## DATA PROCESSING

**STEP I**

Normalization and background correction using frozen robust multichip average (fRMA)

**OUTPUT**

***m x n* Matrix**
***m*** Probesets ($\log_2$) normalized intensities for ***n*** samples

**STEP II**

Principal components analysis (PCA) and clustering to visualize and identify outlier samples

Filtered ***n – y*** samples

**STEP III**

Determine and filter present/absent/marginal calls using approach of Presence Absence calls with Negative Probesets (PANP)

Filtered ***m - x*** probesets

**STEP IV**

Perform surrogate variable analysis (SVA) to identify systematic measured and unmeasured sources of heterogeneity; use linear regression of probeset intensities with SVA covariates to obtain residuals with sources of heterogeneity removed

***(m - μ) x n Matrix***
Probeset level residual Intensities after correcting for expression heterogeneity

**STEP V**

Assign probesets to Refseq genes ***G***, using approach of JetSet

***G x n Matrix***
Gene level residual intensities

## DATA ANALYSIS

Analyze gene level data from all studies together (mega-analysis) using **random effects linear regression with crossed random effects** (lme4 package in R) to account for within study and within subject correlation (when subjects are used across multiple studies)