DNA Microarrays
Post Analysis
BCB 504: Applied Bioinformatics

Matt Settles

University of Idaho
Bioinformatics and Computational Biology Program

February 6, 2013

Outline

Gene Ontology I

The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. T he project provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data. Read more about the Gene Ontology.

The Gene Ontology project provides an ontology of defined terms representing gene product properties. The ontology covers three domains: **cellular component**, the parts of a cell or its extracellular environment; **molecular function**, the elemental activities of a gene product at the molecular level, such as binding

Gene Ontology II

or catalysis; and **biological process**, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

For example, the gene product cytochrome c can be described by the molecular function term oxidoreductase activity, the biological process terms oxidative phosphorylation and induction of cell death, and the cellular component terms mitochondrial matrix and mitochondrial inner membrane.

Further, The GO ontology is structured as a directed acyclic graph, and each term has defined relationships to one or more other terms in the same domain, and sometimes to other domains. The GO

Gene Ontology III

vocabulary is designed to be species-neutral, and includes terms applicable to prokaryotes and eukaryotes, single and multicellular organisms.

Pathways I

Most common pathway resource is **KEGG**, The Kyoto Encyclopedia of Genes and Genomes. However their FTP resource is no longer free as of Aug 2011, but their website and API is still free.

WikiPathways is another resource gaining some traction. Pathways are intended to represent interactions betweens genes (proteins).

Most pathway resources have significant amount of manual curation to them.

Microarray Analysis I

Gene Ontology or Pathway Analysis is a data reduction technique used in order to summarize, or generalize results in a microarray experiment. It is not uncommone to either have hundreds, or even thousands, of differentially expresssed genes in an experiment. To summarize results on a gene by gene basis is not feasible and instead it is common to make more general statement based on Gene Ontologies or Pathways.

The biological question being asked is typically:
Are there Gene Ontologies (or Pathways) significantly over-represented (or under-represented) in my comparison as would be expected by chance.

Microarray Analysis II

There are many R packages (Bioconductor) for both Pathway and Gene Ontology analysis, many of which work for both kinds of data.
Bioconductor Ontology based packages: `http://bioconductor.org/help/search/index.html?q=ontology`
Bioconductor Pathway based package: `http://bioconductor.org/help/search/index.html?q=pathway`
SigPathway package example:
`http://www.bioconductor.org/packages/2.12/bioc/vignettes/sigPathway/inst/doc/sigPathway-vignette.pdf`

Weighted Gene Co-Expression Network Analysis (WGCNA) I

Many techniques employed for network based analysis (co-expression of genes), but my favorite is Dr. Horvath's (UCLA) approach.
`http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/Rpackages/WGCNA/`

Briefly, computes the correlation matrix (expression) between all genes in a dataset, manipulates the correlation so that conform to a scale free discribution (sfn). Then Computes the topological overlap measure (TOM), ie connections are enhanced in a friend of a friend manner and translates this to a distance measure. Clusters (networks) of genes are determined by a dynamic cutting routine

Weighted Gene Co-Expression Network Analysis (WGCNA) II

based on the heirarchical tree and k-means clustering results (hybrid-approach). These networks are then reduced by looking at their eigengenes (eigenvectors) and finnaly experimental parameters (phenotypes) can be associated with particular networks. In addition, genes within a network ranked by their connectivity, genes with high connectivity (central nodes) are assumed to be more important than other genes and may possibly represent gene targets.

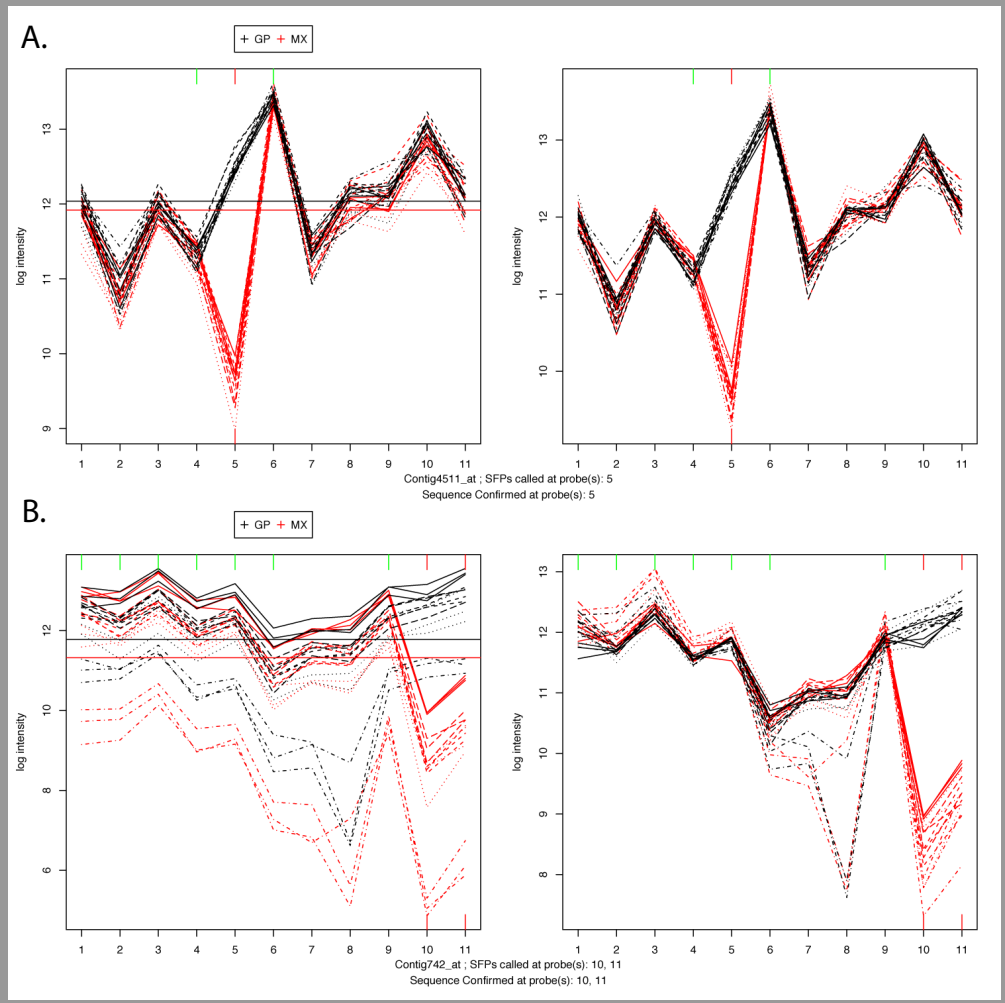
The WGCNA approach does however require a large number of samples, where the experimental conditions are expected to

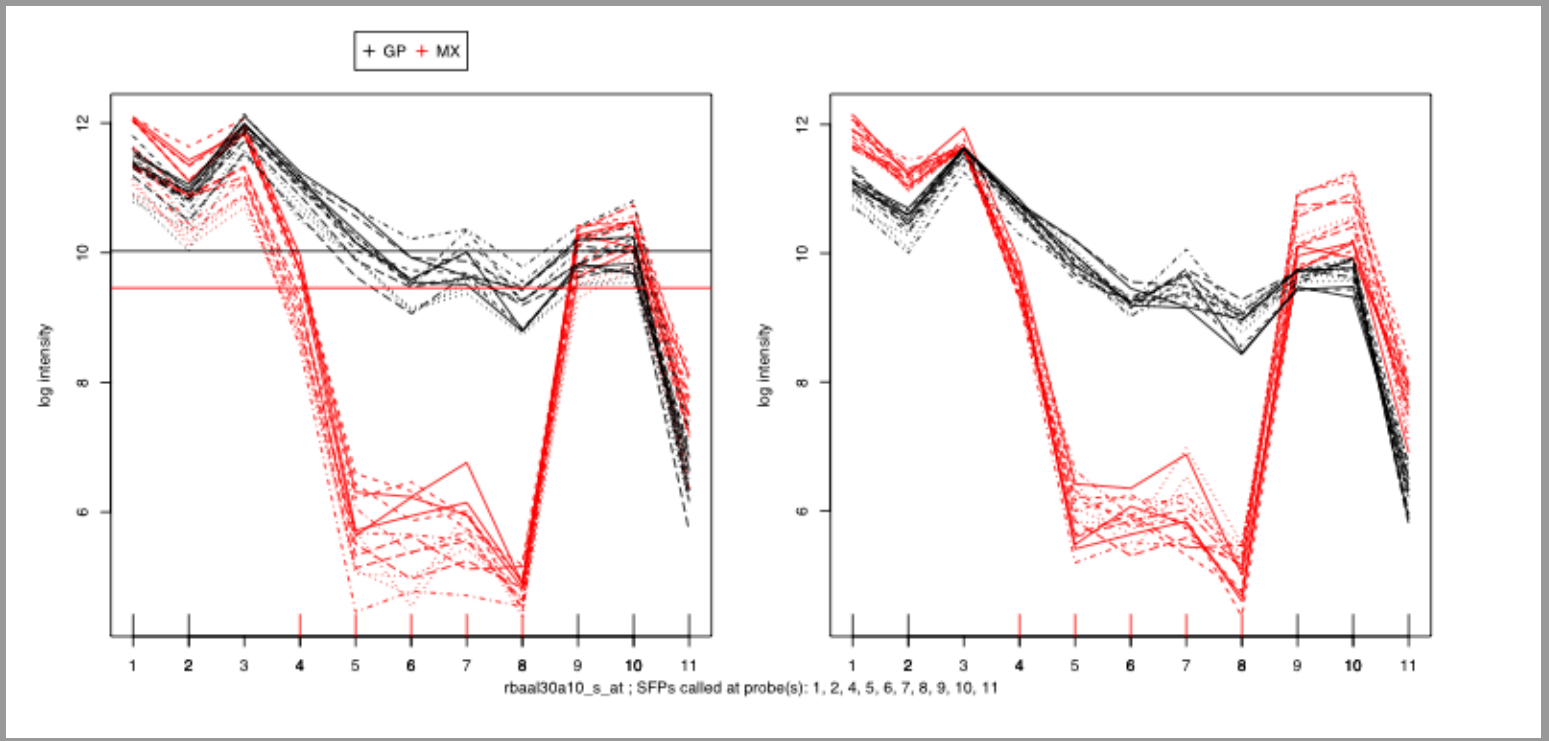
Weighted Gene Co-Expression Network Analysis (WGCNA) III

preturb gene expression. In other workds to develop networks you need to induce a pattern of gene expression across samples.

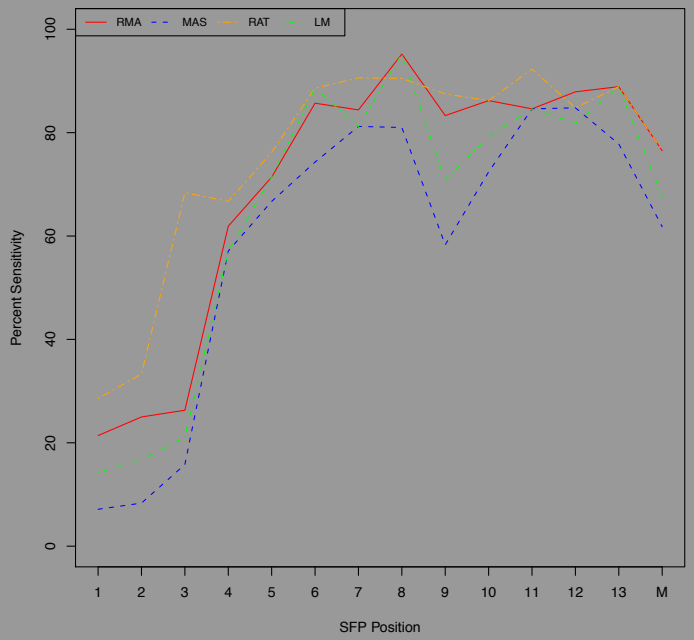
Single Feature Polymorphisms (SFPs)

When a short-oligonucleotide probe is designed at a position with a genomic or transcriptional polymorphism, the hybridization efficiency is reduced. SFPs are statistical differences in the probe level hybridization efficiency between two populations caused by an underlying genetic or transcriptional polymorphism. They are detected by comparing microarray probe level intensity signals, a proxy value for hybridization efficiency, between two populations. When hybridizing gDNA, SFPs are induced by single-nucleotide polymorphisms (SNPs) and small insertions/deletions (INDELS). When hybridizing mRNA, SFPs can also be induced by splicing variation and polyadenylation differences.

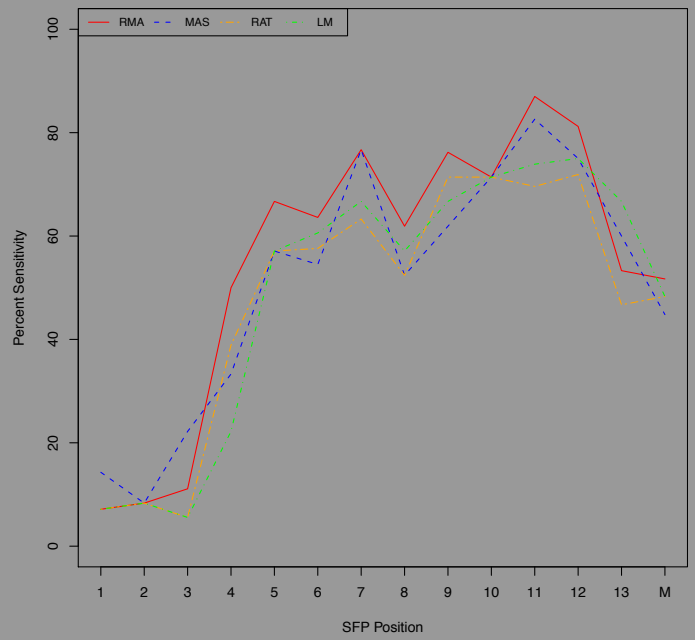

A.
+ GP + MX
log intensity
Contig4511_at ; SFPs called at probe(s): 5
Sequence Confirmed at probe(s): 5
B.
+ GP + MX
log intensity
Contig742_at ; SFPs called at probe(s): 10, 11
Sequence Confirmed at probe(s): 10, 11

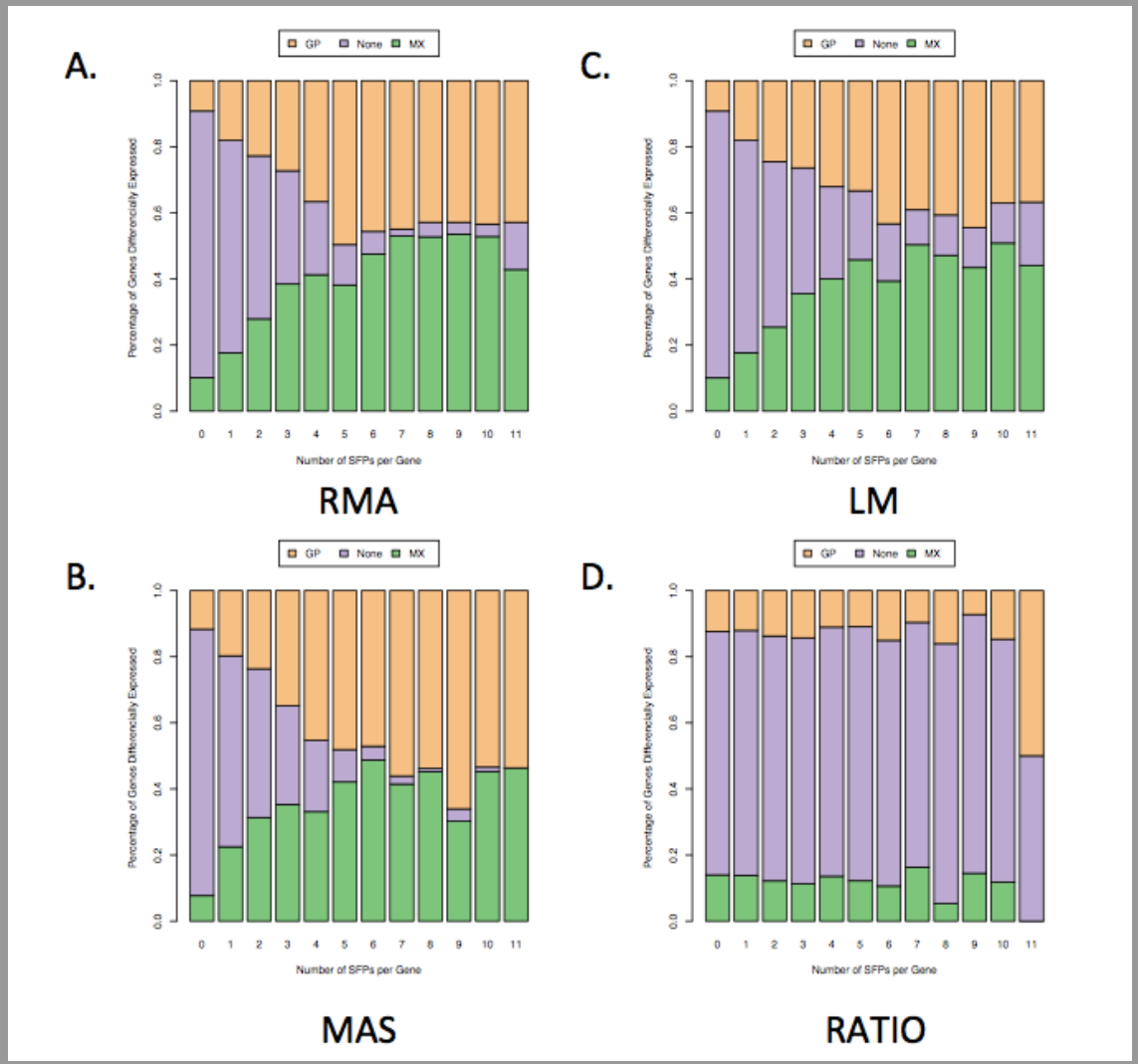

+ GP + MX
log Intensity
12
10
8
6
1
2
3
4
5
6
7
8
9
10
11
rbaal30a10_s_at ; SFPs called at probe(s): 1, 2, 4, 5, 6, 7, 8, 9, 10, 11

BB3



E-TABM-113




A.
GP
None
MX
Percentage of Genes Differencially Expressed
Number of SFPs per Gene
RMA
C.
LM
B.
MAS
D.
RATIO