Iddo/Ivan/Iscott's Microarray Data Examination

Meta+Host

Data Harmonization

Project

Codelink Array Design & QC

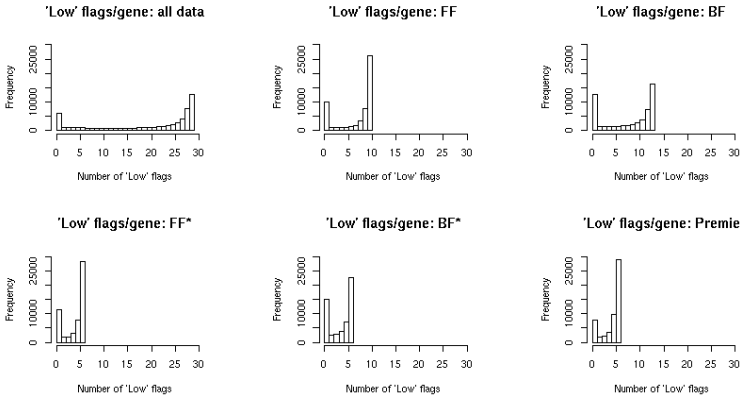
- 29 samples (10 FF, 7 BF, 6 BMS, 6 Premie)
- Probes 54359 (963 are fiducial & +/- controls)
- Removed Error flagged measurements (M/C/I/P):

Sample	FF	BF	Premie	Replicates
1	1359 (2.5%)	1303 (2.4%)	1046 (2.0%)	*1239 (2.3%)
2	1276 (2.4%)	1326 (2.5%)	949 (1.8%)	1287 (2.4%)
3 (FF) (BF)	*1342 (2.5%)	*1210 (2.3%)	1022 (1.9%)	*1241 (2.3%)
4 (FF)	*1219 (2.3%)	1239 (2.3%)	1097 (2.1%)	*1127 (2.1%)
5 (FF) (BF)	*1183 (2.2%)	*1140 (2.1%)	944 (1.9%)	*1377 (2.6%)
6 (FF)	*1198 (2.2%)	1155 (2.2%)	1030 (1.9%)	
7	1178 (2.2%)	1642 (3.1%)		
8 (FF) (BF)	*1146 (2.1%)	*1255 (2.3%)		
9	1173 (2.2%)	1240 (2.3%)		
10 (FF) (BF)	*1019 (1.9%)	*1355 (2.5%)		
11 (BF)		*1311 (2.5%)		
12		1244 (2.3%)		
13 (BF)		*1219 (2.3%)		

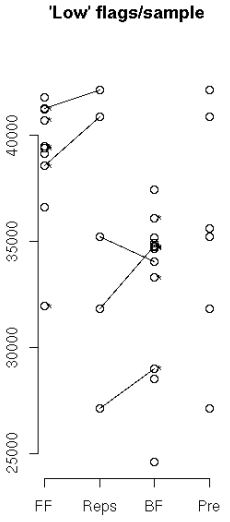
*Also have metagemome sequencing samples

Codelink Array “Low” flags

- “Low” flag: Signal mean < Bkgd_median + 1.5*Bkgd_stdev
- More “Low” flags in FF compared to BF



Further breakdown of “Low” flags

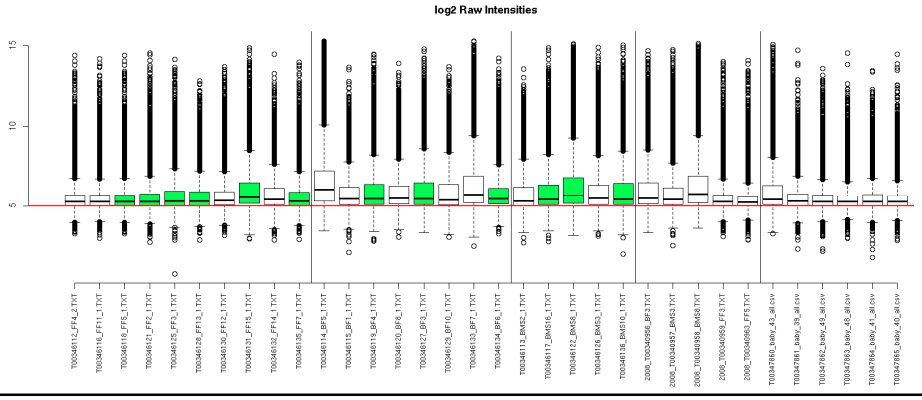


Sample	FF	BF	Premie
1	41255 (77.2%)	24601 (46%)	35638 (66.7%)
2	41784 (78.2%)	36081 (67.5%)	41019 (76.8%)
3 (FF) (BF)	*41267 (77.2%)	*34790 (65.1%)	42721 (80.0%)
4 (FF)	*40723 (76.2%)	35191 (65.9%)	43776 (81.9%)
5 (FF) (BF)	*38567 (72.2%)	*33293 (62.3%)	42205 (79.0%)
6 (FF)	*39486 (73.9%)	34945 (65.4%)	42903 (80.3%)
7	39158 (73.3%)	28535 (53.4%)	
8 (FF) (BF)	*31946 (59.8%)	*36120 (67.6%)	
9	36640 (68.6%)	37445 (70.1%)	
10 (FF) (BF)	*39399 (73.7%)	*34665 (64.9%)	
11 (BF)		*29010 (54.3%)	
12		34033 (63.7%)	
13 (BF)		*34747 (65.0%)	

*Also have metagemome sequencing samples

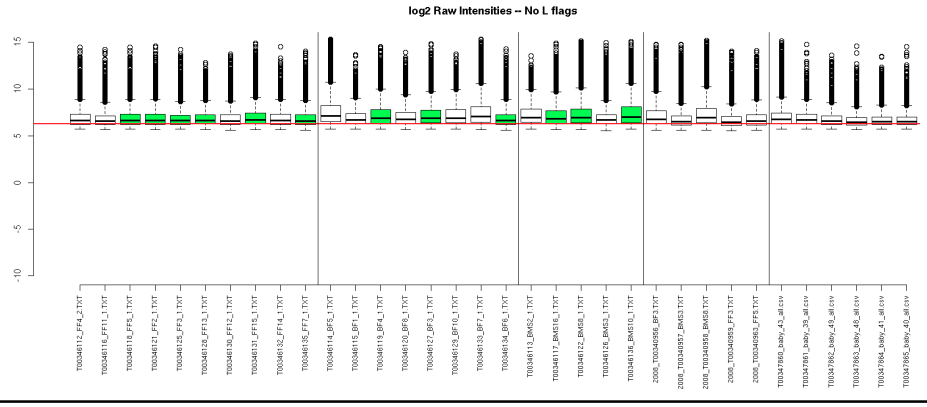
Expression values distributions

- Data is shifted so all values are positive, then log2'd
- Codelink Normed Val = Raw/median(Raw)… just line this up at 0
- More "Low" flags in FF because FF have lower distributions. duh.



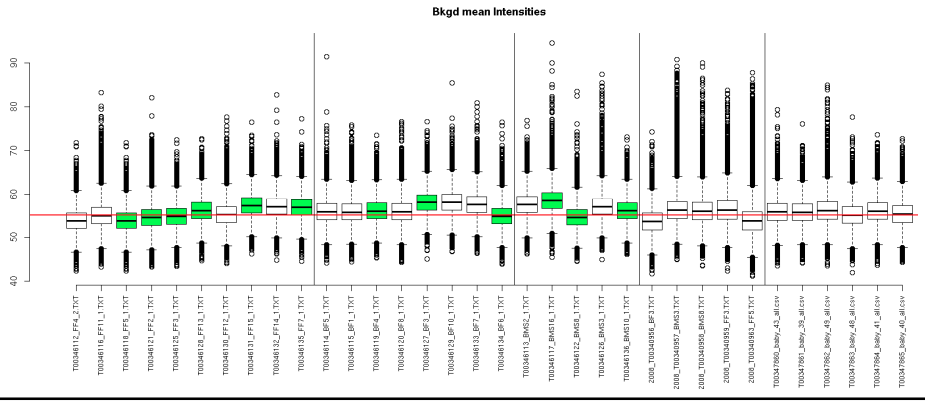
Raw Data without "Low" flags

- The general shift difference between FF and BF naturally remains even after "low" flagged genes are removed
- Losing data: ~75% FF, ~65% BF, and ~78% premie



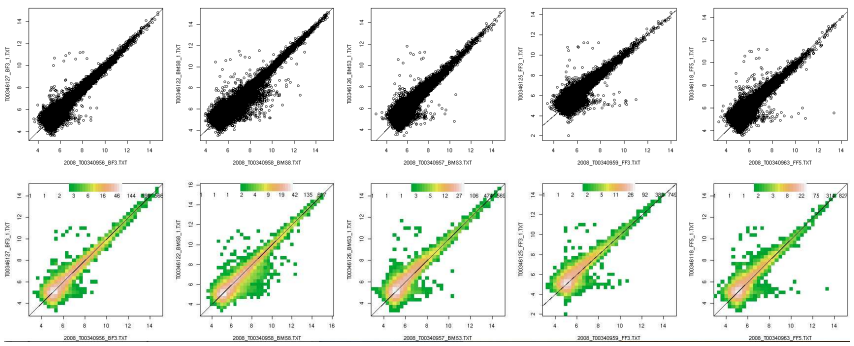
Background noise on slides

- Raw Intensity is actually Spot_mean - Bkgr_median
- Spot_mean - Bkgr_median is a global shift normalization
- Background noise doesn't have the same FF/BF shift pattern



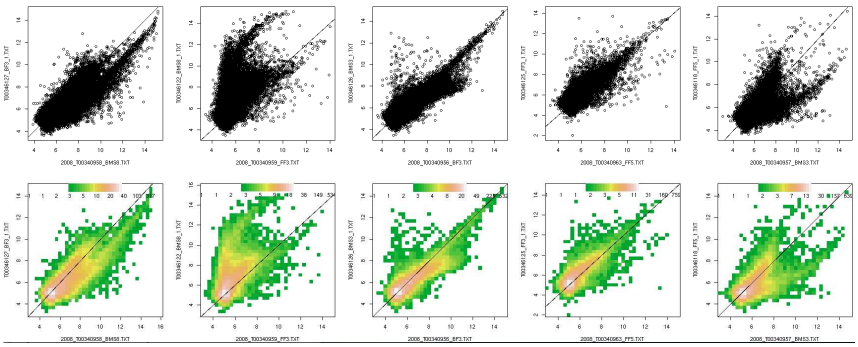
Replication

- Replication means we might say strange results are genuine
- Higher expressions are more replicable
- The shift difference between FF and BF may be real
- (4 of the 5 replicates have been metagenome sequenced)



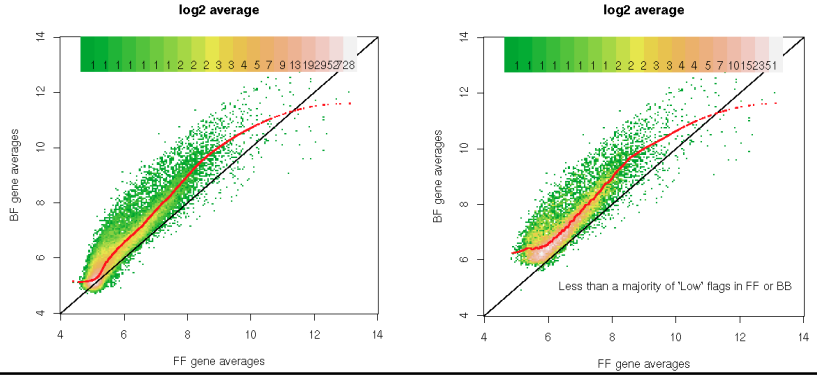
Sanity Check

- Replicates versus wrong samples
- Yeah! …when looking like crap is a chance to celebrate!

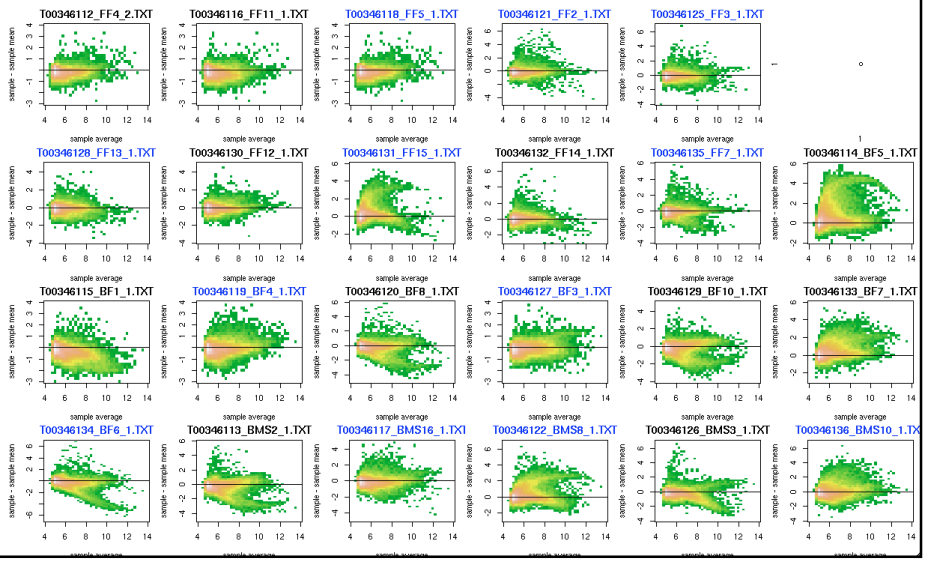


What does the MA plot look like

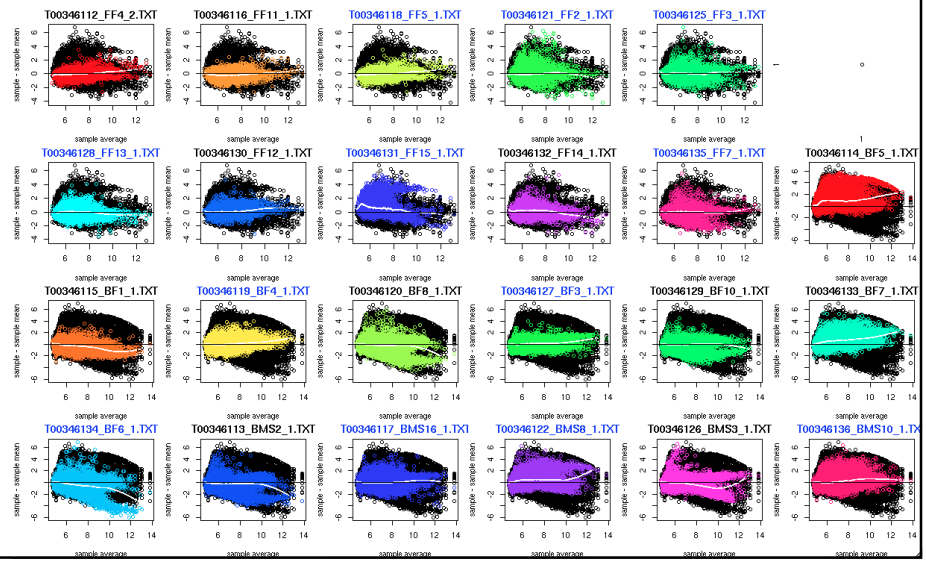
- BF babies genes tend to be higher expressing than FF, still
- Usually, a picture like this results in a loess normalization…
- Regardless, there's a group of exceptionally up expressed genes
- No obviously down expressed genes… ?



Gene expression in BF babies seems highly variable



Gene expression in BF babies seems highly variable



Conclusions (input welcome)

- Higher measurements are more replicable
- Genes are more highly expressed in BF than FF
- Subset of genes strongly driven by BF: Gene activity depends on BF environment
- Try several data sets
  - No normalization
  - Loess normalization
  - Chen's normalization

Other notes

- We're prototyping… data analysis is fallible
- Sample T00346118_FF5_1.TXT being the same as T00346112_FF4_2.TXT
- Not sure about doing outlier detection?
- Need to do one data processing QC step to make sure outlier BF patters are really there

Single gene prediction via phylum composition and future steps

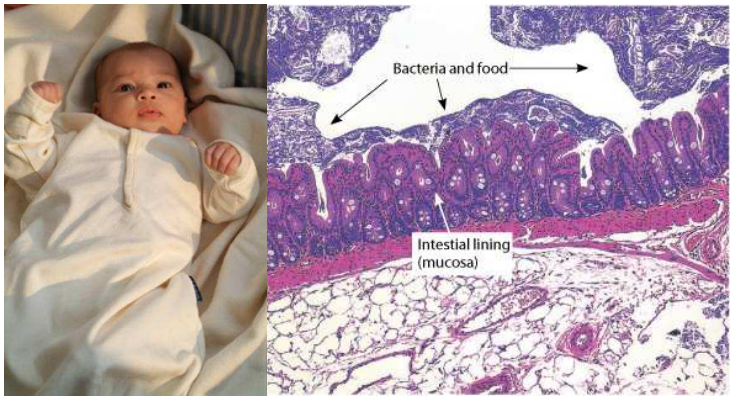
Meta+Host transcriptome

Data Harmonization

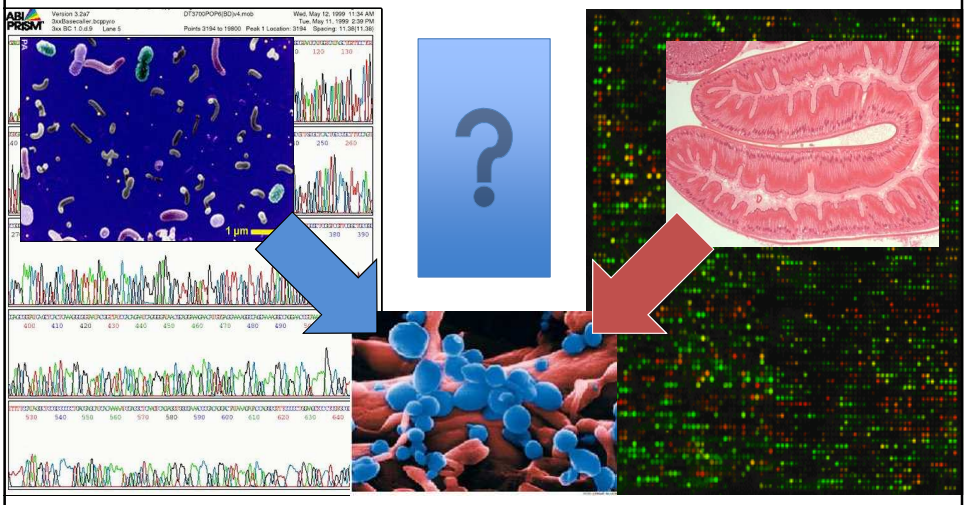
Project

Big Picture

- FF/BF Babies – Noninvasive Examination – Metagenome Sequencing – Host Microarray



Combine this info!!
Get something useful!?

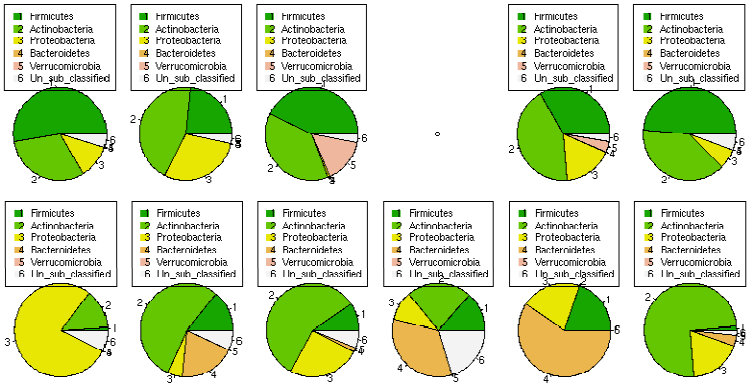


What we've got

- 12 fecal samples (6BF/6FF)
- Microarray of eukaryotic mRNA
- 454 Shotgun/16S sequencing of metagenome
- Big question: *Relate bacterial functional pathways to host functional pathways via array/sequence mRNA data*
- Little question: *Relate bacterial phylum distribution to single gene expression*

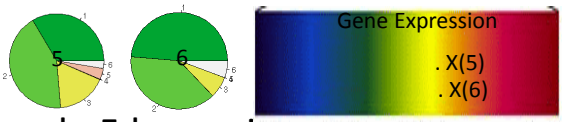
Too broad? High level Phylum data

- Phylum makeup classifies FF/BF (top/bottom)
- GE versus FF/BF signatures is just DE analysis



Too narrow? Fine level Phylum data

- Subtle differences in phylum sample makeup **ARE NOT** predictive of fine levels of gene expression differences between samples.
- E.g.: FF samples 5 and 6 have similar Phylum breakdowns:

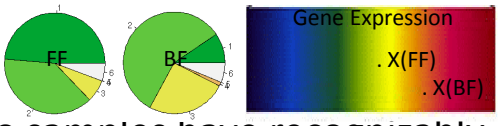


For gene X, sample 5 has a similar, slightly higher, expression measure then sample 6...

We WON'T attempt to explain this by Phylum

In perspective? Predict coarse scale GE via Phylum distribution

- Coarse differences in samples phylum makeup **ARE** predictive of coarse differences in gene expressions between samples.
- E.g.: Samples FF 6 and BF 3 have distinct Phylum makeups:

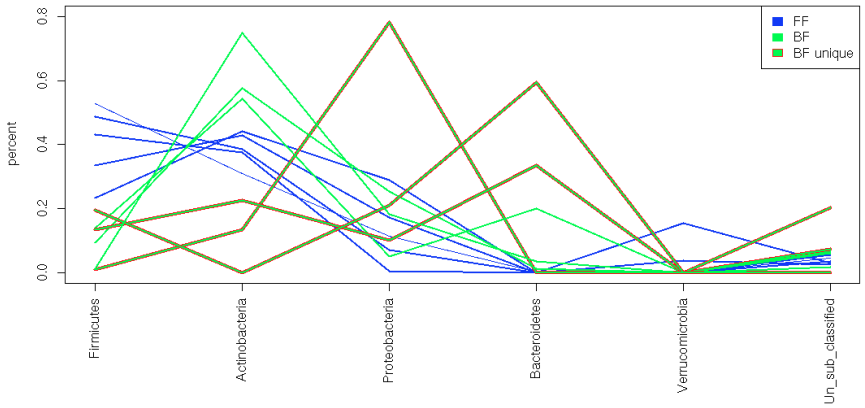


For gene X, the two samples have recognizably distinct expression levels...

We WILL attempt to explain this by Phylum

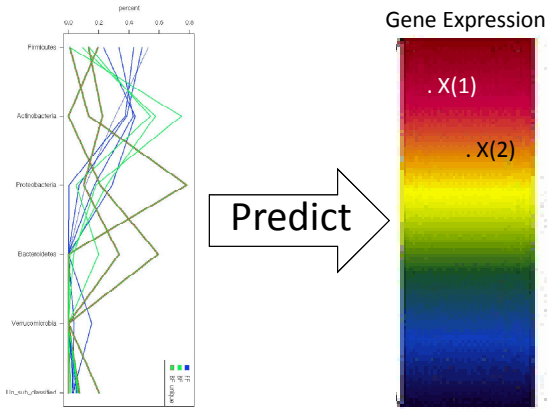
Phylum signatures

- 5 distinct signatures: 6FF, 3BF, BF', BF'', BF'''
- Signatures are "coarse" phylum measures



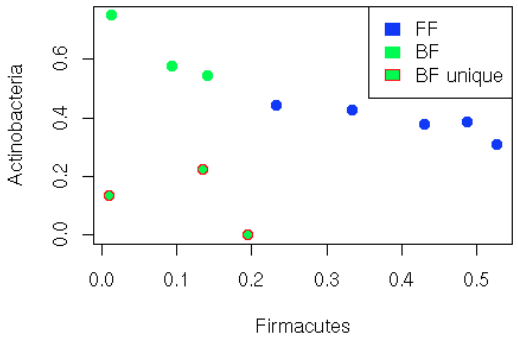
Baby question strategy

- Coarse phylum signature should be predictive of gross scale gene expression measurements



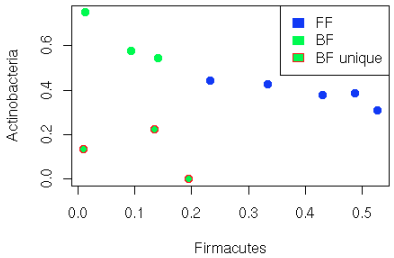
Most informative phylum

- Two strongest identifying characteristics:
  - firmicutes: distinguishes FF/BF
  - Actinobacteria: distinguishes BF/BF' samples



First try prediction methodology

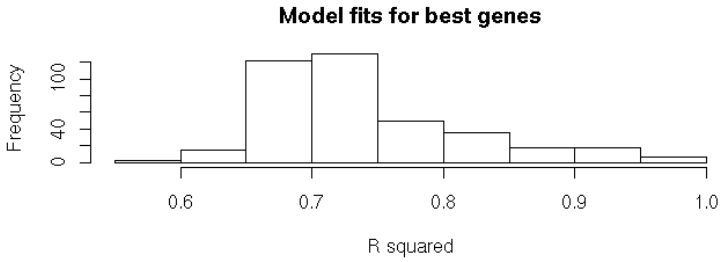
- Firmicutes (F) and Actinobacteria (A) strongly informative phylum signature



- Linear model prediction:
  single gene expression by F, A, and BF/FF

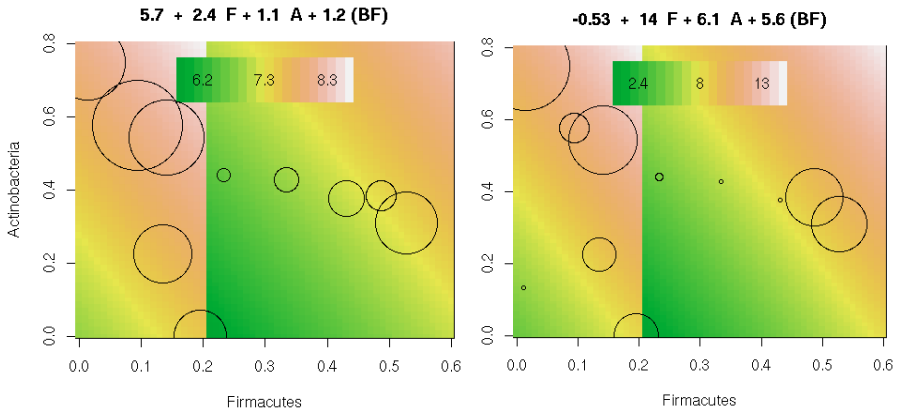
Preliminary Results

- Raw data only – no further processing
- Used 16853 genes that had less than a majority of "low" flags in either FF or BF
- 394 genes were usefully predicted by
  – firmicutes, Actinobacteria, BF versus FF



Linear Model: firmicutes, Actinobacteria, BF Indicator

- 1st order – plane with shift for BF babies
- Left R^2=.88; Right R^2=.62



Quantitatively relating (single) gene expression to Phylum distribution

- GE(X) = B_0 + F*B_f + A*B_a + (BF)*B_(bf)
- F: FF babies have higher %, so they'll be more changed (+/-) more by Beta_f
- A: FF babies have average %, BF signature babies have high %, and BF' signature babies have low %: Strength of effect (+/-) will be analogous
- BF babies have higher expression levels…

Notes

- Using %firmicutes and %Actinobacteria as continuous predictors utilizes more phylum distribution information than, e.g., firmicutes = {high/low}, Actinobacteria = {high/med./low}
- Phylum signatures are a simplex (n=6) – not all can be used as predictors – we used strongest
- 4 parameters and 12 data points leave 8 df
- Does not blatantly overfit the data (e.g., like using the 5 distinct signatures as categories)

Ideas for future Analyses

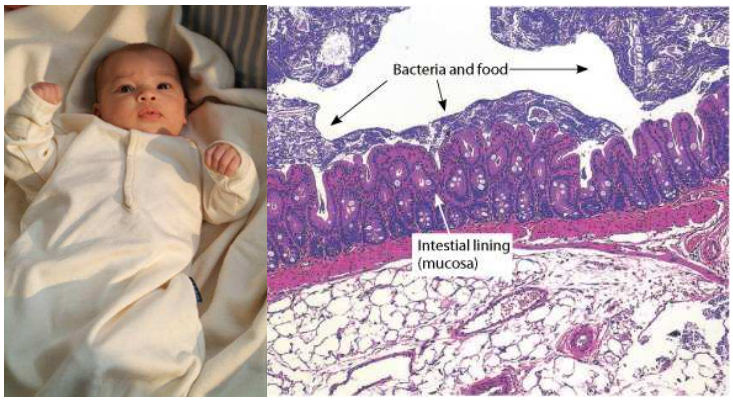
Meta+Host transcriptome

Data Harmonization

Project

Big Picture

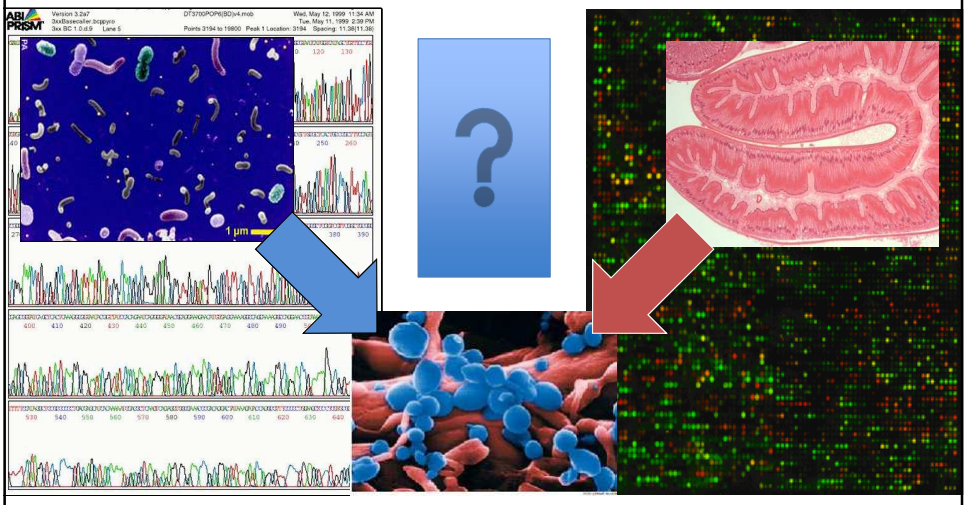
- FF/BF Babies – Noninvasive – Metagenome Sequencing – Host Microarray



Available data

- 12 fecal samples (6BF/6FF)
- Microarray of Host eukaryotic mRNA
  - H(a): Univariate (single) gene expression data
  - H(b): Multivariate (group) gene expression data
  - H(c) (Subset of) functional composition data [?]
- 454 Shotgun/16S sequencing of Metagenome
  - M(a): (Subset of) phylum composition
  - M(b): (Subset of) functional composition data

Combine this info!!
Get something useful!?

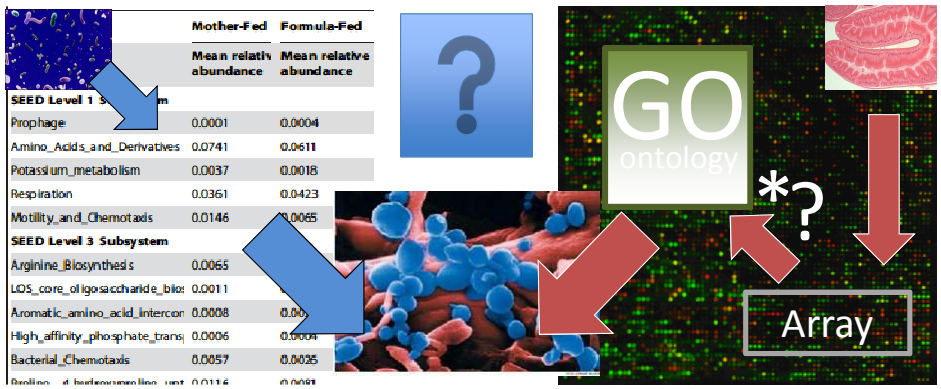


Possible analyses

- (1) Relate bacterial functional pathways to host functional pathways
- (2) Use bacterial information + host information to classify FF/BF
- (3) Relate bacterial functional pathways to subset of host gene expression
- (4) Relate bacterial phylum distribution to subset of host gene expression

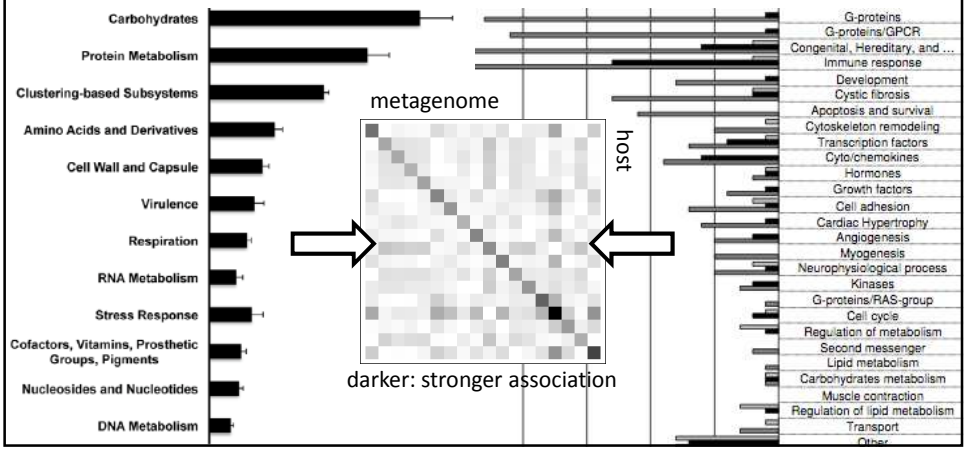
(1) Functional Analysis

- Can we quantitatively relate Bacterial and Host functionally activated pathways?



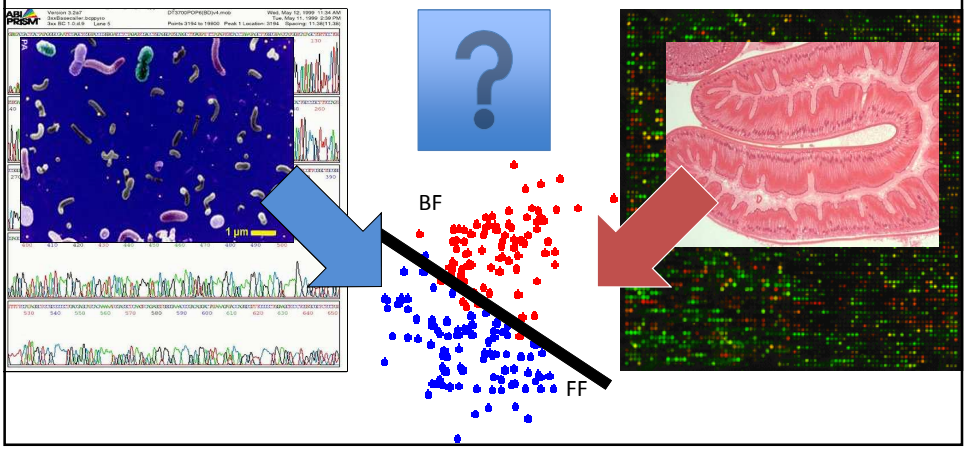
(1) Functional Analysis

- Correlate functional categories: covariance, Bayes, mutual information, or CoD methods



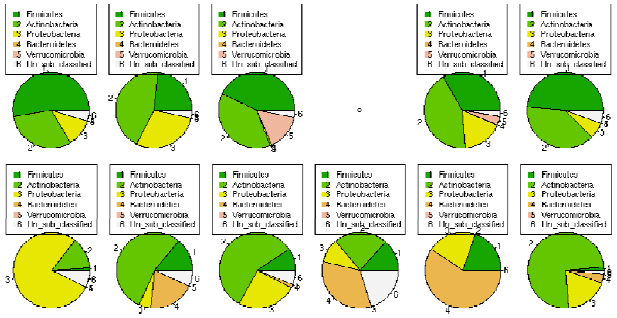
(2) BF/FF classification

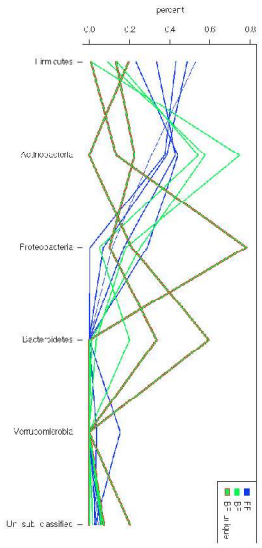
- Use *both* metagenome *and* host information in BF/FF classification, e.g., using LDA



(2) BF/FF classification

- E.g., phylum distinguishes FF/BF
- Right: FF(blue); BF(green/brown)
- Below: FF (top); BF (bottom)



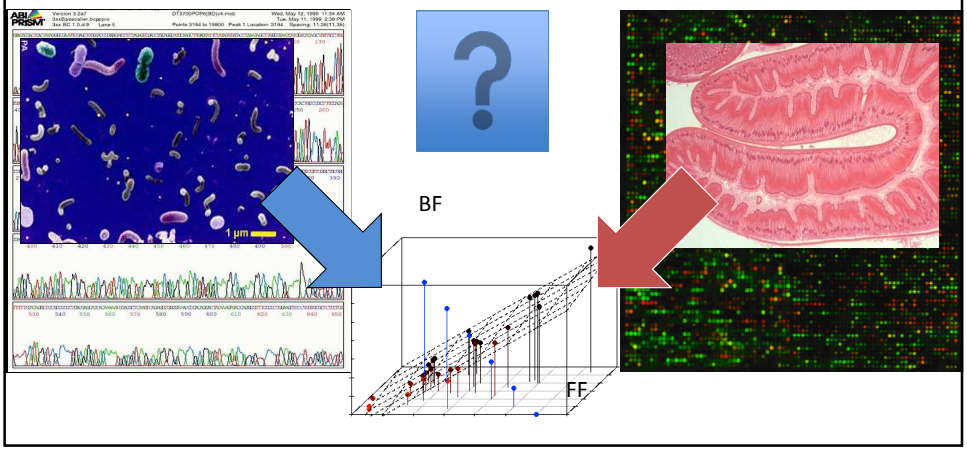


(3) Multivariate analysis

- The analytical concept here is analogous to that of (1), the functional analysis.
- we relate the high dimensional variable of, e.g., distribution of metagenome functional pathways to a set of gene expression values.
- Another method for relating multivariate variables to multivariate variables is canonical correlation analysis.

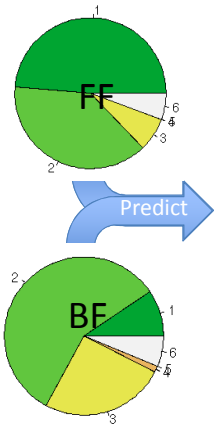
(4) Univariate prediction

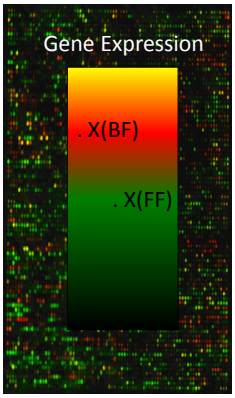
- Predict univariate gene expression using phylum/function metagenome characteristics

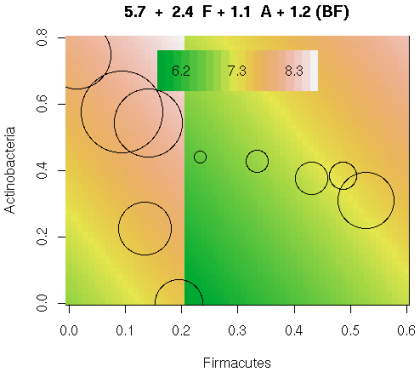


(4) Univariate prediction

- Phylum composition % predicts expression. Nonlinear MI and CoD methods available.

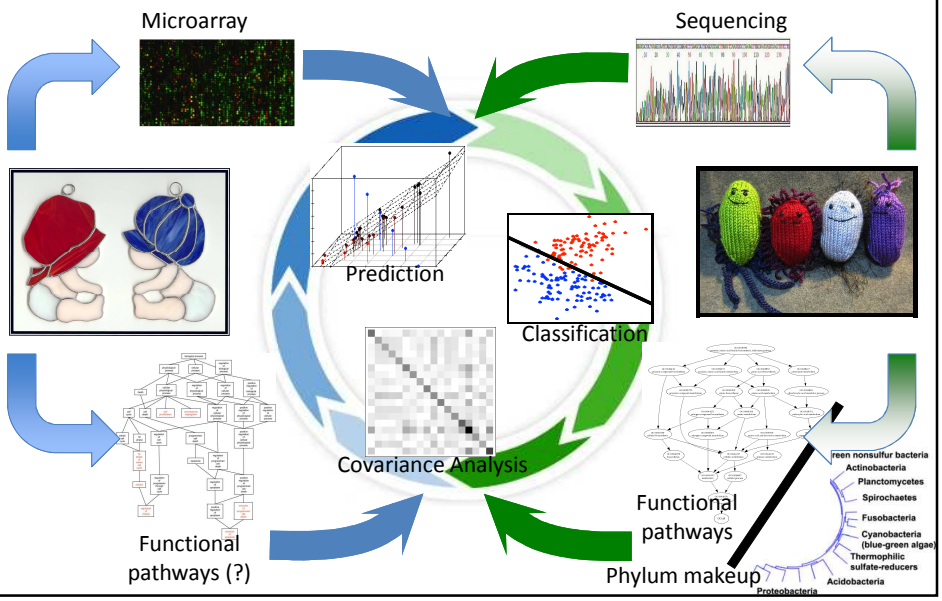




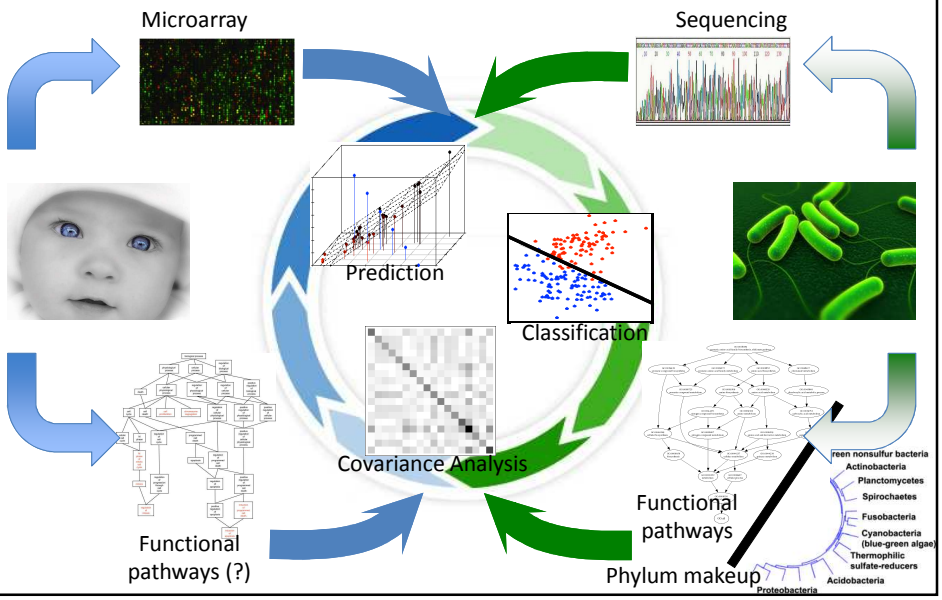


Gradient is surface of prediction
Circle size is gene expression

Summary



Data Harmonization

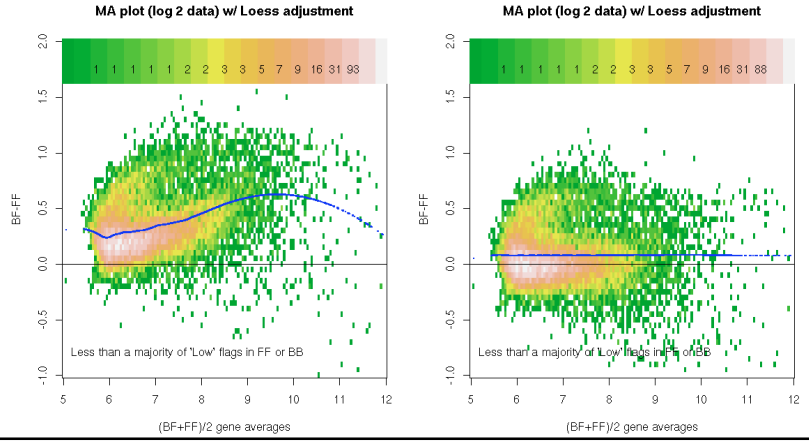


Roadmap

a. ~~Try various normalization for analysis (4)~~
b. Repeat analysis (4) using microbiome metabolic pathways in place of phylum %'s
c. Begin to examine analysis (2) en route
d. Examine potential roles for PCA analysis en route
e. Begin analysis (3) by examining pairwise correlation structure between microbiome pathways and host mRNA expression

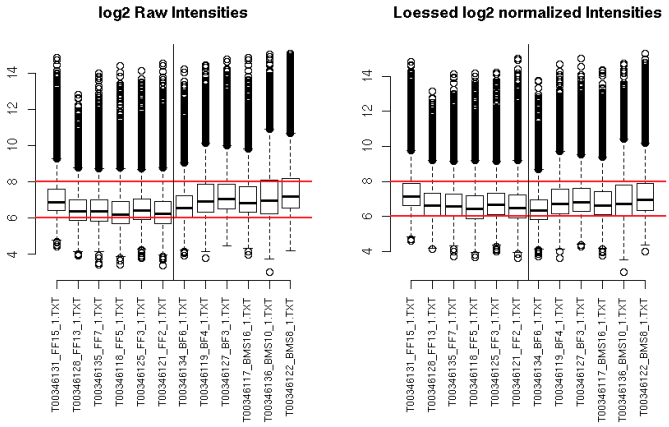
Normalized microarray

- These are the two data sets we're using. The loess normalization is artificially shifted up



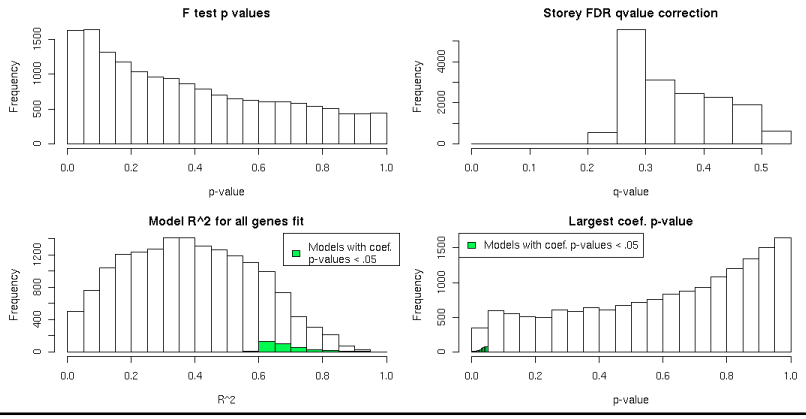
Normalized microarray

- FF is kept higher by personal preference: the loess line (blue) is arbitrary shifted up from 0



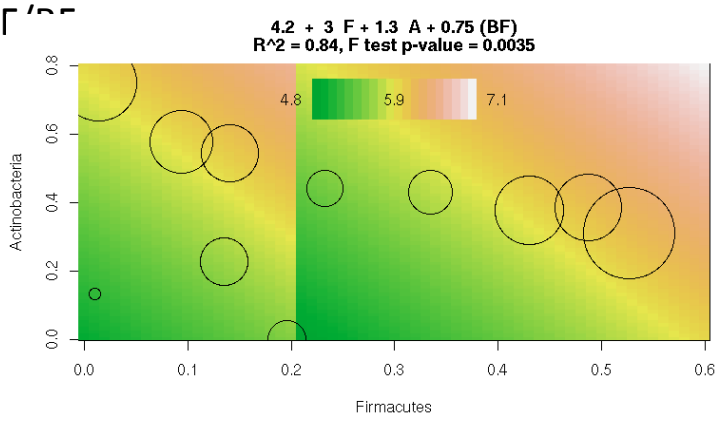
Model Fitting Results

- No significance w/ multiple testing correction. But there are still some suggestions of findings



Model interpretation

- Fits relationships between expression and %F./A. that are internally consistent across FT/BF



Questions/Comments

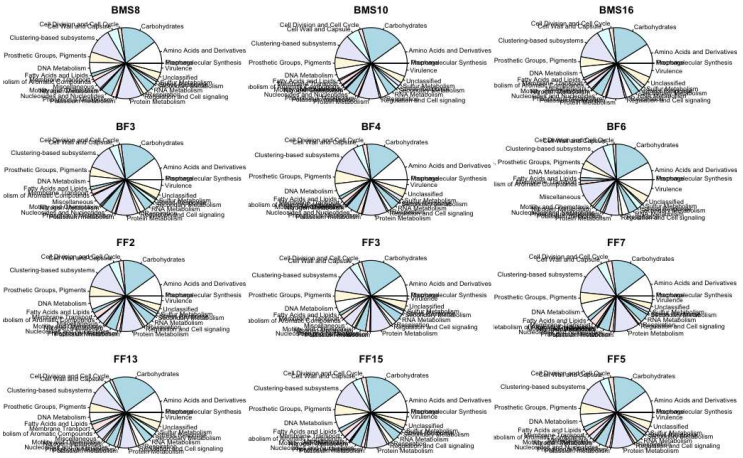
- From lab meeting:
- (a) Phylum is too coarse to be predictive
- (b) Does the interpretation of the model coefficients make sense? What does it mean?
- Thoughts:
- (a) Yes, phylum is probably to coarse
- (b) Alternative specifications address different hypothesis about the data relationships

Need to get to functional stuff

- maximum e-value 10^-5
- minimum alignment length of 100.
- Three min % identity thresholds: 80, 70 & 60.

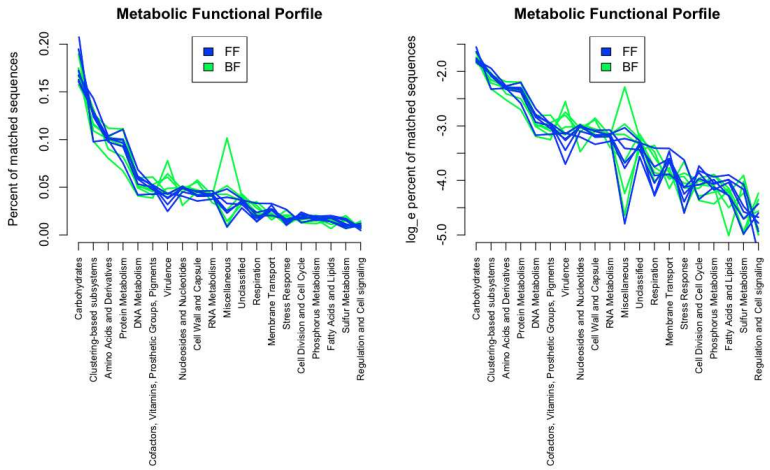
Metabolic Profile: SEED level 1

- Lots of small slivers… everybody gets some pie



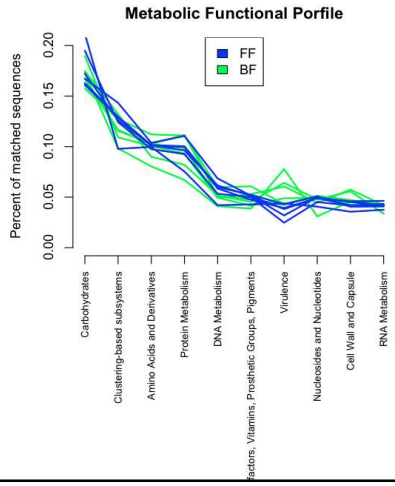
What are the metabolic profiles?

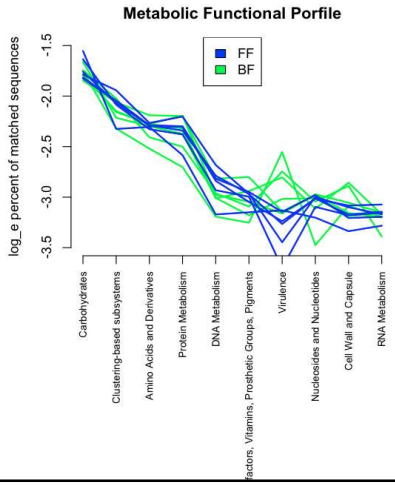
- Don't appear to be obvious FF/BF differences



Dimension Reduction #1

- For PCA, we have 12 samples, so we'll use 10



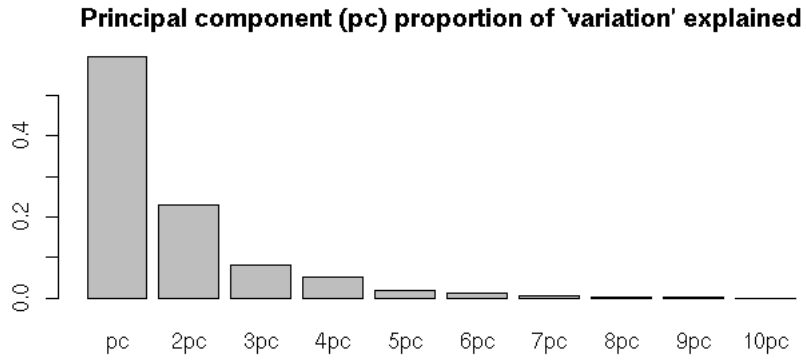


Singular Value Decomposition (SVD): Dimensionality Reduction

- X=UDV' : ([nXp] = [nXl][lXr][rXp])
- U' is a change of basis for X – puts axis structure along the spatial directions of greatest variation
- This provides dimensionality reduction since the new basis induced by U' may be of smaller dimension (l<p) than the original basis.
- Proportion of ``variation'' explained in by the ith dimension of the new basis is proportional to Sqrt(D(i))= ith EigenValue(XX')
- Derivation notes: U and V are orthogonal, i.e., UU'=U'U=I, so, U'X=DV' and XX'=WDD'W'

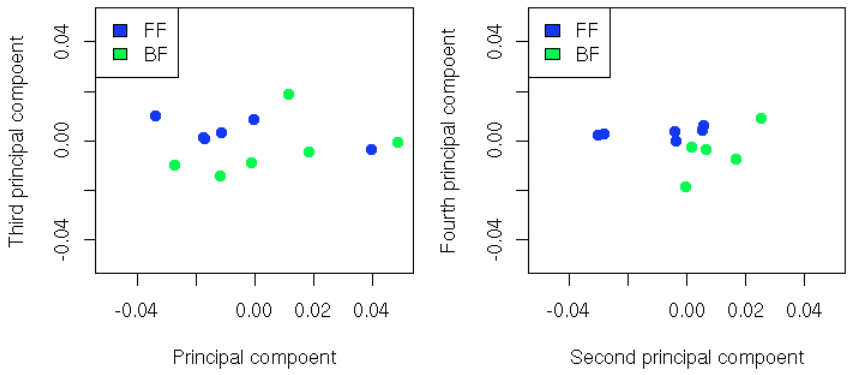
Principal Components

- Used p=10 most common (on average) metabolic functions and did a SVD



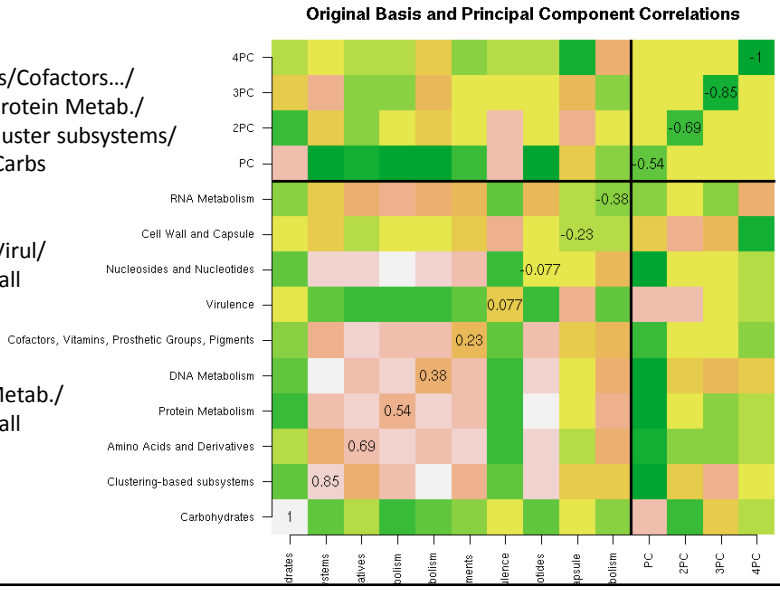
Classification with SVD PCs

- There's probably something there in the 2nd and 4th principal components



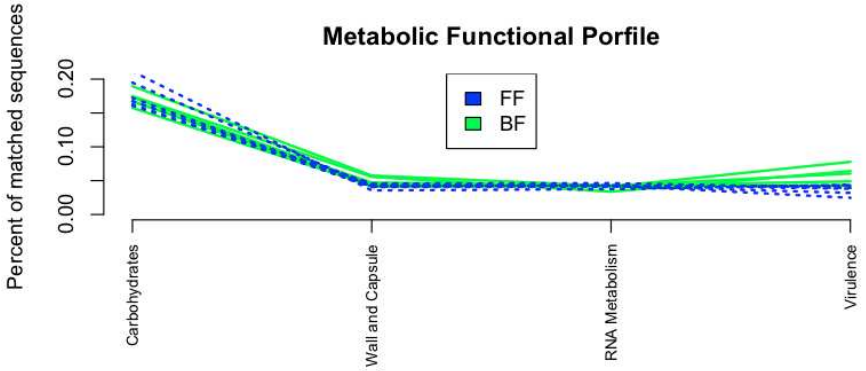
What are these PCs anyway?

- 1st PC: Nuclo's/Cofactors…/ DNA-Protein Metab./ AAs/Cluster subsystems/ Virul/Carbs
- 2nd PC: Carb/Virul/ Cell Wall
- 4th PC: RNA Metab./ Cell Wall



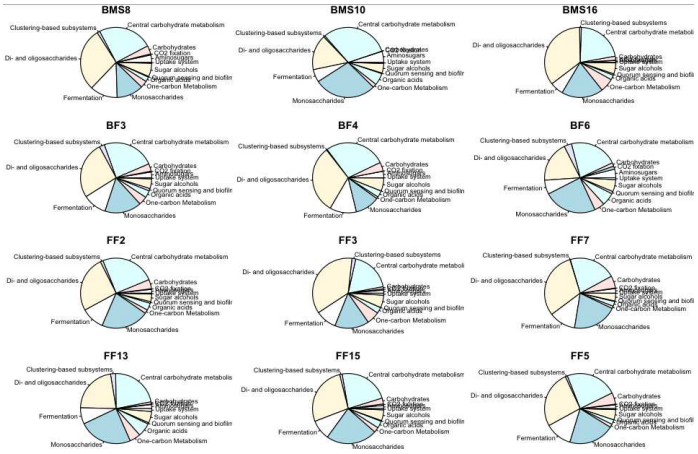
Dimension Reduction #2

- 2nd and 4th PCs were respectively driven by:
  - `Carbohydrates' and `Virulence', and
  - `Cell Wall and Capsule' and `RNA Metabolism'



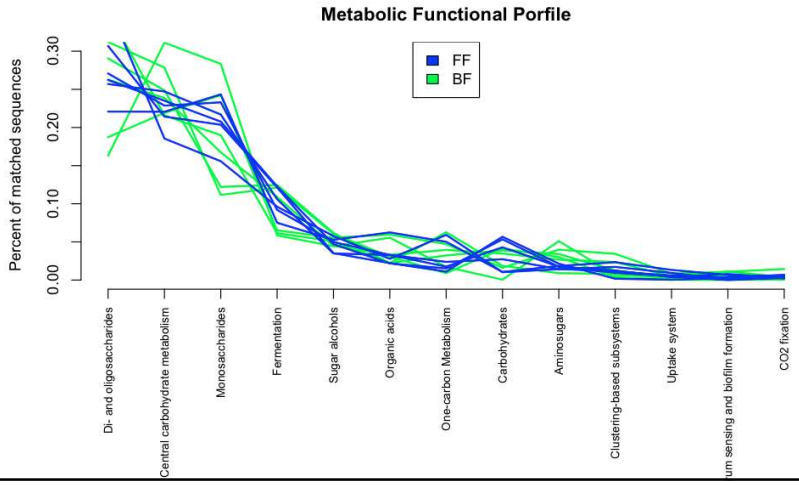
SEED level 2

- Carbohydrate metabolic profile



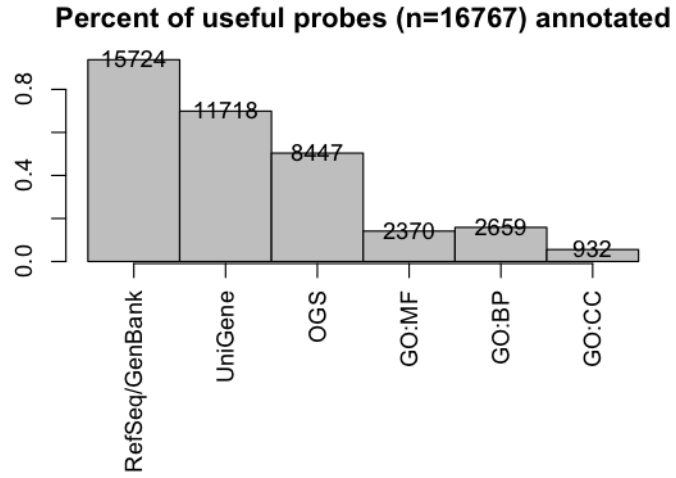
SEED level 2

- Again, not much classification power



Host Annotation

- CodeLink provided some annotation

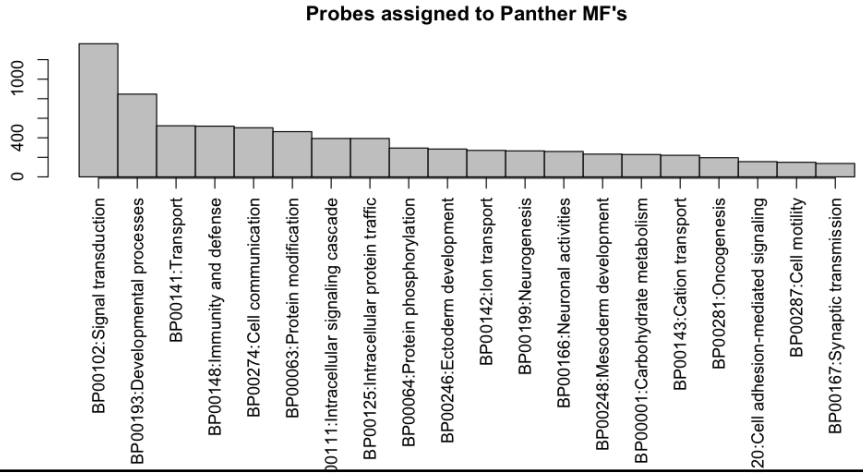


BP00148:Immunity and defense

- 71.5% (4974) of OGS annotated genes were present in Panther Biological Processes
- Of those, 7.5% (519) were categorized as BP00148:Immunity and defense
- 517 (? case issues?) were found back on codelink, in 630 total probes
- This subset of probes was taken forward to the next step of processing

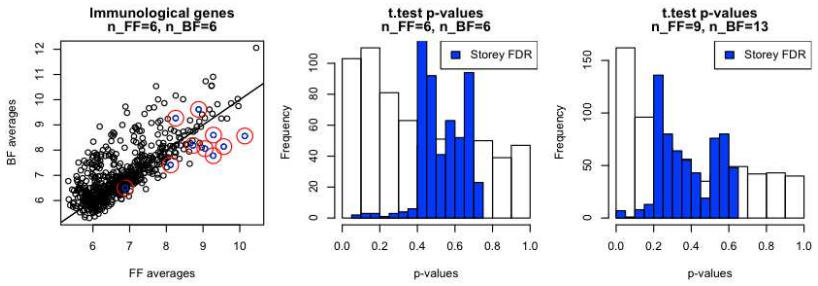
Panther Molecular Function

- Probes can be assigned Molecular Functions



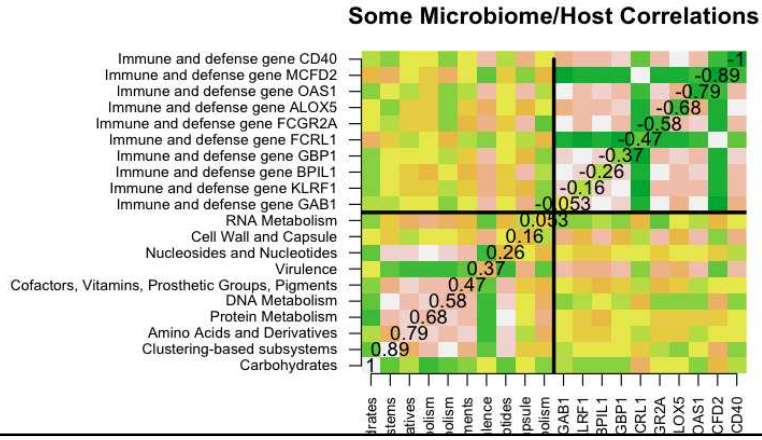
BP00148:Immunity and defense

- There are a handful of interesting Immunological genes relative to FF/BF
- Circled points are smallest 10 p-values



Correlation Structure

- Top 10 microbiome functional categories versus 10 most prominent host immune genes



Principal Components

- Iddo, are we getting good amount of mapping? Iddo, different data bases are different… should we be concerned?
- GO/KEGG/David/Ensemble … start looking at correlation structure… Manasvi/Jennifer
- Abstract Feb 10 (under 2 weeks)
- DeNovo searching versus positive control
- Keep flow diagrams – restrict to immunology
- DoBy

Principal Components

- DoBy
- X^2 Homogeneity test w/in treatment
- Loess/bayes nonparametric curve estimation (On just bacterial side… phylum or metabolic profile)… is dirichlet sample interesting.

Work Flow --

- Here's the plan…