Gene signature selection to predict survival benefits from adjuvant chemotherapy in NSCLC patients

[1,2]Keyue Ding, Ph.D.

Nov. 8, 2014

[1]NCIC Clinical Trials Group, Kingston, Ontario, Canada
[2]Dept. Public Health Sciences, Queen's Univ. Kingston, Ontario, Canada

Outline

- Introduction: Rationale and Objectives
- Microarray Data
- Data preprocessing
  - Normalization
  - Adjusting batch effect
- Predictive gene signature selection
  - Statistical methods
  - Analysis procedure
  - Results
- Summary

Introduction

- Early stage non-small cell lung cancer (NSCLC)
  - Surgery is standard treatment
  - 35-50% will relapse within 5 years even after complete resection

- Adjuvant chemotherapy
  - Clinical trials demonstrate modest benefit: 4-15% for 5-yr survival
  - (Meta-analysis showed a 8.9% 5-yr survival benefit from cisplatin-vinorelbine)
  - Clinical trial results respect to treatment effect of entire population
  - May only benefit to a group patients
  - May cause serious adverse effects and detrimental effects

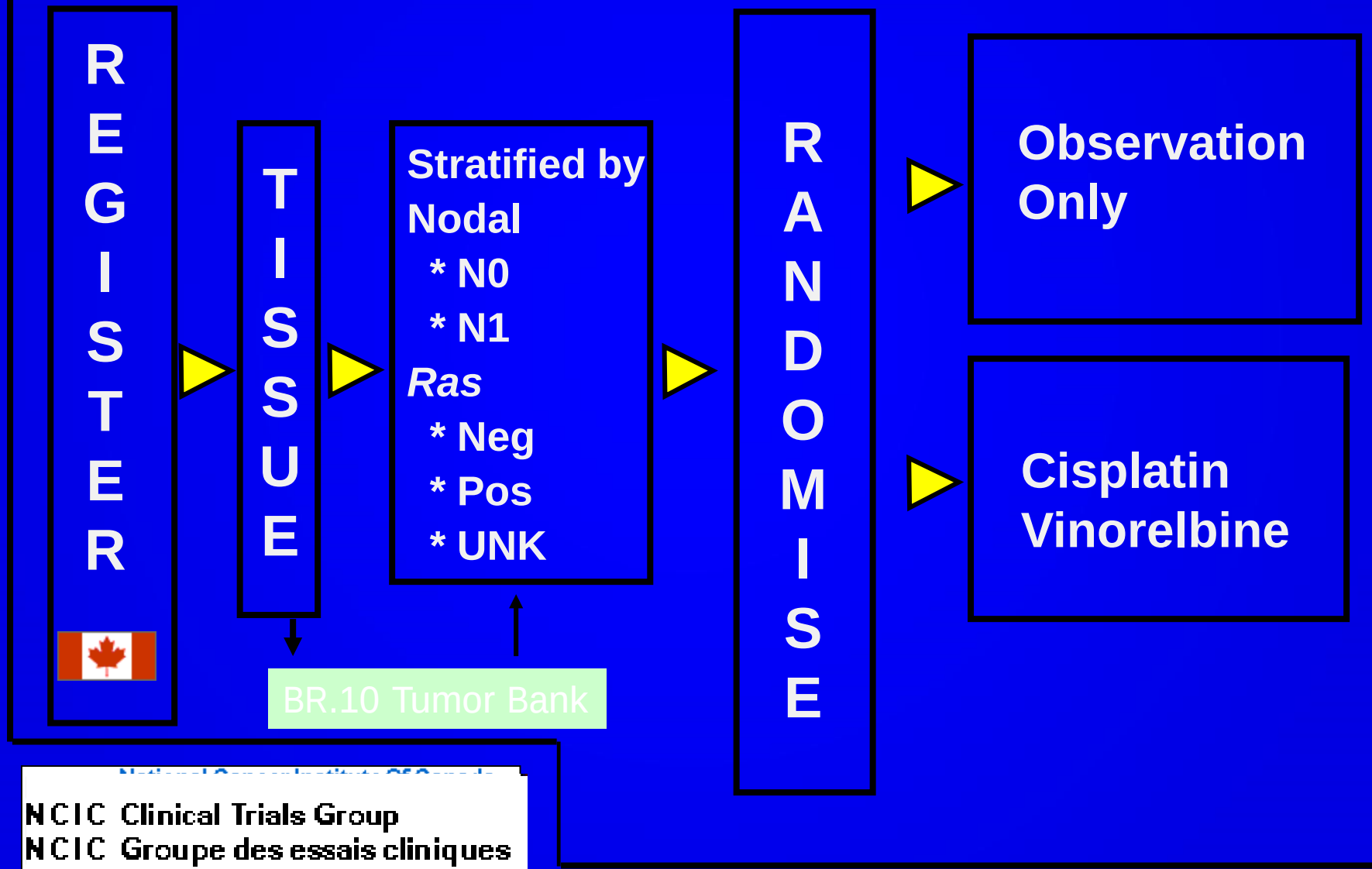
Introduction

- Tumor sample routinely collected accompanying cancer clinical trials
- Pretreatment tumor sample profiles possess the information about the disease and its sensitivity to therapy
- Affymetrix microarray: Genome-wide measurement of expression levels
- Statistical analysis can extract information to predict patients outcome and response to treatment

- Objective

- Using microarray gene expression profiling to identify a gene signature which classifies patients who benefit most from the chemotherapy in early stage resected NSCLC patients

NCIC CTG JBR.10



NCIC Clinical Trials Group
NCIC Groupe des essais cliniques

Snap-frozen Tumor Samples Available for Microarray Studies

Number of Patients	Total
In the trial HR: 0.69, 95% C.I. (0.52, 0.91), p = 0.04. (IB: HR: 0.94, II, HR: 0.59)	482 (240 obs. 242 Chemo)
Available frozen tissue with consent for future studies	169
Microarray studies completed	133

Observation = 62	Adjuvant chemo = 71

Gene microarray data

Microarrays:

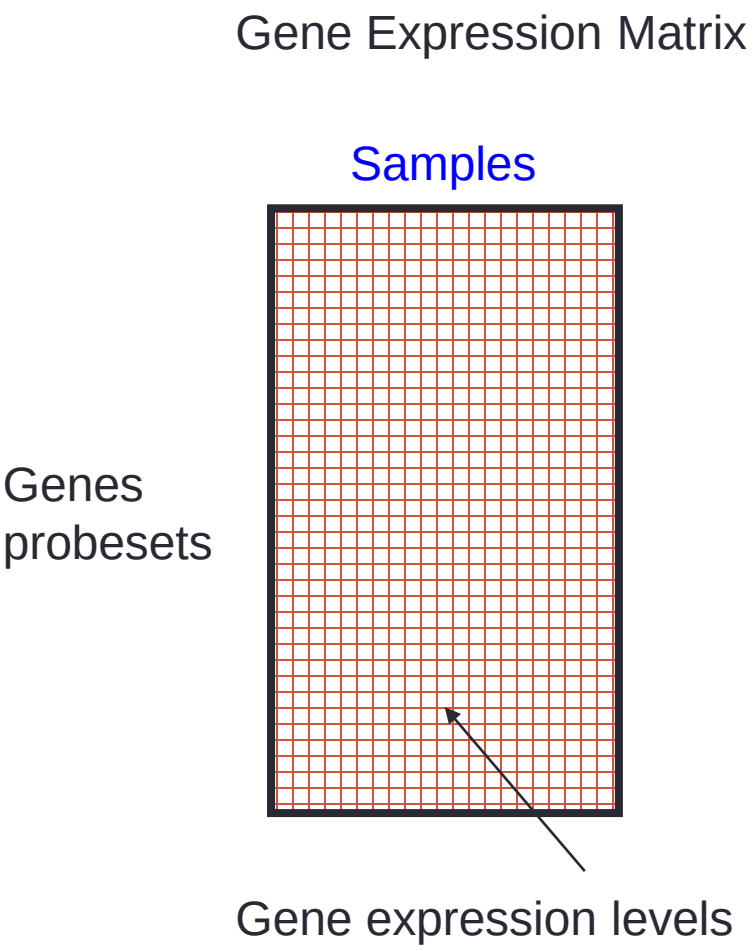
- Tools used to measure the presence and abundance of gene expression in tissue.
- microarray technologies provide a powerful tool by which the expression patterns of *thousands* of genes can be monitored simultaneously

Gene Expression:

- The degree to which a gene is active in a certain tissue of the body, measured by the amount of mRNA in the tissue.
- Gene expression depends on environment!
- Gene expression varies with time !

Gene Expression Matrices

- In a gene expression matrix, rows represent genes and columns represent measurements from different experimental conditions measured on individual arrays.
- The values at each position in the matrix characterise the expression level (absolute or relative) of a particular gene under a particular experimental condition.



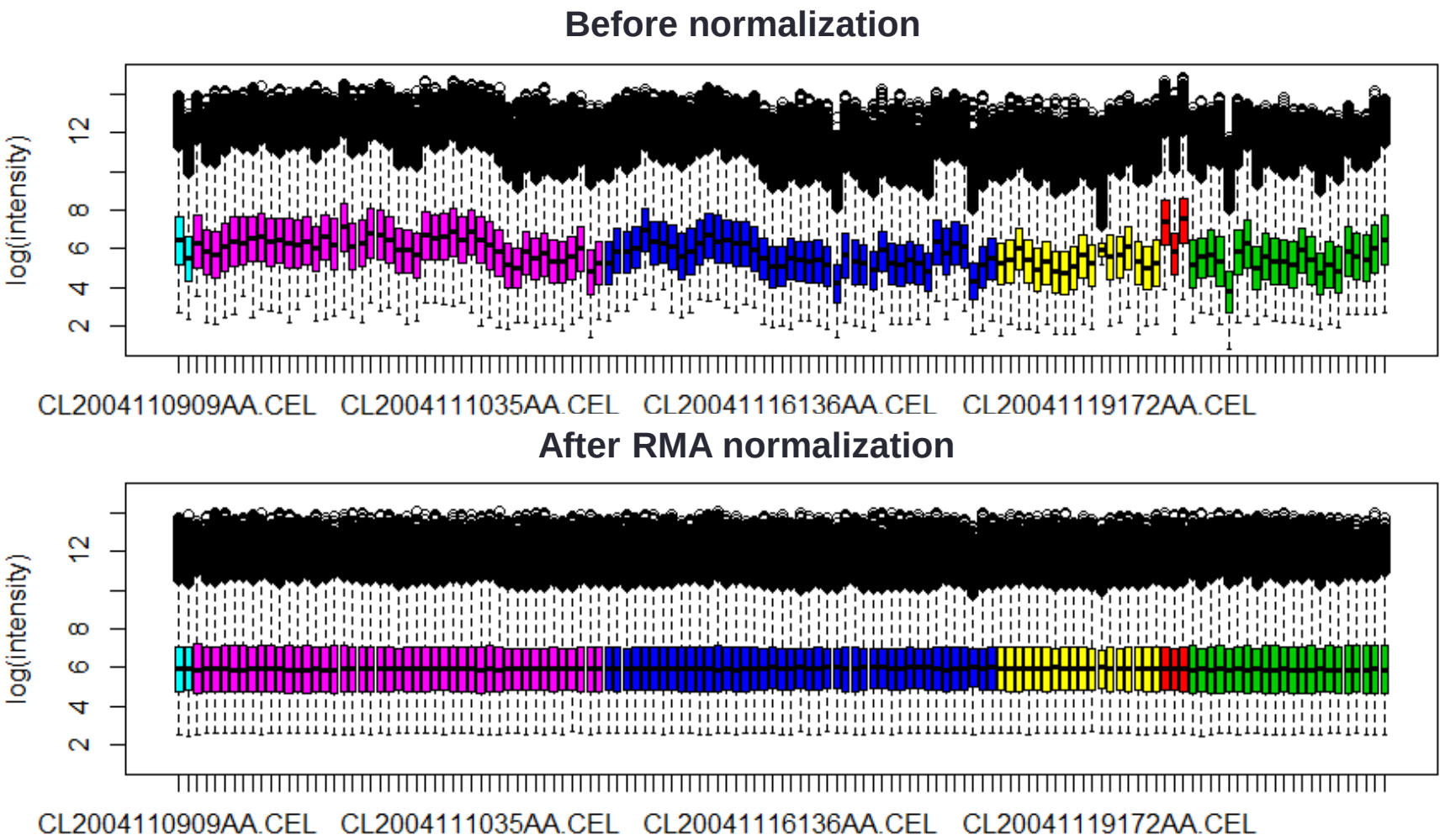
Microarray data preprocessing

- Preprocessing
  - Normalization
  - Adjusting batch effect
- Microarray samples
  - BR10. clinical trial: 133 microarray samples
  - Affymetrix U133A microarrays
  - Each array chip contains ~ 20,000 gene probesets
  - Processed from probe results file: '*.cel' file
- Analysis tools
  - BRB-Array Tool (by NCI biometric research branch)
  - R based Bioconductor genome analysis packages

Normalization

- Why?
  - Microarray data is highly noisy - intensity imbalance between RNA samples
  - Due to technical reason, not biological difference of samples
  - Purpose: adjust gene expression values of all genes so that the ones that are not really differentially expressed have similar values across the arrays
  - Normalisation is a general term for a collection of methods that are directed at reasoning about and resolving the systematic errors and bias introduced by microarray experimental platforms
- Steps
  - Background correction: remove local artifacts and noise
  - Normalization: remove array effects so the arrays are comparable
  - Summarization: combines probe intensities across arrays
- Methods: RMA, GC-RMA, MAS 5.0 … …

Normalization - single array boxplot



Batch effect

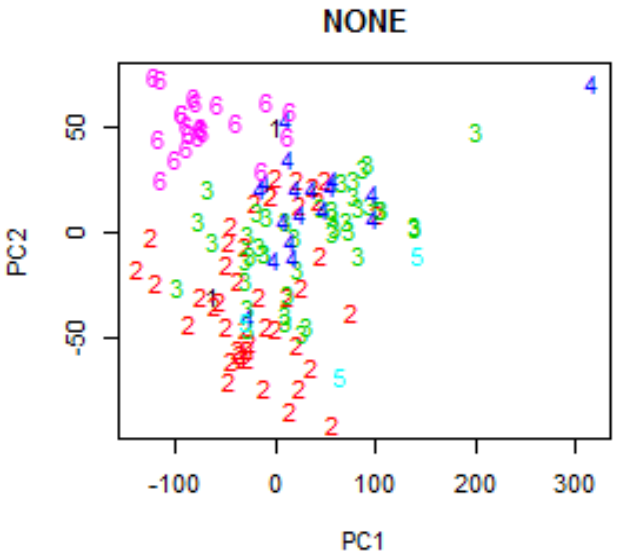
- Systematic technical differences when samples are processed and measured in different batches (e.g. processing dates)
- Unrelated to any biological variation, recorded during experiment

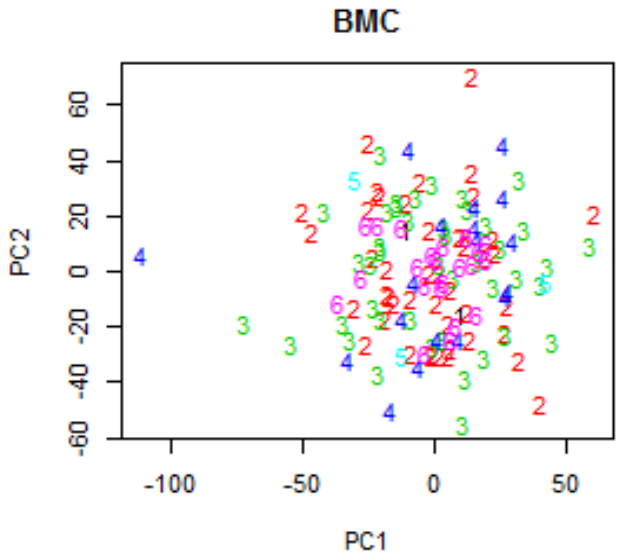
- Methods (Location-scale)
  - Apply models to adjust the gene probesets to have similar mean and variance in each batch
  - BMC, COMBAT, GENENORM, DWD … …

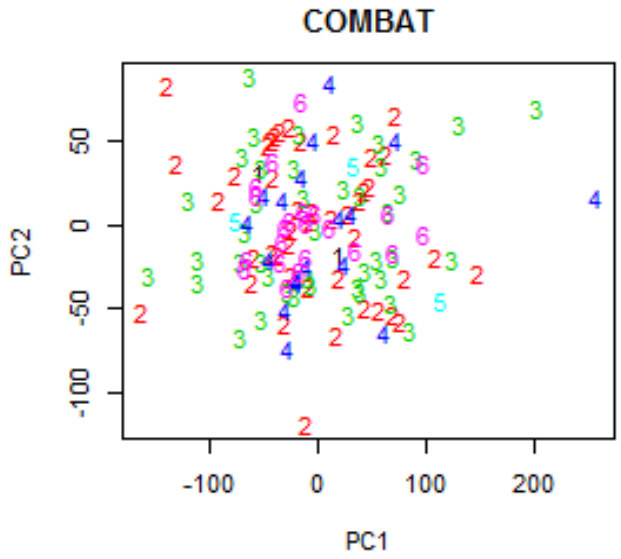
- Total 133 samples and 6 batches

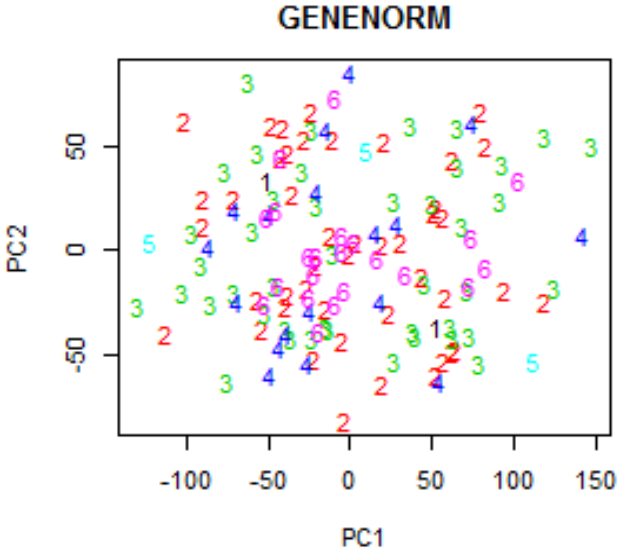
Batch ID	1	2	3	4	5	6
Batch name	1109	1110	1116	1119	1130	0603
number of arrays	2	45	43	18	3	22

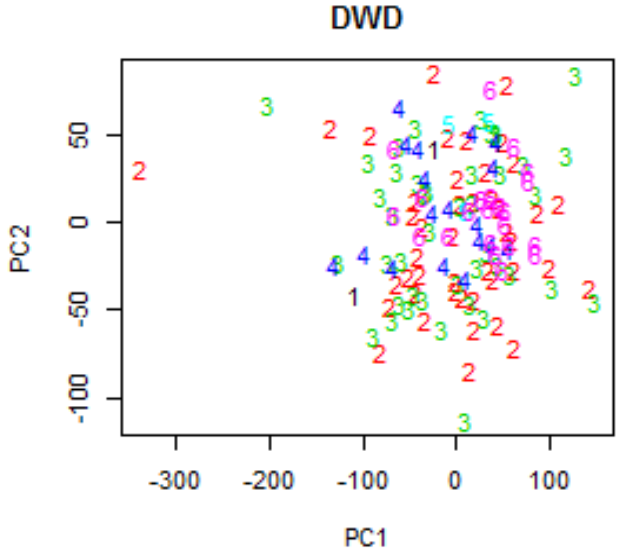
Batch effect – principal component plots











Predictive gene signature selection

- Purpose: Selection a group of genes that classify patients who are most benefit from the received treatment
- Main issues
  - High dimensional covariates (p >> n) ---- variable selection
  - Treatment – covariates interaction

    presence of main effects:
    - Increase the difficulty to detect treatment – covariates interaction
    - Increase the number of covariates

Predictive gene signature selection

- Informative gene selection

-- Non-informative filtering: exclude probesets that ave low variance, and low intensity (expression levels)

-- Informative filtering: Uni-probeset, study treatment, and their interaction term included, keep probesets with predictive potential, with small p-value for the interaction term

- Multi-genes that are predictive of treatment effect: Rank probesets based on the predictive p-value (p-value of the interaction term) in uni-probeset analysis.
- Multi-genes signature selection: modified covariates without main effects (Tian et al, JASA accepted March, 2014).

Tian L, Alizadeh A, Gentles J, Tibshiran R. A Simple method for detecting interactions between a treatment and a large number of covariates. arXiv:1212.2995 [stat.ME]. Dec 2012

Modified covariates method

- Modified covariate: $W(Z)^* = W(Z) \cdot \frac{T}{2}$

  - Z : covariates
  - *T:* treatment

  $W(Z)$: standardized Z

  T = 1 chemotherapy
  T = -1 observation

- Cox regression model using modified covariate

$$h(t|Z,T) = h_0(t)e^{\gamma.W(Z)^*}$$

- $\hat{\gamma}.W(z)^*$ can be used to stratify patients for individualized treatment selection

Variable selection

- Least square model
  - High variance, poor prediction, especially *p* is large
  - instable, not suitable for *p* >> *n* cases
- L_1 penalized model – Lasso (Tibshirani, 1996)
  - Bias-variance trade off to improve prediction accuracy
  - Provides sparse solutions: useful for variable selection in n << p case.
  - Limitation
    - Selects at most *n* variables before it saturates
    - For a group of highly correlated variables, only select one variable from a group and ignore others
- L_2 penalized model – Ridge regression
  - – Removes the limitation on the number of selected variables;
  - – Encourages grouping effect; select correlated variables
  - – Stabilizes the L_1 regularization path.

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 1996; 58: 267–288

Variable selection

- Elastic net (Zou, 2005)

L_2 penalty L_1 penalty

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

  - L_1 penalty: generates a sparse model for variable selection
  - L_2 penalty:
    - remove the limitation on number of selected variables
    - encourage group selection, and stabilized L_1
  - Tuning parameters: (λ_2, α) where $\alpha = \frac{\lambda_1}{\lambda_1+\lambda_2}$, $\alpha \in [0,1]$
    - (λ_2, α) : tuned by in a grid search with min cross validation error rule
    - α: (α = 0.1. was chosen).

Zou, Hui; Hastie, Trevor (2005). Regularization and variable selection via the elastic net. *J. Royal. Stat. Soci, Series B*: 301–320

Gene signature selection procedure

- Microarray preprocessing
  - RMA normalization / DWD adjusting batch effect
- Divide samples into training & test sets
  - Have similar survival experience (stratified by disease stage & histology)
  - Training set is used to select predictive gene signature
- Gene probesets pre-selection
  - Non-informative filtering: Filtered out 1/3 gene probesets with low variance across samples, and mean intensity < 4.
  - Informative filtering: Fit Cox's model with modified covariate without main effect
    - Pre-select gene probesets with absolute estimate of interaction effect no less than 0.4. (662 gene probesets remain)

Gene signature selection

- Predictive gene signature selection
  - Fit multivariable Cox's model with modified covariates based on preselected gene probesets
  - Elastic net for variable selection
  - Bootstrap samples and fit above model 1000 times, and rank probe according the frequency they appeared in the model
  - PCA to synthesize information of the most often selected probesets (k from 1 to 150).

Gene signature selection

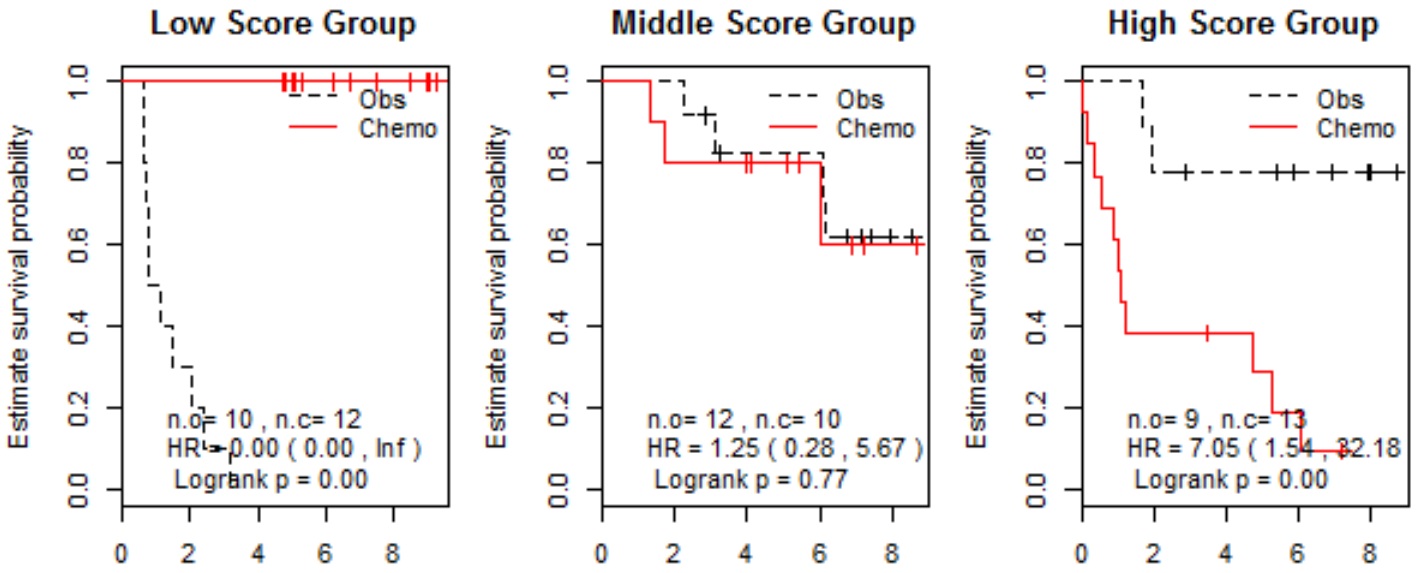
- 10 folds cross-validation
- Fit Cox's model with treatment, PC1 and their interaction terms, and generate cross validation predictive scores: B1+B3*PC1
  - B1: coefficient of treatment estimate
  - B3: coefficient of treatment and PC1 interaction estimate
- Classify patients into low, middle and high groups using CV predictive score
- Predictive gene signature: a group a gene probsets that best separate low score group of patients by treatment arms (min p-value)
- 34-gene probesets were selected.

Predict treatment effect

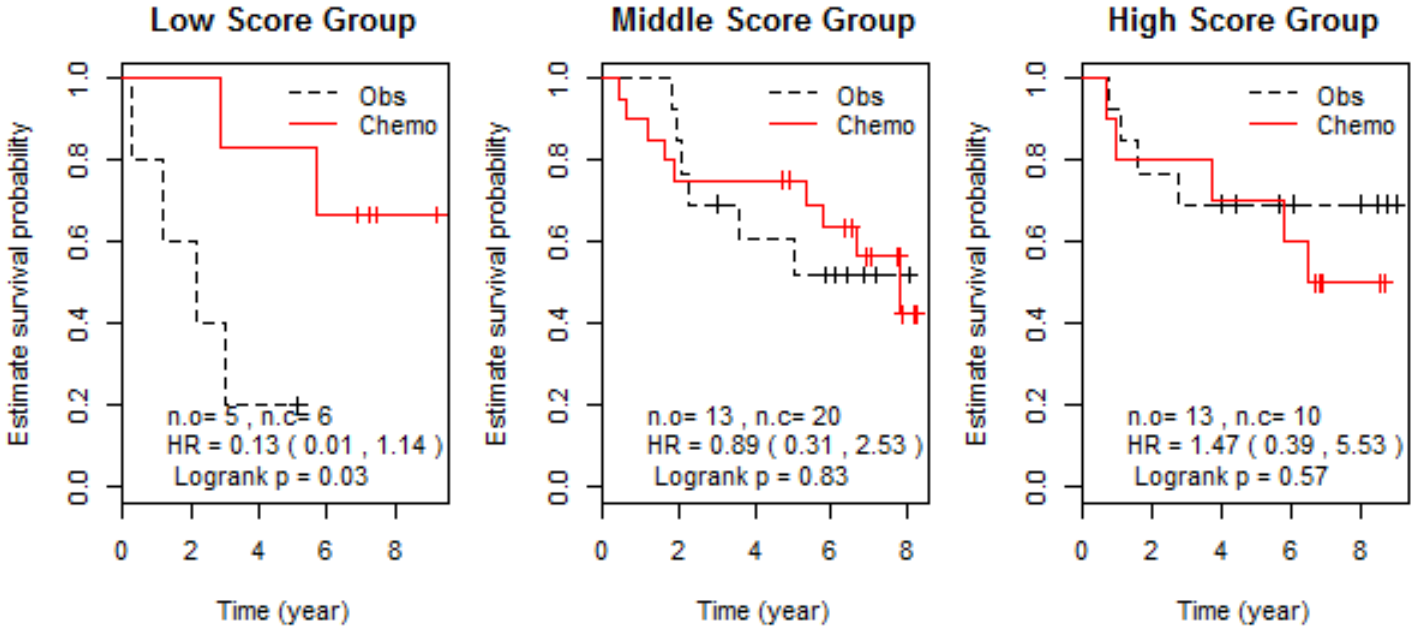
Validate the signature in the testing set

- Generate predictive scores of patients in training set based on selected gene signature using (B3*PC1)
- Classify patients into low, middle and high predictive score groups using 1/3 and 2/3 quantiles of predictive scores as cut-off points
- Generate predictive scores of patients in test data set based on the information in training set:
  - Coefficient of loading matrix of PC1
  - Estimate coefficient of the interaction term of treatment and PC1
- Classify test set patients into low, middle and high predictive score groups using the cut-off points in the training set
- Low predictive score group benefits from chemo therapy

Training set



Testing set



Overall survival of 133 patients in predictive score groups based on 34-gene signature

Predictive score = 0.816 X PC1

Cut-off points:

1/3 quantile: -0.734

2/3 quantile: 0.810

Loading matrix of training dataset

Probeset	PC1 loading coef.	Probeset	PC1 loading coef.
Probeset_1	0.135	Probeset_18	-0.066
Probeset_2	0.153	Probeset_19	-0.083
Probeset_3	0.236	Probeset_20	0.197
Probeset_4	-0.185	Probeset_21	0.262
Probeset_5	-0.080	Probeset_22	-0.169
Probeset_6	0.120	Probeset_23	0.185
Probeset_7	-0.071	Probeset_24	0.206
Probeset_8	-0.199	Probeset_25	0.254
Probeset_9	-0.145	Probeset_26	0.132
Probeset_10	-0.091	Probeset_27	-0.034
Probeset_11	-0.075	Probeset_28	-0.131
Probeset_12	0.235	Probeset_29	-0.072
Probeset_13	0.148	Probeset_30	0.159
Probeset_14	0.108	Probeset_31	-0.208
Probeset_15	0.171	Probeset_32	0.264
Probeset_16	0.250	Probeset_33	-0.212
Probeset_17	-0.215	Probeset_34	0.170

Summary

- Microarray raw data of 133 BR10. samples were preprocessed by normalization and adjusting batch effect.
- Predictive gene probesets were selected using Cox's model fitted by modified covariates of bootstrap samples without main effect, and elastic net for variable selection.
- A 34-gene signature separates patients in low predictive score group between two treatment arms, and the patients in low score group are benefit to chemotherapy.

Acknowledge:

This is the joint work with Ms. Li Liu