Principal Variance Components Analysis for Quantifying Variability in Genomics Data

Tzu-Ming Chu
SAS Institute Inc.

Outlines

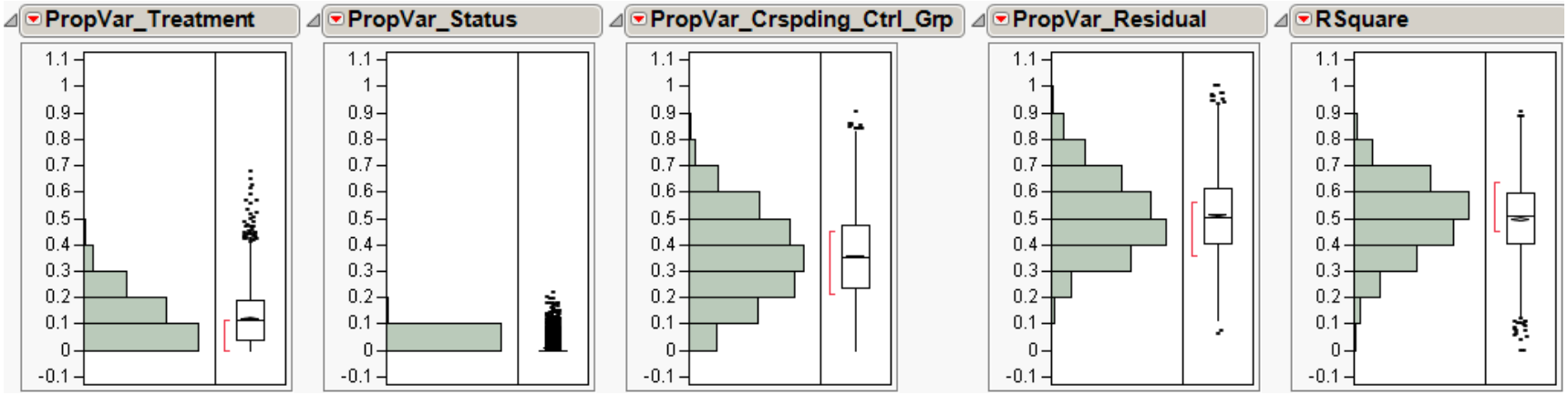
- Motivation
- PVCA (PCA + VCA)
- Example
  - Mouse Lung Tumorigenicity Data
  - Grouped Batch Profile Normalization (GBP)
- Discussion

Motivation - ANOVA

- Estimating variations of variance components among SAMPLEs
  - ANOVA
  - Some ANOVA approaches applied in microarrays
    - Kerr *et al*. 2000 (Full-genes Model)
    - Wolfinger *et al.* 2001 (Gene-by-Gene Mixed Model)
    - Chu *et al*. 2002 (Probe-Level Mixed Model)
    - Feng *et al*. 2006 (Empirical Bayesian Mixed Model)

Motivation - ANOVA

- ANOVA issues
  - Lack of gene by gene correlations
  - Quantification of variations among genes



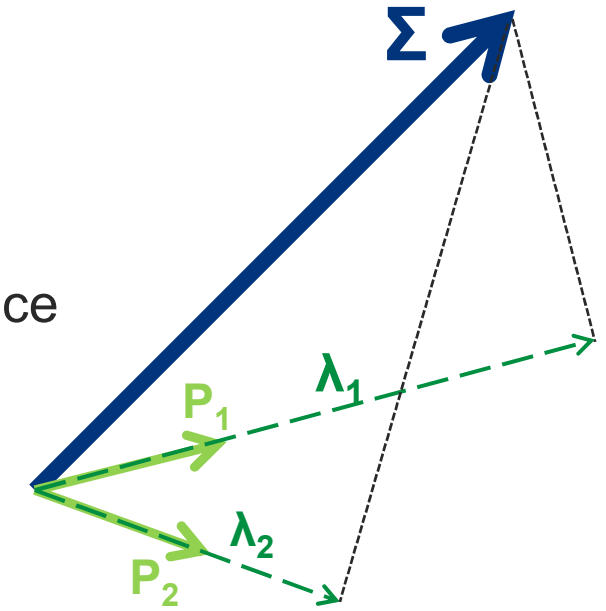
Y = Crspding_Ctrl_Grp + Treatment + Status + Residual

Principal Variance Components Analysis

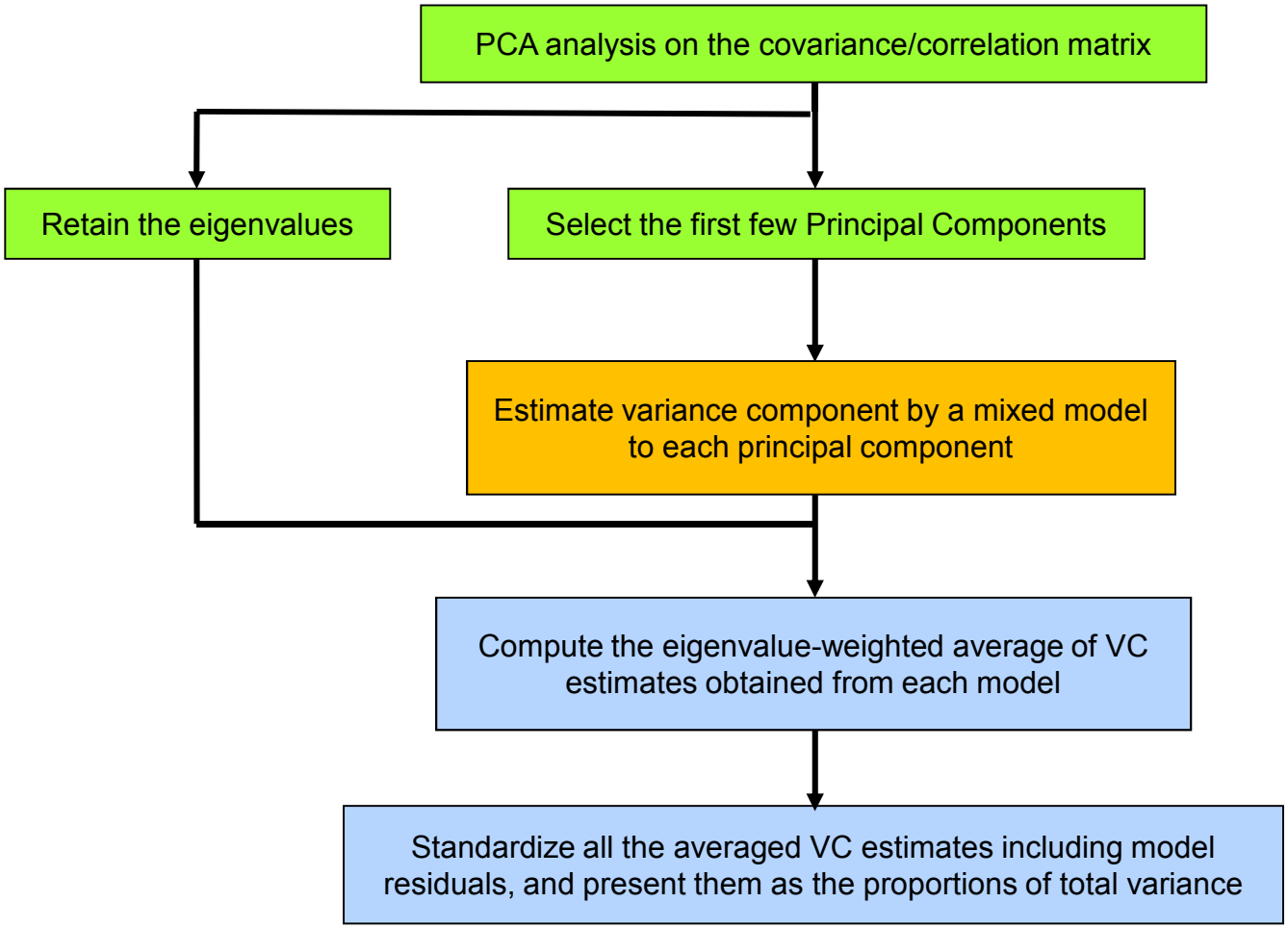
- PCA – the foundation of PVCA
  - Variance-Covariance Matrix
  - Mutually orthogonal among principal components
- VCA
  - Mixed Model on individual principal component
- Weighted proportion of variations
  - Eigen values
- Refer to Li *et al*. 2009, www.wiley.com/WileyCDA/WileyTitle/productCd-0470741384.html

PVCA - PCA

- Projection onto eigenvectors
  - $\Sigma P = \Sigma[P_1, P_2, \cdots, P_n]$
    $= [\Sigma P_1, \Sigma P_2, \cdots, \Sigma P_n]$
    $= [\lambda_1 P_1, \lambda_2 P_2, \cdots, \lambda_n P_n]$
  - Eigenvalue λ_i represents the variance associated with the i^{th} principal component



PVCA – Analysis Flow Chart



Mouse Lung Tumorigenicity Data

- Develop microarray-based biomarkers to predict the results from a rodent cancer bioassay following a short term chemical exposure (NTP 1996)
- Data from www.ncbi.nlm.nih.gov (GSE17933)
- 26 chemicals from 4 studies over 3 years (2005-7)
- 158 Affymetrix Mouse Genome 420 2.0 GeneChips
- One of the six MAQCII data

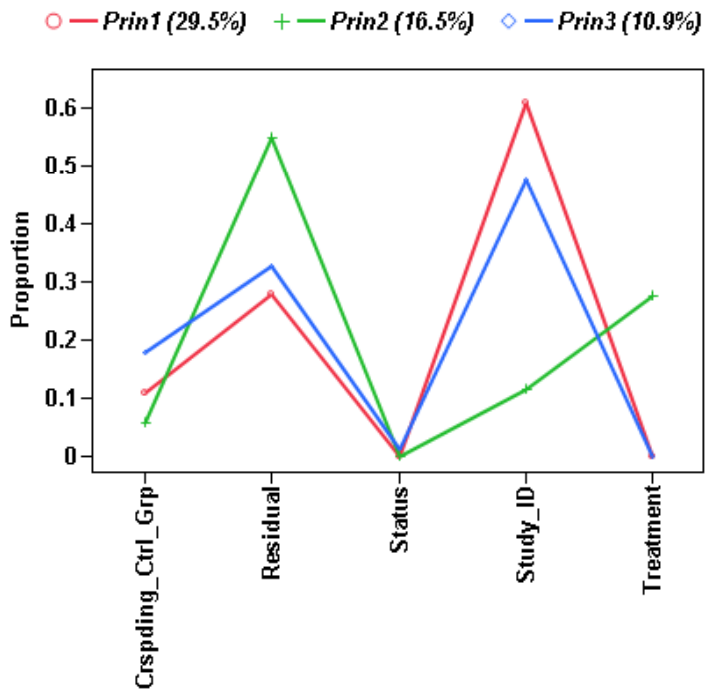
Mouse Lung Tumorigenicity Data

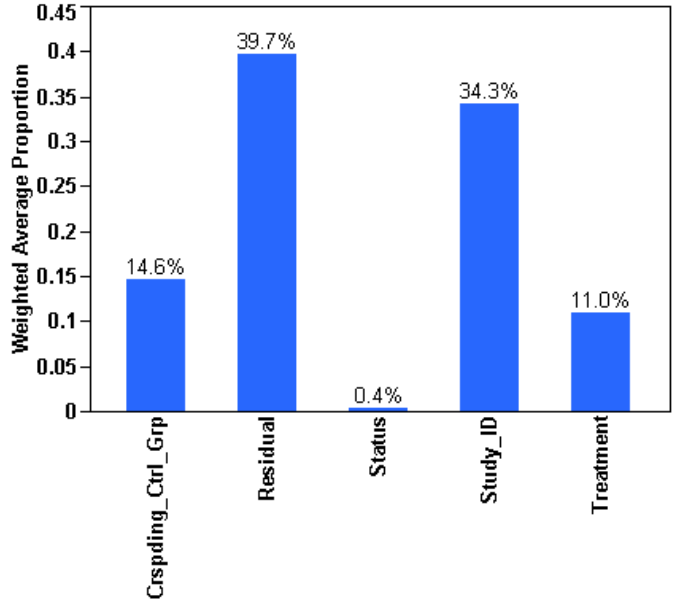
Batch	Animal Study ID	Animal Batch	Vehicle Group[a]	Controls	CA	NCA	Compounds
1	5003	1	FCON	1	1	2	4
2	5003	1	CCON	1	1	0	2
3	6004	1	FCO2	1	1	3	5
4	6004	2	CCO2	1	2	0	3
5	6004	3	CCO3	1	1	0	2
6	6004	4	WCO1	1	1	1	3
7	6014	1	CCO4	1	0	2	3
8	6014	2	FCO1	1	2	0	3
9	6014	3	ACO1	1	0	2	3
10	7008	1	CCO5	1	0	1	2
11	7008	2	ACO3	1	1	1	3
12	7008	3	CCO6	1	1	0	2
13	7008	4	ACO4	1	2	0	3
14	7008	5	ACO5	1	1	0	2

[a]The first three letters in the abbreviation define the vehicle and route used for administration (FCO = food; CCO = corn oil by gavage; WCO = water by gavage; ACO = air by inhalation). The last number or letter refers to the specific batch the vehicle was used. CA=Carcinogenicity.

PVCA Results

Y = Crspding_Ctrl_Grp + Study_ID+ Treatment + Status + Residual





- Average across 12 principal components
- 80.95% of total variation

Grouped Batch Profile (GBP) Normalization

- Batch profile estimation
  - Based on control arrays
- Batch profile grouping
  - Clustering on similar profiles
- Batch profile correction
- Batch profile scoring

GBP – Batch Profile Estimation

- Estimation based on control arrays
  - Assuming additive model

$$Y = \alpha(s) + \beta(b) + e$$

  - Model for Control Arrays

$$Y_c = \alpha(c) + \beta(b) + e$$

  - For simplicity

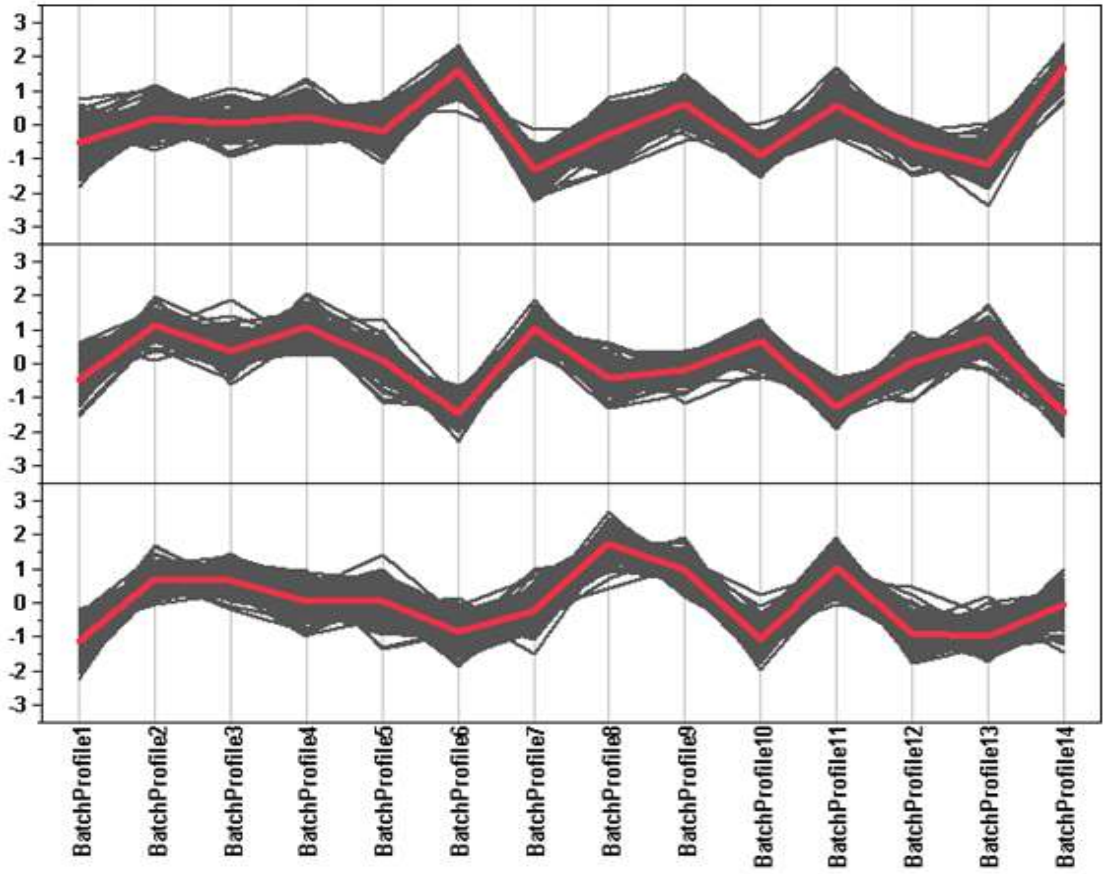
$$Y_{ci} = \mu_c + \beta_i + e$$

  - Estimated Batch Profile

$$\widehat{\beta_i} = \bar{Y}_{ci} - \bar{\bar{Y}}_{ci}$$

GBP – Profile Clusters

- Clustering on standardized profiles



GBP – Batch Profile Correction

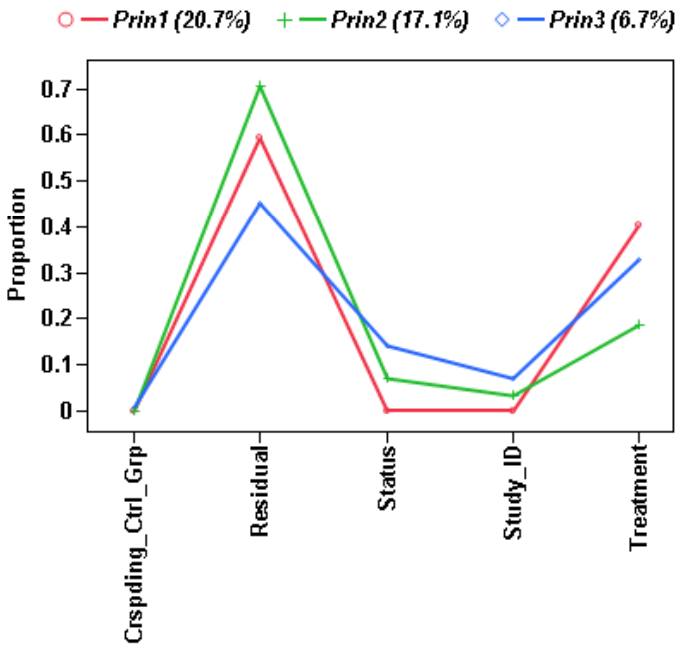
- Extract mean profile from each cluster to be the representative profile, $\widetilde{\boldsymbol{\beta}_g}$
- Correct the corresponding representative profile from data

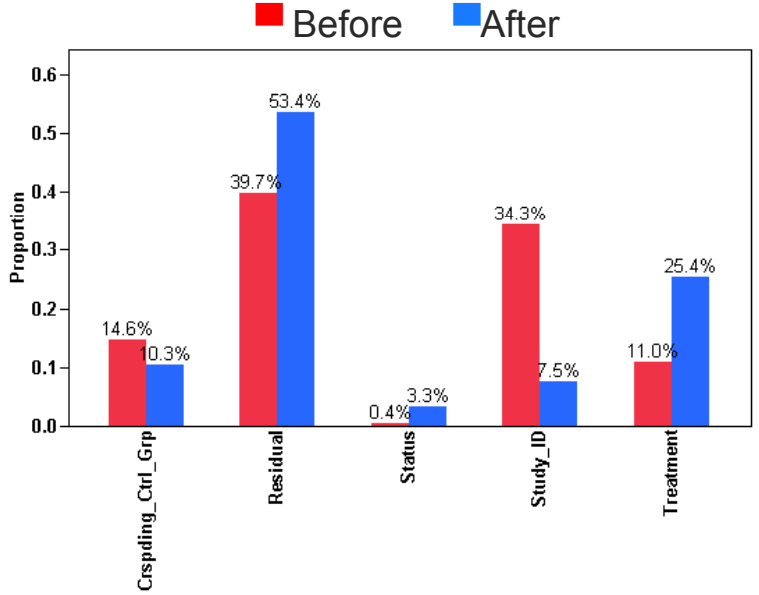
$$\widetilde{\boldsymbol{Y}} = \boldsymbol{Y} - \widetilde{\boldsymbol{\beta}_g}$$

- Scoring new batches

$$\widehat{\boldsymbol{\beta}_j} = \overline{\boldsymbol{Y}}_{cj} - \overline{\overline{\boldsymbol{Y}}}_{ci}$$

PVCA – GBP Results



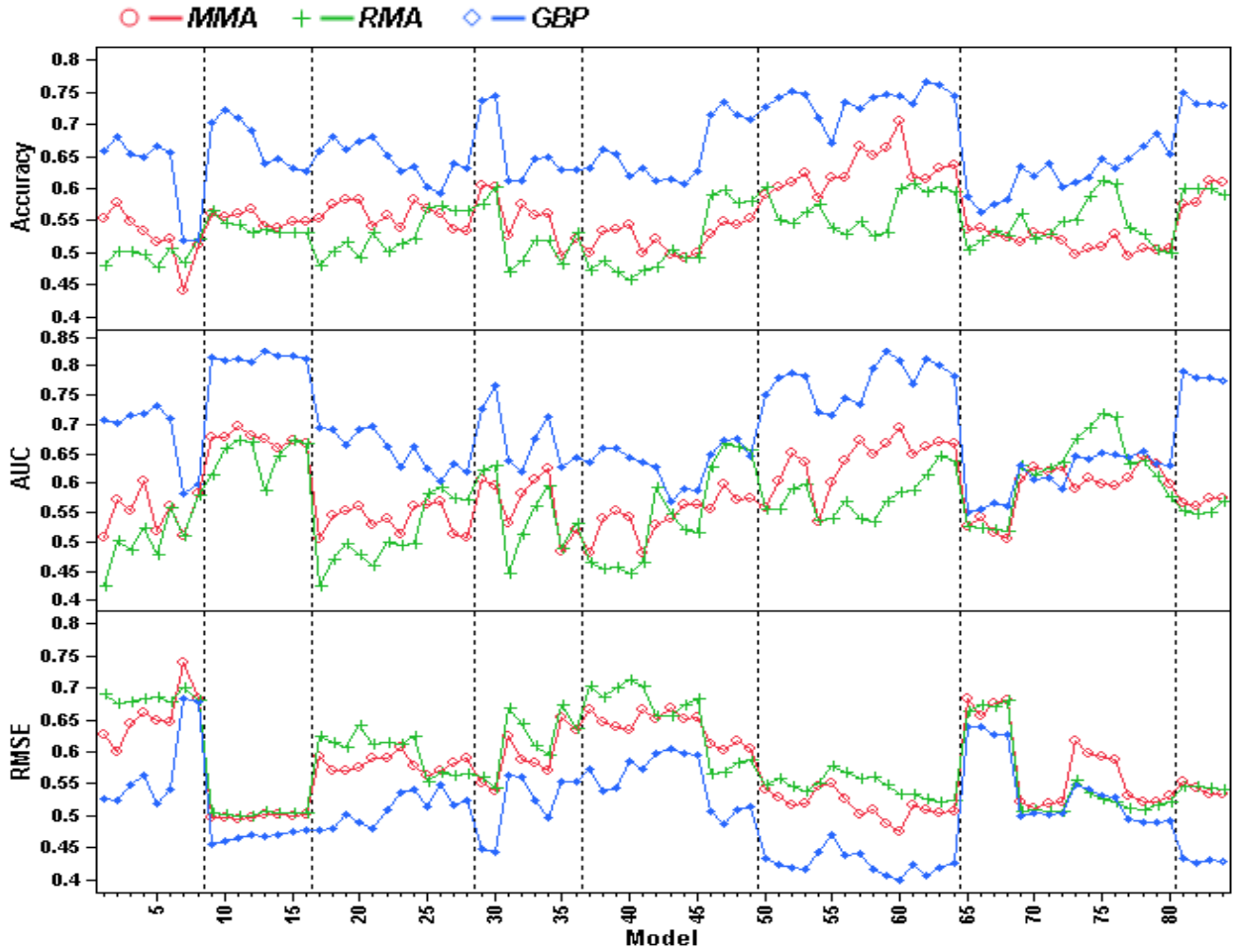


- Average across 19 principal components
- 80.96% of total variation

Prediction of Carcinogenicity

- 84 prediction models from 8 types of models (Discriminant Analysis, Distance Scoring, General Linear Model, K Nearest Neighbors, Logistic Regression, Partial Least Squares, Partition Trees, and Radial Basis Machine)
- Randomly hold out 3 batches from 14 batches as validation data
- Repeat 50 times and assess on Accuracy, AUC, and RMSE
- Apply on 3 normalized data (MMA,RMA,GBP)
- Refer to Chu *et al*. 2009 for details

Cross Validation Results



Discussion

- PVCA provides estimations of variance component among arrays without excluding covariance between genes
- PVCA can be a useful tool for exploring data
- PVCA can be a good assessment tool for normalization to quantify how much of variation changed across variance components
- R and SAS code is available at www.niehs.nih.gov/research/resources/software/pvca

Acknowlegements

- Jianying Li (UNC/Chapel Hill)
- Pierre Bushel (NIEHS)
- SAS Colleagues

  Russ Wolfinger

  Wenjun Bao

  Li Li

  Shannon Conners

SEEING IS BELIEVING