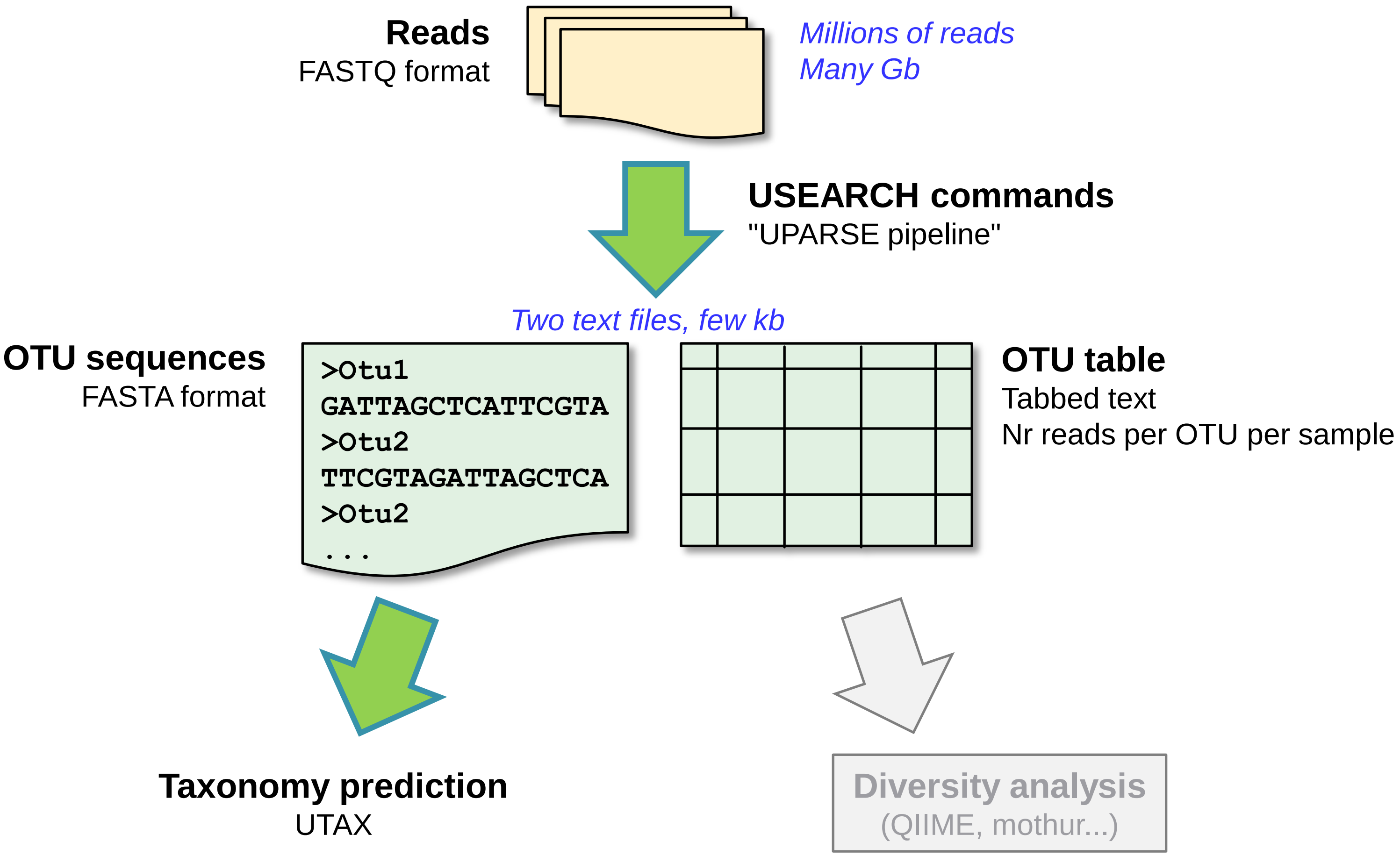
CLUSTERING

STAMPS 2016

Robert Edgar

Independent scientist
robert@drive5.com
www.drive5.com

OTU analysis



Reads
FASTQ format

Millions of reads
Many Gb

USEARCH commands
"UPARSE pipeline"

Two text files, few kb

OTU sequences
FASTA format

```
>Otu1
GATTAGCTCATTCGTA
>Otu2
TTCGTAGATTAGCTCA
>Otu2
...
```

OTU table
Tabbed text
Nr reads per OTU per sample

Taxonomy prediction
UTAX

Diversity analysis
(QIIME, mothur...)

Naive clustering

- Mock community with 20 species
- Cluster reads at 97% using UCLUST
- Thousands of "OTUs"
  - terrible result...
  - clusters are **noise!**

The magic number 97

Q. Why cluster at 97%?

a) Everybody does it

(true)

b) 97 is a happy prime

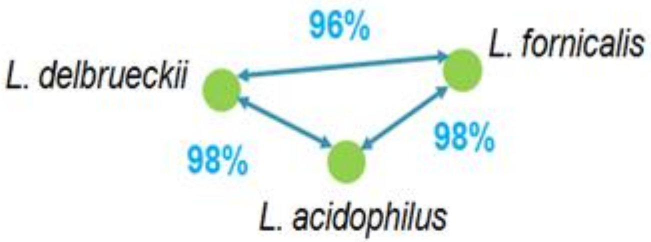
(*true -- look it up!*)

c) 97% clusters are species

(*not true*)

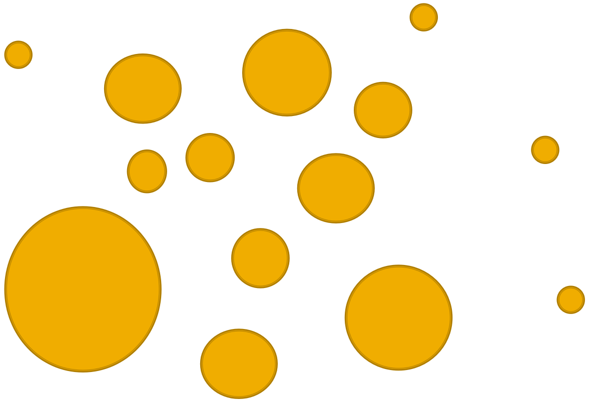
97% is not species

- Reasonable rule of thumb for **full-length** 16S
  - Paralogs in a single species usually >97%
    - But paralogs can be as low as 89%
  - Different strains usually >97%
  - Different species usually <97%
    - But not always, e.g. *Lactobacillus*



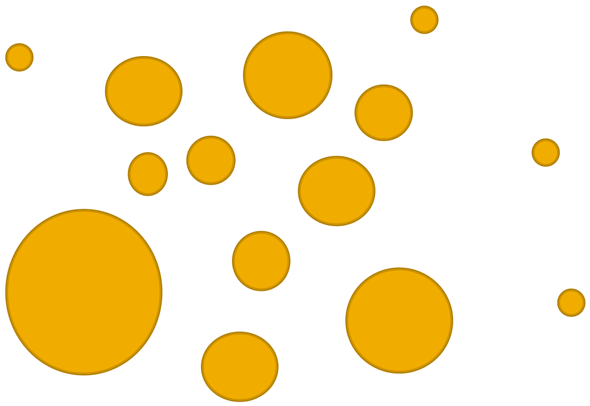
- Not so good for **short tags** like V4
  - Different species often have identical V4 tags
  - 10% genera in RDP14 have pair of identical V4s

16S: reality vs. clusters

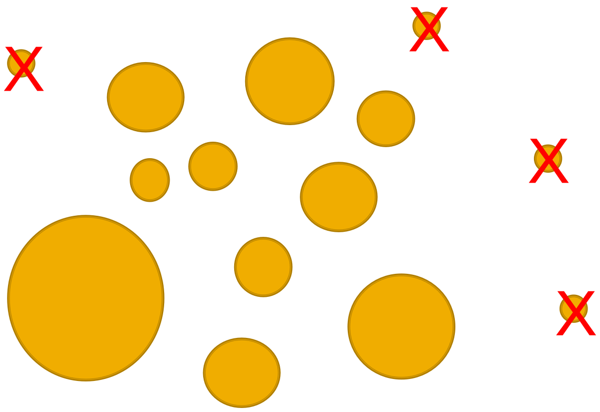


REALITY
Ecologically distinct strains,
size of blob = abundance

16S: reality vs. clusters

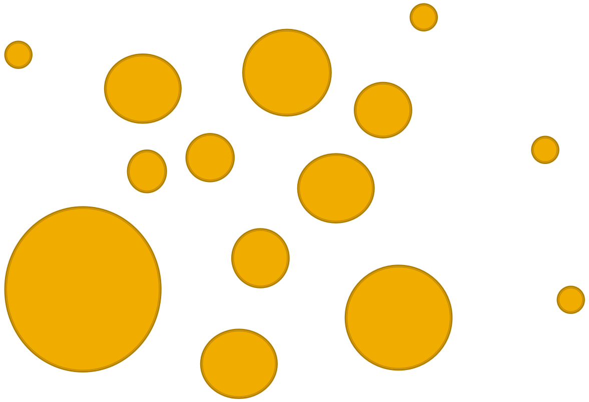


Reality

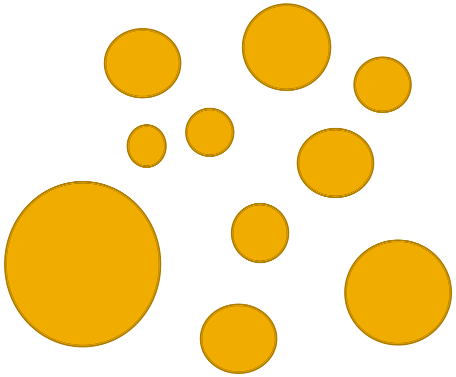


Rare strains not sampled

16S: reality vs. clusters

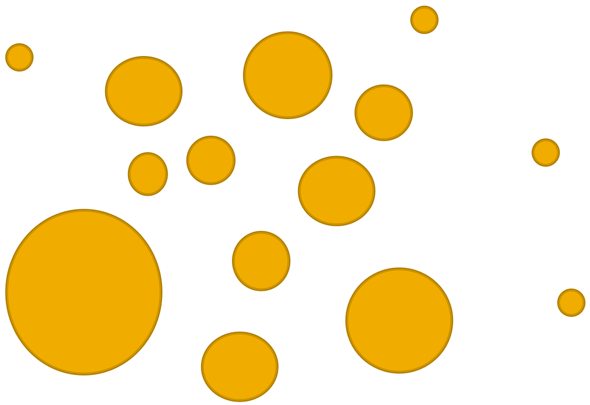


Reality

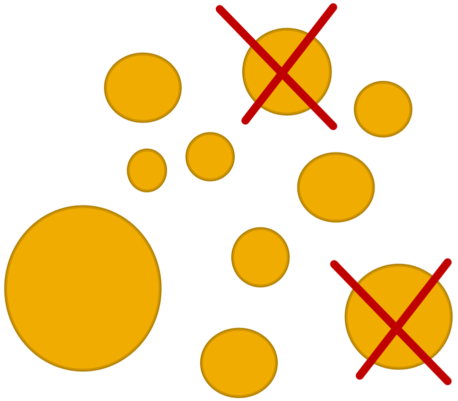


Rare strains not sampled

16S: reality vs. clusters

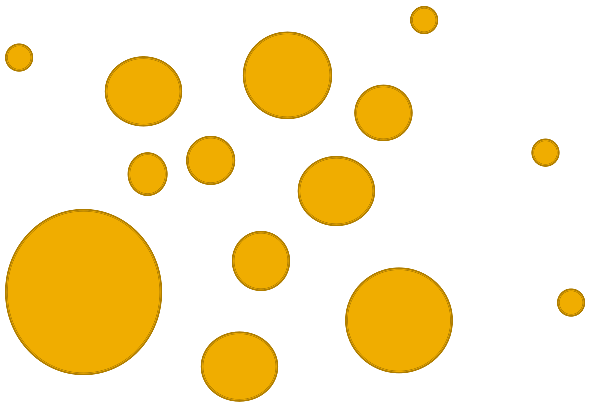


Reality

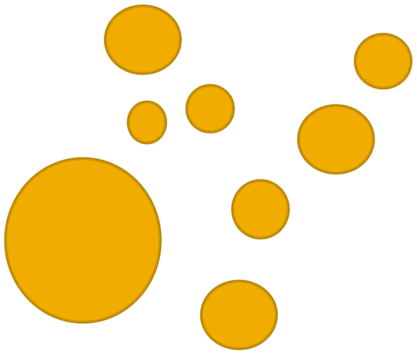


10-15% don't match
"universal primers"

16S: reality vs. clusters

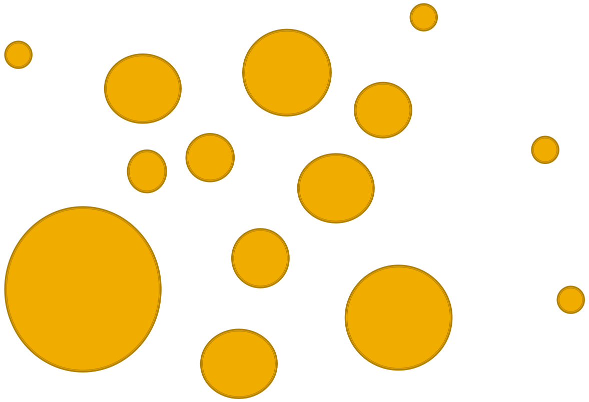


Reality

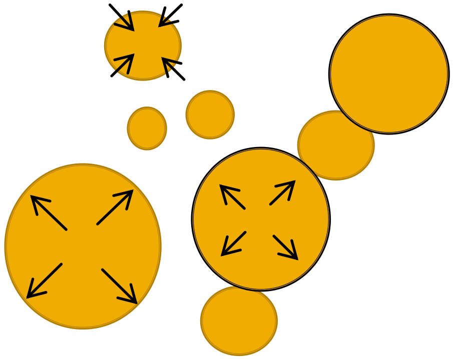


10-15% don't match
"universal primers"

16S: reality vs. clusters

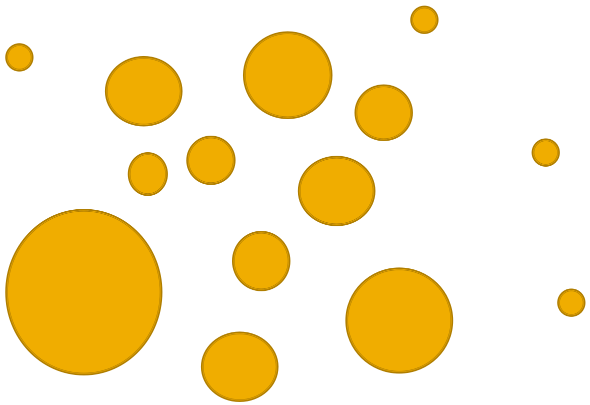


Reality

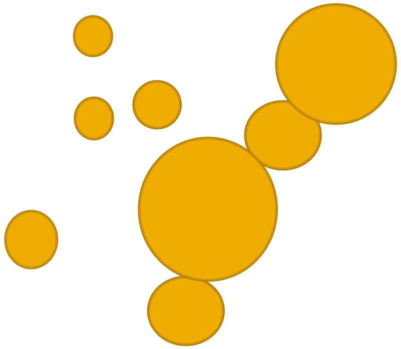


16S copy number varies
from 1 to 15 or so

16S: reality vs. clusters

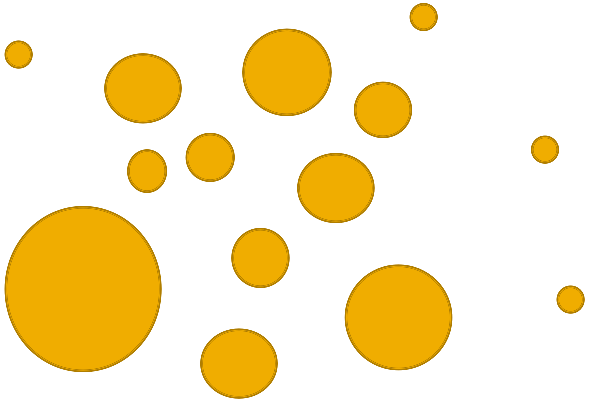


Reality

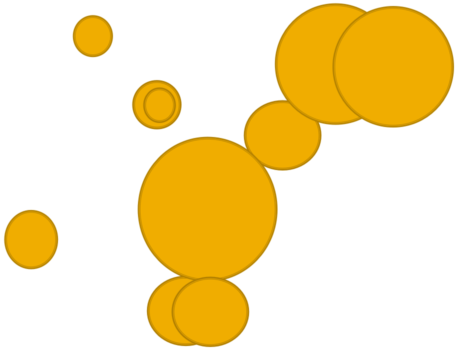


16S copy number varies
from 1 to 15 or so

16S: reality vs. clusters

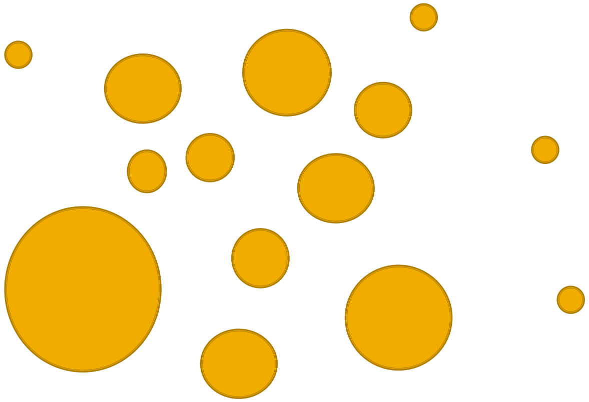


Reality

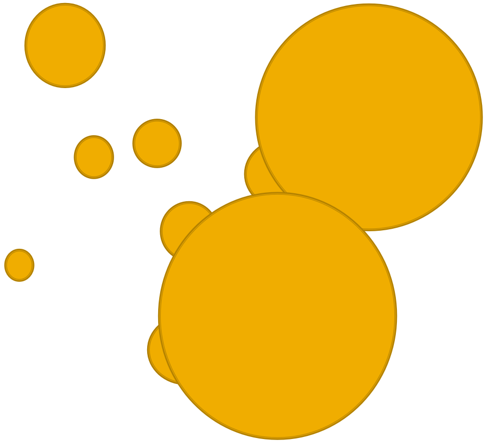


Clusters **split** (paralogs <97% similar) and **merge** (species >97% similar)

16S: reality vs. clusters

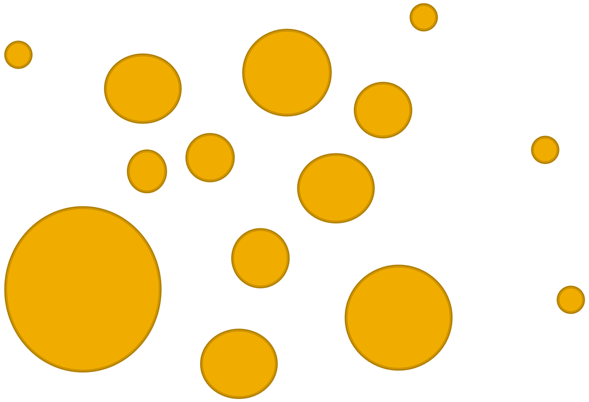


Reality

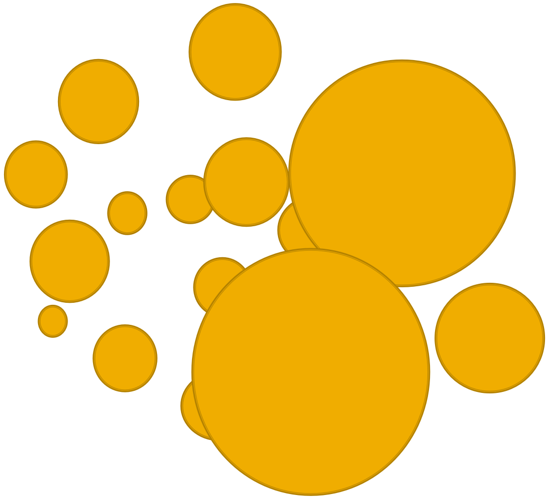


Amplification bias

16S: reality vs. clusters

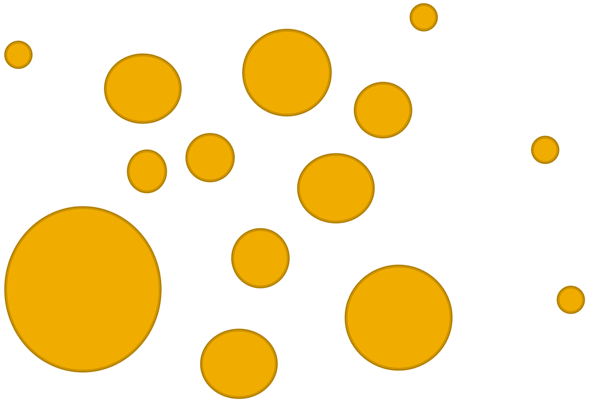


Reality

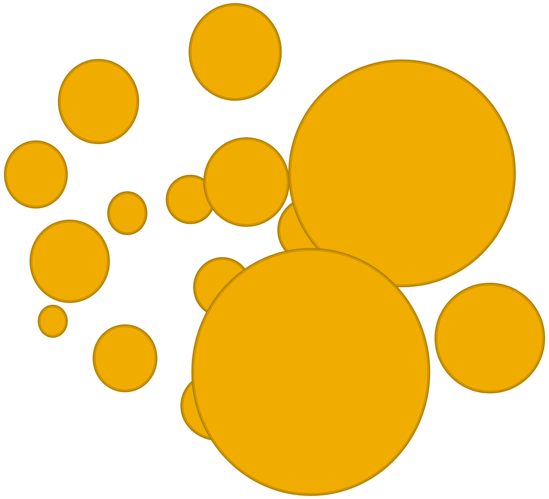


Polymerase errors,
chimeras, read errors,
contaminants

16S: reality vs. clusters



Reality



"OTUs"

Lump or split?

- One genome can contain many 16s genes
  - from one to 10+ typical
- Paralogs may be <100% identical
  - as low as 89%
- Any clustering %id will lump and split
  - Even in ideal scenario where no errors
- Clustering %id often motivated by "species"
  - I disagree

Lump or split?

- Lumping can obscure biological signals
- Splitting preserves information
  - e.g., better to distinguish strains than lump together
- Given all correct sequences
  - no reason to cluster
  - can estimate number of species from number of uniques
    - if needed, but usually not a very interesting or useful question
- Answer: split!
  - Resolve as many distinct genes as possible

Ideal analysis

- *Input*: Reads
- *Output*: Biological sequences
  - **All** biological sequences
  - **Nothing but** biological sequences

Achievable analysis

- Find subset of correct sequences >3%
  - Because ~3% is practical limit for detecting errors
- Sane motivation for 97% clustering
- Should resolve as much detail as possible
  - For any gene 16S, ITS, COI…
  - Regardless of typical intra-species variation
  - Individuals, strains, species, genera… are all **informative**
  - …and are **valid** OTUs!

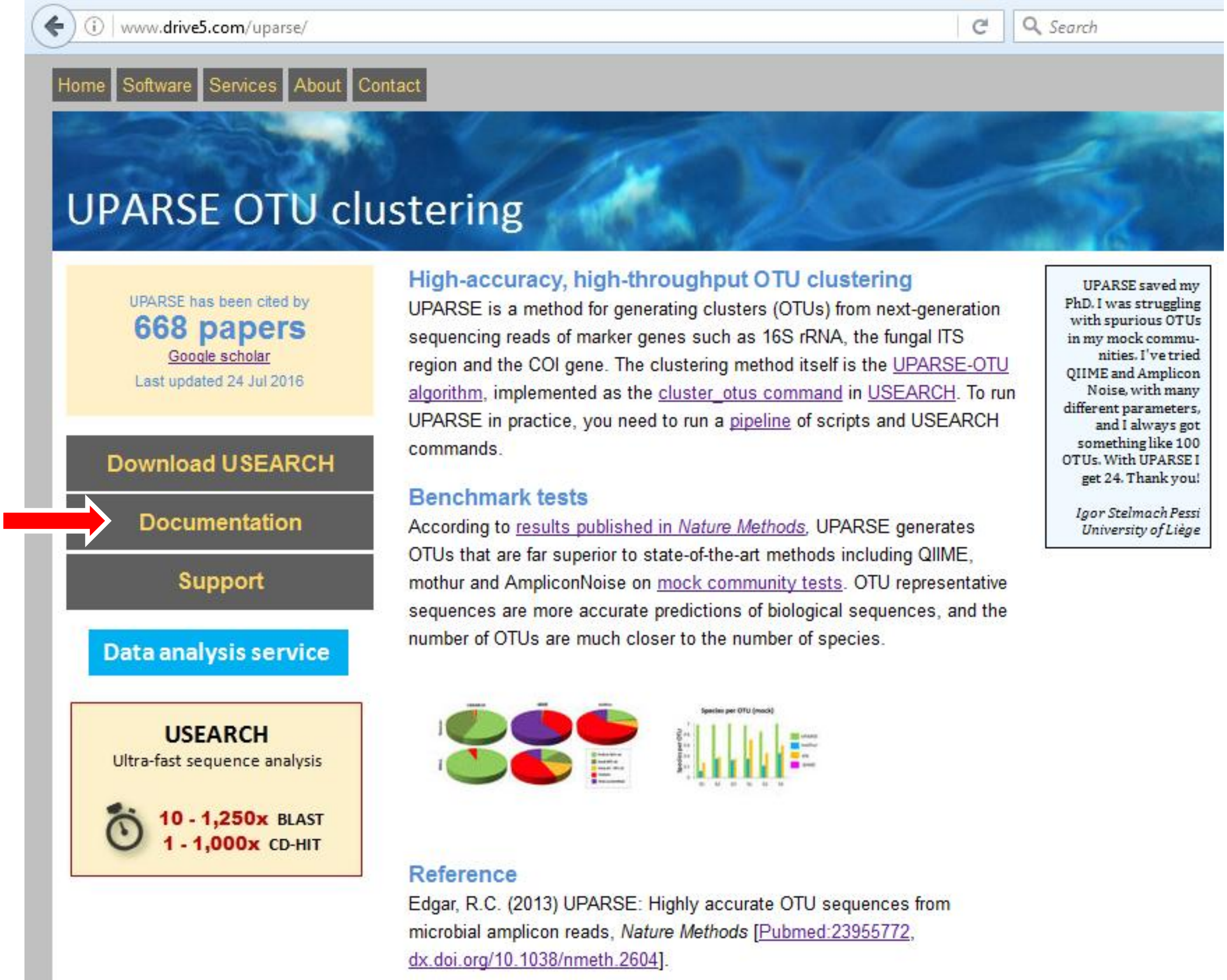
Future is (almost) here!

- Denoising can resolve sequences to ~1 diff
  - DADA2
  - UNOISE2 (coming soon in USEARCH v9)
- Other high-resolution methods
  - "oligotyping" (Eren *et al*. ISME 2015)
  - "sub-OTU resolution" (derep.) (Tikhonov *et al*. ISME 2014)
- Denoising close to ideal analysis
  - **all** biological sequences, and **nothing but**

Reads → OTUs with USEARCH

- Pre-process reads
  - Paired read assembly (with updated Q scores)
  - Expected error filtering (suggest $E < 1$, $E^* = 0$)
  - Discard singletons (optional, but highly recommended)
  - Dereplicate -- find uniques & abundances
  - Sort uniques by decreasing abundance
- Clustering: UPARSE-OTU algorithm
  - Edgar *Nat. Meth.* 2013
  - **cluster_otus** command

drive5.com/uparse



UPARSE OTU clustering

UPARSE has been cited by
668 papers
Google scholar
Last updated 24 Jul 2016

Download USEARCH

Documentation

Support

Data analysis service

USEARCH
Ultra-fast sequence analysis
10 - 1,250x BLAST
1 - 1,000x CD-HIT

High-accuracy, high-throughput OTU clustering

UPARSE is a method for generating clusters (OTUs) from next-generation sequencing reads of marker genes such as 16S rRNA, the fungal ITS region and the COI gene. The clustering method itself is the UPARSE-OTU algorithm, implemented as the cluster_otus command in USEARCH. To run UPARSE in practice, you need to run a pipeline of scripts and USEARCH commands.

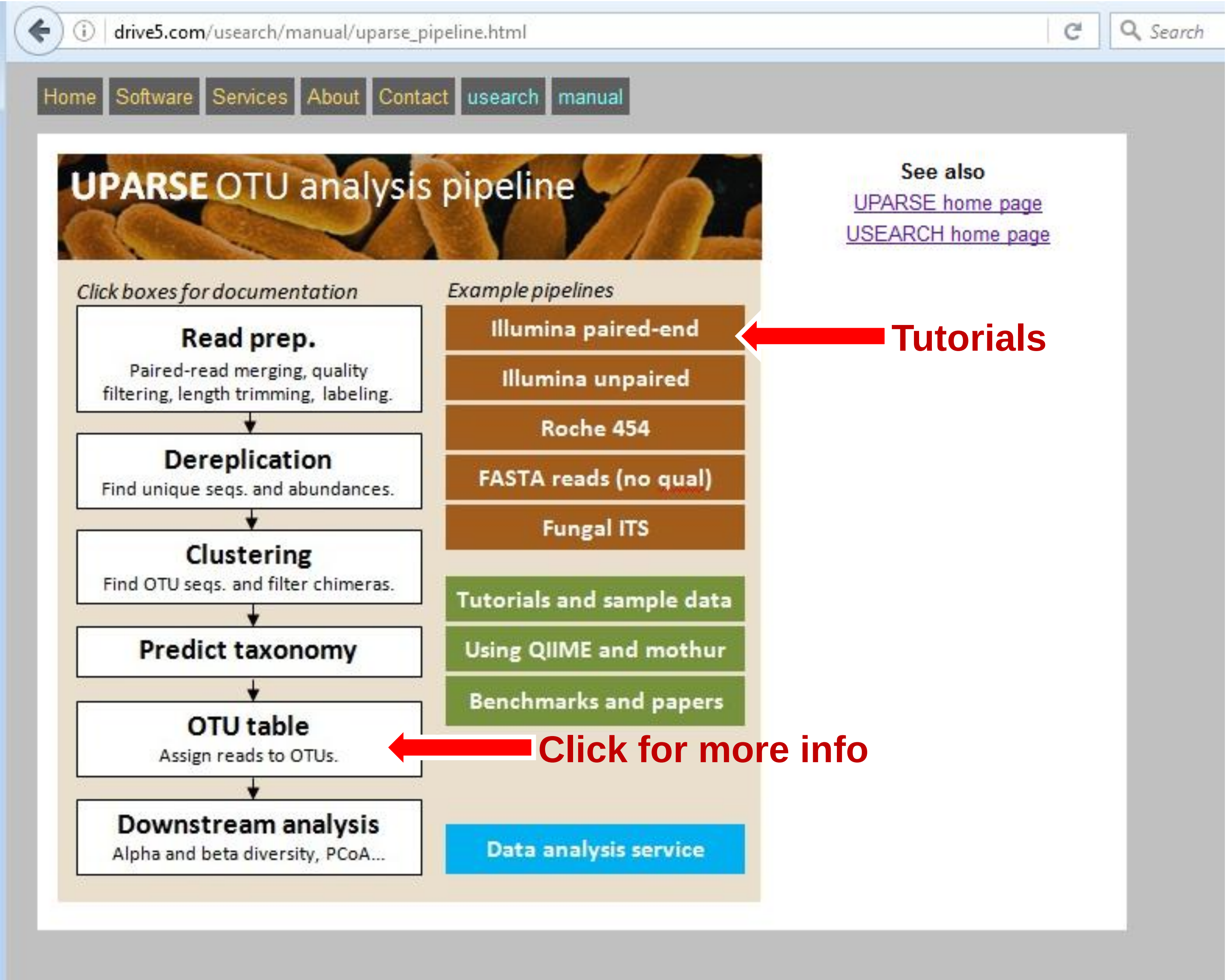
Benchmark tests

According to results published in *Nature Methods*, UPARSE generates OTUs that are far superior to state-of-the-art methods including QIIME, mothur and AmpliconNoise on mock community tests. OTU representative sequences are more accurate predictions of biological sequences, and the number of OTUs are much closer to the number of species.

Reference

Edgar, R.C. (2013) UPARSE: Highly accurate OTU sequences from microbial amplicon reads, *Nature Methods* [Pubmed:23955772, dx.doi.org/10.1038/nmeth.2604].

> UPARSE saved my PhD. I was struggling with spurious OTUs in my mock communities. I've tried QIIME and Amplicon Noise, with many different parameters, and I always got something like 100 OTUs. With UPARSE I get 24. Thank you!
>
> *Igor Stelmach Pessi*
> *University of Liège*



drive5.com/usearch/manual/uparse_pipeline.html

Search

Home | Software | Services | About | Contact | usearch | manual

UPARSE OTU analysis pipeline

See also

UPARSE home page

USEARCH home page

Click boxes for documentation

Read prep.
Paired-read merging, quality filtering, length trimming, labeling.

Dereplication
Find unique seqs. and abundances.

Clustering
Find OTU seqs. and filter chimeras.

Predict taxonomy

OTU table
Assign reads to OTUs.

Downstream analysis
Alpha and beta diversity, PCoA...

Example pipelines

- Illumina paired-end
- Illumina unpaired
- Roche 454
- FASTA reads (no qual)
- Fungal ITS

- Tutorials and sample data
- Using QIIME and mothur
- Benchmarks and papers

- Data analysis service

Tutorials

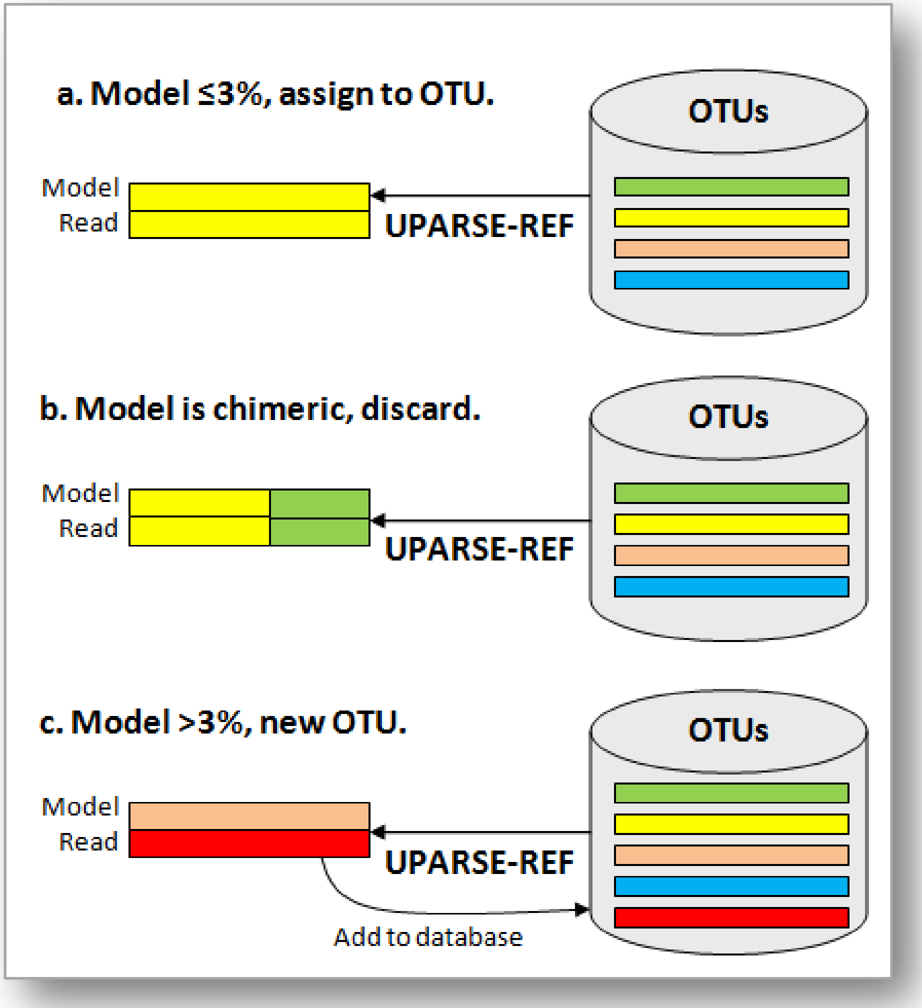
Click for more info

UPARSE-OTU

Process uniques in decreasing abundance order.

Compare each sequence with OTUs found so far.

Construct "model" by max. parsimony (fewest events)



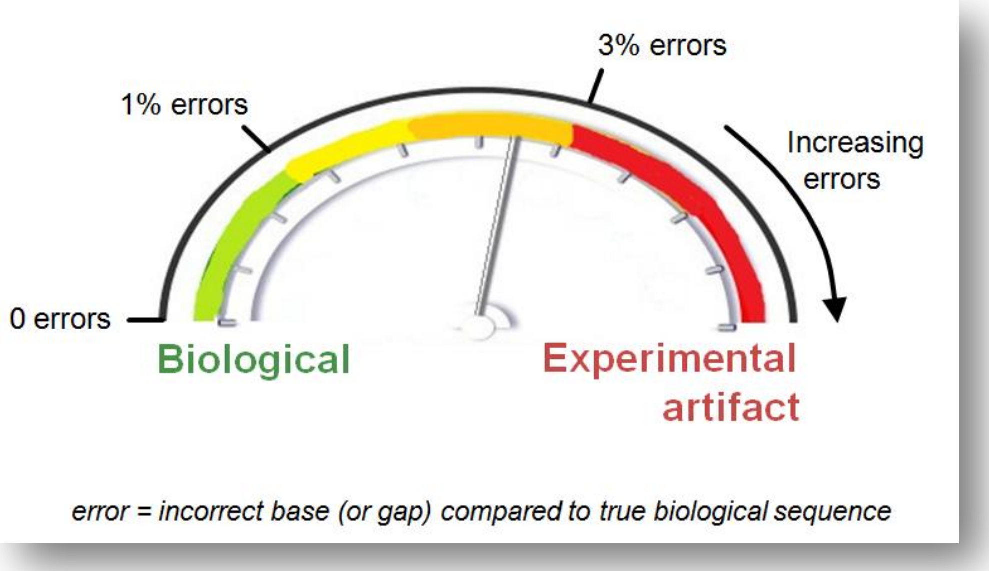
≤ 3% could be sequencing error, chimera or correct -- don't need to distinguish.

Chimeras >3% diverged can be found accurately

Otherwise, new OTU

Benchmark test

- OTUs should be biological sequences
- Other criteria are possible, perhaps…
  - but should be clearly defined!
  - Nr. OTUs = nr. species popular but <u>not valid</u>



error = incorrect base (or gap) compared to true biological sequence

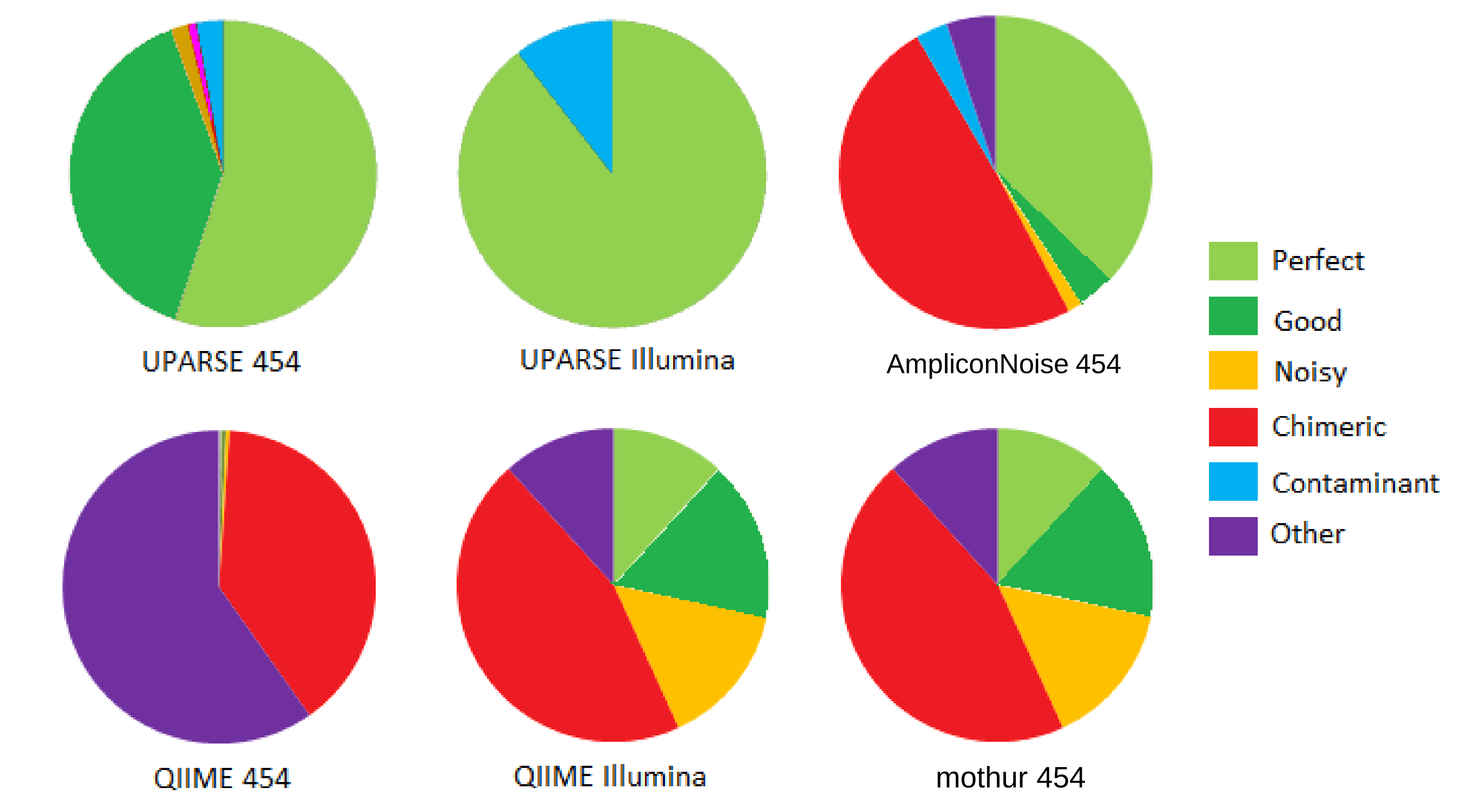
OTU classification

Color	Category	Description
🟩	Perfect	100% identical to biological sequence.
🟩	Good	≥99% identical to biological sequence.
🟨	Noisy	≥97% identical to biological sequence.
🟥	Chimera	"Bad" chimera >3% from biological sequence
⬜	Contaminant	Sequence found in large ref. db.
🟪	Other	None of the above. Could be a novel contaminant, or -- much more likely -- have >3% errors.

16S mock community data

- HMP mock communities
- 21 species
- Even and Staggered mixes
- 454 Titanium and Illumina MiSeq 2x250
- Community & ref db. by Haas *et al.*
  - Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome res.* (2011)

Results on HMP mock datasets



Edgar *Nat. Meth.* (2013)

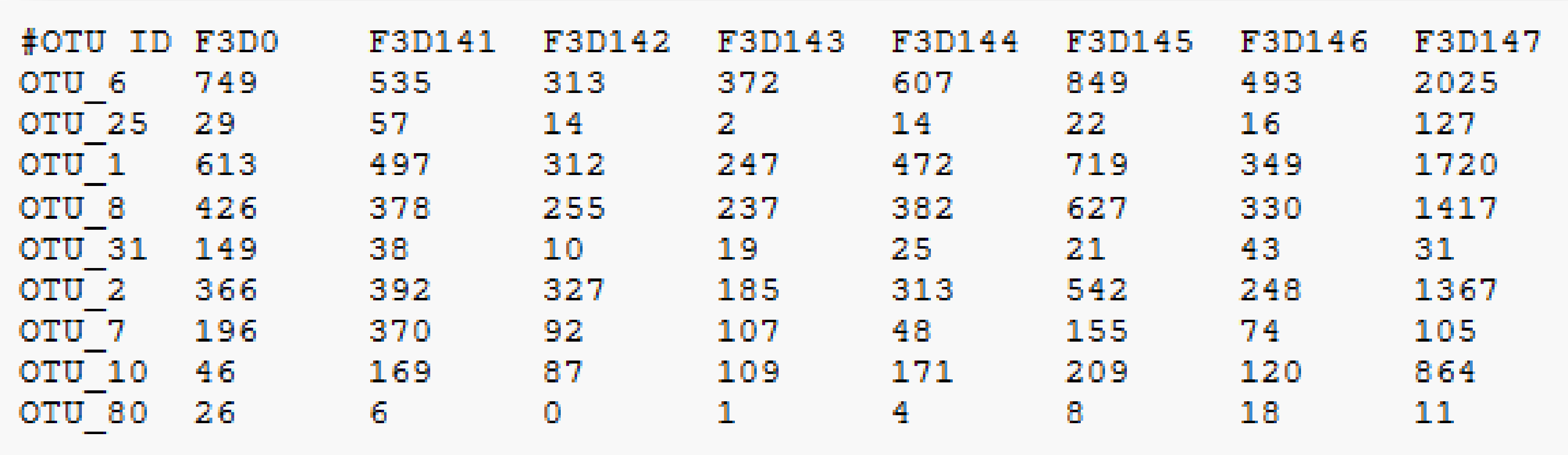
OTU table

- Matrix of OTUs vs. samples
- Value is nr. of reads

	Sample1	Sample2	Sample3	...
Otu1	1,023	455	992	...
Otu2	324	622	12	...
Otu3	871	29	321	...
....	...	....	...	...

QIIME "classic" tabbed text

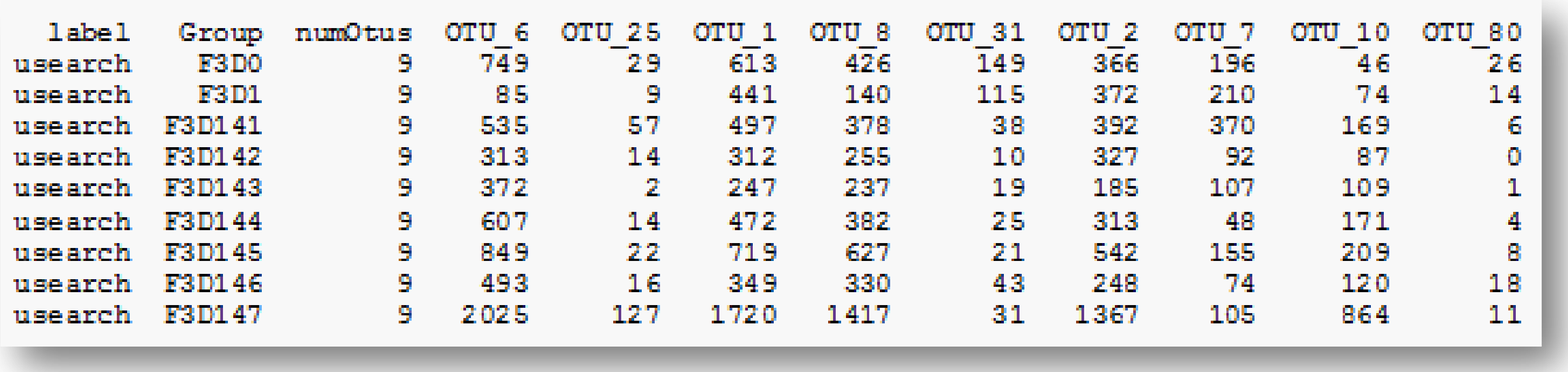
Tab-separated text
Rows are OTUs, columns are samples
Simple, intuitive and convenient
Use cut, grep etc., load into spreadsheet…



```
#OTU ID F3D0    F3D141  F3D142  F3D143  F3D144  F3D145  F3D146  F3D147
OTU_6   749     535     313     372     607     849     493     2025
OTU_25  29      57      14      2       14      22      16      127
OTU_1   613     497     312     247     472     719     349     1720
OTU_8   426     378     255     237     382     627     330     1417
OTU_31  149     38      10      19      25      21      43      31
OTU_2   366     392     327     185     313     542     248     1367
OTU_7   196     370     92      107     48      155     74      105
OTU_10  46      169     87      109     171     209     120     864
OTU_80  26      6       0       1       4       8       18      11
```

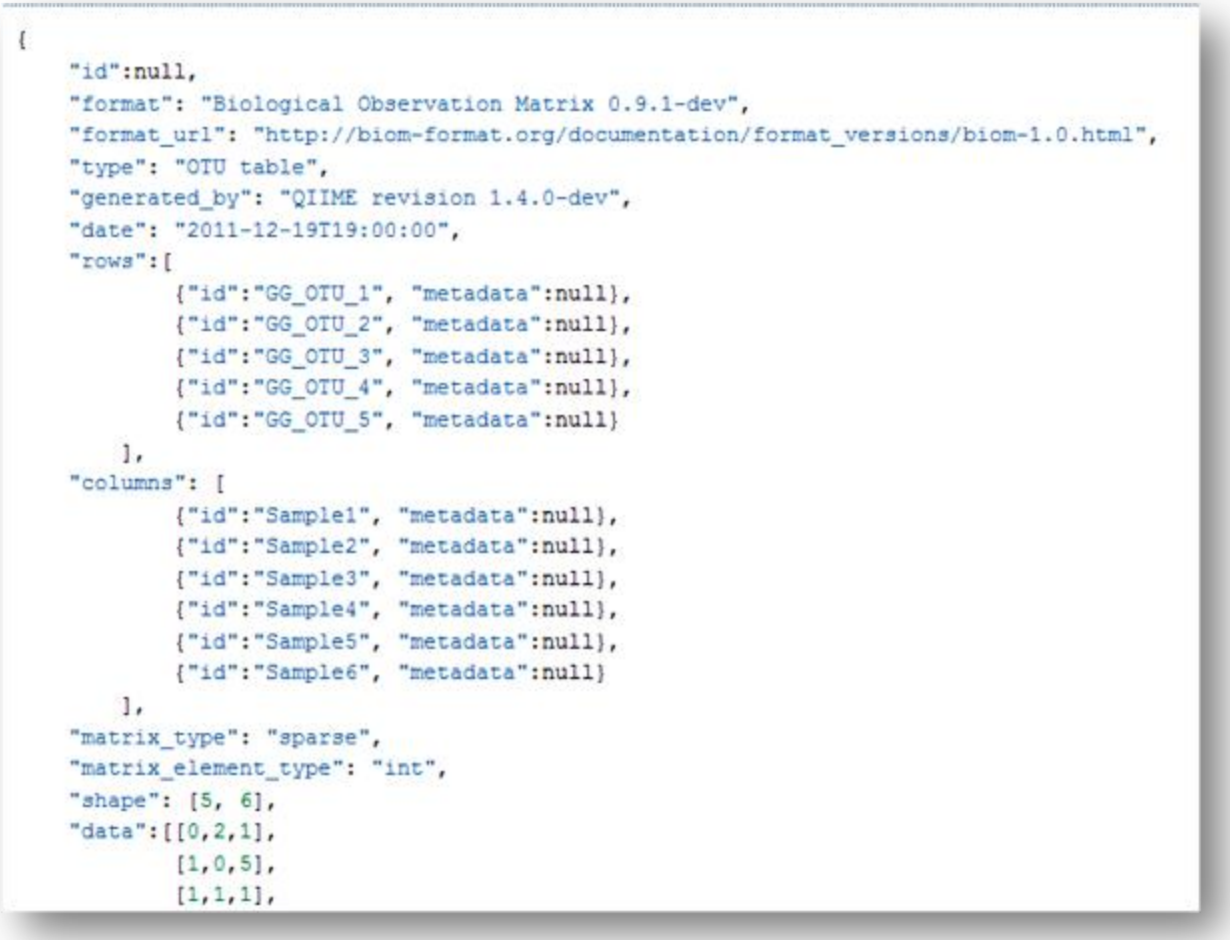
mothur "shared" file

Tab-separated text
Rows are samples ("groups"), columns are OTUs



```
   label     Group  numOtus  OTU_6  OTU_25  OTU_1  OTU_8  OTU_31  OTU_2  OTU_7  OTU_10  OTU_80
usearch       F3D0        9    749      29    613    426     149    366    196      46      26
usearch       F3D1        9     85       9    441    140     115    372    210      74      14
usearch     F3D141        9    535      57    497    378      38    392    370     169       6
usearch     F3D142        9    313      14    312    255      10    327     92      87       0
usearch     F3D143        9    372       2    247    237      19    185    107     109       1
usearch     F3D144        9    607      14    472    382      25    313     48     171       4
usearch     F3D145        9    849      22    719    627      21    542    155     209       8
usearch     F3D146        9    493      16    349    330      43    248     74     120      18
usearch     F3D147        9   2025     127   1720   1417      31   1367    105     864      11
```

BIOM v1 (JSON)



```
{
    "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":null},
            {"id":"GG_OTU_2", "metadata":null},
            {"id":"GG_OTU_3", "metadata":null},
            {"id":"GG_OTU_4", "metadata":null},
            {"id":"GG_OTU_5", "metadata":null}
        ],
    "columns": [
            {"id":"Sample1", "metadata":null},
            {"id":"Sample2", "metadata":null},
            {"id":"Sample3", "metadata":null},
            {"id":"Sample4", "metadata":null},
            {"id":"Sample5", "metadata":null},
            {"id":"Sample6", "metadata":null}
        ],
    "matrix_type": "sparse",
    "matrix_element_type": "int",
    "shape": [5, 6],
    "data":[[0,2,1],
            [1,0,5],
            [1,1,1],
```

Text, but complex

Hard to work with in scripts

Can't use cut, grep, awk…

BIOM v2 (HDF5)

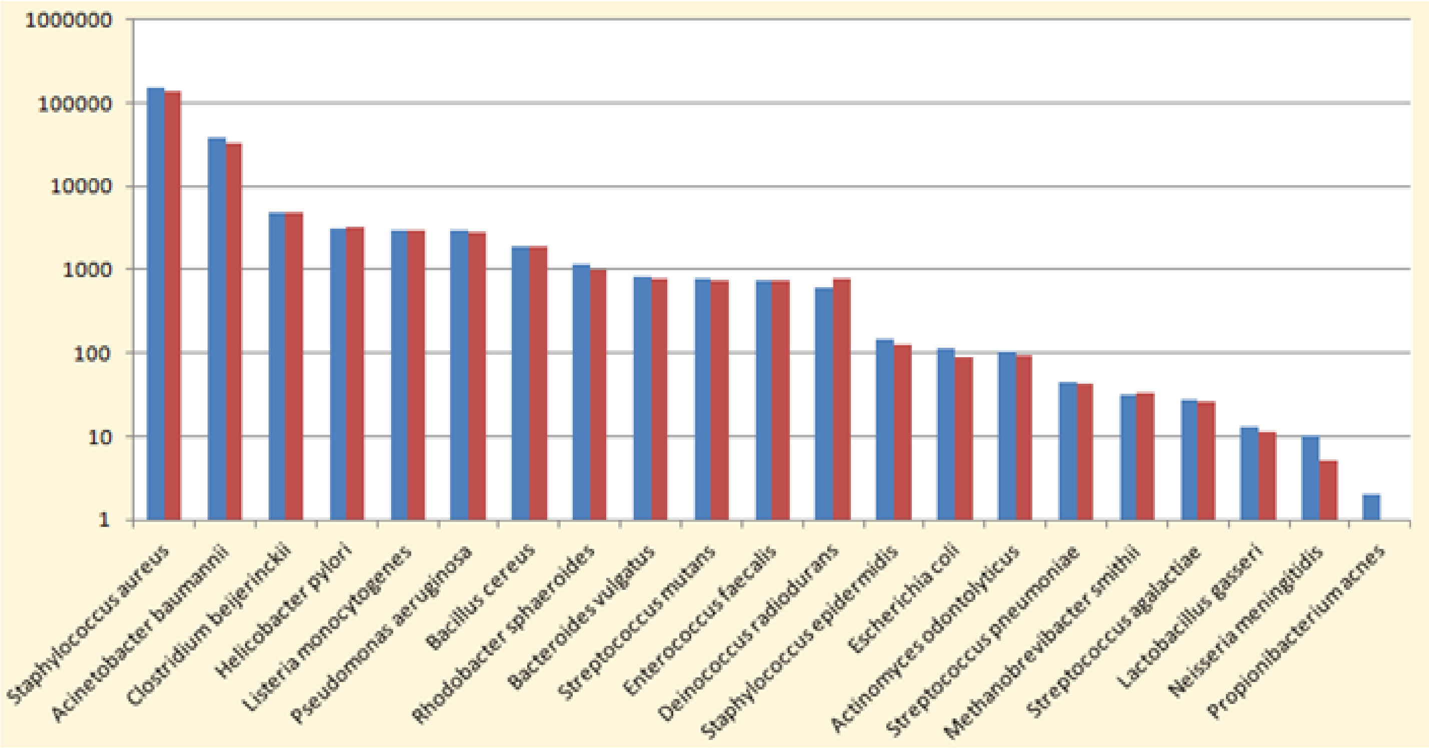
- Totally unrelated to BIOM v1 format
- Not text, opaque binary format
- Motivation: huge OTU tables
  - e.g. Earth Microbiome Project

OTU table values

- Number of reads
  - "Raw"
  - Sub-sampled
    - e.g. to same number reads / sample
  - Rarefied
  - Normalized
- Frequencies
- No standards
  - Minimal software compatibility

Read abundance vs. cells

- Nr reads does **not** predict cell abundance



Read abundance for Even(!) mock community (Bokulich *et al*. 2013)

Metadata

- Taxonomy predictions
- Sample information
  - Healthy / diseased
  - Time / date, location…
  - Temperature, salinity, phase of moon…
- No standards, no software compatibility

Make OTU table with USEARCH

- Clustering gives one sequence for each OTU
  - "Representative sequence", "centroid"
- Align unfiltered reads to OTU sequences
  - database search (usearch_global command)
  - if ≥97%, assign to closest OTU
  - recovers most low-quality & singleton reads
  - almost all unmapped reads have many errors / chimeras
- Outputs one or more formats
  - QIIME classic, mothur shared and / or BIOM v1