Identifying biologically relevant differences between metagenomic communities

Donovan H. Parks and Robert G. Beiko

Journal of Bioinformatics

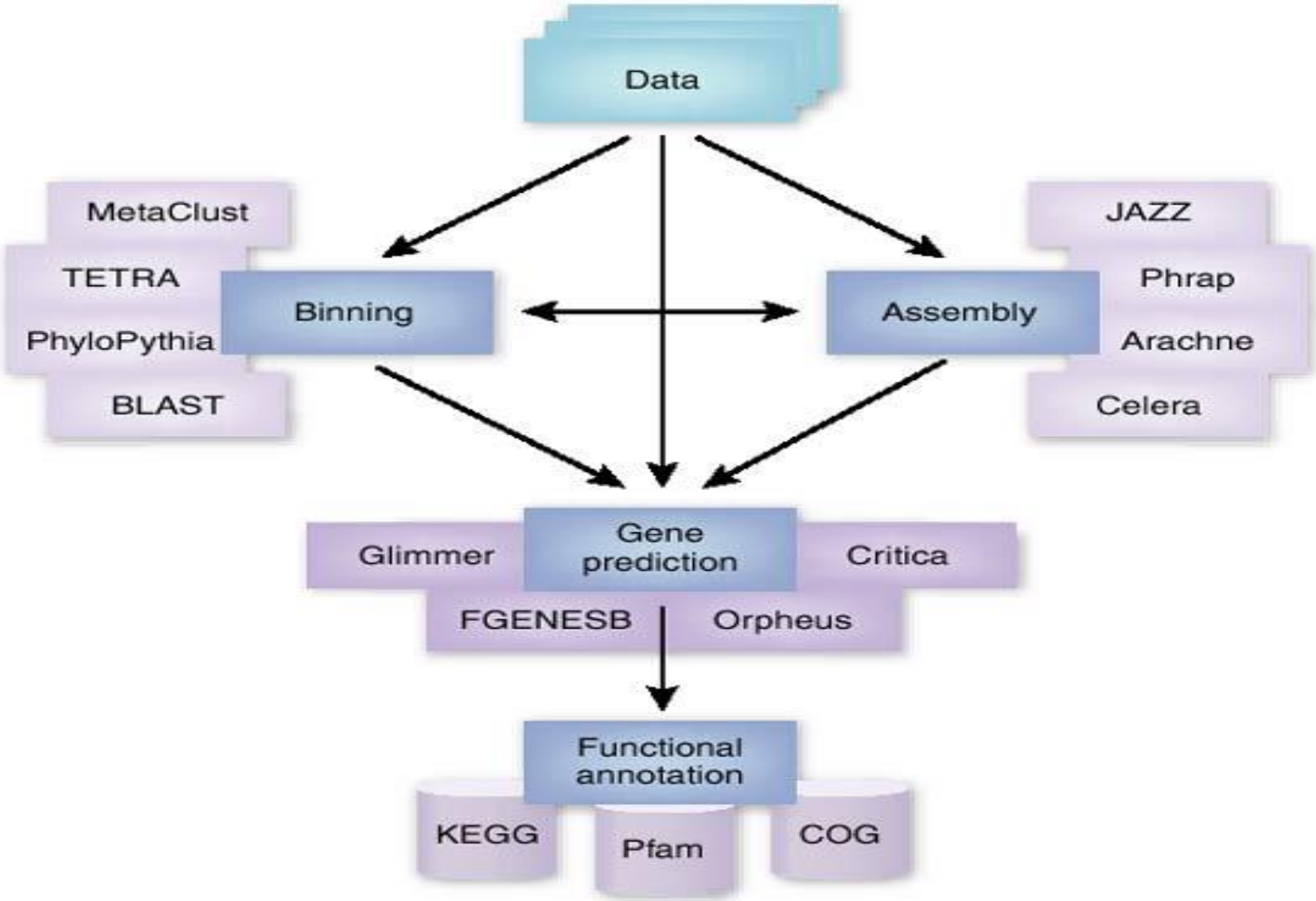
Presented By:

Rajeswari Swaminathan

Contents

- Review
- Problem Statement
- Different Approaches taken
- Implementation
- Results
- Discussion
- Conclusion

Background


Data
MetaClust
TETRA
PhyloPythia
BLAST
Binning
Assembly
JAZZ
Phrap
Arachne
Celera
Glimmer
Gene prediction
Critica
FGENESB
Orpheus
Functional annotation
KEGG
Pfam
COG

Comparing different Metagenomics communities

To see how two metagenomic samples, from different habitats differ from one other with respect to:

- Distribution of sequences for a particular functional profile
- Distribution of organisms across the samples
- Common functionalities across the samples
- Inference on how the environment affects the distribution of functions

Problem statement

Given a functional or taxonomic profile of a pair of metagenome datasets, to identify relevant statistical as well as biological differences between the communities

Approaches taken so far …

Some of the earlier developed tools that performed similar analysis are:

- UniFrac(2005) : This tried to compute microbial diversity based on phylogenetic information
- XIPE – TOTEC(2006): This tool identified statistically different subsystems in different metagenome communities and improvised on the significance by incorporating the non parametric bootstrap test into their analysis, along with the traditional p-value analysis.
- ShotgunFunctionalizeR(2009): An R package for identifying and assessing statistical differences between samples for different pathways and subsystems, based on a Poisson model
- MEGAN(2009): An extension of the already existing tool MEGAN for categorizing sequence reads into OTU's. This new tool allows visual as well as statistical comparisons of large metagenome datasets

Limitations of earlier approaches

- Almost all the approaches taken so far have been only based on computing p-values for some test statistic chosen

- Secondly, no thought has been given to fact that statistical significance does not necessarily always reflect biological relevance

- Thirdly, many other important parameters, such as the size of the individual samples, effect sizes, confidence intervals etc. have not been dealt with in detail

The magical solution !!!!

- This paper implements a novel method STAMP (Statistical analysis of Metagenomic Profiles) for comparing pairs of metagenomes
- Provides a couple of different statistical tests (Fisher's exact test, Bootstrap analysis etc.) that can be performed on the data
- The results obtained not only indicate the statistical enrichments of different subsystems across the samples, but also provides deeper biological insights
- STAMP can be downloaded freely from

  http://kiwi.cs.dal.ca/Software/STAMP

Input data to STAMP

- The input data to STAMP can be any count data obtained from a pair of metagenome samples.
- Usually, the interest is on a particular collection of features(that may define a certain subsystem or pathway relevant to the environment)
- Assessment is carried out using contingency tables

	Sample 1	Sample 2	
Sequences in feature	x_1	x_2	$R_1=x_1+x_2$
Sequences in other features	y_1	y_2	$R_2=y_1+y_2$
Total assigned sequences	$C_1=x_1+y_1$	$C_2=x_2+y_2$	$N=C_1+C_2$

Different Statistical analysis performed

The different statistical analysis that can be performed using this tool are:

- Assessing the P-value significance by doing sampling with and without replacements

- Further biological significance can be assessed by considering the magnitude of the observed difference using “effect size statistics”

- Lastly, by obtaining the confidence intervals for the observed P-values and effect sizes

Errors with sampling

- In order to provide a statistical significance to an observed P-value ,rather than calling it a meager sampling error, Rodriguez *et al*. had suggested the use of the bootstrap analysis
- In this analysis, he randomly picked two samples of M sequences from a pooled set containing all sequences
- This process was repeated multiple times for obtaining a null distribution
- The only limitation with this analysis was that with increase or decrease in the value of M, the number of significant features obtained also changed
- So, for an observation to have been caused by sampling error, the individual sequences in the sample sets must also be considered

Effect of sample sizes on the number of significant features identified

Statistically significant subsystems	***M* = 2319**	***M* = 7700**	***M* = 13,221**
p-value ≤ 0.05	48	118	148
p-value ≤ 0.01	34	100	126

Sampling without replacement

- A set of Monte Carlo permutation tests are carried out for modeling the null distribution of a test statistic

- In this case, a permutation of the sequences from a pair of metagenomic samples is considered as drawing sequences with replacement, which produces a hypergeometric distribution

- The P-values can then be directly calculated from the distribution using Fisher's exact test

- Methods using chi-square and G tests have shown small variations for cases in which the sample sizes are unequal

Comparison of statistical hypothesis tests to Fisher's tests

	Chi-square (Diff. b/w prop.)	**Chi-square w/ Yates**	**G-test**	**G-test w/ Yates**	**Permutation**	**Bootstrap**
C_1 = 100 (618)	-16±8.3% [-35%; -0.47%]	17±11% [0.31%; 52%]	-17±8.0% [-42%; 3.2%]	17±11% [-5.3%; 55%]	2.0±23% [-69%; 121%]	-11±11% [-48%; 24%]
C_1 = 200 (2204)	-12±6.9% [-41%; 1.9%]	13±8.9% [-8.1%; 51%]	-13±7.1% [-41%; 4.3%]	12±9.0% [-7.8%; 54%]	0.8±21% [-72%; 119%]	-8.0±9.9% [-46%; 31%]
C_1 = 500 (2677)	-10±6.6% [-39%; 13%]	11±8.6% [-7.5%; 65%]	-10±7.8% [-41%; 5.5%]	11±9.3% [-8.8%; 60%]	1.1±21% [-72%; 118%]	-6.9±9.5% [-41%; 35%]
C_1=1000 (2782)	-9.9±6.9% [-39%; 13%]	11±8.8% [-7.3%; 64%]	-9.7±8.2% [-41%; 5.7%]	11±9.4% [-9.1%; 60%]	1.4±21% [-71%; 118%]	-6.5±9.6% [-46%; 36%]
C_1=10,000 (2863)	-9.6±7.1% [-38%; 13%]	10±9.2% [-7.1%; 63%]	-9.3±8.4% [-40%; 6%]	11±9.4% [-9.4%; 60%]	0.9±21% [-82%; 133%]	-6.2±9.8% [-42%; 44%]
C_1=100,000 (2873)	-9.6±7.2% [-38%; 13%]	10.8±9.2% [-7.1%; 63%]	-9.3±8.4% [-40%; 6%]	11±9.4% [-9.4%; 60%]	1.2±17% [-72%; 120%]	-6.3±10% [-41%; 43%]

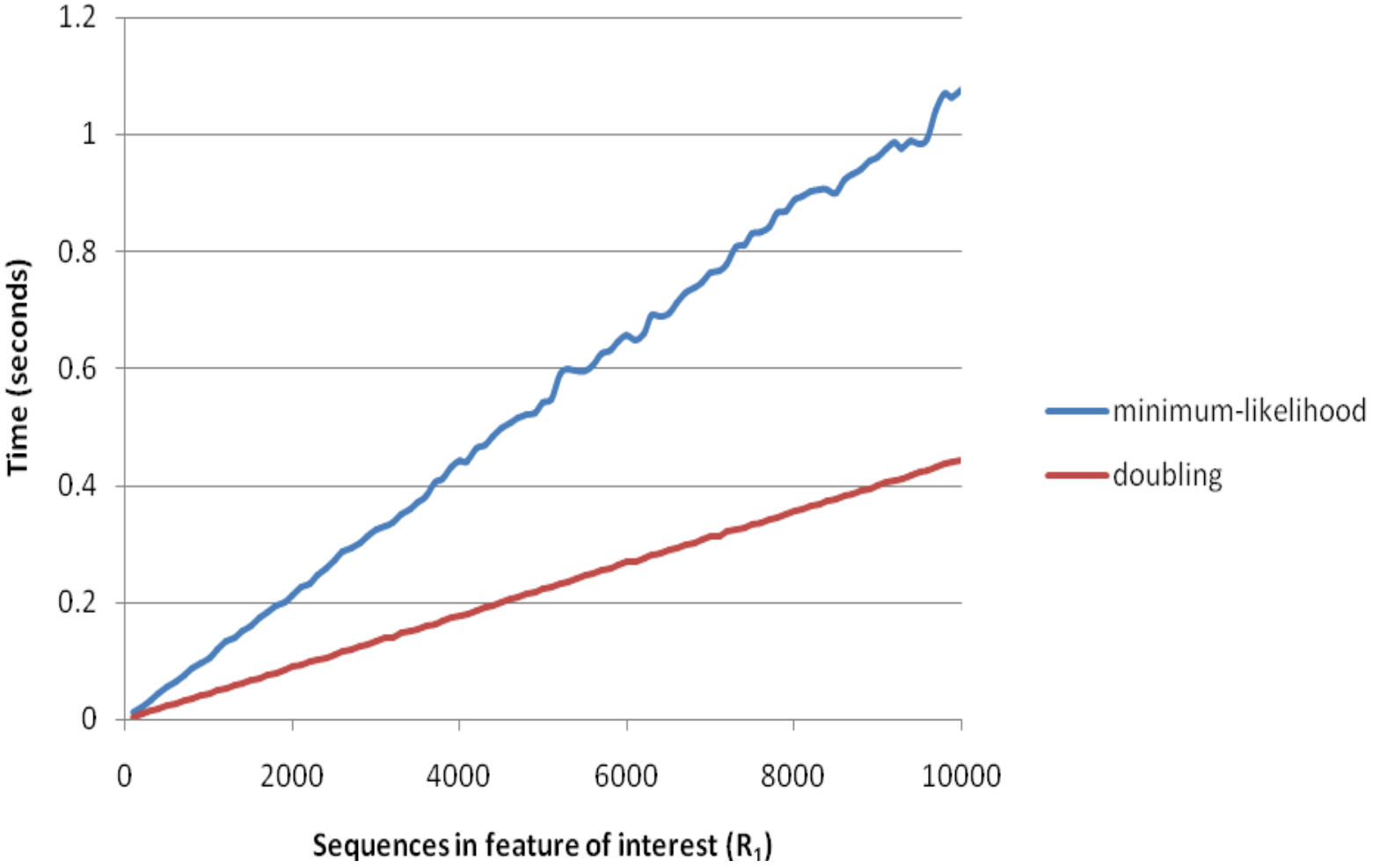
Similar comparison for smaller sequence set per feature

	Chi-square (Diff. b/w prop.)	Chi-square w/ Yates	G-test	G-test w/ Yates	Permutation	Bootstrap
C_1 = 100 (40)	-23±15% [-61%; -1.4%]	30±24% [-2.7%; 127.77%]	-32±28% [-86%; 2.2%]	23±20% [-22%; 65%]	8.0±25.25% [-34 %; 105%]	-15±18% [-49%; 30%]
C_1 = 200 (40)	-23±15% [-61%; -1.3%]	30±24% [-2.7%; 126%]	-31±28% [-860%; 3.0%]	23±20 % [-22%; 65%]	8.4±24.75% [-33%; 100%]	-14±18% [-49%; 25. %]
C_1 = 500 (40)	-23±15% [-60%; -1.2%]	30±24% [-2.7%; 125%]	-31±28% [-85%; 3.4%]	23±23% [-22%; 65%]	9.2±22.61% [-3.1%; 98%]	-12±16% [-48%; 23%]
C_1=1000 (40)	-23±15% [-60%; -1.2%]	30±24% [-2.5%; 125%]	-31±28% [-85%; 3.5%]	23±20% [-22%; 65%]	8.8±22.45% [-7.6%; 95%]	-12±16% [-48%; 29%]
C_1=10,000 (40)	-23±15% [-60%; -1.2%]	30±24% [-2.7%; 125%]	-31±28% [-85%; 3.6%]	23±20% [-22%; 65%]	8.7±23.50% [-32%; 100%]	-14±19% [-49%; 41%]
C_1=100,000 (40)	-23±15% [-60%; -1.2%]	30±24% [-2.7%; 125%]	-31±28% [-85%; 3.7%]	23±20% [-22%; 65%]	-10±28.18% [-49%; 115%]	-16±17% [-48%; 28%]

Comparison applied to real metagenome datasets

	Fisher's exact test	Chi-square (Diff. b/w prop.)	Chi-square w/ Yates	G-test	G-test w/ Yates
De Novo Purine Biosynthesis	3.81e-04	6.72e-04 (76.5%)	8.93e-04 (134.6%)	2.53e-04 (-33.6%)	3.60e-04 (-5.3%)
Asp-Glu-tRNA transamidation	4.00e-04	8.95e-04 (123.6%)	1.28e-03 (218.8%)	2.43e-04 (-39.2%)	3.94e-04 (-1.5%)
Chorismate Synthesis	1.30e-03	5.27e-04 (-59.4%)	8.39e-04 (-35.4%)	1.37e-03 (5.7%)	2.00e-03 (54.0%)
Thiamin biosynthesis	1.58e-03	4.46e-04 (-71.7%)	8.12e-04 (-48.5%)	1.46e-03 (-7.2%)	2.32e-03 (47.4%)
Ribosome SSU bacterial	2.22e-03	3.05e-03 (37.5%)	4.26e-03 (92.2%)	1.18e-03 (-46.8%)	1.83e-03 (-17.3%)
Respiratory dehydrogenases 1	3.35e-03	4.41e-03 (31.8%)	6.23e-03 (86.2%)	1.76e-03 (-47.5%)	2.78e-03 (-16.9%)
Proteolysis in bacteria	1.02e-02	1.21e-02 (18.4%)	1.60e-02 (55.8%)	7.28e-03 (-28.9%)	1.02e-02 (-0.8%)
Resistance to fluoroquinolones	1.12e-02	1.26e-02 (12.6%)	1.68e-02 (50.7%)	7.25e-03 (-35.1%)	1.04e-02 (-6.9%)
Fatty acid metabolism	2.99e-02	2.74e-02 (-8.3%)	3.48e-02 (16.1%)	1.98e-02 (-33.7%)	2.62e-02 (-12.6%)
Transcription factors bacterial	3.34e-02	3.46e-02 (3.8%)	4.38e-02 (31.2%)	2.58e-02 (-22.7%)	3.39e-02 (1.5%)

Number of sequences as a measure of the execution time for Fisher's test



Sampling with replacement strategy

- Real world metagenome datasets cannot work on the assumption that on resampling, we would always get the same number of sequences from the two samples

- To relax this, random samples are now generated by sampling with replacements proposed by Rodriguez *et al.* (2006)

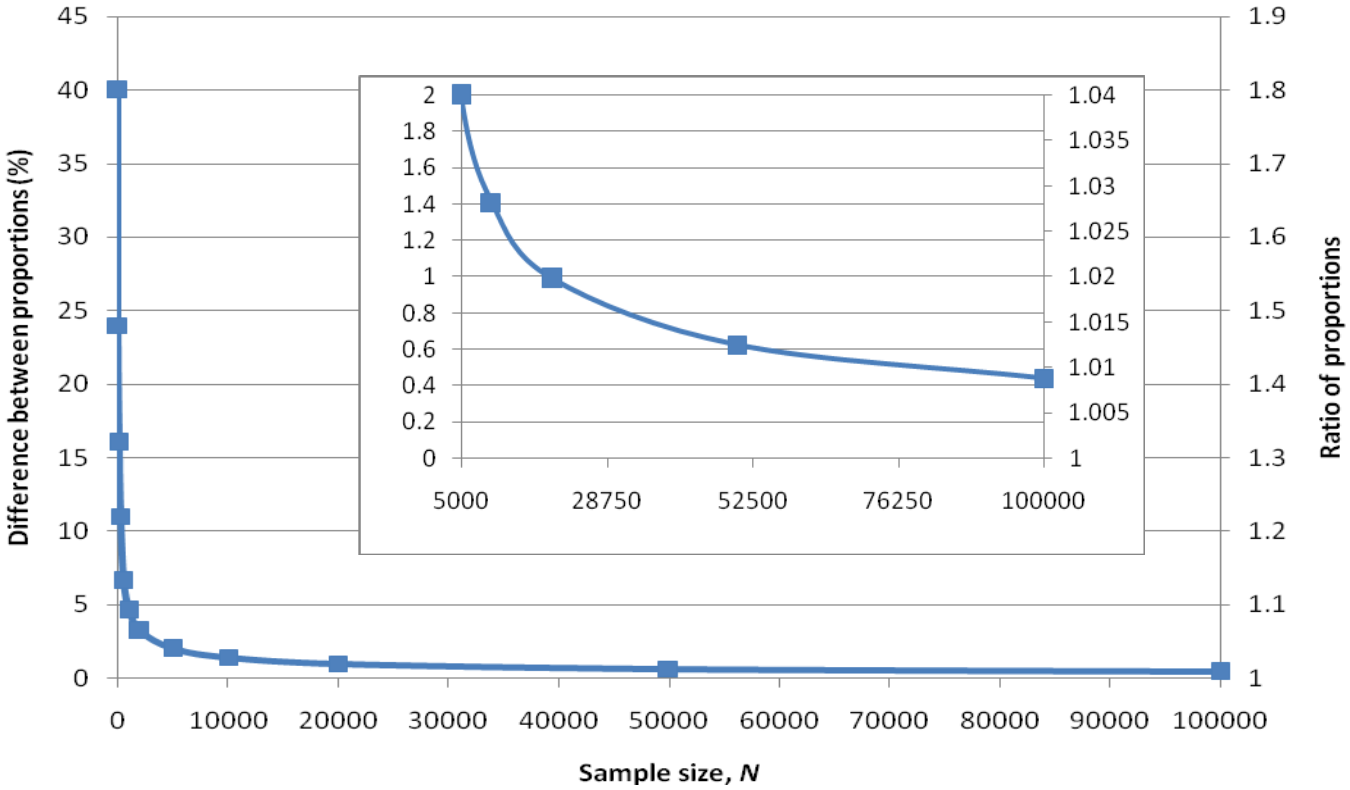
- Thus, in this case we can assume the number of sequences drawn from individual samples is fixed, but the sequences assigned to each feature can vary

- This gives rise to a binomial distribution and the z- test can be used to estimate the P-values.

Considering Effect Size to better understand statistical and biological relevance

- Along with using P-values to understand the significance of the observed difference, it is also necessary to quantitatively measure the magnitude of the difference
- This is done using the concept of "Effect sizes"
- Effect Size = True value – hypothesized value
- This is an important parameter because many a times, a small effect may be significant if the sample sizes are large and vice versa
- The effect size measures utilized in this tool are:
  - Difference of Proportion
  - Ratio of Proportion
  - Odds Ratio

Effect size required to achieve statistical significance for increasing sample sizes



Confidence Intervals

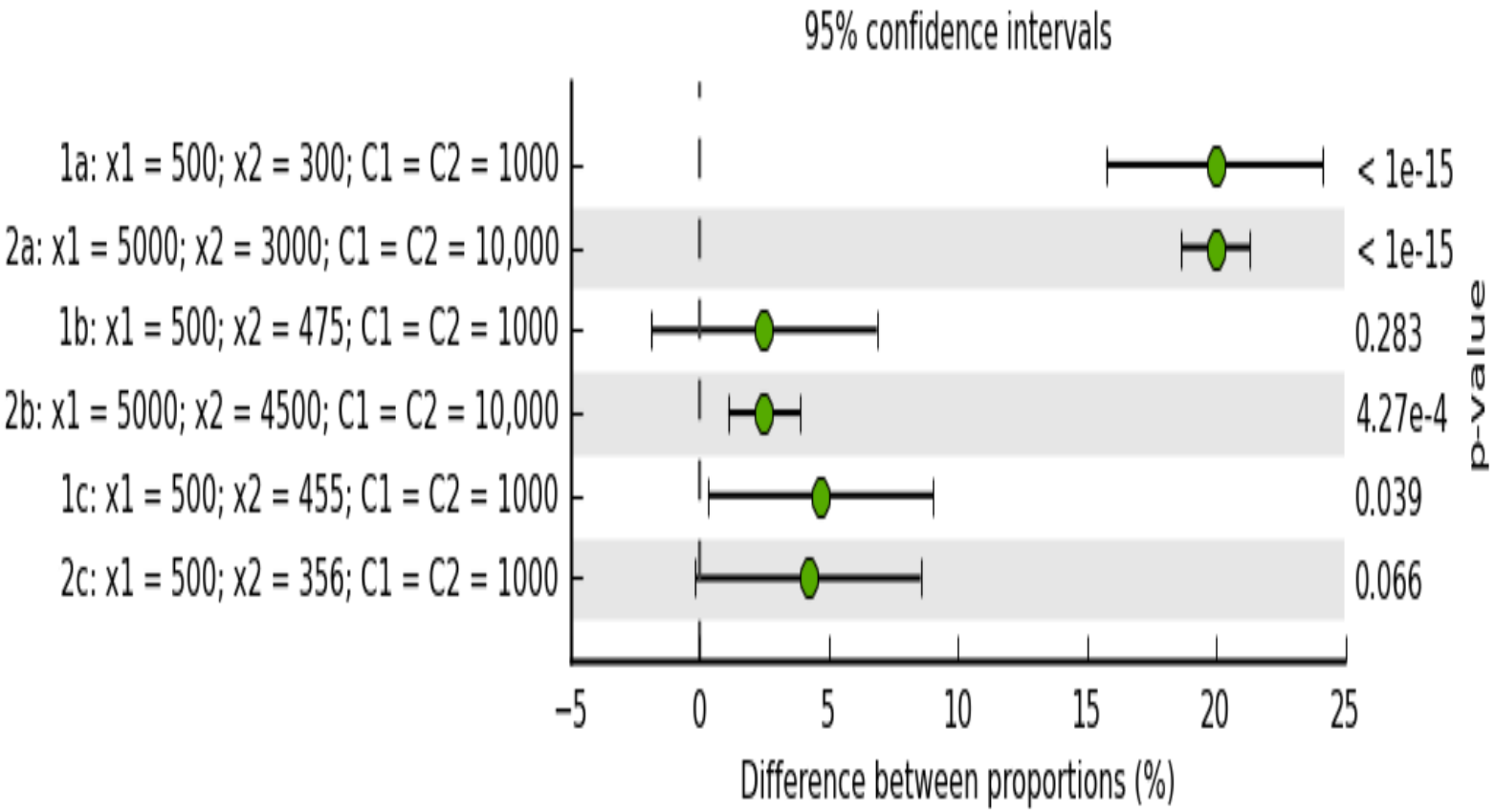
- This is another statistic used to assess the significance of the results obtained(both statistical and biological)

- Confidence intervals indicate a range of effect size values for different sampling events that have a probability of producing the true result

- The most important reason for choosing CI's as opposed to P-value assessments is that CI's vary considerably with sample sizes

- Thus, CI's help in identifying those cases where there is a marginal difference between the features in the 2 samples

Benefit of considering effect sizes and confidence intervals



Multiple test hypothesis correction

- A typical metagenomic profile would contain several hundreds of features and accordingly the p values need to be modified to obtain a consensus interpretation

- In order to avoid bias in the results obtained, certain correction methods are applied to the results

- Bonferroni correction is based on detecting family wise error rates

- Benjamini –Hochberg method is based on identifying the false positives for each feature

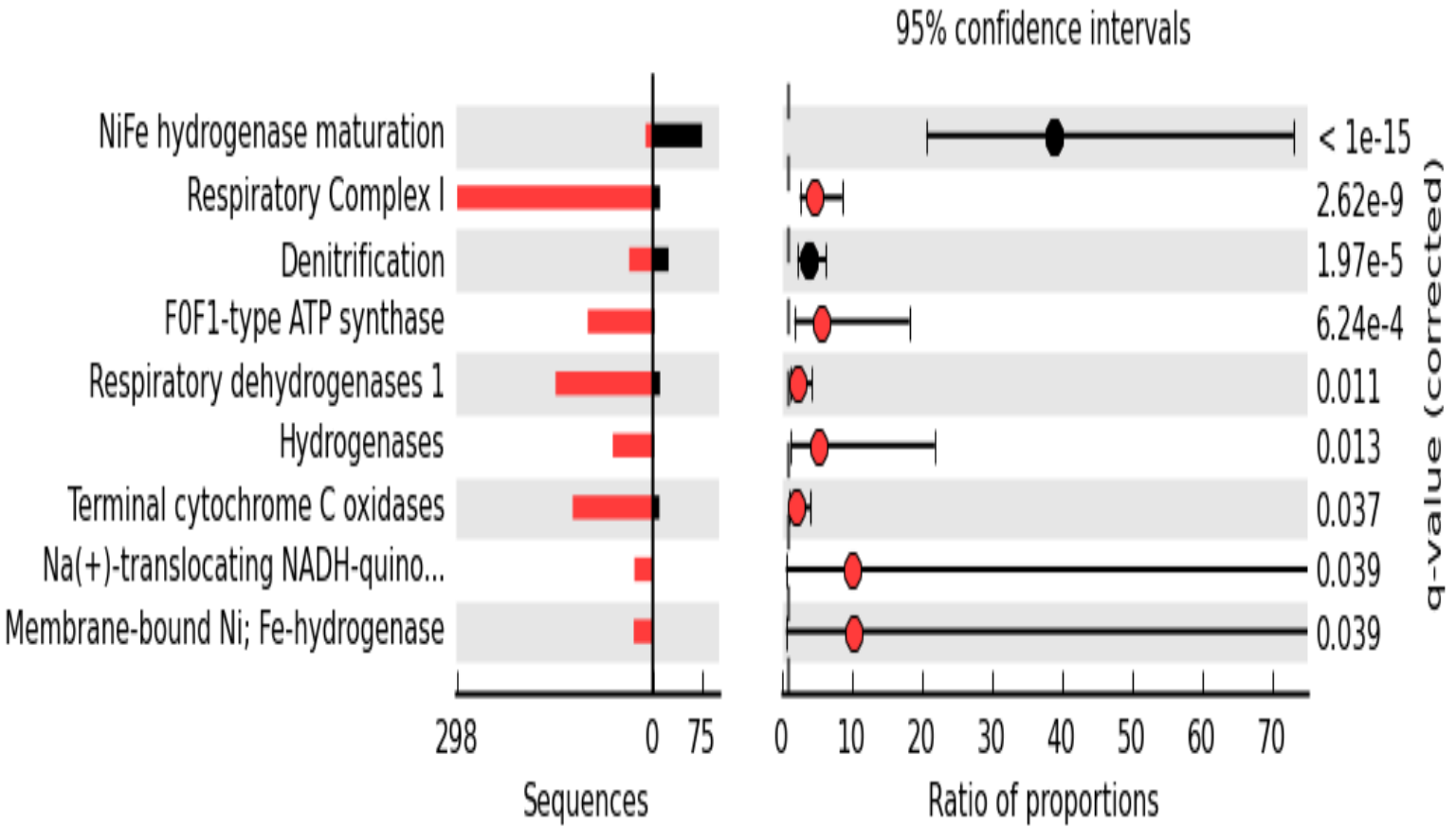
Comparison of available comparative metagenomic software

Program	GUI	p-values	ES	CI	MTC
STAMP	yes	yes	yes	yes	FWER,FDR
XIPE-TOTEC	no	no	no	no	no
ShotgunFunctionalizeR	no	yes	no	no	FWER,FDR
MEGAN	yes	yes	no	no	FWER
IMG/M	yes	yes	no	no	no

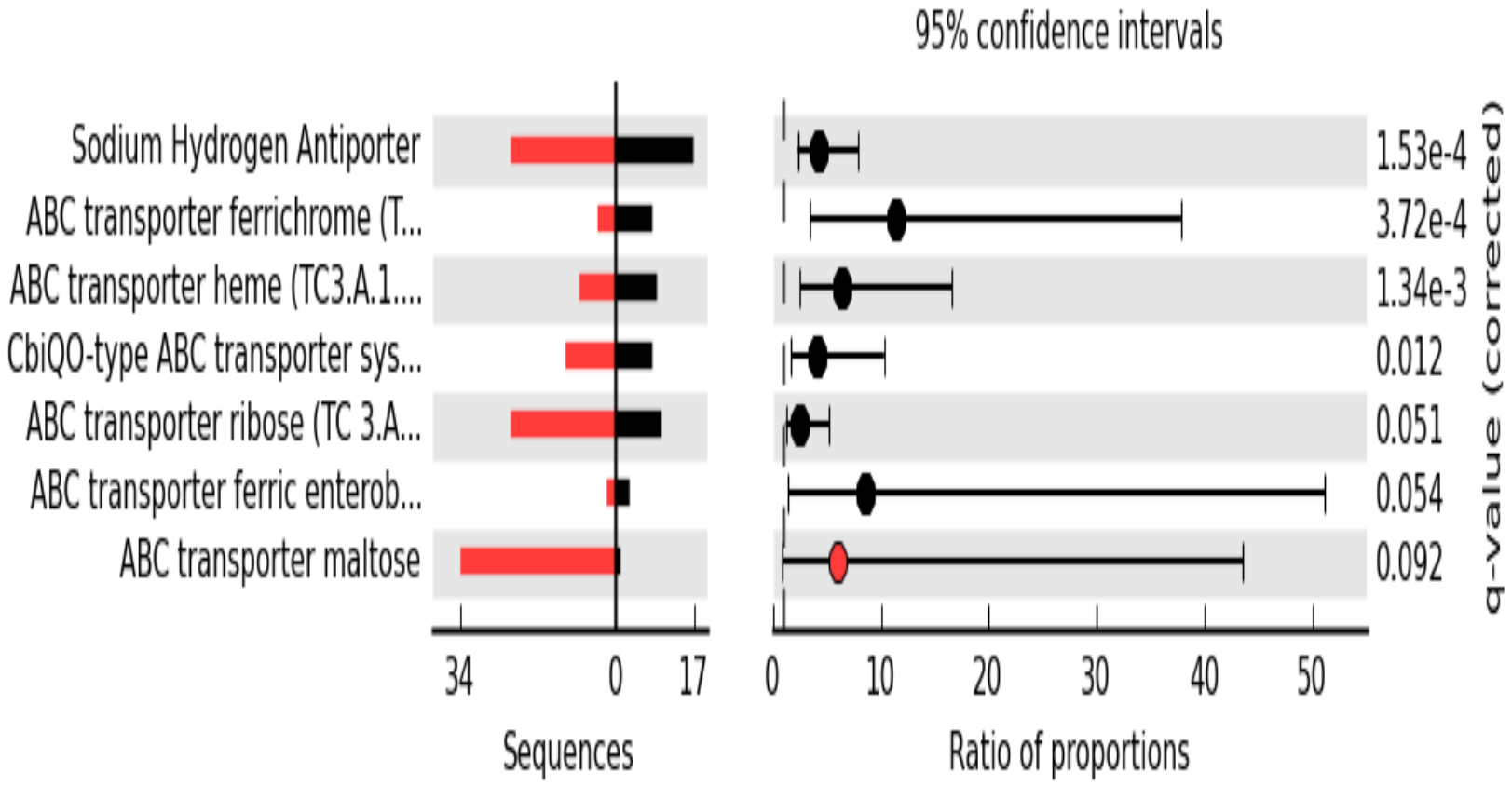
Implementation of STAMP on Soudan Iron Mine dataset

- The dataset from two distinct parts of the Soudan Iron Mine were obtained from the paper by Edwards *et al*
- In the original paper, the unassembled reads from the dataset were originally compared to the SEED database using the XIPE-TOTEC method
- The results are compared with those obtained for the same dataset using STAMP
- Fisher's Test in STAMP identifies 11% fewer statistically significant systems than XIPE-TOTEC with replicate sample sizes of M = 5000
- Edward's paper reported a total of 69 statistically different subsystems between the 2 communities as opposed to 60 using STAMP

Distribution of the respiratory systems in the two samples from Soudan Iron Mines



Distribution of membrane transport systems in the two samples from Soudan Iron mine



Study of the Accumulibacter strains from two distinct environments

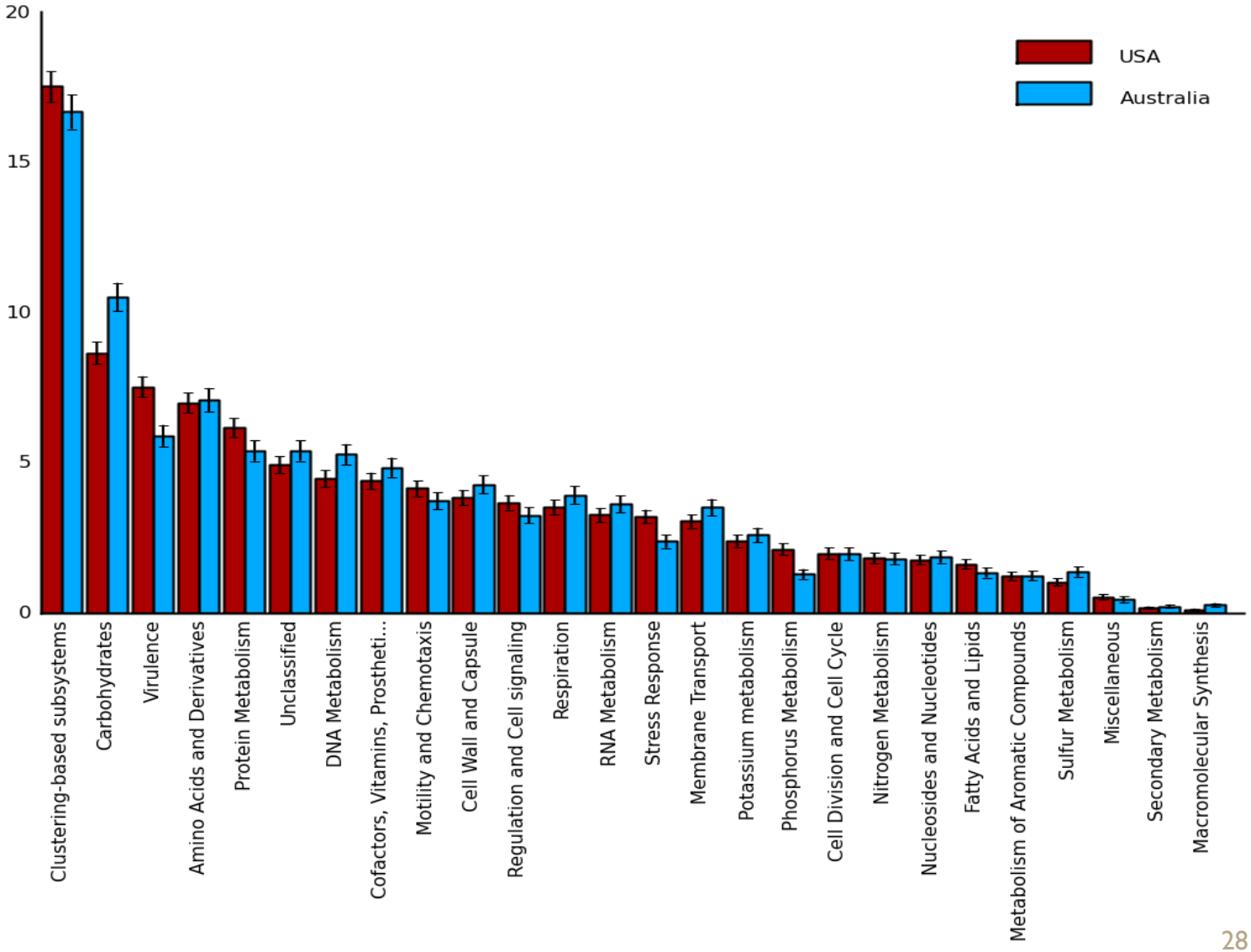
- Enhanced biological phosphorus removal (EBPR) is a process in which excess inorganic phosphate by employing micro organisms

- In order to understand the performance of these systems over time and location, function profiles of *Accumulibacter* strains from two distinct locations are studied

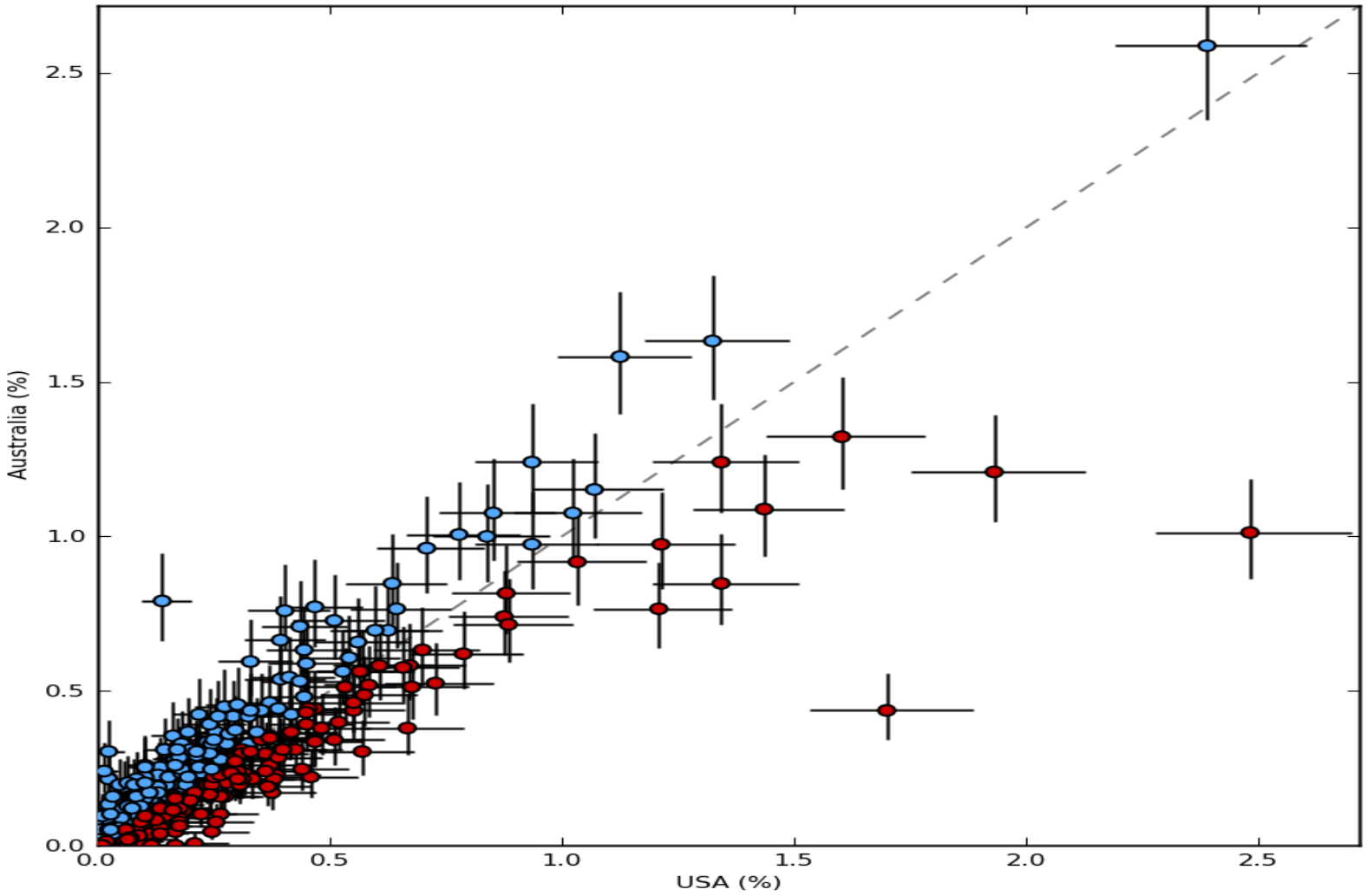
- The dataset was obtained from Garcia Martin *et al* 2006

- The *A. phosphatis* strains genes were assigned to 26 functional classes, containing a total of 491 SEED subsystems

Proportion of genes in A .phosphatis assigned to the 26 functional classes

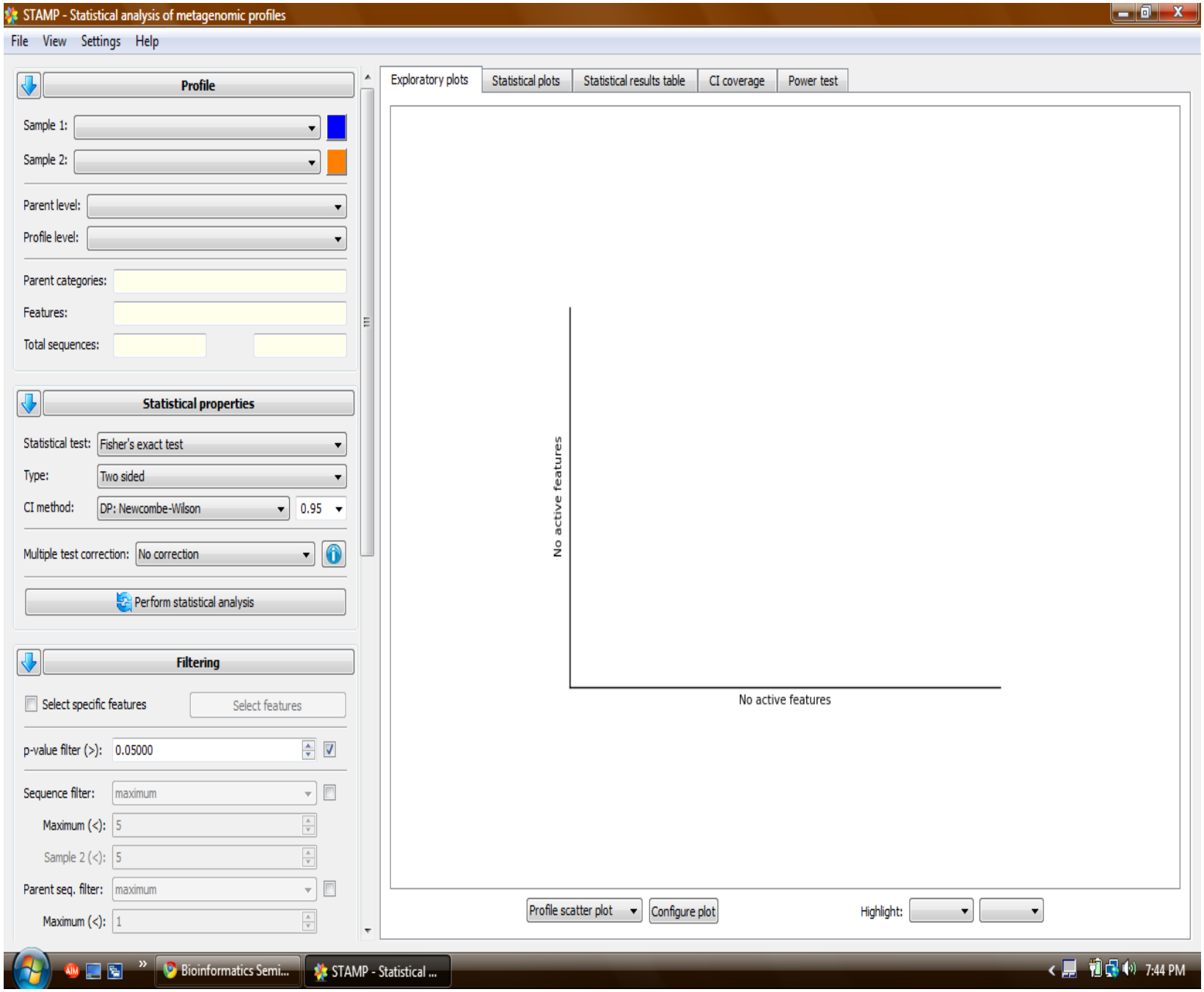


Scatter plot indicating the proportion of genes assigned to the 491 SEED subsystems



Comparison of the significant features obtained using the different techniques

	No multiple test correction	Storey's FDR	Bonferroni FWER
Fisher's exact test (minimum-likelihood)	116 (17.45)	77 (3.85)	22 (0.05)
Fisher's exact test (doubling)	120 (17.45)	93 (4.65)	21 (0.05)
Chi-square	108 (17.45)	71 (3.55)	18 (0.05)
Chi-square with Yates'	118 (17.45)	95 (4.75)	28 (0.05)
G-test	107 (17.45)	73 (3.65)	20 (0.05)
G-test with Yates'	108 (17.45)	71 (3.55)	20 (0.05)
Permutation	118 (17.45)	87 (4.35)	23 (0.05)
Bootstrap	116 (17.45)	80 (4.00)	28 (0.05)


STAMP - Statistical analysis of metagenomic profiles
File View Settings Help
Profile
Sample 1:
Sample 2:
Parent level:
Profile level:
Parent categories:
Features:
Total sequences:
Statistical properties
Statistical test: Fisher's exact test
Type: Two sided
CI method: DP: Newcombe-Wilson 0.95
Multiple test correction: No correction
Perform statistical analysis
Filtering
Select specific features
Select features
p-value filter (>): 0.05000
Sequence filter: maximum
Maximum (<): 5
Sample 2 (<): 5
Parent seq. filter: maximum
Maximum (<): 1
Exploratory plots
Statistical plots
Statistical results table
CI coverage
Power test
No active features
No active features
Profile scatter plot
Configure plot
Highlight:
Bioinformatics Semi...
STAMP - Statistical ...
7:44 PM

Discussion..

- Metagenomics relies heavily on the “guilt by association” paradigm

- As it is important to compute a reasonable result, so is important to identify sources of error in those results and how they can influence the interpretation of those results

- STAMP tries to do so by incorporating very naïve statistical techniques of CI and Effect size, which were completely unconsidered in all previous methods used

- These make STAMP a valuable tool to aid in interpreting the statistical results obtained more efficiently

Thank you