Lessons in Tumour Classification using Gene Expression Microarrays

Shelley Bull

Samuel Lunenfeld Research Institute

University of Toronto

Co-authors: Sarah Colby,

Wenqing He, Pingzhao Hu, Xiang Sun

The First Canadian Workshop on Statistical Genomics,

National Program in Complex Data Structures

Fields Institute, 4 September 2003

Outline

- Background and Study Design
- Primary Analyses of Human Tumour Studies
  - Methods, Results
- Further Comparisons of Methods
- Lessons Learned and On-going Work

- Background and Study Design
- Primary Analyses of Human Tumour Studies
  - Methods, Results
- Further Comparisons of Methods
- Lessons Learned and On-going Work

Node-negative Breast Cancer

Molecular factors for disease prognosis and targetted therapy

Cohort of 1500 (1987-1998), 870 specimens

I. Confirmatory studies

Sequential evaluation of molecular genetic factors: *her/neu-erbB2* amplification, *p53*, *p27*

II. Exploratory studies

Application of microarray technology to identify patterns of expression predictive of disease course

Musculo-Skeletal Neoplasia

Molecular factors for metastatic disease and targetted therapy

Canadian Sarcoma Group and pediatric Hospital for Sick Children tumour banks

Multiple tumour types - MFH sarcoma, osteosarcoma, ...

I. Confirmatory

Multidrug resistance (*MDR1*), *p53*

II. Exploratory

Molecular classification

Design Features

Observational Studies: Human tissue specimens

Aims: Class comparison – differential expression

Class prediction – classifying new samples

Gene co-expression – novel gene classes

Levels of Replication:

Biological * Generalize to a population, Increase precision

Between Individuals within a Group

Technical * Reduce measurement error, systematic effects

Repeated Arrays within Individuals - dye-swaps

Repeated Measures Within Arrays - duplicate spots

Microarray Features

Microarrays: 19k cDNA spot arrays

- Duplicate spots, side by side
- Tumour and control samples cy3/cy5 dye labelled
- Common reference control – mixture of cancer cell lines
- Indirect design: $\log(T_1/C) - \log(T_2/C)$

Pre-Processing:

- Local background subtraction
- Log base 2 ratio of tumour to control
- Subarray based median location adjustment and IQR scaling
- Imputation of missing data

- Background and Study Design
- Primary Analyses of Human Tumour Studies
  - Methods
- Further Comparisons of Methods
- Lessons and On-going Work

Each array: quality, intensity, normalization - diagnostic plots

Select array sets

Analytic Strategy

Global expression

Single gene differential expression

Multiple gene tumour classification

Dimension reduction Gene clustering, etc.

Validation

Approaches to Ensure Validity:

- Internal
  - Replication, Reproducibility studies, Diagnostic plots
  - Statistical
    - Permutation for multiple testing
    - Cross-validation/bootstrap for prediction
    - Bootstrap for cluster reliability assessment
    - Assess validity & power by statistical experimentation
  - Molecular
    - Confirmation by PCR, immunohistochemistry
- External
  - Confirmation in independent samples

Supervised Analysis - Comparison of two groups of tumours

- Single gene differential expression
  - permutation p-values, false discovery rates
- Multiple gene tumour classification
  - “honest” tumour class prediction using CV
- Clustering of selected genes
- Refinement of multigene classification
  - dimension reduction

Classification Methods

- Linear discriminant analysis (LDA)
  - Covariance matrix for genes has too many parameters
- Diagonal linear discriminant (DLDA)
  - Assumes that genes are uncorrelated
- Compound covariate predictor (CCP)
  - Weighted linear combination of mean differences
- Nearest centroid (NC)
- Nearest neighbour (1-NN, 3-NN)
- Support vector machine (SVM)

Dudoit et al (2002), JASA

Cross-validation Methods

- Prediction accuracy of class membership of tumours used to develop a classifier will be overoptimistic, ie. Misclassification error will be too small compared to an independent sample
- CV - Modification of the idea of having a training sample to construct a classifier and an independent test sample to assess it
- Algorithm: Divide dataset into *k* disjoint subsets
  - In training set of *k-1* subsets: *select genes*, build classifier
  - In k^{th} test set: apply classifier, compare to known class
  - Repeat for each of *k* subsets
  - Estimate overall classification accuracy/error in test sets

- Background and Study Design
- Primary Analyses of Human Tumour Studies
  - Node-negative Breast Cancer
- Further Comparisons of Methods
- Lessons and On-going Work

ANN Breast Cancer - Global Gene Expression

- Dataset Assembly
  - 103 patient tumours, 30 with = 2 replicates
  - total of 143 arrays
  - 12,851 genes
  - clinical data: patient outcome, pathological and molecular tumour characteristics
- Purpose of the Analysis
  - Differential expression between tumour groups
  - “Short” list of genes for molecular validation
  - Classification accuracy for statistical validation

Supervised Analysis - Two Group Comparisons:

With versus without lymphatic invasion (n=37/66)

- Single gene differential expression
  - by multiple t-tests (BRB Tools, SAM)
- Number of genes selected (BRB Tools)
  - $p < 0.01$ 4,774
  - $p < 10^{-5}$ 1,146
  - $p < 10^{-6}$ 615
  - $p < 10^{-8}$ 139
- False discovery rate (SAM)
  - median FDR is 3/2,576 (0.11%) and the 90th percentile is 12/2,576 (0.45%)

Clustering of
139 genes and
103 tumours

Tumours

Genes

Supervised Analysis - Two Group Comparisons:

With versus without lymphatic invasion (n=37/66)

- Multiple gene tumour classification
  - 139 genes selected with $p < 10^{-8}$
- *Apparent* accuracy of < 90%
- Cross-validation: leave-one-out *with* selection
- Methods (BRB Tools)
  - compound covariate predictor 81%
  - diagonal linear discriminant analysis 82%
  - 3-nearest neighbour, nearest centroid 85%, 79%
  - support vector machine 80%

- Background and Study Design
- Primary Analyses of Two Human Tumour Studies
  - MFH Soft-tissue Sarcoma
- Further Comparisons of Methods
- Lessons and On-going Work

Sarcoma - Global Gene Expression

- Dataset Assembly
  - 47 MFH patients (malignant fibrous histiocytoma)
  - 45 of 47 tumours with 2 dye-swap replicates
  - total of 92 arrays
  - 19,200 genes
  - clinical characteristics: presence of metastases, stage (size, depth, grade)
- Purpose of the analysis
  - Differential expression between tumour groups
  - “Short” list of genes for molecular validation
  - Classification accuracy for outcome prediction

Supervised Analysis - Two Group Comparisons

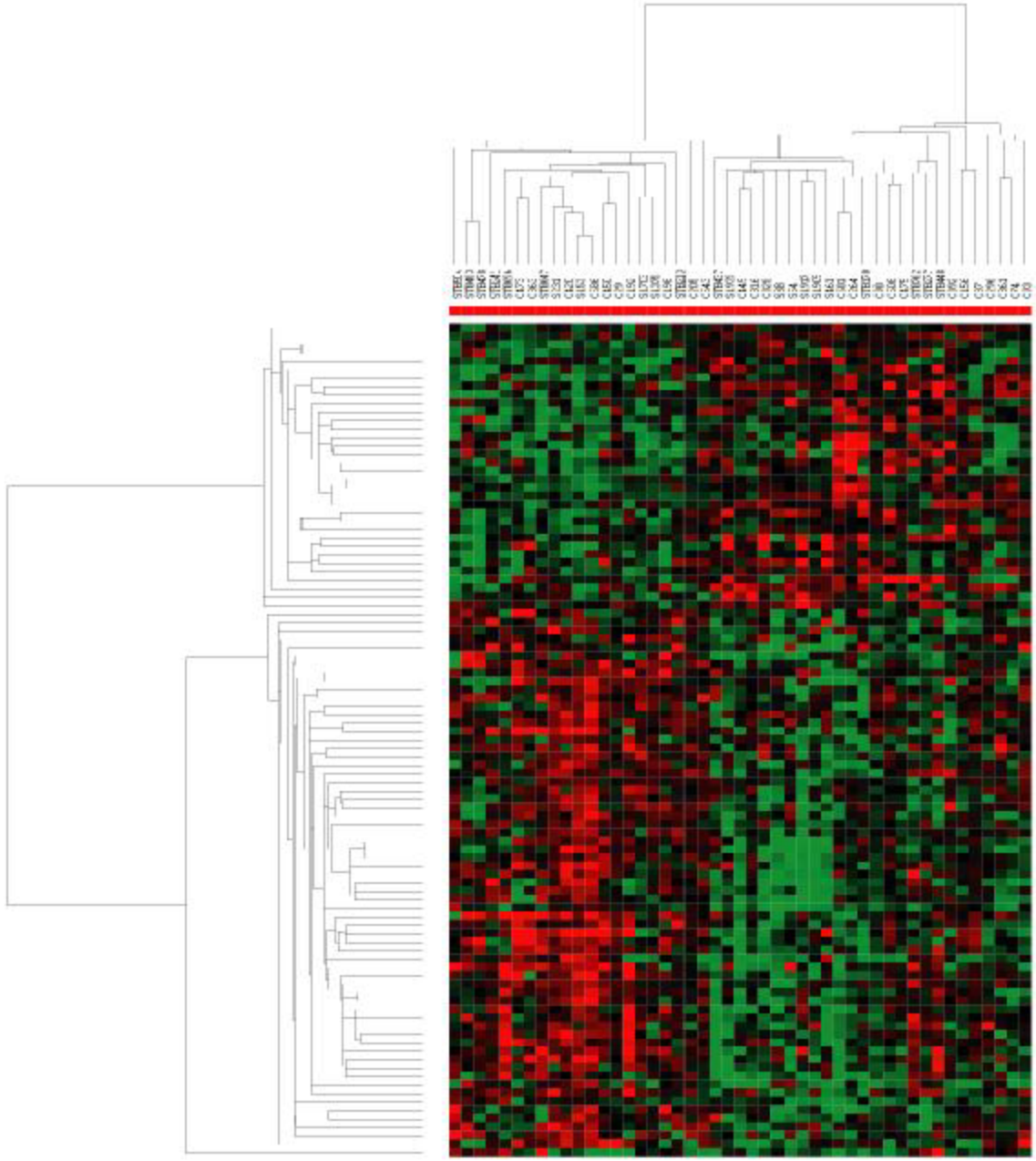
With versus without metastasis (n=24/23)

- Single gene discrimination
  - by multiple t-tests (BRB Tools, SAM)
- Number of genes selected (BRB Tools)
  - $p < 0.01$ 196
  - $p < 0.005$ 99
  - $p < 0.001$ 18
  - $p < 5 * 10^{-4}$ 6
- False discovery rate (SAM)
  - median false discovery rate is 102/274 (37%) and the 90th percentile is 135/274 (49%)

Clustering for 99 genes and 47 tumors

Tumours

Genes



Supervised Analysis - Two Group Comparisons:

With versus without metastasis (n=24/23)

- Multiple gene tumour classification
  - 99 genes selected with $p < 0.005$
  - *Apparent* accuracy of 98%
- Cross-validation: leave-one-out *with* selection
- Methods (BRB Tools)
  - compound covariate predictor 68%
  - diagonal linear discriminant analysis 68%
  - 3-nearest neighbour, nearest centroid 58%, 64%
  - support vector machine 68%

- Background and Study Design
- Primary Analyses of Human Tumour Studies
  - Methods, Results to date
- Further Comparisons of Methods
- Lessons and On-going Work

Issues re Tumour Classification Methods

- Bias vs variance trade-off in leave-one-out CV versus 10-fold CV or .632+ bootstrap
- Information in the discriminant score
  - ie. prob of group membership
  - use of ROC curve (sensitivity, specificity)
- Gene selection is *univariate*, not multivariate
  - how many genes needed for accurate classification
  - can correlation among genes be used to improve classification accuracy or reduce variability

Following Ambroise and McLachlan (2002), PNAS

Fig. 1a

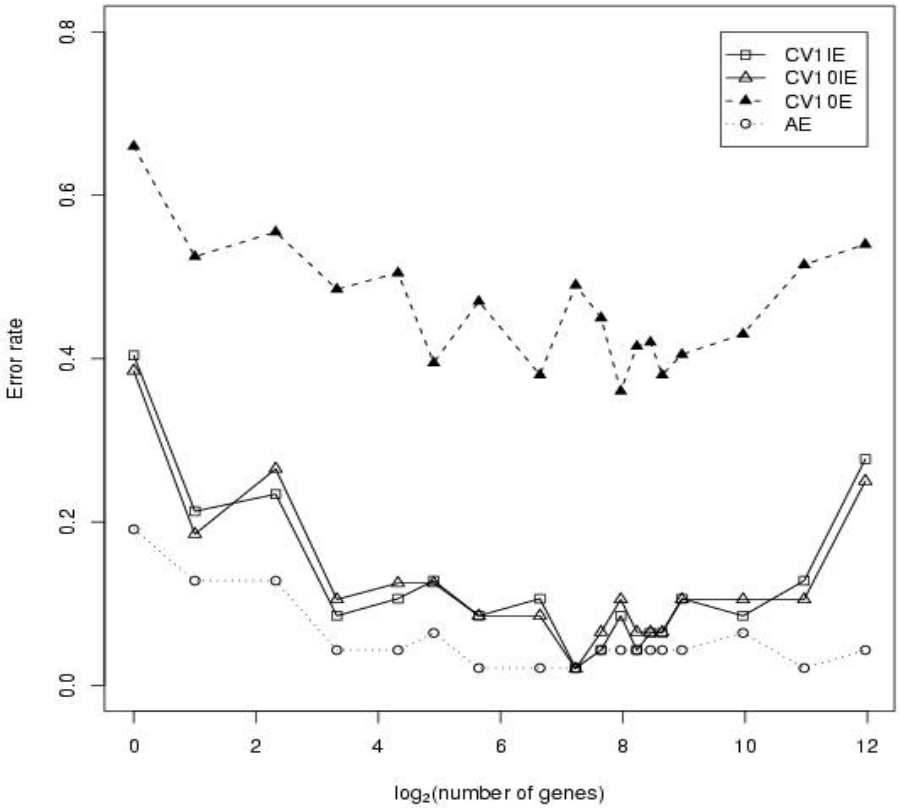


Fig. 1b

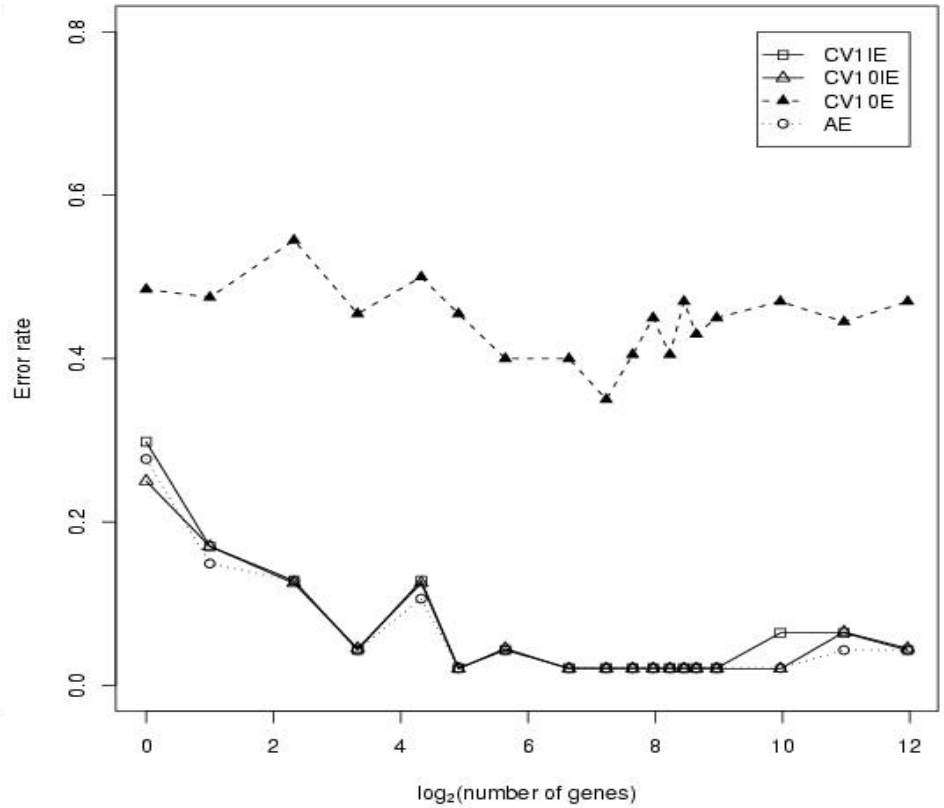


Fig. 1a & 1b | Error rates of the 3 nearest-neighbour predictor (Fig. 1a) and the compound covariate predictor (Fig. 1b) based on the sarcoma dataset.

Fig. 2a

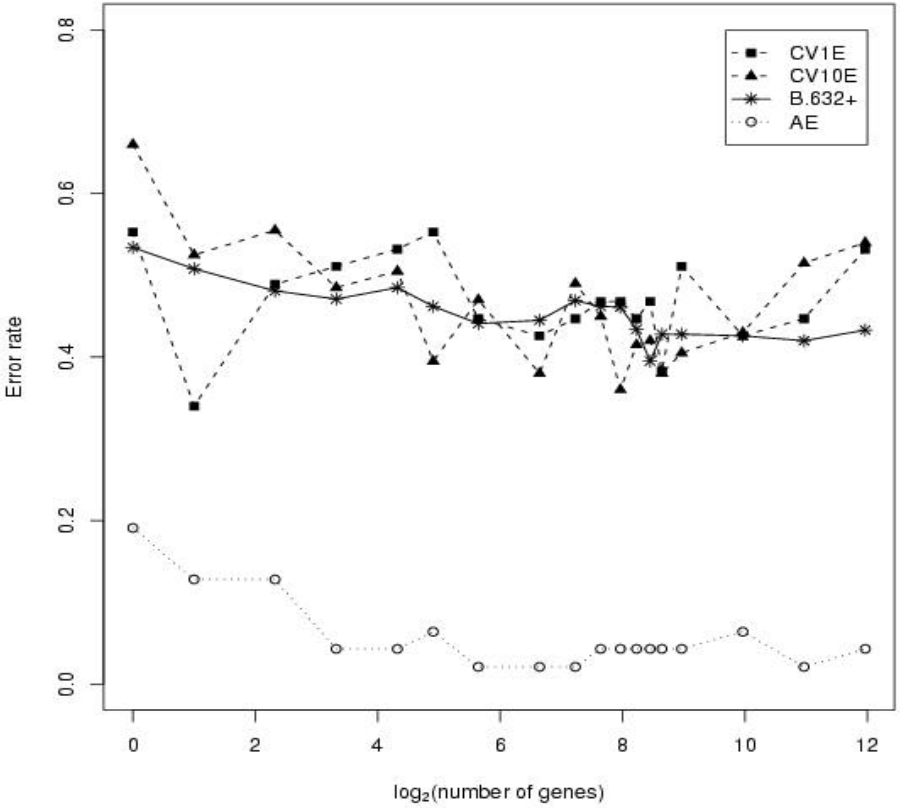


Fig. 2b

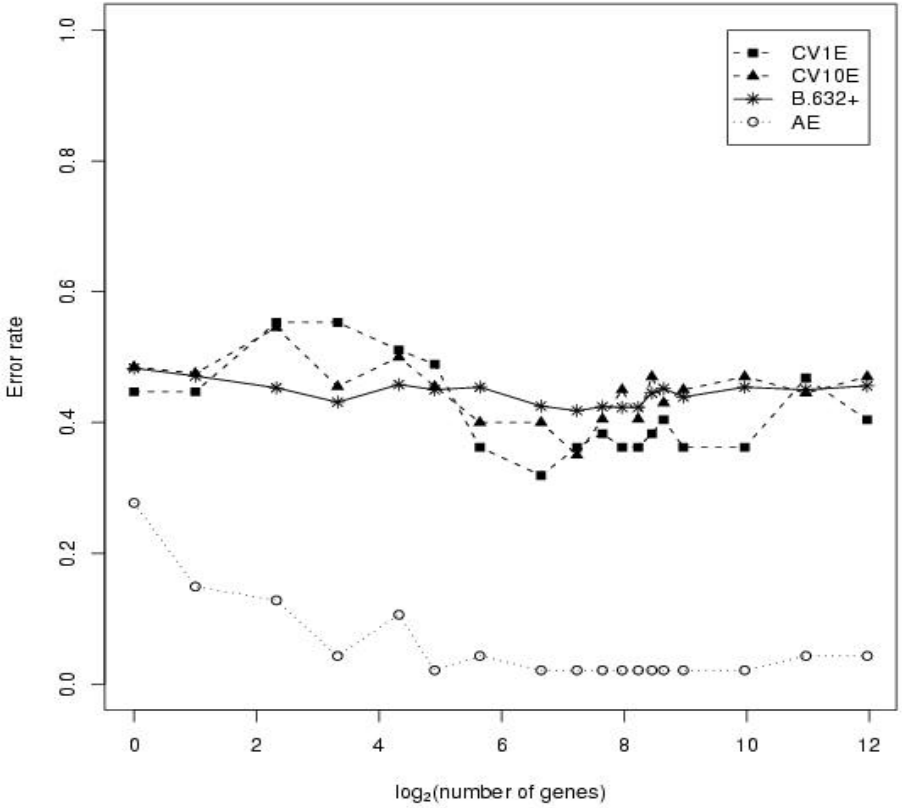


Fig. 2a & 2b | Error rates of the 3 nearest-neighbour predictor (Fig. 2a) and the compound covariate predictor (Fig. 2b) based on the sarcoma dataset.

Fig. 3a

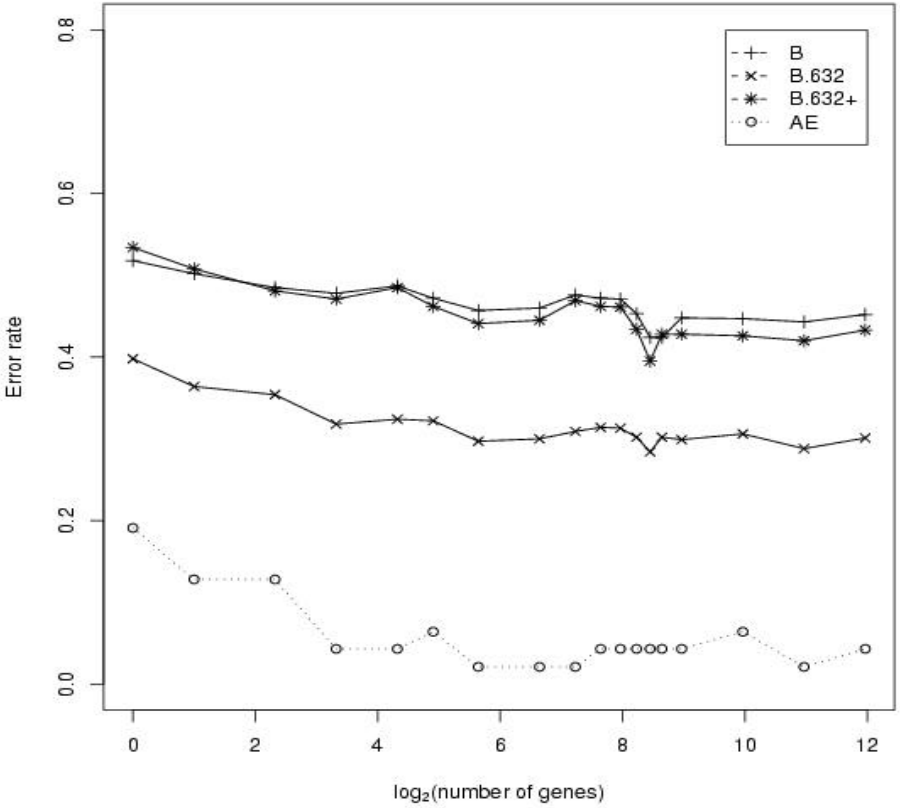


Fig. 3b

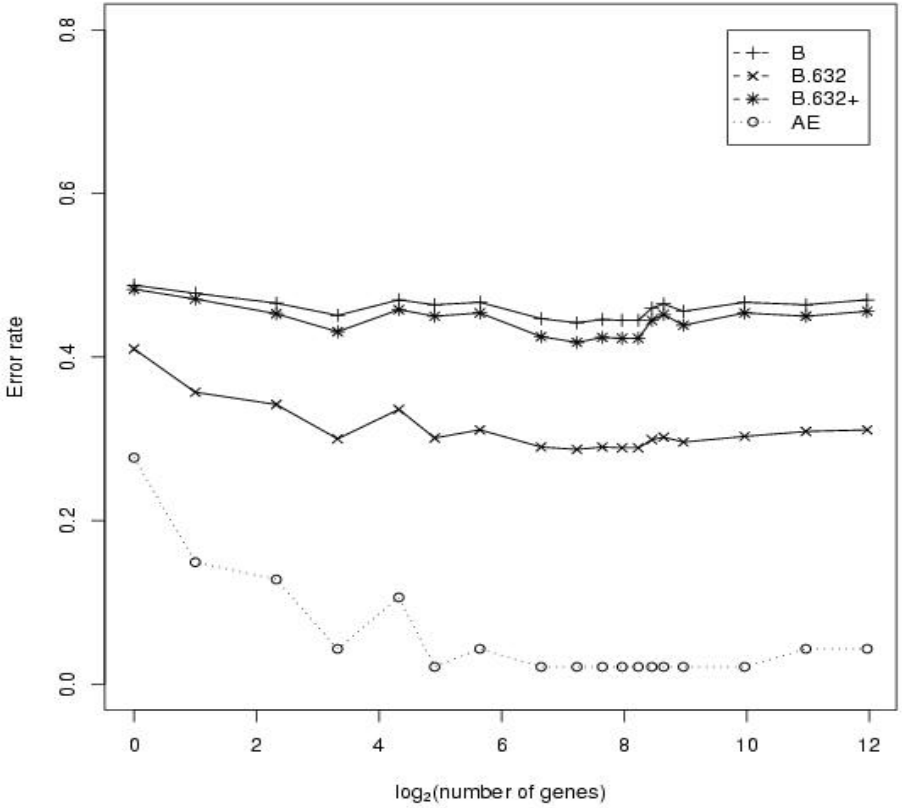


Fig. 3a & 3b | Error rates of the 3 nearest-neighbour predictor (Fig. 3a) and the compound covariate predictor (Fig. 3b) based on the sarcoma dataset.

Issues re Tumour Classification Methods

- Bias vs variance trade-off in leave-one-out CV versus 10-fold CV or .632+ bootstrap
- Information in the discriminant score
  - ie. prob of group membership
  - use of ROC curve (sensitivity, specificity)
- Gene selection is *univariate*, not multivariate
  - how many genes needed for accurate classification
  - can correlation among genes be used to improve classification accuracy or reduce variability

Alternative Classification Method

- Block diagonal linear discriminant analysis (BLDA)
  - Assumes an exchangeable correlation structure within gene clusters, zero correlation between clusters
  - Use of SVD for matrix inversion shows that this serves as a form of within cluster averaging
- Two-step algorithm:

(1) select genes one-at-a-time using univariate methods and statistical criteria

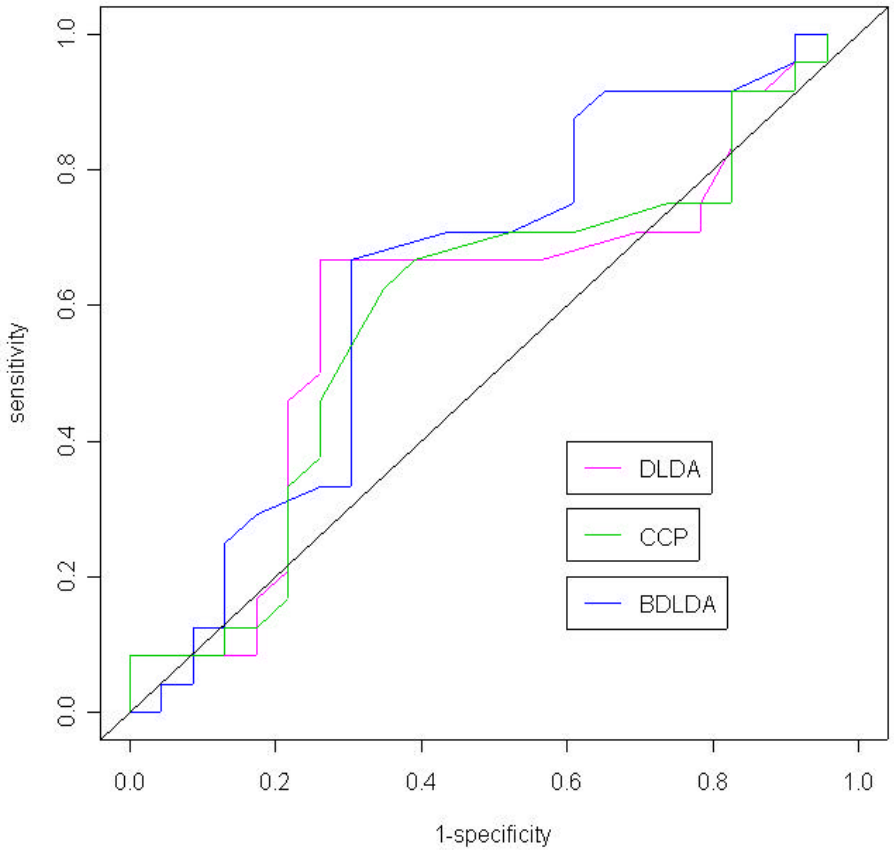
(2) option 1: cluster selected genes

option 2: treat selected genes as the “seeds” of a cluster, include additional genes that are highly correlated with the selected gene

ROC curves - 10-fold CV

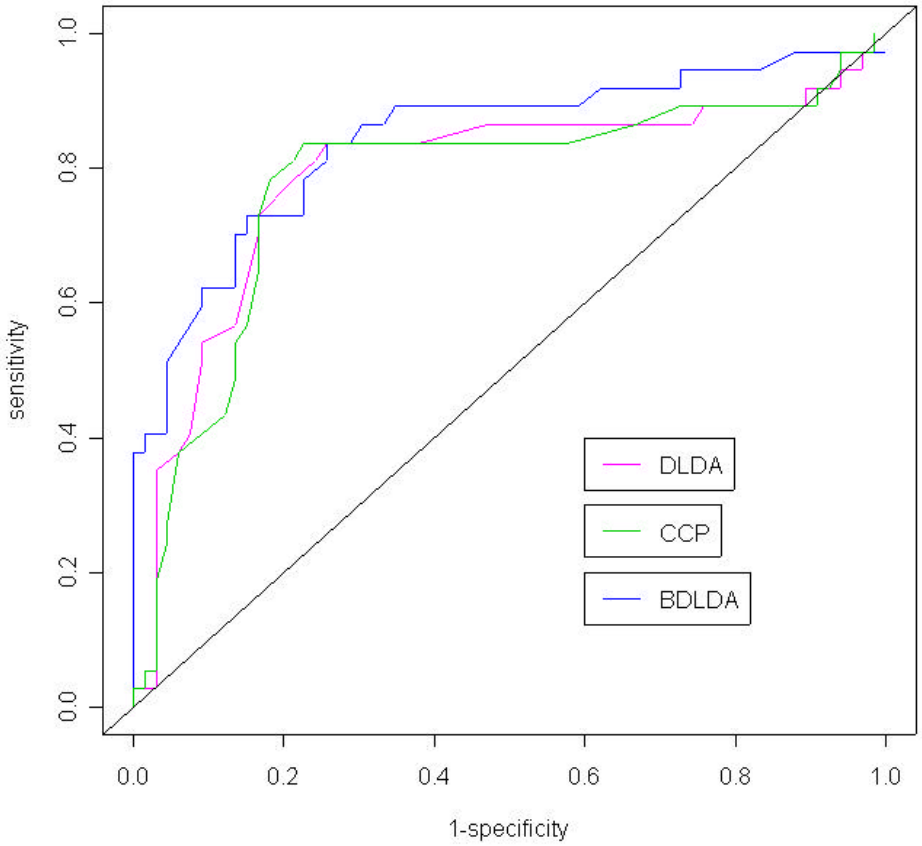
Sarcoma

	AUC
DLDA	61.3
CCP	60.2
BLDA	66.2



Breast Cancer

	AUC
DLDA	79.4
CCP	78.2
BLDA	84.6



- Background and Study Design
- Primary Analyses of Two Human Tumour Studies
  - Methods, Results to date
- Further Comparisons of Methods
- Lessons Learned and On-going Work

Lessons Learned

- Computational and data handling issues should not be underestimated
- Existing microarray specific tools (BRB, SAM, R) are a great asset in getting started
- Overfitting, even with simple methods, needs to be properly addressed, especially with small sample sizes
- Different methods tend to misclassify the same observations in leave-one-out CV
- Leave-one-out and 10-fold CV more variable
- Some prediction problems are more difficult - patient outcomes vs tumour characteristics, heterogeneous disease

On-going Work

- Cross-validation techniques
  - characterization of tumours that are “difficult” to classify, use of covariate data
- Use of gene clustering in classification
- Criteria to assess normalization methods and filter genes - sensitivity analyses
- Comparison of multi-gene classification and clustering methods
  - Construction of artificial datasets for statistical experiments, based on own data

Acknowledgements

- Collaborators

  Irene Andrulis Jay Wunder
  Jim Woodgett Nalan Gokgoz
  Lucine Collins Sasha Eskandarian

- Software

  BRB Array Tools - R Simon, NCI
  SAM - R Tibshirani, Stanford
  R/Bioconductor

- Funding

  National Cancer Institute of Canada,
  Terry Fox Program Project
  CIHR - IHRT in Musculo-Skeletal Neoplasia
  NCE - MITACS