DNA metabarcoding

A fancy novel way to do ecological research

PD Dr. Alexander Keller
Universität Würzburg

F1 Bioinformatics

Let's download the raw data first!
(takes some minutes)

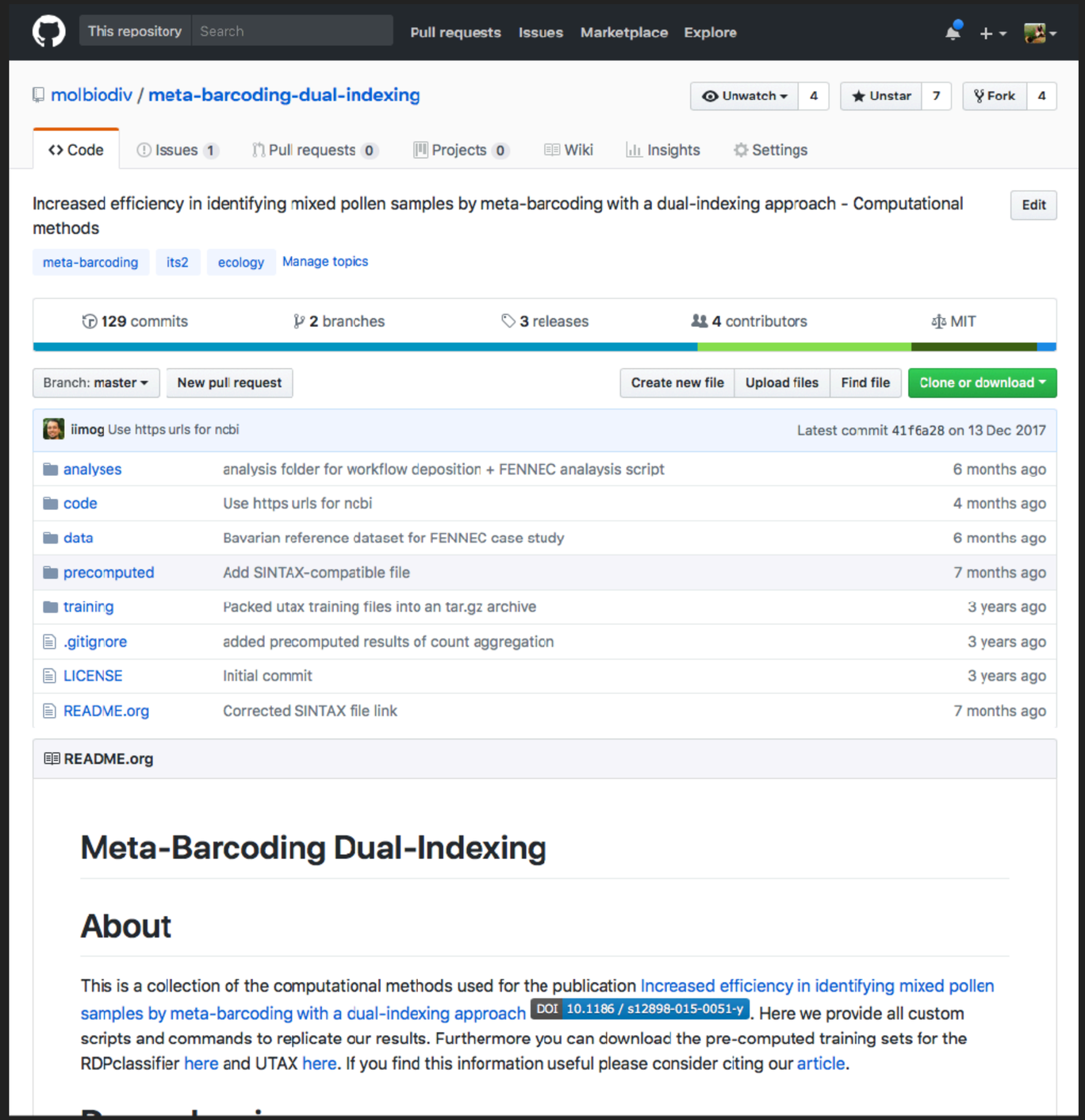
GITHUB

https://github.com/F1_MBC

1. These slides: F1_MBC.pdf
2. Command Line Script: F1_MBC.sh
3. R Script: F1_MBC.R

Based on:

https://github.com/molbiodiv/meta-barcoding-dual-indexing



ANALYSIS OF POLLEN SAMPLES

GETTING THE SEQUENCE DATA INFORMATION

```
wget http://www.ebi.ac.uk/ena/data/warehouse/filereport\?
accession\=PRJEB8640\&result\=read_run\&fields\=study_accession,secondary_study_accession,sample_
accession,secondary_sample_accession,experiment_accession,run_accession,tax_id,scientific_name,in
strument_model,library_layout,fastq_ftp,fastq_galaxy,submitted_ftp,submitted_galaxy\&download\=tx
t -O reads.tsv
```

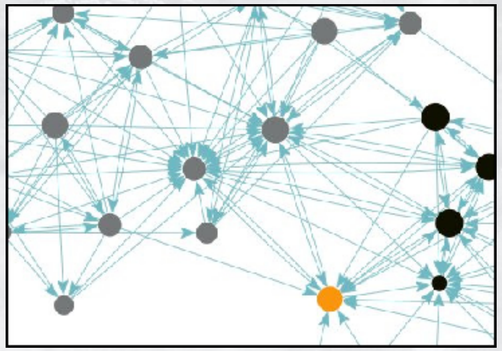
WE ONLY WANT 20 OF THESE

```
head reads.tsv > reads_subset.tsv
tail reads.tsv >> reads_subset.tsv
less -S reads_subset.tsv
#press q to exit the text reader
```

DOWNLOAD THE RAW DATA

```
for i in $(cut -f13 reads_subset.tsv | grep fastq.gz | perl -pe 's/;/\n/')
do
    wget $i
done
```

size -
morphological characters -
degradation +
taxonomic expertise -
previous knowledge -
…



Identify and understand
Biodiversity

DNA sequencing

Preparation

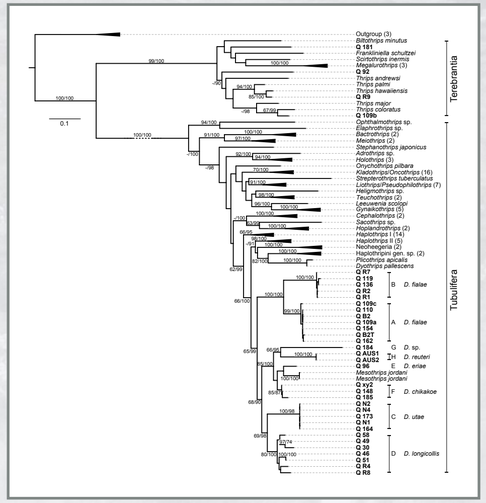
Amplification
short genomic marker

Sequencing

```
ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT
```

Sample

Phylogeny

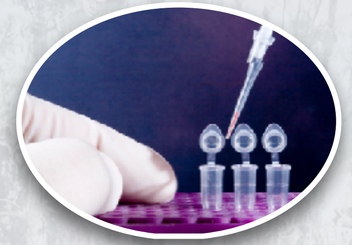


DNA-Barcoding

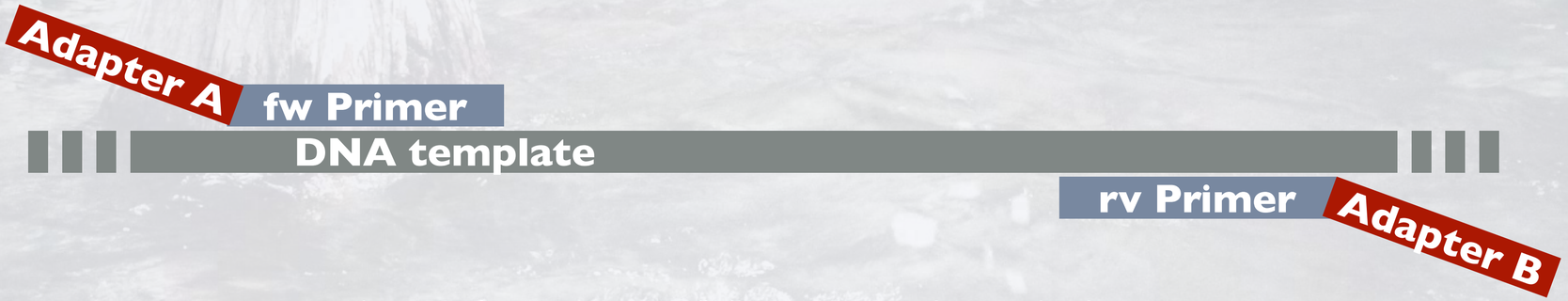
Build up/Compare with a
reference library for identification

Identity
Evolutionary History

Sequenz	ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT	<X%
Referenz 1	ACGTAGTTTT ACGTTCGGTT AGCTCGTTCC CCGTGCCGTT	>X%
Referenz 2	GCGTGGTTTT ACGTTCTGTT AGCTGGTTCC GCGTTCCGTT	



Amplification
short genomic target
Library Prep



Attachment of Sequencing-Adapter

Mixed Sample

Preparation

Amplification
short genomic target
Library Prep

Next-Generation
Sequencing

ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT
ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT
ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT
ACGTAGTTTT ACGTTCGGTT AGCTCGTTCC CCGTGCCGTT
GCGTGGTTTT ACGTTCTGTT AGCTGGTTCC GCGTTCCGTT
ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT
ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT
ACGTAGTTTT ACGTTCGGTT AGCTCGTTCC CCGTGCCGTT
GCGTGGTTTT ACGTTCTGTT AGCTGGTTCC GCGTTCCGTT
ACGTAGTTTT ACGTTCGGTT AGCTCGTTCC CCGTGCCGTT
GCGTGGTTTT ACGTTCTGTT AGCTGGTTCC GCGTTCCGTT
ACGTAGTTTT ACGTTCGGTT AGCTCGTTCC CCGTGCCGTT
ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT
ACGTAGTTTT ACGTTCGGTT AGCTCGTTCC CCGTGCCGTT
GCGTGGTTTT ACGTTCTGTT AGCTGGTTCC GCGTTCCGTT
ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT
GCGTGGTTTT ACGTTCTGTT AGCTGGTTCC GCGTTCCGTT
ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT
ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT
ACGTAGTTTT ACGTTCGGTT AGCTCGTTCC CCGTGCCGTT
GCGTGGTTTT ACGTTCTGTT AGCTGGTTCC GCGTTCCGTT
ACGTAGTTTT ACGTTCGGTT AGCTCGTTCC CCGTGCCGTT
GCGTGGTTTT ACGTTCTGTT AGCTGGTTCC GCGTTCCGTT
ACGTAGTTTT ACGTTCGGTT AGCTCGTTCC CCGTGCCGTT
•••

Preparation

Amplification
short genomic target
Library Prep

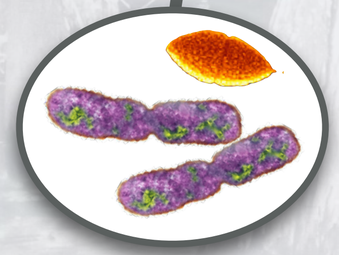
Next-Generation
Sequencing

Clustering
of similar sequences

```
2441 x ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT
  13 x ACGTAGTTTT ACGTTCGGTT AGCTCGTTCC CCGTGCCGTT
 185 x GCGTGGTTTT ACGTTCTGTT AGCTGGTTCC GCGTTCCGTT
```

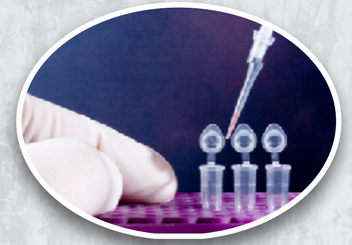
Meta-Barcoding

```
Sequenz    ACGTAGTTTT ACGTTCCGTT AGCTCGTTCC CCGTGCCGTT ]<X%
Referenz 1 ACGTAGTTTT ACGTTCGGTT AGCTCGTTCC CCGTGCCGTT ]>X%
Referenz 2 GCGTGGTTTT ACGTTCTGTT AGCTGGTTCC GCGTTCCGTT
```

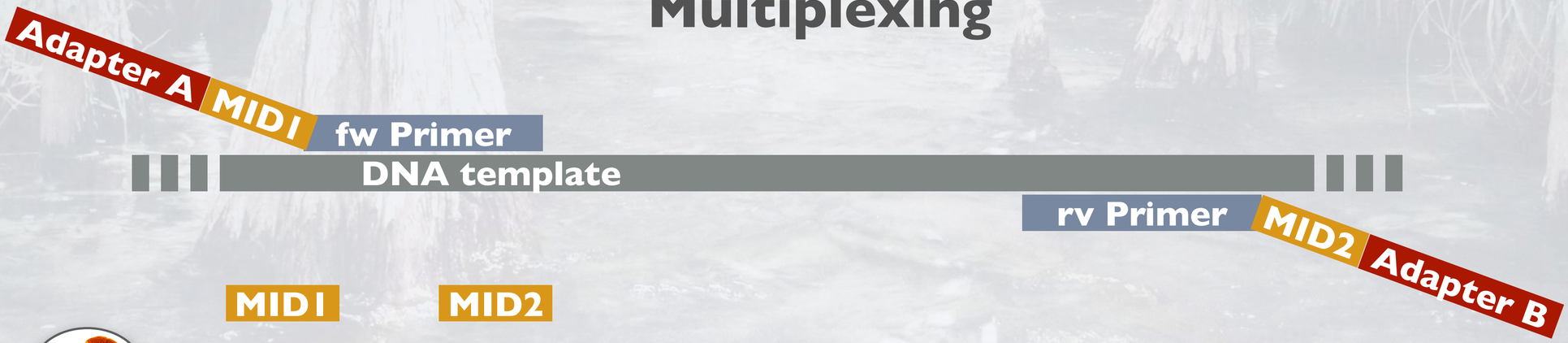


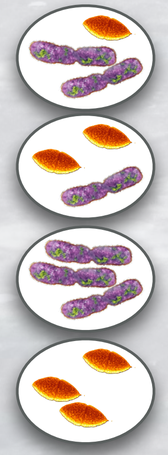
Mixed Sample

Identity
Diversity
Composition
Evolutionary History



Amplification
short genomic target
Library Prep
Multiplexing





MID1	MID2
ACGT	GCGC
ACGT	AGAT
TTAA	GCGC
TTAA	AGAT

Sample Multiplex-Identifiers

24 x 16 MID combinations = 384 mixed samples
for one sequencing chip

requires bioinformatic demultiplexing

Pollen-MB for 480 bee nests

two solitary bee species
6 months coverage
30 sampling sites (same for both)

goal: comparative dietary analysis

Osmia truncorum

Osmia bicornis

Sickel *et al. BMC Ecol (2015) 15:20*
DOI 10.1186/s12898-015-0051-y

BMC Ecology

METHODOLOGY ARTICLE Open Access

CrossMark

Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach

Wiebke Sickel, Markus J Ankenbrand, Gudrun Grimmer, Andrea Holzschuh, Stephan Härtel, Jonathan Lanzen, Ingolf Steffan-Dewenter and Alexander Keller*

CHALLENGES:

1. GETTING THE DATA ✓
2. MERGING FORWARD AND REVERSE READS
3. DATA QUALITY FILTERING
4. TAXONOMIC CLASSIFICATION
   4.1. DIRECT
   4.2. (HIERARCHIC)
5. DATA ANALYSIS
   5.1. LOADING DATA TO R
   5.2. BASIC PLOTS

X. A LOT OF REFORMATTING IN BETWEEN

2. MERGING FORWARD AND REVERSE READS
3. DATA QUALITY FILTERING

SETTING VARIABLES

```
data=$(pwd)
u9=/storage/full-share/mbd/usearch9                    # USEARCH 9.0 binary
u8=/storage/full-share/mbd/usearch8                    # USEARCH 8.0 binary
f=/storage/full-share/mbd/fastq-join/fastq-join        # FASTQ-JOIN binary
p=/storage/full-share/mbd/python_scripts               # Folder of python scripts
s=/storage/full-share/mbd/seqfilter/bin/SeqFilter      # SeqFilter binary
```

SETTING UP ANALYSIS

```
mkdir -p $data/raw
mkdir -p $data/joined

gunzip *.gz
ls -l

#store file suffixes as variables for forward and reverse reads

RF='_R1_001.fastq';
RR='_R2_001.fastq';

# get sample names
ls  $data/*$RF | sed "s/^.*\/\([a-zA-Z0-9_.-]*\)$/\1/g" | sed "s/$RF//" > samples.txt
head samples.txt
```

2. MERGING FORWARD AND REVERSE READS
3. DATA QUALITY FILTERING

SETTING UP ANALYSIS

```
for file in `cat samples.txt` ;
do
    echo "Processing >>> $file <<<";
    echo "..join ends";

    # joining forward and reverse reads
    $f $data/$file$RF $data/$file$RR -o $data/joined/$file.%.fq

    # keep R1 files for those that do not join, perhaps they are long enough and of good quality
    cat $data/joined/$file.join.fq $data/joined/$file.un1.fq > $data/joined.$file.fastq

    # move original files
    mv $data/$file$RF $data/raw/; mv $data/$file$RR $data/raw/

    # filter reads with high expected error rate, that are too short or have Ns
    echo "..filter";
    $u -fastq_filter $data/joined.$file.fastq -fastq_maxee 1 -fastq_minlen 200 -fastq_maxns 1 \
        -fastaout filter.$file.fasta
    rm $data/joined.$file.fastq

    # rename sequence names to match their original sample
    echo "..parse";
    python $p/fasta_number.py filter.$file.fasta $file. > parsed1.$file.fasta
    cat parsed1.$file.fasta | sed "s/_L001//g" | sed "s/\./_/g " > parsed2.$file.fasta
    rm filter.$file.fasta
    rm parsed1.$file.fasta

done ### a good time for questions now!
```

4. TAXONOMIC CLASSIFICATION
4.1. DIRECT

GET REFERENCE DATABASE

```
wget https://github.com/molbiodiv/meta-barcoding-dual-indexing/blob/master/data/ \
viridiplantae_bavaria_2015.fa
```

CLASSIFICATION

```
# combine files of all samples to a single file to be searched
cat parsed2.* > all.fasta

# clean up temporary files
rm -r raw/ joined/ parsed* joined.* filter.*

# convert to a barcoding format readable by usearch
cat all.fasta | sed -e "s/^>\([a-zA-Z0-9-]*\)_\(.*$\)$/>\1_\2;barcodelabel=\1/" \
   > all.bc.fasta

# find best direct hits for all filtered sequences with more than 97% identity
$u -usearch_global all.bc.fasta -db viridiplantae_bavaria_2015.fa -id 0.97 -uc  \
   output_BV3.uc -fastapairs output_BV3.fasta  -strand plus
```

4. TAXONOMIC CLASSIFICATION
4.2.(HIERARCHIC)

GET REFERENCE DATABASE

```
wget https://github.com/iimog/meta-barcoding-dual-indexing/raw/v1.1/training/utax/\
utax_trained.tar.gz

tar xzvf utax_trained.tar.gz
rm utax_trained.tar.gz

$u -makeudb_usearch utax_trained/viridiplantae_all_2014.utax.fa -output utax_trained/\
viridiplantae_all_2014.utax.udb
```

HIERARCHIC CLASSIFICATION

```
# get names and sequences of those without hits, classify them herarchical with utax
afterwards

grep "^N[[:space:]]" output_BV3.uc | cut -f 9 > output_BV3.nohit

$s all.bc.fasta --ids output_BV3.nohit --out all.bc.BV3.nohit.fasta

$u -utax all.bc.BV3.nohit.fasta -db utax_trained/viridiplantae_all_2014.utax.udb -
utax_rawscore -tt utax_trained/viridiplantae_all_2014.utax.tax -utaxout
all.bc.BV3.nohit.utax

# another good time for questions now!
```

4. TAXONOMIC CLASSIFICATION
4.2.(HIERARCHIC)

HIERARCHIC CLASSIFICATION

```
less -S all.bc.BV3.nohit.utax

# create a pseudo.uc file of the hierarchical classification and filter low qual scores
perl -ne
'($id, $tax, $sign)=split(/\t/);
@tmp=split(/,/, $tax);
@tax=();
foreach $t (@tmp){
    $t=~s/__/:/g;
    $t=~s/ +/_/g;
    $t=~s/\(([\d.]+)\)//;
    if($1 < 27){last;}
    push @tax, $t;
    $t=~/_(\d+)$/;
    $taxid=$1;
}
next if(@tax == 0);
print "H".("\t"x8)."$id\tt$taxid;tax=".join(",", @tax).";\n"
' all.bc.BV3.nohit.utax | sed "s/,s:.*$//g" > all.bc.BV3.nohit.pseudo.uc

less -S all.bc.BV3.nohit.pseudo.uc
```

X. A LOT OF REFORMATTING IN BETWEEN

PREPARING DATA FOR R

```
#combine direct hit .uc and the hierarchically classified pseudo.uc
cat output_BV3.uc all.bc.BV3.nohit.pseudo.uc > combined.uc

#convert output format to TAX/OTU-Table (community matrix including taxonomic lineages)
python $p/uc2otutab.py  combined.uc > combined.txt

less -S combined.txt

# split this into a otu table and a tax table
cat combined.txt | sed "s/;tax=.*;//g;s/:/_/g" | sed "s/;tax=[^\t]*\t/\t/g" > combined.otu
cat combined.txt | cut -f 1 | sed "s/;tax=/,/g;s/:/_/g" | sed "s/;//" | sed "s/
    OTUId/,Kingdom,Phylum,Class,Order,Family,Genus,Species/" > combined.tax

head combined.otu
head combined.tax
tail combined.tax

# create a mapping file with bee species information
cat reads_subset.tsv | cut -f8,13 | sed -e "s/ftp.sra.ebi.ac.uk.*\///" -e "s/_.*//" -e "s/\(.*\)
    \t\(.*\)/\2\t\1/" -e "s/ /_/" > MapFile.txt
sed  -i "s/scientific/#SampleID\tBeeSpecies/" MapFile.txt

less MapFile.txt
```

CHALLENGES:

1. GETTING THE DATA ✓
2. MERGING FORWARD AND REVERSE READS ✓
3. DATA QUALITY FILTERING ✓
4. TAXONOMIC CLASSIFICATION ✓
   4.1. DIRECT ✓
   4.2. (HIERARCHIC) ✓
5. DATA ANALYSIS
   5.1. LOADING DATA TO R
   5.2. BASIC PLOTS

X. A LOT OF REFORMATTING IN BETWEEN ✓

X. A LOT OF REFORMATTING IN BETWEEN

SETTING UP ENVIRONMENT

```
# load packages
library(phyloseq)
library(ggplot2)
library(vegan)

# this package is not installed on the server globally, install locally then:
install.packages("bipartite", dependencies=T, repos="https://cloud.r-project.org")
library(bipartite)
```

LOADING DATA INTO R

```
data.otu = otu_table(read.table("./combined.otu", sep="\t", header=T, row.names=1),
    taxa_are_rows=T)
data.tax = tax_table(as.matrix(read.table("./combined.tax", sep=",", header=T, row.names=1,
    fill=T)))
data.samp = import_qiime_sample_data("MapFile.txt")

# combining to a single object
dataset.comp = merge_phyloseq(data.otu, data.tax, data.samp)
```

DIVERSITY PLOTS

```
pdf(file="plot_diversity.pdf")
    plot_richness(dataset.comp, x="BeeSpecies", measures=c(„Shannon“,"Observed")) +geom_boxplot()
dev.off()
```

ORDINATION

```
ordi = ordinate(dataset.comp.rel.filter, method="NMDS","bray", k=2)

pdf(file="plot_ordination.pdf")
    plot_ordination(dataset.comp.rel.filter, ordi,color="BeeSpecies")+
        geom_point(size=6) +
        theme_bw()
dev.off()
```

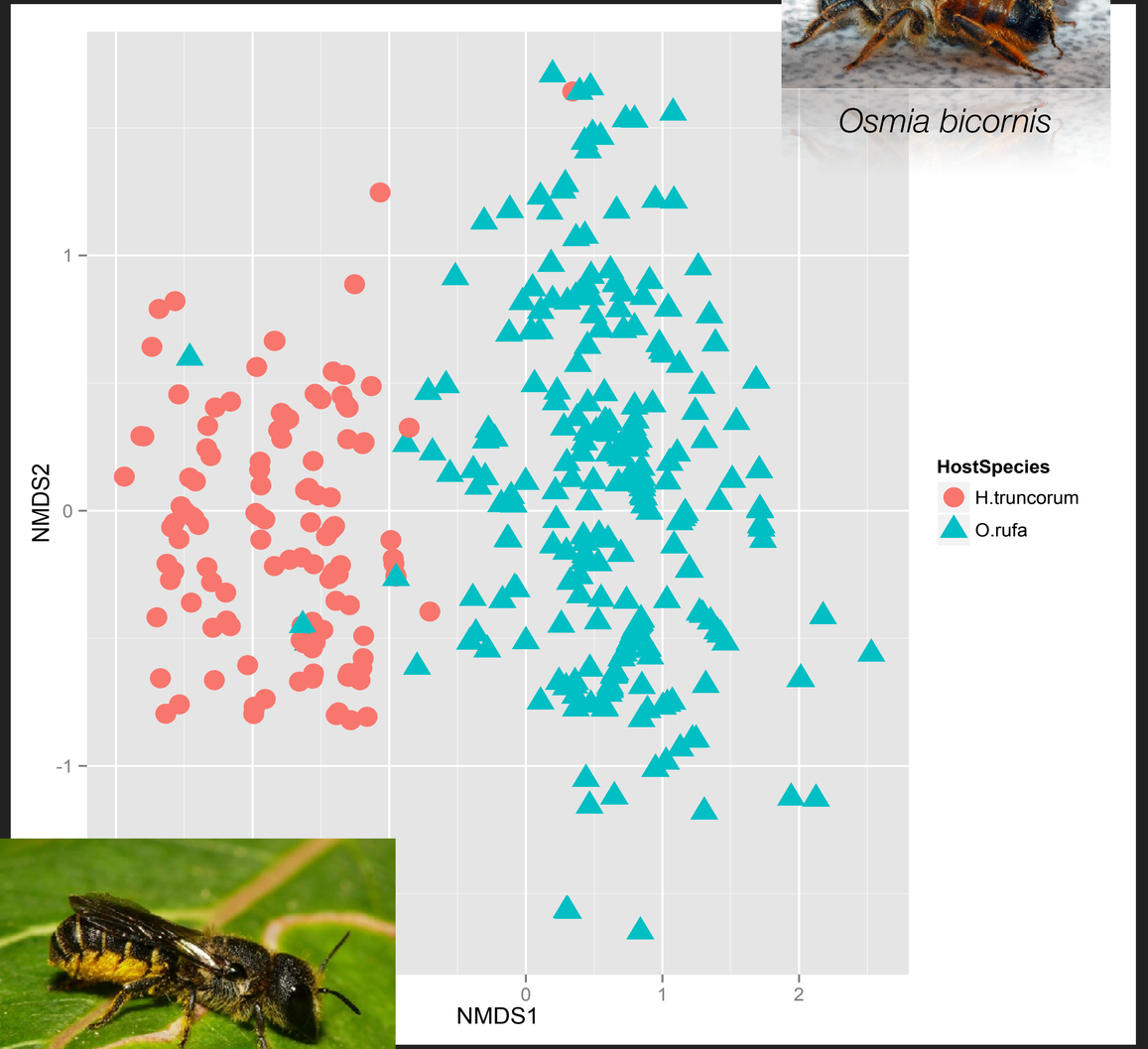
BARPLOT

```
pdf(file="plot_barplot.pdf")
    plot_bar(dataset.comp.rel.filter,x="Family", fill="BeeSpecies")
dev.off()
```

INTERACTION NETWORK

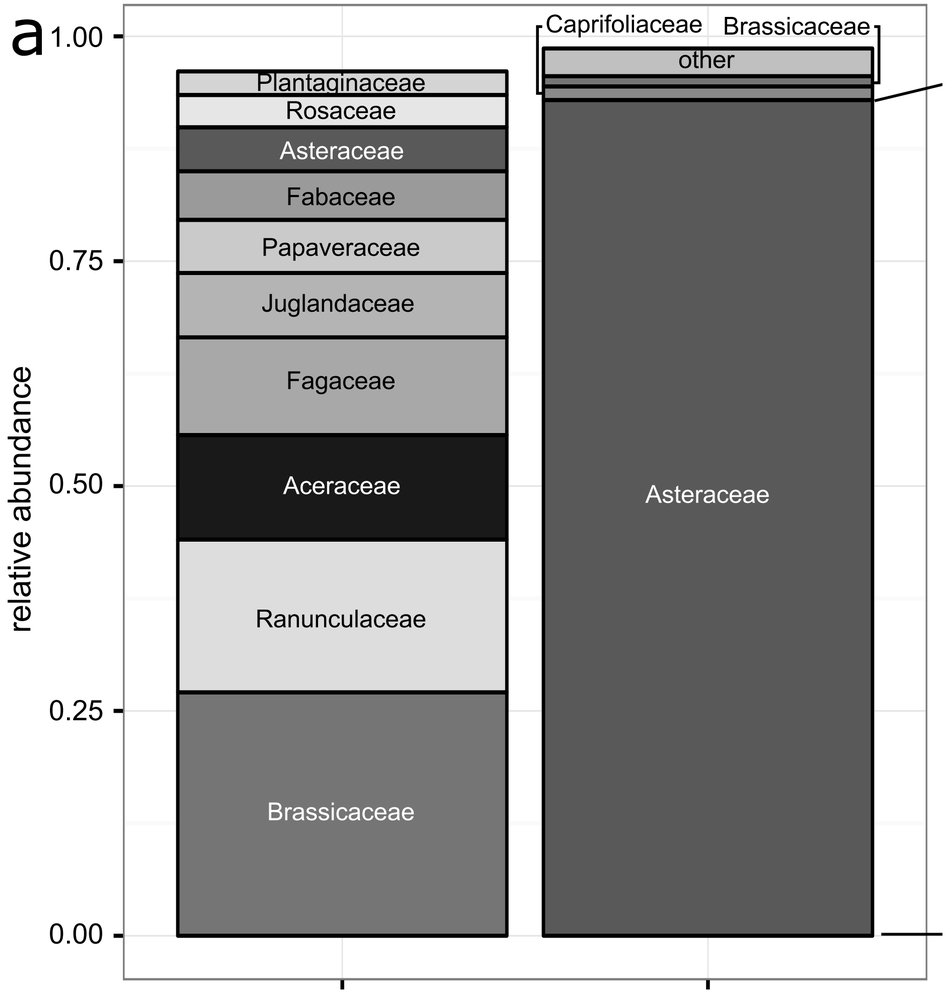
```
dataset.comp.family <- tax_glom(dataset.comp.rel.filter, taxrank="Family")
taxa_names(dataset.comp.family) <- tax_table(dataset.comp.family)[,"Family"]

dataset.bees= merge_samples(dataset.comp.family, "BeeSpecies", fun=mean)

pdf(file=„plot_network.pdf")
    plotweb(round(t(data.frame(otu_table(dataset.bees)))*100))
dev.off()
```

Osmia bicornis



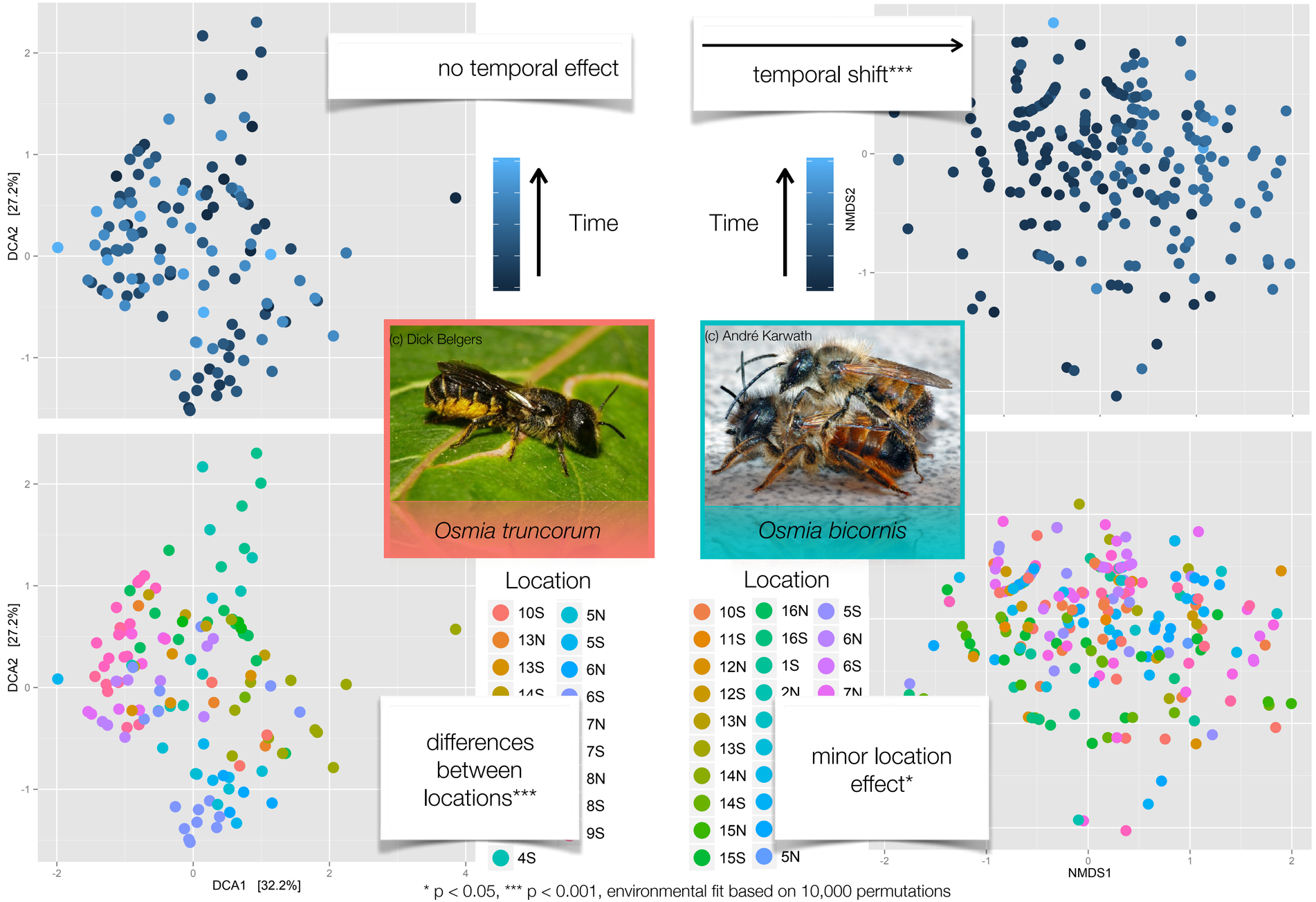
Osmia truncorum

a



Osmia bicornis

Osmia truncorum


no temporal effect
temporal shift***
Time
Time
DCA2 [27.2%]
NMDS2
(c) Dick Belgers
Osmia truncorum
(c) André Karwath
Osmia bicornis
Location
10S
13N
13S
14S
4S
5N
5S
6N
6S
7N
7S
8N
8S
9S
differences between locations***
Location
10S
11S
12N
12S
13N
13S
14N
14S
15N
15S
16N
16S
1S
2N
5N
5S
6N
6S
7N
minor location effect*
DCA1 [32.2%]
NMDS1
* p < 0.05, *** p < 0.001, environmental fit based on 10,000 permutations