TEL AVIV UNIVERSITY

UPARSE: highly accurate OTU sequences from microbial amplicon reads

Robert C Edgar

Seminar in Computational Methods in Metagenomics and Microbiome Research
Spring Term 2019

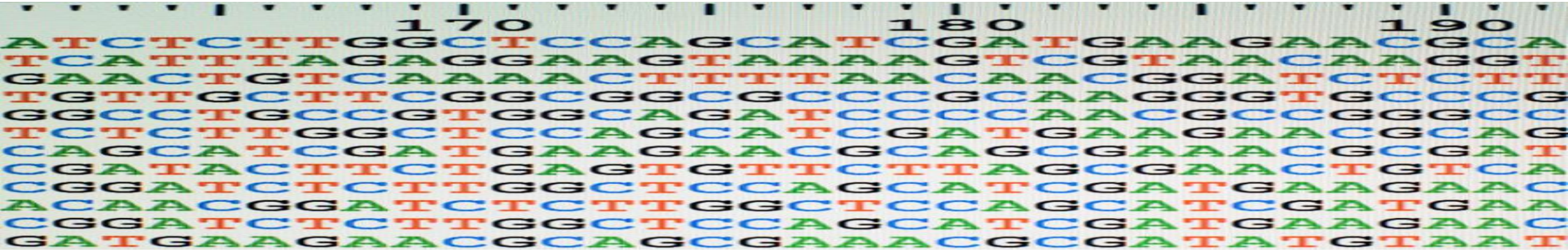
Name: Gal Cohen

E-mail: galcohen@mail.tau.ac.il

בית הספר למדעי המחשב על שם בלבטניק
The Blavatnik School of Computer Science

Next Generation Sequencing (NGS)

- **A catch-all term used to describe a number of different modern sequencing technologies.**
- **Made DNA and RNA sequencing much faster and cheaper then ever before.**
- **Revolutionized the study of genomics and molecular biology!**



Next Generation Sequencing (NGS)

Some of the different technologies:

- **Illumina (Solexa) sequencing**
- **Roche 454 sequencing**
- **Ion torrent: Proton / PGM sequencing**
- **SOLiD sequencing**

Each one has its pros and cons!

What do we do with those letters?

Our goal is to characterize microbial community structure and function.

How do we do that?

- **Organize the sequences into groups.**
- **Call those groups OTUs (operational taxonomic units) in order to confuse the common CS student**
- **OTUs are intended to correspond to taxonomic clades or monophyletic groups.**

Sounds easy?

- **The data is full of artifacts!**
- **To make things worse – there are many different types of artifacts.**
- **Different techniques to deal with each them:**
    1. **Quality filtering of reads**
    2. **Denoising of flowgrams**
    3. **Chimera filtering**
    4. **clustering**

The problem is…

- **Just like research – no matter how hard you try, those problems won't leave your dataset.**
- **Solution A:**
  1. **Get angry**
  2. **Blame everything you can think of (but yourself)**
  3. **Leave the field**

TEL AVIV UNIVERSITY

Or

בית הספר למדעי המחשב על שם בלבטניק
The Blavatnik School of Computer Science

Use the UPARSE pipeline!

- **Constructing OTUs *de novo* from next-generation reads .**
- **Achieves high accuracy in biological sequence recovery.**
- **Improves richness estimates on mock communities.**
- **Highly robust to variations in the input data.**
- **Low computational resource requirements.**
- **Published by only one author – respect!**

Our Rivals

There were several different pipelines at the time the paper was published

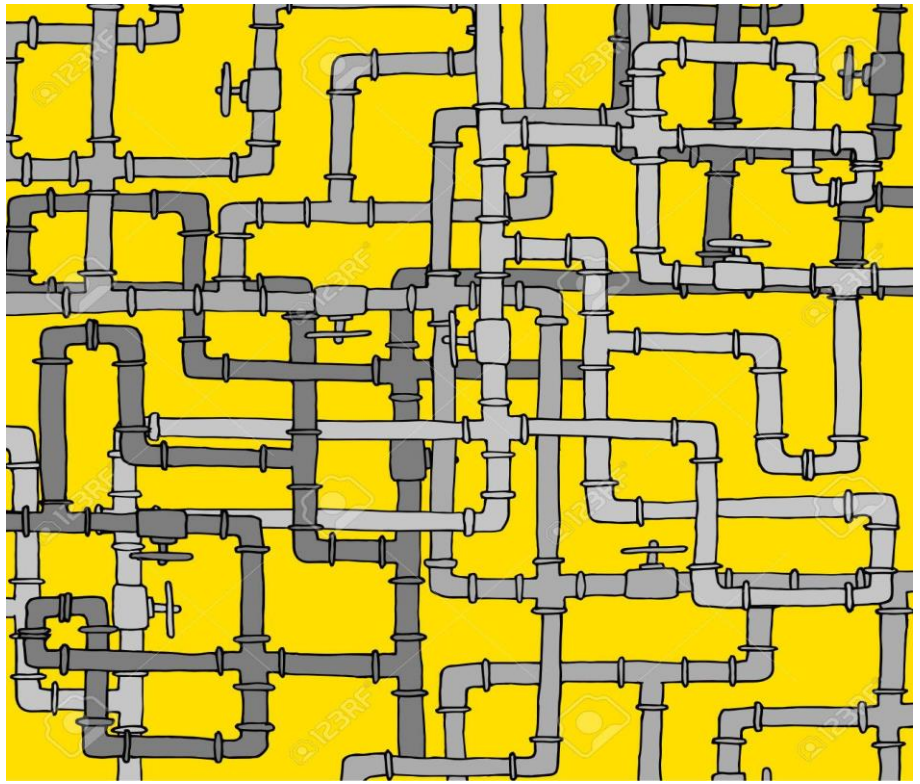
- **QIIME**
- **MOTHUR**
- **AmpliconNoise**

Each pipeline has its own pros and cons and they are all still widely used today.

UPARSE Workflow

Our pipeline include several steps:

1. **Merging of paired reads**
2. **Read quality filtering**
3. **Length trimming**
4. **Dereplication**
5. **Discarding singletons**
6. **OTU clustering**



Step 1: Merging of Paired Reads

1. **Ask for help from the proffesor**

Step 2: Read Quality Filtering

1. **Set your minimum quality score (Q_{min}=16 Default) at the beginning**
2. **The quality score used called "Phred Quality Score"**
3. **Impose minimal quality score for all bases in the read.**

The last step is done by truncating at the first read base with $Q < Q_{min}$

This is done on reads in FASTQ format

FASTQ format - stores both the sequence and its corresponding quality scores

Phred Quality Score

- **A quality score of a base, also known as Q score.**
- **An integer value representing the estimated probability of an error, i.e. that the base is incorrect.**
- **If P is the error probability then:**

$$Q = -10 * \log_{10}(P)$$

- **For example, if Phred assigns a Q score of 30 (Q30) to a base, this is equivalent to the probability of an incorrect base call 1 in 1000 times**

Step 3: Length Trimming

- **Step 2 produced reads with variable lengths – might cause problems.**
- **For example if we have one read which is an exact match to the prefix of a longer read.**
- **Simple solution – truncate reads at fixed length (L)**
- **Discard reads that were shorter**

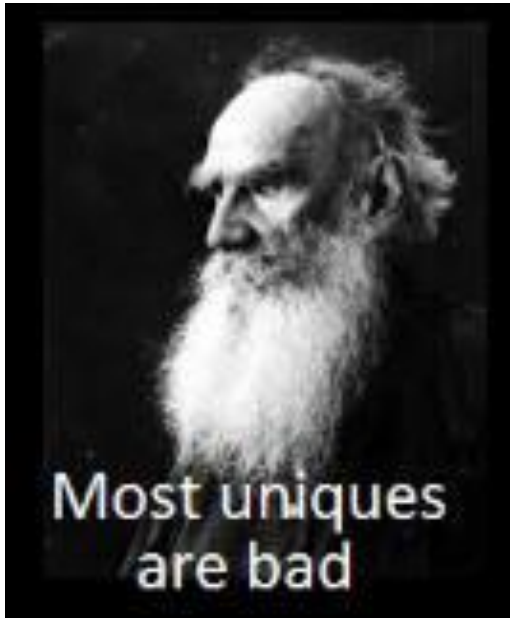
Step 4: Merging of Identical Reads (dereplication)

Dereplication is the removal of duplicated sequences

- **Identify the set of unique read sequences**
- **Record the number of occurrences for each sequence.**
- **As all reads has the same length this is very trivial.**

Step 5: Discarding Singletons

- A singleton is a read with a sequence that is present exactly once
- Expected to have at least one error
- If errors are independent and randomly distributed they are not likely to be correct
- Discard them as they will probably induce spurious OTUs
- Singletons can be retained for later clustering with new reads



Step 6: UPARSE-OUT Clustering Method

- **A new greedy algorithm for OTU clustering was introduced**
- **It uses a single representative sequence to define each cluster (OTU)**
- **Initial steps:**
  1. **Initialize an empty database of OTU sequences**
  2. **Consider unique read sequences in order of decreasing abundance**
  3. **Move to slide number 19**

UPARSE-OTU Algorithm (cont.)

4. **If the read matches an existing OTU within the identify threshold (default 97%): update abundance**

5. **Otherwise: construct a model of the read with UPARSE-REF algorithm with the current database as reference**

6. **If chimeric: discard the read**

7. **Else: add the read to the database as a new OTU representative**

UPARSE-REF Algorithm

We have an OTU database and a read that does not “fit” to any representative in it.

There are two options:

1. **It was forged by several OTUs (chimeric)**
2. **It is a brand new OTU representetive!**

We should try to figure out what is the shortest way for it occur from our database via amplifications.

The above mentioned model is the most parsimonius explanation of the read from the database

$$\Phi(S,M) = d(S,M) + (m-1)$$

UPARSE-REF Algorithm

The calculation is done dynamically –

$$\Phi_{j+1,k} = min_{k'=1\ldots N}\{\Phi_{jk'} + d_{j+1,k} + \mathbf{1}(k' \neq k)\}$$

If the model was not chimeric – the read most be a new OTU

Conclusion

- **The UPARSE pipeline produce a much more reasonable number of OTUs compared to the other platforms.**
- **Substantial improvement in OTU construction.**
- **Requires less computational resources.**