Todd Lane

Kunal Poorey

Kunal Poorey

Systems Biology Department
Sandia National Labs

Capturing the pond crash signature for algal cultivation optimization

Supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Why Algae ?

- Food
  - Spirulina, Chlorella, Irish moss, Sea lettuce etc.
  - Algae oil (Omega-3 and Omega-6)
- Fertilizer and agar
- Pollution control
  - Waste water treatment, CO_2 reduction
- Energy production
  - biodiesel, biobutanol, biogasoline, methane, ethanol
  - Fast growth, does not compete with agriculture
- Other usage
  - cosmetics animal feed, bioplastic, pharmaceutical ingredients
- Job Creation





Hurdles in Algae Cultivation

- Short term areal production of 30-50 g/m^2/day is high

BUT

- Annualized areal production rates of 13.2 g/m^2/day:
  ANL, NREL, PNNL 2012

- We need to be higher to meet Economic threshold: 25-50 g/m^2/day annualized

- Conditions in production ponds are not found in nature

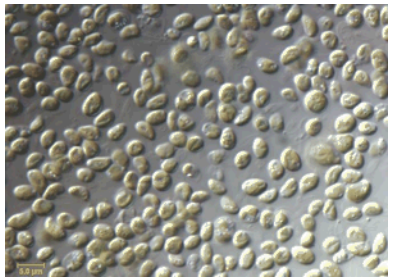
- Pond crashes responsible for loss of 10-30% of annualized production

- The reasons for failure are poorly understood
  - Low resources and low technical sophistication

We need : High productivity, resilient algae culture and, low operation cost

Algae crash agents

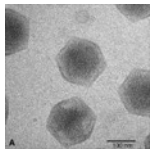
Algae



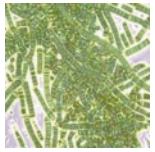
+

Agent

Viruses

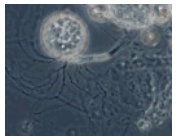


Bacteria



Predators

Fungi



+

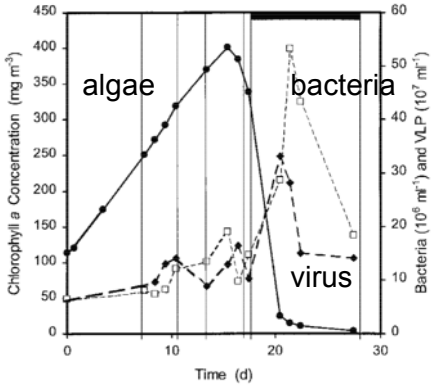
Environment

Environment (Temp, salinity, pH, CO2, nutrients)

=

Collapse

CRASH



Herman Gons et al., Antonie van Leeuwenhoek, 81: 319-326, 2002.

Patterson & Laderman, 2001.

- Adapted from Todd Lane

"Perhaps the most worrisome component of the large-scale algal cultivation enterprise is the fact that algal predators and pathogens are both pervasive and little understood."

- DOE Draft Algal Biofuels Technology Roadmap (2009)

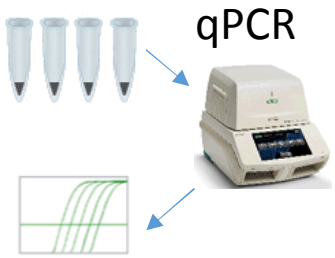
Detection of biological agents

Goal : For sustainable reliable production / cultivation of algae.

- Find probiotic communities and deleterious species affecting biomass yield
- New methods to detect deleterious species

Current detection methods: Microscopy and qPCR sampling

- Disadvantages
  - Prior expertise needed
  - Sampling bias in identifications



Sequencing techniques for detection of biological agents

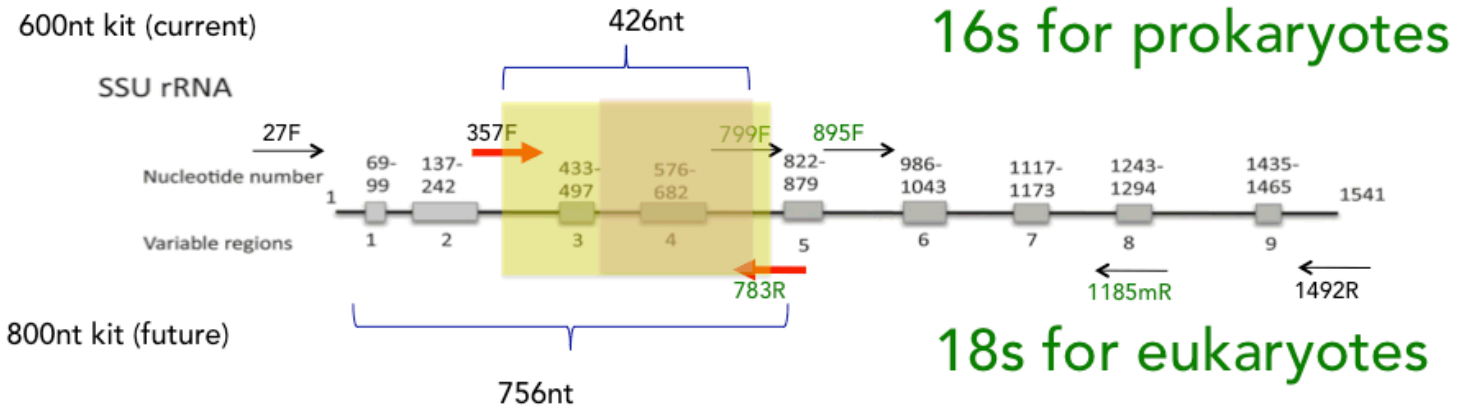
Advantages

Exploiting the diversity in ssu rRNA gene with DNA sequencing for species classification

- Advantages:
  - No prior knowledge needed to classify.
  - No sample based bias
  - Time saver
  - Able to classify a larger variety of organisms

Phylogenetic “Tree of Life” for SSU rRNA

Entries in Silva SSU database
597,607

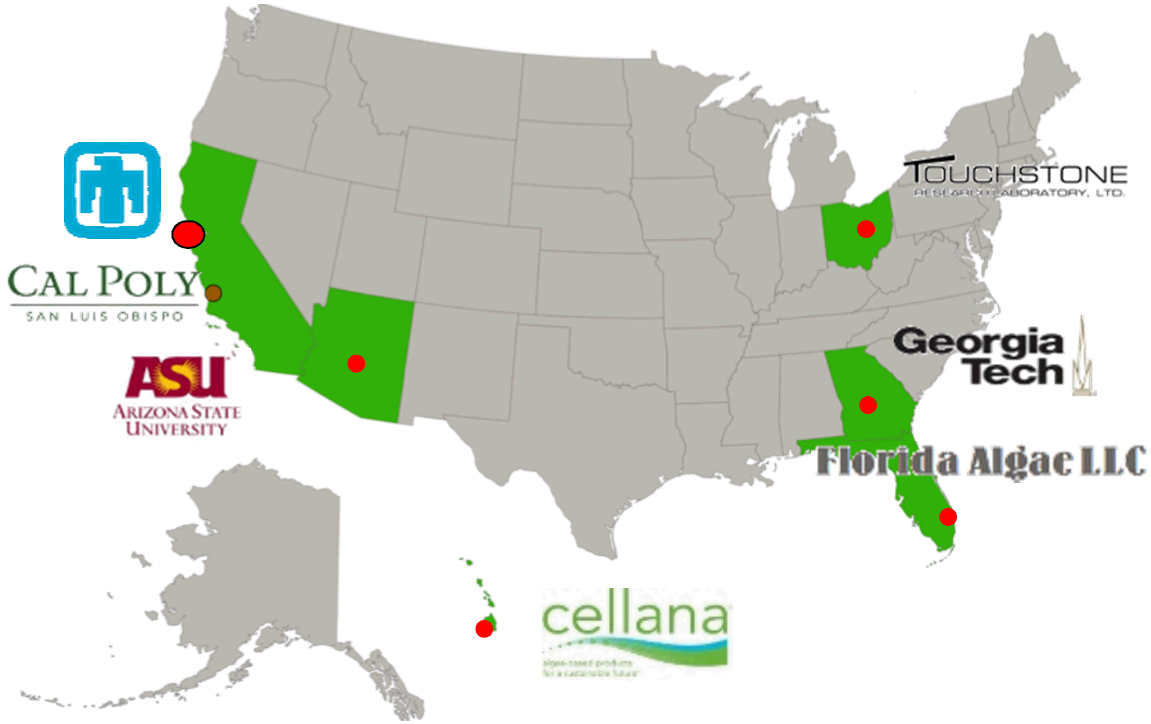


16s for prokaryotes

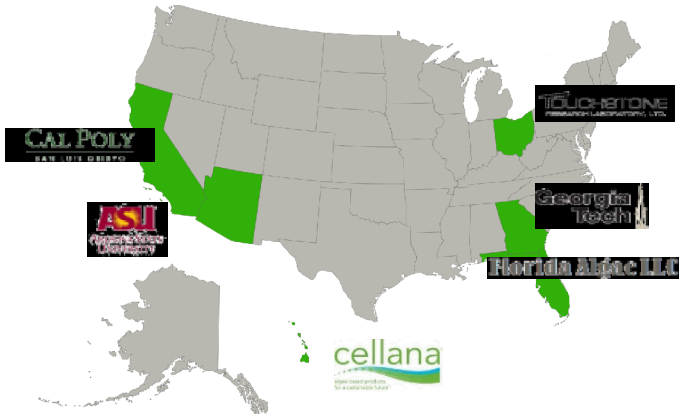
18s for eukaryotes

Sandia's Role in the ATP^3 Consortium:
genetic identification of pond crash agents

- 6 X 1000 L replicate ponds in 5 geographically distinct locations
- *Nannochloropsis oceanica*
- *Chlorella sp* DOE 1412
- *Desmodesmus*



Research Strategy



Collecting samples for various time-points and storing in RNA later and shipping at -4°C

↓

Sample prep and sequencing

↓

Analysis through MAGPie to obtain community structure

↓

Metadata associated with samples, Strain, location, weather, pond health, productivity → **Machine learning**

↓

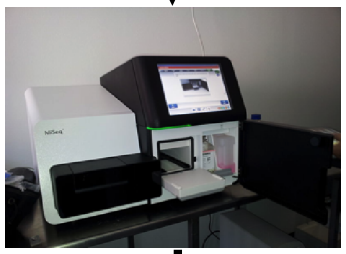
Pond crash signature

We sequenced and analyzed ~1200 sequences for a full year of operations of ATP^3 ponds in 2014

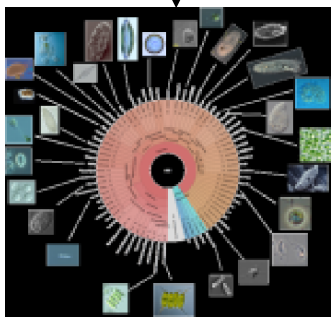
SSU rRNA Amplicon
sequencing and analysis workflow

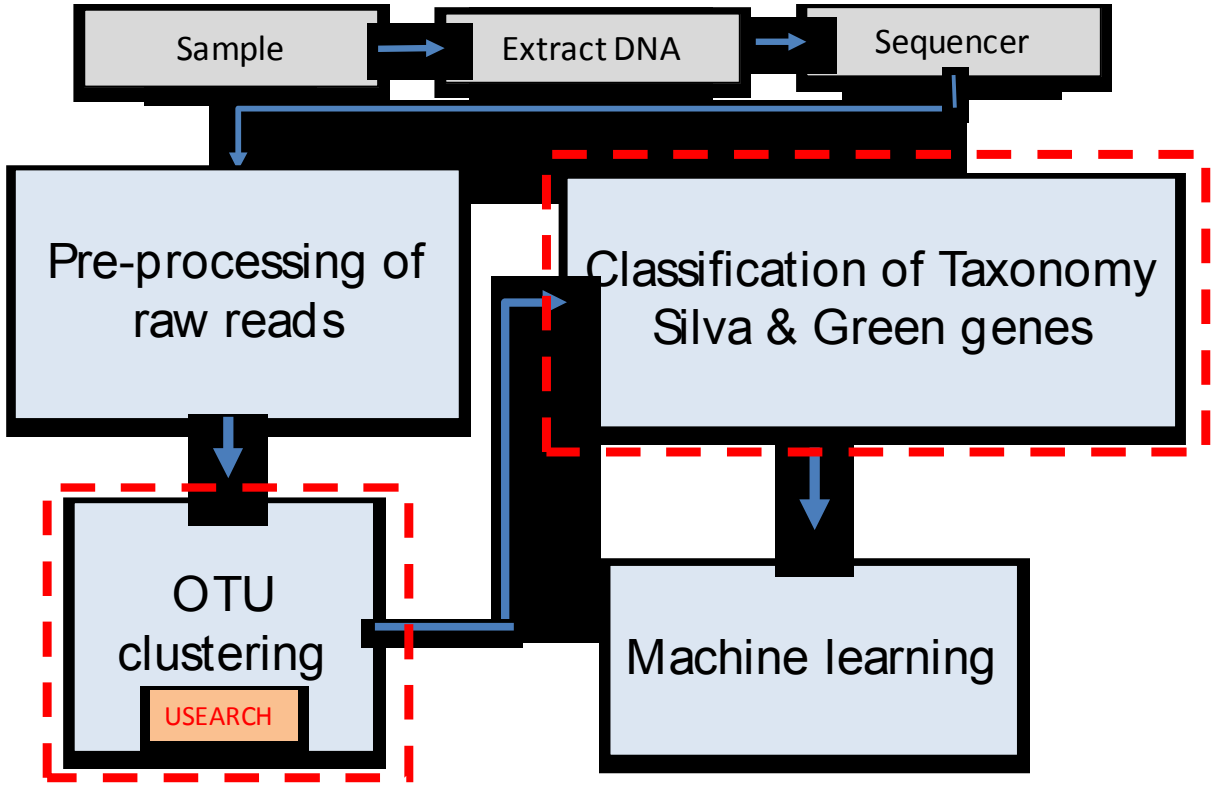
MAGPie

Pond



Microbiome





Classification accuracy		Illumina	
MAGPie		65%	→ Genus
100%	→ Genus	30%	→ Species
70%	→ Species		

Output data structure

Map of sequence reads to the 16s ssu DNA locus

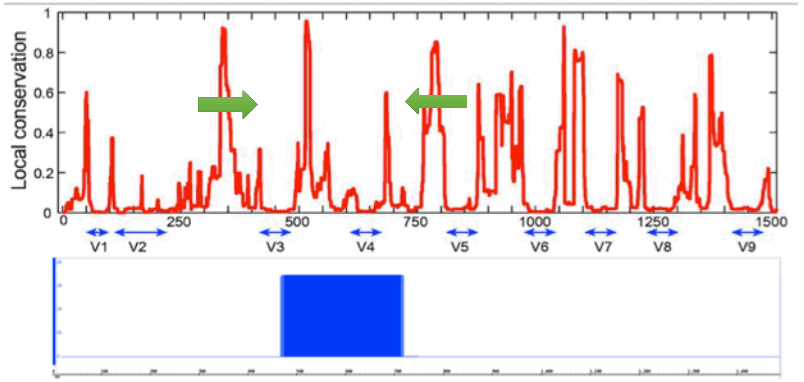


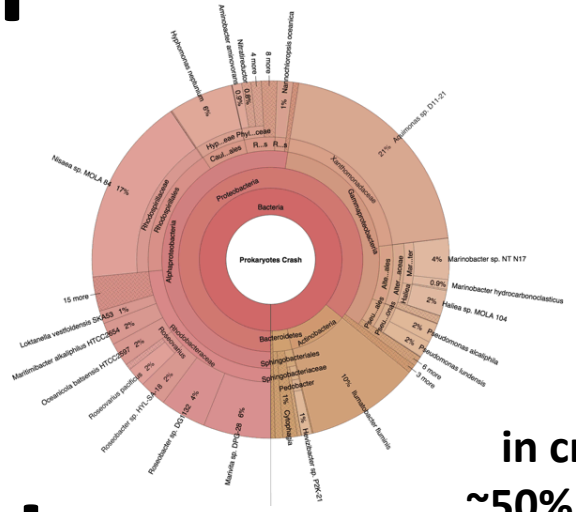
Table of classified OTUs and their abundances can be displayed on websites as well

taxid in DB	Taxanomy	taxa	Pond 1	Pond 2	Pond 3	Pond 4	Pond 5	Pond 6
U41092.1.1792	Eukaryota;Eustigmatophyceae;Eustigmatales;Monodopsidacea	Nannochloropsis granulata;	13379	1	1	1	19644	21286
Y14950.1.1794	Eukaryota;Viridiplantae;Chlorophyta;Trebouxiophyceae;Chlore	Chlorella sp. Yanaqocha RA1;	3375		216	113	677	202
JF834543.1.1228	Eukaryota;Bacillariophyta;Bacillariophyceae;Thalassiophysales;	Amphora sp. PP-2011;		1		1	1	1
DQ059583.1.1784	Eukaryota;Spirotrichea;Urostylida;Holostichidae;Holosticha;Ho	Holosticha diademata;	1					
EF165112.1.1687	Eukaryota;Chrysophyceae;Chromulinales;Chromulinaceae;Och	Ochromonas cf. gloeopara;	1		1		8	
FR865727.2.1770	Eukaryota;Viridiplantae;Chlorophyta;Chlorophyceae;Sphaeropl	Scenedesmus armatus;	12		27	10	1	4
AY520448.1.1735	Eukaryota;Bicosoecida;Cafeteriaceae;Cafeteria;Cafeteria minim	Cafeteria minima;		6	13	3	1	17
AY622191.1.1753	Eukaryota;Metazoa;Arthropoda;Ostracoda;Podocopida;Cyclocy	Cypria crenulata;			36	12	30	3
HM161745.1.1787	Eukaryota;Synurophyceae;Ochromonadales;Ochromonadaceae	Poterioochromonas sp. Y4;	11		7			3
L27634.3.1794	Eukaryota;Labyrinthulomycetes;Thraustochytriaceae;Labyrinth	Labyrinthuloides minuta;						
EU106848.1.1693	Eukaryota;Bicosoecida;Caecitellus;Caecitellus sp. RCC1072;	Caecitellus sp. RCC1072;		2	14		21	63
M74497.1.1789	Eukaryota;Viridiplantae;Chlorophyta;Chlorophyceae;Sphaeropl	Hydrodictyon reticulatum;	7		12	11	20	
JN120201.1.1541	Eukaryota;Oligohymenophorea;Opisthonectidae;Opisthonecta	Opisthonecta henneguyi;						
DQ514856.1.1722	Eukaryota;Bacillariophyta;Coscinodiscophyceae;Stephanodisca	Cyclotella meneghiniana;			5		1	
EU032356.1.1751	Eukaryota;Oligohymenophorea;Pleuronematida;Cyclidiidae;Cyc	Cyclidium glaucoma;	1		9	1	8	1
JF489982.1.1752	Eukaryota;Eustigmatophyceae;Eustigmatales;Monodopsidacea	Nannochloropsis oceanica;						1
EU039885.1.1700	Eukaryota;Colpodea;Colpodida;Hausmanniellidae;Bresslauides	Bresslauides discoideus;						
U49911.1.1808	Eukaryota;Metazoa;Rotifera;Monogononta;Ploimida;Brachioni	Brachionus plicatilis;			11	2	9	
AF352222.1.2260	Eukaryota;Viridiplantae;Streptophyta;Zygnemophyceae;Desmi	Closterium moniliferum;						
HQ912576.1.1744	Eukaryota;Bacillariophyta;Coscinodiscophyceae;Stephanodisca	Cyclotella meneghiniana;						

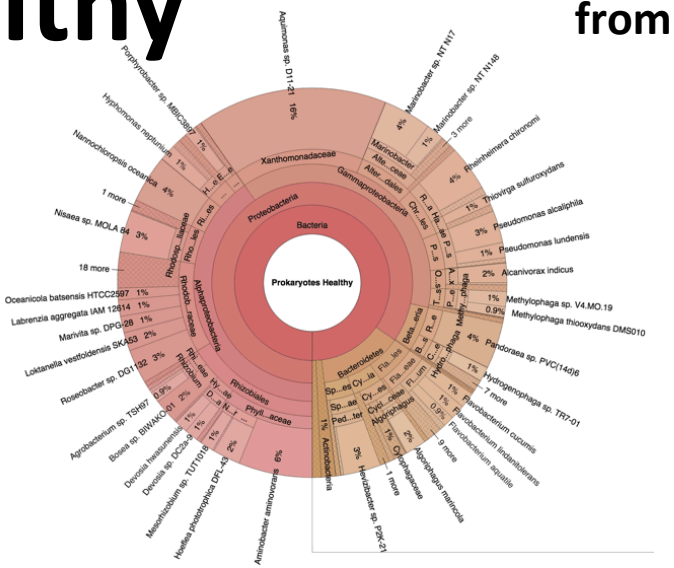
Can we apply machine learning on this data to learn more ?

Diversity in crash ponds

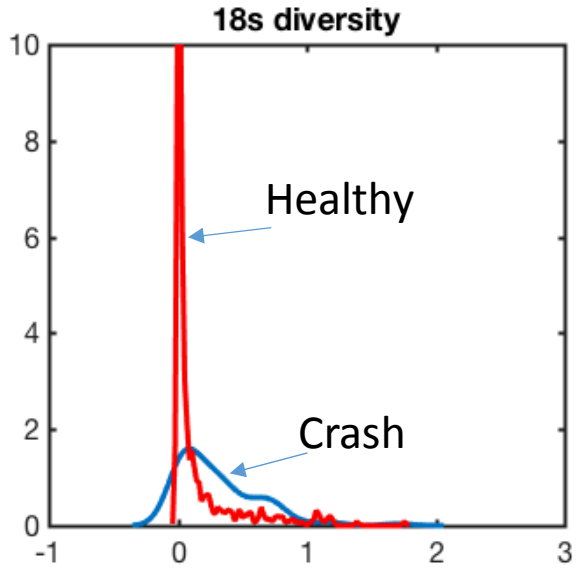
Crash

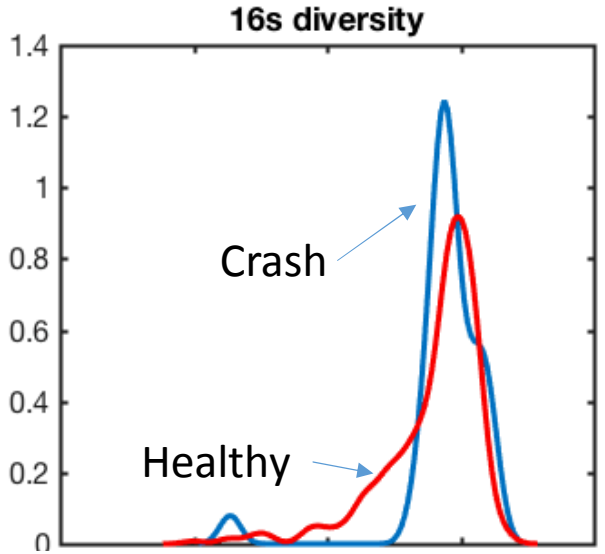


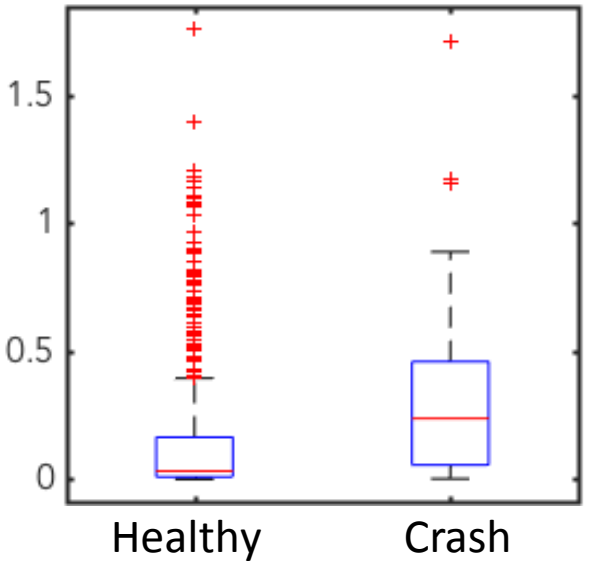
Healthy

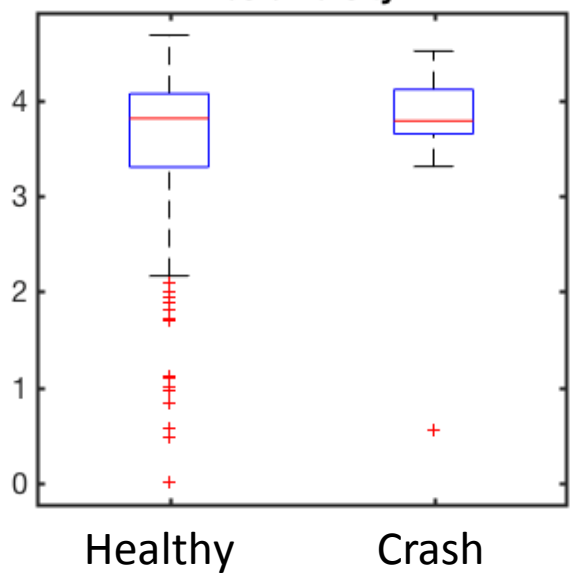


in crash samples ~50% of reads came from 3 species



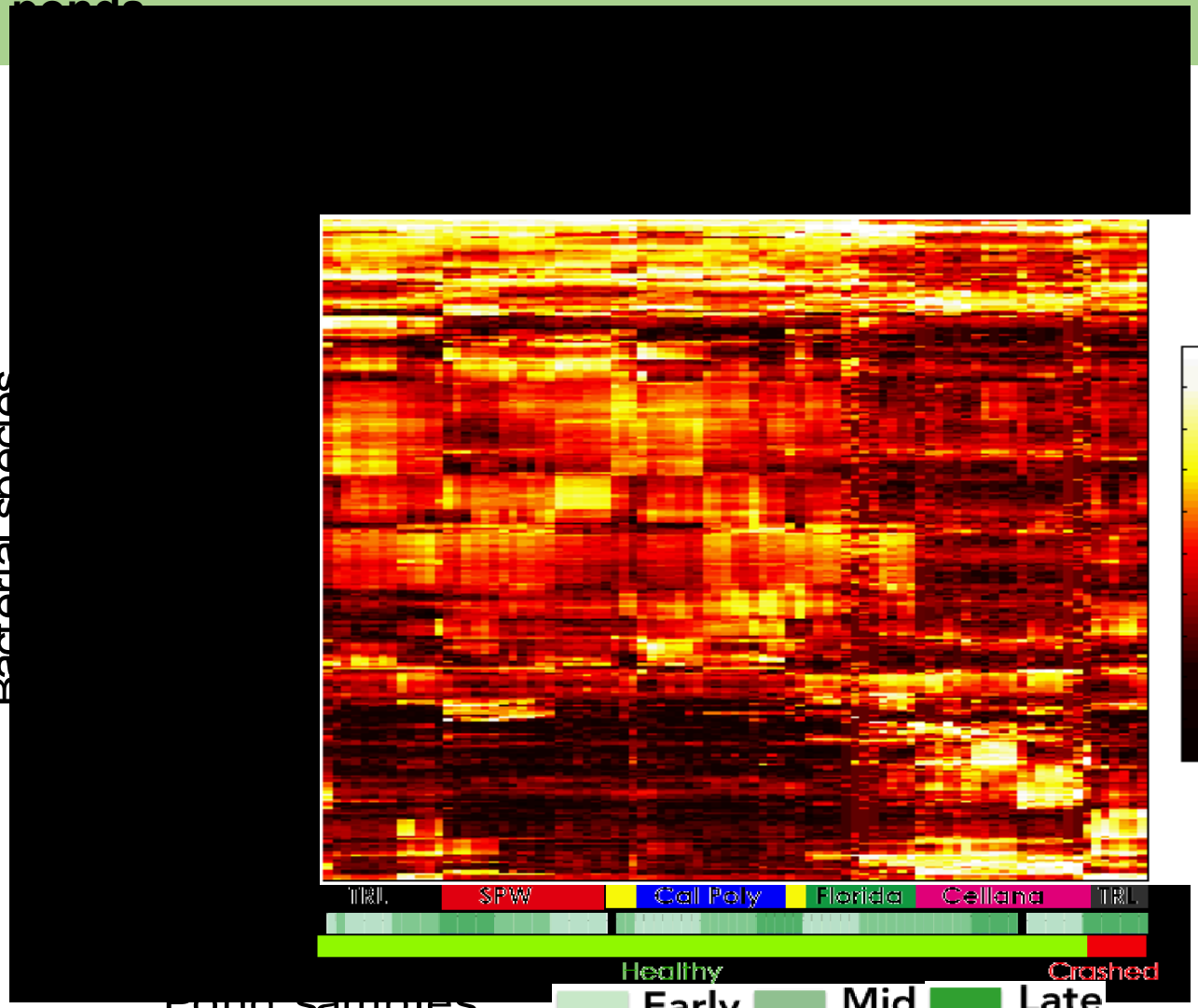






Clustering analysis:

Highlights the community structure of the various geographical locations and health of the ponds

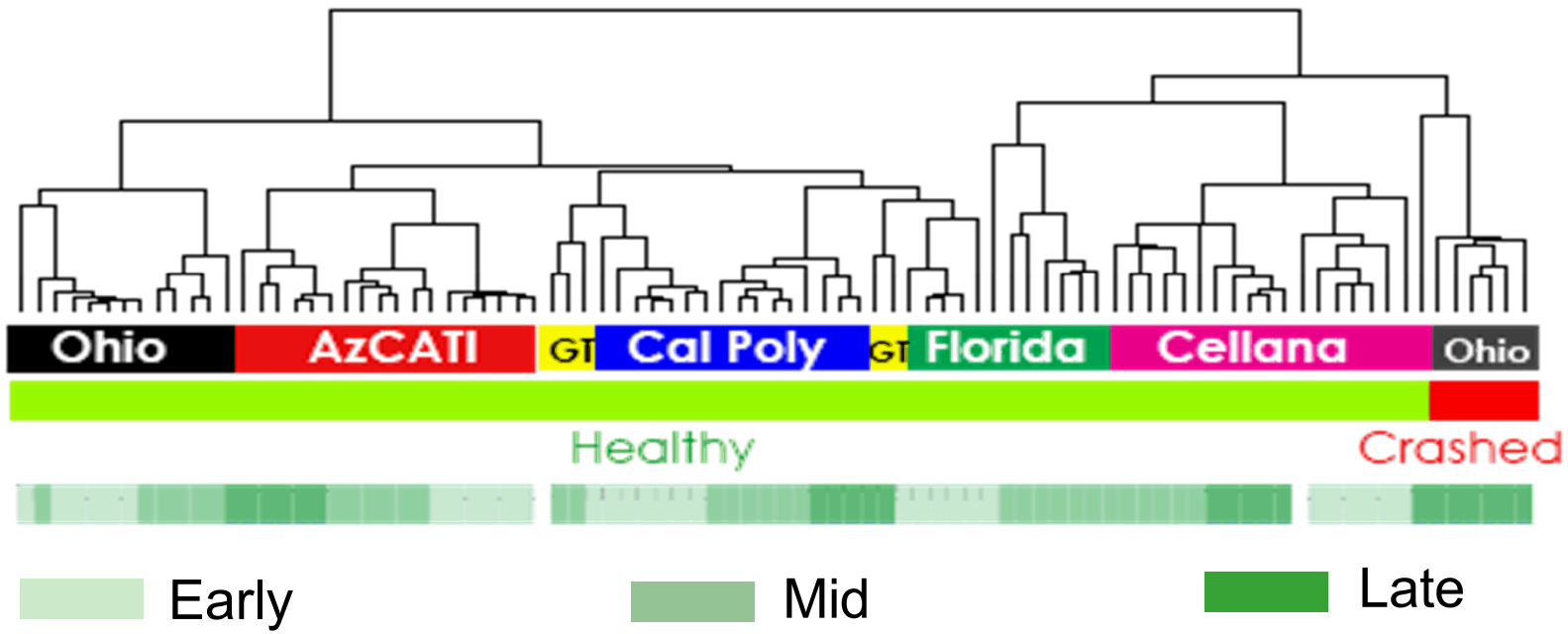


16s data for Spring 2014 all sites

Heatmap of clustered samples and species

Sites cluster together

Community structure gradually changed with time



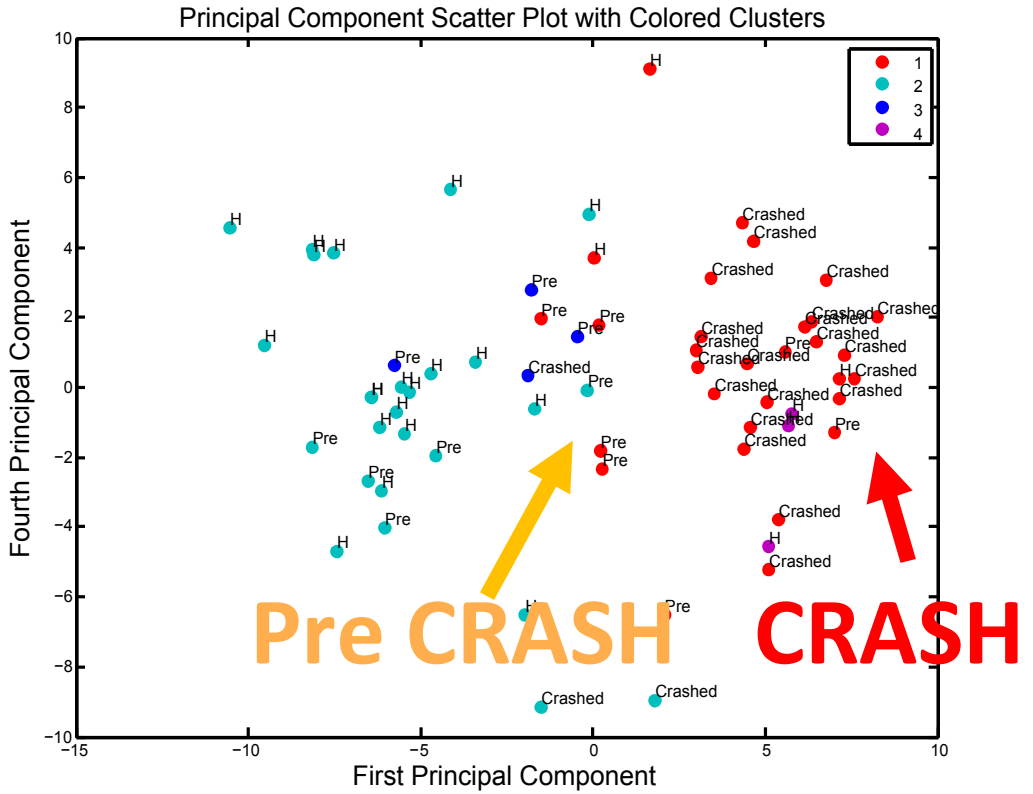
And within same sites the community structure gradually changed with time

Clustering highlights pre-crash signature

Clustering analysis on the ponds could differentiate the healthy and crashed ponds.

Ash-Free Dry Weight

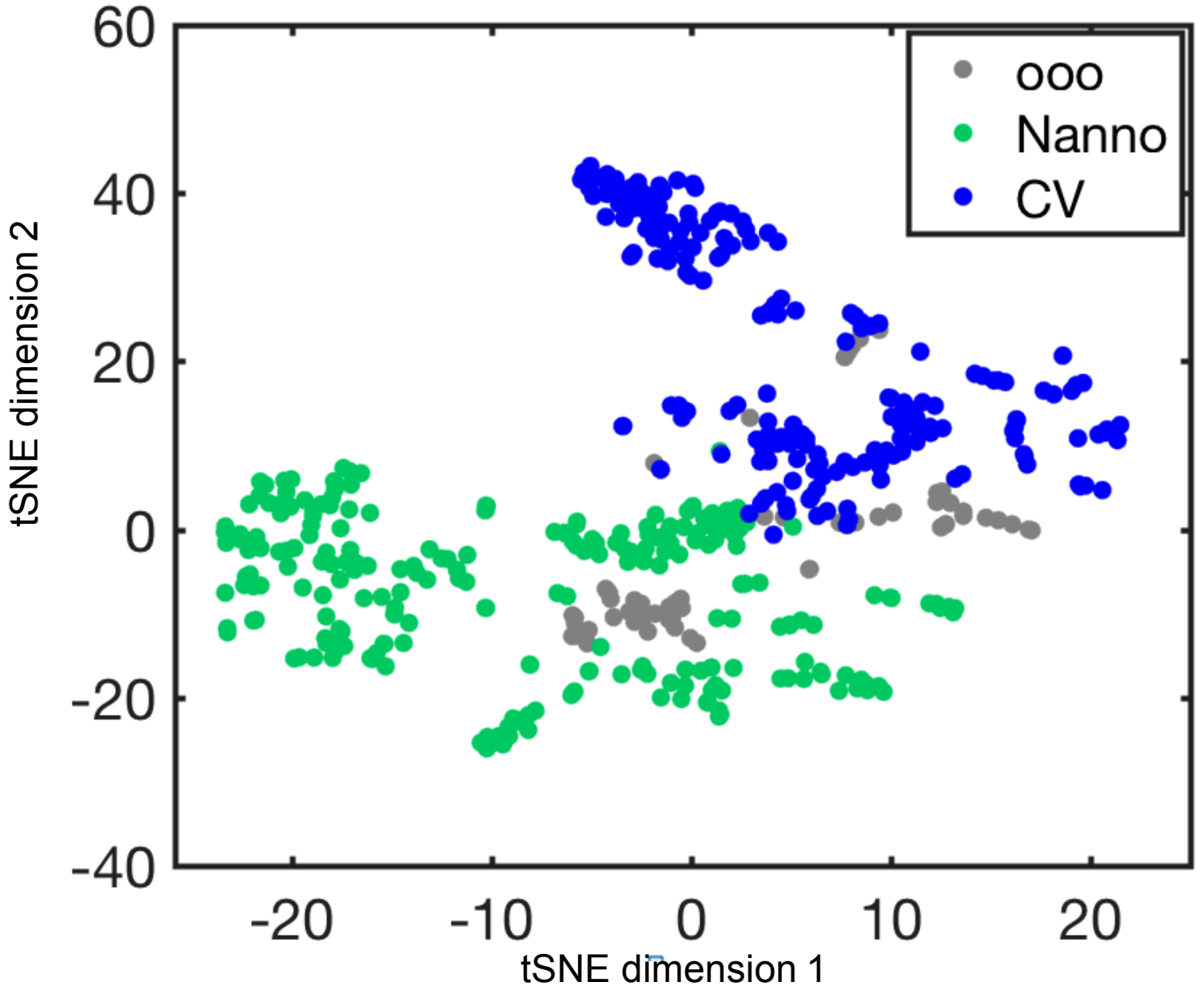
Principle component analysis : crashed ponds have a different microbiome signature



tSNE-plots

The clusters obtained from the tSNE transformation explain most of the variations in the microbial ecology of the samples. Hence when we apply the decision tree over these datasets we can only identify few key features of biomarkers of crashes

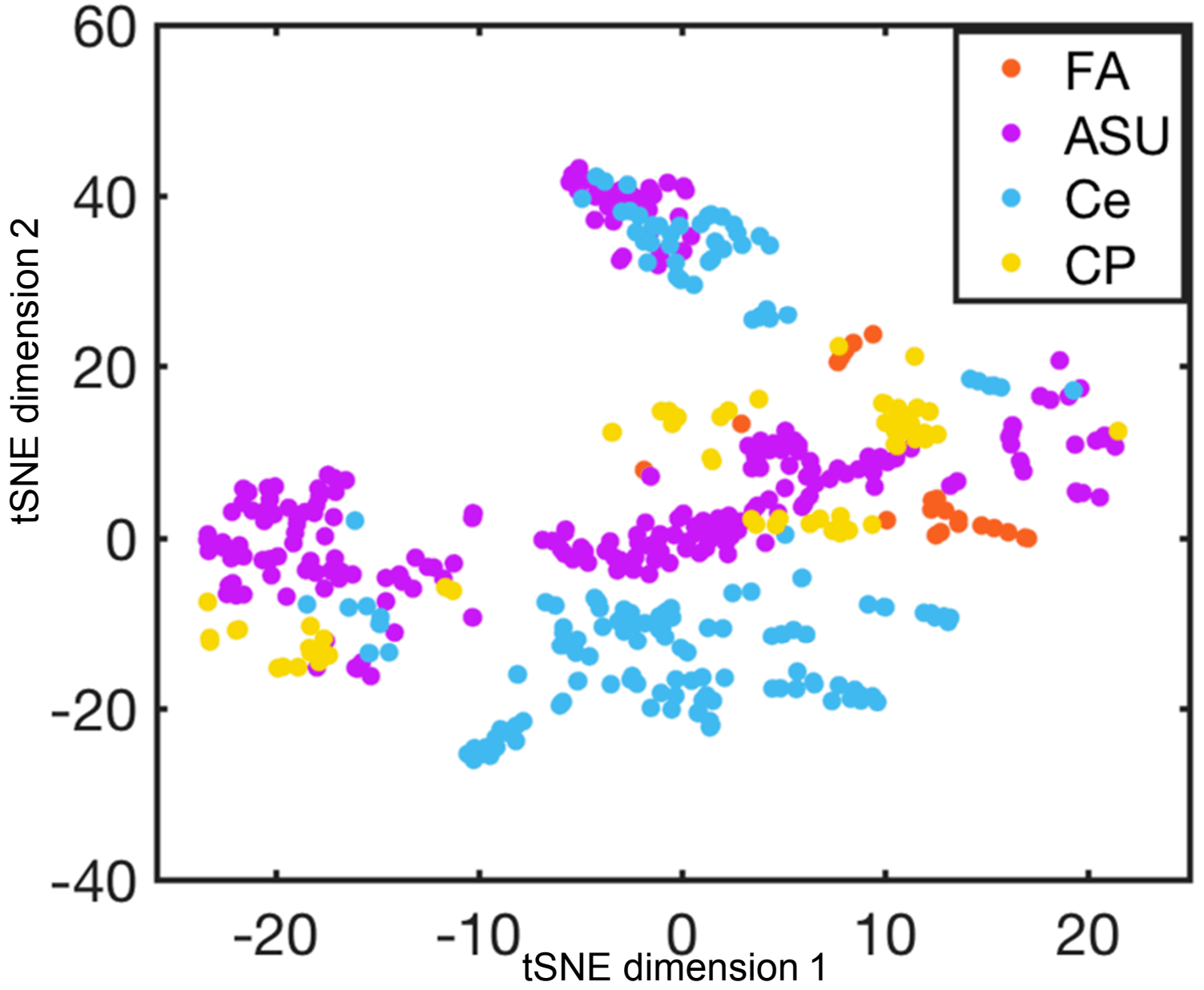
"tSNE (t-distributed stochastic neighbor embedding) algorithm is a machine learning algorithm for dimensionality reduction. It is a nonlinear dimensionality reduction technique that is particularly well-suited for embedding high-dimensional data into a space of two or three dimensions, which can then be visualized in a scatter plot.)"



tSNE-plots

The clusters obtained from the tSNE transformation explain most of the variations in the microbial ecology of the samples. Hence when we apply the decision tree over these datasets we can only identify few key features of biomarkers of crashes

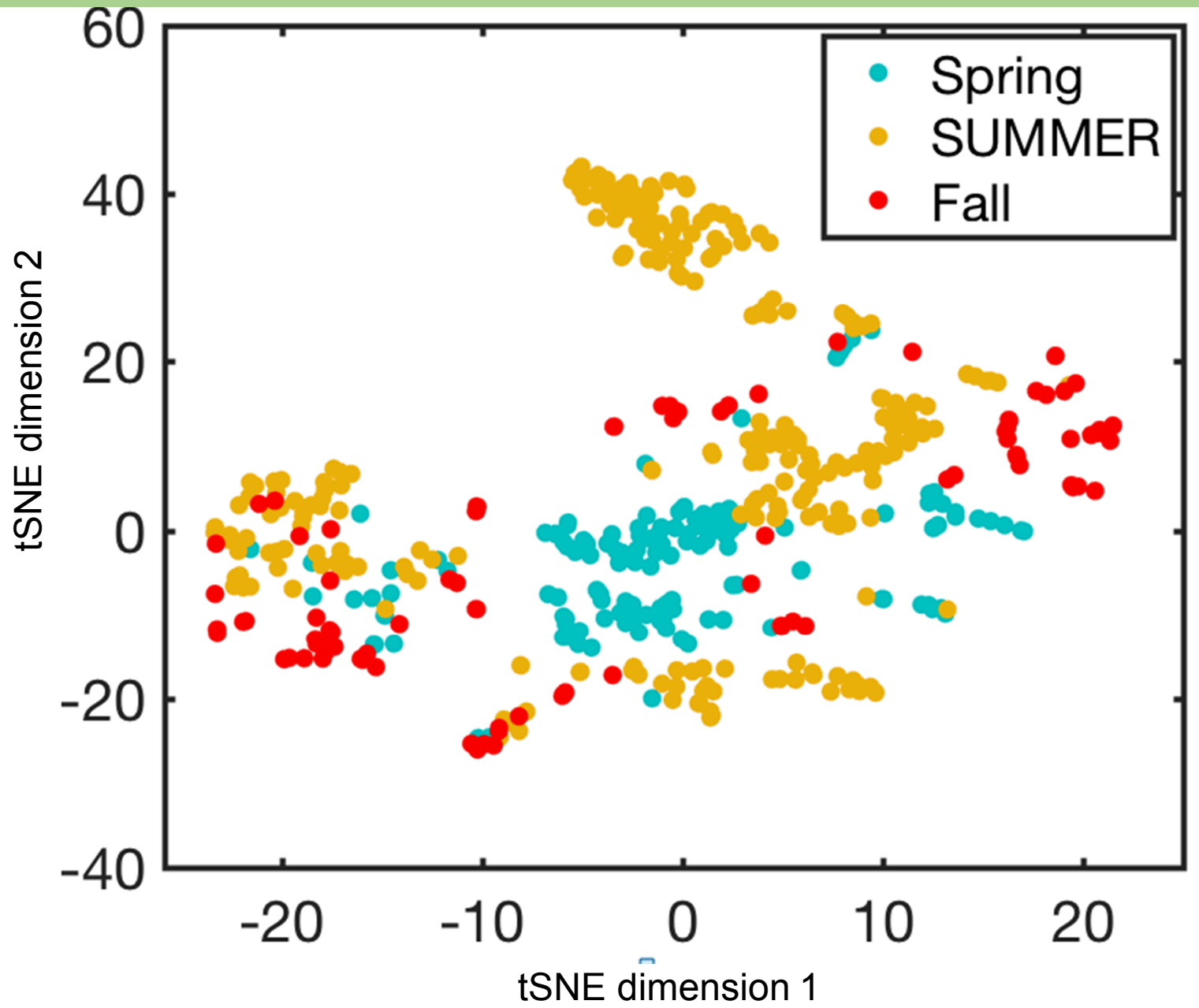
"tSNE (t-distributed stochastic neighbor embedding) algorithm is a machine learning algorithm for dimensionality reduction. It is a nonlinear dimensionality reduction technique that is particularly well-suited for embedding high-dimensional data into a space of two or three dimensions, which can then be visualized in a scatter plot.)"

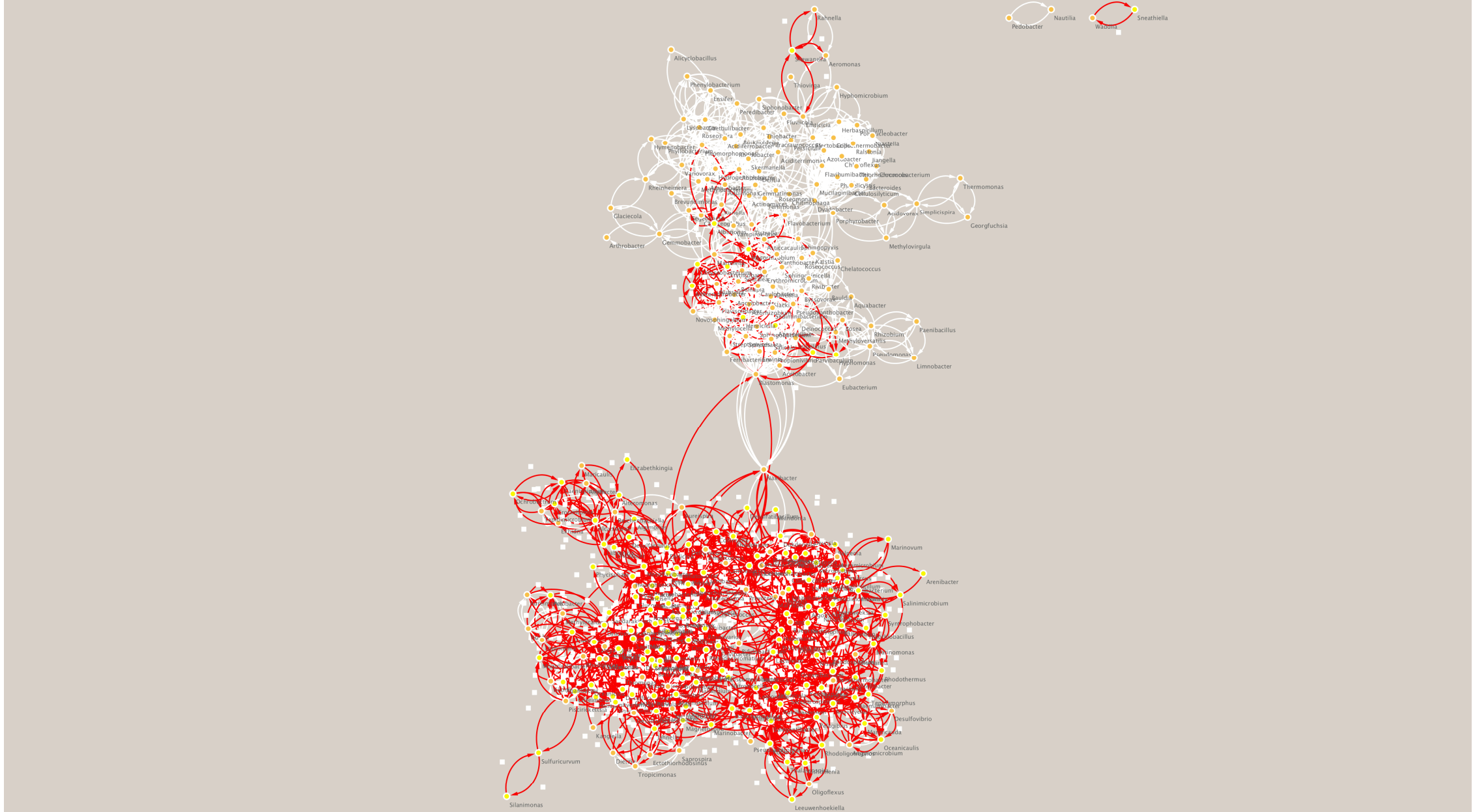


tSNE-plots

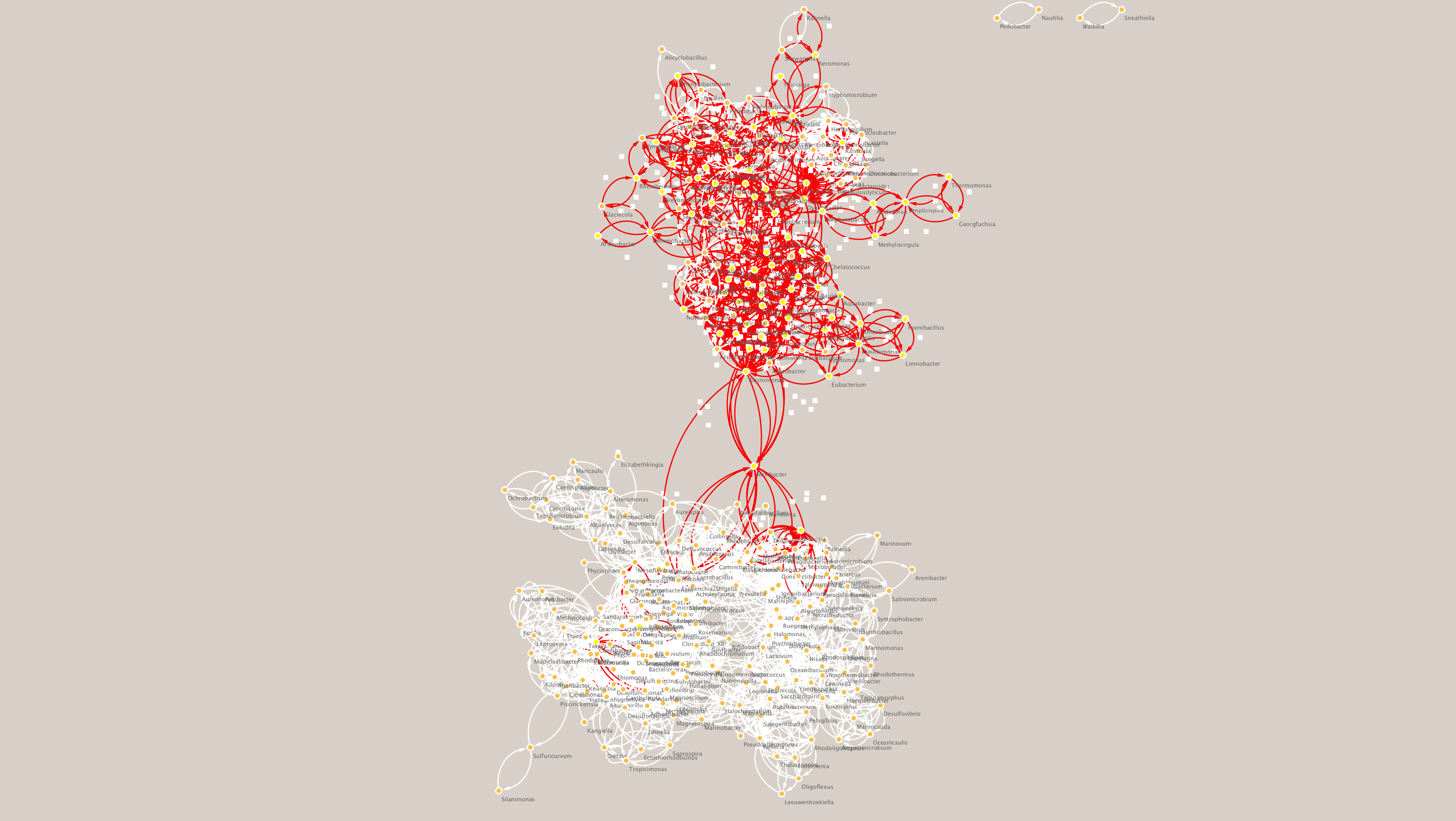
The clusters obtained from the tSNE transformation explain most of the variations in the microbial ecology of the samples. Hence when we apply the decision tree over these datasets we can only identify few key features of biomarkers of crashes

"tSNE (t-distributed stochastic neighbor embedding) algorithm is a machine learning algorithm for dimensionality reduction. It is a nonlinear dimensionality reduction technique that is particularly well-suited for embedding high-dimensional data into a space of two or three dimensions, which can then be visualized in a scatter plot.)"





Microbial Network in ATP^3 Ponds

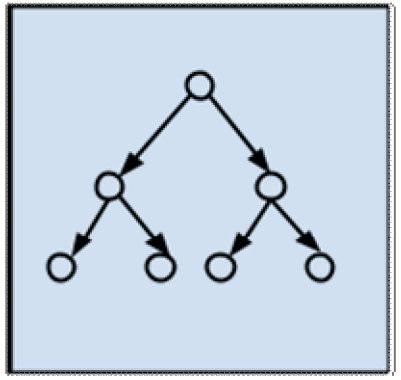


Microbial Network in ATP^3 Ponds

Machine learning on pond crashes

	Species 1	Species 2	Species 3	….	Species M	Result
Pond1_Site1	34	456	2	….	78	Crash
Pond2_Site2	0	2	45	….	900	Healthy
Pond3_Site3	765	0	4	….	22	Healthy
Pond4_Site4	44	334	11	….	12	Healthy
Pond5_Site5	73	543	7	….	5	Crash
….	…	….	….	….	….	….
Pond6_SiteN	456	100	233	….	33	Healthy

CART
Classification
Regression Tree

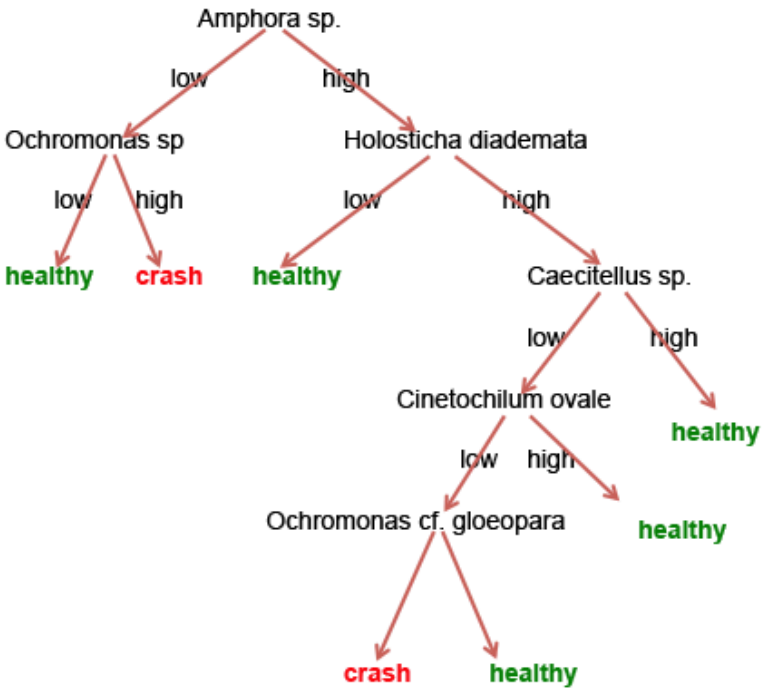


Decision Tree
Algorithms

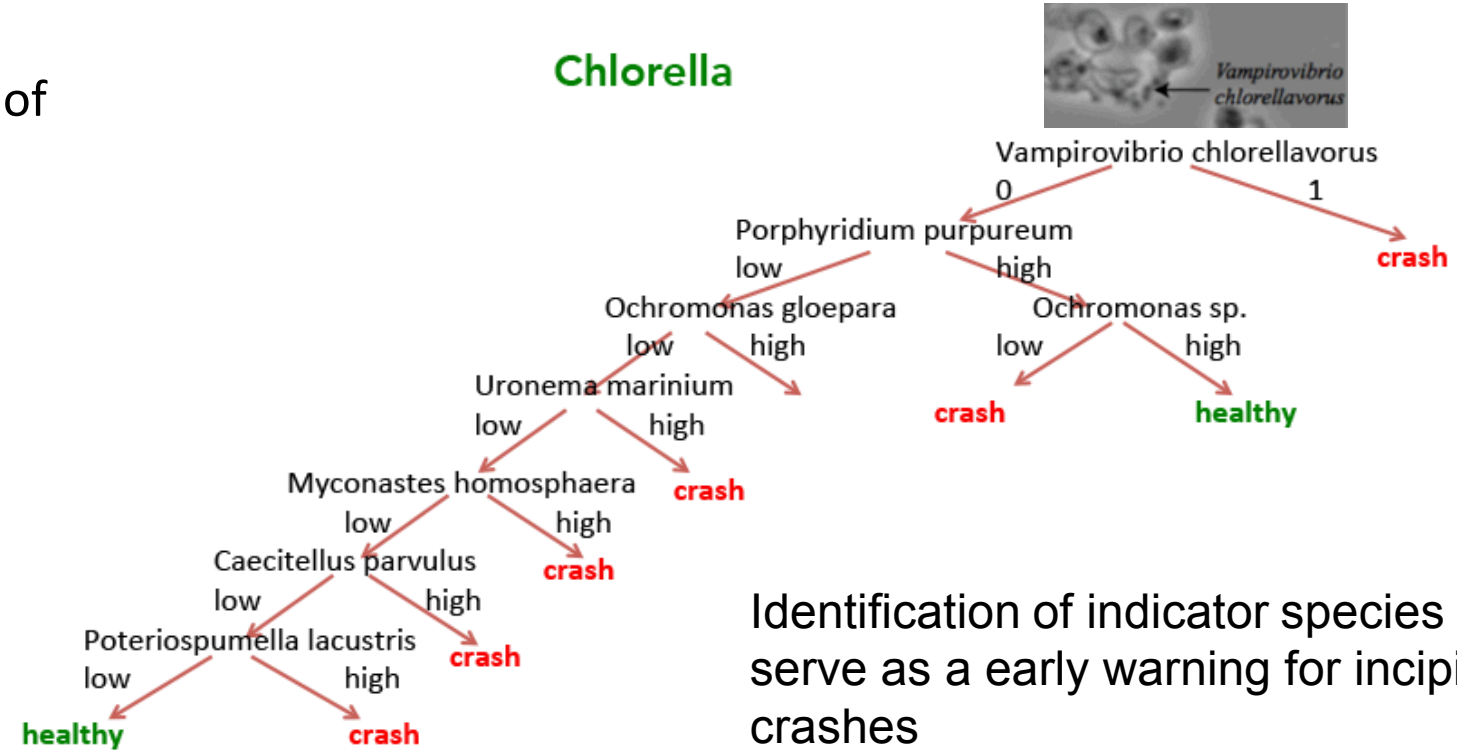
Applying data driven approaches: Decision Tree reveals pond crash signature

Cross-validated tree showing the species determining the signature of crash ponds

Nannochloropsis



Chlorella



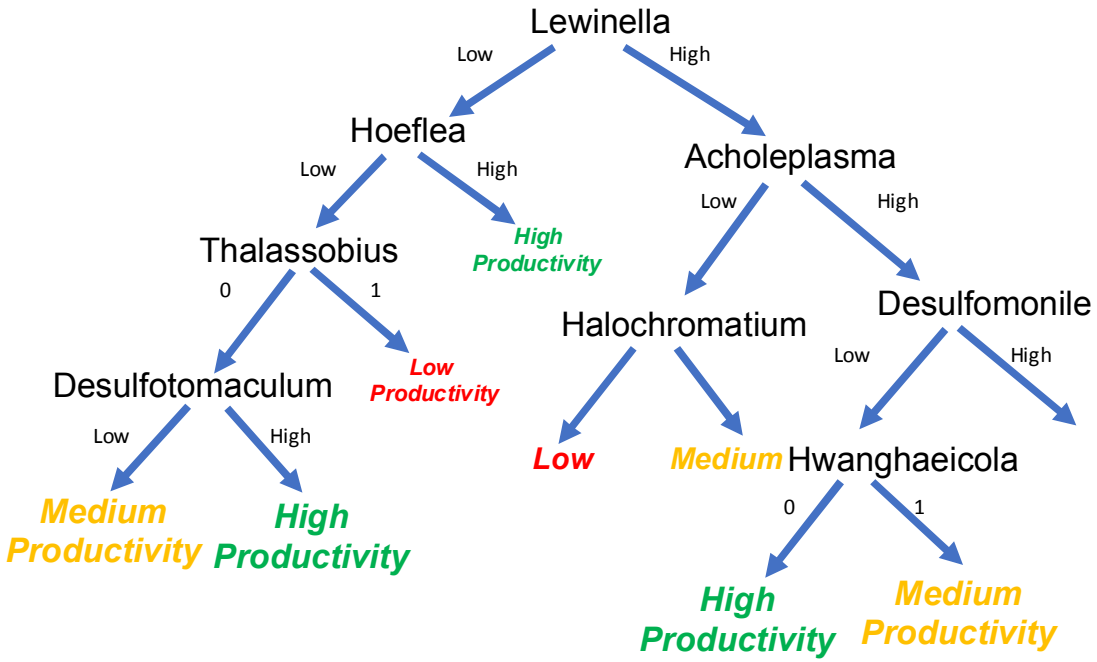
Identification of indicator species may serve as a early warning for incipient crashes

Model Accuracy	Crash Prediction	Productivity Prediction
Chlorella V	87%	71%
Nanno O	91%	65%

Applying data driven approaches: Decision Tree reveals pond productivity

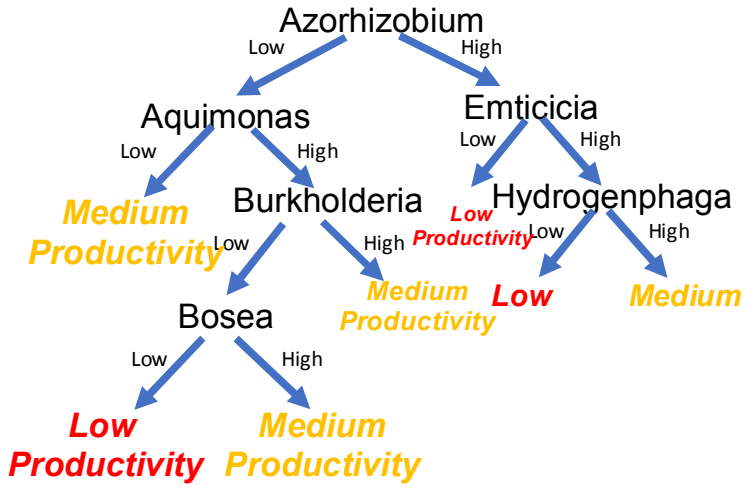
A

Nanno O



B

Chlorella V



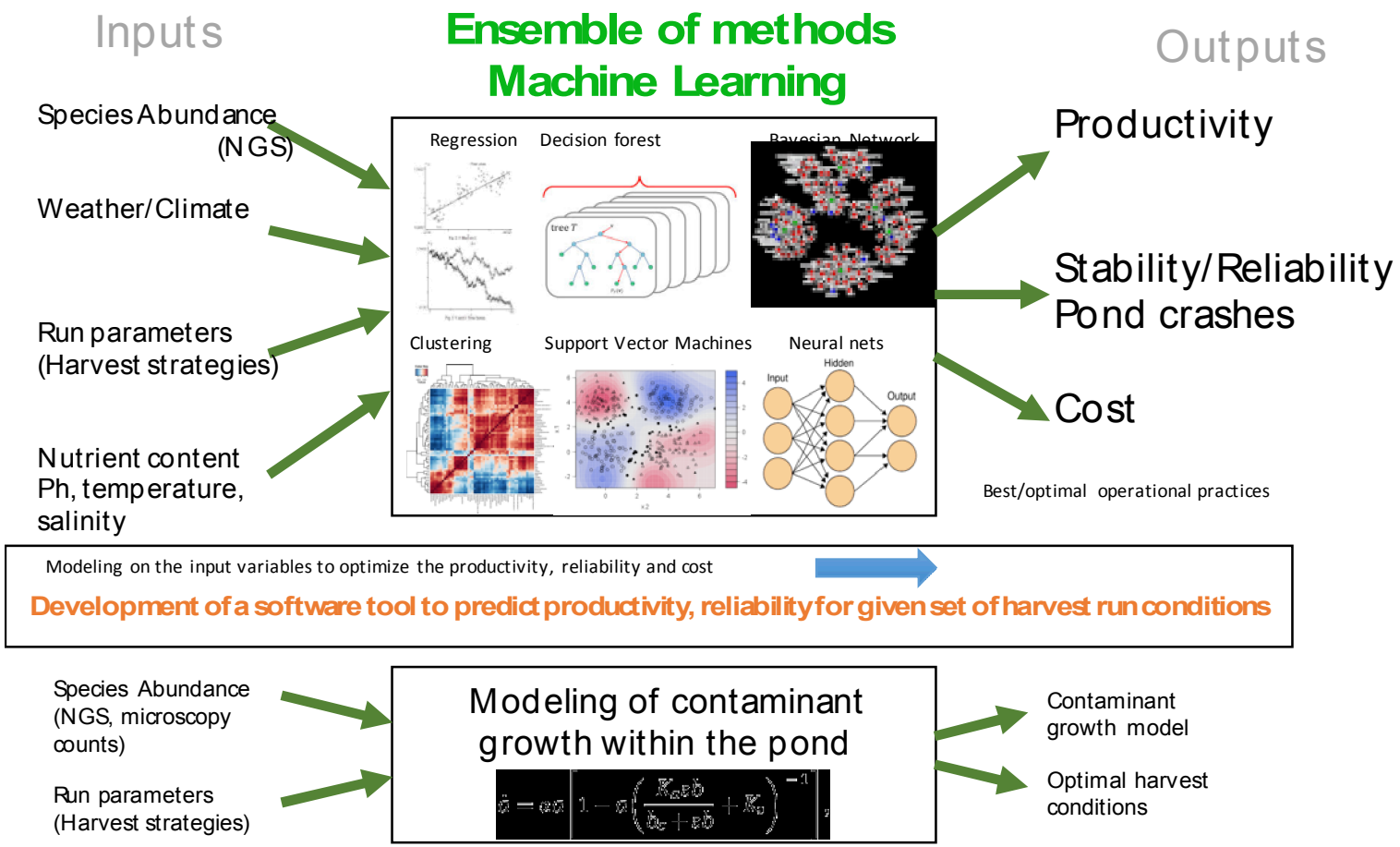
Model Accuracy	Crash Prediction	Productivity Prediction
Chlorella V	87%	71%
Nanno O	91%	65%

Summary

- Workflow for quick turnaround of amplicon sequencing 16s and 18s
- New bioinformatics software pipeline
  - fast and efficient and producing user friendly outputs.
  - pipelines used and developed:
    - MAGPie (Metagenomics and Amplicon Sequencing Pipeline)
    - & RapTOR (Rapid Threat Organism Recognition)
- Identified Eukaryotic and Prokaryotic community structure associated with pond crashes
- Further work on better predictive and insightful model is underway

Further work

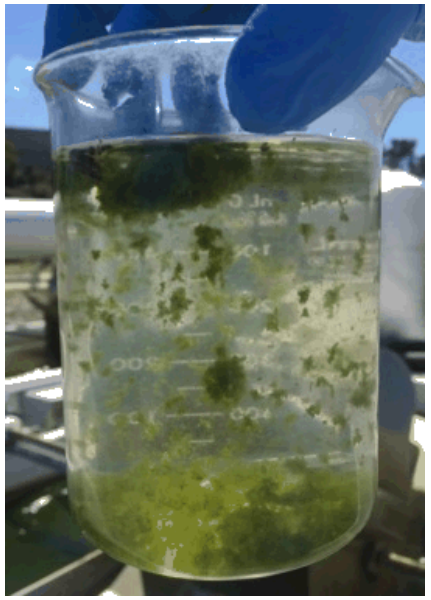
Further work has being carried out on LDRD (PI RW Davis) and DISCOVR AOP (TW Lane)



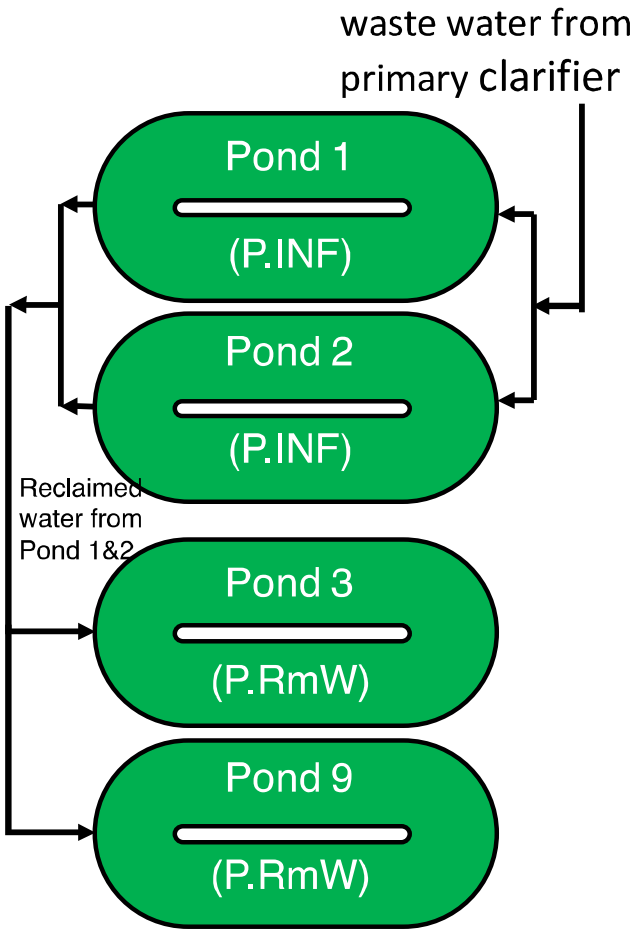
Algal Biomass Yield: Cal Poly SLO

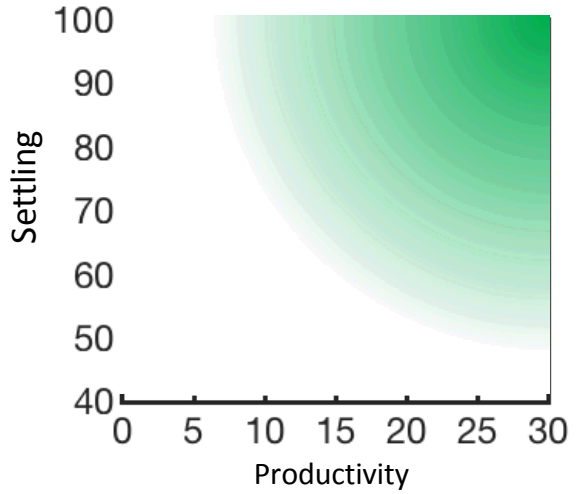
Algae cultivation using waste water

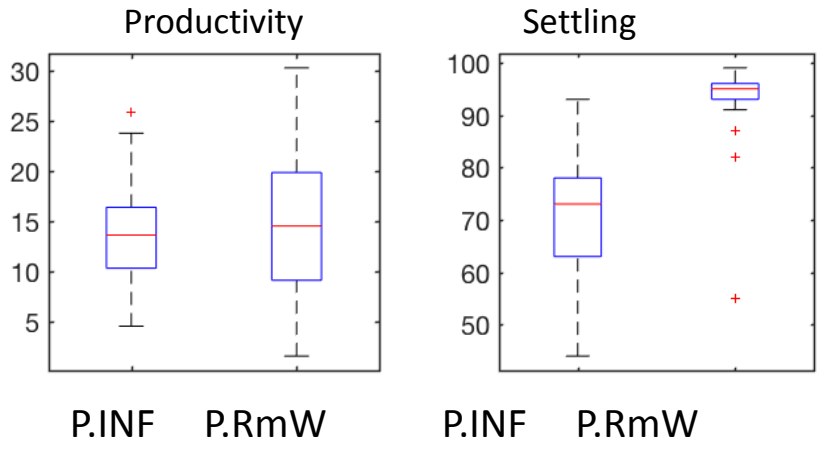
- Delhi CA sewage treatment plant
- Freshwater natural algal assemblages grown on primary effluent
- Original Oswald pond system
- ~1000L raceways (mBio Eng)



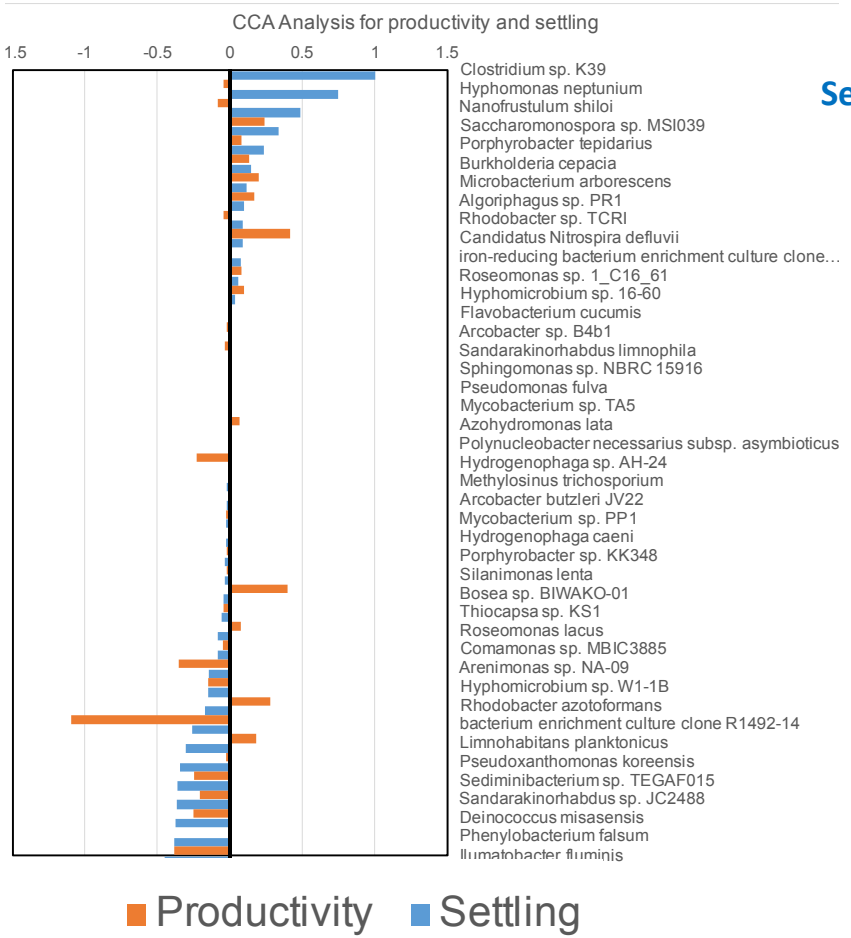
Algae Cultivation from Municipal Waste Water at SLO

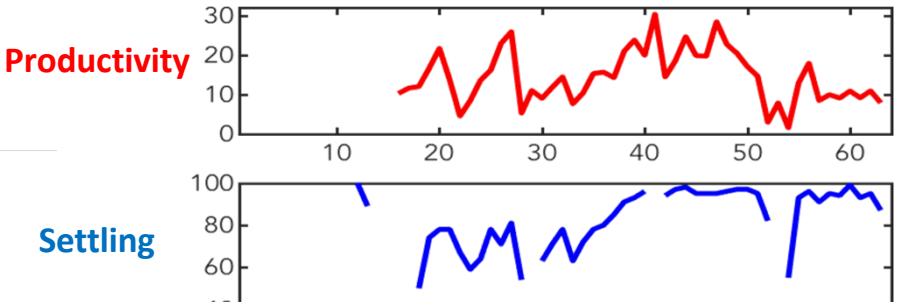


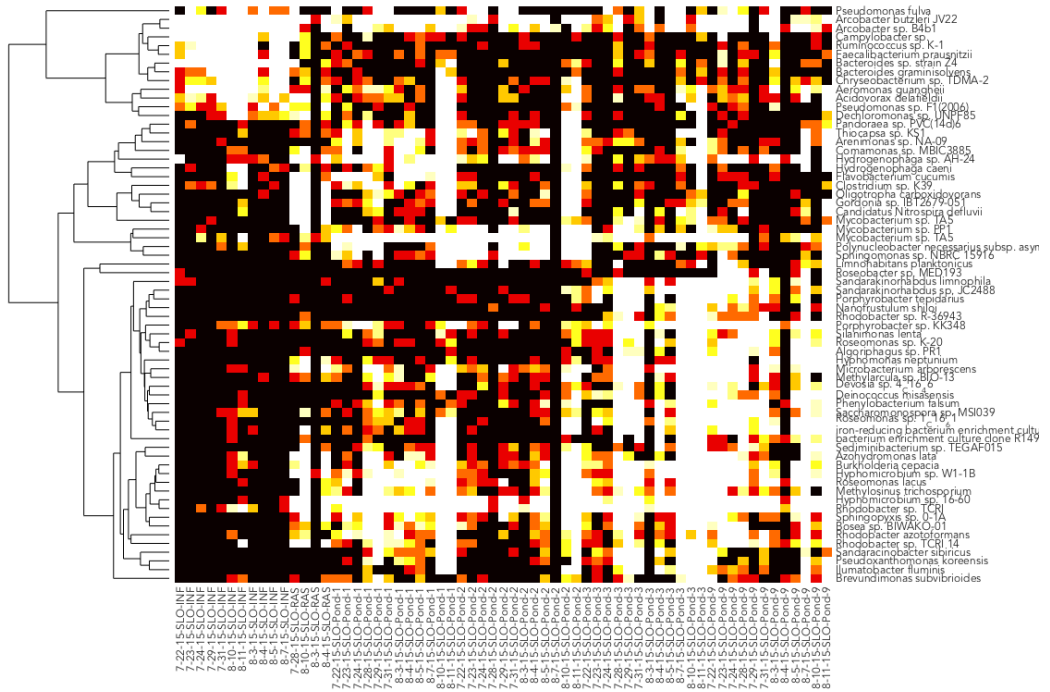




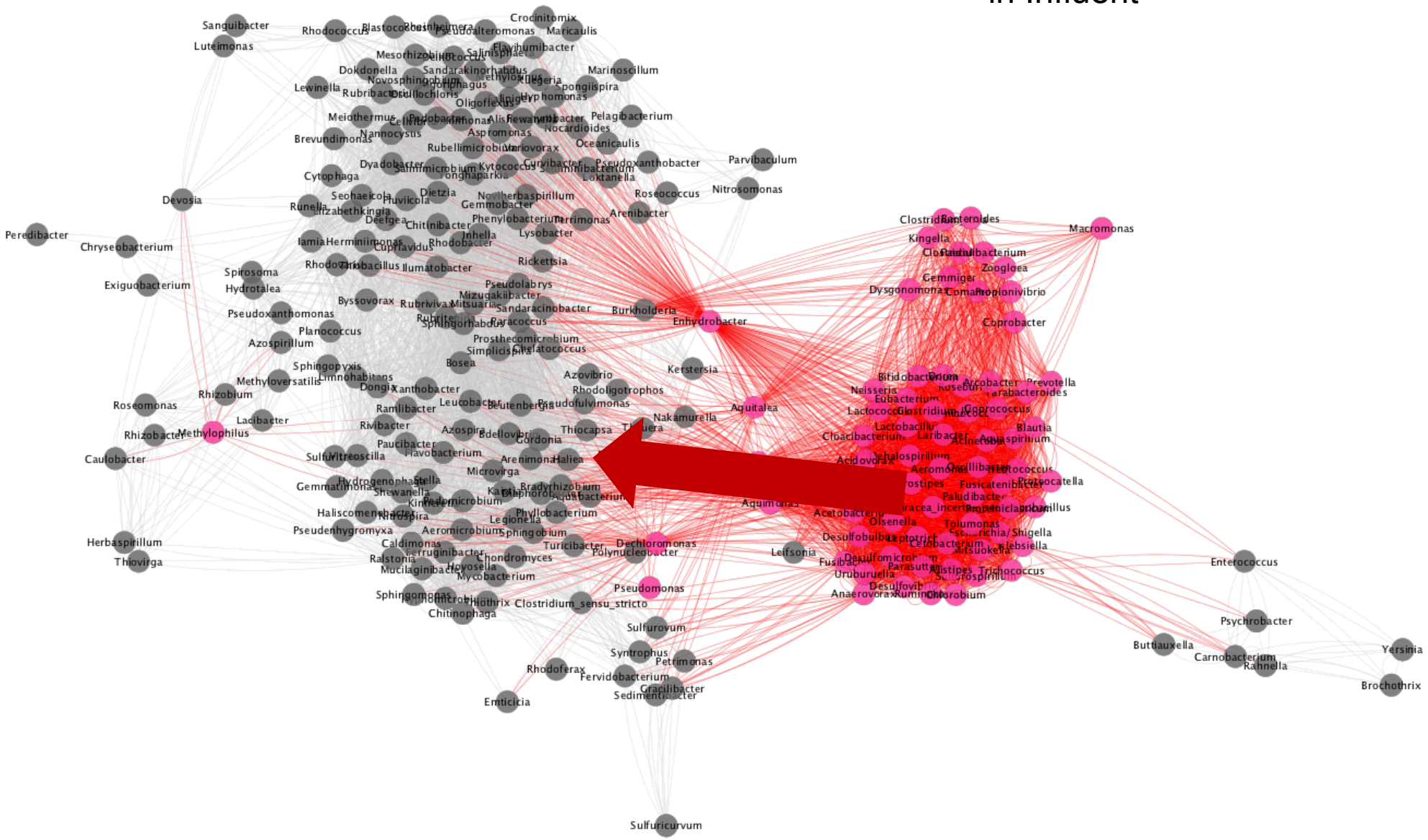
Canonical Correspondence Analysis of productivity and settling for SLO site



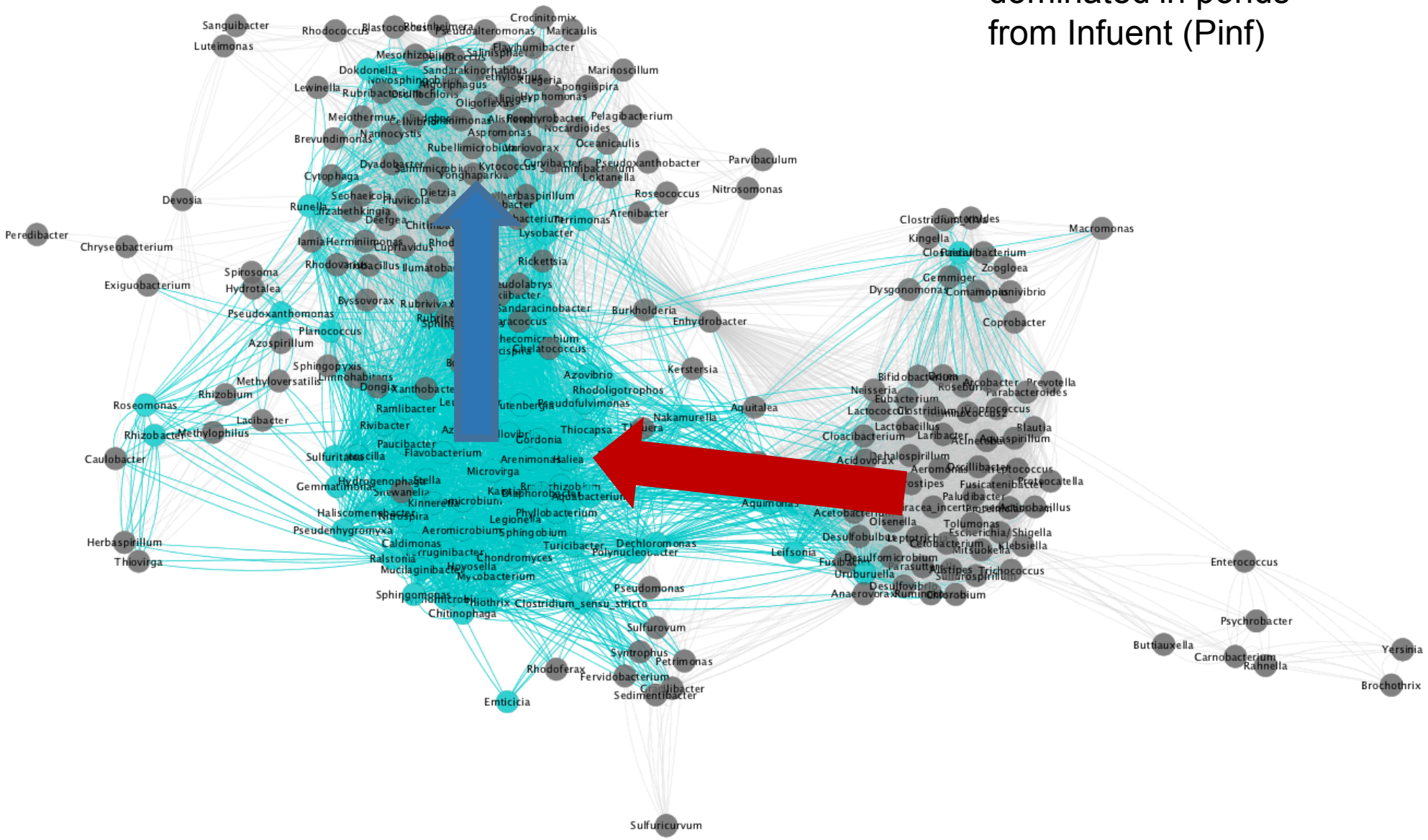




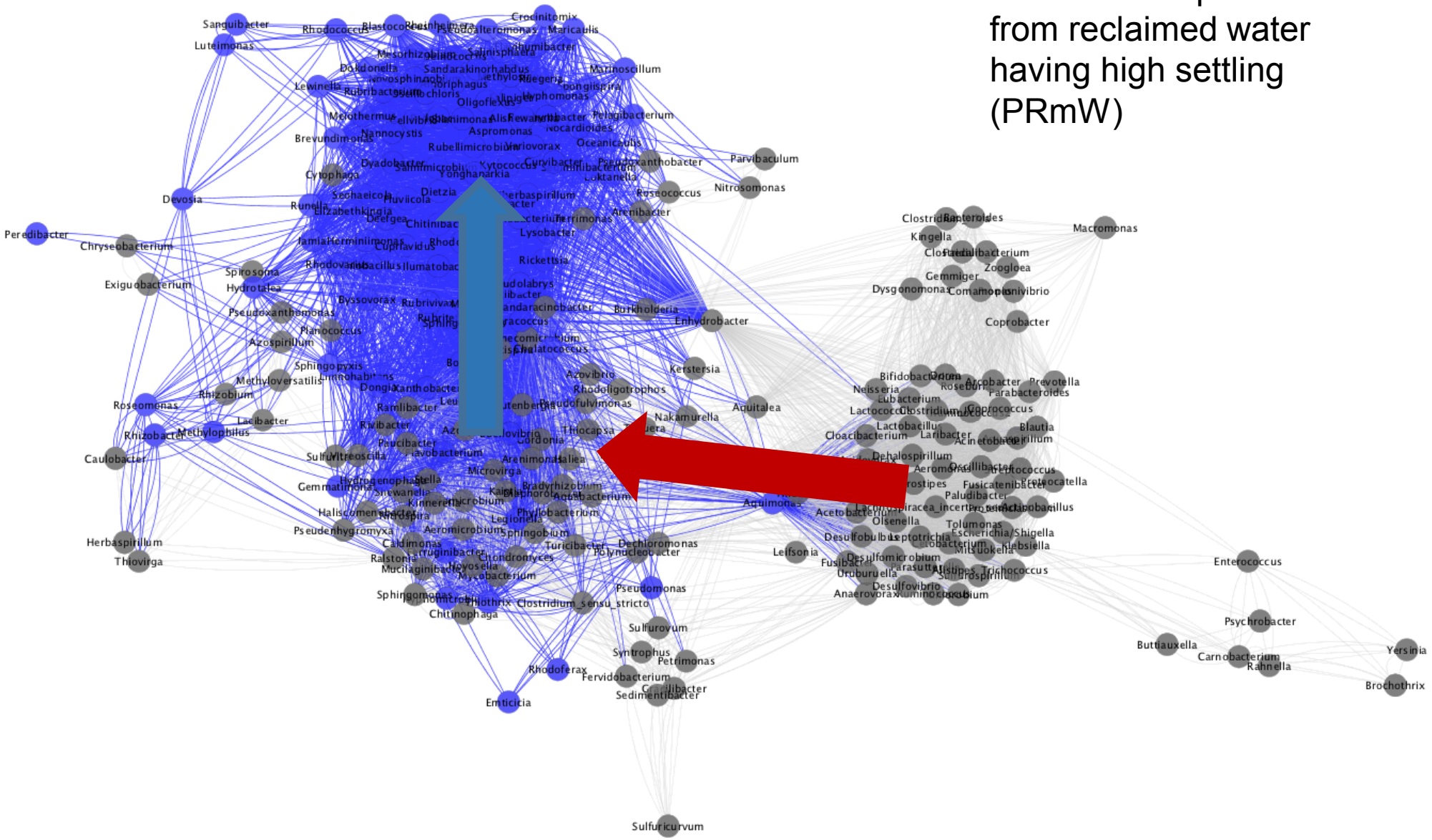
Microbial community in Influent

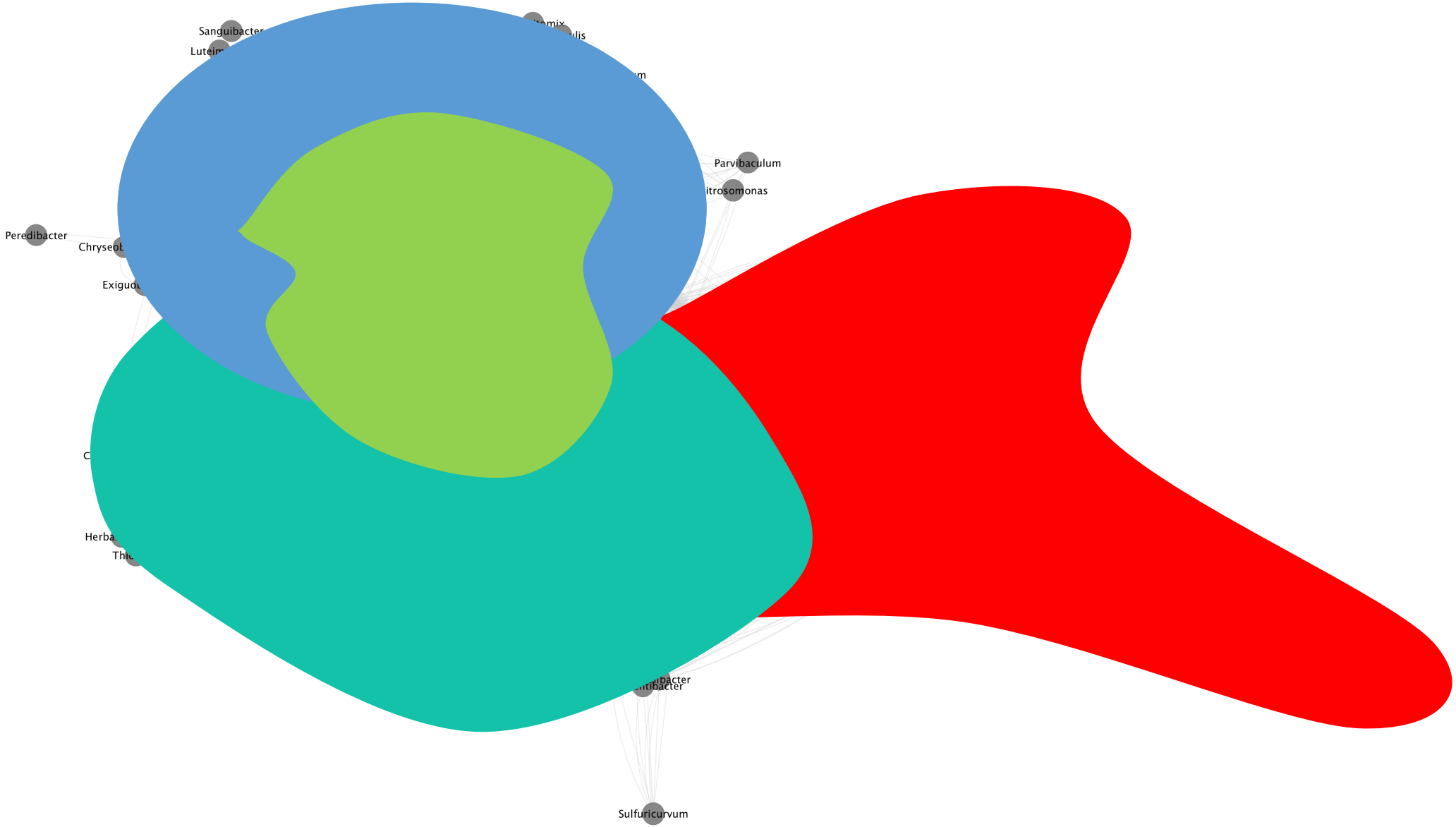


Microbial community dominated in ponds from Infuent (Pinf)



Microbial community dominated in ponds from reclaimed water having high settling (PRmW)




Parvibaculum
Peredibacter
Sulfuricurvum

Supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Thank You