Supporting material accompanying research article:

**Trace element supplementation is associated with increases in fermenting bacteria in biogas mono-digestion of grass silage.**

FitzGerald, J.A1,2., Wall, D.M 2,3., Jackson, S.A1., Murphy, J.D2,3. and Dobson, A.D.W.1,2

1 School of Microbiology, University College Cork, Ireland.

2 The MaREI Centre, Environmental Research Institute, University College Cork, Ireland

3 School of Engineering, University College Cork, Ireland

\* Corresponding author:

Email ID: a.dobson@ucc.ie

**Table of Contents**

1. Data Statement
2. Sequencing Summary & Sample Coverage

2.1 Pyrosequencing

2.2 Sequencing Coverage

1. Reagents

3.1 TENP-P Extraction Buffer

3.2 10% CTAB Buffer

1. Extraction Protocol

4.1 Preparation

4.2 Extraction

4.3 Recombining Technical Replicates

1. PCR Recipe
2. Thermocycler Recipe
3. Bioinformatics

7.1 Terminal Code

7.2 R Code

7.3 LEfSe Code

1. **Data Statement**

The sequence data from this study were submitted as a single SFF file to the ENA databases under accession number PRJEB22994 and are accessible at the URL <http://www.ebi.ac.uk/ena/data/view/PRJEB22994>.

1. **Sequencing Summary & Sample Coverage**

**1.1 Pyrosequencing**

Pyrosequencing generated 202,157 reads across the 14 samples; 120,816 reads (60%) passed quality filtering (length >350bp, q-score threshold of 25) with an average length of 417bp, providing an average of 8,940 +/-5,575 reads per sample. Sequence clustering (98% identity) and taxonomy assignment (95% identity) gave 344 genus-level OTUs (109 families, 65 orders, 51 classes, 25 phyla). 4,351 (2% of total) reads could not be resolved to known taxa and were grouped as 'No Relatives'; as these reads could not sensibly be included in any of the analyses they were removed from the dataset.

**1.2 Sequencing Coverage**

Rarefaction curves indicated the reactor communities are well described by this study’s coverage (Supporting Figure 1). Clustering through correlation analysis (CA) showed no overt segregation of samples by sequence depth (Supporting Figure 2); instead, communities showed the strongest correlation with SMY (constrained correlation analysis, CCA), demonstrating a strong correlation between biomethane production and the microbial communities, see Figure 3 A.

1. **Reagents**

**2.1 TENP-P Extraction Buffer**

1.25ml TRIS @ 1M

1ml EDTA @ 0.5M

2.5ml PVPP @ 10% w/v

12.5ml NaCl @ 5M

bring to 25ml w. TE buffer

**2.2 10% CTAB BUFFER**

0.7M NaCl

1. **Extraction Protocol**

**3.1 Preparation:**

* Anaerobic reactor sludges frozen at -80\*C were rapidly defrosted through immersion of the containers in a water bath at 50\*C. Once sufficiently defrosted, sludges were coarsely homogenised with a sterilised rod to ensure sedimentation did not affect sampling.
* For each of the 14 timepoints described in the main text, 3 technical replicates of approx. 1g wet-weight sludge biomass were removed to a 2ml ependorf tube (i.e. 42 tubes total, containing 1g biomass each). All 42 tubes were processed identically in successive batches (6 tubes per batch), using the same methodology and batch of stock reagents (see recipe).
* Note that volume constraints will produce four 1.5ml ependorf tubes for every one 2ml tube started with.
* TENP solution was made up just prior to use from pre-made stock solutions. Hot plate set to 60\*C.

**3.2 Extraction:**

1. Each sample was spun at 1000G for 1 minute to separate sample into supernatant and supernatant-and-pellet for ease of processing: equal volumes (approx. 500\*ul) were partitioned to 2ml ependorf tubes, and treated as separate samples.
2. To each sample (i.e. each 2ml tube), 800\*ul of TENP extraction buffer was added. The pipette head was used to disrupt the biomass pellet as necessary.
3. 12\*ul of proteinase K was added to each sample, mixed by inversion, and incubated at 60\*C for two hours.
4. 150\*ul of 10% w/v CTAB buffer was added, and each sample was incubated for a further 15 minutes at 60\*C.
5. Samples were spun at 1000G for two minutes, and then separated to separate 2ml tubes in equal volumes (supernatants only; supernatant and biomass) of approx 750\*ul each.
6. 800\*ul of 25:24:1 Phenol:Chloroform:Iso-amyl alcohol (PCI) was added to each tube (i.e. slightly more than equal volume of lysed sample).
7. Samples were spun at max available centrifuge speed at 0\*C for ten minutes.
8. Sample supernatant (i.e. aqueous portion) was carefully removed to clean 2ml ependorf tubes with all due deference paid to the interface layer. Sample tube with the remaining interface and organic fraction was disposed of appropriately.
9. A further 750\*ul of PCI were added to the supernatant in the fresh 2ml tube, mixed by inversion, and spun at maximum available centrifuge speed at 0\*C for 30 minutes.
10. Aqueous supernatant was carefully removed from the centrifuged tube, and pipetted to a sterile 1.5ml ependorf tube. The organic fraction was disposed of appropriately.
11. To the aqueous sample, 0.2 volumes (~140\*ul) of Na-Acetate was added, and mixed by inversion. Then, 0.8 volumes (~450\*ul) of iso-propanol was added, mixed by inversion, and spun at highest possible centrifuge speed for 100 minutes at 0\*C.
12. Supernatant was discarded, and pelleted material was cleaned by adding cold 70% volume/volume ethanol to the tube, centrifuging for 10 minutes, followed by careful decanting of the ethanol. This was repeated three times.
13. The nucleic acid pellet was air-died under the draft of an open bunsen burner for 3o minutes, after which 100\*ul of TE was added and the pellet left re-disolve. Extracts were stored at 8\*C.

**3.3 Recombining Technical Replicates:**

Each of the 14 timepoints was extracted in three technical replicates. Each replicate processed was subdivided at steps 1 and 5 above, giving 4 100\*ul extracts per replicate. For each timepoint, this produced 12 100\*ul nucleic acid extracts. Each timpeoint was pooled at 25\*ul per 100\*ul fraction (i.e. to 300\*ul total per timepoint), concentration was standardised using a spectrophotometer, and then stored at 8\*C until use. Raw extracts were frozen for eternity at -20\*C.

1. **PCR Recipe**

Component vol per 25µl reaction:

Sterile Water 15.2µl

10x Buffer 2.5µl

dNTPs 2.5µl

Forward Primer 1.25µl

Reverse Primer 1.25µl

Polymerase (Taq) 0.2µl

Template DNA 0.5µl

DMSO 1µl

BSA 0.625µl

1. **Thermocycler Program**

step temperature duration

initial denaturing 94°C 5m

33 cycles:

 denature 94°C 40s

 annealing 55°C 40s

 extenstion 72°C 1m

 final extension 72°C 5m

1. **Bioinformatics**

**7.1 Terminal code (Ubuntu 16.04 zsh shell)**

## initiate qiime:

source activate qiime

## set dir

cd /home/user/

## provide demultiplexed reads in one large fasta file

demultiplex\_fasta.py -b 10 -m '/home/user/TE\_MAP.TXT' -f '/home/user/TE\_READS.FASTA' -q '/home/user/TE\_READS.QUAL' -o ‘/home/user/DEMULT’

## manually make individual empty file for each of the 14 samples: TE1.FASTA - TE14.FASTA

## separate reads to one file per sample, using -A argument to grab sequence line following matching header

for i in $(ls TE\*); do grep $i -A 1 '/home/user/DEMULT/demultiplexed\_seqs.fna' | cat >> $i; done

## remove '\n--' artefact; this didn't work on CLI, so had to use Ctrl+H in gedit!

# for i in $(ls DW\*); do sed 's/\n--//g' $i > $i; done

## submit to ARB-SILVA

## done

**7.2 R Code**

## dw\_mk.ii images

## S T A G E D R E S S I N G

# for either Windows or BioUbuntu etc., change dir to contain the ####supplementary files###XYZ#######:

#setwd("/home/user/dw")

#setwd("C:/Users/user/dw")

## packages as necessary

#install.packages('tidyr','phyloseq','ggplot2','DESeq2','ancom.R','vegan','RColorBrewer','scales')

# # wave at the cavalry

# source('https://bioconductor.org/biocLite.R')

# biocLite()

## queue lights

library('tidyr')

library('phyloseq')

library('ggplot2')

library('DESeq2')

library('vegan')

library('RColorBrewer')

library('scales')

## 3, 2, 1:

## ##

# #

# #

## Import, shape, analyse and finally export data: ##

# #

# #

## ##

## M E T A D A T A

# mapfile contains the metadata: here take oportunity to add some more

dw\_slv\_map <- '../dw\_map.txt'

dw\_slv\_map <- import\_qiime\_sample\_data(dw\_slv\_map)

## additional data

dw\_slv\_map$Treat <- c('silage','silage','silage','silage','silage','silage','silage','silage','silage w.TE','silage w.TE','slurry-silage','slurry-silage','slurry-silage','slurry-silage')

dw\_slv\_map$TE <- c('low-TE','low-TE','low-TE','low-TE','low-TE','low-TE','low-TE','low-TE','hi-TE','hi-TE','hi-TE','hi-TE','hi-TE','hi-TE')

# Beff values calculcated from BMP values given in Wall et al., 2014

# used in mk.iii of manuscript, but changed to using SMY for CCA in mk.iv thus now deprecated

dw\_slv\_map$Beff <- dw\_slv\_map$SMY/c(400,400,400,400,400,400,400,400,400,400,348,348,348,348)

## I M P O R T S I L V A O T U

### REMAKE OTU FILE? SPLIT/ APPEND ACC#?

## import from LEfSe output, and convert to phyloseq object

#dw\_slv\_tax<-(separate(data = dw\_slv\_input, col = 'taxonomy', into = c(Domain","Phylum","Class","Order","Family","Genus","OTU"), sep = ";"))[,16:21]

# file tidied before addition to R

dw\_slv\_input<-read.table('dw\_splibbed---ssu---fingerprint----Total---sim\_95---tax\_silva---td\_20.csv',sep='\t',header=TRUE,row.names = 1)

dw\_slv\_ab <- dw\_slv\_input[,1:14]

# cut out NO RELATIVES (row 345), as cannot contribute to analysis

dw\_slv\_ab <- dw\_slv\_ab[1:344,]

dw\_slv\_ab <- otu\_table(dw\_slv\_ab,taxa\_are\_rows = TRUE)

# cut out NO RELATIVES (row 345), as cannot contribute to analysis

dw\_slv\_tax<-(dw\_slv\_input[1:344,15:21])

dw\_slv\_tax<-(dw\_slv\_input[,15:21])

dw\_slv\_tax <- as.matrix(dw\_slv\_tax,rownames.force = TRUE)

dw\_slv\_tax<-tax\_table(dw\_slv\_tax)

# set row and column names

row.names(dw\_slv\_tax) <- row.names(dw\_slv\_input)

colnames(dw\_slv\_tax) <- c('Domain','Phylum','Class','Order','Family','Genus','OTU')

dw\_slv<-merge\_phyloseq(dw\_slv\_ab,dw\_slv\_tax,dw\_slv\_map)

# add sequencing depth metadata via phyloseq

sample\_data(dw\_slv)[,'SeqDepth']<-sample\_sums(dw\_slv)

# view phyloseq object

dw\_slv

## T I D Y U P O T U s

# tidy MBA03 OTU

tax\_table(dw\_slv)['silva4\_253',] <-c('Bacteria','Firmicutes','Clostridia','MBA03','MBA03 F.','MBA03 G.','silva4\_253')

# tidy vadinBC27 OTU

tax\_table(dw\_slv)['silva4\_051',] <-c('Bacteria','Bacteroidetes','Bacteroidia','Bacteroidales','Rikenellaceae','vadinBC27 group','silva4\_051')

# Abbreviate accession from Ruminococcaceae UCG-012 as clunks display

tax\_table(dw\_slv)['silva4\_236',] <-c('Bacteria','Firmicutes','Clostridia','Clostridiales','Ruminococcaceae','Rum.\'aceae UCG-012','silva4\_236')

# Abbreviate accession from Christensenellaceae R-7 as clunks display

tax\_table(dw\_slv)['silva4\_113',] <-c('Bacteria','Firmicutes','Clostridia','Clostridiales','Christensenellaceae','Chri.\'aceae R-7 group','silva4\_236')

## T R A N S F O R M - R E L A T I V E A B U N D A N C E

# rel.bun

dw\_slv\_rb <- transform\_sample\_counts(dw\_slv, function(x)x/sum(x))

## S U B S E T

## subset OTUs for display purposes:

## agglomerate OTUs with which have less than 2% relative abundance as they cannot be displayed

# arbitrary rel.abundance cutoff: >2%, in 23% of samples

prun.dw\_s.02.23 = genefilter\_sample(dw\_slv\_rb, filterfun\_sample(function(x) x >=0.02), A=2) #0.23\*nsamples(dw\_slv\_rb))

dw\_s.02.23 = prune\_taxa(prun.dw\_s.02.23, dw\_slv\_rb)

dw\_sun.02.23 = prune\_taxa(!prun.dw\_s.02.23, dw\_slv\_rb)

# flatten all pruned families for plotting

tax\_table(dw\_sun.02.23)[,'Family'] <- c('F:Misc & NA')

dw\_s.sun <- merge\_phyloseq(dw\_s.02.23,dw\_sun.02.23)

## P L O T T I N G

pdf('dw\_trace\_element\_supplements.pdf',width=12, height=9)

## P R O C E S S - note change in order of samples to Timepoint

z.rm <- sample\_data(dw\_slv)[with(sample\_data(dw\_slv),order(Timepoint)),]

z.rm[,17:19]<-NULL

z.rm[,1:5]<-NULL

z.rm[,'TE']<-NULL

z.rm[,'w\_teor']<-NULL

z.rm[,'NH3']<-NULL

z.rm[,'CH4\_pc']<-NULL

z.rm[,'Timepoint']<-NULL

z.rm[,'Week']<-NULL

plot.ts(z.rm[1:10,1:8],main='Process for G samples')

plot.ts(z.rm[10:14,1:8],main='Process for SG samples')

## S U M M A R Y S T A T S ##

# diversity metrics:

plot\_richness(dw\_slv, x = "Treat", color = "SeqDepth", shape = 'Reactor')

# corrleation analysis of sequence depth by community structure:

plot\_ordination(dw\_slv,ordinate(dw\_slv,'CCA'),color='SeqDepth',shape='Treat',title='CA of dw showing sequencing depth',label='Timepoint',type='biplot')

# raw read abundances, coloured by Phylum

plot\_bar(dw\_slv,'Timepoint',fill='Phylum',title='Sample Absolute Read Abundance') + facet\_grid(~Treat,scales='free\_x',space='free\_x')

# relative abundances, coloured by Order

plot\_bar(dw\_s.sun ,'Timepoint',fill='Family',title='Relative Read Abundance (taxa <2% grouped)') + facet\_grid(~Treat,scales='free\_x',space='free\_x')

## A B U N D A N C E S

f2.a <- psmelt(dw\_s.02.23)

f2.b <- psmelt(dw\_sun.02.23)

#plot the agglomerated OTUs for reference of phylum abundance

## pool all non->2% to dummy phylum, append to bottom of melted frame

f2.b[,27] <- c('Phyla <2%')

f2.b[,28] <- c('Classes <2%')

f2.b[,29] <- c('Orders <2%')

f2.b[,30] <- c('Families <2%')

f2.b[,31] <- c('Genera <2%')

f2.c <- rbind(f2.a, f2.b)

# make abundance a percentage rather than a proportion

f2.d <- f2.c ; f2.d$Abundance <- (f2.d$Abundance)\*100

# improve colours

z.col <- colors(distinct=TRUE)[1:135] ; z.col <- c(z.col,colors(distinct=TRUE)[233:501])

z.col <- sample(z.col,40) # pick 40 random colours

show\_col(z.col)

z.col42 = c("#781156","#A51876","#D21E96","#E43FAD","#EA6CC0","#F098D3", # each block of 6 = a hue

 "#114578","#185EA5","#1E78D2","#3F91E4","#6CABEA","#98C4F0",

 "#117878","#18A5A5","#3FE4E4","#6CEAEA","#98F0F0","#117845",

 "#18A55E","#1ED278","#3FE491","#6CEAAB","#98F0C4","#787811",

 "#A5A518","#D2D21E","#E4E43F","#EAEA6C","#F0F098","#F7F7C5",

 "#784511","#A55E18","#D2781E","#E4913F","#EAAB6C","#F0C498",

 "#781122","#A5182F","#D21E2C","#E43F5B","#EA6C81","#F098A7")

show\_col(z.col42)

z.col12 <- c('#A51876','#E43FAD','#1E78D2','#094582','#117878','#3FE4E4',

 '#18A55E','#3FE491','#b8f9d3','#D2D21E','#D2781E','#E43F5B',

 'grey50')

show\_col((z.col12))

f2 <-

 ggplot(f2.c, aes(x=Timepoint,y=Abundance,fill=Genus)) +

 geom\_bar(aes(fill=Genus), stat ="identity", position="stack", colour="black") +

 facet\_grid(Family~Treat,scales='free',space='free', switch = 'y') + # on top of the rest, switch labels for y axis

 scale\_fill\_manual(values=z.col12) +

 theme(strip.text.y = element\_text(angle = 0, size =12), axis.text.y = element\_text(angle = 0, size=9)) +

 theme(strip.text.x = element\_text(size = 14), axis.text.x = element\_text(angle = 270, size=11)) +#, strip.text.y = element\_text(size = 12))# +

 #stuff from MIHUA's rmarkdown

 theme(panel.background = element\_rect(fill=NA, colour=NA)) + # rm backgrnd

 theme(panel.grid.major.y = element\_line(colour='grey75')) + # horiz lines 1

 theme(panel.grid.minor.y = element\_line(colour='grey85')) + # horiz lines 2

 theme(panel.grid.major.x = element\_blank()) + # rm vert lines

# theme(legend.position = 'none')

 ggtitle("AD of Grass Silage / Slurry: OTUs >2% grouped by Family")

f2

# abundance value overplotting can be taken care of as SVG externally (i.e. in Inkscape)!

# for reference

ggplot(f2.c, aes(x=Timepoint,y=Abundance,fill=Genus)) +

 geom\_bar(aes(fill=Genus), stat ="identity", position="stack", colour="black") +

 facet\_grid(Order~Treat,scales='free\_x',space='free\_x') +

 theme(strip.text.y = element\_text(angle = 0), axis.text.x = element\_text(angle = 270, size=11)) +

 theme(strip.text.x = element\_text(size = 14), strip.text.y = element\_text(size = 14)) +

 ggtitle("the previous version of Rel Ab testing")

## plot all the agglomerated taxa (<0.02% relative abundance), for comparison.

f2.b <- psmelt(dw\_sun.02.23) # reset f2.b

te\_agglom <- ggplot(f2.b, aes(x=Timepoint,y=Abundance,fill=Class)) +

 geom\_bar(aes(fill=Class), stat ="identity", position="dodge", colour="black") +

 facet\_grid(Phylum~Treat,scales='free\_x',space='free\_x' +

 theme(strip.text.y = element\_text(angle = 270))) + theme(strip.text.y = element\_text(angle = 0), axis.text.x = element\_text(angle = 270)) +

 theme(legend.position="none") +

 ggtitle("Relative Abundance Overview of 'Minor' OTUs (<2%)")

te\_agglom

## Ladies and Gentlemen; T H E A R C H A E A!

# everyone's favourite phylogenetic neighbour

# Archaeal community abundance

te\_arch <- psmelt(subset\_taxa(dw\_slv\_rb,Domain=='Archaea'))

te\_arch\_plot <- ggplot(te\_arch, aes(x=Timepoint,y=Abundance,fill=Genus)) +

 geom\_bar(aes(fill=Genus), stat ="identity", position="stack", colour="black") +

 facet\_grid(Family~Treat,scales='free\_x',space='free\_x') +

 theme(strip.text.y = element\_text(angle = 0), axis.text.x = element\_text(angle = 270)) +

 ggtitle("Archaeal OTUs by Family/Orders") +

 theme(strip.text.x = element\_text(size = 14), strip.text.y = element\_text(size = 14))

te\_arch\_plot

## ... \*clap\*, \*clap\*.

## D I F F E R E N T I A L A B U N D A N C E T E S T S - L E f S e

##export for testing

## tried doing this intelligently, doing it reliably instead via export to Excel

#write.table(tax\_table(dw\_slv),'dw\_slv\_tax.txt',sep='\t')

#write.table(otu\_table(dw\_slv),'dw\_slv\_otu.txt',sep='\t')

## import LEfSe output for plotting

dw\_lefse <- as.data.frame(read.table('dw\_te\_edited.txt',sep='\t',header=TRUE,row.names = 1))

#colnames(dw\_lefse) <- c('taxa','KW.RST','Setup','LDA','p.val') # already in file, ported with header/row.names above

dw\_lefse <- dw\_lefse[with (dw\_lefse, order(Taxa)),]

## table com, binned with LefSe output in excel to show the relevant stuff: many higher clade are merely

## representations of their OTU, but at a higher confidence due to bootstrapping: these do not contribute

## but do clutter - smooth out!

#dw\_lefse\_te<-read.table('dw\_noted\_otus\_tabular.txt',sep='\t',header=TRUE,row.names = 1)

dw\_lefse\_te<-read.table('dw\_te\_edited2.txt',sep='\t',header=TRUE,row.names = 1)

##

## kept getting bad orders of numeric data: FIX: sort by column A, then by coulmn B:

dw\_lefse\_te <- dw\_lefse\_te[with(dw\_lefse\_te, order(Condition, -KW.RST)), ]

## still need to stop ggplot reordering: so 'fix' the order by factorising:

dw\_lefse\_te$Taxa <- factor(dw\_lefse\_te$Taxa, levels = dw\_lefse\_te$Taxa)

ggplot(dw\_lefse\_te, aes(x=Taxa, y=KW.RST, fill= Condition, shape=Condition)) +

 theme(axis.text = element\_text(size=13), legend.text=element\_text(size=13), axis.title = element\_text(size=13)) +

 labs(title='Taxa Associations with Reactor Setup') +

 scale\_size\_continuous(name='LDA Effect Size (log10)') +

 geom\_point(aes(size=LDA)) +

 coord\_flip() +

 guides(fill=guide\_legend(title.theme=element\_text(size=21, face= 'bold', colour='green'))) +

 theme(strip.text.x = element\_text(size = 15)) + facet\_grid(~Condition) +

 labs(y='Magnitude of Difference (H score)') +

 theme(panel.background = element\_rect(fill=NA, colour=NA)) + # rm backgrnd

 theme(panel.grid.major.x = element\_line(colour='grey80')) + # horiz lines 1

 theme(panel.grid.major.y = element\_line(colour='grey80')) + # horiz lines 1

 scale\_shape\_manual(values = c(22, 21, 24)) +

 scale\_fill\_manual(values = c(brewer.pal(9,'YlOrRd')[8],brewer.pal(3,'Greens')[3],brewer.pal(3,'YlGnBu')[3])) #+

# theme(panel.grid.minor.y = element\_line(colour='grey85')) #+ # horiz lines 2

# theme(panel.grid.major.x = element\_blank()) # rm vert lines

## V E G A N C C A P L O T

# alteration: use SMY rather than Beff

# vegan friendly map: DF

cca\_map <- as.data.frame(as.matrix(dw\_slv\_map))

# fix SMY values as numeric values for CCA

cca\_map$SMY<-as.numeric(as.character(cca\_map$SMY))

# fix order of rows for map, as sample #04 is timepoint #12 due to barcode use (not important, just a little awkward)

cca\_map<-cca\_map[order(rownames(cca\_map)),]

# Weeks as samplenames for clarity

rownames(cca\_map)<-cca\_map$Timepoint

##from OTU\_table to sorted, coloured plot:

veg.3up=as.data.frame(t(otu\_table(dw\_slv))) # 14 X 344 df of the OTU abundances from phyloseq

rownames(veg.3up)<-cca\_map$Timepoint # match the rownames in cca\_map

#veg.5.Fam=c(tax\_table(dw\_slv)[,5]) # the Fam slot data for dw\_slv

# names of OTUs significant in LEfSe are changed to a dummy value for LDA score bins (a-h), so can be plotted by vegan.

# colouring by veg.rassoclev below plots that category as a colour in conjunction with veg.9cols

# non-significant OTUs simply relabelled 'unassoc. taxa'

veg.rassoc<-read.table('dw\_noted\_otus.txt',sep='\t',header=FALSE,row.names = 1)

# transpose so maches otu table (OTUs as col), to DF, and use to rename OTU table columns::

veg.rassoc<-as.data.frame(t(veg.rassoc))

veg.3up2 <- as.matrix(veg.3up)

colnames(veg.3up2) <-as.matrix(veg.rassoc[1,1:344]) # couldnt make this work any other way

veg.3up <- as.data.frame(veg.3up2)

# vector to carry association levels as factors:

veg.rassoclev <-c('R-G: LDA ≦3.25',

 'R-G-TE: LDA ≦3.25',

 'R-G-TE: LDA 3.26-3.5',

 'R-G-TE: LDA 3.6-4.0',

 'R-G-TE: LDA >4.0',

 'R-SG: LDA ≦3.25',

 'R-SG: LDA 3.26-3.5',

 'R-SG: LDA 3.6-4.0',

 'unassoc. taxa')

# setup CCA object

veg.cca3c<-vegan::cca(veg.3up~SMY,cca\_map)

# set plotting scale

veg.scl<-c(-2)

# the colour and the shape

# make three sets; green red blue for setups

# 14 colours for 14 timepoints

veg.14cols<-c(brewer.pal(9,'YlOrRd')[2:9],brewer.pal(7,'Greens')[6:7],brewer.pal(9,'YlGnBu')[5:9])

veg.chars <- c(22, 21, 24)

veg.2col <- c('rosybrown1','lightblue') # for ordiellipse bg

# OTUs

v.blues <- brewer.pal(9,'Blues')

v.greens <- brewer.pal(9,'Greens')

veg.9cols <- c('maroon')

veg.9cols<-c(veg.9cols,v.greens[7:9],v.greens[4])

veg.9cols <- c(veg.9cols,v.blues[7:8],v.blues[5])

veg.9cols <- c(veg.9cols,'gray85')

#show\_col(veg.9cols)

# plot that stuff out!

veg.phy <- plot(veg.cca3c, type='n', display='species', scaling=veg.scl, main=' ') # plot empty

with (cca\_map, ordiellipse(veg.cca3c, Reactor, col=veg.2col, draw='polygon', kind='sd', label=TRUE, cex=1.2, alpha=0)) # SD ellipse

with (veg.3up, points(veg.cca3c, display='spec', col = 'gray30', bg=veg.9cols[as.factor(colnames(veg.3up))], pch=21, cex=0.9)) # colour by fav families

with (cca\_map, points(veg.cca3c, display='sites', col='black', bg=veg.14cols[cca\_map$Timepoint], pch=veg.chars[as.factor(cca\_map$Treat)],cex=2)) # reactor points and colours

with(cca\_map,text(veg.cca3c,display='bp',col='black',cex=1.4))

with(veg.phy, legend("topleft", legend = levels(as.factor(veg.rassoclev)), bty = "n", col = 'gray30', pch = 21, pt.bg = veg.9cols, cex=1.2, pt.cex=1))

with(cca\_map, legend("topright", legend = levels(as.factor(dw\_slv\_map$Timepoint)), bty = "n", col = 'gray30', pch = veg.chars[as.factor(dw\_slv\_map$Treat)], pt.bg = veg.14cols, cex=1.2, pt.cex=1.4))

# end workflow

dev.off()

## ##

# #

# #

## Specifically recreate the paper figures: ##

# #

# #

## ##

# FIGURE 1A, 1B

# Created in excel, not reproduced here. Data supplied as tabular table in supporting info.

# FIGURE 2

f2 <-

 ggplot(f2.c, aes(x=Timepoint,y=Abundance,fill=Genus)) +

 geom\_bar(aes(fill=Genus), stat ="identity", position="stack", colour="black") +

 facet\_grid(Family~Treat,scales='free',space='free', switch = 'y') + # on top of the rest, switch labels for y axis

 theme(strip.text.y = element\_text(angle = 0, size =12), axis.text.y = element\_text(angle = 0, size=11)) +

 theme(strip.text.x = element\_text(size = 14), axis.text.x = element\_text(angle = 0, size=11)) +#, strip.text.y = element\_text(size = 12))# +

 #stuff from MIHUA's rmarkdown

 theme(panel.background = element\_rect(fill=NA, colour=NA)) + # rm backgrnd

 theme(panel.grid.major.y = element\_line(colour='grey75')) + # horiz lines 1

 theme(panel.grid.minor.y = element\_line(colour='grey85')) + # horiz lines 2

 theme(panel.grid.major.x = element\_blank()) + # rm vert lines

 # theme(legend.position = 'none')

 ggtitle("AD of Grass Silage / Slurry: OTUs with Rel.Abundances above 2% by Genus/Orders")

f2

# FIGURE 3a

ggplot(dw\_lefse\_te, aes(x=Taxa, y=KW.RST, fill= Condition, shape=Condition)) +

 theme(axis.text = element\_text(size=13), legend.text=element\_text(size=13), axis.title = element\_text(size=13)) +

 labs(title='Taxa Associations with Reactor Setup') +

 scale\_size\_continuous(name='LDA Effect Size (log10)') +

 geom\_point(aes(size=LDA)) +

 coord\_flip() +

 guides(fill=guide\_legend(title.theme=element\_text(size=21, face= 'bold', colour='green'))) +

 theme(strip.text.x = element\_text(size = 15)) + facet\_grid(~Condition) +

 labs(y='Magnitude of Difference (H score)') +

 theme(panel.background = element\_rect(fill=NA, colour=NA)) + # rm backgrnd

 theme(panel.grid.major.x = element\_line(colour='grey80')) + # horiz lines 1

 theme(panel.grid.major.y = element\_line(colour='grey80')) + # horiz lines 1

 scale\_shape\_manual(values = c(22, 21, 24)) +

 scale\_fill\_manual(values = c(brewer.pal(9,'YlOrRd')[8],brewer.pal(3,'Greens')[3],brewer.pal(3,'YlGnBu')[3])) #+

# FIGURE 3B

# ordihull alpha set to zero!

veg.phy <- plot(veg.cca3c, type='n', display='species', scaling=veg.scl, main=' ') # plot empty

with (cca\_map, ordiellipse(veg.cca3c, Reactor, col=veg.2col, draw='polygon', kind='sd', label=TRUE, cex=1.2, alpha=0)) # SD ellipse

with (veg.3up, points(veg.cca3c, display='spec', col = 'gray30', bg=veg.9cols[as.factor(colnames(veg.3up))], pch=21, cex=0.9)) # colour by fav families

with (cca\_map, points(veg.cca3c, display='sites', col='black', bg=veg.14cols[cca\_map$Timepoint], pch=veg.chars[as.factor(cca\_map$Treat)],cex=2)) # reactor points and colours

with(cca\_map,text(veg.cca3c,display='bp',col='black',cex=1.4))

with(veg.phy, legend("topleft", legend = levels(as.factor(veg.rassoclev)), bty = "n", col = 'gray30', pch = 21, pt.bg = veg.9cols, cex=1.2, pt.cex=1))

with(cca\_map, legend("topright", legend = levels(as.factor(dw\_slv\_map$Timepoint)), bty = "n", col = 'gray30', pch = veg.chars[as.factor(dw\_slv\_map$Treat)], pt.bg = veg.14cols, cex=1.2, pt.cex=1.4))

## ##

# #

# #

## Specifically recreate Supporting figures: ##

# #

# #

## ##

## SF1: Rarefaction curve: produced by silva, not in R.

## SF2: CA with respect to Sequencing Depth

plot\_ordination(dw\_slv,ordinate(dw\_slv,'CCA'),color='SeqDepth',shape='Treat',title='CA of dw showing sequencing depth',label='Timepoint',type='biplot')

## SF3: Archaeal Community Abundance

te\_arch <- psmelt(subset\_taxa(dw\_slv\_rb,Domain=='Archaea'))

te\_arch\_plot <- ggplot(te\_arch, aes(x=Timepoint,y=Abundance,fill=Genus)) +

 geom\_bar(aes(fill=Genus), stat ="identity", position="stack", colour="black") +

 facet\_grid(Family~Treat,scales='free\_x',space='free\_x') +

 theme(strip.text.y = element\_text(angle = 0), axis.text.x = element\_text(angle = 270)) +

 ggtitle("Archaeal OTUs by Family/Orders") +

 theme(strip.text.x = element\_text(size = 14), strip.text.y = element\_text(size = 14))

te\_arch\_plot

## un-SF, unmentioned: Diversity indices, coloured by Sequencing depth

#plot\_richness(dw\_slv, x = "Treat", color = "SeqDepth", shape = 'Reactor') #+ geom\_boxplot()

**7.3 LEfSe Code**

cd /home/jfg/BioApps/LEfSe/

mkdir DW\_SILVA\_run

cd DW\_SILVA\_run

# c= main class, s= sub-class, u=subject. o = value to standardise all abundances to

python2.7 /home/jfg/BioApps/LEfSe/format\_input.py /home/jfg/Desktop/lefseinput.txt FORMATTED.in -c 1 -s 2 -u 3 -o 1000000

# l = LDA cutoff

python2.7 /home/jfg/BioApps/LEfSe/run\_lefse.py /home/jfg/BioApps/LEfSe/DW\_SILVA\_run/FORMATTED.in /home/jfg/BioApps/LEfSe/DW\_SILVA\_run/RESULT.res -l 3.0

python2.7 /home/jfg/BioApps/LEfSe/plot\_res\_mod.py /home/jfg/BioApps/LEfSe/DW\_SILVA\_run/RESULT.res /home/jfg/BioApps/LEfSe/DW\_SILVA\_run/IMAGE.png --dpi 300

python2.7 /home/jfg/BioApps/LEfSe/plot\_cladogram\_mod.py /home/jfg/BioApps/LEfSe/DW\_SILVA\_run/RESULT.res /home/jfg/BioApps/LEfSe/DW\_SILVA\_run/RESULT.IMAGE.png --format png --dpi 300

mkdir biomarkers\_images

python2.7 /home/jfg/BioApps/LEfSe/plot\_features\_mod.py /home/jfg/BioApps/LEfSe/DW\_SILVA\_run/FORMATTED.in /home/jfg/BioApps/LEfSe/DW\_SILVA\_run/RESULT.res biomarkers\_raw\_images/

echo '\n\n n o w s c r e a m d o w n t h e h o u s e '