A method for finding molecular signatures from gene expression data

Ramón Díaz-Uriarte
`rdiaz@cnio.es`

`http://bioinfo.cnio.es/˜rdiaz`

Unidad de Bioinformática

Centro Nacional de Investigaciones Oncológicas (CNIO)

(Spanish National Cancer Center)

Workshop on Statistical Analysis of Gene Expression Data, Wye, UK, July 2003

Introduction

- *Molecular signatures* or *Gene expression signatures* are a key feature in many papers in cancer research. For instance, Alizadeh et al., 2000; Golub et al., 1999; Huang et al., 2002; Pomeroy et al. 2002; Ramaswamy et al., 2003; Rosenwald et al., 2002; Shipp et al., 2002; Yeoh et al., 2002.
- A possible definition: "(...) a group of genes expressed in a specific cell lineage or stage of differentiation or during particular biological response." (Rosenwald et al., 2002, N. Eng. J. Med., 346, p. 1942)
- Often used as independent variables to model clinically relevant information (cancer vs. healthy, survival time, etc).
- Provide insight into biological mechanisms and processes and have potential diagnostic use.
- However, searching for molecular signatures often done using a very diverse and ad-hoc methodology.

What we want:

- Find groups of genes [“group of genes” = “signature component”] so that genes within a group are tightly coexpressed, and the set of groups do a decent predictive job.
- Nice if a similar procedure can be applied to different types of dependent (phenotypic) data (e.g., class membership, survival data, expression of a relevant protein).
- Should help gain some understanding, not necessarily find **The** best predictor (flexibility to play around with trade-offs).

What is, operationally, a signature?

What is, operationally, a signature?

Based on the literature, it seems reasonable that the following conditions should hold:

What is, operationally, a signature?

Based on the literature, it seems reasonable that the following conditions should hold:

- Genes of a signature component should show tight co-expression. Thus, the component or profile should capture most of the variability in the genes that are part of that signature component or profile.

What is, operationally, a signature?

Based on the literature, it seems reasonable that the following conditions should hold:

- Genes of a signature component should show tight co-expression. Thus, the component or profile should capture most of the variability in the genes that are part of that signature component or profile.
- For a given classification/prediction problem only a few signature components should be needed to get a decent predictor.

What is, operationally, a signature?

Based on the literature, it seems reasonable that the following conditions should hold:

- Genes of a signature component should show tight co-expression. Thus, the component or profile should capture most of the variability in the genes that are part of that signature component or profile.
- For a given classification/prediction problem only a few signature components should be needed to get a decent predictor.
- Signature components could have many genes.

A method: key elements

A method: key elements

- **Tight co-expression:** We can use *Principal Components Analysis* to characterize a signature component. The first principal component of a PCA on the genes that belong to a signature component should capture a large fraction of the total variance of those genes.

A method: key elements

- **Tight co-expression:** We can use *Principal Components Analysis* to characterize a signature component. The first principal component of a PCA on the genes that belong to a signature component should capture a large fraction of the total variance of those genes.
- **Few signature components:** Add new components only if justified. (Is this always what we want to do?).

A method: key elements

- **Tight co-expression:** We can use *Principal Components Analysis* to characterize a signature component. The first principal component of a PCA on the genes that belong to a signature component should capture a large fraction of the total variance of those genes.
- **Few signature components:** Add new components only if justified. (Is this always what we want to do?).
- **Signature components could have many genes:** Retain as many genes per component as possible.

A method: key elements

- **Tight co-expression:** We can use *Principal Components Analysis* to characterize a signature component. The first principal component of a PCA on the genes that belong to a signature component should capture a large fraction of the total variance of those genes.
- **Few signature components:** Add new components only if justified. (Is this always what we want to do?).
- **Signature components could have many genes:** Retain as many genes per component as possible.
- **Predictor:** Build a predictor using the signature components (1st PCs). **Diagonal Linear Discriminant Analysis** (*DLDA*).

- We have: $\mathbf{Y}\ (n \times q)$, $\mathbf{X}\ (n \times p)$, $p \gg n$.
- We want: $\mathbf{Y},\ \mathbf{X}^*\ (n \times k),\ k < n$.
- $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_{pr1,1}, \mathbf{x}_{pr1,2}, \mathbf{x}_{pr1,3}, \ldots, \mathbf{x}_{pr2,1}, \mathbf{x}_{pr2,2}, \ldots]$
- $\mathbf{X}^* = [\mathbf{pr}_1, \mathbf{pr}_2, \ldots, \mathbf{pr}_k]$.
- $\mathbf{pr}_i$ is the ith signature component or profile, and it is the 1st PC of a PCA on genes $\mathbf{x}_{pr_i,1}, \mathbf{x}_{pr_i,2}, \ldots, \mathbf{x}_{pr_i,m_i}$.
- Each gene belongs to either one (and only one) signature component or to none.
- $\mathbf{Y} = f(\mathbf{X}^*) + \epsilon$.

Addition of signature components

- Find a *seed gene* for a new signature component:

Addition of signature components

- Find a *seed gene* for a new signature component:
  - For each gene i among available genes: $model_i = DLDA(previous.components + gene_i)$

Addition of signature components

- Find a *seed gene* for a new signature component:
    - For each gene i among available genes: $model_i = DLDA(previous.components + gene_i)$
    - Select as *seed gene*: i with lowest (cross-validated) error, $pred.error.model_i$.

Addition of signature components

- Find a *seed gene* for a new signature component:
  - For each gene i among available genes: $model_i = DLDA(previous.components + gene_i)$
  - Select as *seed gene*: i with lowest (cross-validated) error, $pred.error.model_i$.
  - Add new signature component if $pred.error.model_i < last.pred.error - c_1 s.e.$ ($c_1 = 1$).

Addition of signature components

- Find a *seed gene* for a new signature component:
  - For each gene i among available genes: $model_i = DLDA(previous.components + gene_i)$
  - Select as *seed gene*: i with lowest (cross-validated) error, $pred.error.model_i$.
  - Add new signature component if $pred.error.model_i < last.pred.error - c_1 s.e.$ ($c_1 = 1$).
- **Initial signature component**: all genes with abs. corr. with seed gene > r_{seed} (e.g., $r_{seed} = 0.65$).
  - These are the candidate genes to belong to that component.
  - But this initial signature component might not fulfill previous requirements (%var, predictive performance).
  - Examine if elimination of genes is needed.

Elimination of genes from components

Elimination of genes from components

- Eliminate, one by one, genes with smallest abs. corr. with 1st PC until % variance of 1st PC $> v$ (e.g., 85% or 75%).

Elimination of genes from components

- Eliminate, one by one, genes with smallest abs. corr. with 1st PC until % variance of 1st PC $> v$ (e.g., 85% or 75%).
- Ensure that predictive accuracy cannot be improved by removing any gene from signature component:
  - For each gene, i, in current signature component: $model_i = DLDA(previous.components + current.component_{-i})$.
  - Eliminate gene i from signature component if (cross-validated) $pred.error.model_i < last.pred.error - c_2 s.e.$ ($c_2 = 1$).
  - Repeat until no gene is eliminated.

Bootstrap

We use the bootstrap to asses stability of results and measure prediction error (.632+ rule).

- Take B (= 100) bootstrap samples, and for each one run the above procedure.
- *Common genes*: genes that are returned in at least 20% of the samples.
- For each run, eliminate from the signature components those genes that are not in common genes to obtain "clean signature components".
- *Consensus signature components* are obtained as the (most inclusive) union of all "clean signature components" with a non-zero intersection.

Can we recover signatures?

- Simulation study.
- Generate signature data from a multivariate normal distribution.
- Correlation between genes within a signature component: 0.9. Between genes among signature components: 0. (i.e.,

$$\Sigma = \begin{bmatrix} a & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & a & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & a \end{bmatrix},$$

$$a = \begin{bmatrix} 1 & 0.9 & \cdots & 0.9 \\ 0.9 & 1 & \cdots & 0.9 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

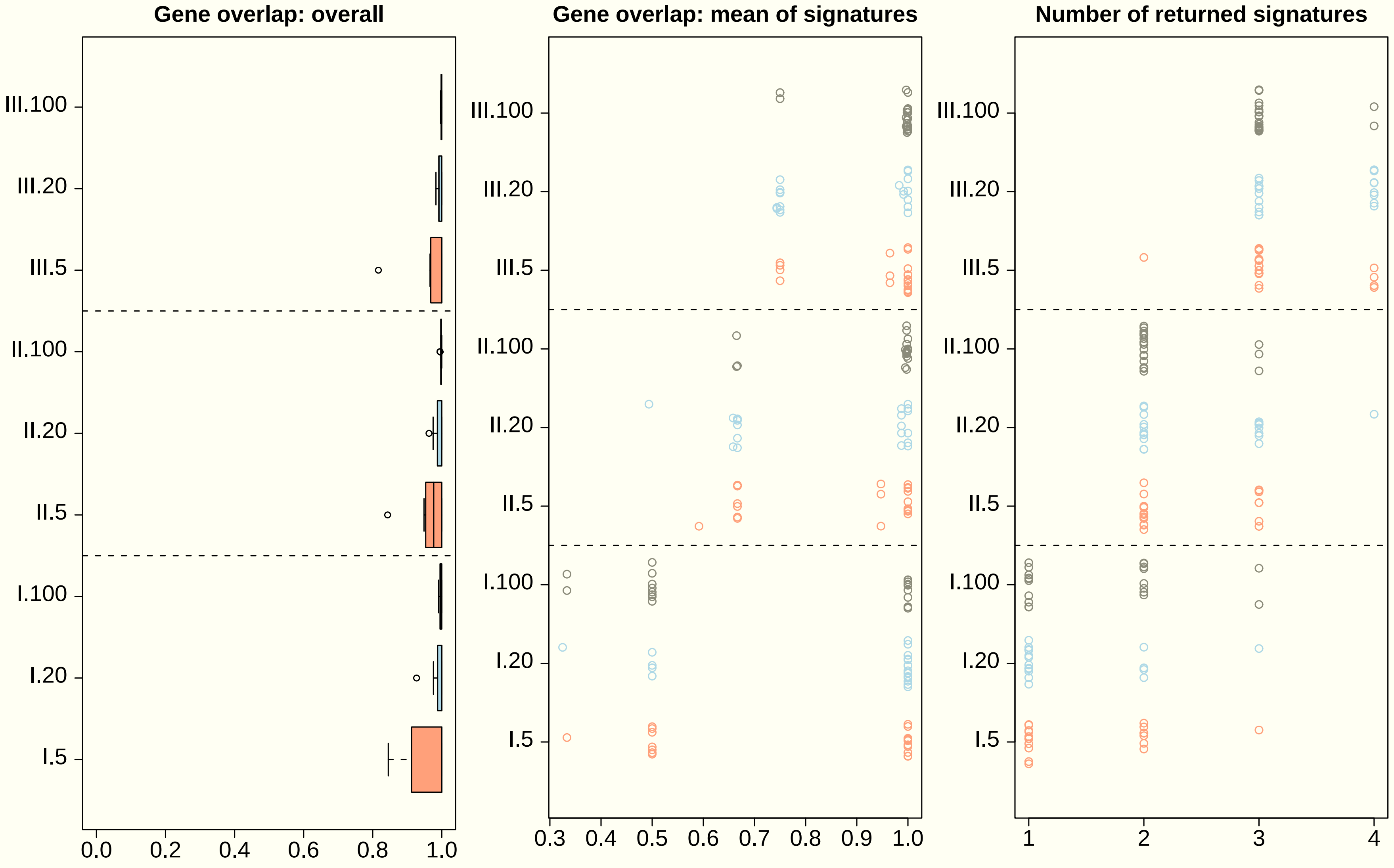
).

- Means of classes set so that:
  - unconditional prediction error rate of a DLDA with a gene from each signature component is approx. 5%;
  - each signature component has the same relevance in separation.
- Number of signature components: {1, 2, 3}.
- Number of classes: {2, 3, 4}.
- Number of genes per signature component: {5, 20, 100}.
- Add another 4000 $N(0, 1)$ variables to matrix of covariates.
- Number of subjects: 25 per class.
- Generate 20 data sets and run procedure.

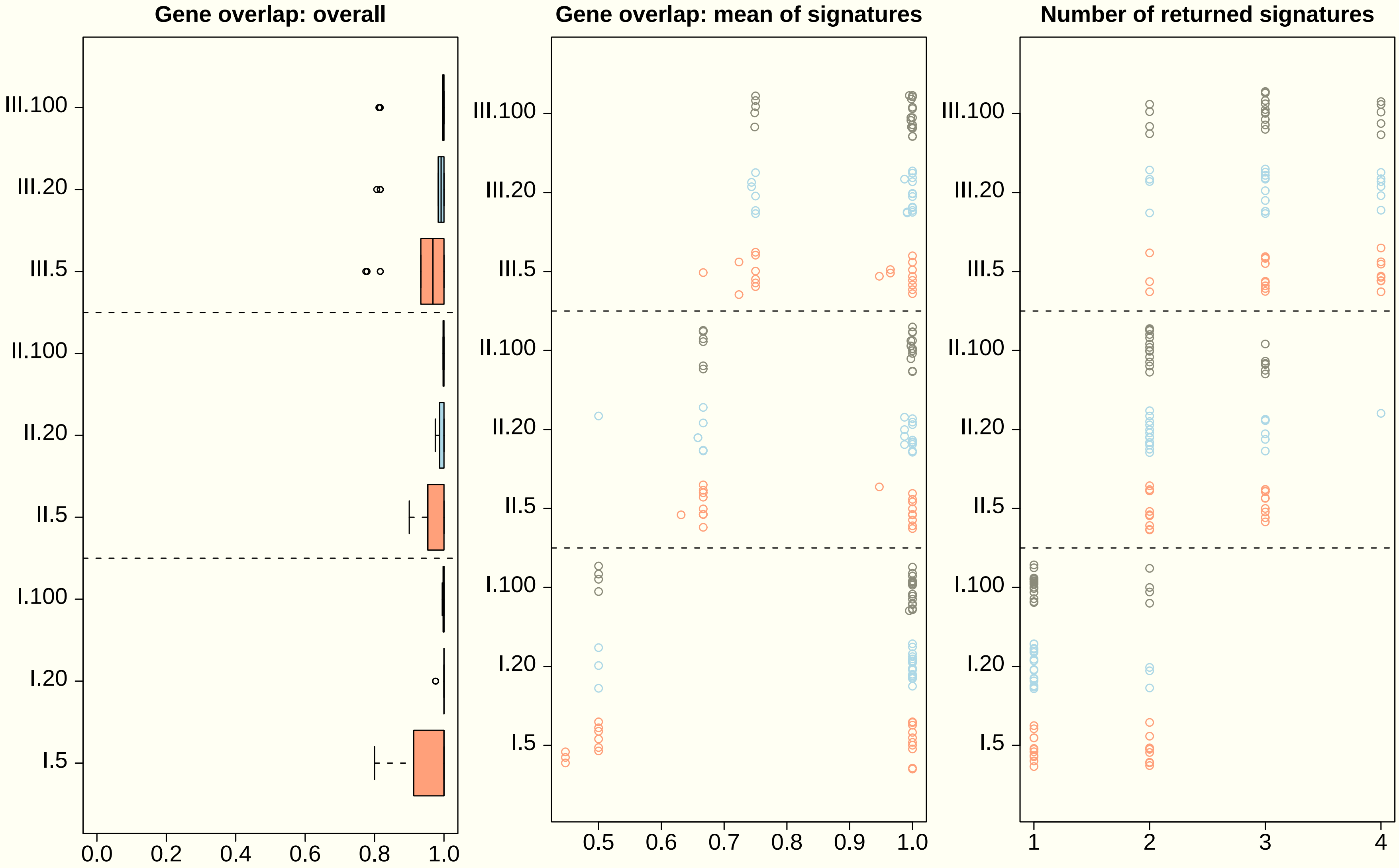
Class means

- One signature component:
  - Two classes: $\mu_1 = -1.65, \mu_2 = 1.65$.
  - Three classes: $\mu_1 = -3.58, \mu_2 = 0, \mu_3 = 3.58$.
  - Four classes: $\mu_1 = -3.7, \mu_2 = 0, \mu_3 = 3.7, \mu_4 = 7.4$.
- Two signature components:
  - Two classes:$\mu_1 = [-1.18, -1.18], \mu_2 = [1.18, 1.18]$.
  - Three classes: $\mu_1 = [0, 0], \mu_2 = [3.88cos(15), 3.88sin(15)], \mu_3 = [3.88cos(75), 3.88sin(75)]$.
  - Four classes: $\mu_1 = [1, 1], \mu_2 = [4.95, 1], \mu_3 = [1, 4.95], \mu_4 = [4.95, 4.95]$.
- Three signature components:
  - Two classes:$\mu_1 = [-0.98, -0.98, -0.98], \mu_2 = [0.98, 0.98, 0.98]$.
  - Three classes: $\mu_1 = [2.76, 0, 0], \mu_2 = [0, 2.76, 0], \mu_3 = [0, 0, 2.76]$.
  - Four classes: $\mu_1 = [2.96, 0, 0], \mu_2 = [0, 2.96, 0], \mu_3 = [0, 0, 2.96], \mu_4 = [2.96, 2.96, 2.96]$

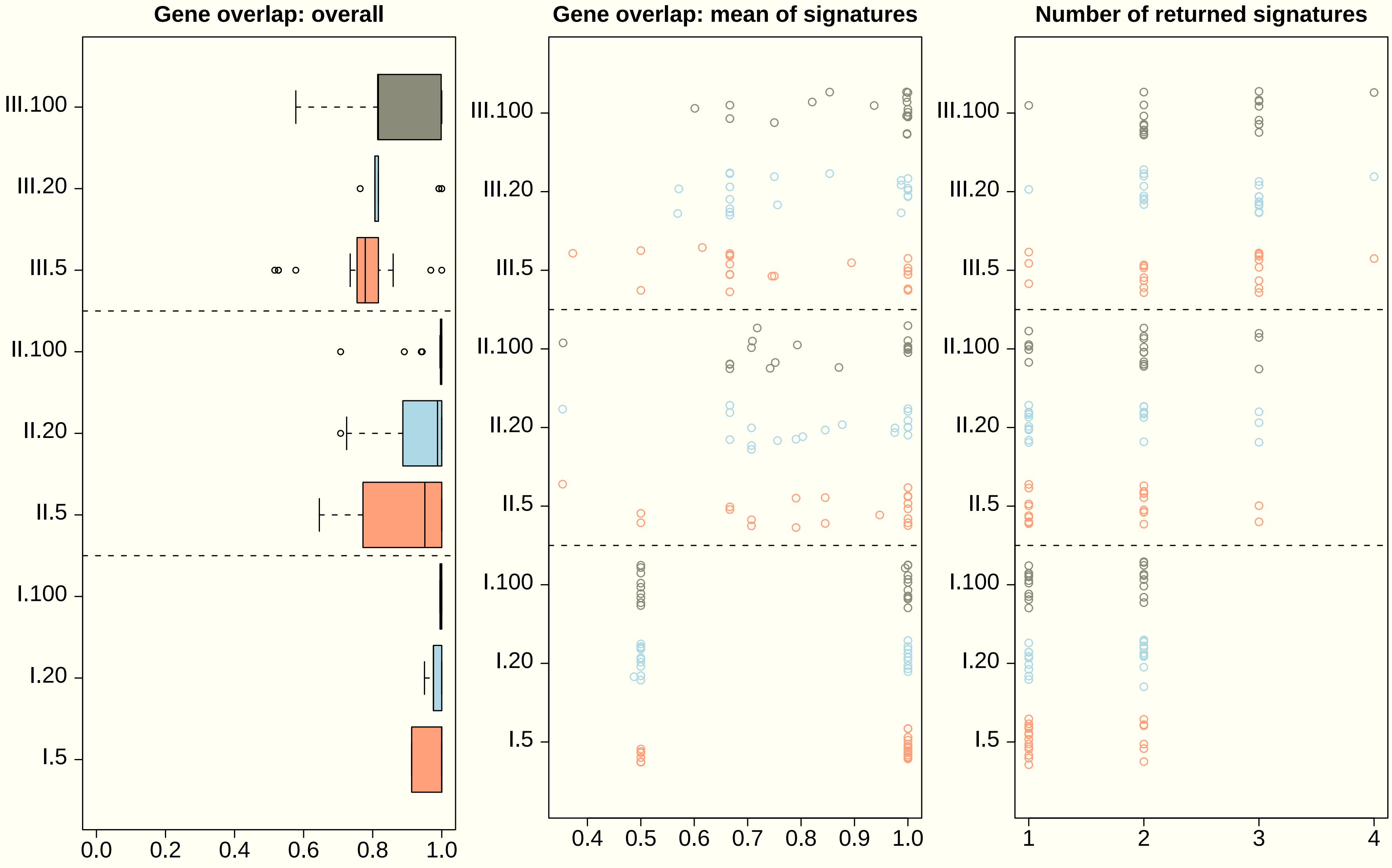
Four classes



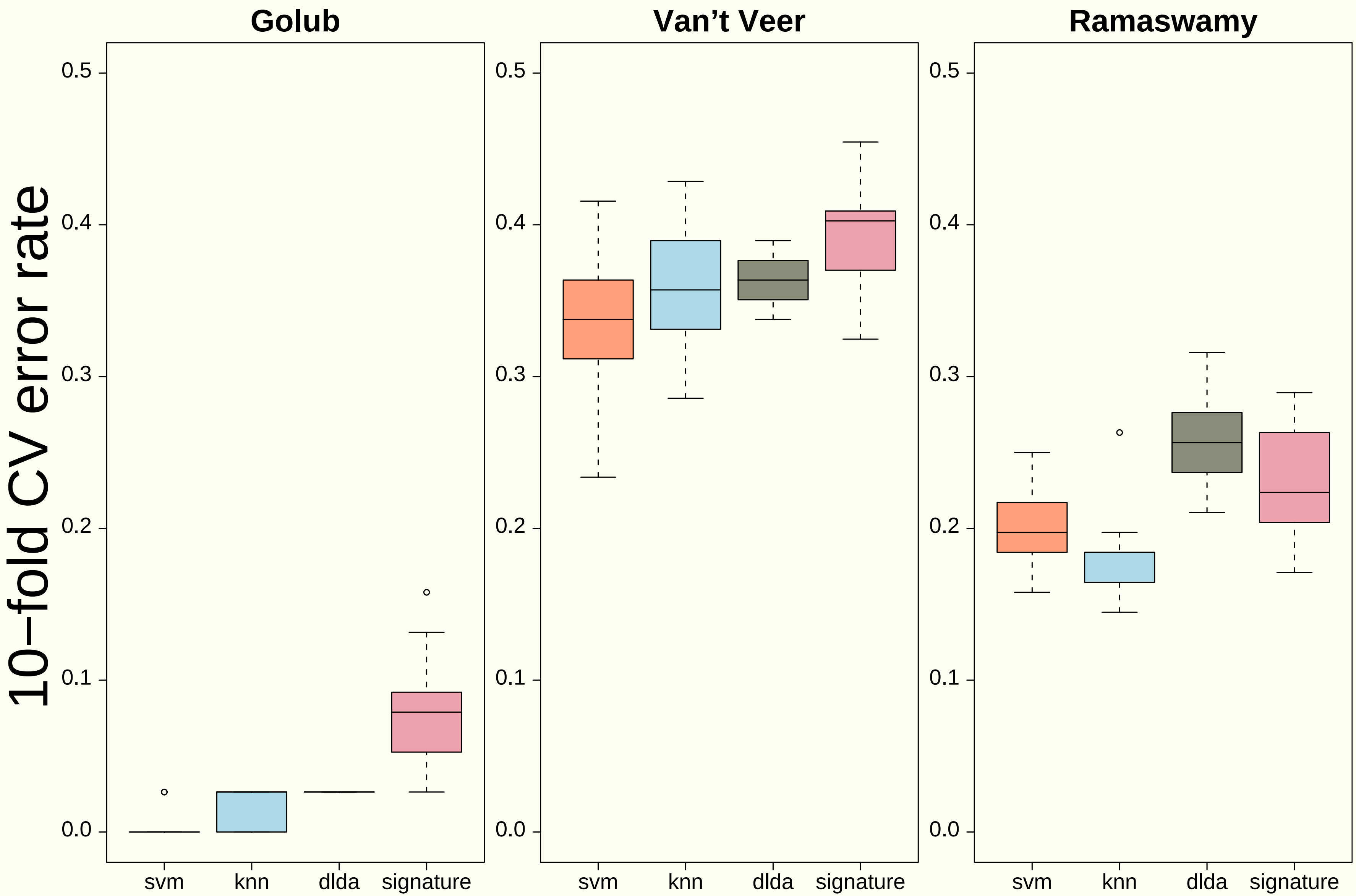
Three classes



Two classes


Gene overlap: overall
Gene overlap: mean of signatures
Number of returned signatures
III.100
III.20
III.5
II.100
II.20
II.5
I.100
I.20
I.5
0.0
0.2
0.4
0.6
0.8
1.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1
2
3
4

Comparison with standard methods



Stability of results?

- Models on all data:
  - Golub: 1st PC > 85%: 1 comp. (1 gene); 1st PC > 75%: 1 comp. (6 genes); 1st PC > 70%: 2 comp. (11 and 19 genes).
  - van't Veer: 1 comp. (1 gene); 1 comp. (3 genes); 5 comp. (13 genes); . . .
  - Ramaswamy: 1 comp. (1 gene); 2 comp. (10 and 1 genes); 2 comp. (2 genes); . . .

Stability of results?

- Models on all data:
  - Golub: 1st PC > 85%: 1 comp. (1 gene); 1st PC > 75%: 1 comp. (6 genes); 1st PC > 70%: 2 comp. (11 and 19 genes).
  - van't Veer: 1 comp. (1 gene); 1 comp. (3 genes); 5 comp. (13 genes); . . .
  - Ramaswamy: 1 comp. (1 gene); 2 comp. (10 and 1 genes); 2 comp. (2 genes); . . .
- Bootstrap:
  - 1st PC > 85% var:
    - Golub: 1 comp. with 19 genes;
    - van't Veer and Ramaswamy: no common genes;
  - 1st PC > 75% var:
    - Golub: 1 comp. with 48 genes;
    - van't Veer: no common genes;
    - Ramaswamy: 1 comp. of 2 genes;

Discussion

- ?
  - Appropriate threshold for % var., correlations, etc?
  - Entry of a component, given previous components?
  - Within-group heterogeneity?
  - PCA: between vs. within group patterns.

Discussion

- ?
  - Appropriate threshold for % var., correlations, etc?
  - Entry of a component, given previous components?
  - Within-group heterogeneity?
  - PCA: between vs. within group patterns.
- Easily extended:
  - Other classifiers (e.g., logistic regression, knn, svm).
  - Other dependent variables: survival analysis.

Related to

- Partial Least Squares (and, to a lesser extent, Principal Components Regression).
- Factor analysis with oblique rotations to obtain clusters of variables (SAS's PROC VARCLUS).
- "Supergenes" or "metagenes" of West et al.
- . . .

Conclusion

- The logic of this method follows directly from what are considered biologically relevant signature characteristics.

Conclusion

- The logic of this method follows directly from what are considered biologically relevant signature characteristics.
- Results that are biologically relevant and interpretable and can complement other approaches.

Conclusion

- The logic of this method follows directly from what are considered biologically relevant signature characteristics.
- Results that are biologically relevant and interpretable and can complement other approaches.
- This method defines a framework that allows us to find signatures regardless of the type of dependent variable.

Conclusion

- The logic of this method follows directly from what are considered biologically relevant signature characteristics.
- Results that are biologically relevant and interpretable and can complement other approaches.
- This method defines a framework that allows us to find signatures regardless of the type of dependent variable.
- Easy to implement and R code available.

Acknowledgements

- A. Pérez and M. A. Piris for emphasizing the importance of molecular signatures.
- Bioinformatics Unit, CNIO, for discussion.
- C. Lázaro-Perea for discussion.