DNA Microarrays
Quality Assurance and Preprocessing
BCB 504: Applied Bioinformatics

Matt Settles

University of Idaho
Bioinformatics and Computational Biology Program

January 23, 2012

Outline

Components of Raw Data

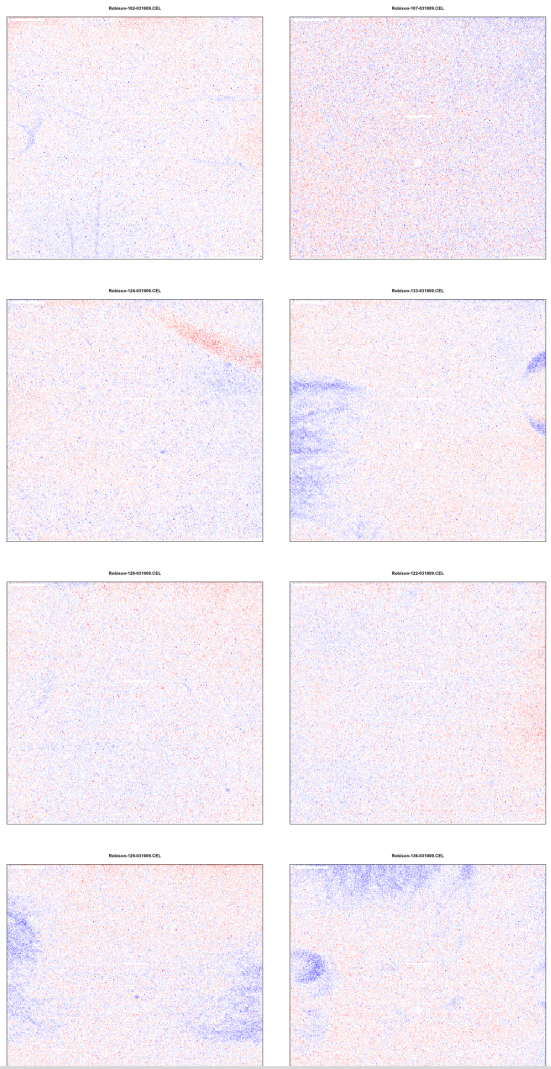
- Quantified intensities for each microarray probe (spot), usually identified by X,Y coordinate
  - Affymetrix: Binary CEL files
  - Nimblegen: plain text pair or xys files
- Microarray description file, associates probes to probesets and genes
  - Affymetrix: CEL description files or CDF files (Bioconductor provides)
  - Nimblegen: Nimblegen description files or NDF Files (must create your own packages)

Quality Assurance

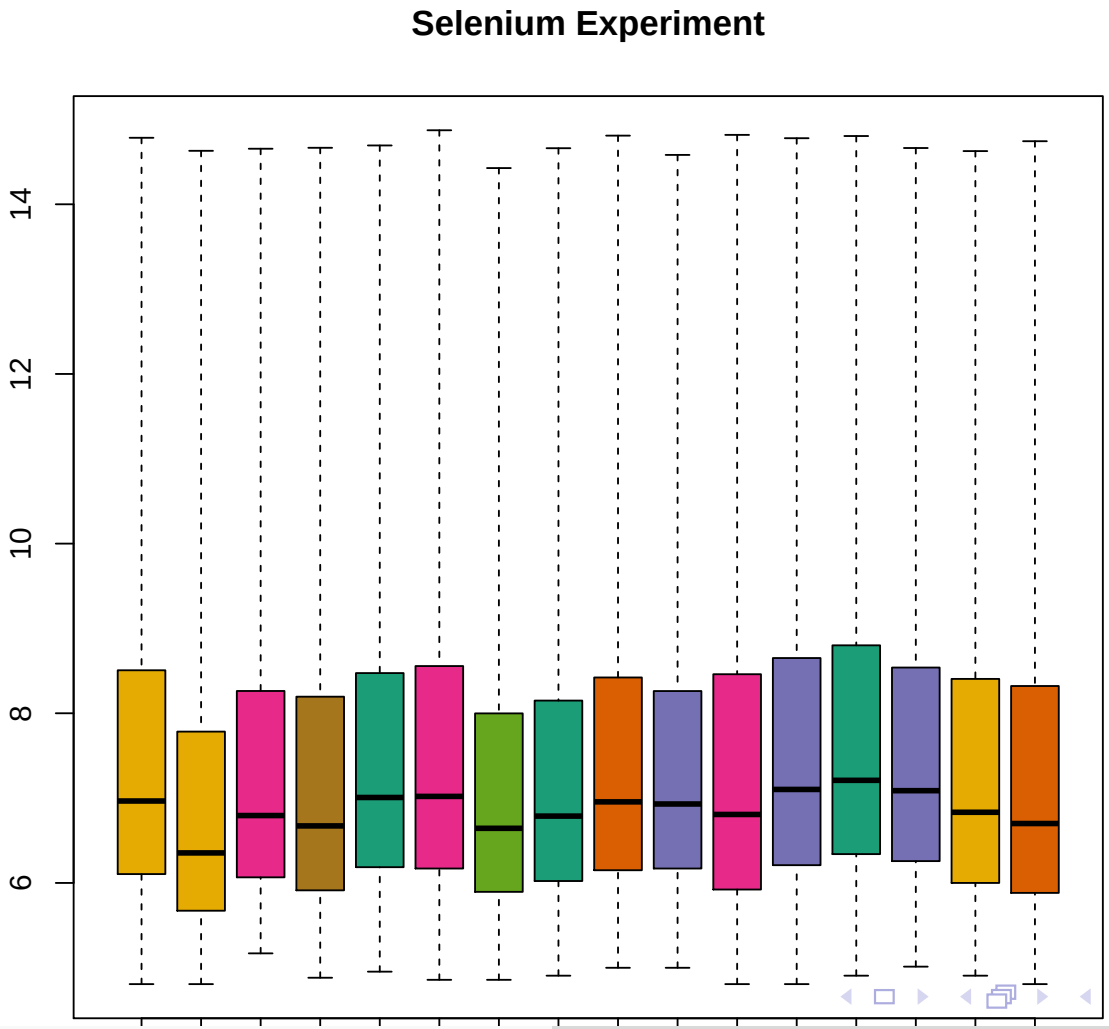
1. Visual check of chip pseudo-images.
2. Consistency across microarrays.
3. Observed patterns meet expectations.

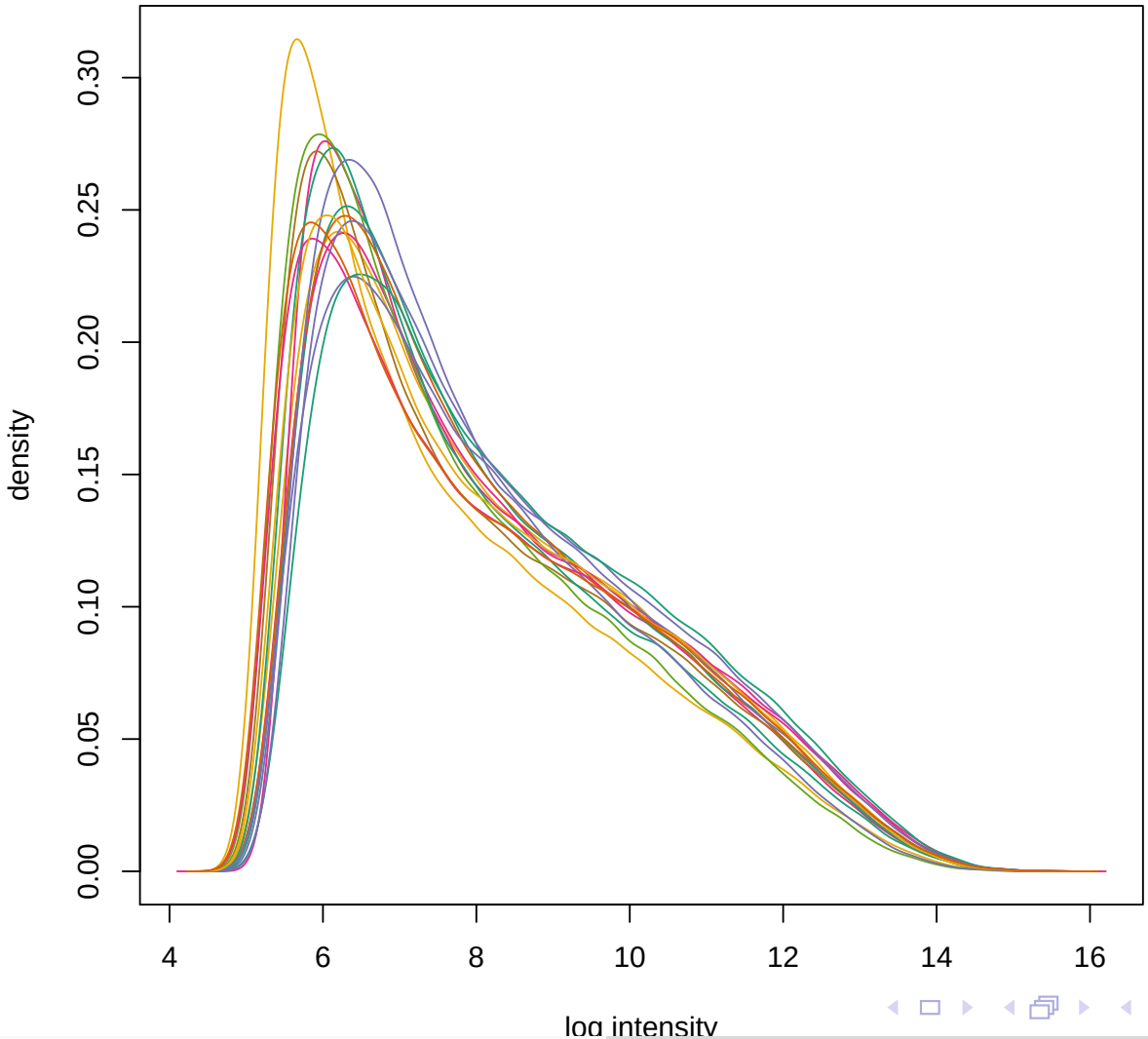
Pseudo-chip images

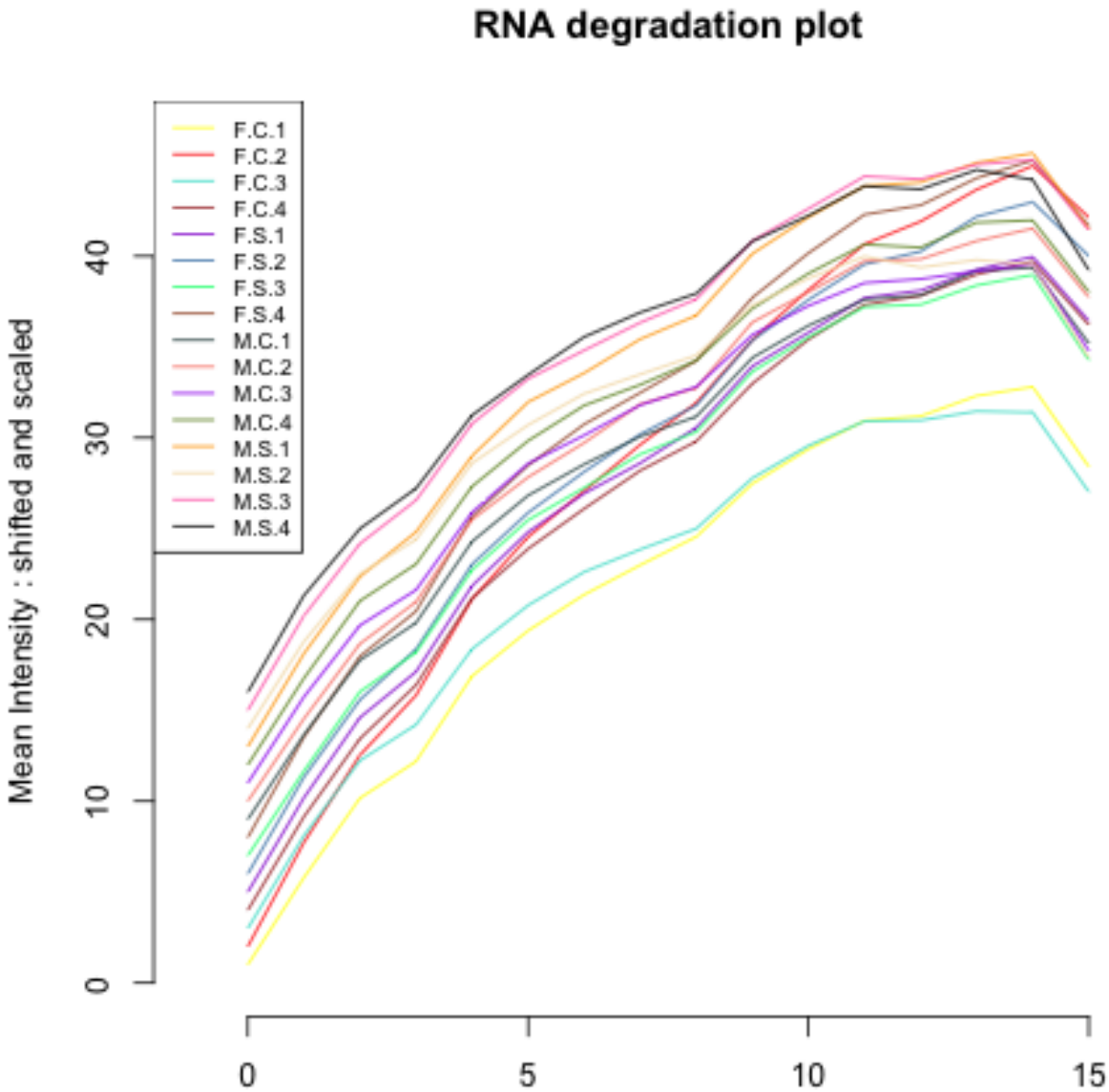
- Pseudo-chip images are created from the data themselves and not from the original scanned images
- Can be created from the raw intensities, log intensities, model fitting residuals and other
- Looking for large anomalies



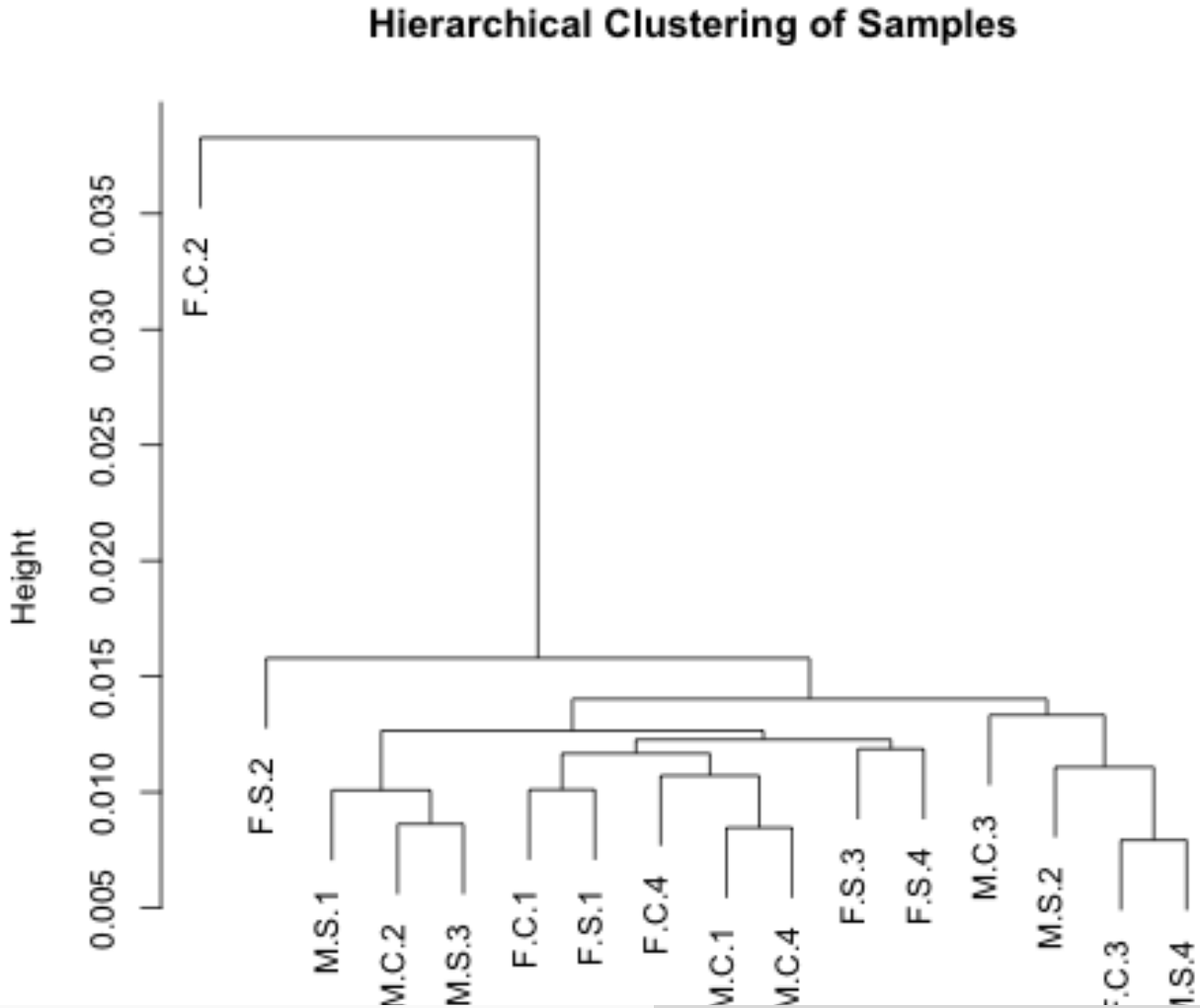
Consistency across microarrays




density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
4
6
8
10
12
14
16
log intensity


RNA degradation plot
Mean Intensity : shifted and scaled
0
10
20
30
40
0
5
10
15
F.C.1
F.C.2
F.C.3
F.C.4
F.S.1
F.S.2
F.S.3
F.S.4
M.C.1
M.C.2
M.C.3
M.C.4
M.S.1
M.S.2
M.S.3
M.S.4

Expectations



Preprocessing

The goals of preprocessing Affymetrix microarray data are three fold:

1. To remove variation due to technical sources, while preserving variation from biological sources.
2. To normalize a set of microarrays in order to make them comparable.
3. To produce summarized expression values for each probe set.

Background correction Deviations from actual expression levels are introduced by many sources including scanner artifacts, non-specific binding (hybridization), RNA quality, reagents, etc.. All of which can be considered as background noise.

Probe level normalization The role of normalization is to compensate for the technical effects, while preserving the effects due to the biology.

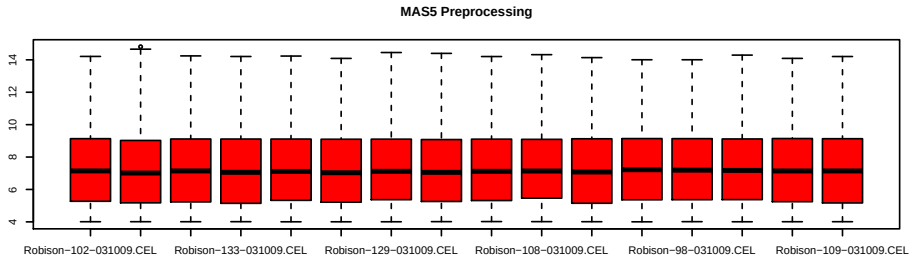
PM Correction PM correction routines are designed to account for non-specific binding by use of the MM probes.

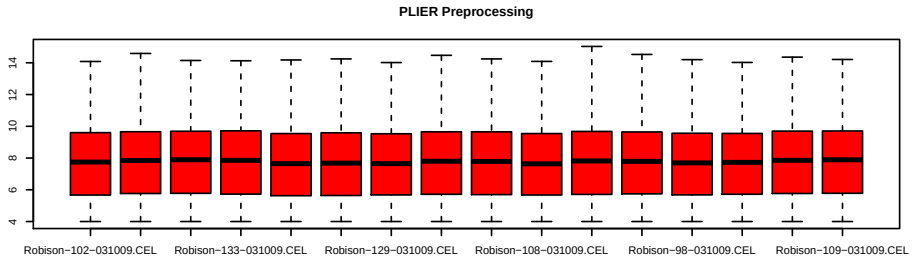
Probe set summarization Probe set summarization produces a single value for a probe set that is the estimated expression level for the probe set (i.e. gene).

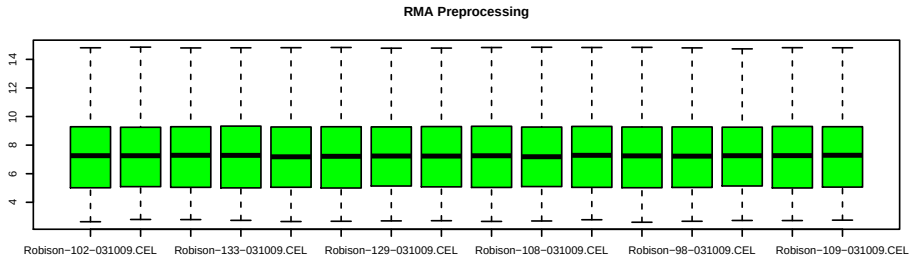
Probe set normalization Same purpose as Probe level normalization but on the probe sets.

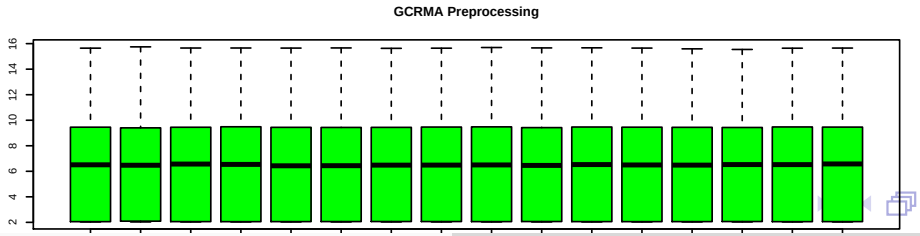
Table: Breakdown of the preprocessing steps for the MAS5, Plier, RMA, GCRMA, and dChip preprocessing pipelines.

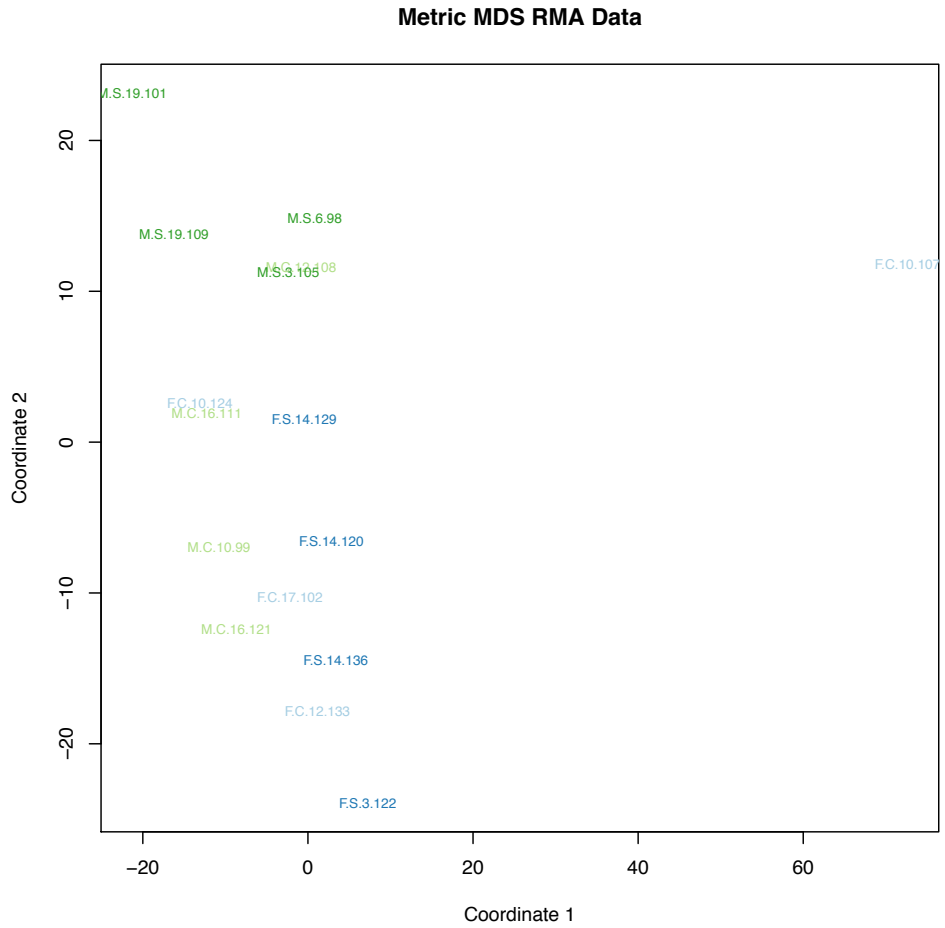
	MAS5	**Plier**	**RMA**	**GCRMA**	**dChip**
Background Correction	weighted average	none	RMA (global model)	GCRMA (model based)	none
Probe Level Normalization	none	quantile normalization	quantile normalization	quantile normalization	invariant set
PM Correction	ideal mismatch	none	none	none	subtract MM none
Probe set Summarization	Tukey biweight	Plier	median polish	median polish	MBEI
Probe set Normalization	mean scaled	none	none	none	none


MAS5 Preprocessing
4
6
8
10
12
14
Robison−102−031009.CEL
Robison−133−031009.CEL
Robison−129−031009.CEL
Robison−108−031009.CEL
Robison−98−031009.CEL
Robison−109−031009.CEL


PLIER Preprocessing
4
6
8
10
12
14
Robison−102−031009.CEL
Robison−133−031009.CEL
Robison−129−031009.CEL
Robison−108−031009.CEL
Robison−98−031009.CEL
Robison−109−031009.CEL


RMA Preprocessing
4
6
8
10
12
14
Robison−102−031009.CEL
Robison−133−031009.CEL
Robison−129−031009.CEL
Robison−108−031009.CEL
Robison−98−031009.CEL
Robison−109−031009.CEL


GCRMA Preprocessing
2
4
6
8
10
12
14
16


Metric MDS RMA Data
M.S.19.101
M.S.6.98
M.S.19.109
M.S.3.105
F.C.10.107
F.C.10.124
M.C.16.111
F.S.14.129
M.C.10.99
F.S.14.120
F.C.17.102
M.C.16.121
F.S.14.136
F.C.12.133
F.S.3.122
Coordinate 2
20
10
0
−10
−20
−20
0
20
40
60
Coordinate 1

- Experiment_Directory
  - Raw_Data
  - Design_Files
  - Figures
  - Tables
  - Data
  - Analysis.txt and/or Analysis.R
  - targets.txt

targets.txt - an annotated data frame

An annotated data frame allows you to manage the samples for the experiment and includes at minimum sample names with associated raw data files. The framework however is flexible and allows for more information about samples to be included.

See targets.txt file in example data

Public Repositories for Microarrays Data

GEO http://www.ncbi.nlm.nih.gov/geo/

ArrayExpress http://www.ebi.ac.uk/arrayexpress/