Bayesian multiple testing and q-values for genetics applications

Oleg Makhnin
New Mexico Tech
Dept. of Mathematics

September 28, 2012

Bioinformatics applications require simultaneous testing of multiple quantitative traits. For example, the studies of gene expression are common, for which two or more groups are compared (say, Normal and High Fat diets) and we wish to identify which genes are expressed differently in these groups.

A Bayesian framework is developed using which we can borrow information between different traits tested, and perform multiple testing more efficiently than using a simple Bonferroni approach. Additionally, we discuss the concept of q-values, or "False Discovery Rates" that can be obtained naturally using our approach.

Multiple testing

Bayesian approach and q-values

Application to microbial genome testing

Bioinformatics applications require simultaneous testing of multiple quantitative traits. For example, the studies of gene expression are common, for which two or more groups are compared (say, Normal and High Fat diets) and we wish to identify which genes are expressed differently in these groups.

A Bayesian framework is developed using which we can borrow information between different traits tested, and perform multiple testing more efficiently than using a simple Bonferroni approach. Additionally, we discuss the concept of q-values, or "False Discovery Rates" that can be obtained naturally using our approach.

Multiple testing

Bayesian approach and q-values

Application to microbial genome testing

Bioinformatics applications require simultaneous testing of multiple quantitative traits. For example, the studies of gene expression are common, for which two or more groups are compared (say, Normal and High Fat diets) and we wish to identify which genes are expressed differently in these groups.

A Bayesian framework is developed using which we can borrow information between different traits tested, and perform multiple testing more efficiently than using a simple Bonferroni approach. Additionally, we discuss the concept of q-values, or "False Discovery Rates" that can be obtained naturally using our approach.

Multiple testing

Bayesian approach and q-values

Application to microbial genome testing

Bioinformatics applications require simultaneous testing of multiple quantitative traits. For example, the studies of gene expression are common, for which two or more groups are compared (say, Normal and High Fat diets) and we wish to identify which genes are expressed differently in these groups.

A Bayesian framework is developed using which we can borrow information between different traits tested, and perform multiple testing more efficiently than using a simple Bonferroni approach. Additionally, we discuss the concept of q-values, or "False Discovery Rates" that can be obtained naturally using our approach.

Multiple testing

Bayesian approach and q-values

Application to microbial genome testing

Hypothesis testing

- Comparing Sample 1 and Sample 2. For example, Sample 1 = rats on High-fat diet (HFD, "treatment"), Sample 2 = rats on Regular diet ("control").

  Some genes might be involved in Metabolic Syndrome that is caused by high-fat diet. Which genes?

Hypothesis testing

- Comparing Sample 1 and Sample 2. For example, Sample 1 = rats on High-fat diet (HFD, "treatment"), Sample 2 = rats on Regular diet ("control").

  Some genes might be involved in Metabolic Syndrome that is caused by high-fat diet. Which genes?

- Hypothesis test tries to separate the random fluctuations in data from the real changes (presumably) caused by HFD.

  Testing if some condition (in our case, gene expression level) has changed from Sample 1 to Sample 2.

P-value

P-value is a measure of how extreme the change in each gene is.

$$\text{P-value}_i = P(\textit{observed difference}_i | \textit{no change})$$

Declare the change (in gene i) statistically significant when p-value$_i < \alpha$, for a given threshold α.

α= false positive rate (FPR). For N experiments with no change at all, in approximately $N\alpha$ we will declare that a change has occurred!

Multiple tests: corrections needed

- • Why a correction is needed: too many false positives?
  If, say, $N = 1000$ tests are run, each with a false positive rate (FPR) of 0.05, then we'll observe $\approx 1000(0.05) = 50$ false positives for the entire experiment

Multiple tests: corrections needed

- • Why a correction is needed: too many false positives?
  If, say, $N = 1000$ tests are run, each with a false positive rate (FPR) of 0.05, then we'll observe $\approx 1000(0.05) = 50$ false positives for the entire experiment
- • Bonferroni correction
  Thus, run each test (gene) with a smaller positive rate. If the final FPR of α is needed, run each test with a FPR of α/N.

Q-value

Q-value (Storey and Tibshirani, 2003) is the opposite of p-value:

$$\text{P-value} \approx \text{P}(\text{Test positive} \mid \text{no change}) \mapsto \text{False Positive Rate}$$

$$\text{Q-value} = \text{P}(\text{no change} \mid \text{Test positive}) \mapsto \text{False Discovery Rate (FDR)}$$

Like with p-value, declare genes with q-value $< \alpha$ "statistically significant".

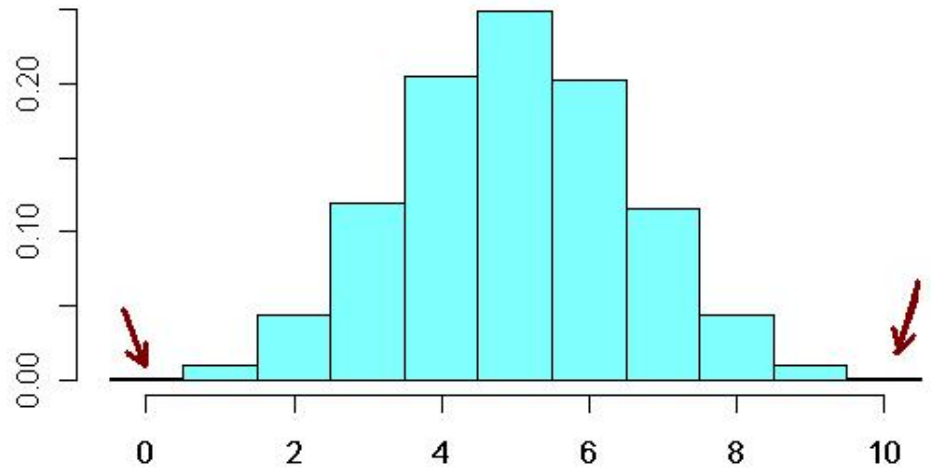
Q-value may be more desirable to practitioners: "Which portion of genes I have 'discovered' are bogus?"

P-value example

- Suppose we toss a coin 10 times, and observe 0 Heads. Are we inclined to believe that $P(Heads) = 1/2$?

P-value example

- Suppose we toss a coin 10 times, and observe 0 Heads. Are we inclined to believe that $P(Heads) = 1/2$?



In our case, p-value
$= P(0\ Heads) + P(10\ Heads) = 2 * 0.5^{10} \approx 0.002$
Since this is unlikely, we claim that the coin must be unfair.
In 1000 tosses, we'll have 2 false positives.

Q-value example

On the other hand, q-value will depend on what we know about the person tossing. If we trust the person, we estimate $P(\textit{fair coin}) = 0.9$

A little calculation:

$P(0 \textit{ or } 10 \textit{ Heads}) = P(0 \textit{ or } 10 \textit{ Heads}|\textit{fair}) * P(\textit{fair}) +$
$+P(0 \textit{ or } 10 \textit{ Heads}|\textit{unfair}) * P(\textit{unfair}) = 0.002 * 0.9 + 0.1 * 0.1$
≈ 0.012

$P(\textit{unfair}|0 \textit{ or } 10 \textit{ Heads}) \approx 0.002/0.012 = 1/6$

Hidden variable method

Idea 1: introduce *hidden variables* h_i which indicate whether the change occurred.

$$h_i = 1, \quad \text{if change in Gene } i$$
$$h_i = 0, \quad \text{if no change in Gene } i$$

Bayesian Markov Chain Monte Carlo computation through Gibbs sampler produces samples of all unknown variables and parameters. Our method allows for easy estimation of q-values:

$$q_i = \#\{h_{i,m} = 0, m = 1, ..., M\}/M$$

where $h_{i,m}$, is the mth sample from the hidden variable h_i.

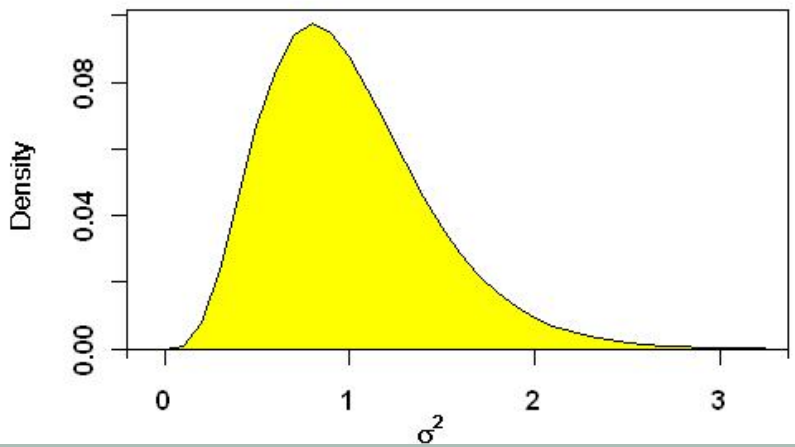
M is the Monte Carlo sample size (do not confuse with the sample size n for our data and with number of genes N). Higher M leads to more precise estimates of q_i.

Dealing with σ

Each gene has (potentially) different variability σ_i. For small sample sizes, it is hard to estimate σ_i well. Leads to highly inefficient Bonferroni method.

Earlier work (Dapo) assumed that σ_i's are the same for all genes $i, i = 1, ..., N$, but it's an oversimplification.

Idea 2: pool the estimates of σ_i from different genes i. This means that σ_i have some common distribution to be estimated.



Microbial study

PCE (perchloroethylene) used in dry cleaning

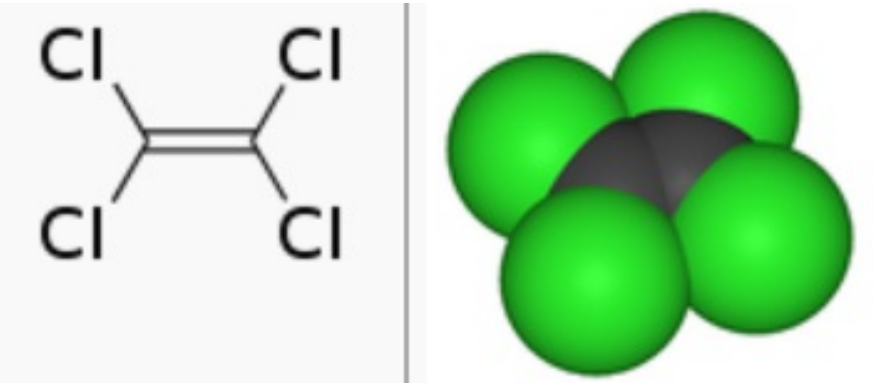


Image: Wikipedia

A well contaminated with PCE was injected with oil to stimulate microbial growth. The microbes break down PCE. We measure the species' abundance before and after remediation.

In this example, we are concerned not with individual genes, but rather whole microbial species. The data are counts of RNA segments (extracted from well water) that are attributed to different species.

Model

Let $L_{i,k}^{\texttt{before}}$ and $L_{i,k}^{\texttt{after}}$ be the expression levels (counts of RNA segments) for species i, sample k, from before and after the remediation.

The data here come from *matched pairs* design, so we'll take as model input $X_{i,k} := \log(L_{i,k}^{\texttt{after}} / L_{i,k}^{\texttt{before}})$
Logging is done to make data more Normally distributed.

Model

Let $L_{i,k}^{\texttt{before}}$ and $L_{i,k}^{\texttt{after}}$ be the expression levels (counts of RNA segments) for species i, sample k, from before and after the remediation.

The data here come from *matched pairs* design, so we'll take as model input $X_{i,k} := \log(L_{i,k}^{\texttt{after}} / L_{i,k}^{\texttt{before}})$
Logging is done to make data more Normally distributed.

- $X_{i,k}$ are Normal with mean μ_i and variance ${\sigma^2}_i$
  shortly $X_{i,k} \sim \mathcal{N}(\mu_i, {\sigma^2}_i)$

Model

Let $L_{i,k}^{\texttt{before}}$ and $L_{i,k}^{\texttt{after}}$ be the expression levels (counts of RNA segments) for species i, sample k, from before and after the remediation.

The data here come from *matched pairs* design, so we'll take as model input $X_{i,k} := \log(L_{i,k}^{\texttt{after}} / L_{i,k}^{\texttt{before}})$
Logging is done to make data more Normally distributed.

- $X_{i,k}$ are Normal with mean μ_i and variance ${\sigma^2}_i$
  shortly $X_{i,k} \sim \mathcal{N}(\mu_i, {\sigma^2}_i)$
- $\mu_i = 0$ if no change for species i (i.e. $h_i = 0$)
  and $\mu_i \sim \mathcal{N}(0, \tau^2)$ otherwise

Model

Let $L_{i,k}^{\texttt{before}}$ and $L_{i,k}^{\texttt{after}}$ be the expression levels (counts of RNA segments) for species i, sample k, from before and after the remediation.

The data here come from *matched pairs* design, so we'll take as model input $X_{i,k} := \log(L_{i,k}^{\texttt{after}} / L_{i,k}^{\texttt{before}})$
Logging is done to make data more Normally distributed.

- $X_{i,k}$ are Normal with mean μ_i and variance ${\sigma^2}_i$
  shortly $X_{i,k} \sim \mathcal{N}(\mu_i, {\sigma^2}_i)$
- $\mu_i = 0$ if no change for species i (i.e. $h_i = 0$)
  and $\mu_i \sim \mathcal{N}(0, \tau^2)$ otherwise
- ${\sigma^2}_i$ have Inverse Chi-square distribution, shortly
  ${\sigma^2}_i \sim \mathit{Inv}\chi^2(\nu_0, {\sigma^2}_0)$

Microbial study: results

1594 out of 1900 most abundant species had q-value below 0.05, **1415** had q-value below 0.01.

Other model parameters: (estimates $\pm$ standard errors)

Center of σ-distribution $s_0 = 0.109 \pm 0.004$

Degrees of freedom (tightness) of σ-distribution $\nu_0 = 2.67 \pm 0.22$

Probability of change $p_{DE} = 0.949 \pm 0.008$

Average change size (*log* scale) $\tau = 1.34 \pm 0.02$

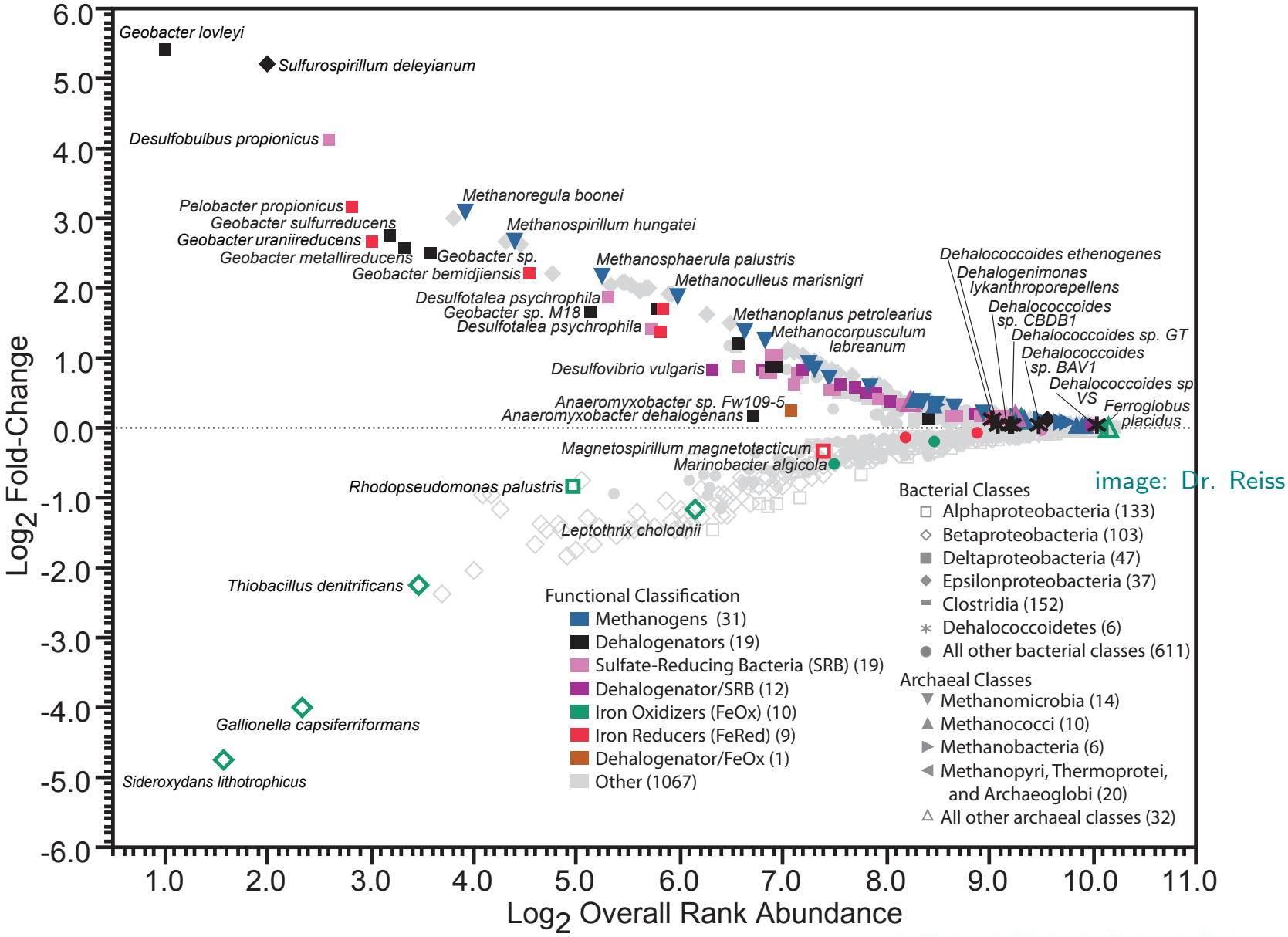
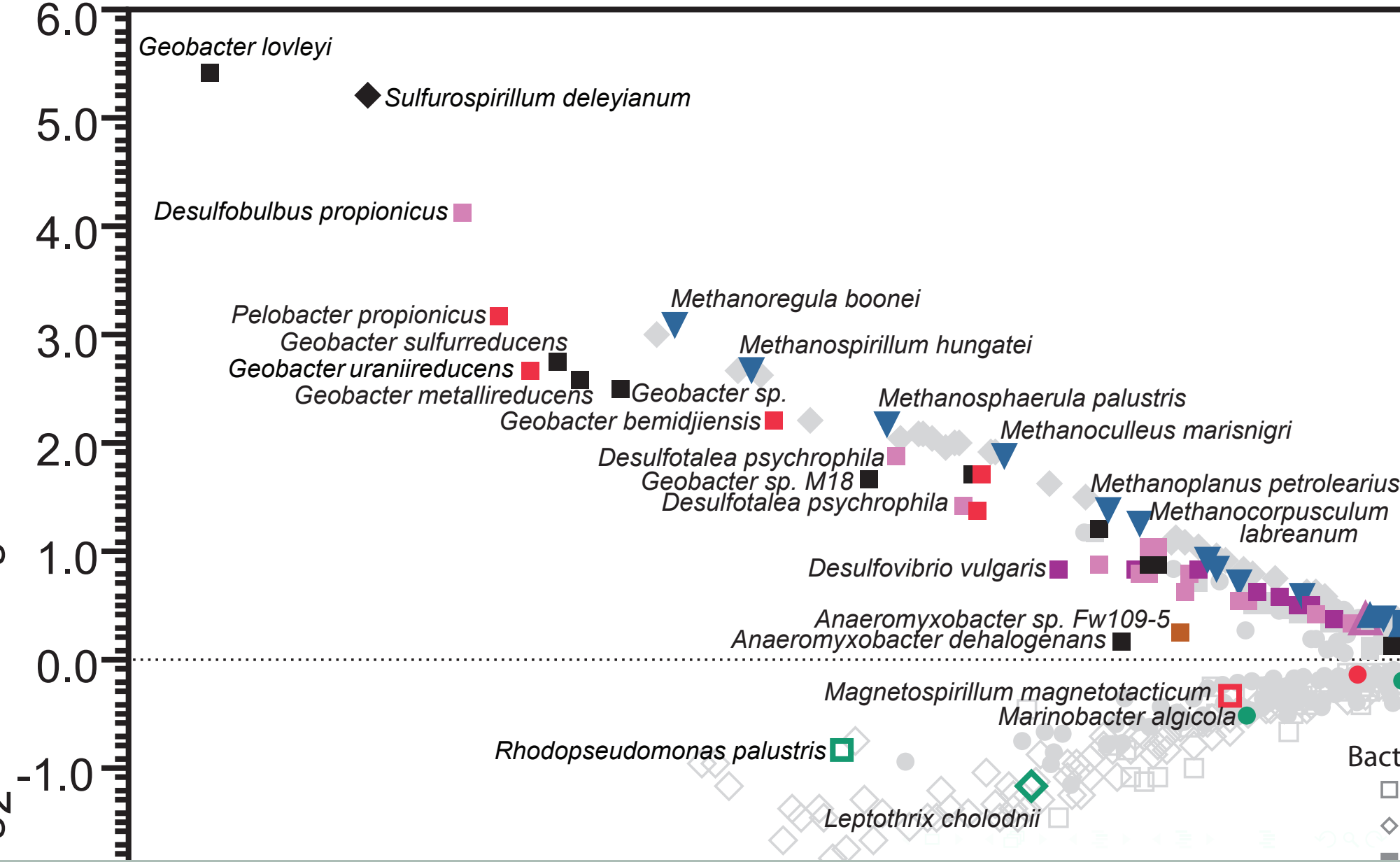
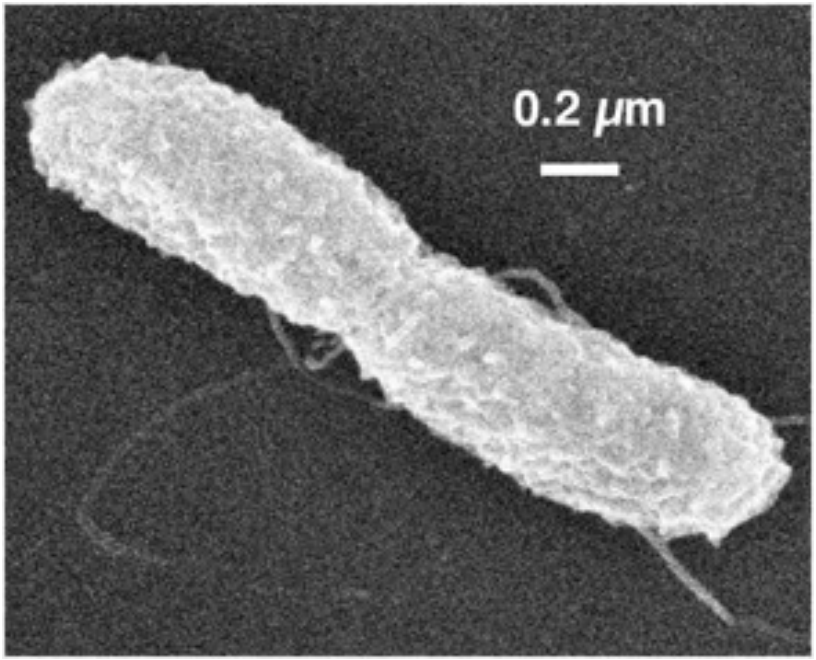


image: Dr. Reiss


6.0
5.0
4.0
3.0
2.0
1.0
0.0
-1.0
Geobacter lovleyi
Sulfurospirillum deleyianum
Desulfobulbus propionicus
Pelobacter propionicus
Geobacter sulfurreducens
Geobacter uraniireducens
Geobacter metallireducens
Methanoregula boonei
Methanospirillum hungatei
Geobacter sp.
Geobacter bemidjiensis
Methanosphaerula palustris
Methanoculleus marisnigri
Desulfotalea psychrophila
Geobacter sp. M18
Desulfotalea psychrophila
Methanoplanus petrolearius
Methanocorpusculum labreanum
Desulfovibrio vulgaris
Anaeromyxobacter sp. Fw109-5
Anaeromyxobacter dehalogenans
Magnetospirillum magnetotacticum
Marinobacter algicola
Rhodopseudomonas palustris
Leptothrix cholodnii
Bact



Geobacter lovleyi

QUESTIONS?

THANK YOU!

see www.nmt.edu/~olegm/talks/BMT