

BIO9905MERG1 - Bioinformatics for DNA-metabarcoding and environmental Sequencing

Welcome!



Galina Gusarova



Organization

- Mix of lectures + hands on sessions
- Basic → advanced (different starting points)
- Hands on: Work two and two together → Better learning outcome + need less capacity at Abel
- High variation in experience. Experienced students → help less experienced ones!
- We will connect to Abel and run analyses there → security issues (passwords)
- Kitchen upstairs – room 4213 (tea, coffee, etc.)
- Pizza Monday evening at ~5pm.
- Contact: haavarka@ibv.uio.no or 99697116



Report

- Should describe a factual or fictive DNA metabarcoding study.
- You should describe the various analytical steps and especially emphasise the bioinformatics parts.
- You should not only write **how** the data are analyzed but also **why** you select to analyse the data in a certain way
- Report: Provide a minimum 3 page report with deadline 28 April

Schedule

	Monday 25.3	Tuesday 26.3	Wednesday 27.3	Thursday 28.3	Friday 29.3
9-10					
10-11					
11-12					
12-13	Lunch break	Lunch break	Lunch break	Lunch break	Lunch break
13-14					
14-15					
15-16					
16-17					
17-18	Pizza				

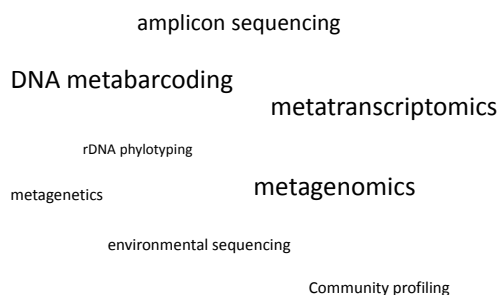
Theme	Teachers
Introduction	Håvard Kausenrud + Robert Lytle
Case studies	Sandy Maurice, Pål Trossvik/Eric De Mulock, Anders Krabberød
Connect and line intro	Anders Krabberød, m.fl.
Introduce the data	Luis Morgado, Anders Krabberød, m.fl.
Demultiplexing + VSEARCH + Swarm	Frederic Mahu/Torstein Rognes/Marit Bjorbakmo
QIIME2	Luis Morgado/Sunil Mundra
LRU/post processing	Sunil Mundra/Luis Morgado
Evolutionary Placement Algorithm	Anders Krabberød/Marit Bjorbakmo
Test out on your own? (Not compulsory)	

Somewhat flexible program – will be small breaks in-between

BIO9905MERG1 - Bioinformatics for **DNA-metabarcoding** and environmental Sequencing

Welcome!

Some confusing / important terms



Some important terms

- Metabarcoding Sequence variation in a single locus (e.g. 16S)
- Metagenomics Genome wide sequence variation
- Metatranscriptomics cDNA sequence variation

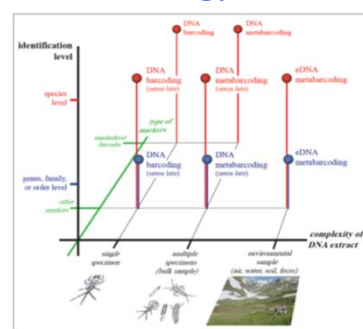
Some important terms

- Metabarcoding Sequence variation in a single locus (e.g. 16S) **Who are there?**
- Metagenomics Genome wide sequence variation **Which genes (and who) are there?**
- Metatranscriptomics cDNA sequence variation **Who are active and doing what?**

Terminology

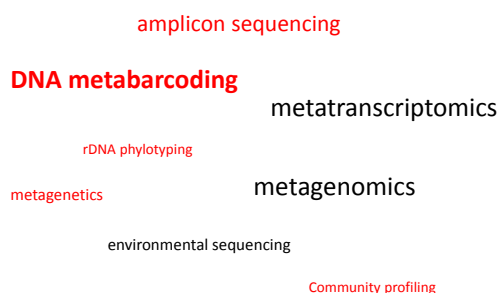


Pierre Taberlet



Taberlet et al. (2012) *Molecular Ecology*, 21, 1789-1793.

Some confusing / important terms



DNA metabarcoding

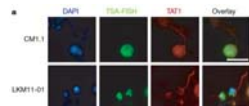
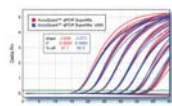
- Research questions:
 - Who are there?
 - How many taxa / operational taxonomic units (alpha/gamma diversity)?
 - Compositional differences (beta diversity)?
 - Who are common – who are rare?
 - Co-occurrence patterns → Interactions
 - Which processes and drivers are shaping the communities?
 - Ecosystem processes: What are they doing..? → indirect evidence



FunGuild

DNA metabarcoding

- A largely descriptive field of science
- Often functions as a first step looking into largely uncharacterized habitats and study systems
- Often generates hypotheses that can be addressed more in-depth with more taxon specific approaches



DNA metabarcoding: From 'wild west' towards an established approach?



- Primary phase
- Poor replication
- Lack of controls
- Lack of insight into important biases
- Poor bioinformatics approaches
- Etc.

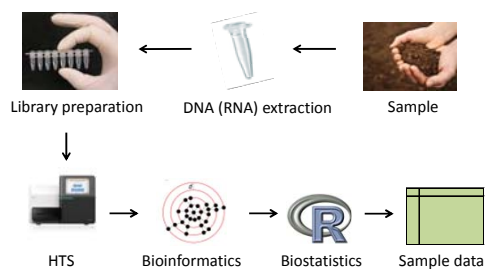
Editorial note for Molecular Ecology

DNA metabarcoding - need for robust experimental designs to draw sound ecological conclusions

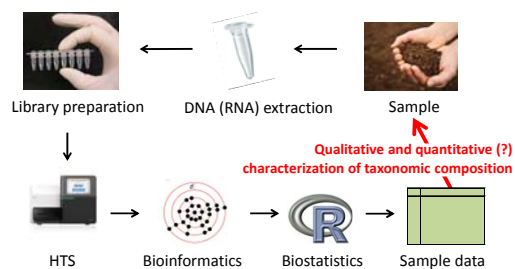
Authors
Lucie Zager¹, Aurélie Bonin¹, Inger G. Alsos¹, Miklós Balint^{1,2}, Holly Bär¹, Frédéric Beyer¹,
Anthony A. Charrier¹, Simon Croft¹, Eric Cousin¹, Bruce E. Douglas^{1,3,4}, Marta De Barbo¹, Ina A.
Dicker¹, Alex J. Doubilet¹, Gentile Francesco Fiorini^{1,5}, Noah Fries^{1,6,7}, Luca Fumagalli¹,
M. Thomas P. Gilbert^{1,8}, Simon Jarman^{1,9}, An Jaapponen¹, Mircea Kussend¹, Ludovic Orlando¹,
Adam Parns^{1,10}, Jan Pavlovskis^{1,11}, Lello Tedesco¹, Philip Francis Thomas¹, Erik
Widénen^{1,12,13}, Pierre Taberlet¹

Established scientific approach with a set of widely accepted guidelines

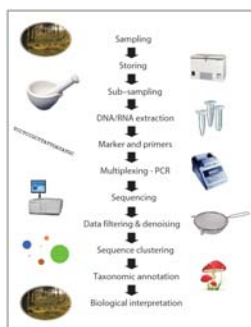
General workflow in DNA metabarcoding studies



General workflow in DNA metabarcoding studies



DNA metabarcoding - many steps



Lindahl et al. 2013

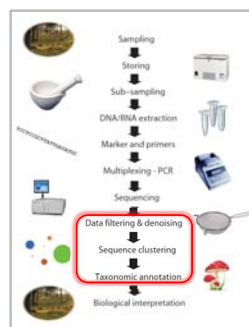


Many
steps...



... to go
wrong

DNA metabarcoding - many steps



Lindahl et al. 2013

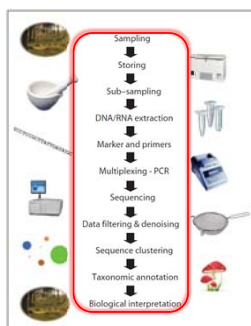


Many steps...



... to go
wrong

DNA metabarcoding - many steps



Lindahl et al. 2013



Many steps...



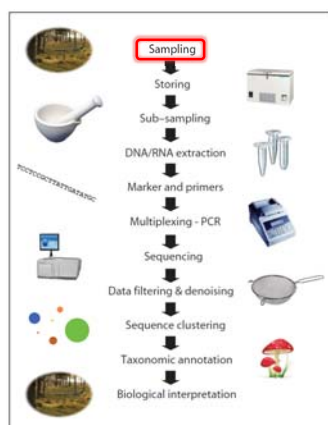
... to go wrong

If new study system – conduct a pilot?

- Which sampling scheme?
- How many replicates?
- Which sequencing depth?
- Which sequencing technique?
- Etc.



→ Depends on the alpha, beta and gamma diversity (which you don't know)..



Sampling strategy

- Must be planned based on the aims and current knowledge about the study system
- Randomization! Experimental subjects must be allocated to treatment groups at random
- **Proper replication!** The more times something is repeated, the greater the confidence of ending up with a genuine result



«Replicate or lie»

Option: Replicate or lie

Number of articles

% with replicates

Appl Environ Microbiol 60 23

Environ Microbiol 47 15

FEMS Microbiol Ecol 29 24

ISME J 23 13

Microbiol Ecol 22 9

Total 181 18

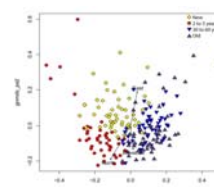
Clone library analysis and pyrosequencing

	Number of articles	% with replicates
Appl Environ Microbiol	60	23
Environ Microbiol	47	15
FEMS Microbiol Ecol	29	24
ISME J	23	13
Microbiol Ecol	22	9
Total	181	18

It doesn't help that you are dealing with HTS data if you don't replicate properly!

Biological + technical replicates!!

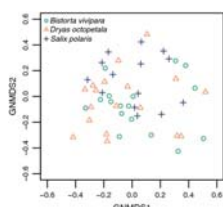
- Some samples should be analyzed multiple times
- The point is to reveal the variability (experimental error) of the analysis technique, thus allowing one to set limits for what is meaningful / significant data



Davey et al. 2015

Biological + technical replicates!!

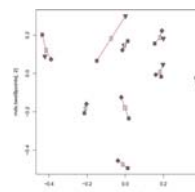
- Some samples should be analyzed multiple times
- The point is to reveal the variability (experimental error) of the analysis technique, thus allowing one to set limits for what is meaningful / significant data



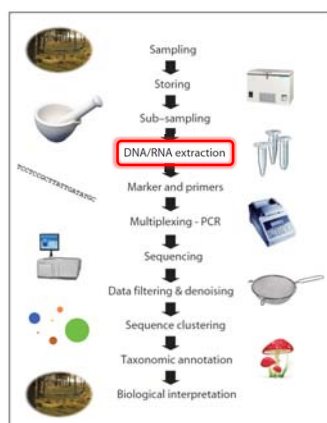
Botnen *et al.* 2015

Biological + technical replicates!!

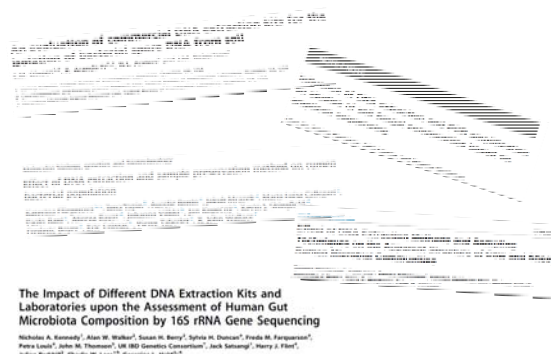
- Some samples should be analyzed multiple times
- The point is to reveal the variability (experimental error) of the analysis technique, thus allowing one to set limits for what is meaningful / significant data



Davey *et al.* 2015



DNA extraction



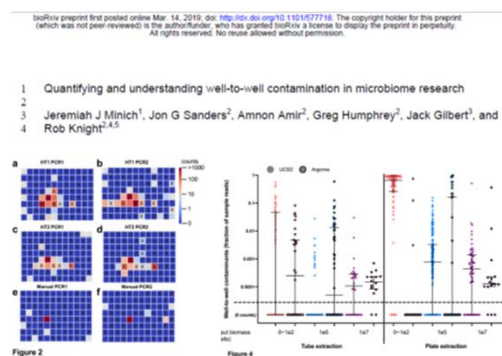
DNA extraction

- Extraction protocols should yield high and uniform amounts of DNA, while the concentration of PCR inhibitors is minimized
- Same protocol should be used for all samples!
- If no proper literature or experimental evidence are available on your study system / substrate – conduct a pilot?!
- Extraction negatives!



- MoBio Power Soil?
- FastDNA kit for Soil?
- EZNA Soil kit?
- CTAB + cleanup kit?

DNA extraction



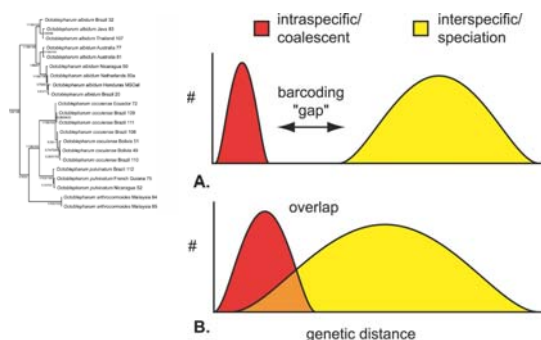
Markers in DNA metabarcoding

- The ideal marker:
 - Have primer sites that are shared by all target organisms and easy to amplify
 - Be of appropriate length for efficient amplification and sequencing
 - Should preferably be of similar length
 - No intragenomic variation (i.e. no paralogs)
 - Be possible to align
 - Have high interspecific variation
 - Have low intraspecific variation
- No known markers meet all these requirements!

Markers in DNA metabarcoding

- The ideal marker:
 - Have primer sites that are shared by all target organisms and easy to amplify
 - Be of appropriate length for efficient amplification and sequencing
 - Should preferably be of similar length
 - No intragenomic variation (i.e. no paralogs)
 - Be possible to align
 - Have high interspecific variation
 - Have low intraspecific variation
- No known markers meet all these requirements!

The barcoding gap

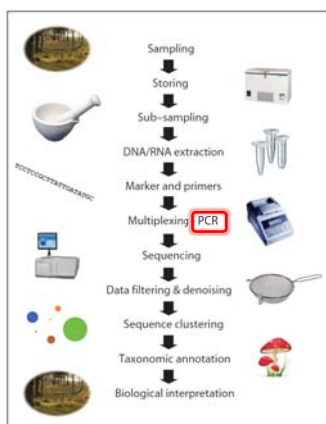


How conserved/variable are the marker?

- 18S: Low variability, low intraspecific variation, low interspecific variation
- 16S: Low variability, low intraspecific variation, low 'interspecific' variation
- ITS: High variability, (relatively) high intraspecific variation, (relatively) high 'interspecific' variation



- Affect how some of the bioinformatics analyses should be conducted!! (i.e. no single way)

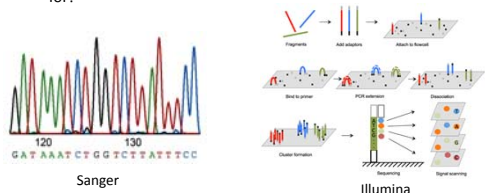


PCR

- Different relevant factors during PCR:
 - Which polymerase enzyme (proofreading or not)?
 - Which RAMP speed?
 - How many cycles?
 - Which annealing temperature?
 - Multiple/replicate PCR reactions followed by pooling?
 - PCR negatives!

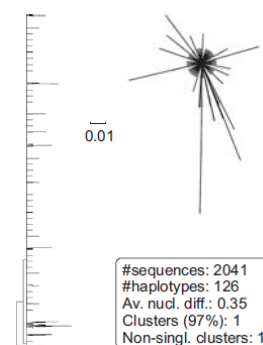
PCR-induced errors

- PCR mutations: polymerase enzyme introduces an erroneous nucleotide
 - Dependent on the technology whether these become «visible»
 - In classic (direct) Sanger sequencing such errors become «diluted»
 - In methods where your final sequence is derived from a single PCR-amplified DNA template they become visible and must be corrected for!



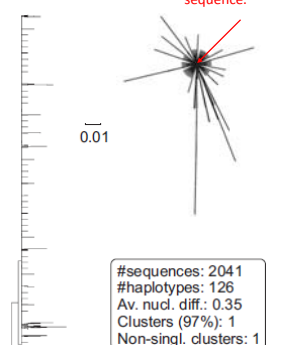
PCR-induced errors

Ecology and Evolution
 Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi



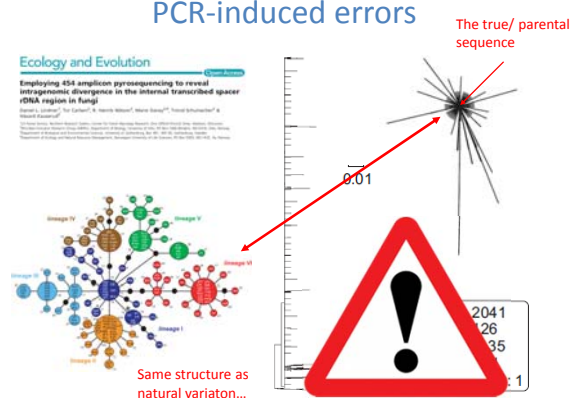
PCR-induced errors

Ecology and Evolution
 Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi

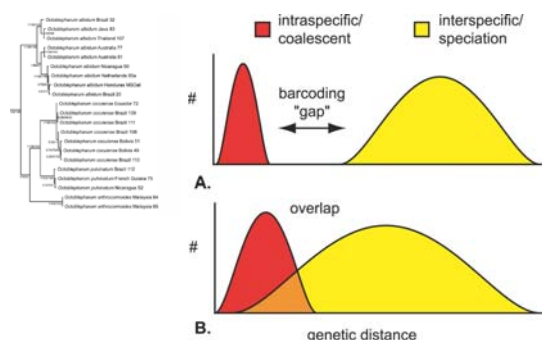


PCR-induced errors

Ecology and Evolution
 Employing 454 amplicon pyrosequencing to reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi

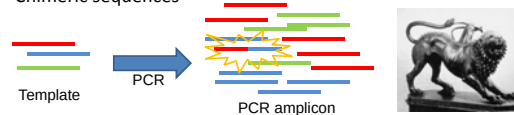


The barcoding gap



PCR-induced errors

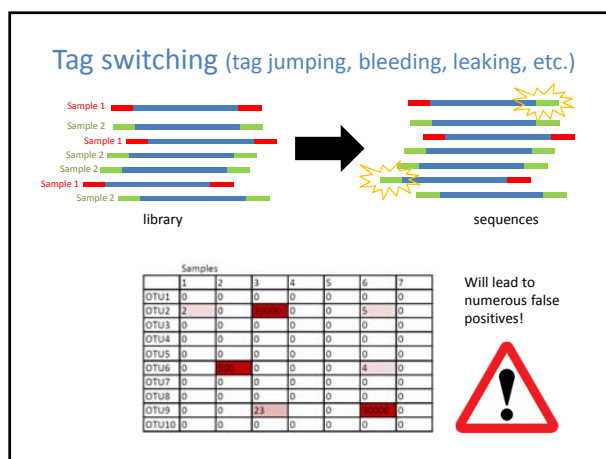
- Chimeric sequences



MOLECULAR ECOLOGY
 ITS all right man: investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities

The level of chimeric sequences depends on how variable the marker is!

Can reduce the problem with PCR settings



Tag switching

Accurate multiplexing and filtering for high-throughput amplicon sequencing
Philippe F. "Ligandated France" and Prevalence J. J.

Up to 28.2% of the unique sequences correspond to undetectable (critical) mistakes in single- or saturated double-tagging libraries.

"We found that an average of 2.6% and 2.1% of sequences had tag combinations, which could be explained by tag jumping..."

Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies

DOI: 10.1016/j.molsec.2018.04.001

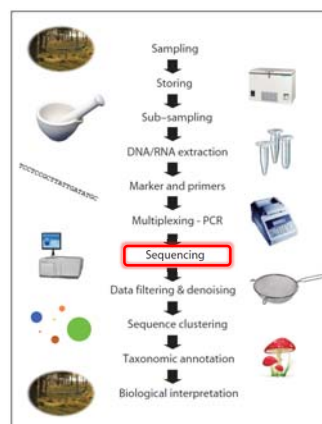
Tag switching

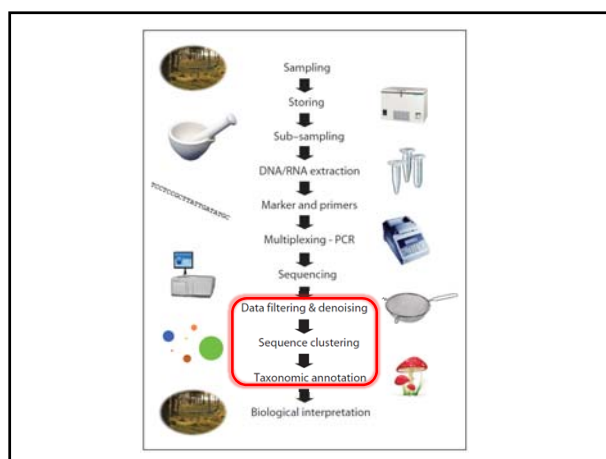
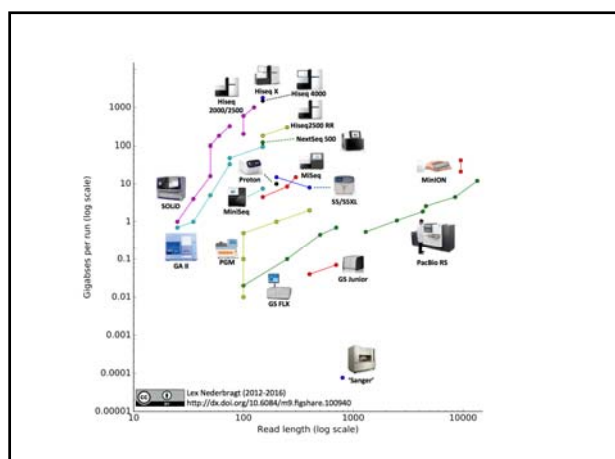
- The problem can be reduced or controlled for by:
 - Tagging in both ends with unique tag combinations
 - Rinse the PCR amplicons thoroughly
 - Avoid PCR steps during the final library preparations steps before sequencing (i.e. when adaptors are introduced)
 - Include positive and negative controls → identify the level of switching/leakage
 - (Drastic) *ad hoc* solution: Process your OTU/sample matrix by removing low frequency occurrences

	1	2	3	4	5	6	7
OTU1	0	0	0	0	0	0	0
OTU2	0	0	0	0	0	5	0
OTU3	0	0	0	0	0	0	0
OTU4	0	0	0	0	0	0	0
OTU5	0	0	0	0	0	0	0
OTU6	0	0	0	0	0	4	0
OTU7	0	0	0	0	0	0	0
OTU8	0	0	0	0	0	0	0
OTU9	0	0	23	0	0	0	0
OTU10	0	0	0	0	0	0	0

Library preparation

- Normalization step before pooling of amplicons into a library by:
 - DNA concentration measurements?
 - Normalization kit (plate)?



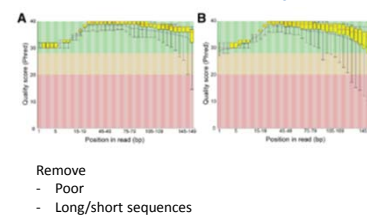


Bioinformatics – main steps

- Quality control
- Produce contigs
- Demultiplexing
- Dereplication
- OTU construction
- Chimera checking
- Taxonomic annotation
- Removal of non-target organisms
- Cleaning of tag bleeding
- OTU modifications
- Positive negatives
- Singleton removal
- Data transformation

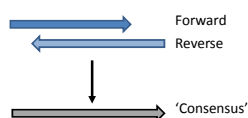
Bioinformatics – main steps

- Quality control
- Produce contigs
- Demultiplexing
- Dereplication
- OTU construction
- Chimera checking
- Taxonomic annotation
- Removal of non-target organisms
- Cleaning of tag bleeding
- OTU modifications
- Positive negatives
- Singleton removal
- Data transformation



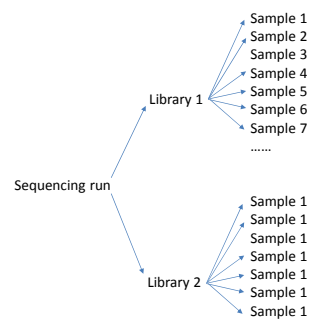
Bioinformatics – main steps

- Quality control
- Produce contigs
- Demultiplexing
- Dereplication
- OTU construction
- Chimera checking
- Taxonomic annotation
- Removal of non-target organisms
- Cleaning of tag bleeding
- OTU modifications
- Positive negatives
- Singleton removal
- Data transformation

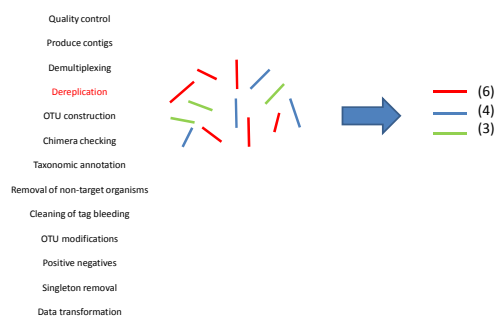


Bioinformatics – main steps

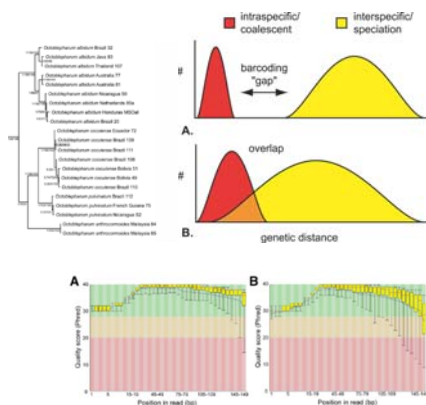
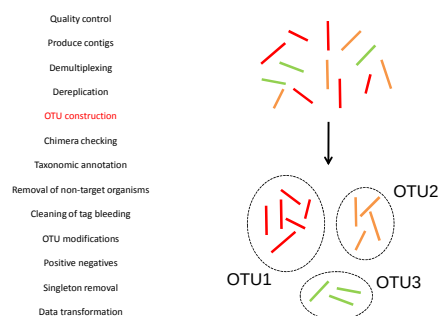
- Quality control
- Produce contigs
- Demultiplexing**
- Dereplication
- OTU construction
- Chimera checking
- Taxonomic annotation
- Removal of non-target organisms
- Cleaning of tag bleeding
- OTU modifications
- Positive negatives
- Singleton removal
- Data transformation



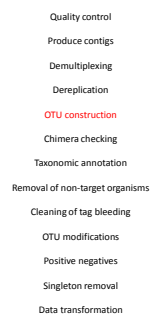
Bioinformatics – main steps



Bioinformatics – main steps

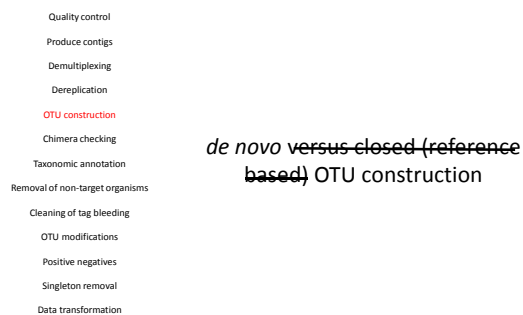


Bioinformatics – main steps

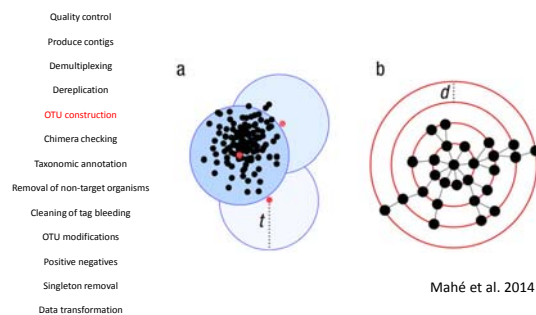


de novo versus closed (reference based) OTU construction?

Bioinformatics – main steps

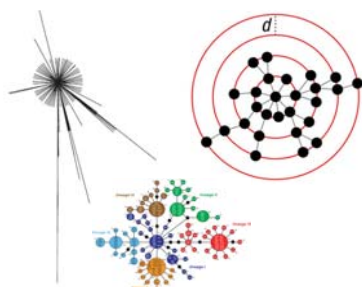


Bioinformatics – main steps



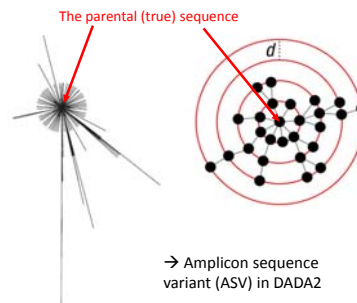
Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation



Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation



How conserved/variable are the marker?

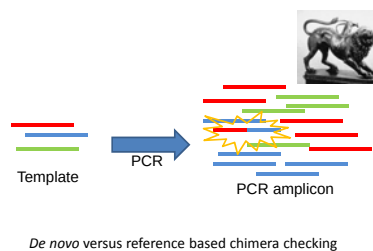
- 18S: Low variability, low intraspecific variation, low interspecific variation
- 16S: Low variability, low intraspecific variation, low 'interspecific' variation
- ITS: High variability, (relatively) high intraspecific variation, (relatively) high 'interspecific' variation



- Affect how some of the bioinformatics analyses should be conducted!! (i.e. no single way)

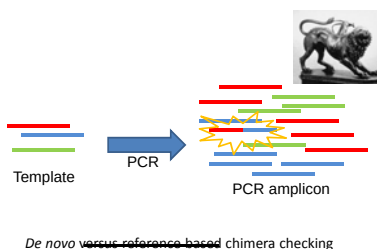
Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation



Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation



The level of chimeric sequences depends on how variable the marker is! → Be aware of false positives in the tests

Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation

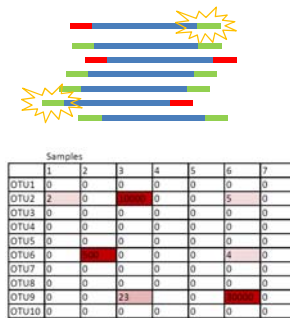


Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation

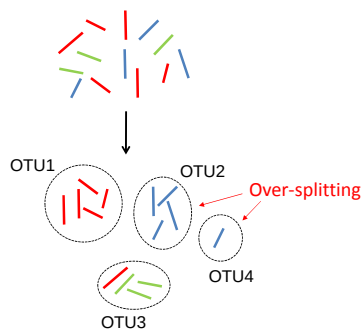
Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation



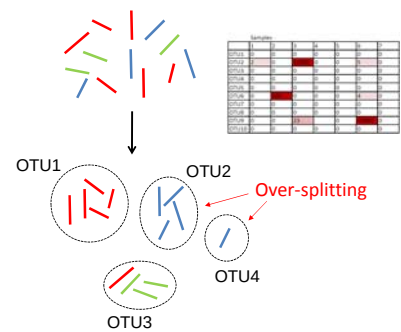
Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation



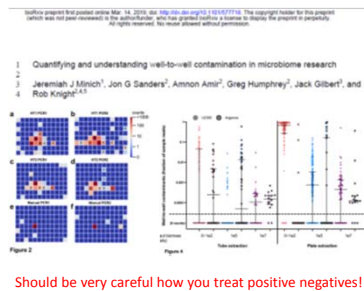
Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation



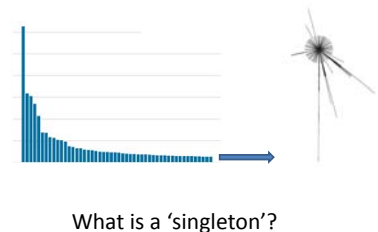
Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation

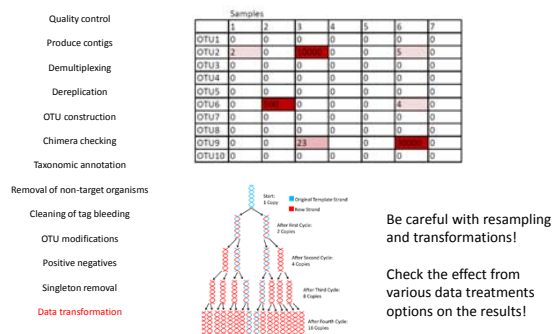


Bioinformatics – main steps

Quality control
Produce contigs
Demultiplexing
Dereplication
OTU construction
Chimera checking
Taxonomic annotation
Removal of non-target organisms
Cleaning of tag bleeding
OTU modifications
Positive negatives
Singleton removal
Data transformation



Bioinformatics – main steps



The importance of controls

- Technical replicates
 - Extraction negatives
 - PCR negatives
 - Positive control(s)
- ➔ Different purposes

Conclusions

- Which methods to use? No general answer – it's context dependent! Must argue for your choices



Editors: ... for Molecular Ecology

DNA metabarcoding - need for robust experimental designs to draw sound ecological conclusions

Authors

Lucie Zinger¹, Aurélie Bonin², Inge G. Aho³, Mikko Buisson⁴, Holly Bak⁵, Frédéric Borel⁶, Anthony A. Chaves⁷, Simon Côté⁸, Eric Coissac⁹, Bruce E. Deagle¹⁰, Maria De Barba¹¹, Ian A. Dickie¹², Alan J. Dunning¹³, Gentile Francesco Ficetola^{14,15}, Noah Fierer^{16,17}, Luca Fanapat¹⁸, M. Thomas P. Gilbert^{19,20}, Susan Jarama^{21,22}, An J. Jongsom²³, Håvard Kassen²⁴, Ludovic Orlando^{25,26}, John Paus^{27,28}, Jan Pawłowski^{29,30}, Lello Tedesco³¹, Philip Francis Thomas³², Erik Willerslev^{33,34}, Pierre Taberlet³⁵