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

1,2Keyue Ding, Ph.D.

Nov. 8, 2014

1NCIC Clinical Trials Group, Kingston, Ontario, Canada
2Dept. 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 Clinical Trials Group
NCIC Groupe des essais cliniques

NCIC CTG JBR.10

REGISTER

TISSUE

Stratified by Nodal
* N0
* N1
Ras
* Neg
* Pos
* UNK

RANDOMISE

Observation Only

Cisplatin Vinorelbine

BR.10 Tumor Bank
<table>
<thead>
<tr>
<th>Number of Patients</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>In the trial</td>
<td>482</td>
</tr>
<tr>
<td>HR: 0.69, 95% C.I. (0.52, 0.91), p = 0.04. (IB: HR: 0.94, II, HR: 0.59)</td>
<td>(240 obs. 242 Chemo)</td>
</tr>
<tr>
<td>Available frozen tissue with consent for future studies</td>
<td>169</td>
</tr>
<tr>
<td>Microarray studies completed</td>
<td>133</td>
</tr>
<tr>
<td>Observation = 62</td>
<td></td>
</tr>
<tr>
<td>Adjuvant chemo = 71</td>
<td></td>
</tr>
</tbody>
</table>

Snap-frozen Tumor Samples Available for Microarray Studies
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

Before normalization

After RMA normalization
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

<table>
<thead>
<tr>
<th>Batch ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch name</td>
<td>1109</td>
<td>1110</td>
<td>1116</td>
<td>1119</td>
<td>1130</td>
<td>0603</td>
</tr>
<tr>
<td>number of arrays</td>
<td>2</td>
<td>45</td>
<td>43</td>
<td>18</td>
<td>3</td>
<td>22</td>
</tr>
</tbody>
</table>
Batch effect – principal component plots

![PC plots for different batch correction methods](images)
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 have 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).

Modified covariates method

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

  $Z$: covariates
  $W(Z)$: standardized $Z$
  $T$: treatment

  $T = 1$ chemotherapy
  $T = -1$ observation

• Cox regression model using modified covariate

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

• $\hat{\gamma} \cdot W(z)^*$ 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.

---

Variable selection

• Elastic net (Zou, 2005) 

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

- L1 penalty: generates a sparse model for variable selection
- L2 penalty:
  • remove the limitation on number of selected variables
  • encourage group selection, and stabilized L1

- Tuning parameters: \((\lambda_2, \alpha)\) where \(\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2}, \quad \alpha \in [0,1]\)
  - \((\lambda_2, \alpha)\) : tuned by in a grid search with min cross validation error rule
  - \(\alpha\): (\(\alpha = 0.1\). was chosen).

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 probesets 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 chemotherapy
Training set

Testing set

Low Score Group

Middle Score Group

High Score Group

Low Score Group

Middle Score Group

High Score Group
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
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