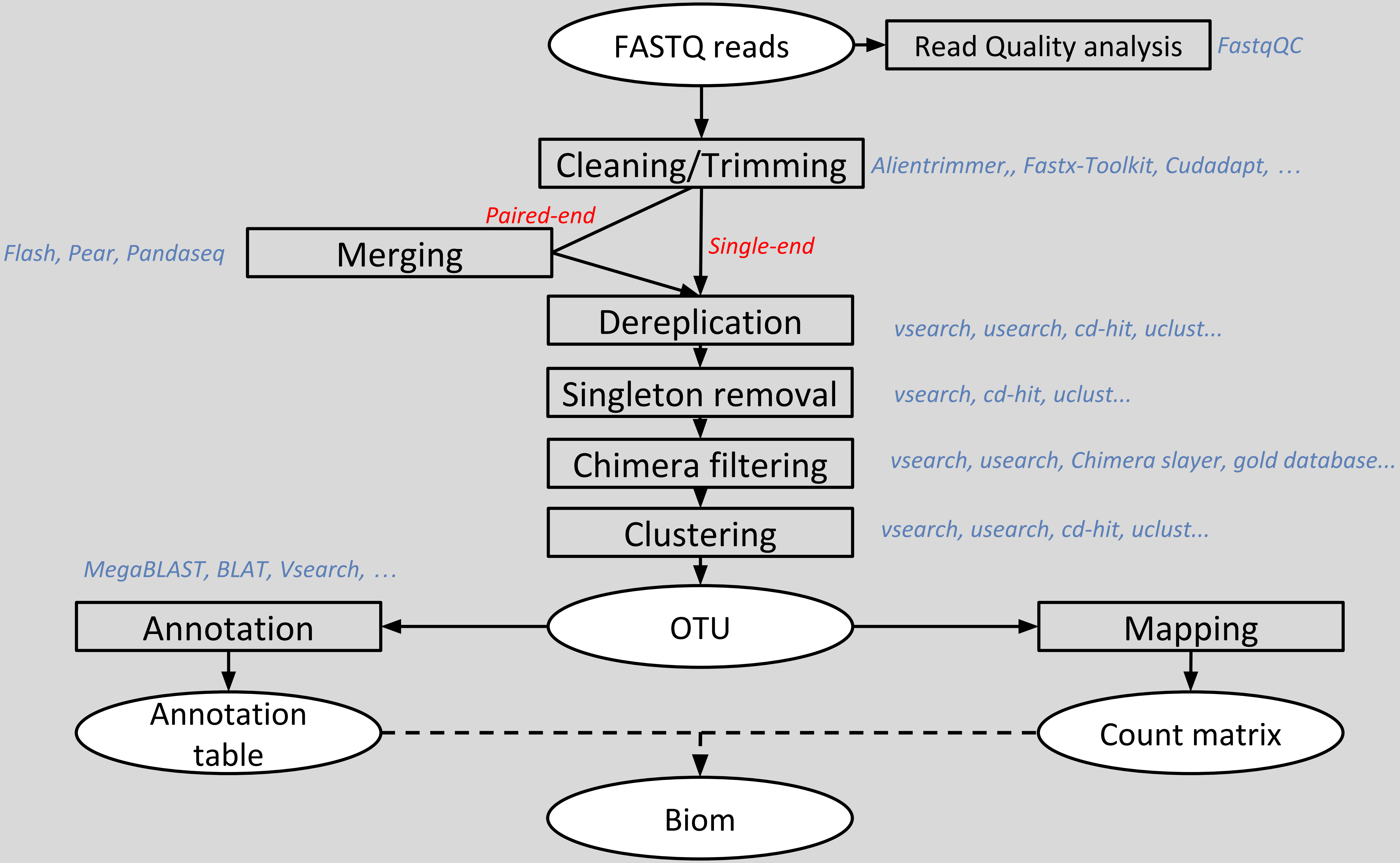


***Sequence quality filtering (trimming, filtering) has huge impact too.**

**Answer is maybe all [**Westcott, Schloss, 2016 PeerJ; Rideout 2014; Schmidt et al. 2015 **]**

**Vsearch seems to stand out in de novo approach**

# AGC-Targeted metagenomics pipeline



FASTQ reads

Read Quality analysis

*FastqQC*

Cleaning/Trimming

*Alientrimmer,, Fastx-Toolkit, Cudadapt, …*

*Paired-end*

*Single-end*

*Flash, Pear, Pandaseq*

Merging

Dereplication

*vsearch, usearch, cd-hit, uclust...*

Singleton removal

*vsearch, cd-hit, uclust...*

Chimera filtering

*vsearch, usearch, Chimera slayer, gold database...*

Clustering

*vsearch, usearch, cd-hit, uclust...*

*MegaBLAST, BLAT, Vsearch, …*

Annotation

OTU

Mapping

Annotation table

Count matrix

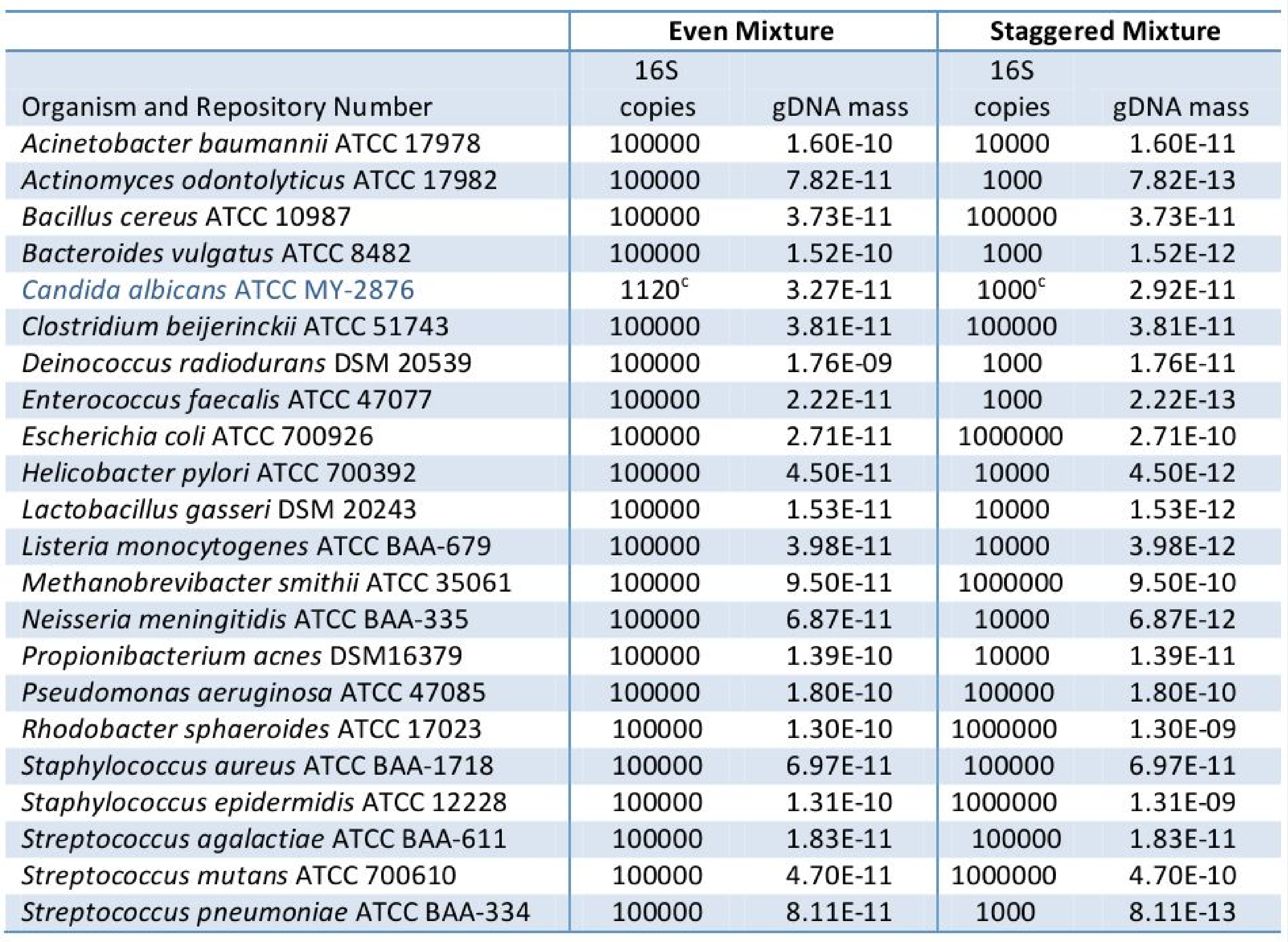
Biom

# TP

**Terminal (in your home folder):**

```
$ cd masque/tp/
$./cahier_handOnmetagenomics.sh
```

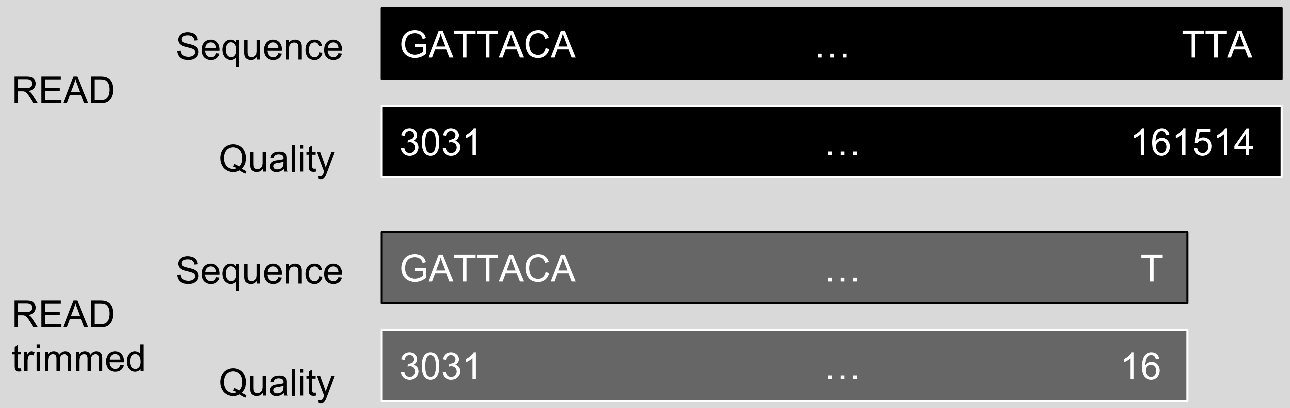
# Mock communities



| | Even Mixture | | Staggered Mixture | |
|---|---|---|---|---|
| Organism and Repository Number | 16S copies | gDNA mass | 16S copies | gDNA mass |
| *Acinetobacter baumannii* ATCC 17978 | 100000 | 1.60E-10 | 10000 | 1.60E-11 |
| *Actinomyces odontolyticus* ATCC 17982 | 100000 | 7.82E-11 | 1000 | 7.82E-13 |
| *Bacillus cereus* ATCC 10987 | 100000 | 3.73E-11 | 100000 | 3.73E-11 |
| *Bacteroides vulgatus* ATCC 8482 | 100000 | 1.52E-10 | 1000 | 1.52E-12 |
| *Candida albicans* ATCC MY-2876 | 1120[c] | 3.27E-11 | 1000[c] | 2.92E-11 |
| *Clostridium beijerinckii* ATCC 51743 | 100000 | 3.81E-11 | 100000 | 3.81E-11 |
| *Deinococcus radiodurans* DSM 20539 | 100000 | 1.76E-09 | 1000 | 1.76E-11 |
| *Enterococcus faecalis* ATCC 47077 | 100000 | 2.22E-11 | 1000 | 2.22E-13 |
| *Escherichia coli* ATCC 700926 | 100000 | 2.71E-11 | 1000000 | 2.71E-10 |
| *Helicobacter pylori* ATCC 700392 | 100000 | 4.50E-11 | 10000 | 4.50E-12 |
| *Lactobacillus gasseri* DSM 20243 | 100000 | 1.53E-11 | 10000 | 1.53E-12 |
| *Listeria monocytogenes* ATCC BAA-679 | 100000 | 3.98E-11 | 10000 | 3.98E-12 |
| *Methanobrevibacter smithii* ATCC 35061 | 100000 | 9.50E-11 | 1000000 | 9.50E-10 |
| *Neisseria meningitidis* ATCC BAA-335 | 100000 | 6.87E-11 | 10000 | 6.87E-12 |
| *Propionibacterium acnes* DSM16379 | 100000 | 1.39E-10 | 10000 | 1.39E-11 |
| *Pseudomonas aeruginosa* ATCC 47085 | 100000 | 1.80E-10 | 100000 | 1.80E-10 |
| *Rhodobacter sphaeroides* ATCC 17023 | 100000 | 1.30E-10 | 1000000 | 1.30E-09 |
| *Staphylococcus aureus* ATCC BAA-1718 | 100000 | 6.97E-11 | 100000 | 6.97E-11 |
| *Staphylococcus epidermidis* ATCC 12228 | 100000 | 1.31E-10 | 1000000 | 1.31E-09 |
| *Streptococcus agalactiae* ATCC BAA-611 | 100000 | 1.83E-11 | 100000 | 1.83E-11 |
| *Streptococcus mutans* ATCC 700610 | 100000 | 4.70E-11 | 1000000 | 4.70E-10 |
| *Streptococcus pneumoniae* ATCC BAA-334 | 100000 | 8.11E-11 | 1000 | 8.11E-13 |

*The NIH HMP Working Group, 2009 (Genome Research)

# Trimming



READ

Sequence GATTACA … TTA

Quality 3031 … 161514

READ trimmed

Sequence GATTACA … T

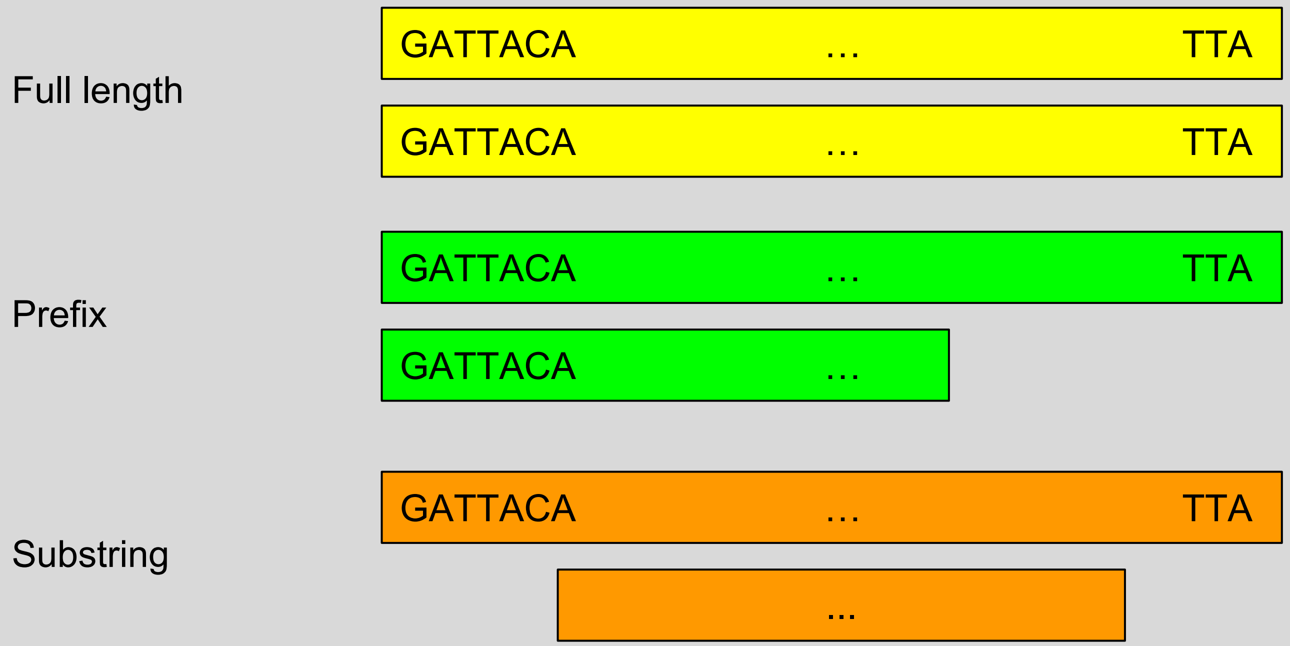
Quality 3031 … 16

```
vsearch --fastq_filter sample --fastqout sample_filt --fastq_truncqual 16 --fastq_trunclen 250
```

**I do not recommend Vsearch trimming, tp only !**

[http://drive5.com/uparse/](http://drive5.com/uparse/), [https://github.com/torognes/vsearch](https://github.com/torognes/vsearch)

# Dereplication and Singleton removal



Full length

GATTACA … TTA

GATTACA … TTA

Prefix

GATTACA … TTA

GATTACA …

Substring

GATTACA … TTA

…

```
vsearch --derep_fulllength sample -output sample_drep -sizeout
```

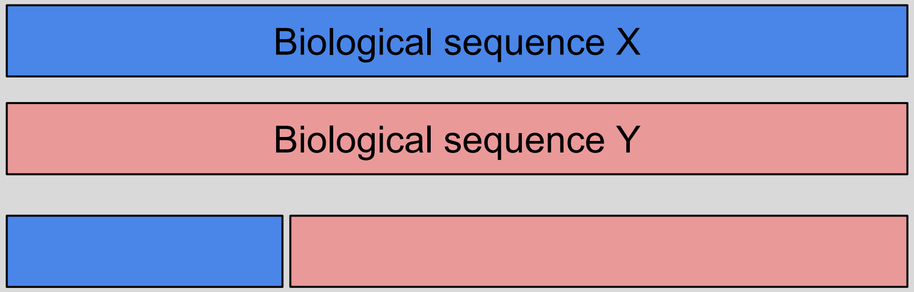
*Abundance

```
vsearch -sortbysize sample_drep -output sample_nosing  -minsize 2
```

We could be more stringent

http://drive5.com/uparse/, https://github.com/torognes/vsearch

# Chimera filtering

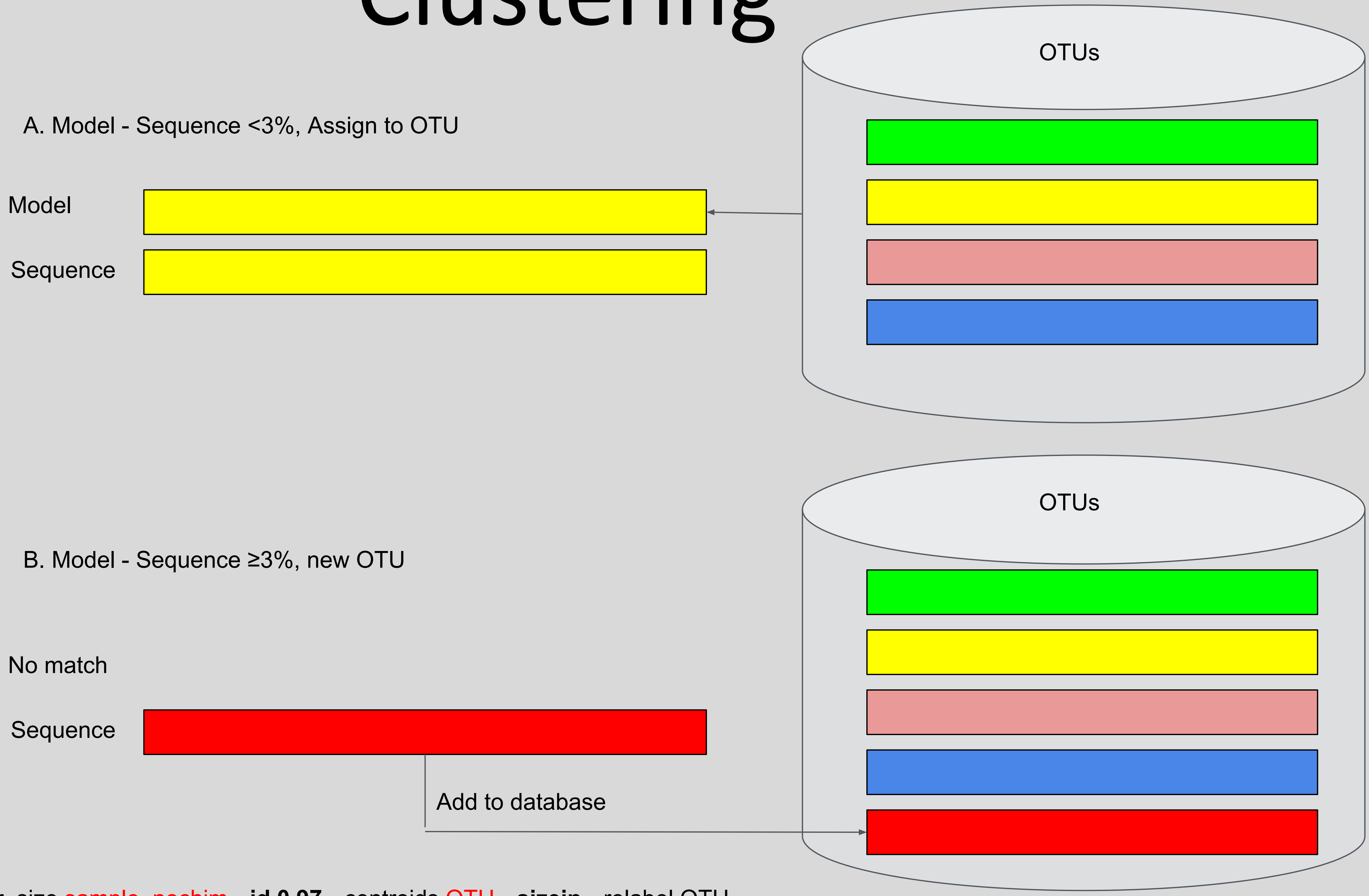


Chimera formed from X and Y

```
vsearch --uchime_denovo sample_nosing --chimeras sample_chim --nonchimeras sample_nochim
```

http://drive5.com/uparse/, https://github.com/torognes/vsearch

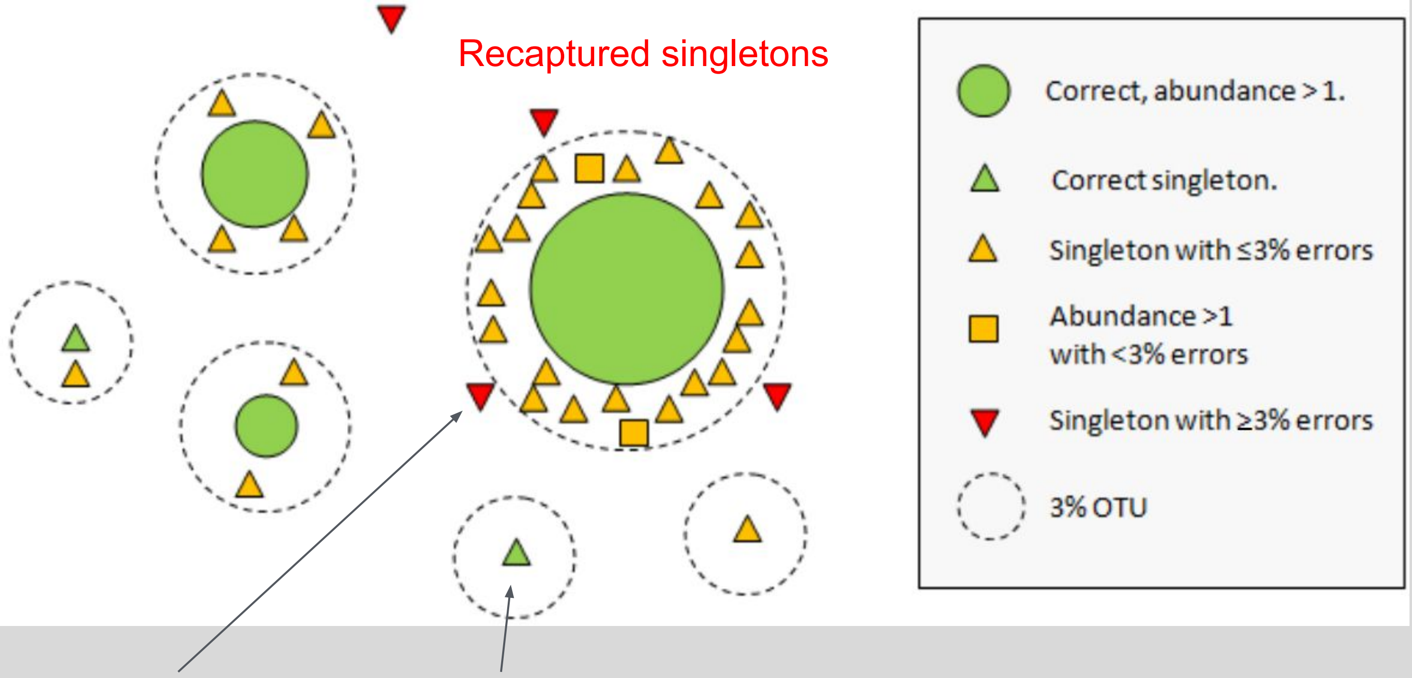
# Clustering


OTUs
A. Model - Sequence <3%, Assign to OTU
Model
Sequence
OTUs
B. Model - Sequence ≥3%, new OTU
No match
Sequence
Add to database

vsearch --cluster_size sample_nochim **--id 0.97** --centroids OTU **--sizein** --relabel OTU_

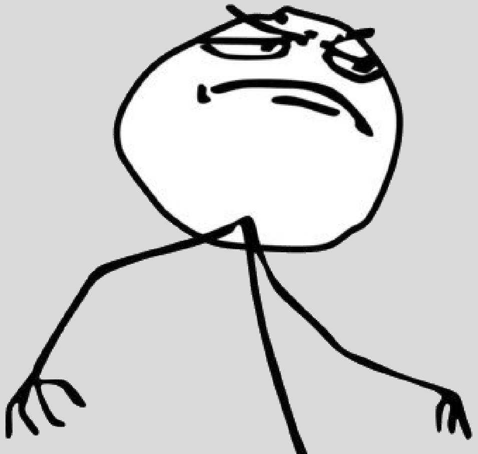
http://drive5.com/uparse/, https://github.com/torognes/vsearch

# Mapping



“Bad” singleton Isolated singleton

```
vsearch -usearch_global sample -db OTU  --id 0.97 -uc map

uc2otutab.py map > otu_table
```



*Edgar et al. 2013 (Nature methods)

# Taxonomical annotation

**SILVA [Pruesse, et al. 2007]:**

- 597,607 sequences (last update 04/2016)
- Small (16S/18S, SSU) and large subunit (23S/28S, LSU)
- Bacteria, Archaea and Eukarya
- Non redundant (Uclust 99% id)
- Based on EMBL-bank

silva
high quality ribosomal RNA databases

green genes
16S rRNA gene database and workbench compatible with ARB
greengenes.lbl.gov

**Greengenes [DeSantis et al. 2006]:**

- 1,262,986 sequences (last update 05/2013)
- Small subunit (16S/18S, SSU)
- Bacteria and Archaea
- Non redundant (Uclust 99% id)
- Based on Genbank

**Ribosomal Database Project [Maidak et al. 1994]:**

- 3,224,600 + 108,901 sequences (last update 05/2015)
- Small subunit (16S/18S, SSU) and Fungal 28S
- Bacteria, Archaea and Fungi

rdp

# Taxonomical annotation

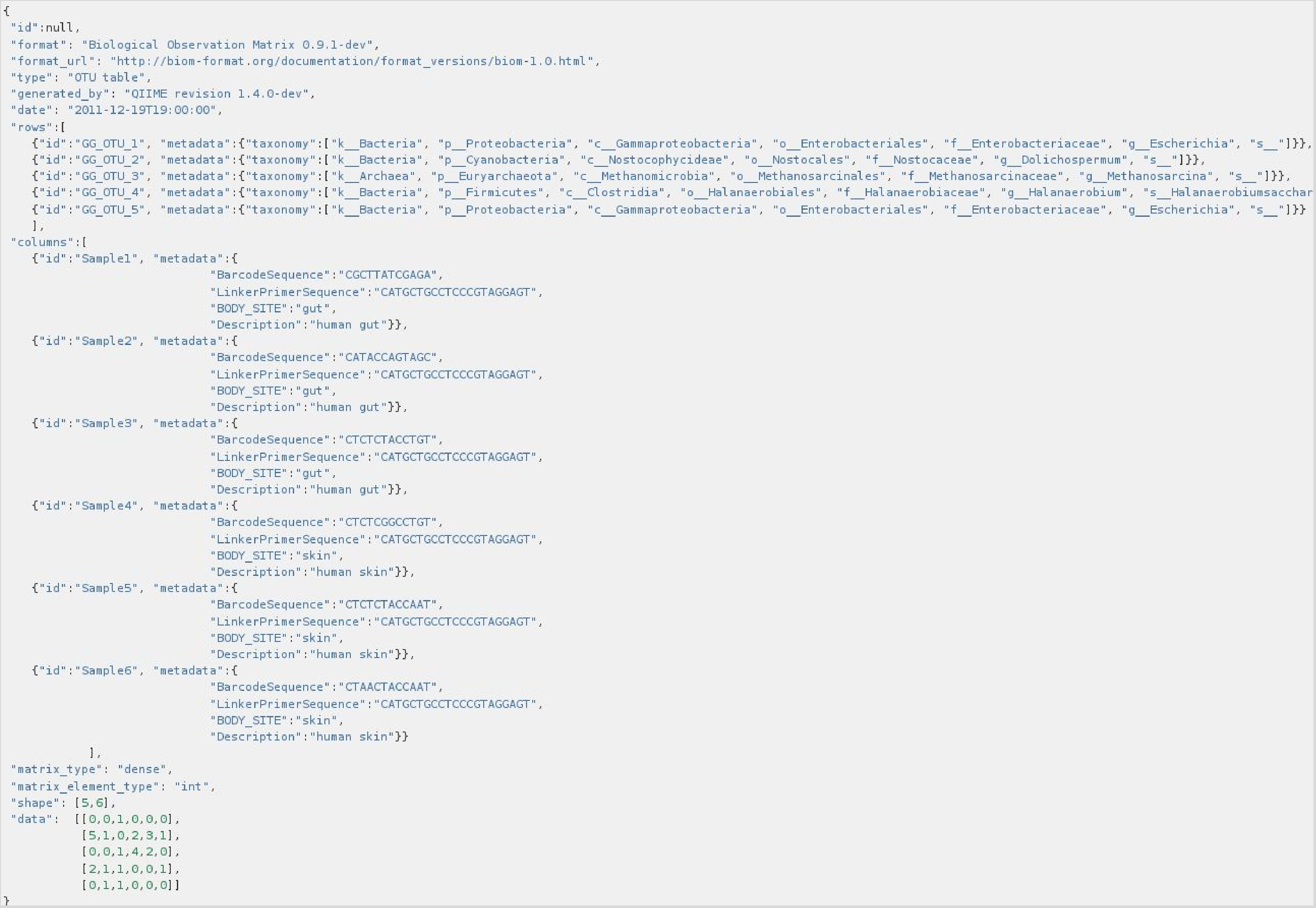
```
vsearch --usearch_global OTU --db database --id 0.9  --blast6out annotation --alnout alignment

get_taxonomy.py -i annotation -u OTU -d database -o annotation_table -ob annotation_biom
```

# BIOM format

**Motivation:**

- Encapsulation of the whole project (count table, annotation, metadata...)
- Efficient storage
- Compatibility between softwares

# BIOM format version 1.0



Annotation

Metadata

Count

```
{
 "id":null,
 "format": "Biological Observation Matrix 0.9.1-dev",
 "format_url": "http://biom-format.org/documentation/format_versions/biom-1.0.html",
 "type": "OTU table",
 "generated_by": "QIIME revision 1.4.0-dev",
 "date": "2011-12-19T19:00:00",
 "rows":[
    {"id":"GG_OTU_1", "metadata":{"taxonomy":["k__Bacteria", "p__Proteobacteria", "c__Gammaproteobacteria", "o__Enterobacteriales", "f__Enterobacteriaceae", "g__Escherichia", "s__"]}},
    {"id":"GG_OTU_2", "metadata":{"taxonomy":["k__Bacteria", "p__Cyanobacteria", "c__Nostocophycideae", "o__Nostocales", "f__Nostocaceae", "g__Dolichospermum", "s__"]}},
    {"id":"GG_OTU_3", "metadata":{"taxonomy":["k__Archaea", "p__Euryarchaeota", "c__Methanomicrobia", "o__Methanosarcinales", "f__Methanosarcinaceae", "g__Methanosarcina", "s__"]}},
    {"id":"GG_OTU_4", "metadata":{"taxonomy":["k__Bacteria", "p__Firmicutes", "c__Clostridia", "o__Halanaerobiales", "f__Halanaerobiaceae", "g__Halanaerobium", "s__Halanaerobiumsacchar
    {"id":"GG_OTU_5", "metadata":{"taxonomy":["k__Bacteria", "p__Proteobacteria", "c__Gammaproteobacteria", "o__Enterobacteriales", "f__Enterobacteriaceae", "g__Escherichia", "s__"]}}
    ],
 "columns":[
    {"id":"Sample1", "metadata":{
                             "BarcodeSequence":"CGCTTATCGAGA",
                             "LinkerPrimerSequence":"CATGCTGCCTCCCGTAGGAGT",
                             "BODY_SITE":"gut",
                             "Description":"human gut"}},
    {"id":"Sample2", "metadata":{
                             "BarcodeSequence":"CATACCAGTAGC",
                             "LinkerPrimerSequence":"CATGCTGCCTCCCGTAGGAGT",
                             "BODY_SITE":"gut",
                             "Description":"human gut"}},
    {"id":"Sample3", "metadata":{
                             "BarcodeSequence":"CTCTCTACCTGT",
                             "LinkerPrimerSequence":"CATGCTGCCTCCCGTAGGAGT",
                             "BODY_SITE":"gut",
                             "Description":"human gut"}},
    {"id":"Sample4", "metadata":{
                             "BarcodeSequence":"CTCTCGGCCTGT",
                             "LinkerPrimerSequence":"CATGCTGCCTCCCGTAGGAGT",
                             "BODY_SITE":"skin",
                             "Description":"human skin"}},
    {"id":"Sample5", "metadata":{
                             "BarcodeSequence":"CTCTCTACCAAT",
                             "LinkerPrimerSequence":"CATGCTGCCTCCCGTAGGAGT",
                             "BODY_SITE":"skin",
                             "Description":"human skin"}},
    {"id":"Sample6", "metadata":{
                             "BarcodeSequence":"CTAACTACCAAT",
                             "LinkerPrimerSequence":"CATGCTGCCTCCCGTAGGAGT",
                             "BODY_SITE":"skin",
                             "Description":"human skin"}}
            ],
 "matrix_type": "dense",
 "matrix_element_type": "int",
 "shape": [5,6],
 "data":  [[0,0,1,0,0,0],
           [5,1,0,2,3,1],
           [0,0,1,4,2,0],
           [2,1,1,0,0,1],
           [0,1,1,0,0,0]]
}
```

# BIOM format

## Motivation:

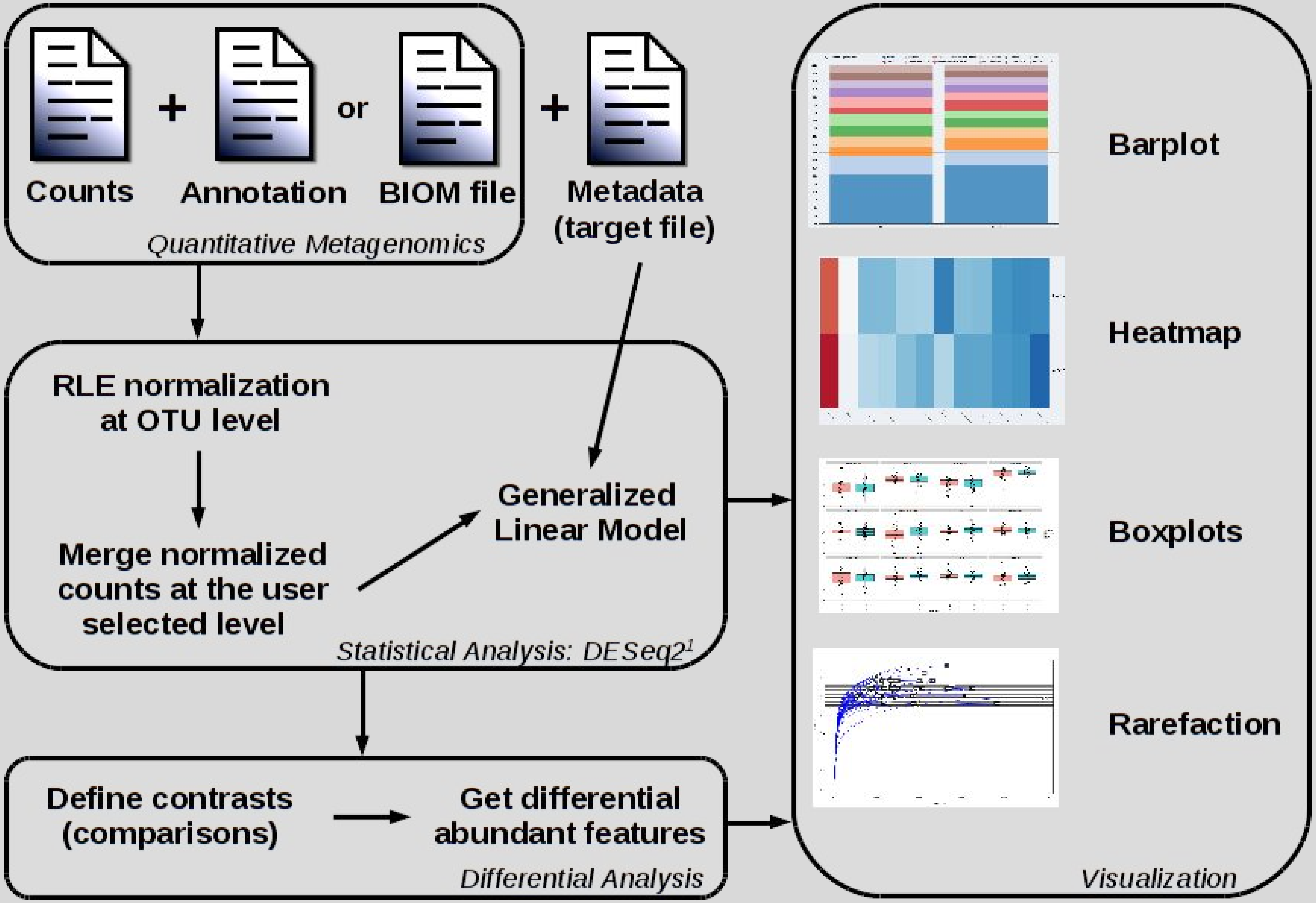
- Efficient storage
- Encapsulation of the whole project (count table, annotation, metadata...)
- Compatibility between softwares

## Cons:

- Not human readable
- 3 different versions of BIOM format (1.0, 2.0, 2.1)
- Not strict enough in the version 1.0
- BIOM library does not provide good support of every version

```
biom convert -i otu_table -o biom --table-type="OTU table" --to-json
biom add-metadata -i biom -o annotated_biom --observation-metadata-fp annotation_table_biom --observation-header id,taxonomy --sc-separated taxonomy
```

# Differential analysis : SHAMAN



Counts + Annotation or BIOM file + Metadata (target file)

*Quantitative Metagenomics*

RLE normalization at OTU level

Merge normalized counts at the user selected level

Generalized Linear Model

*Statistical Analysis: DESeq2*[1]

Define contrasts (comparisons)

Get differential abundant features

*Differential Analysis*

Barplot

Heatmap

Boxplots

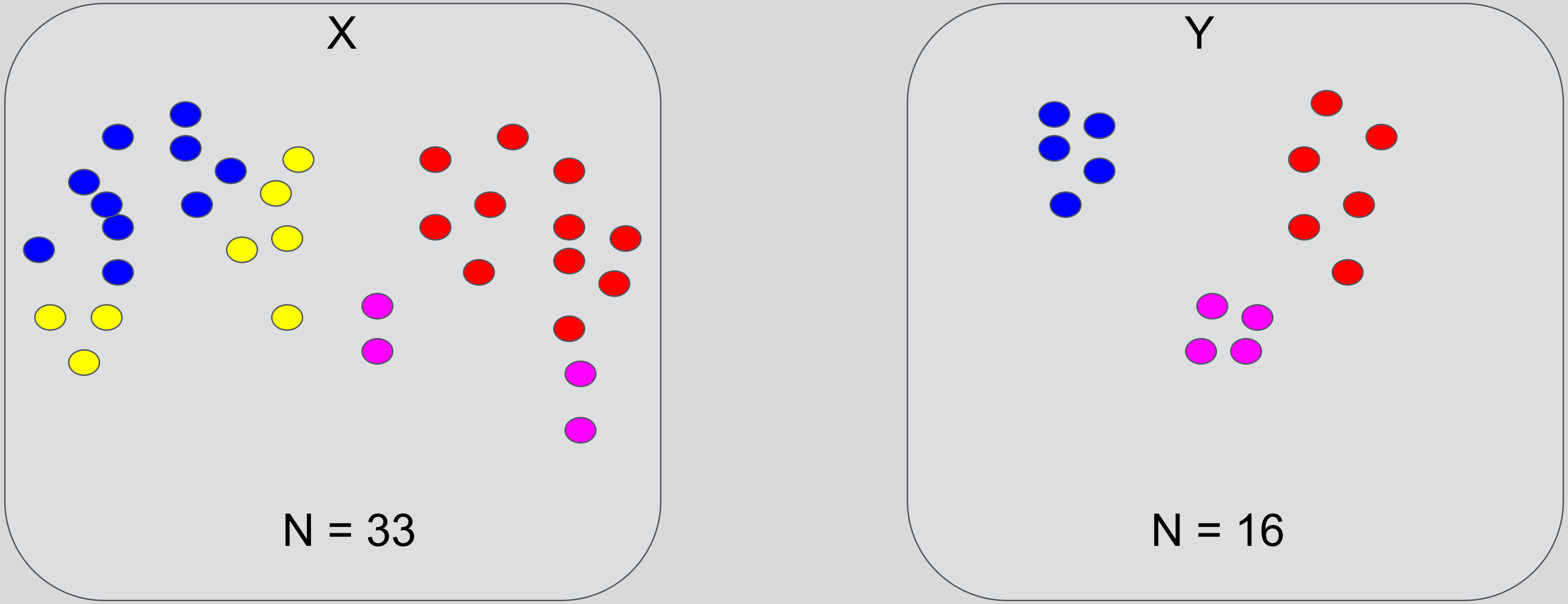
Rarefaction

*Visualization*

[1] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol.

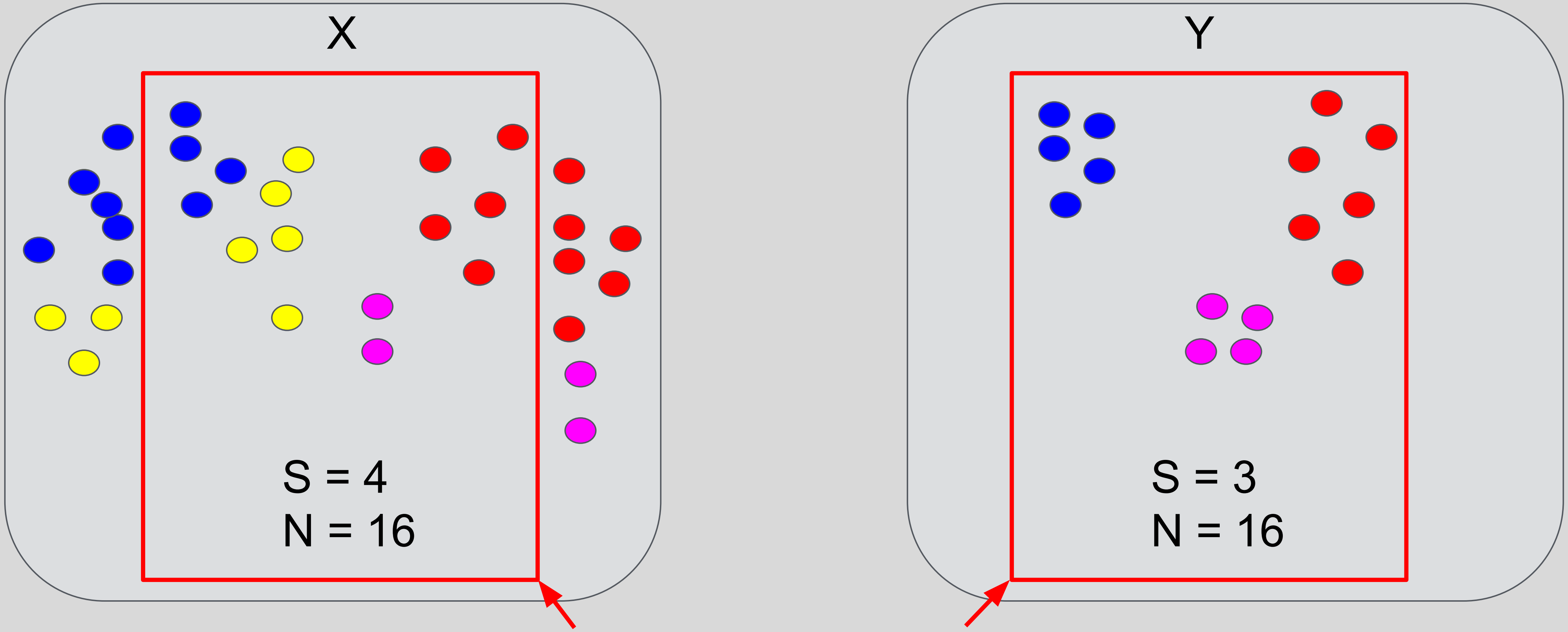
http://157.99.181.67:3838/shaman/

# Diversity



- **N: Total count of individual**
- **Rarefaction: a type of normalisation : Rarefy to the same number of individual**

# Diversity

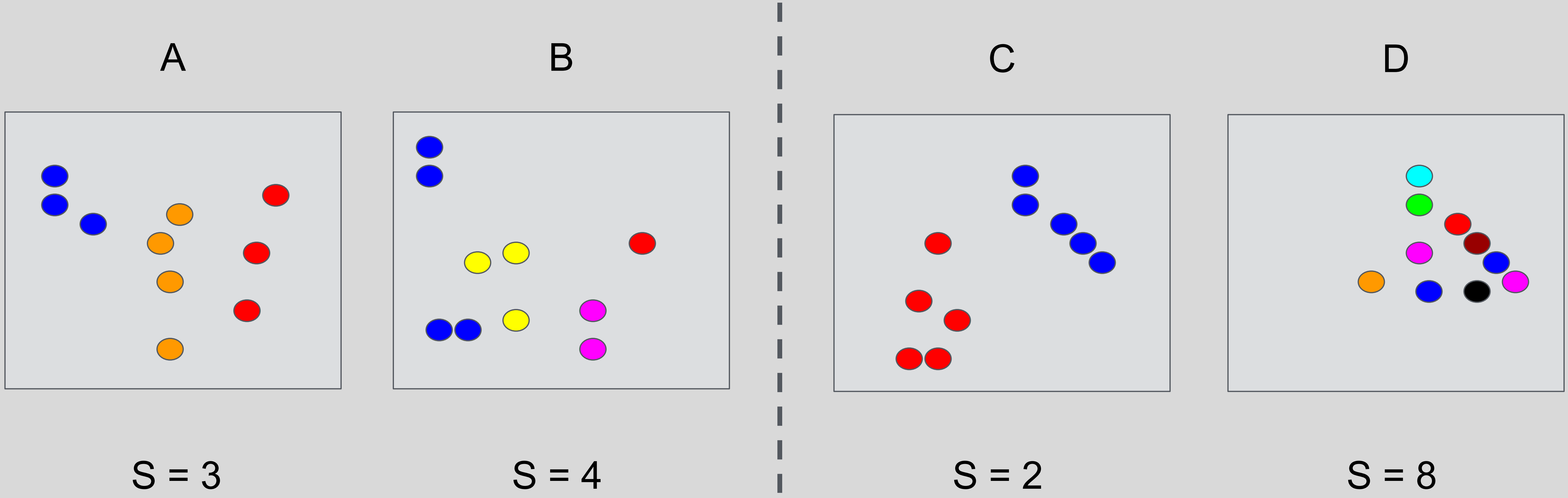


RAREFACTION = “DOWNSIZING”

- N: Total count of individual
- Rarefaction: a type of normalisation : Rarefy to the same number of individual
- S = number of species = richness = number of object > 0

# Alpha diversity

CONDITION 1 CONDITION 2



A B C D

S = 3<br>
N= 10

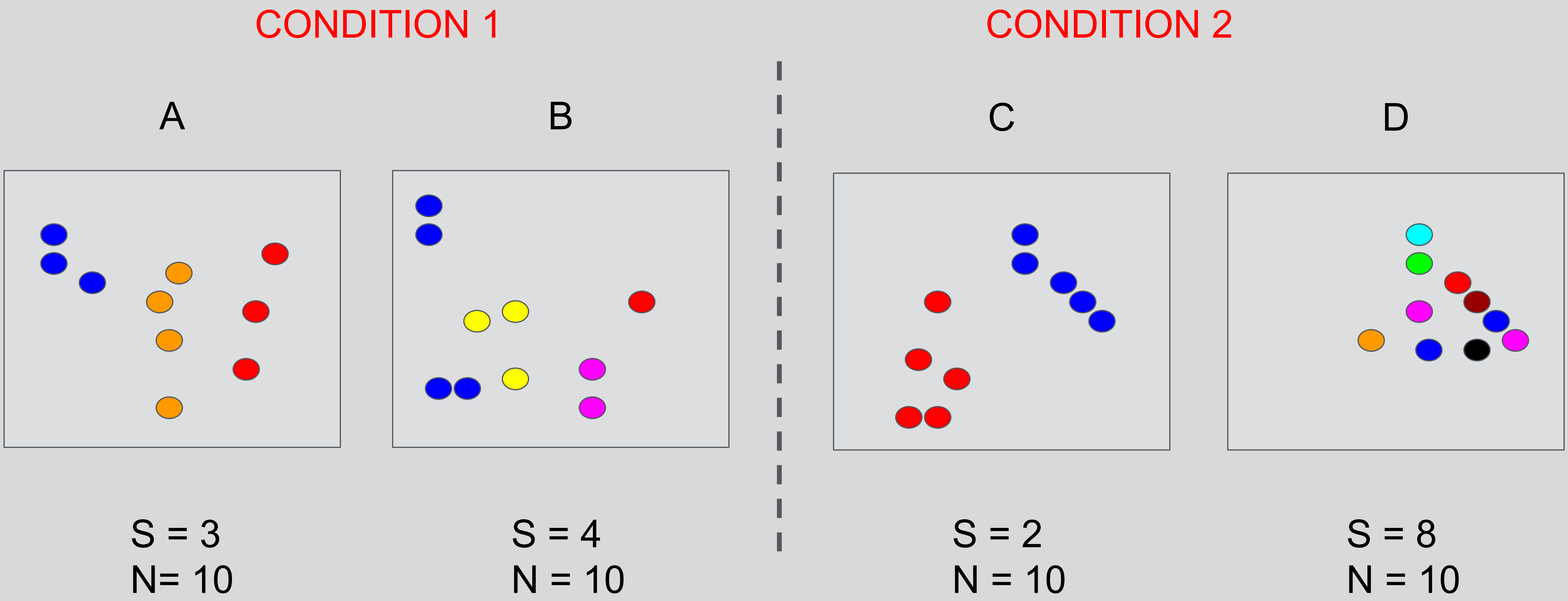
S = 4<br>
N = 10

S = 2<br>
N = 10

S = 8<br>
N = 10

- **S = number of species = richness = number of object > 0**
- **Alpha diversity:**
  - **Condition 1 : $\alpha_1$ = mean($S_A$, $S_B$) = 3.5**
  - **Condition 2 : $\alpha_2$ = mean($S_C$, $S_D$) = 5**

# Gamma diversity



CONDITION 1

CONDITION 2

A

B

C

D

S = 3
N= 10

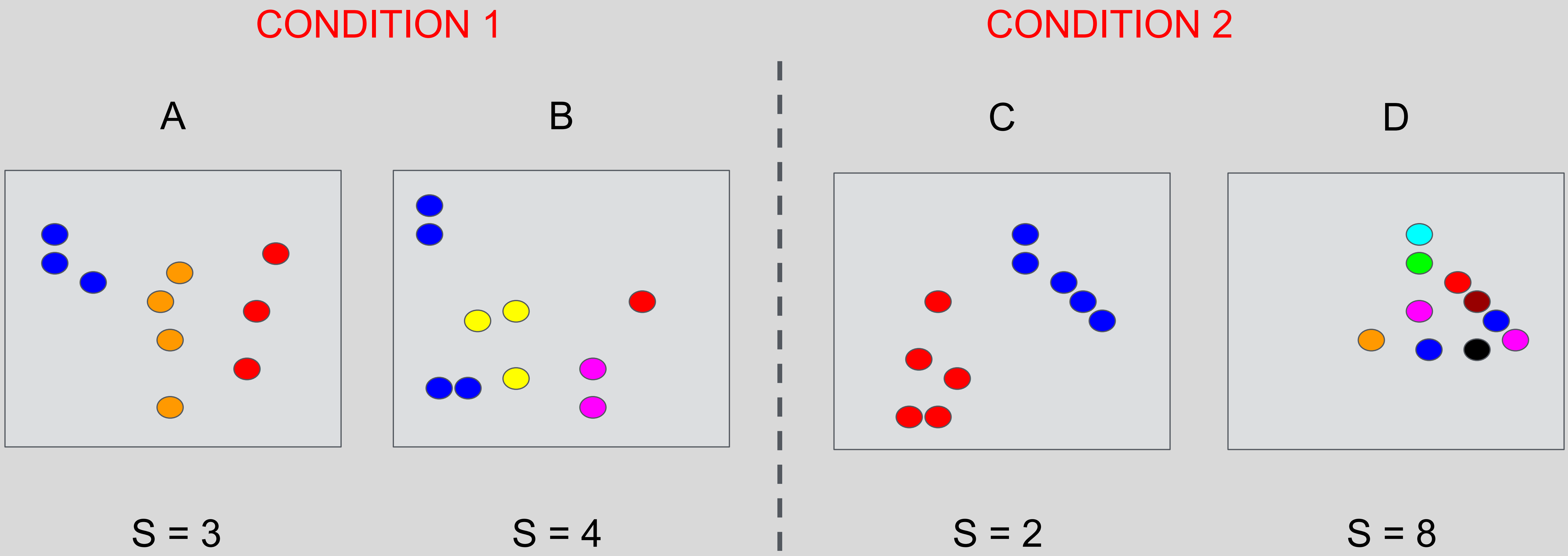
S = 4
N = 10

S = 2
N = 10

S = 8
N = 10

- **S = number of species = richness = number of object > 0**
- **Gamma diversity:**
  - **Condition 1 : $\gamma_1 = S_1 = 5$**
  - **Condition 2 : $\gamma_2 = S_2 = 8$**

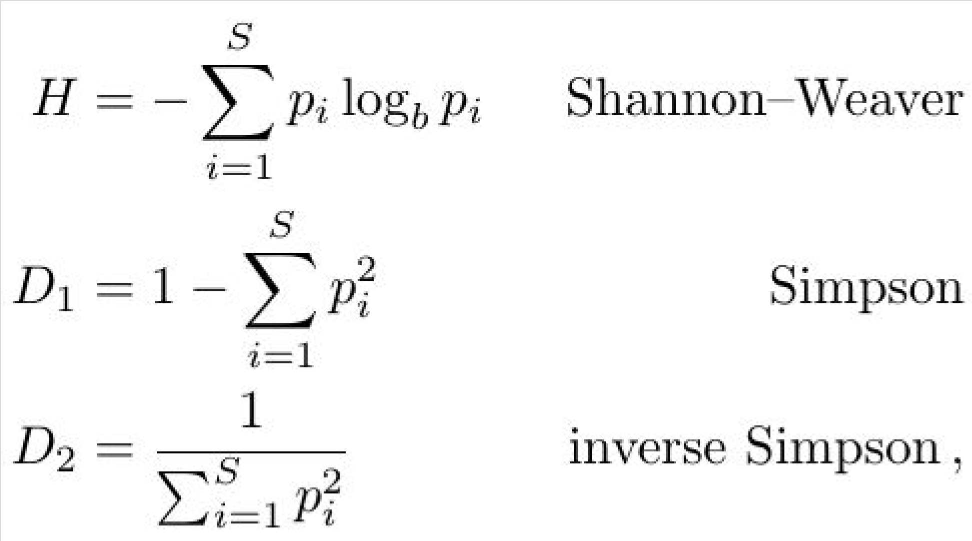
# Beta diversity



| A | B | C | D |
|---|---|---|---|
| S = 3<br>N= 10 | S = 4<br>N = 10 | S = 2<br>N = 10 | S = 8<br>N = 10 |

- **S = number of species = richness = number of object > 0**
- **Beta diversity:**
  - **Condition 1 : $\beta_1 = \dfrac{\gamma_1}{\alpha_1} - 1 = 0.43$**
  - **Condition 2 : $\beta_2 = 0.6$**

# Other diversity measures



$$H = -\sum_{i=1}^{S} p_i \log_b p_i \qquad \text{Shannon–Weaver}$$

$$D_1 = 1 - \sum_{i=1}^{S} p_i^2 \qquad \text{Simpson}$$

$$D_2 = \frac{1}{\sum_{i=1}^{S} p_i^2} \qquad \text{inverse Simpson}\,,$$

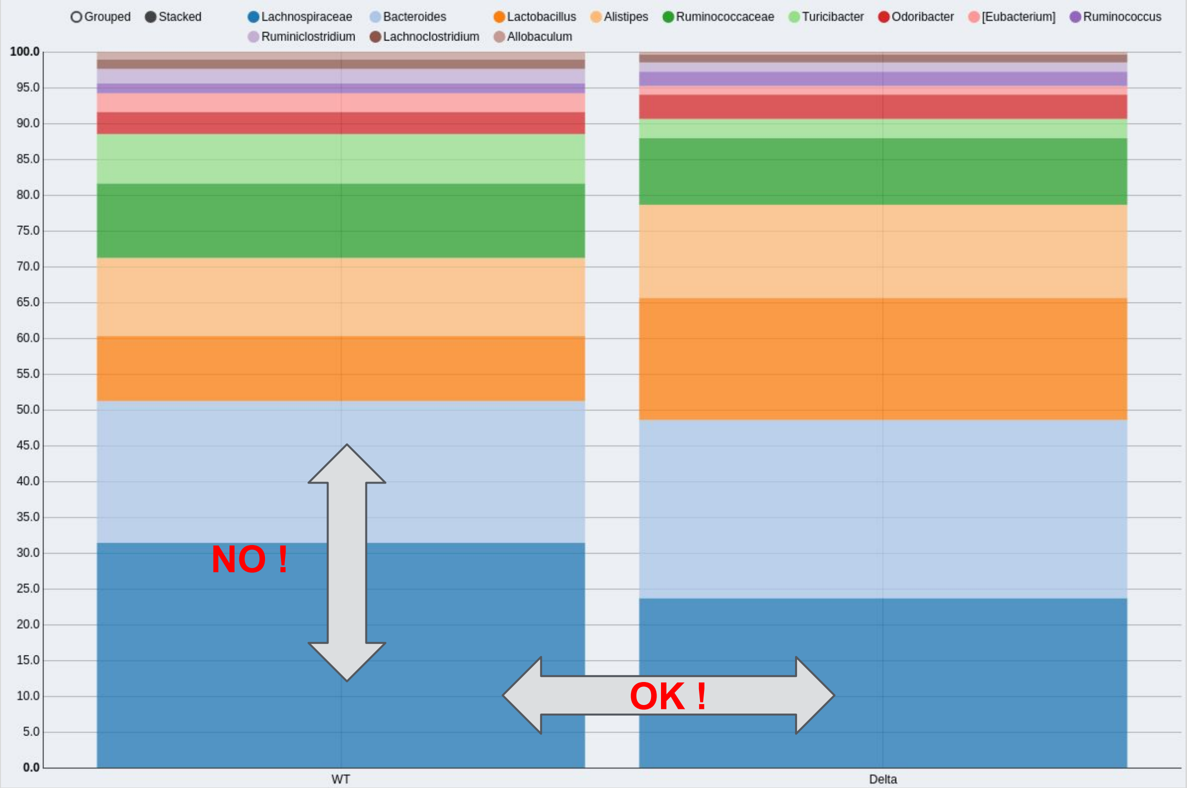
$P_i$ proportion of species i and S number of species

*Okasen J., Vegan R packages

# 16S rRNA limits

**Motivation:**

- Copy number varies in the genomes (from 1 to 15)*
- 16S sequence variants in the same specie and even genome** -> impact diversity



*Vetrovsky and Baldrian 2013 (Plos One), Klappenbach et al. 2001 (Nuc Acid Res), **Acinas et al. 2004 (J Bacteriol), ***Stoddard et al. 2015 (Nucl Acid Res)

# 16S rRNA limits

**Motivation:**

- Copy number varies in the genomes (from 1 to 15)*
- 16S sequence variants in the same specie and even genome** -> impact diversity

**Solutions:**

- rrnDB: ribosomal RNA operon copy number database***
- Good clustering and differential analysis
- Whole Genome Sequencing

*Vetrovsky and Baldrian 2013 (Plos One), Klappenbach et al. 2001 (Nuc Acid Res), **Acinas et al. 2004 (J Bacteriol), ***Stoddard et al. 2015 (Nucl Acid Res)

# 16S analysis at Pasteur

**Available:**

- MASQUE pipeline on bic and on tars
  - module use /pasteur/projets/Matrix/modules
  - module add masque/0.1 -> bic
  - module add masque/0.2 -> tars
- SHAMAN
- Galaxy : FROGS

*Vetrovsky and Baldrian 2013 (Plos One), Klappenbach et al. 2001 (Nuc Acid Res), **Acinas et al. 2004 (J Bacteriol), ***Stoddard et al. 2015 (Nucl Acid Res)