

BIOINFORMATICS

BIO9905MERG1 - SPRING 2019

LUIS N. MORGADO

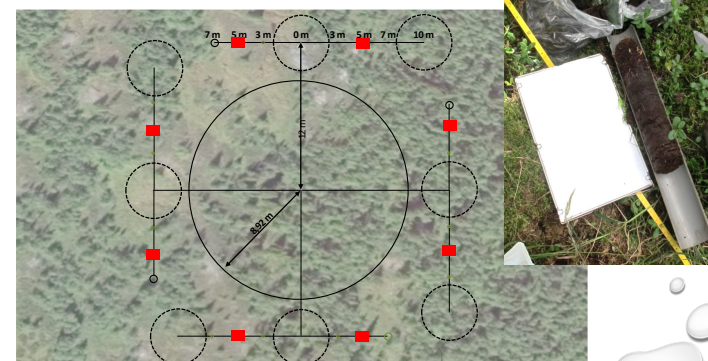
OVERVIEW

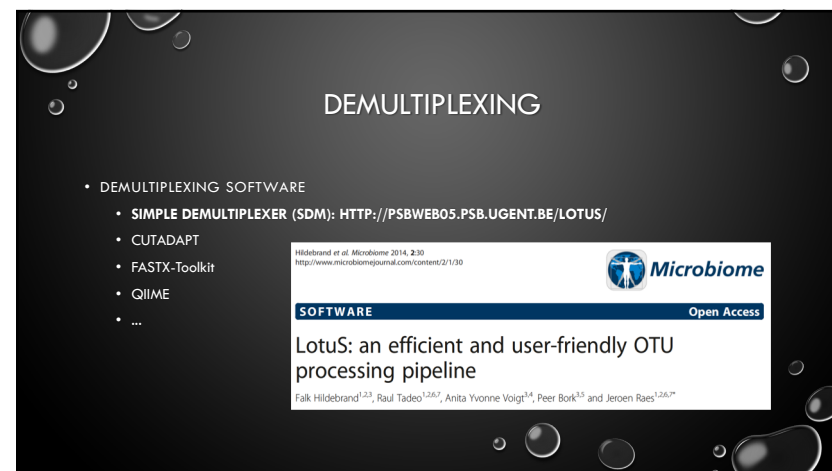
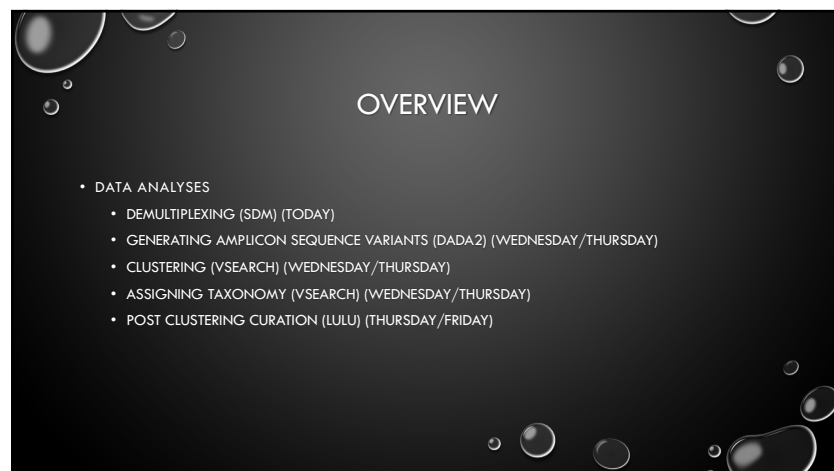
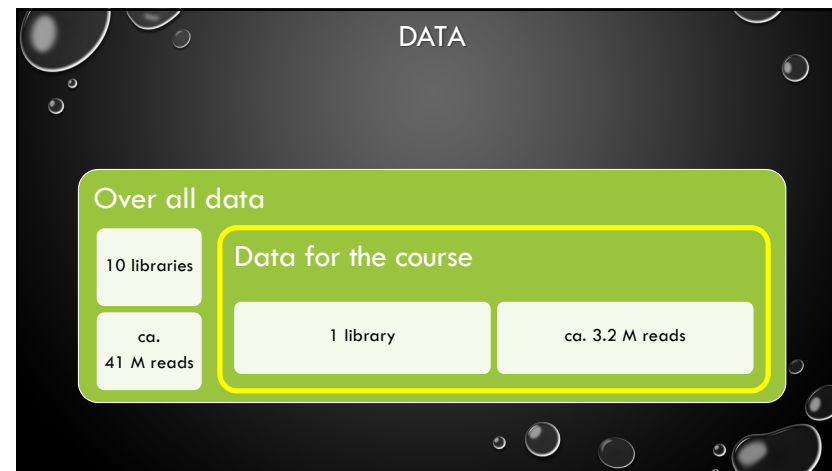
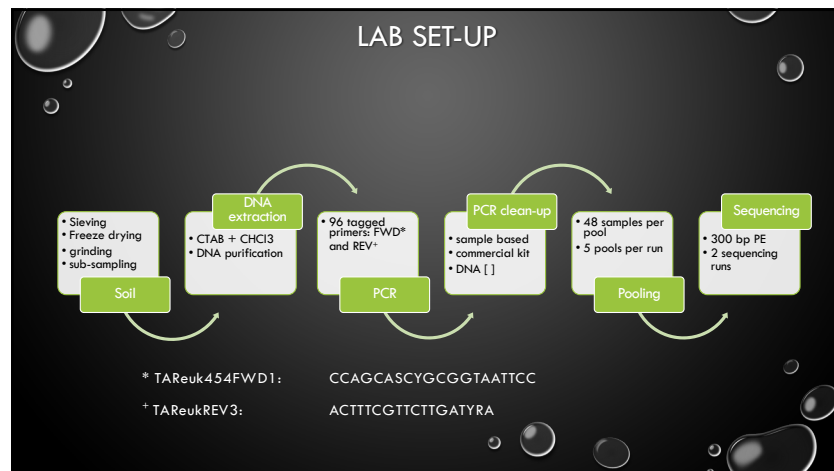
- INTRODUCTION TO THE DATASET (TODAY)
- DATA ANALYSES
 - DEMULTIPLEXING (SDM) (TODAY)
 - GENERATING AMPLICON SEQUENCE VARIANTS (DADA2) (WEDNESDAY/THURSDAY)
 - CLUSTERING (VSEARCH) (WEDNESDAY/THURSDAY)
 - ASSIGNING TAXONOMY (VSEARCH) (WEDNESDAY/THURSDAY)
 - POST CLUSTERING CURATION (LULU) (THURSDAY/FRIDAY)

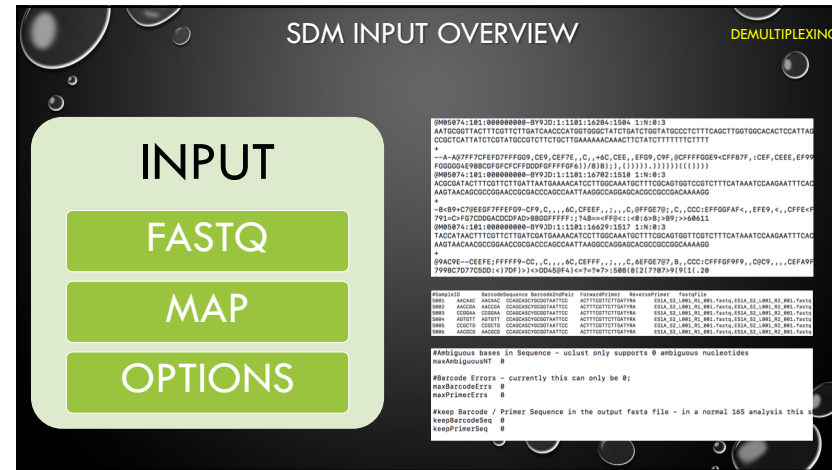
EXPERIMENTAL DESIGN

- Boreal forest age gradient from 2 to 245 years
- Sample = organic soil layer
- Nested sampling design (ca. 500 samples)
 - 62 plots
 - 8 subplots

EXPERIMENTAL DESIGN







GENERAL COMMAND TO RUN SDM FOR DEMULTIPLEXING R1 AND R2 INDEPENDENTLY

FOR A SINGLE LIBRARY

- SDM -I FASTQ_XX_R1.FASTQ, XX_R2.FASTQ -MAP MAP_XXX.TXT -OPTIONS XXX.TXT -PAIRED 2 -SAVEDEMULTIPLEX 1 -SYNCREADPAIRS T -O_DEMULTIPLEX FOLDER_FOR_OUTPUT -LOG TEXT_FILE_FOR_THE_LOG

FOR MULTIPLE LIBRARIES

- SDM -I_PATH PATH_TO_FOLDER_WITH_FASTQS -MAP MAP_YYY.TXT -OPTIONS XXX.TXT -PAIRED 2 -SAVEDEMULTIPLEX 1 -SYNCREADPAIRS F -O_DEMULTIPLEX FOLDER_FOR_OUTPUT -O_FASTQ OUTPUT1.FASTQ,OUTPUT2.FASTQ -LOG TEXT_FILE_FOR_THE_LOG

- # RUN SDM
- DEMULTIPLEXING
- Log in to abel...
 - Make a folder to host the data and the scripts
 - `Mkdir raw_data`
 - `###` Navigate to this folder
 - `Cd raw_data`
 - `###` Get the path to this folder
 - `Pwd`
 - `####` Copy the fastq files and script for demultiplexing into that folder
 - `cp -r /work/tmp/BIO9905MERG1/luisnm/data/*.`
 - `cp -r /work/tmp/BIO9905MERG1/luisnm/demult_scripts/*.`

RUN SDM

DEMULTIPLEXING

- Load the software you need:
 - Start a qlogin session first
 - Type: `source /cluster/software/BIO9905/set.Source`
 - Test if SDM is working type: `SDM`
 - Look at the options available `SDM --help_options`
- Run the demultiplexing with the option that I prepared


```
sdm -i fastq ES1A_S2_L001_R1_001.Fastq.es1a_s2_L001_r2_001.Fastq -map map_es1a_abel.Txt -
options options_no_filters.txt -paired 2 -savedemultiplex 1 -syncreadpairs T -o_demultiplex demult_samp -
log demult_log.Txt
```

Folder missing? Create a folder for the output file before the previous command

```
mkdir demult_samp
```

RUN SDM

DEMULTIPLEXING

- Look at the output:
 - How many reads are kept per file
 - And per sample
- Is it what we expected?
- If you want you can play with the options and check the output (create a separate folder for this)



Fast and accurate sample inference from amplicon data with single-nucleotide resolution

WHAT IS DADA2?

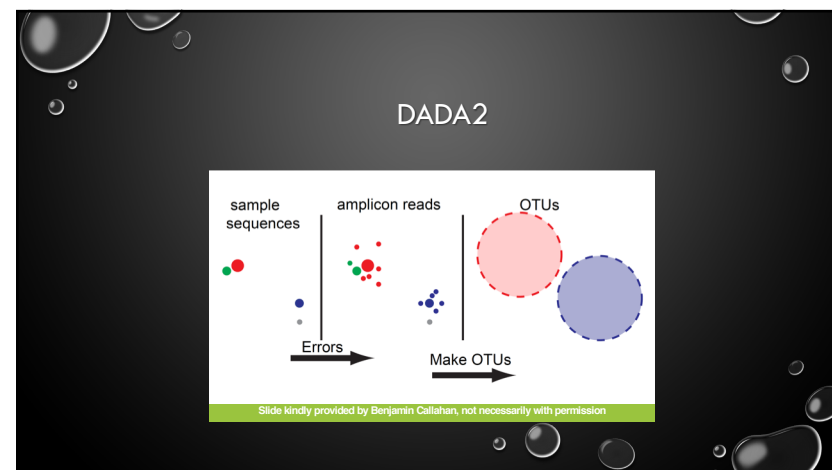
It's an open-source software with an algorithm for modeling and correcting illumina-sequenced amplicon errors

<https://benjineb.github.io/dada2/index.html>

DADA2: High-resolution sample inference from illumina amplicon data

Benjamin J Callahan¹, Paul J McMurdus¹, Michael Holmes¹, Andrew W Hart¹, Aaron D A Johnson¹, Brian P Huttenhower¹

NATURE METHODS | VOL.13 NO.7 | JULY 2016 | 581



DADA2

3. Abundance p-value

- This step relies and quantifies the notion that sequence i is too abundant to be explained by errors in amplicon sequencing taking into account the error rate.

The number of amplicon reads with sequence i that will be produced from sample sequence j is Poisson distributed with expectation equal to an error rate λ_{ji} multiplied by the expected reads of sample sequence j

$$p_A(j \rightarrow i) = \frac{1}{1 - \rho_{\text{pois}}(n_j \lambda_{ji}, 0)} \sum_{a=a_i}^{\infty} \rho_{\text{pois}}(n_j \lambda_{ji}, a)$$

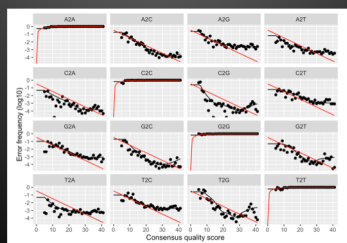
4. Partition algorithm

- Grouping unique sequences -> Dereplication with abundance + consensus profile
- All unique sequences into 1 partition – most abundant as the center
- All unique sequences are then compared to the center of their partition
 - error rates are calculated and stored and
 - the abundance p-value is calculated for each unique sequence
 - If the smallest p-value is below **OMEGA_A** a new partition is formed with the unique sequence with the smallest p-value at the center.
 - All unique sequences are compared to the center of the new partition.
 - After a new partition is formed every unique sequence is allowed to join the partition most likely to have produced it (based on the highest expected number of that unique sequence)
- Then there is an iteration in this procedure until all abundance p-values are greater than **OMEGA_A** (i.e., all unique sequences are consistent with being produced by amplicon sequencing the center of their partition)
- The composition of the sample is then the central sequences and the corresponding total abundance.

DADA2

5. Error model parametrization

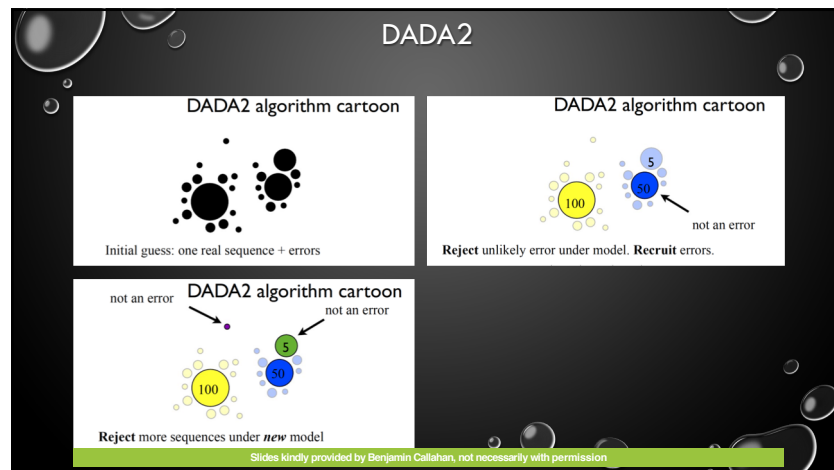
- DADA2 depends on a parameterized error model, the 16 x 41 transition probabilities, for example, $p(A \rightarrow C, 35)$, but if parameters are not known a priori then DADA2 can estimate them from the data.



DADA2

6. Estimation until consistency

- Alternates sample inference (conditional on the parameters of the error model) with parameter estimation (conditional on the inferred sample composition) until convergence, at which point jointly consistent estimates of the error parameters and sample composition are reported.



DADA2

- How is this different from other bioinformatic pipelines?
 - Most pipelines do not correct for sequencing errors

	DADA2	OTUs
Abundance	yes	Ranks only?
Sequence differences	yes	Ranks only?
Quality	yes	no
Error model	yes	no

DADA2

- What is the output?
 - DADA2 outputs an amplicon sequence variants table (analog to an OTU table)
 - But contains amplicon sequence variants, which reflect real biological variation of the data

ATTAACGAGATTATAACGAGTACGAATA...

is a **consistent label**

OTU85 is **not**

Consistent labels
allow samples to be
independently processed

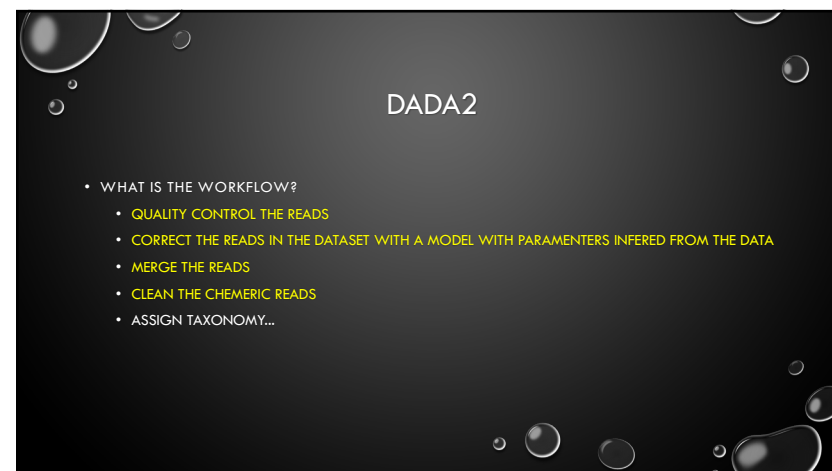
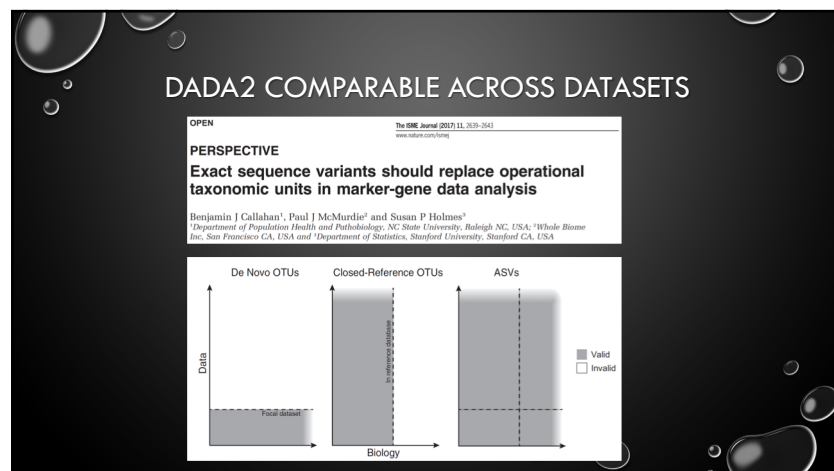
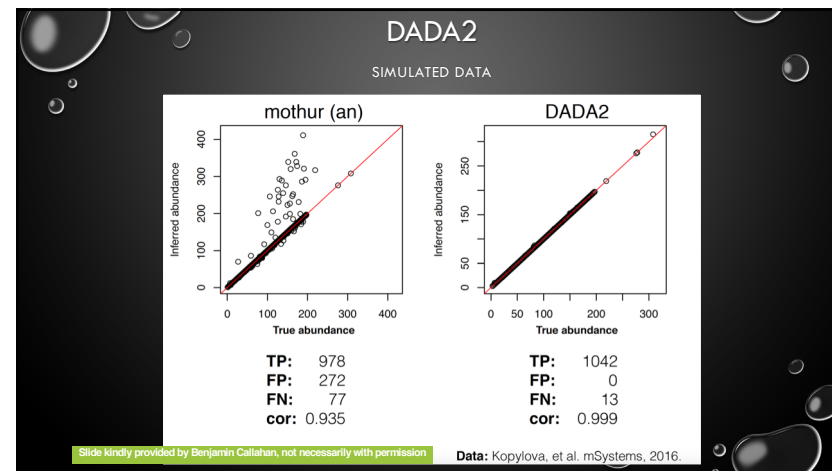
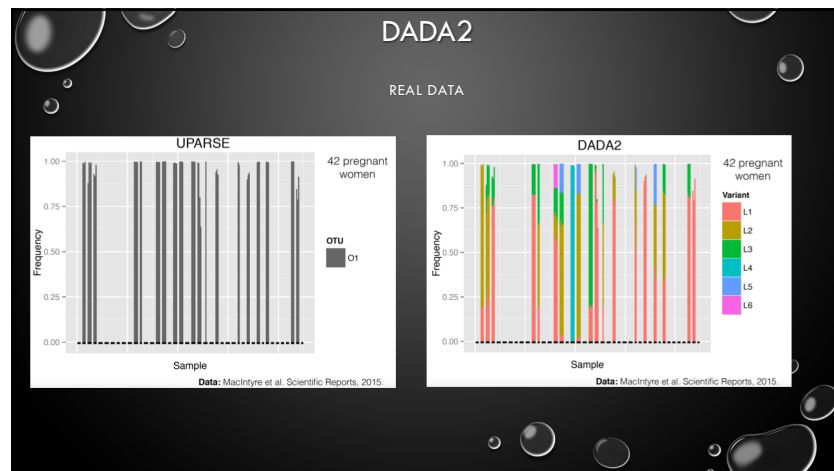
Sequence Tables

Cross-study comparison

Slides kindly provided by Benjamin Callahan, not necessarily with permission

DADA2

- WHY IS IT GOOD?
 - EXTENSIVE BENCHMARKING POINTS TO:
 - HIGH RESOLUTION OF BIOLOGICAL DIFFERENCE AT THE LEVEL OF 1 – 2 NUCLEOTIDES
 - FEWER FALSE POSITIVES THAN OTHER METHODS
 - LINEAR SCALING WITH SAMPLE NUMBER AND FLAT MEMORY REQUIREMENTS
 - OPEN SOURCE
 - COMPARABLE ACROSS DATASETS
 - GREAT FORUM FOR ADVICE AND TROUBLESHOOTING



DADA2 HANDS-ON

[HTTPS://BENJJNEB.GITHUB.IO/DADA2/TUTORIAL.HTML](https://benjjneb.github.io/dada2/tutorial.html)

RUN DADA2

- 1ST PART OF THE SCRIPT
 - CHECK IF THE PRIMERS ARE STILL IN THE READS
- 2ND PART OF THE SCRIPT
 - CHECK THE QUALITY OF THE READS FOR EACH SAMPLE
 - QUALITY FILTER THE R1 AND R2 READS INDEPENDENTLY
 - CONTROL IF THERE ARE ANY PRIMERS LEFT IN THE READS
- 3RD PART OF THE SCRIPT
 - DEREPLICATE THE READS
 - BUILD THE ERROR MODEL TO CORRECT THE READS
 - RUN THE MODEL IN THE DEREPLICATED DATA

RUN DADA2

- 4TH PART OF THE SCRIPT
 - GENERATE AMPLICON SEQUENCE VARIANTS TABLE
 - SAVE THE FILES
- 5TH PART OF THE SCRIPT
 - CHECK FOR CHIMERIC SEQUENCES IN THE DATA
 - CONTROL HOW MANY CHIMERAS WERE FOUND AND EXCLUDED
 - TRACE THE READS THROUGH THE ABOVE STEPS
- 6TH PARTH OF THE SCRIPT
 - SAVE THE FINAL DATA

CLUSTERING OUTPUT OF DADA2

- 7TH PART OF THE SCRIPT
 - OUTSIDE R
 - NAVIGATE TO THE FOLDER WITH THE CHIMERA FREE OUTPUT EXTRACTED FROM DADA2
 - CLUSTER AT 98% SIMILARITY THE AMPLICON SEQUENCE VARIANTS
- 8TH PART OF THE SCRIPT
 - GET TAXONOMICAL ASSIGNMENT TO THE CLUSTERS
 - THIS TAKES QUITE SOME TIME
 - MERGE THE RESULTS WITH THE OTU TABLE
- YOU CAN FURTHER FILTER YOUR OTUS WITH LULU ALGORITHM...

SITES WHERE YOU CAN GET MORE INFORMATION ABOUT DADA2

Problems and solutions

- <https://github.com/benjjneb/dada2/issues>

Manual, complementary information and another annotated tutorial

- <https://www.bioconductor.org/packages/release/bioc/manuals/dada2/man/dada2.pdf>
- <https://benjjneb.github.io/dada2/pool.html#pooling-for-sample-inference>
- https://astrobiomike.github.io/amplicon/dada2_workflow_ex

Benchmarking

- <https://benjjneb.github.io/dada2/sota.html>