Metapathways Workshop: Functional Space Reduction

Analyzing metagenomic data from different oceanic provinces

Metapathways: Steps

1. Quality Control and ORF prediction

2. ORF annotation
   A. Functional Annotation
   B. Taxonomic Annotation

3. Analyses

4. ePGDB construction

Metapathways Input

- .fasta file

File Path ▾ : ~/MetaPathways_1_0/input/jgi_illumina/4096453.combine

4096453.combined_unique.fa

```
>Sequence0000000001
GGTGCTGAGGTAATTCATCTTGGTCATAATCGCTCAGTGCATGAAATA
AAGGTGGTCACATGGAGTTTTTCAAATATATGTATGATATGCTTGAAG
TCCAGAGGAAGTAAAAGAACTAATGAATTATGGTATTACCAGAATTTA
ATGGAGAAATCTGATTTTCCCACAGGAGAAAGTCTAAAAATTGATTTA
CAGAAAATTATCCTGAAATAGCAGCACCCATGATAAAAGAGCTGATAA
TGCTGGAAAATCTTCACTTGTGGATGAAATTGTAAGACGCTTTCTTGA
AAATCTGGAGGTGCACTTCTTGGCGATCGTATAAGGATGAATGCAATT
GTATTTCAAATTATGTTAAGGATGCTGAAATAATTGCTCAGGCCGCAG
AATAATTGATCATAGTGATGTTTCATTCTATGTTATGACCCCTGAATA
```

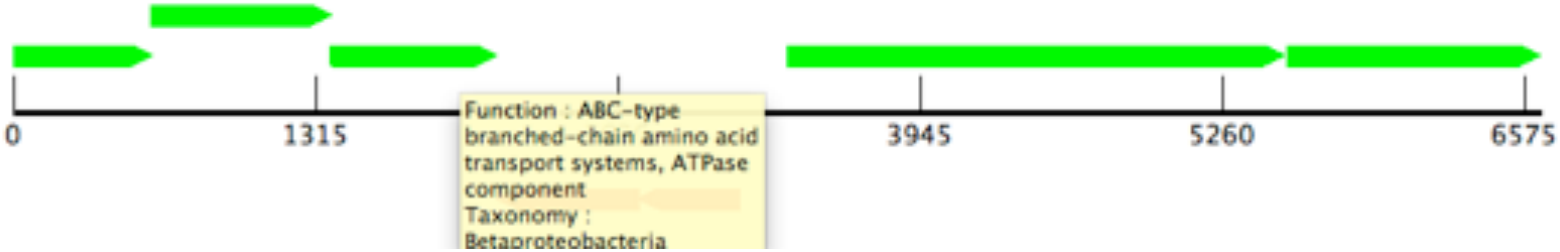
- Template Param File

File Path ▾ : ~/MetaPathways_1_0/input/jgi_illumina/Sep_2013/illumina_param.txt

illumina_param.txt

```
##V.1   do not remove this line
# MetaPathways v1.0
# Kishori M. Konwar, Niels W. Hanson
# Parameter File

INPUT:format fasta
# e.g. fasta gbk-annotated gbk-unannotated gff-annotated gff-unannotated

# Quality Control  parameters
quality_control:min_length  180
quality_control:delete_replicates  yes

# ORF prediction parameters
orf_prediction:algorithm    prodigal
orf_prediction:min_length   60

# ORF annotation parameters
annotation:algorithm last
# e.g. blast or last
annotation:dbs    metacyc-v5-2011-10-21,kegg-pep-2011-06-18,cog-2007-10-30,refseq_protein_2012-11-13,MDM_SAG_proteins
# e.g. annotation:dbs    cog,kegg,refseq,metacyc
annotation:min_bsr  0.4
annotation:max_evalue 0.000001
annotation:min_score 20
annotation:min_length 60
annotation:max_hits 5

# rRNA annotation parameters
# e.g. rRNA:refdbs
rRNA:refdbs GREENGENES_gg16S-2012-11-06,SSURef_111_NR_tax_silva-2012-11-06,LSURef_111_tax_silva_2012
rRNA:max_evalue 0.000001
rRNA:min_identity 20
rRNA:min_bitscore 50

# pathway tools parameters
ptools_settings:taxonomic_pruning no

# grid settings
grid_engine:batch_size 200
grid_engine:max_concurrent_batches 400
grid_engine:walltime 10:00:00
grid_engine:RAM 10gb
grid_engine:user myusername
grid_engine:server mygrid.domain.com

# pipeline execution flags
# e.g. yes, skip, redo
metapaths_steps:PREPROCESS_FASTA    yes
metapaths_steps:ORF_PREDICTION  yes
metapaths_steps:GFF_TO_AMINO    yes
metapaths_steps:FILTERED_FASTA  yes
metapaths_steps:COMPUTE_REFSCORE    yes
metapaths_steps:BLAST_REFDB  yes
metapaths_steps:PARSE_BLAST yes
metapaths_steps:SCAN_rRNA   yes
metapaths_steps:STATS_rRNA  yes
metapaths_steps:SCAN_tRNA   yes
metapaths_steps:ANNOTATE    yes
metapaths_steps:PATHOLOGIC_INPUT    yes
metapaths_steps:GENBANK_FILE    yes
metapaths_steps:CREATE_SEQUIN_FILE  yes
metapaths_steps:CREATE_REPORT_FILES yes
metapaths_steps:MLTREEMAP_CALCULATION   skip
metapaths_steps:MLTREEMAP_IMAGEMAKER    skip
metapaths_steps:PATHOLOGIC  redo
```

1. Quality Control and ORF prediction



- Finds regions in nucleotide sequence which code for an open reading frame (ORF)
- ORFs predicted using PRODIGAL
- Conversion from nucleotide seq to AA sequence e.g. ATG ⟹ MET

Default lengths: 180 nucleotides / 60 AAs

/Output: /pre-processed

/orf_prediction

2. ORF annotation

A. Functional Annotation

B/LAST to compare AA sequences in your query to proteins in user-defined reference databases

e.g. Databases used in NESAP Illumina Analysis:

KEGG, COG, RefSeq, MetaCyc, MDM-SAG-proteins

Also could use CAZy, EggNog, or any other protein database

2. ORF annotation

B. Taxonomic Annotation

Nucleotide sequences are queried against reference nucleotide databases (e.g. SILVA and Greengenes) to identify ribosomal genes in sample metagenomes

- Functional and taxonomic info are combined to generate input files for ePGDB creation

/Output: /genbank (.annotated.gff) → *ePGDB creation*
/blast_results (.blast.parsed.txt)
/results/rRNA (rRNA.stats.txt)
/results/annotation_tables (.fxn_and_taxa_table.txt)

3. Analyses

A. tRNA Scan to identify relevant tRNAs

/Output: /results/tRNA

B. Least Common Ancestor for taxonomic binning

/Output: /results/LCA

C. ML TreeMap

/Output: /results/mltreemap

D. ePGDB creation

Input file: .annotated.gff

/Output: /ptools/

Current Metapathways Outputs

① Individual Sample Level

① Auxilary Sample Statistics

- Nucleotide length distribution, AA length distributions

② Functional and Taxonomic Table

③ Taxonomic distribution of sample in NCBI tree

④ Functional characterization at KEGG & COG levels

★ ⑤ Pathway Table

② Multi-Sample Comparison Level

★ ① Master pathway table

Sample Pathway Table

- Created after pipeline run using:

extract_pathway_table_from_pgdb.pl

/Output: /results/pgdb/XXX.basepathways.txt

File Path : ~/MetaPathways_1_0/output/fos_ends/Sep13_2013/a4_10/results/pgdb/a4_10.basepathways.txt

a4_10.basepathways.txt

PWY_NAME	PWY_COMMON_NAME	NUM_REACTIONS	NUM_COVERED_REACTIONS	ORF_COUNT					
PWY0-1312	acetate formation from acetyl-CoA I	2	1	3	a4_10_46_0	a4_10_1224_0	a4_10_5932_0		
PWY-6961	L-ascorbate degradation II (bacterial, aerobic)	8	1	3	a4_10_1904_1	a4_10_11058_0	a4_10_7394_0		
PWY-5794	malonate degradation I (biotin-independent)	3	1	3	a4_10_5943_0	a4_10_11134_0	a4_10_2927_0		
PWY-5162	2-oxopentenoate degradation	3	2	6	a4_10_342_1	a4_10_11189_0	a4_10_8802_0	a4_10_9514_0	a4_10_802_0 a4_10_37

ROWS
pathways

COLUMNS

(1) Pathway short name
(2) Pathway long name
(3) # of reactions needed to complete a pathway (from MetaCyc)
(4) # of distinct reactions covered in a sample
(5) # of ORFs found in a pathway in a sample
(6) Names of ORFs found in each pathway in each sample

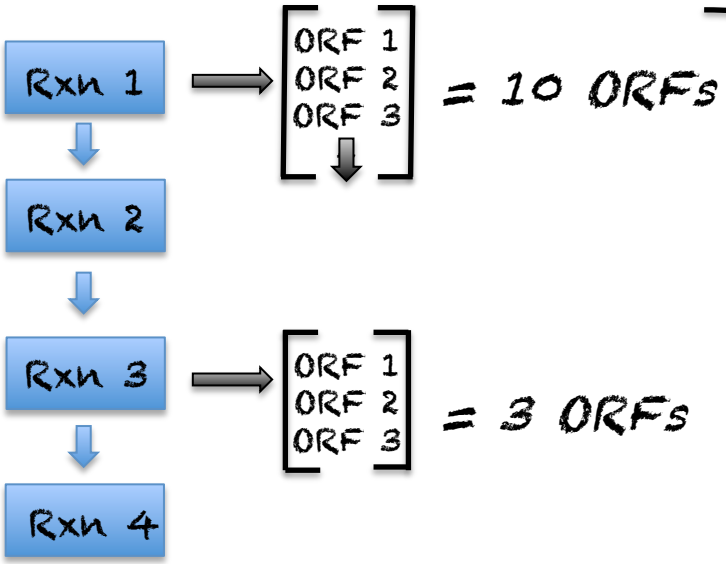
Example: Sample Pathway Table

ROWS
pathways

COLUMNS

(1) Pathway short name
(2) Pathway long name
(3) # of reactions needed to complete a pathway (from MetaCyc)
(4) # of distinct reactions covered in a sample
(5) # of ORFs found in a pathway in a sample
(6) Names of ORFs found in each pathway in each sample

Example: sample A, pathway X



Total 13 ORFs covering 2 distinct reaction steps

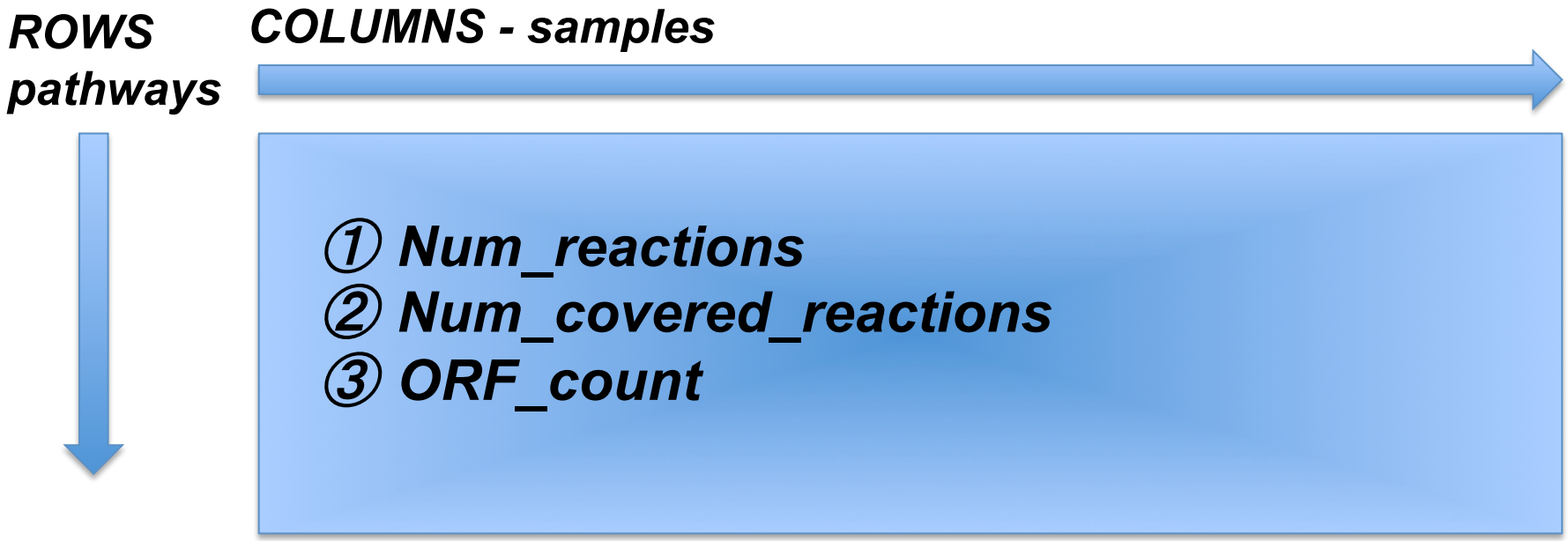
For Sample A, pathway X

Column (3) = 4
Column (4) = 2
Column (5) = 13
Column (6) = ORF 1, ORF 2, ...

Master Pathway Table

- Created after pipeline run using:

./make_table.pl



Text files that you can import to program of choice for analysis (e.g. R, Matlab, Excel)

A pathways level analyses of the NESAP metagenomes

Run through Metapathways:

☑ 49 Saanich and Line P fosmid ends

☑ 91 Saanich and Line P Illumina samples

Template Param File Statistics:

LAST algorithm

Libraries used: *KEGG-pep-2011-06-18*

COG 2007-10-30

Metacyc v5 2012-10-21

RefSeq protein 2012-11-13

MDM-SAG proteins

Master Pathway Table

- Created after pipeline run using:

/make_table.pl

ROWS
pathways

COLUMNS - samples

① Num_reactions
② Num_covered_reactions
③ ORF_count

Matlab Workflow I

1. Import master tables

2. Plot number of nucleotide sequences & ORFs before, after, and difference for each sample

- *Calculate variability in ORFs after across the samples to decide effect of relativizing by library size*

3. Calculate variability in pathway length (number of reactions) for each pathway in dataset

- *Done to decide if need to relativize by pathway length*

Note: #2 & 3 only matter if using quantitative data

Matlab Workflow II

To get a quick sense of the data:

4. Calculate variability in:

(i) Total ORFs in the samples down the pathways
(ii) Total ORFs in the pathways across the samples

5. Venn diagrams of shared and exclusive pathways

Matlab Workflow III

Statistical Analyses:

5. Make distance/dissimilarity matrix
6. NMS (aka NMDS)
7. Monte Carlo to check dimensions on NMS
8. Cluster Analysis
9. Define groups using either NMS or cluster analysis
10. Run ISA to determine pathways which are causing the groups to differentiate

Hallam Lab Timeseries: NESAP Metagenomes

- **49 Fosmid End Libraries:**

Saanich = 19 samples from 2004-2007

Line P = 30 samples from 2009-2010

Unassembled

➢ *Can use quantitative ORF counts*

- **91 Illumina Samples:**

Saanich = 48 samples from 2009-2011

Line P = 43 samples from 2008-2010

Assembled by JGI using Velvet

➢ *Cannot use quantitative ORF counts at present*

Hallam Lab Timeseries: NESAP Metagenomes

- **49 Fosmid End Libraries:**

Saanich = 19 samples from 2004-2007

Line P = 30 samples from 2009-2010

Unassembled

➢ *Can use quantitative ORF counts*

- **91 Illumina Samples:**

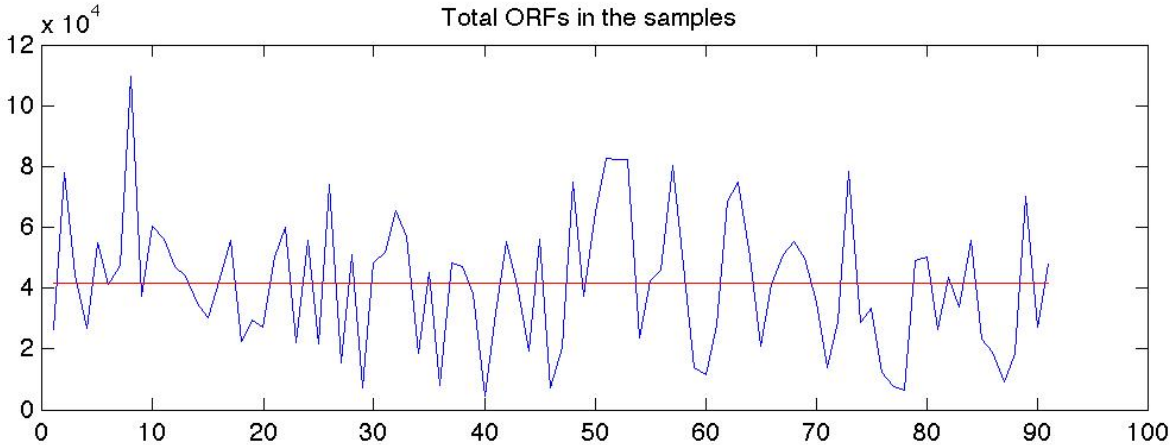
Saanich = 48 samples from 2009-2011

Line P = 43 samples from 2008-2010

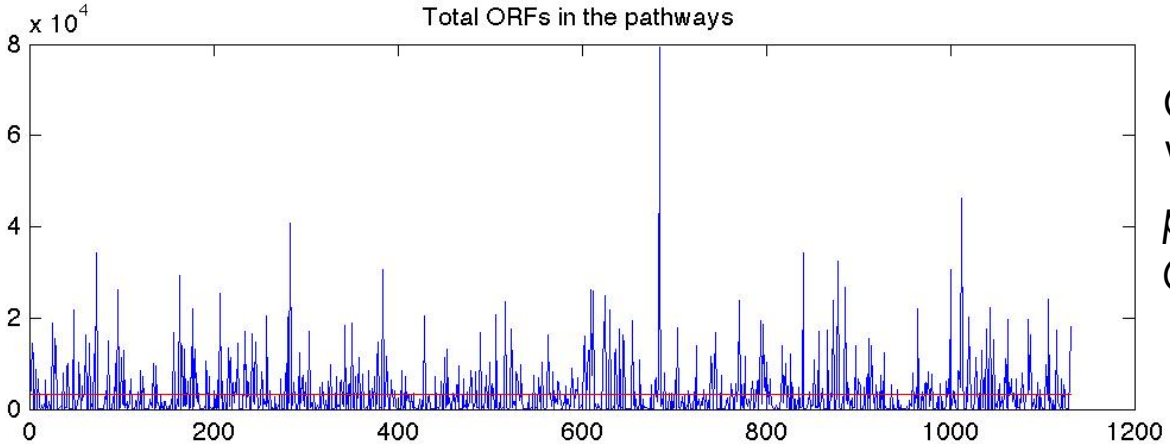
Assembled by JGI using Velvet

➢ *Cannot use quantitative ORF counts at present*

NESAP Illumina Dataset: Quick Look



Coefficent of Variation (CV) 100*(Std/Mean): 52%
Variability among the samples based on total ORFs is moderate

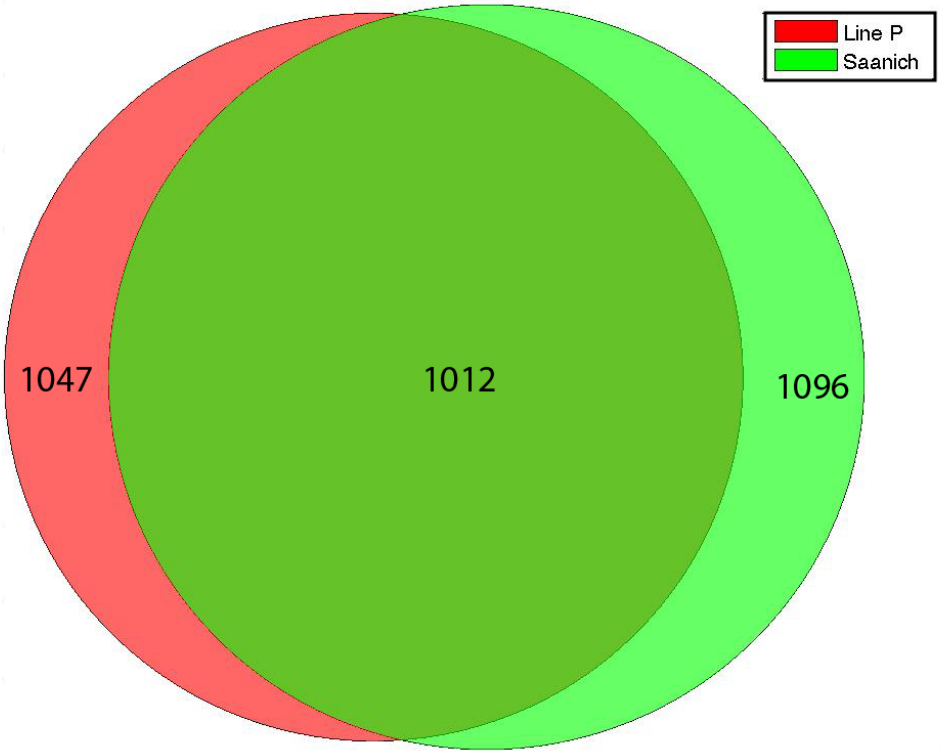


CV = 182%
Variability among the pathways based on total ORFs is large

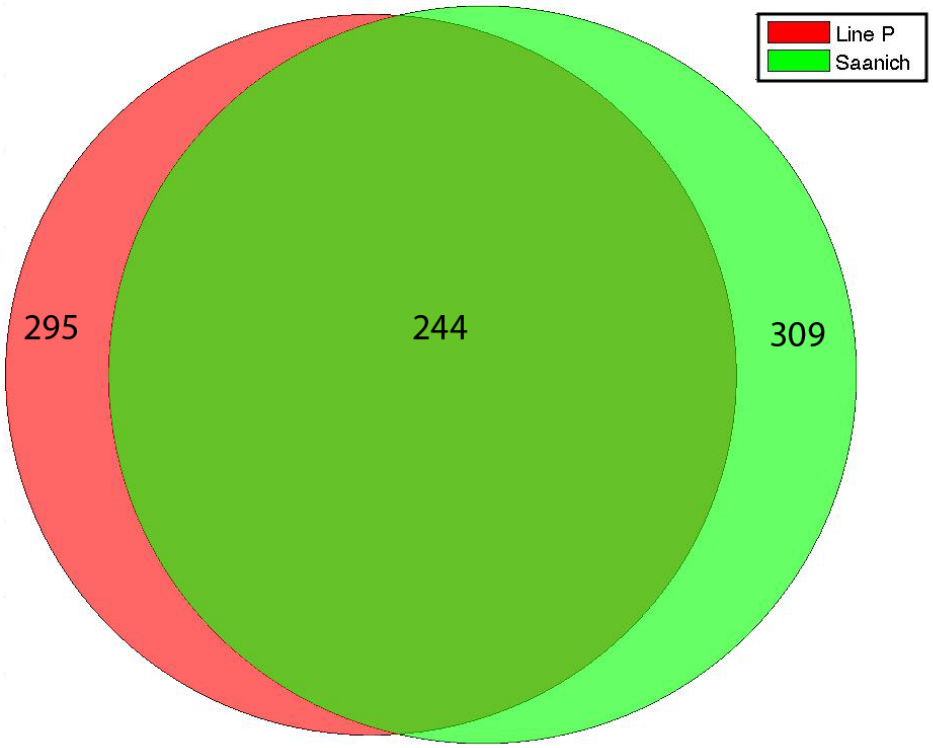
- 1131 pathways resolved
- ~2000 pathways in Metacyc
- Maximum resolved using MetaPathways is 1239
- Illumina dataset is resolving 91.3% of max using MetaPathways

NESAP Illumina Dataset: Shared & Exclusive Pathways

Pathway Distribution:
1012 pathways shared at some point



Core Pathways:
244 pathways always present

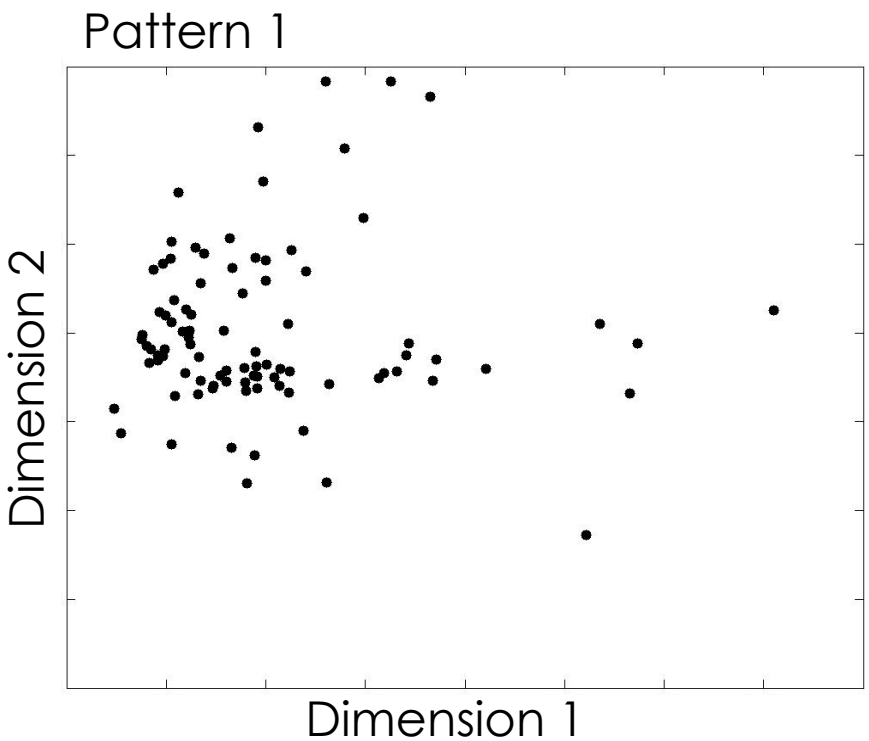


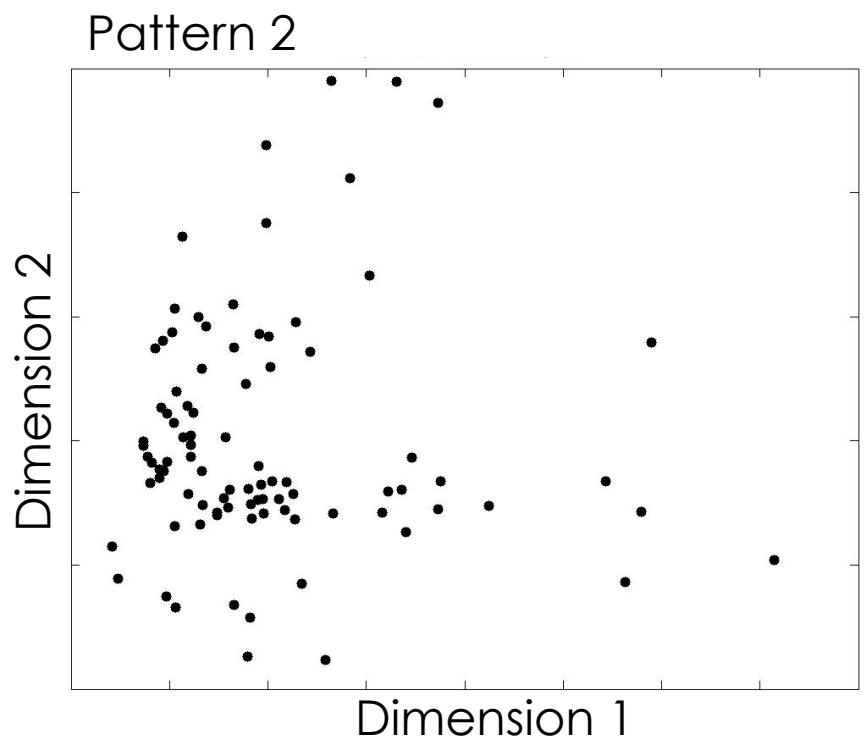
- Follows **119 pathways are NOT shared**
- Of which **35 are exclusive to Line P**
- And **84 are exclusive to Saanich**

NESAP Illumina Dataset: NMS

Many NMS runs consistently reveal a stable configuration, resulting in 1 of 2 similar patterns

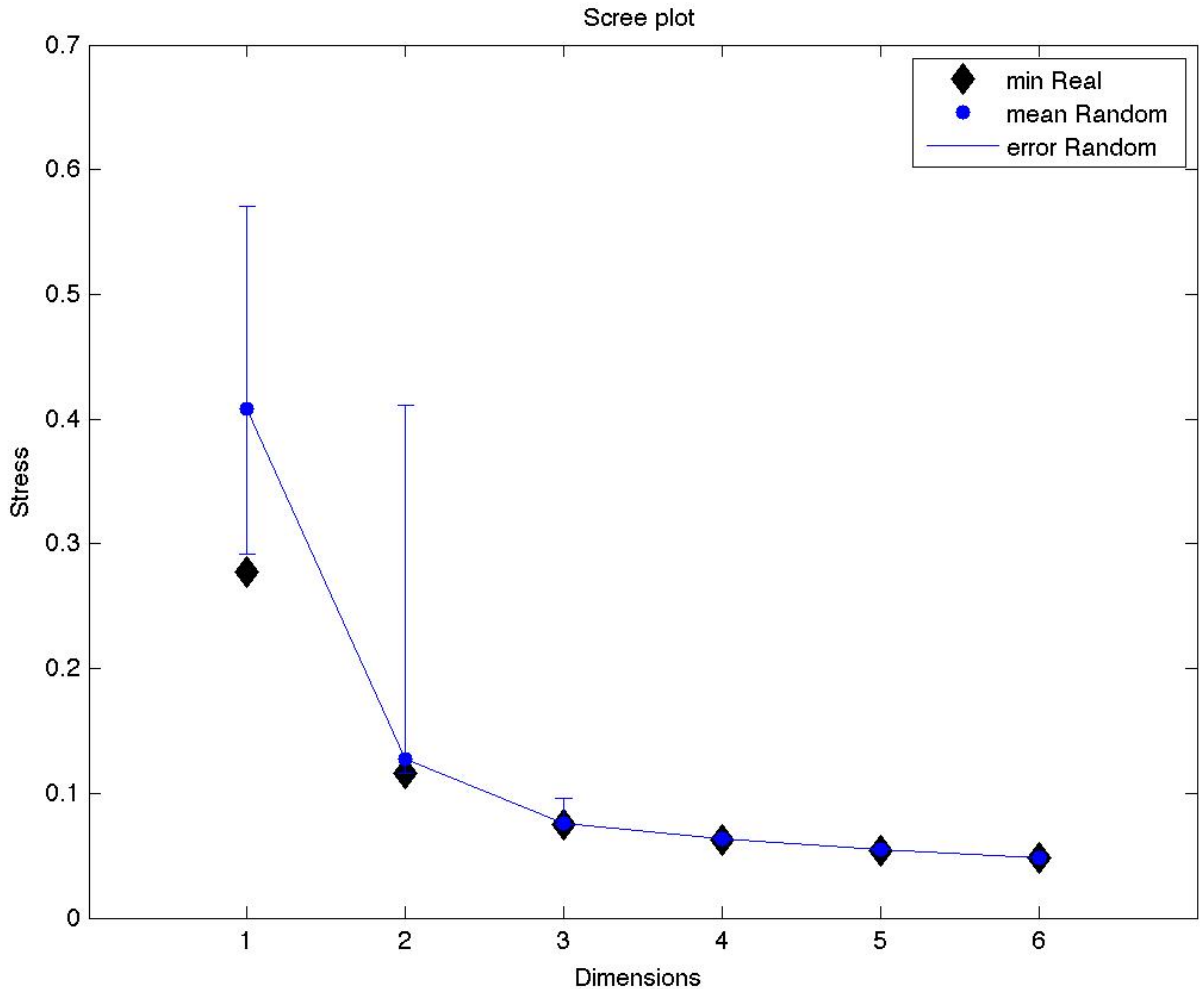
- Average stress 1 x 100 ~ 12 (good/fair)



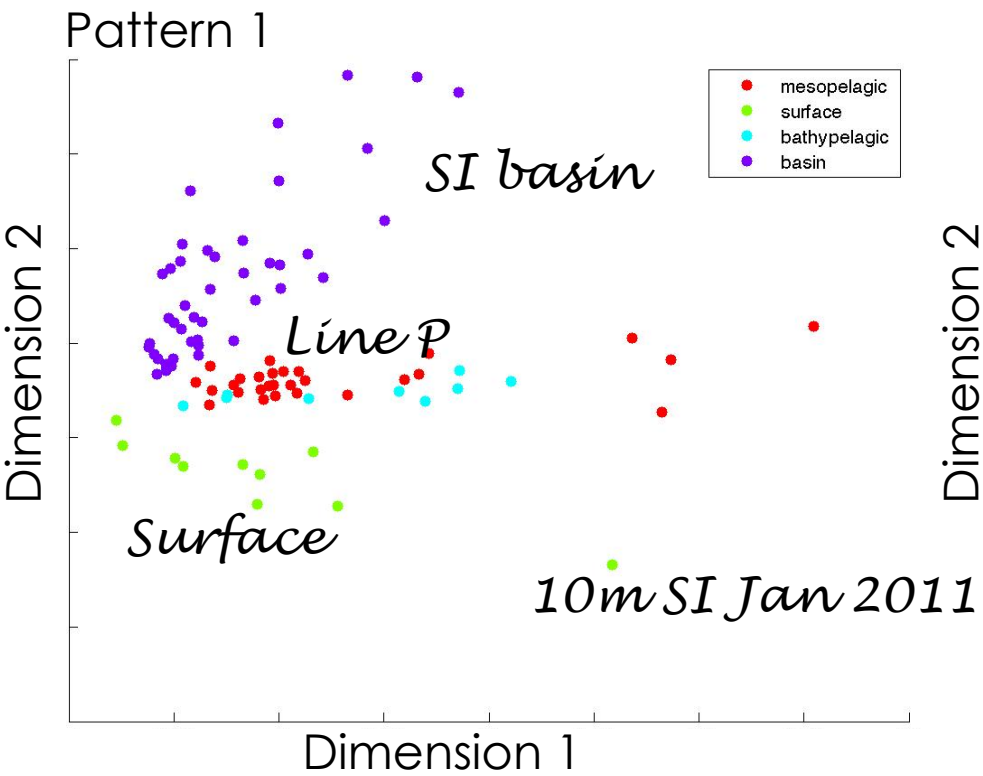


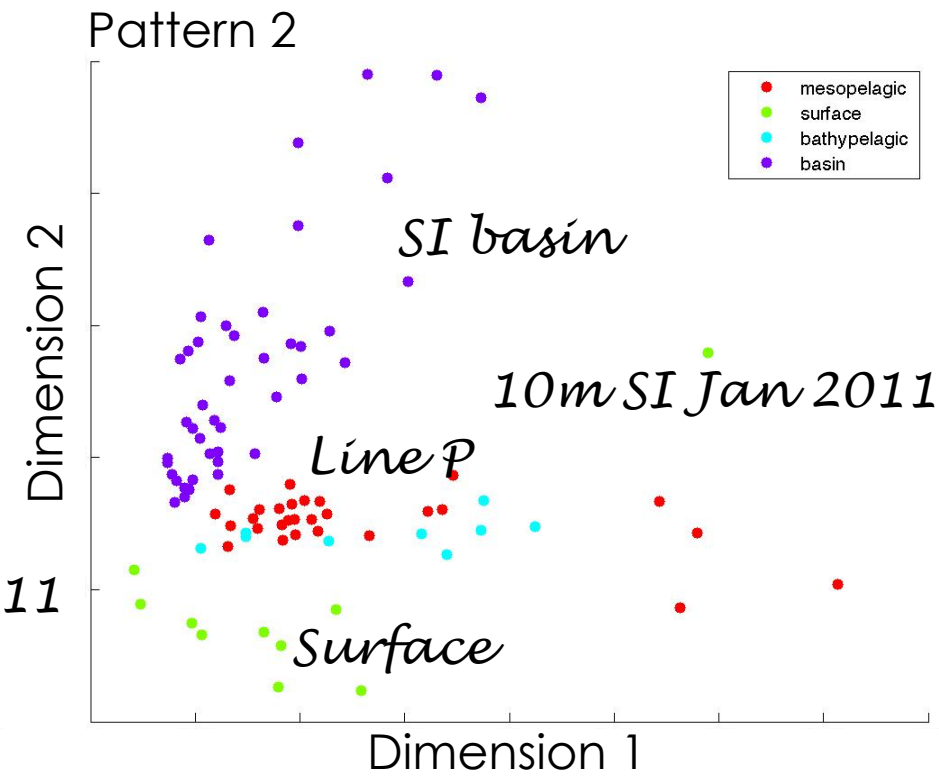
NESAP Illumina Dataset: NMS

Monte Carlo reveals 2 dimensions for jgi illumina produces the greatest reduction in stress:



NESAP Illumina Dataset: NMS Depth



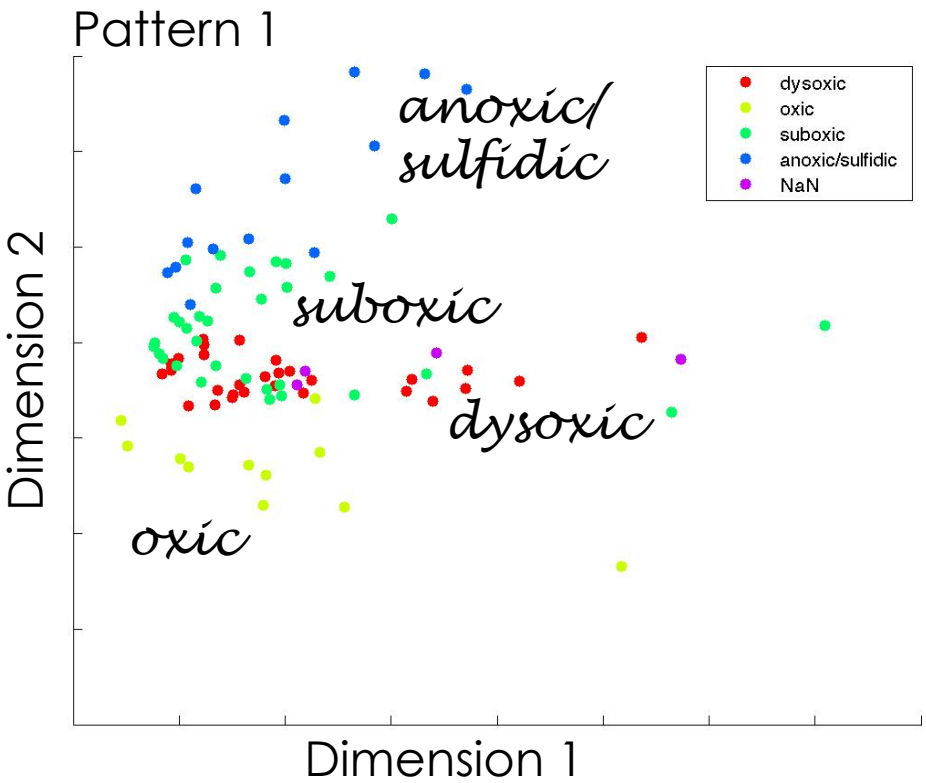


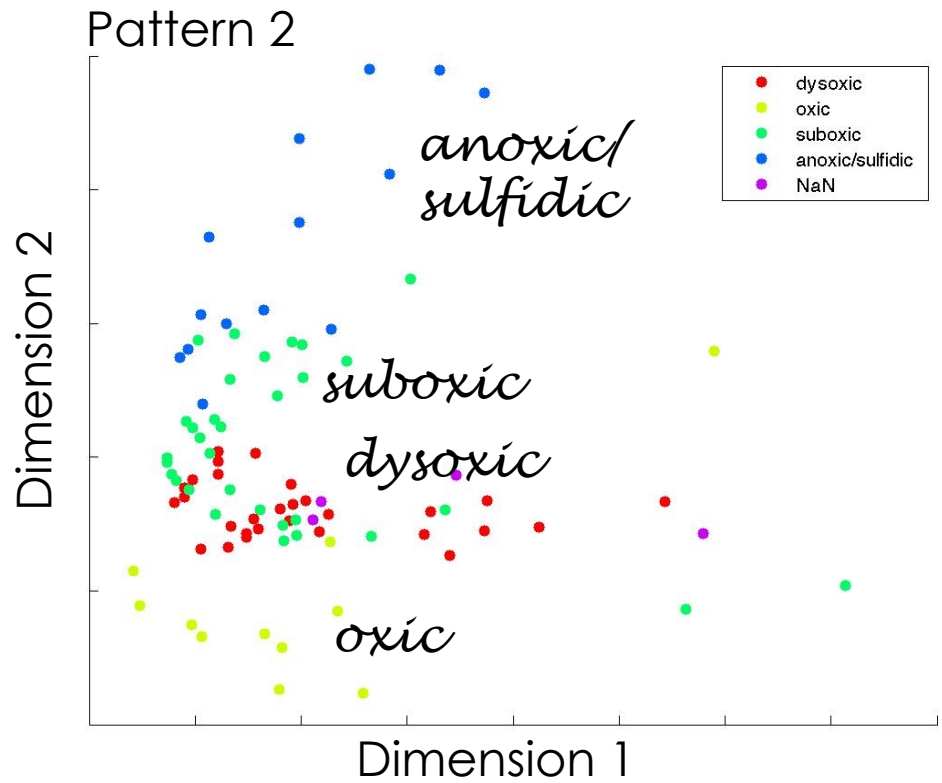
Patterns 1 & 2 give the same results

✔ *Pattern with depth*

Pathways are different between samples at different depth classes

NESAP Illumina Dataset: NMS Oxygen

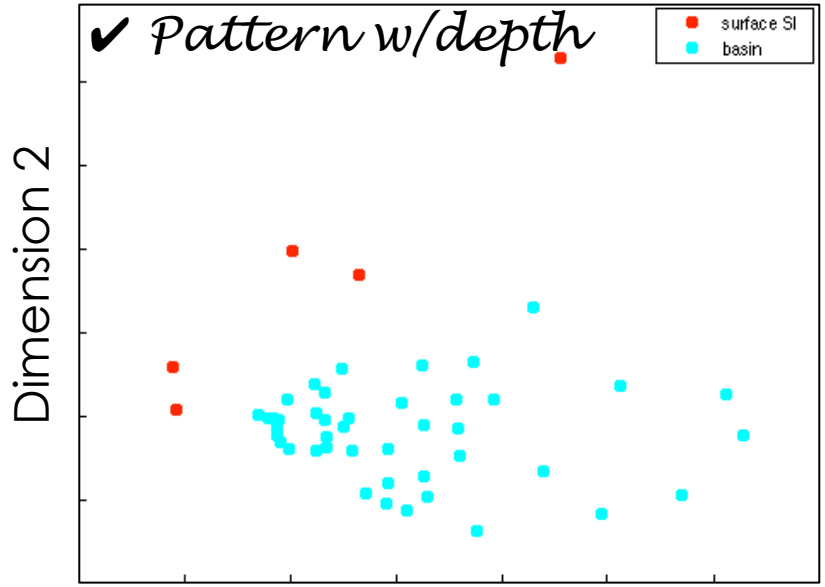


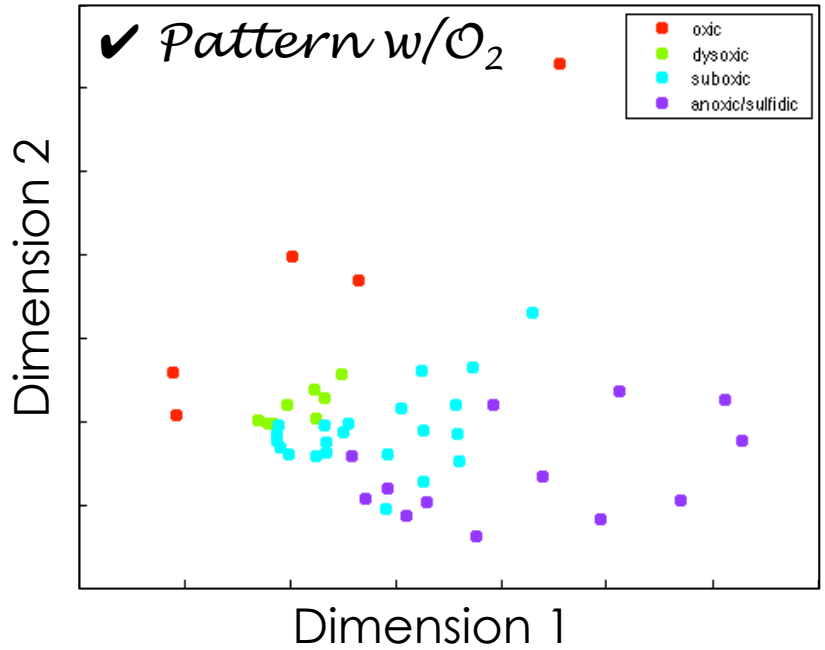


Patterns 1 & 2 give the same results

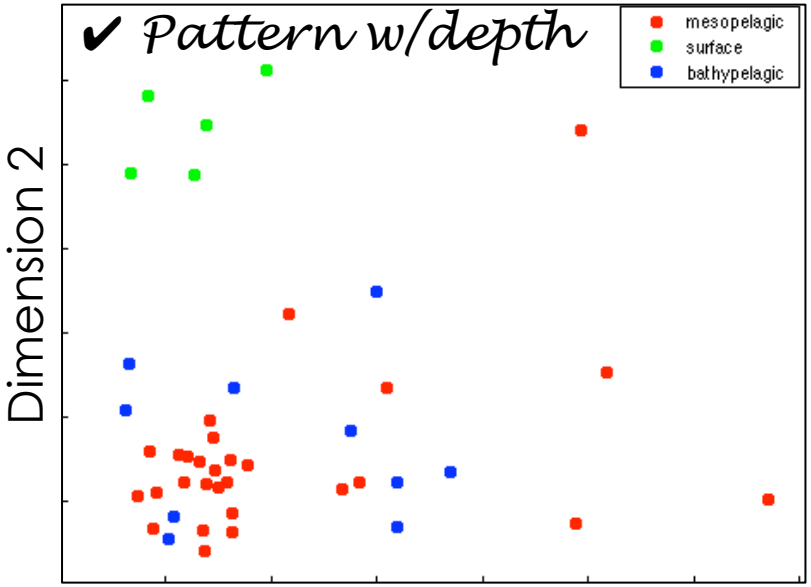
✔ *Pattern with oxygen*

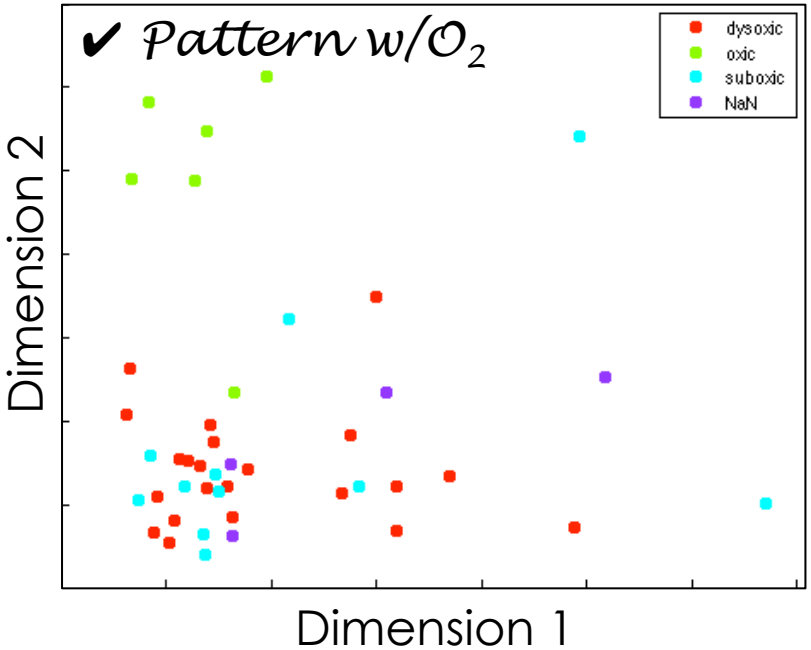
Saanich



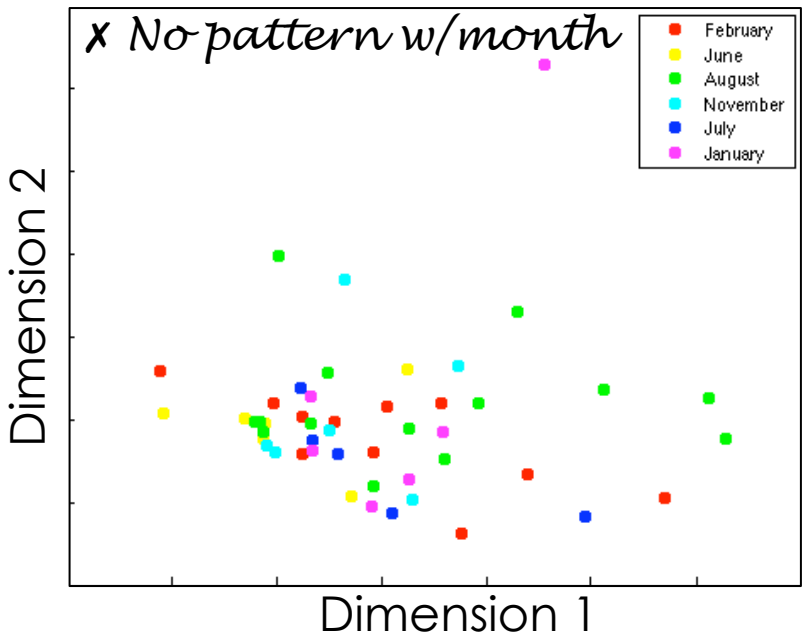


Line P

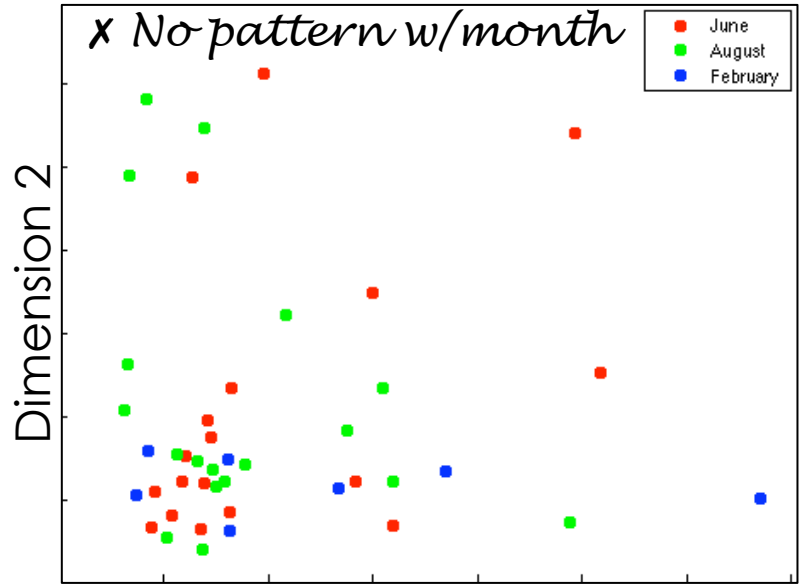


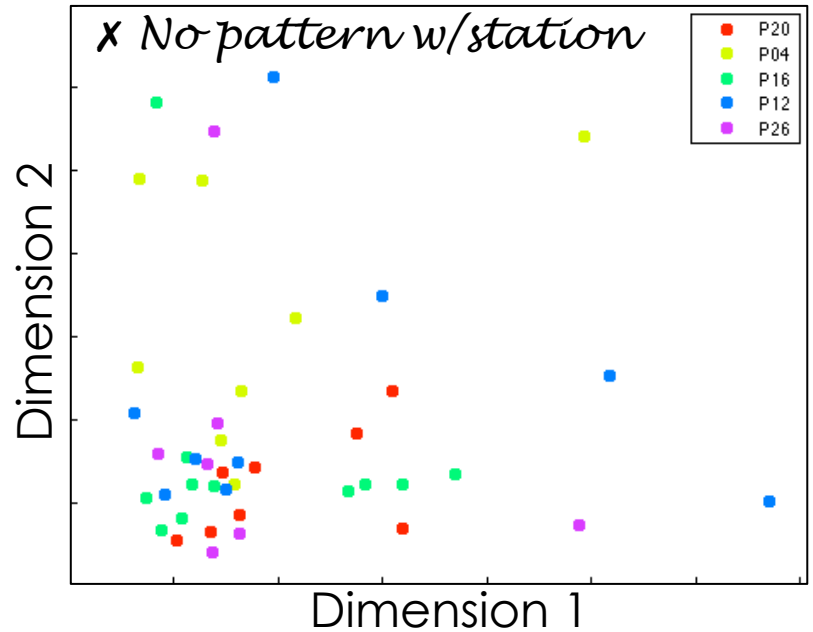


Saanich



Line P





Summary of NMS Trends for NESAP Pathways

Groups	Depth	O_2	Year	Month	Station	NO_3^-	H_2S
LP	✔	✔	✗	✗	✗	possibly	N/A
SI	✔	✔	✗	✗	N/A	possibly	✔
LP + SI	✔	✔	✗	✗	N/A	possibly	N/A

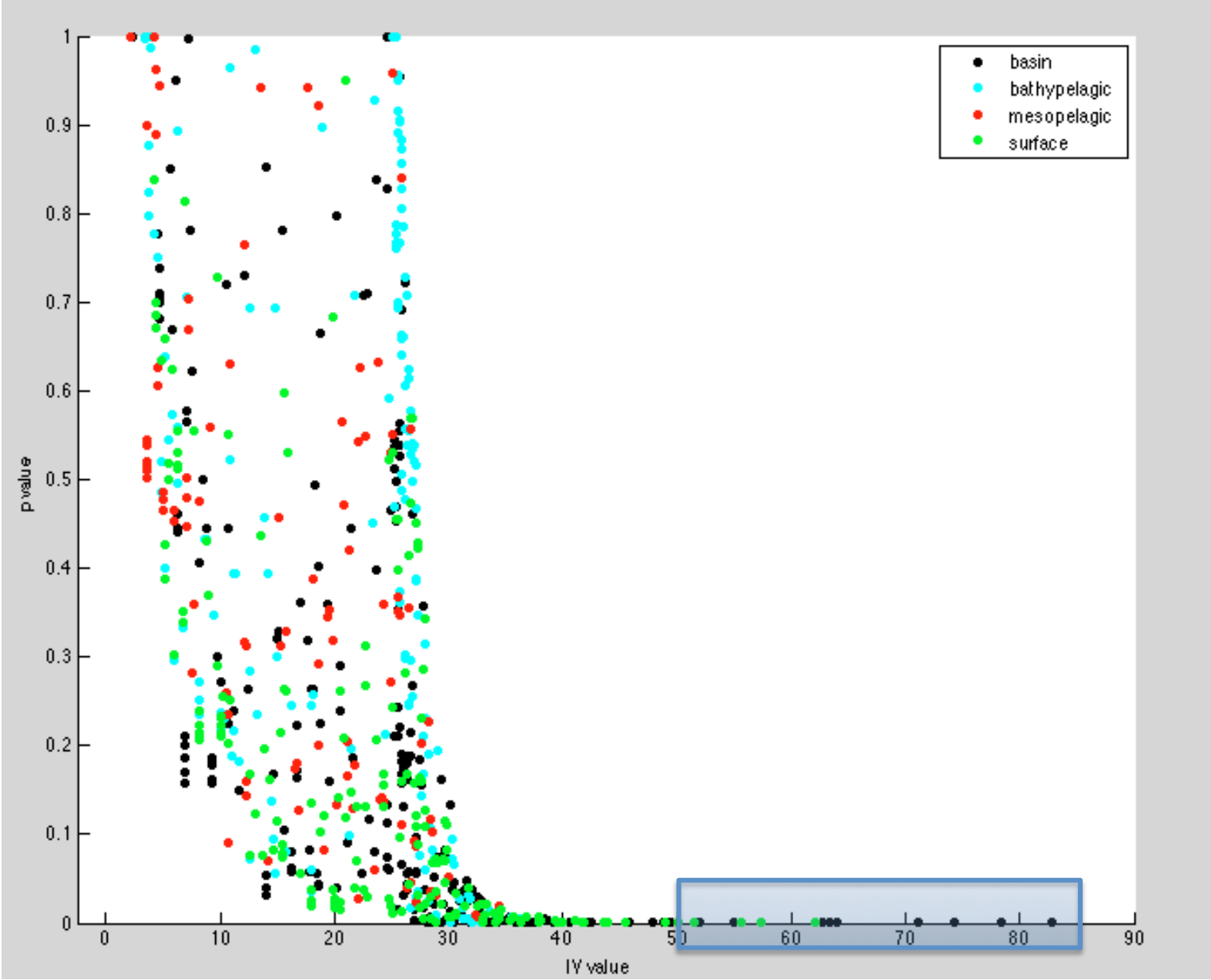
- Pathways are different between samples at different depths & between samples with different oxygen concentrations
- Statistical basis for using groups for ISA based on 4 depth and oxygen classes

NESAP Indicator Species Analyses

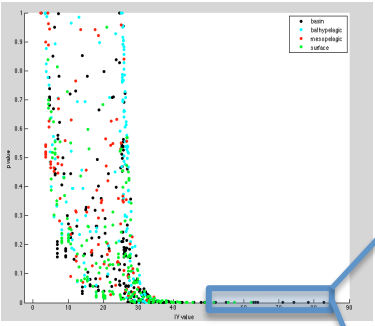
- Will hopefully tell us which of the 1131 pathways are causing the groups to differentiate
- To be a perfect "indicator" pathway for a group, a pathway needs to be both *exclusive* (always present in that group) and *faithful* (only present in that group)

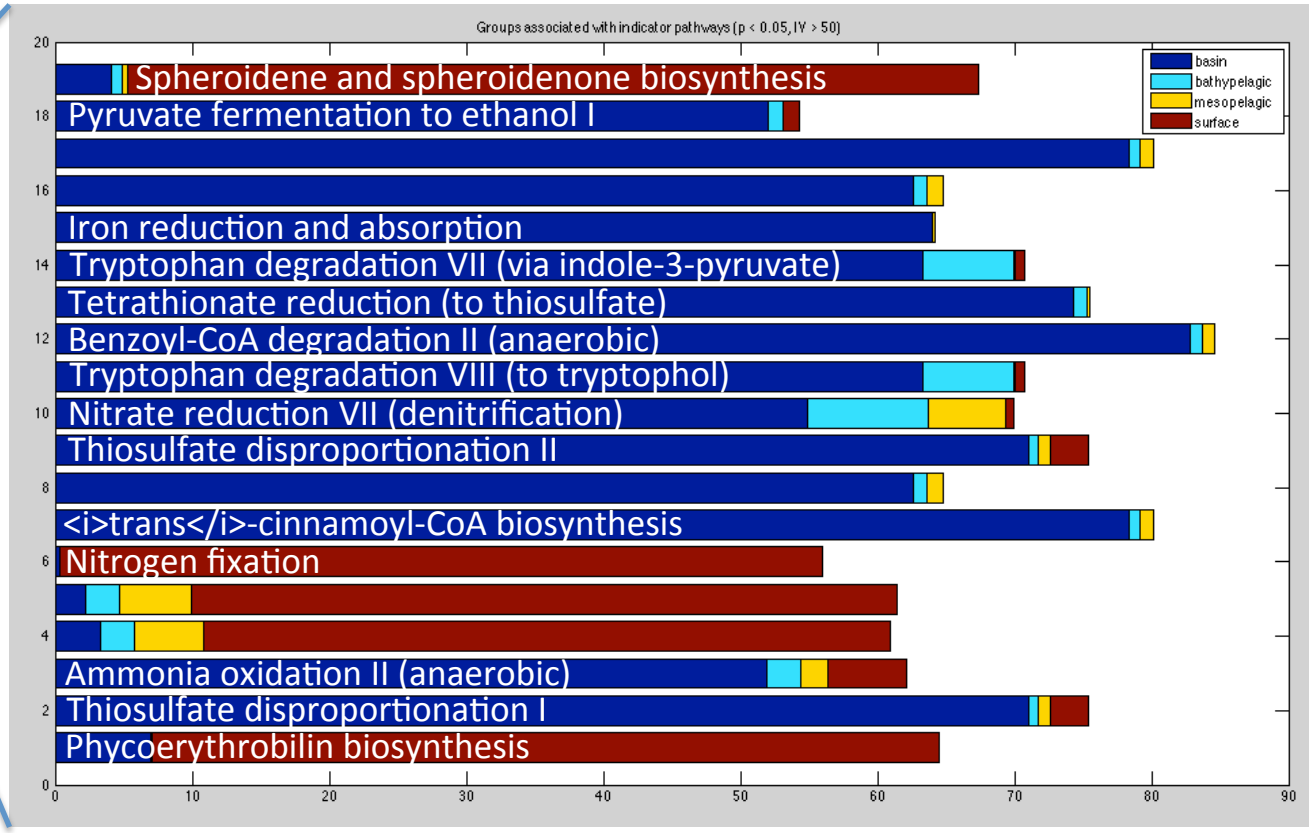
- *We'll see how successful this analysis is considering the amount of shared pathways!*
- *As a caveat, this analysis will be more successful with quantitative data*

Indicator Species Analysis (ISA)



Indicator Species Analysis (ISA)

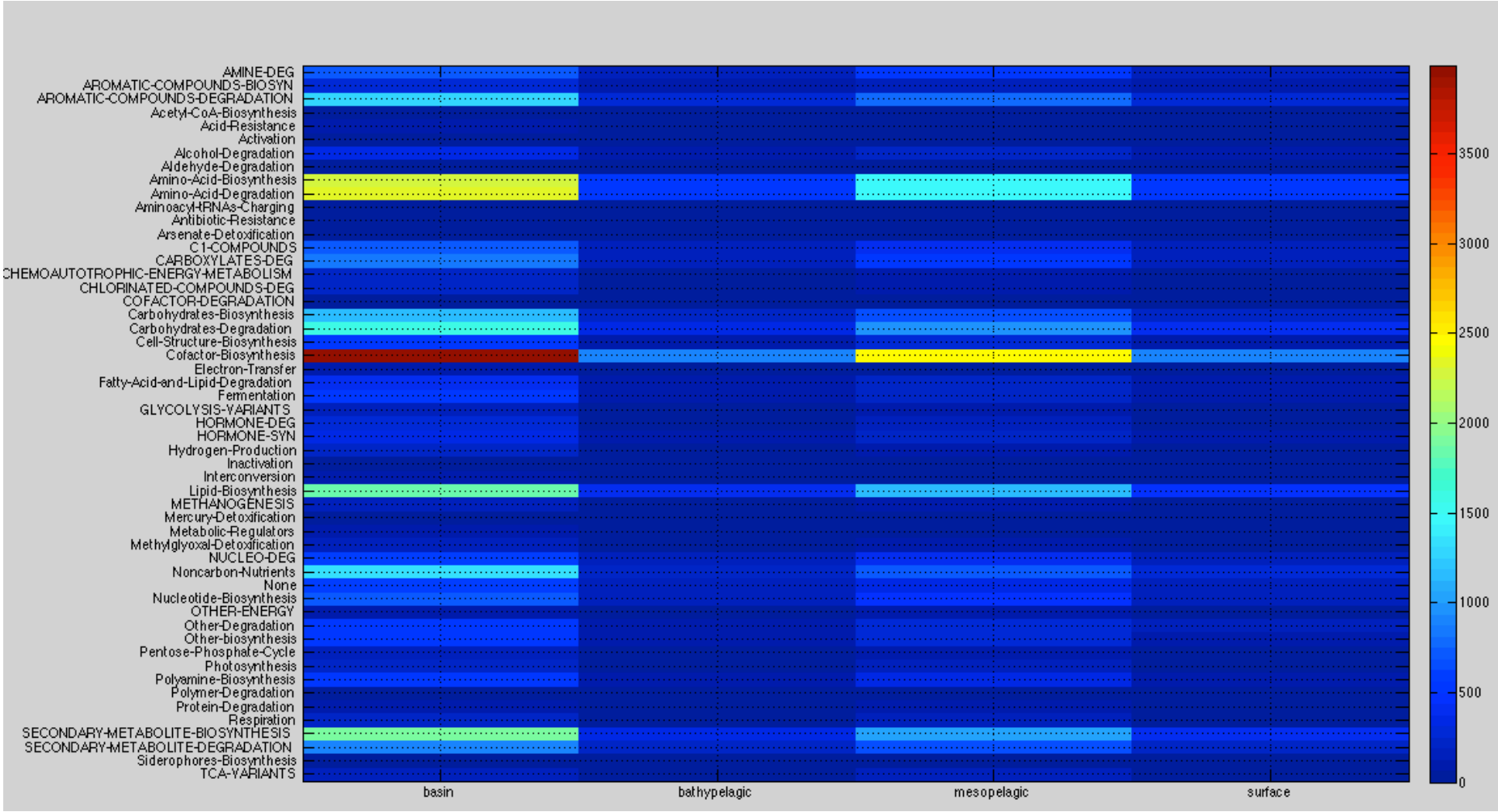




19 'indicator' pathways: 5 for the surface, 14 for the basin

We can consider pathways major metabolic traits associated with two end-member environments (aerobic sunlit, anaerobic dark)

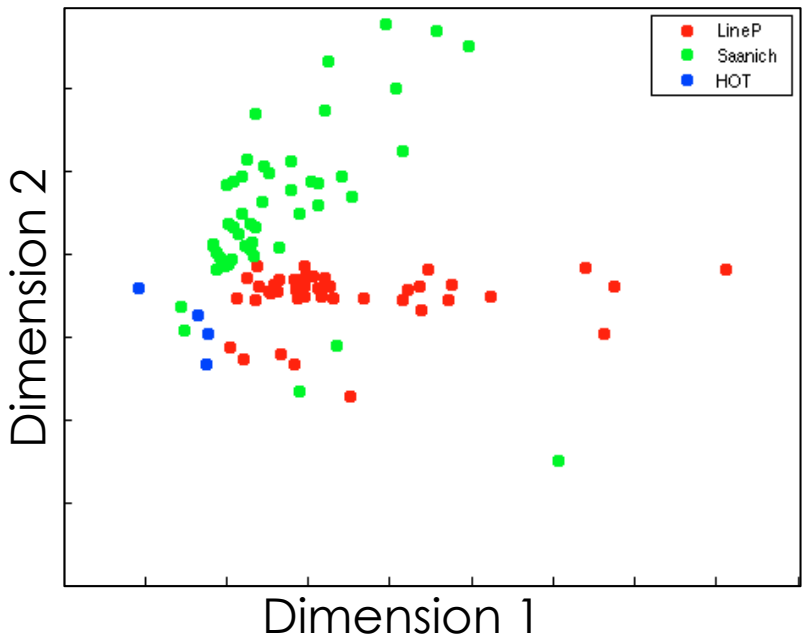
Quick Heatmaps

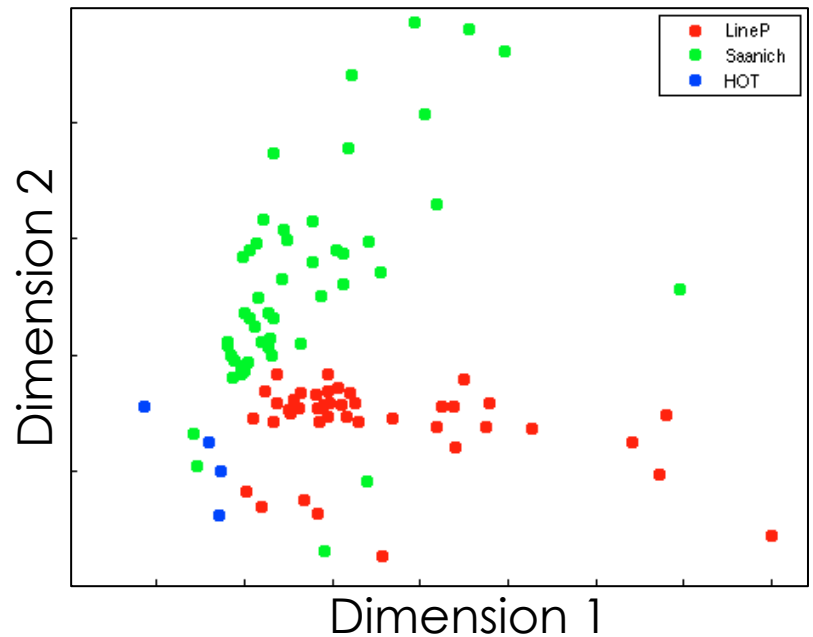


Adding in HOT…

- HOT 454 data from Cruise 186 October 2006
- 4 samples: 25m, 75m, 110m, 500m
- Combined master table (SI + LP + HOT):

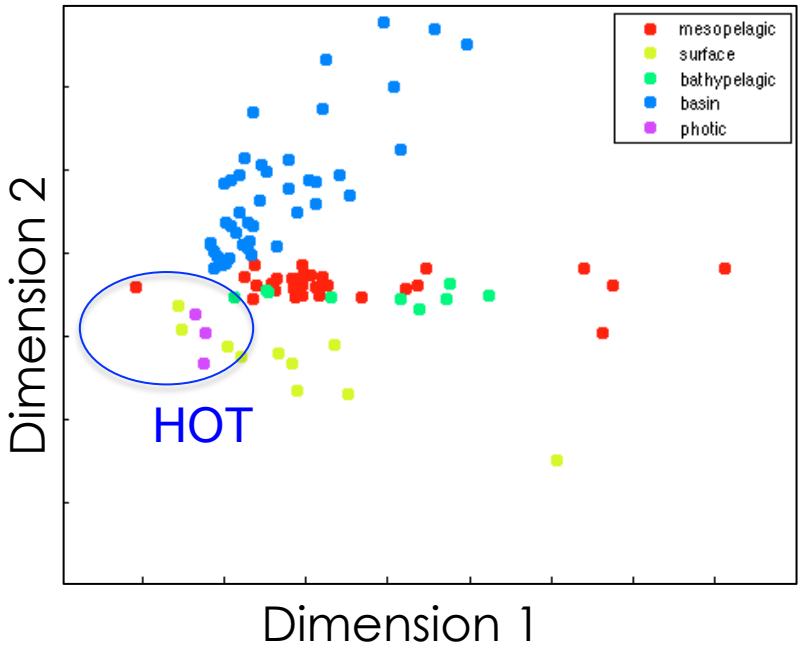
➢ 1144 pathways (+13 unique to HOT)

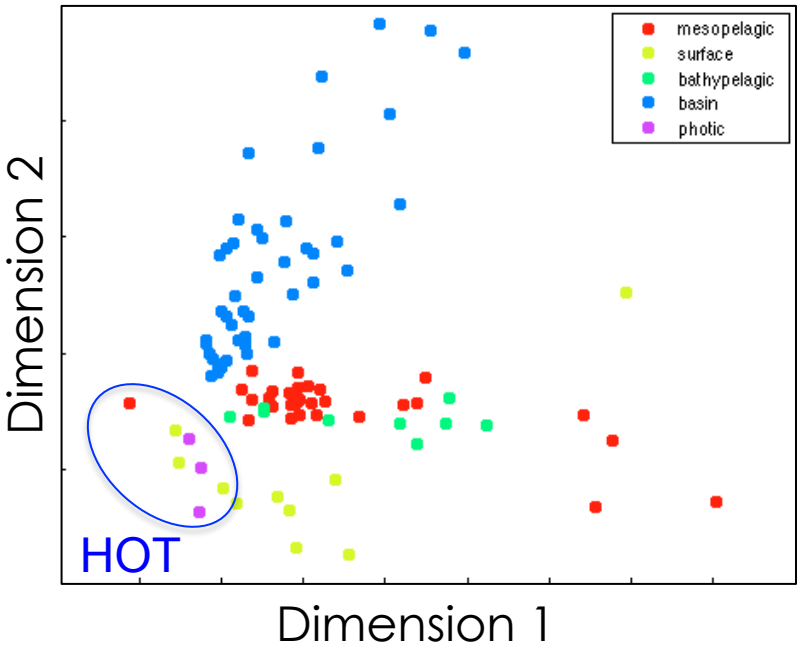




- Again, NMS reveals same patterns 1 & 2
- Stress is also similar, ~12 (good/fair)

Adding in HOT…





- HOT samples grouping with surface NESAP

➢ *Pathways in HOT most similar to those also found in NESAP surface*

Questions?