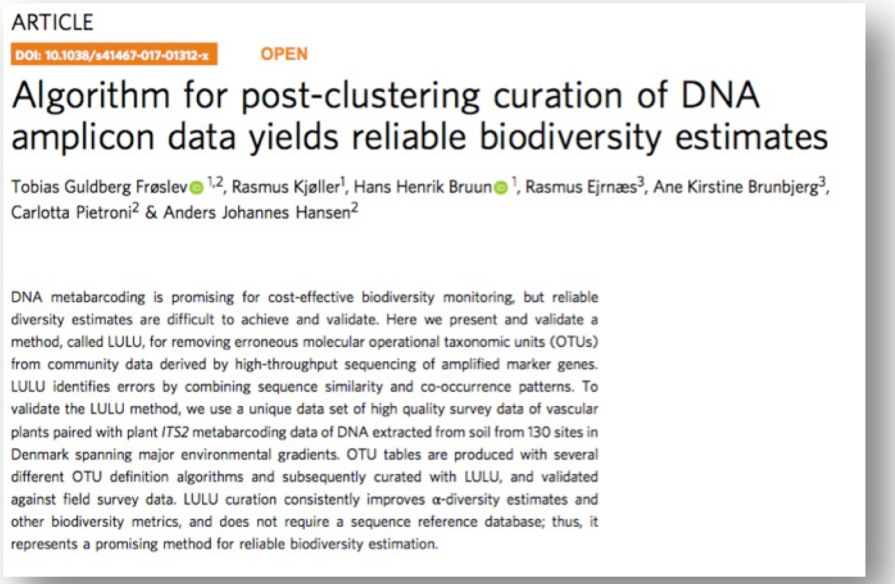


UiO : Department of Biosciences
University of Oslo

LULU – a method for getting better OTU data

Sunil Mundra





ARTICLE

DOI: 10.1038/s41467-017-01312-x OPEN

Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates

Tobias Guldberg Frøslev[1,2], Rasmus Kjøller[1], Hans Henrik Bruun[1], Rasmus Ejrnæs[3], Ane Kirstine Brunbjerg[3], Carlotta Pietroni[2] & Anders Johannes Hansen[2]

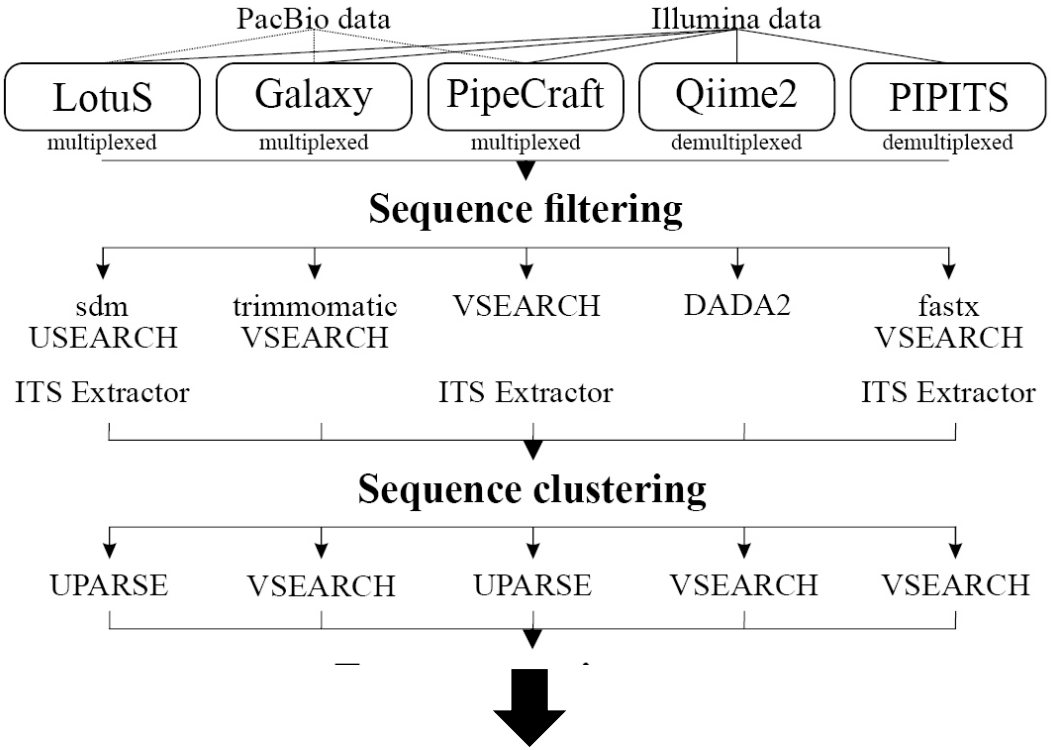
DNA metabarcoding is promising for cost-effective biodiversity monitoring, but reliable diversity estimates are difficult to achieve and validate. Here we present and validate a method, called LULU, for removing erroneous molecular operational taxonomic units (OTUs) from community data derived by high-throughput sequencing of amplified marker genes. LULU identifies errors by combining sequence similarity and co-occurrence patterns. To validate the LULU method, we use a unique data set of high quality survey data of vascular plants paired with plant *ITS2* metabarcoding data of DNA extracted from soil from 130 sites in Denmark spanning major environmental gradients. OTU tables are produced with several different OTU definition algorithms and subsequently curated with LULU, and validated against field survey data. LULU curation consistently improves α-diversity estimates and other biodiversity metrics, and does not require a sequence reference database; thus, it represents a promising method for reliable biodiversity estimation.

Metabarcoding analysis

Name	Description and link
DADA2	Amplicon sequence variant analysis pipeline
	https://benjjneb.github.io/dada2/
Galaxy	Web-based platform, including various analytical tools
	https://usegalaxy.org/
LotuS	Full pipeline for amplicon data
	http://psbweb05.psb.ugent.be/lotus/index.html
mothur	Versatile software suite (designed mostly for 16S rRNA)
	https://www.mothur.org
AMPtk	Full pipeline for amplicon data
	http://amptk.readthedocs.io
OBITools	Versatile software package
	https://git.metabarcoding.org/obitools
PipeCraft	Full pipeline for amplicon data (with graphical user interface)
	https://plutof.ut.ee/#/datacite/10.15156%2FBIO%2F587450
PIPITS	Full pipeline for fungal ITS amplicon data (only for Illumina data)
	https://github.com/hsgweon/pipits
QIIME	Full pipeline for amplicon data (designed mostly for 16S rRNA)
	https://qiime2.org
SEED2	Full pipeline for amplicon data (with graphical user interface; on Windows)
	http://www.biomed.cas.cz/mbu/lbwrf/seed
Microbiology.se	Tools, including ITSx and Metaxa2, for processing ITS, SSU and LSU data
	http://microbiology.se
USEARCH	Versatile software package
	https://www.drive5.com/usearch
VSEARCH	Versatile software package
	https://github.com/torognes/vsearch

Nillson et al. 2019

Metabarcoding analysis



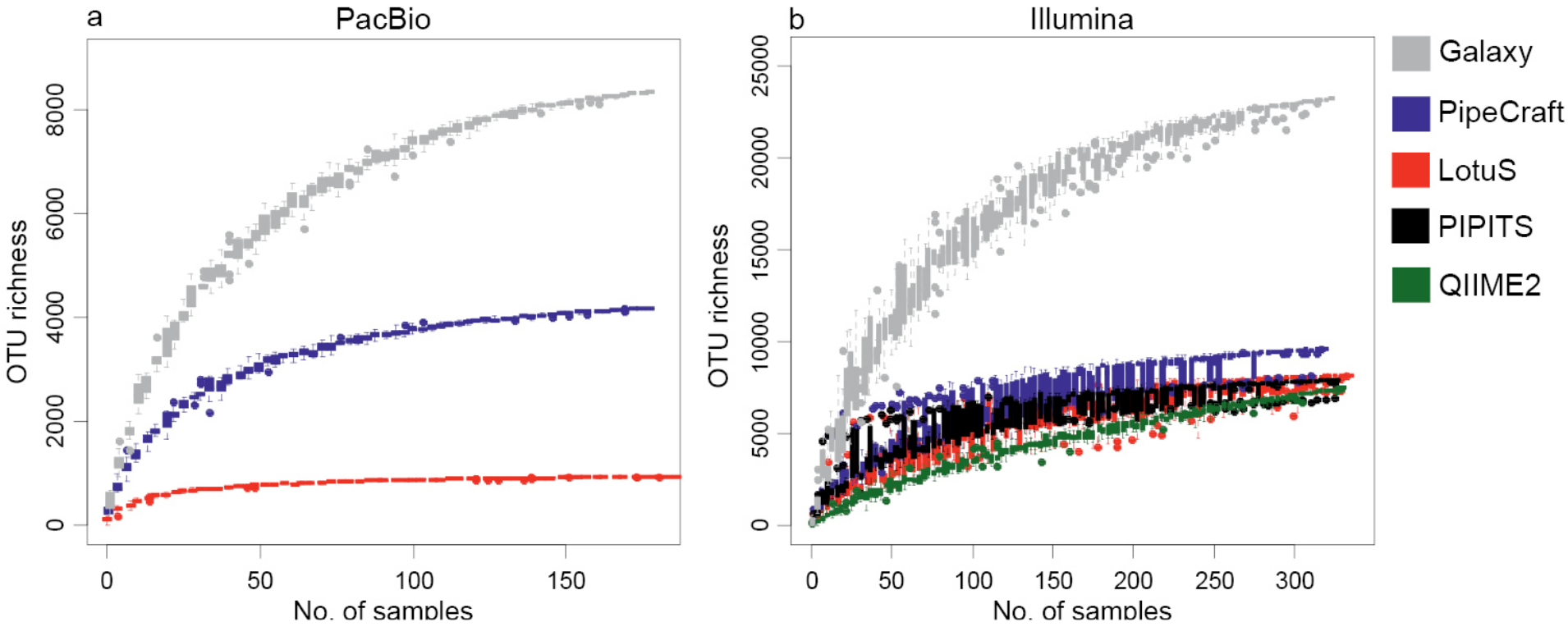
	A	B	C	D	E	F
1		sample01	sample01	sample02	sample02	sample03
2	OTU0001	528	68	1755	773	2167
3	OTU0002	138	68	559	2588	1198
4	OTU0003	36	533	673	351	815
5	OTU0004	1	2618	5	17	19
6	OTU0005	224	237	81	271	313

Anslan et al. 2018

Sample vs species matrix (OTU table)

UiO : Department of Biosciences
University of Oslo

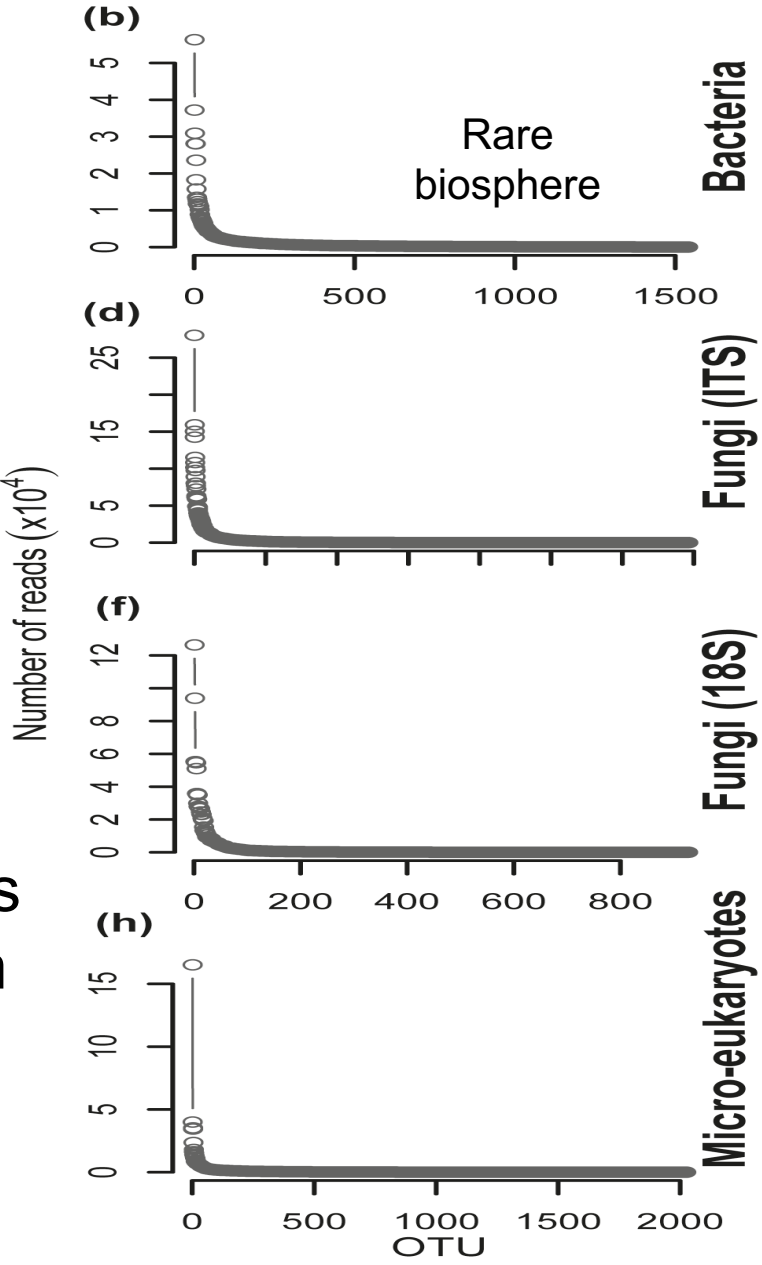
Richness estimation



UiO : Department of Biosciences
University of Oslo

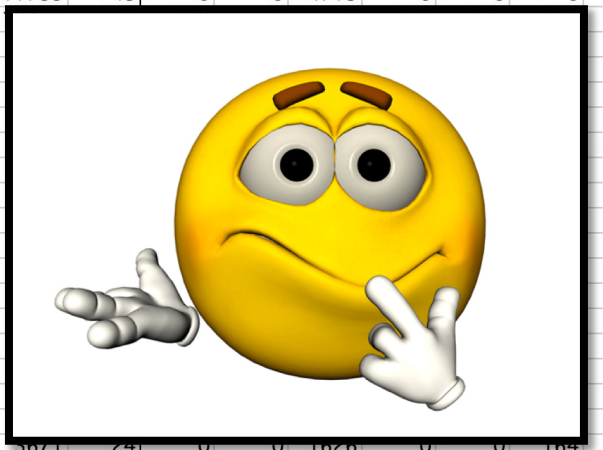
OTUs skewness

- e-DNA Sampling issues
- PCR and sequencing errors
- Intraspecific genetic variation
- Incomplete reference databases (difficult in separating error from true OTUs)



UiO : Department of Biosciences
University of Oslo

taxonomy	references	identity	total	spread	S049	S178	S070	S252	S107	S133	S172	S138	S106	S081	S155	S248	S148	S259	S061	S064	S
s__Archaeorhizomyces+sp	DQ309164_SH019959.07FU	100	63159	87	13665	0	0	1734	1502	6859	9	6076	1998	134	4232	243	2798	0	705	4710	
s__Archaeorhizomyces+sp	DQ309123_SH180488.07FU	100	49267	33	34167	5385	0	0	0	0	0	171	0	0	0	0	0	0	2634	0	
s__Archaeorhizomyces+sp	FR877526_SH180492.07FU	100	32946	29	0	27130	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
s__Archaeorhizomyces+sp	KF412410_SH025554.07FU	94.3	20602	83	0	0	9	2424	1818	2	7502	0	238	1423	0	0	0	89	0	0	
s__Archaeorhizomyces+sp	UDB004160_SH222011.07FU	94.1	19739	74	0	0	10181	49	42	292	0	0	0	0	0	0	10	1684	0	0	
s__Archaeorhizomyces+finlayi	JF836021_SH222246.07FU	77.9	15633	33	0	0	0	6283	5397	0	0	0	149	857	0	0	0	225	0	0	
s__Archaeorhizomyces+sp	HQ211925_SH203825.07FU	89.3	14821	47	0	0	0	321	291	1127	0	0	229	1953	0	0	136	1214	537	0	
s__Archaeorhizomyces+sp	HM069472_SH209756.07FU	100	14095	126	178	325	0	258	214	677	61	972	643	170	416	0	167	35	138	9	
s__Archaeorhizomyces+sp	GQ159981_SH222239.07FU	100	13916	105	2075	356	0	1094	1085	107	10	1050	987	92	989	0	0	72	0	0	
s__Archaeorhizomyces+sp	GQ160152_SH203823.07FU	94.9	11739	45	0	0	4713	0	0	0	0	6	0	0	0	0	0	79	0	0	
s__Archaeorhizomyces+sp	EU292259_SH180923.07FU	99.4									0	384	25	45	2118	348	1704	0	569	32	
s__Archaeorhizomyces+sp	GU174326_SH222009.07FU	99.5									362	6	3	118	0	0	0	134	0	0	
s__Archaeorhizomyces+sp	GQ160105_SH203822.07FU	99.4									0	0	1888	232	0	0	235	0	37	0	
s__Archaeorhizomyces+sp	GU174299_SH019983.07FU	94.8									0	0	115	45	0	0	23	0	0	0	
s__Archaeorhizomyces+sp	UDB004160_SH222011.07FU	92.5									0	0	0	0	0	0	0	9	0	0	
s__Archaeorhizomyces+sp	GQ160152_SH203823.07FU	93.8									3574	6	0	6	0	0	7	287	0	35	
s__Archaeorhizomyces+sp	JN012084_SH222241.07FU	99.4									0	0	539	0	0	2157	0	0	0	0	
s__Archaeorhizomyces+sp	GQ160017_SH004229.07FU	94.3									0	0	0	1395	0	0	0	0	531	0	
s__Archaeorhizomyces+sp	KF617305_SH004487.07FU	89.1									0	14	0	0	0	0	0	280	0	0	
s__Archaeorhizomyces+sp	GQ160019_SH180521.07FU	97.6									600	0	144	463	0	0	0	12	8	0	
s__Archaeorhizomyces+sp	GU174290_SH025552.07FU	99									327	0	0	44	0	0	0	946	0	0	
s__Archaeorhizomyces+sp	GQ160152_SH203823.07FU	94.9									0	0	536	0	0	2941	0	0	0	0	
s__Archaeorhizomyces+sp	GQ160019_SH180521.07FU	87.5									977	0	368	65	66	0	0	0	0	0	
s__Archaeorhizomyces+sp	GU174298_SH222010.07FU	96.2									0	0	0	0	0	0	0	765	0	0	
s__Archaeorhizomyces+sp	JN006469_SH180489.07FU	98.8									0	197	0	0	0	0	0	0	245	0	
s__Archaeorhizomyces+borealis	EU046087_SH204031.07FU	100									0	46	0	0	0	0	0	0	328	0	
s__Archaeorhizomyces+sp	GU174301_SH222037.07FU	94.8									0	0	23	180	0	0	0	0	0	0	
s__Archaeorhizomyces+sp	GU174326_SH222009.07FU	94.7	3671	24	0	0	1626	0	0	164	0	0	0	0	0	0	0	28	0	0	
s__Archaeorhizomyces+sp	GQ160036_SH203826.07FU	100	3332	21	0	0	0	0	0	0	834	4	0	0	0	0	101	0	0	0	
s__Archaeorhizomyces+sp	AY969828_SH024137.07FU	87.8	3294	56	0	0	0	0	0	0	171	3	0	0	4	0	17	3	83	0	
s__Archaeorhizomyces+sp	GQ159981_SH222239.07FU	92.2	3166	13	2	0	0	238	189	0	0	0	45	0	0	785	0	0	0	0	
s__Archaeorhizomyces+sp	JX042816_SH181604.07FU	91.4	2003	22	0	0	770	0	0	420	0	0	0	0	0	0	0	0	0	0	
s__Archaeorhizomyces+sp	FJ552837_SH020116.07FU	87.2	1720	14	0	0	251	0	0	1244	0	0	0	0	0	0	0	0	0	0	
s__Archaeorhizomyces+sp	KF617495_SH004483.07FU	93.7	1424	26	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	
s__Archaeorhizomyces+sp	GU174301_SH222037.07FU	99.5	1414	64	0	0	0	16	11	0	151	26	0	10	1	0	0	0	10	0	
s__Archaeorhizomyces+sp	KF412410_SH025554.07FU	92.6	1401	13	0	0	0	203	153	0	0	0	0	455	0	0	0	0	0	0	
s__Archaeorhizomyces+sp	JF755899_SH004485.07FU	100	1355	39	0	0	0	0	0	53	0	451	0	0	0	0	0	0	20	0	
s__Archaeorhizomyces+sp	GQ160152_SH203823.07FU	94.3	1230	26	0	0	0	286	284	0	15	0	13	85	0	0	0	0	149	12	
s__Archaeorhizomyces+sp	GU055693_SH215128.07FU	91	1216	16	0	0	0	0	0	181	0	0	0	0	0	0	0	0	0	0	



10687 Achaeorhizomyces OTUs!

UiO : Department of Biosciences
University of Oslo



axonomy	ident%	total read count	No samples	S103	S187	S237	S069				
agus grandifolia	91.477	82	3	0	0	0					
agus grandifolia	97.143	18	1	0	0	0					
agus hayatae	97.429	141	7	0	0	9					
agus hayatae	93.429	82	7	0	0	6					
agus hayatae	96.571	67	3	0	0	0					
agus hayatae	97.714	48	5	0	0	0					
agus hayatae	96.857	47	4	0	0	0					
agus hayatae	98.571	28	1	0	28	0					
agus hayatae	95.714	24	2	0	0	0					
agus hayatae	98.571	23	2	0	0	6	0	0	0		
agus hayatae	97.714	17	2	0	0	0	0	0	0		
agus hayatae	96	14					0	0	0		
agus hayatae	96	11					0	2	0		
agus hayatae	90.857	11					0	0	0		
agus hayatae	96.857	8					0	1	0		
agus hayatae	98.571	6					0	0	4		
agus hayatae	96.857	6					0	0	0		
agus orientalis	99.143	5273					0	116	185		
agus orientalis	98.857	3658					0	87	72		
agus orientalis	96	467					0	5	9		
agus orientalis	98.867	93					0	1	2		
agus orientalis	98.571	19					0	0	0		
agus orientalis	98.857	13					0	0	0		
agus orientalis	96.571	11					0	0	0		
agus orientalis	99.143	11					0	0	0		
agus orientalis	99.143	8					0	1	1		
agus orientalis	98.857	3					0	0	0		
agus sylvatica	100	574565					20113	16390	16988	14	
agus sylvatica	100	59106					1607	1479	1078	1	
agus sylvatica	99.433	19599					0	536	226	1	
agus sylvatica	98.867	19211					515	413	493		
agus sylvatica	99.433	9703					642	66	56		
agus sylvatica	98.3	9585					0	255	474		
agus sylvatica	99.433	9425					699	319	164		
agus sylvatica	98.867	5099					0	288	124		
agus sylvatica	99.433	4954					0	85	46		
agus sylvatica	98.3	2732	57	3	215	21	0	43	22		
agus sylvatica	99.15	2567	57	76	86	112	0	94	33		
agus sylvatica	99.433	1455	36	93	17	53	0	15	38		
agus sylvatica	99.433	868	26	44	0	29	0	19	58		
agus sylvatica	97.143	588	34	18	3	56	0	39	11		
agus sylvatica	99.15	499	28	0	20	8	0	11	26		
agus sylvatica	99.433	460	21	25	28	3	0	27	0		
agus sylvatica	96.154	320	20	19	35	9	0	0	0		



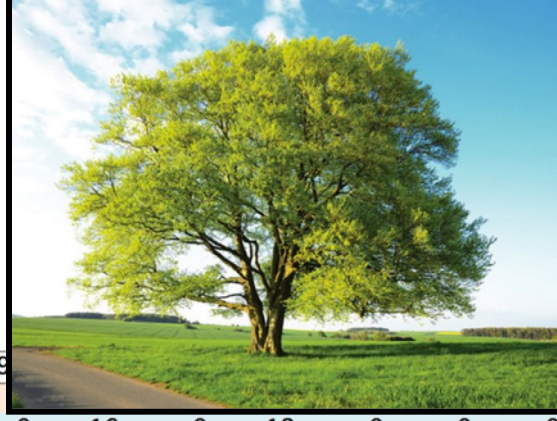
156 Fagus OTUs!

Filtering and trimming

- 97% clustering
- OTU with at least 10 reads in total
- OTU present in at least 2 samples

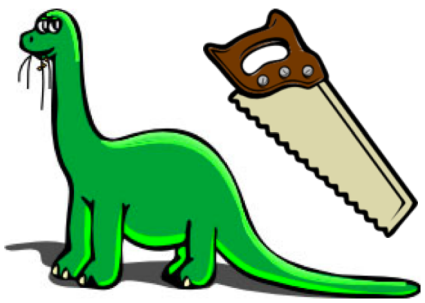
UiO : Department of Biosciences
University of Oslo

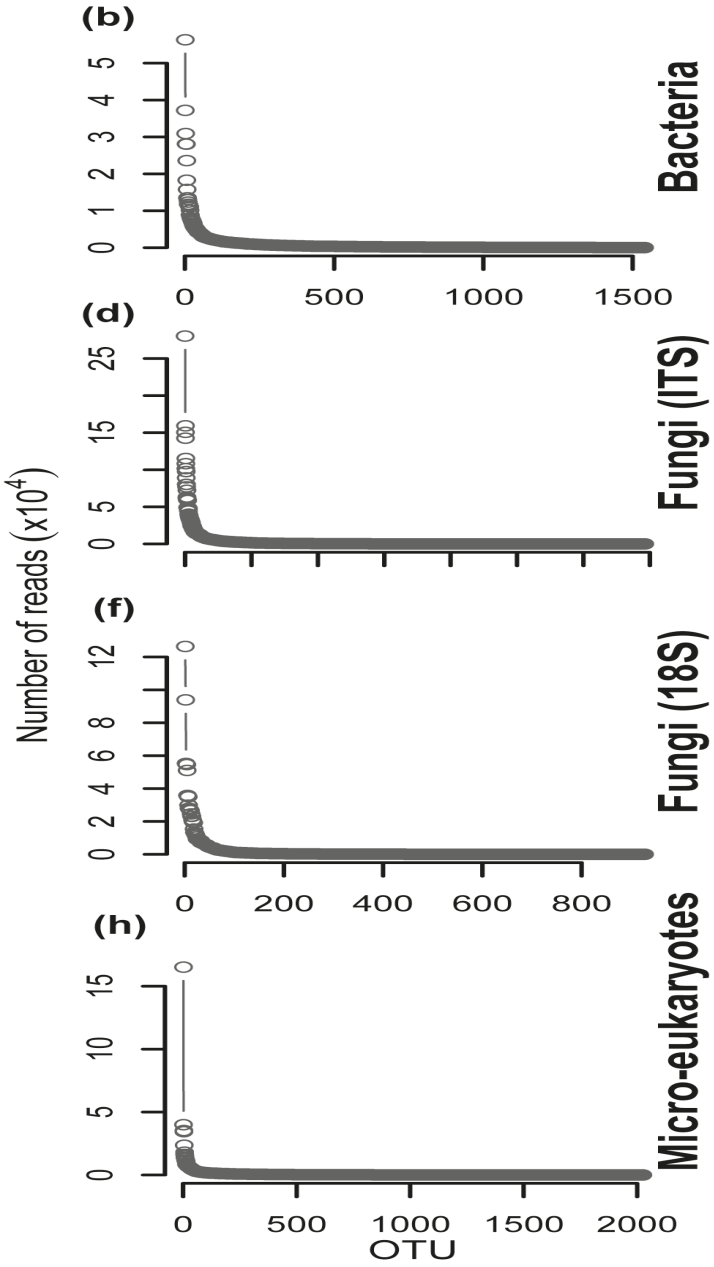
Filtering and trimming



majority_name	pident	total	spread	S187	S237	S103	S069	S101	S376	S211	S050	S374	S079						
Fagus grandifolia	91.477	93	3	0	0	0	0	0	0	0	0	0							
Fagus hayatae	97.429	286	14	0	18	0	3	3	0	25	2	0	0	16	0	12	0	0	0
Fagus hayatae	97.714	159	10	29	8	0	0	0	29	20	0	0	0	0	0	0	0	0	0
Fagus hayatae	93.429	113	6	0	9	0	0	0	15	0	0	0	0	0	0	0	3	0	0
Fagus hayatae	95.714	88	6	0	0	0	0	13	0	0	6	0	0	0	0	9	0	0	0
Fagus hayatae	96	21	3	0	0	0	0	2	0	0	0	0	0	0	0	15	0	0	0
Fagus hayatae	96	20	4	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
Fagus hayatae	90.857	13	2	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fagus orientalis	96	690	30	5	12	10	0	26	17	19	55	26	12	35	0	0	0	42	0
Fagus sylvatica	100	652357	111	24485	24308	23924	23077	19680	19184	18764	18343	16581	16375	15651	15457	15174	14383	14091	13630
Fagus sylvatica	96.884	8661	66	260	269	512	0	245	415	398	285	126	165	137	235	150	157	262	334
Fagus sylvatica	97.143	704	33	5	72	19	0	52	14	12	16	61	61	28	0	63	0	36	0
Fagus sylvatica	96.154	292	19	48	9	19	0	0	0	13	8	0	0	6	0	0	0	0	0
Fagus sylvatica	96.034	284	17	10	17	0	0	16	18	0	0	31	72	0	0	1	0	0	0
Fagus sylvatica	97.143	55	21	1	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0
Fagus sylvatica	97.429	31	8	8	1	0	0	0	0	0	9	0	0	0	0	0	2	1	0
Fagus sylvatica	100	22	5	0	0	0	0	0	0	0	0	0	5	0	1	0	0	11	0
Fagus sylvatica	96.601	11	2	0	0	0	0	0	0	0	10	0	0	1	0	0	0	0	0
Fagus sylvatica	96.884	11	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Still 19 Fagus OTUs!




(b)
Bacteria
(d)
Fungi (ITS)
(f)
Fungi (18S)
(h)
Micro-eukaryotes
Number of reads (x10^4)
OTU



UiO : Department of Biosciences
University of Oslo

LULU – An algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates

- The aim is to retain true α-diversity and taxonomic composition, while discarding the artefactual OTUs from community data derived by HTS of marker genes (metabarcoding).

- Independent of reference database and pipeline/algorithm

LULU – An algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates

- The core mechanism is the identification and merging of 'daughter' OTUs with consistently co-occurring, sequence similar, but more abundant 'parent' OTUs across a multi-sample data set, under the assumption that the 'daughter' OTUs are artefacts.

Fagus sylvatica	100	652357	111	24485	24308	23924	23077	19680	19184	18764	18343	16581	16375	15651	15457	15174	14383	14091	13630

UiO : Department of Biosciences
University of Oslo

LULU – An algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates

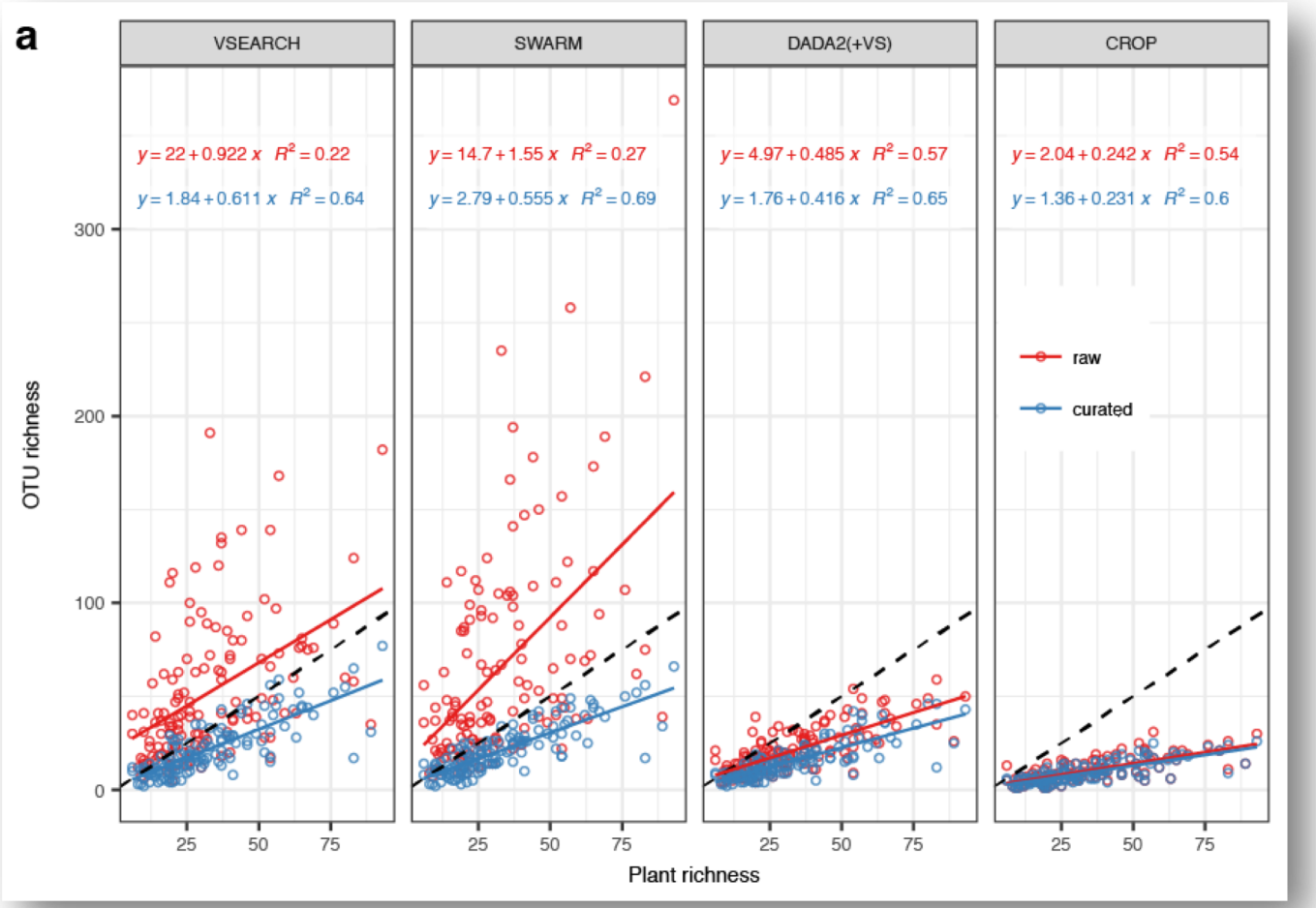
majority_name	pident	total	spread	S187	S237	S103	S069	S101	S376	S211	S050	S374	S079	S239	S352	S273	S030	S200	S303
Fagus grandifolia	91.477	93	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fagus hayatae	97.429	286	14	0	18	0	3	3	0	25	2	0	0	16	0	12	0	0	0
Fagus hayatae	97.714	159	10	29	8	0	0	0	29	20	0	0	0	0	0	0	0	0	0
Fagus hayatae	93.429	113	6	0	9	0	0	0	15	0	0	0	0	0	0	0	3	0	0
Fagus hayatae	95.714	88	6	0	0	0	0	13	0	0	6	0	0	0	0	9	0	0	0
Fagus hayatae	96	21	3	0	0	0	0	2	0	0	0	0	0	0	0	15	0	0	0
Fagus hayatae	96	20	4	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
Fagus hayatae	90.857	13	2	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fagus orientalis	96	690	30	5	12	10	0	26	17	19	55	26	12	35	0	0	0	42	0
Fagus sylvatica	100	652357	111	24485	24308	23924	23077	19680	19184	18764	18343	16581	16375	15651	15457	15174	14383	14091	13630
Fagus sylvatica	96.884	8661	66	260	269	512	0	245	415	398	285	126	165	137	235	150	157	262	334
Fagus sylvatica	97.143	704	33	5	72	19	0	52	14	12	16	61	61	28	0	63	0	36	0
Fagus sylvatica	96.154	292	19	48	9	19	0	0	0	13	8	0	0	6	0	0	0	0	0
Fagus sylvatica	96.034	284	17	10	17	0	0	16	18	0	0	31	72	0	0	1	0	0	0
Fagus sylvatica	97.143	55	21	1	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0
Fagus sylvatica	97.429	31	8	8	1	0	0	0	0	0	9	0	0	0	0	0	2	1	0
Fagus sylvatica	100	22	5	0	0	0	0	0	0	0	0	0	5	0	1	0	0	11	0
Fagus sylvatica	96.601	11	2	0	0	0	0	0	0	0	10	0	0	1	0	0	0	0	0
Fagus sylvatica	96.884	11	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

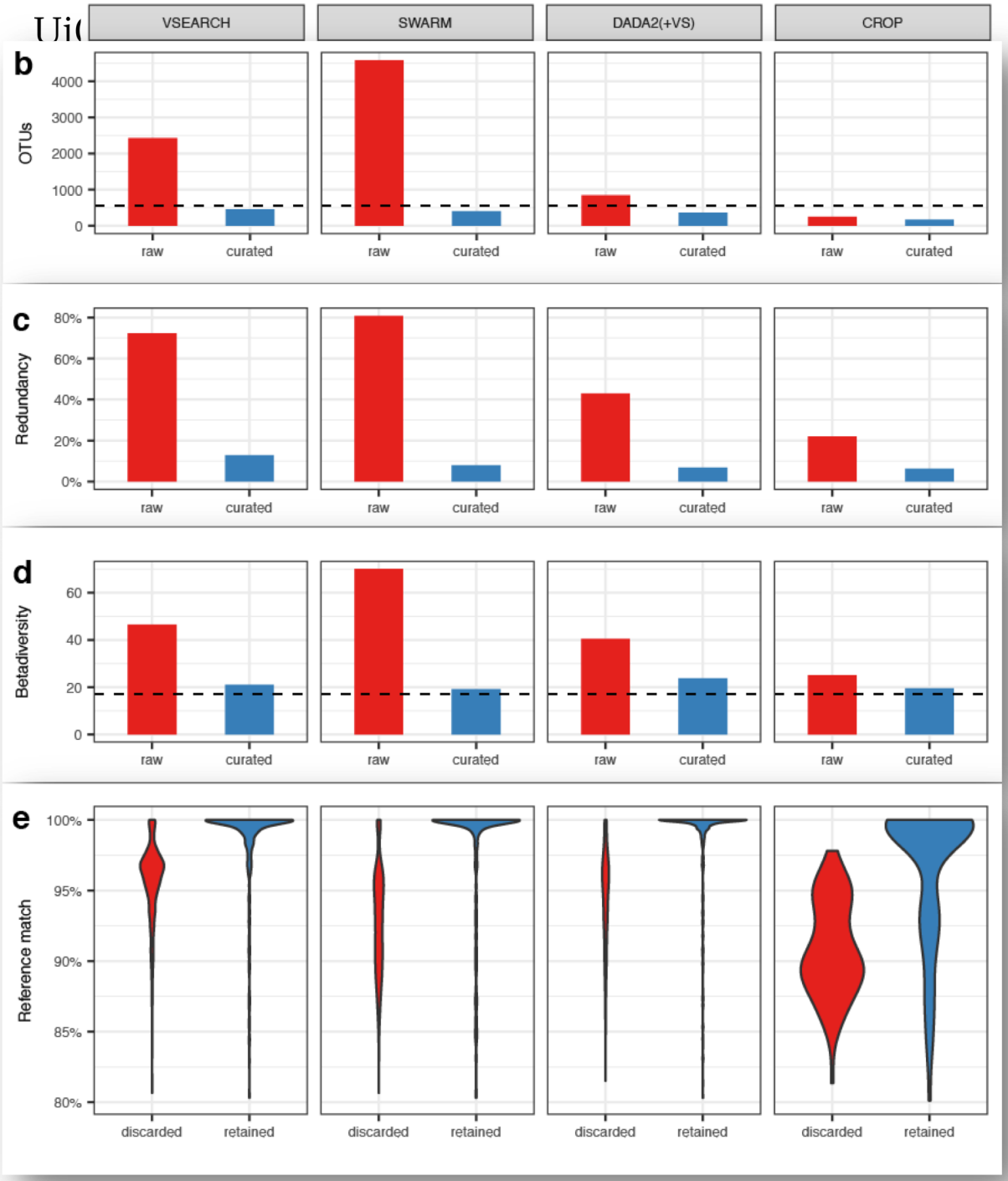
Parent OTU

Daughter OTU

Curation effect: OTU vs plant richness

- Correspondence of OTU (plant *ITS2* sequence data) richness vs. plant richness for each of the 130 sampling sites





Total OTU richness vs Plant richness

Total number of OTUs compared to total plant species recorded (564 species)

Taxonomic redundancy

Percentage of OTUs having taxonomically redundant annotation

Betadiversity

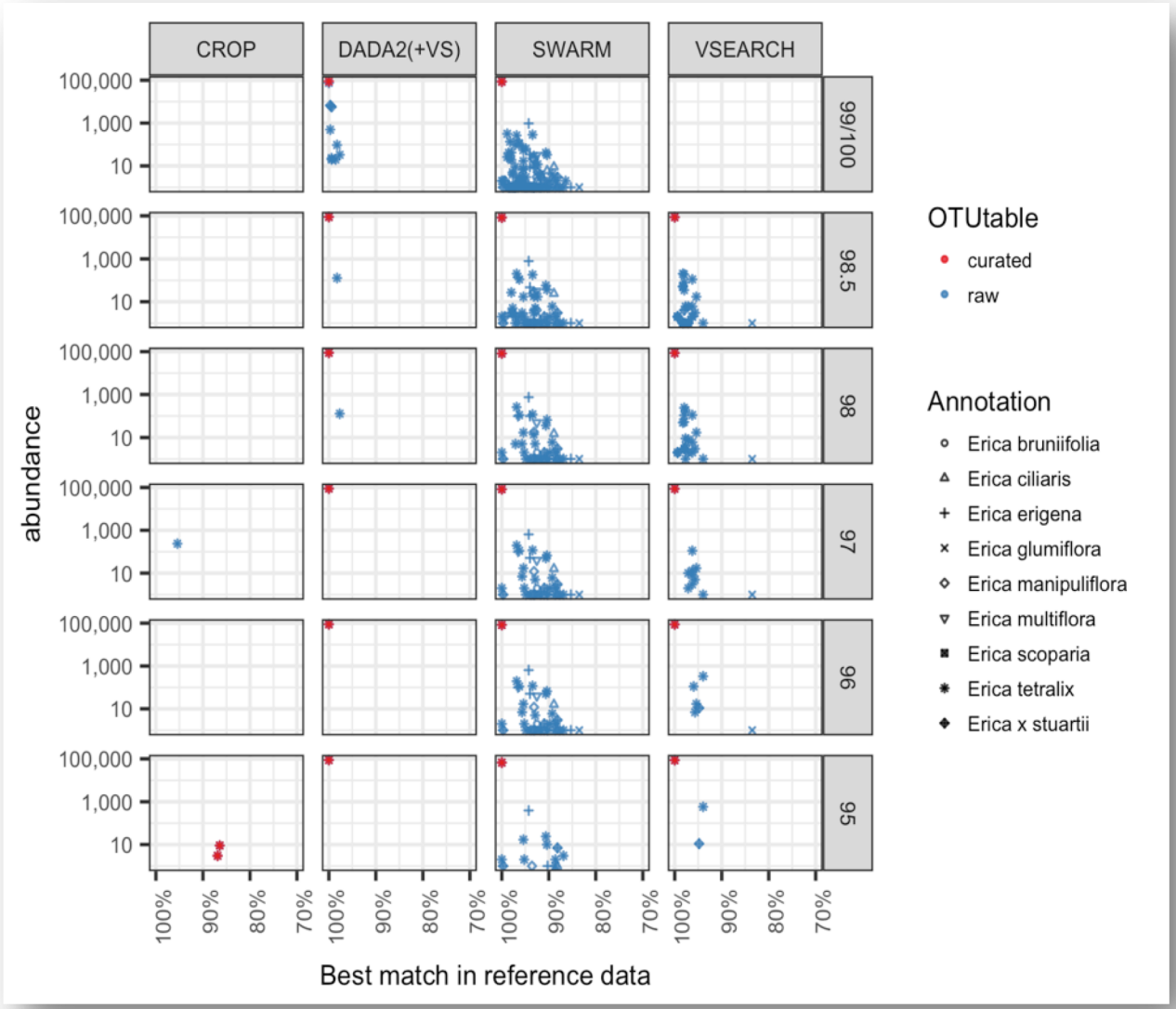
OTU β-diversity (total richness / mean site richness) compared to plant β-diversity (17.23, dashed line)

Best reference match

Distribution of best reference database (GenBank) match for OTUs retained and discarded by LULU

UiO : Department of Biosciences
University of Oslo

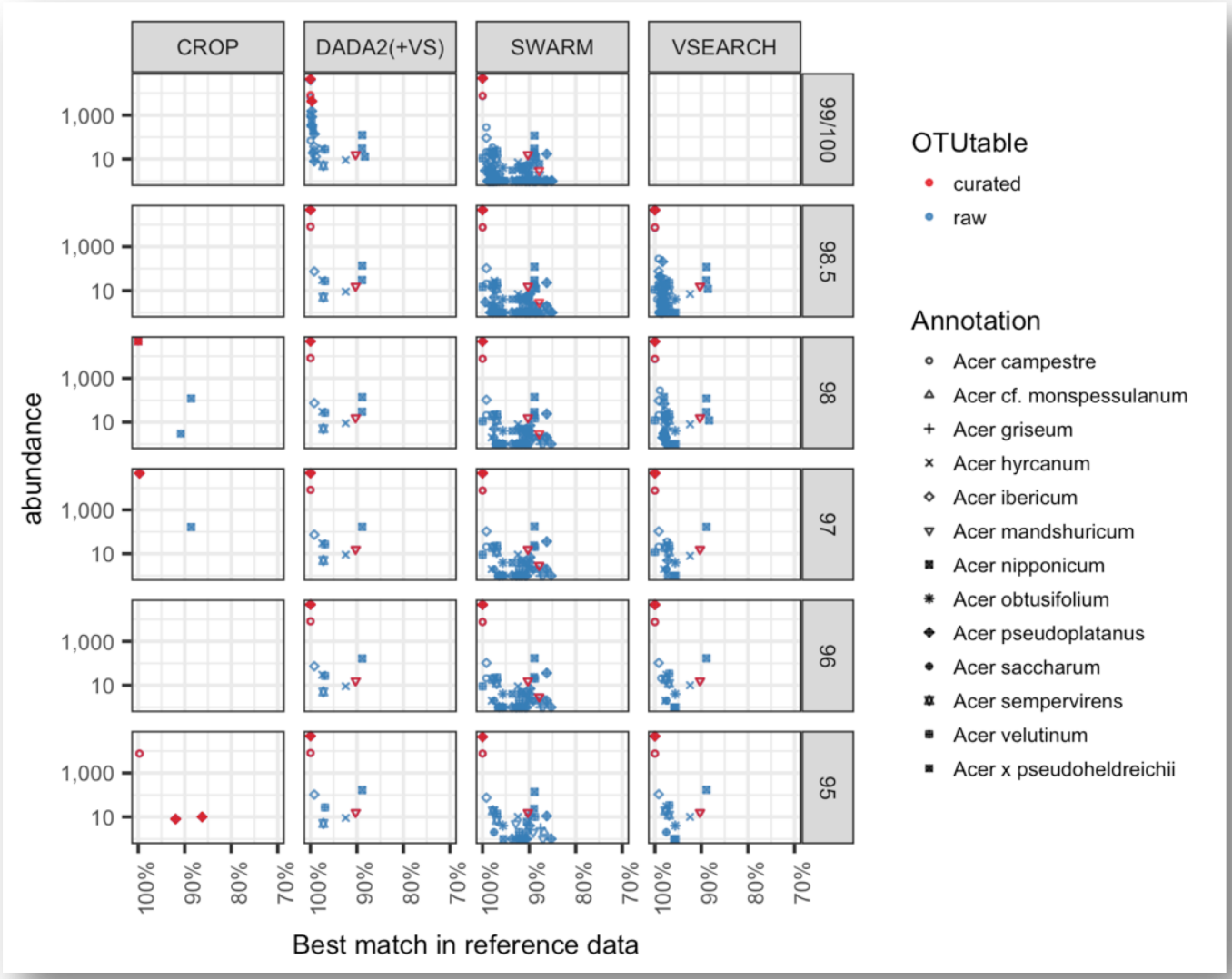
LULU curation effect - *Erica*





UiO : Department of Biosciences
University of Oslo

LULU curation effect - *Acer*









UiO : Department of Biosciences
University of Oslo

LULU –

- Improve biodiversity metrics irrespective of initial clustering algorithm – especially metrics relating to richness

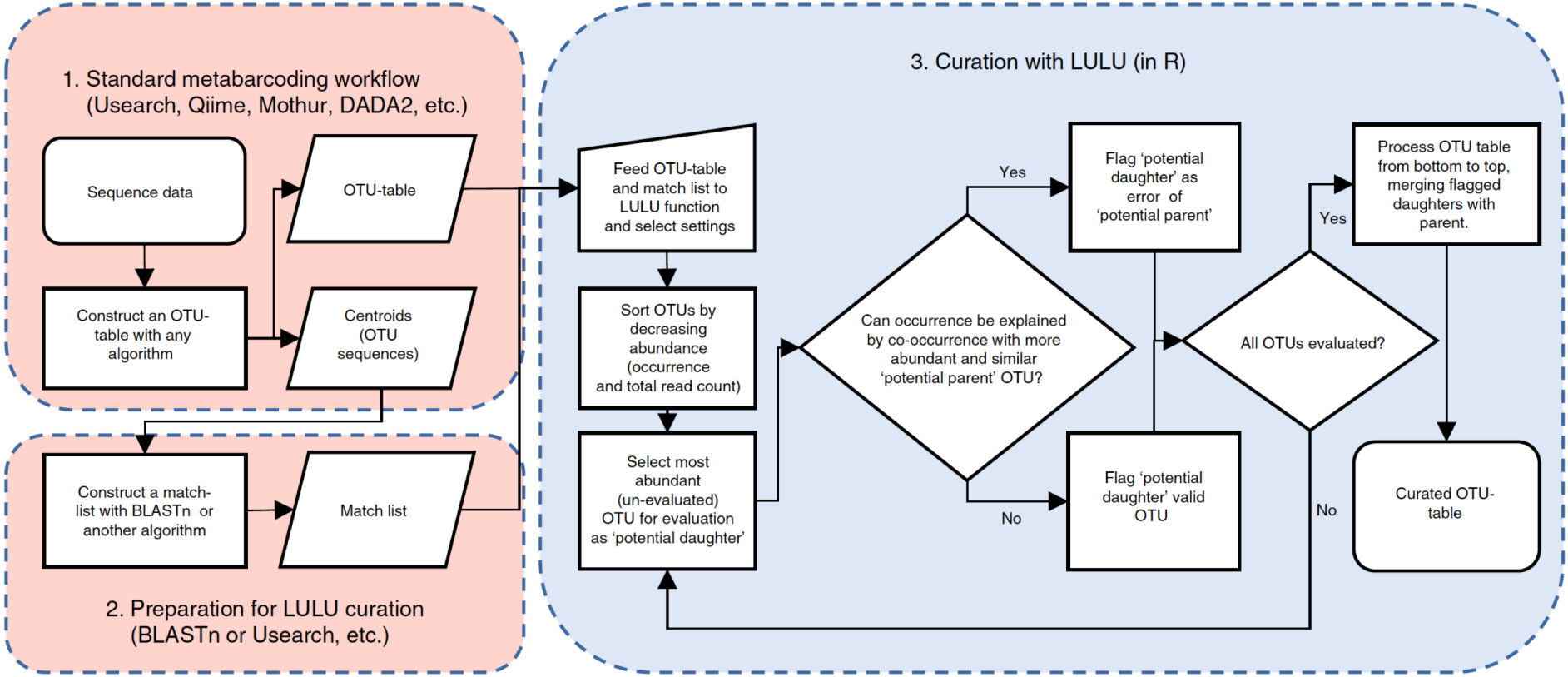
- A fast procedure - easy to learn

- A step-by-step tutorial (and test data) can be found on GitHub: https://github.com/tobiasgf/lulu

Contact: Tobias Guldberg Frøslev
Natural History Museum of Denmark
København K
tobiasgf@bio.ku.dk



LULU – work-flow



LULU – Practical exercise

1. Log-in to the abel:
qlogin --account=ln0002k --mem-per-cpu=3800M --time=05:00:00
source /cluster/software/BIO9905/set.source

2. Download the data
pwd # check the home directory
cp -r /work/tmp/BIO9905MERG1/sunilm . # Download the files
mv sunilm/ Lulu # rename the folder
cd Lulu # Change the folder (there should be two folders vsearch_pipeline and dada2_pipeline)

cd vsearch_pipeline/ # enter in the vsearch_pipeline folder
ls # Check the files, should be 5 files (vsearch_centroids.fasta Functions.R vsearch_OTU_table.txt Lulu_exercise.txt, Boreal_forest_soils_18SV4_48_samples.OTU.filtered.table.xlsx)

LULU – Practical exercise

3. Count number of sequence in fasta file:
grep -c ">" vsearch_centroids.fasta # Should be 5777 OTUs

4. Make Blast database:
module load blast+/2.6.0
makeblastdb -in vsearch_centroids.fasta -parse_seqids -dbtype nucl
ls # Check the files (blast files are generated)

5. Internal similarity search of sequences within the dataset using blastn:
blastn -db vsearch_centroids.fasta -outfmt '6 qseqid sseqid pident' -out match_list.txt -qcov_hsp_perc 80 -perc_identity 84 -query vsearch_centroids.fasta
Parameters: -qcov_hsp_perc – coverage of sequences; -perc_identity – percent identity of the sequences (Process may take 5-6 minutes)

ls # Check the files #match_list.txt is generated

LULU – Practical exercise

6. Run Lulu via R:

```
R   # load the R in Abel
# Open the file “Function.R” with WinSCP in text-editor/Notepad/R/rstudio and copy the script into the R environment of Abel (Line 1-6 first)
# The main Lulu algorithm start from Line 72 - 200 and this all need to be copied together and execute with R

curated_result$curated_count # Count OTUs left after curation #2350
curated_result$discarded_count # Count discarded OTUs #3427

# After completing the script run check the output file named “lulu_curated_otu_table.txt” download and open in excel. Compare reads and OTUs with with input OTUs file.
# Check the log file for details about OTUs similarity with each other

quit() # Close the R environment
```

LULU – Practical exercise

Repeat the analysis from Dada2 output
Navigate to the folder “Dada_pipeline”.
ls #should be 4 files (vsearch_centroids.fasta, Functions.R, dada2_OTU_table.txt, OTU table.xls)
1. Count number of sequence in fasta file:
grep -c ">" dada2_centroids.fasta # Should be 5777 OTUs

2. Make Blast database:
module load blast+/2.6.0
makeblastdb -in dada2_centroids.fasta -parse_seqids -dbtype nucl
ls # Check the files (blast files are generated)

3. Internal similarity search of sequences within the dataset using blastn:
blastn -db dada2_centroids.fasta -outfmt '6 qseqid sseqid pident' -out match_list.txt -qcov_hsp_perc 80 -perc_identity 84 -query dada2_centroids.fasta
ls # Check the files #match_list.txt is generated

LULU – Practical exercise

6. Run Lulu via R:

```
R   # load the R in Abel
# Open the file “Function.R” with WinSCP in text-editor/Notepad/R/rstudio and copy the script into the R environment of Abel (Line 1-6 first)
# The main Lulu algorithm start from Line 72 - 200 and this all need to be copied together and execute with R

curated_result$curated_count # Count OTUs left after curation #2350
curated_result$discarded_count # Count discarded OTUs #3427

# After completing the script run check the output file named “lulu_curated_otu_table.txt” download and open in excel. Compare reads and OTUs with with input OTUs file.
# Check the log file for details about OTUs similarity with each other

quit() # Close the R environment
```

LULU – Practical exercise

Compare the results from both pipeline.

How many OTUs were produced from both the pipeline after bioinformatic analysis?

How many OTUs left after Lulu curation from both the pipeline?

#If there is error (header related issue) while analyzing dada2_pipeline data, change the header in both OTU and centroid file as follows:

#Fasta file correction:

```
sed -i 's/;size=.*//' dada2_centroid.fasta
```

#OTU table correction:

```
sed -i 's/;size=.*\"/\"/' dada2_otu_table.txt
```