# Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis

*Gloria Chi and Marea Cobb*

*March 8, 2015*

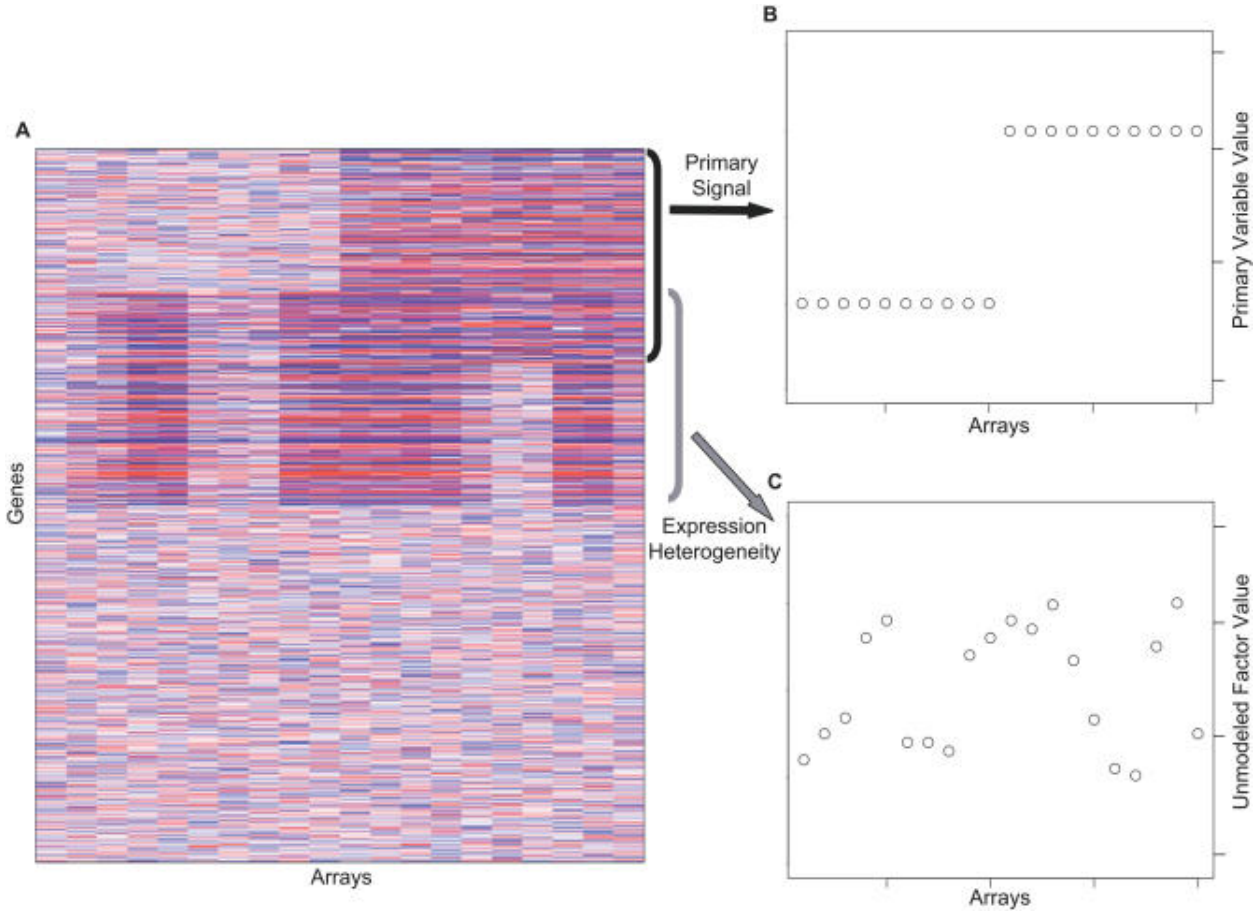
## Gene Expression Students

- Characterize transcriptional variablity
- Models fail to include unmodeled or unmeasured factors
- Noise can lead to a decrease power in detecting association

So, how do we handle background noise?

## Expression Heterogeneity (EH)

- Describes patterns of variation due to unmodeled factors
- Commonly expressed in human expression data and compelx systems
- Sources include technical, environmental, demographic, genetic factors, etc.



## Proposed Solution: Surrogate Variable Analysis (SVA)

- Identifies, estimates, and utilizes the components of EH
- Improves accuracy and consistency in detecting differential expression
- Captures signatures of EH and uses them as covariates in differential expression analysis

## Algoirthm Overview: Step 1

- Remove the signaldue to the primary variables to obtain a residual expression matrix
  - Form estimates $\hat{\mu_i}$ and $\hat{f}_i$ by fitting the model to $x_{ij} = \mu_i + f_i(y_j) + e^*_{ij}$
  - Caclulate residual expression matrix $R$ where (i,j) element is $r_{ij}$
    $r_{i,j} = x_{ij} - \hat{\mu_i} + \hat{f}_i(y_j)$
- Apply a decomposition to the residual expression matrix to identify signatures of EH (identifies signatures of the EH)
  - $d_l$ is the lth orthagonal signatures of EH, "eigenvalue"
  - $k$ is a gene corresponding to the signatures of EH, "eigengene"
  - Calculate a null statistic for each gene

$$T_k = \frac{d_k^2}{\sum_{l=1}^{n-df} d_{0l}^2}$$

- Use a statistical test to determine the singular vectors that represent more variation than is expected by chance
  - calculate a p-value for the eigengene k

$$p_k = \frac{\#T_k^{0b} >= T_k; b = 1, ..., B}{B}$$

## Algorithm Overview: Step 2

- Identify subset of genes driving each signature
- Repeat step 1 including the signature eigenvalues

## Algorithm Overview: Step 3

- For each subset of genes, build a surrogate variable based on the full EH signature
- Build matrix containing all genes associated with the residual eigengene

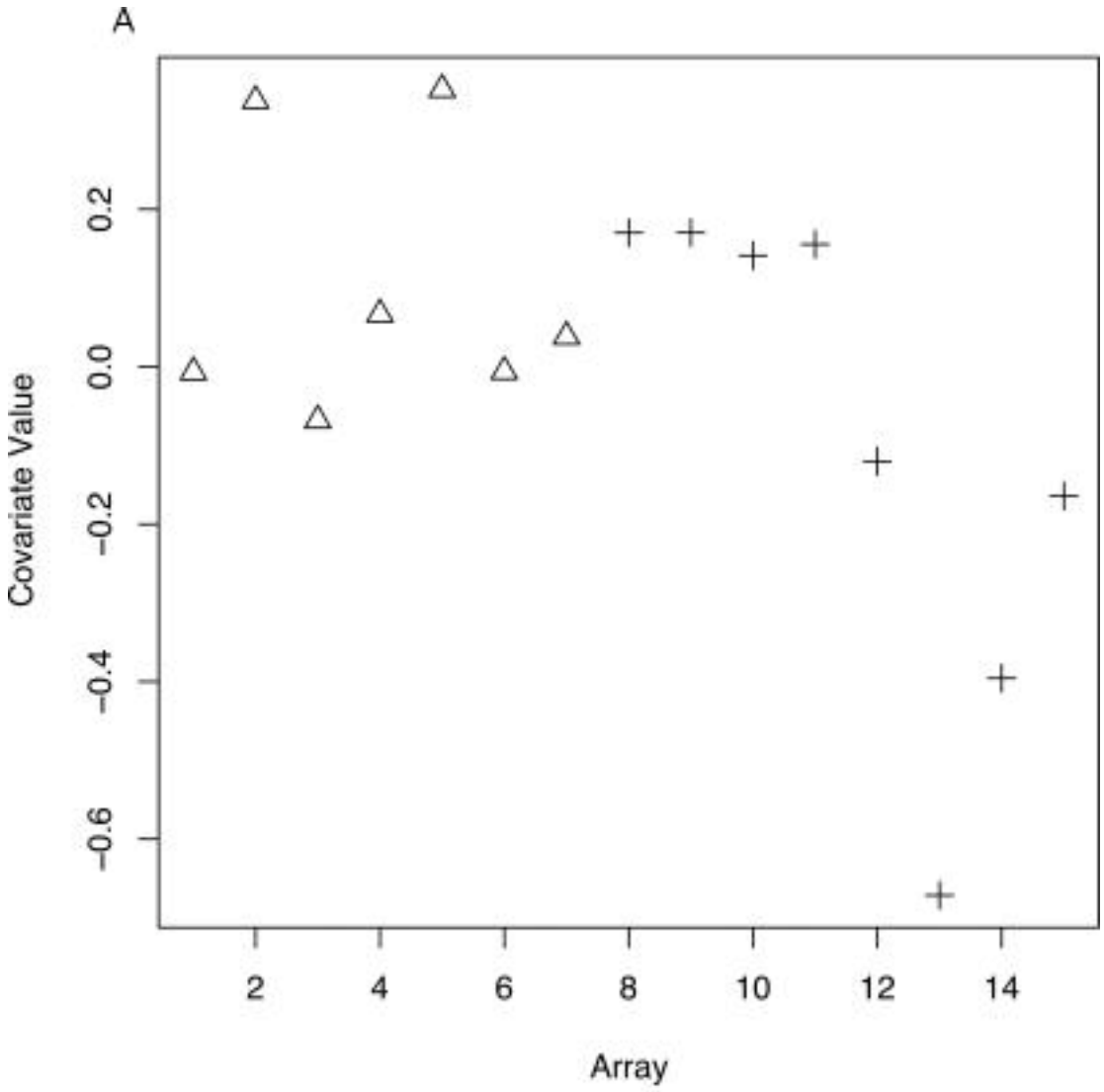
## Algorithm Overview: Step 4

- Include all surrogate variables as covariates in subsequent regression analysis

$$x_{ij} = \mu_i + f_i(y_j) + \sum_{k=1}^{K} \gamma_{ki} \hat{h_{kj}} + e_{ij}$$

## Gene-expression Profiles in Hereditary Breast Cancer. Hedenfalk et, al.

- BRCA1 and BRCA2 tumor samples
- Identify genes that showed differential expression across tumor subtypes




## Outcome

- Accurately estimate the signatures of expression heterogeneity
- Corrects the null distribution of p-values
- Improves estimation of the false discovery rate
- Robust to confounding between the primary variables and surrogate variables

## Extra

Generality Model $x_{ij} = \mu_i + f_i(y_j + e_{ij})$

$\gamma_{li}$ = gene-sepcific coefficient for the lth unmodeled factor

Expression for gene i on array j $x_{ij} = \mu_i + f_i(y_j + \sum_{l=1}^{L} \gamma_{li} g_{li} + e^*_{ij})$

Remove signal of primary variables creating a residual expression matrix

- Normalized expression matrix $X_{mxn} = (x_1, ..., x_m)^T$
- Vector representing primary variable of interest $y = (y_1, ..., y_n)^T$
- Baseline level of expression $\mu$

$$x_{ij} = \mu_i + f_i(y_j + \sum_{l=1}^{L} \gamma_{li} g_{li} + e^*_{ij})$$

$$x_{ij} = \mu_i + f_i(y_j + \sum_{k=1}^{K} \gamma_{ki} g_{ki} + e^*_{ij})$$