Sparse canonical methods for biological data integration: application to a cross-platform study

Kim-Anh Lê Cao, Pascal G.P. Martin, Christèle Robert-Granié, Philippe Besse

Sparse canonical methods for biological data integration: application to a cross-platform study

Kim-Anh Lê Cao[1,2,4*]and Pascal G.P. Martin,[3,4] Christèle Robert-Granié[2], Philippe Besse[1]

Abstract

In the context of integration for systems biology, very few sparse approaches have been proposed so far to select variables in a canonical framework. In this study we propose a canonical mode of a new sparse PLS approach to handle two-block data sets, where the relationship between the two types of variables is known to be symmetric. Sparse PLS has been proposed for either a regression or a canonical mode and includes a built-in procedure to perform variable selection while integrating data. To illustrate the canonical mode approach, we analyzed the NCI60 data sets, where two different platforms (cDNA and Affymetrix chips) were used to study the transcriptome of sixty cancer cell lines.
We compare the results obtained with two other sparse or related canonical approaches: CCA with Elastic Net penalization (CCA-EN) and Co-Inertia Analysis (CIA). The latter does not include a built-in procedure for variable selection and requires a two-step analysis. We stress the lack of statistical criteria to evaluate canonical methods, which makes biological interpretation crucial to compare the different gene lists. We propose comprehensive graphical representations of both samples and variables to facilitate the biologist interpretation.
We show that sPLS and CCA-EN select highly relevant genes, which enable a detailed understanding of the molecular characteristics of several groups of cell lines. These two approaches were found to bring similar results, although they highlighted the same phenomenons with a different priority. On the other hand, CIA tended to select redundant information. These canonical methods seem to be efficient tools to deal with variable selection in the context of high-throughput data integration.

Introduction

When dealing with the integration of high dimensional biological data, the application of linear multivariate models such as Partial Least Squares regression (PLS, Wold, 1966) and Canonical Correlation Analysis (CCA, Hotelling, 1936), are often limited by the size of the data set (ill-posed problems), the noisy and the multicollinearity characteristics of the data and the lack of interpretability (PLS). However, these approaches still remain extremely interesting for this type of problems, first because they allow for the compression of the data

*to whom correspondence should be addressed: Kim-Anh.Le-Cao@toulouse.inra.fr
[1]Institut de Mathématiques, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse, France
[2]Station d'Amélioration Génétique des Animaux UR 631, Institut National de la Recherche Agronomique, F-31326 Castanet, France
[3]Laboratoire de Pharmacologie et Toxicologie UR 66, Institut National de la Recherche Agronomique, F-31931 Toulouse, France
[4]Both authors contributed equally to this work.

into 2 to 3 dimensions for a more powerful and global view, and second as their resulting components and loading vectors capture dominant and latent properties of the studied process. They may hence provide a better understanding of the underlying biological systems, for example by revealing groups of samples that were previously unknown or uncertain.
In this study, we were interested in integrating two high dimensional data sets, where variables of two types are measured on the same individuals or samples. Recent integrative biological studies applied Principal Component Analysis, or PLS (Bylesjö et al., 2007; Vijayendran et al., 2008), but in a regression framework, where prior biological knowledge indicates which type of omic data is expected to explain the other type (for example transcripts and metabolites).
Here, we specifically focus on a canonical framework, when there is either no assumption on the relationship between the two sets of variables (exploratory approach), or when a reciprocal relationship between the two sets is expected (*e.g.* cross platform comparisons).
Few statistical methods can answer this problem. Among them, some are limited by the number of variables (CCA) or do not give straightforward interpretable results when the number of variables is too large (PLS). Some associated approaches have recently been developed to include a built-in selection procedure, so as to allow variable selection in both data sets. These sparse methods adapt lasso penalty (Tibshirani, 1996) or combine lasso and ridge penalties (Elastic Net, Zou and Hastie, 2005) for feature selection in integration studies.
In this study, we propose a sparse canonical approach called "sparse PLS" (sPLS) in the context of integration for systems biology. Methodological aspects and analyses of the sPLS in a regression framework were presented in (Lê Cao et al., 2008). This novel computational method provides variable selection of two-block data set in a one step procedure, for integrating variables of two types.
When applying canonical methods, most validation criteria used in a regression context are not statistically meaningful. Instead, the biological relevancy of the results should be evaluated during the validation process. We therefore compare sparse PLS with two other canonical approaches: penalized CCA adapted with Elastic Net (Waaijenborg et al., 2008), which is a sparse method that was applied to relate gene expression with gene copy numbers in human gliomas, and Co-Inertia Analysis (CIA, Doledec and Chessel, 1994) that was first developed for ecological data, and then for canonical high-throughput biological studies (Culhane et al., 2003). This latter approach does not include feature selection, which has to be performed in a two-step procedure.
This comparative study has two aims. First to better understand the main differences between each of these approaches and identify which method would be appropriate depending on the biological question. Second to highlight how each method is able to reveal the underlying biological processes inherent to the data. This type of comparative analysis renders the biological interpretation mandatory to strengthen the statistical hypothesis, especially when there is a lack of statistical criteria to assess the validity of the results.
We first recall some canonical methods among which the two sparse methods will be compared with CIA on the NCI60 cell lines data set, which is fully described. We propose to use appropriate graphical representations to discuss the results. The different gene lists are assessed, first with some statistical criteria, and then through their biological interpretation. Finally we discuss the pros and cons of each tested approach before concluding.

1 Canonical Methods

We focus on two-block data matrices denoted $X(n \times p)$ and $Y(n \times q)$, where the p variables x^j and q variables y^k are two types of measures performed on the same samples or individuals, $j = 1 \dots p$, $k = 1 \dots q$. Prior biological knowledge on these data allow us to settle into a canonical framework, *i.e.* there exists a reciprocal relationship between the X variables and the Y variables. In the case of high throughput biological data, the large number of variables may affect the exploratory method, due to numerical issues (as it is the case for example with CCA), or lack of interpretability (PLS).
We next recall three types of multivariate methods (CCA, PLS, CIA). For CCA and PLS, we recall their associated sparse approaches that were proposed, either to select variables from each set or to deal with the ill-posed problem commonly encountered in high-throughput biological data.

1.1 Canonical Correlation Analysis

Canonical Correlation Analysis (Hotelling, 1936) studies the relationship between two sets of data. The CCA n-dimensional score vectors (Xa_h, Yb_h) come in pairs to solve the objective function:

$$\arg \max_{a_h' a_h = 1, b_h' b_h = 1} cor(Xa_h, Yb_h), \qquad h = 1 \dots H$$

where the p- and q-dimensional vectors a_h and b_h are called canonical factors, or loading vectors, and h is the CCA chosen dimension.
As $cor(Xa_h, Yb_h) = cov(Xa_h, Yb_h)/\sqrt{var(Xa_h)}\sqrt{var(Yb_h)}$, the aim of CCA is to simultaneously maximize $cov(Xa_h, Yb_h)$ and minimize the variances of Xa_h and Yb_h.
In the $p + q >> n$ framework, CCA suffers from the high dimensionality as it requires the computation of two inverses of the covariance matrices XX' and YY' that are singular. This implies numerical difficulties, since the canonical correlation coefficients are not uniquely defined. One solution proposed by Vinod (1976) was to introduce l_2 penalties in a ridge CCA (rCCA) on the covariance matrices, so as to make them invertible. González et al. (2008b) applied rCCA to post genomic data (Combes et al., 2008) and proposed to choose the optimal penalization parameters with cross-validation.
It is known (Gittins, 1985) that the CCA loadings are not directly interpretable. It is however very instructive to interpret these components by calculating the correlation between the original data set X and $\{a_1, \dots, a_H\}$ and similarly between Y and $\{b_1, \dots, b_H\}$, to project the variables on correlation circles. Easier interpretable graphics are obtained, which readability was improved by González et al. (2008b) in the `R` package `cca`. In our study, rCCA could not be applied as it does not perform feature selection. Furthermore, because of the non direct interpretability of the loadings, a variable selection in a two-step procedure is difficult to perform, as it must be based on correlation circles graphics.

1.2 PLS

Partial Least Squares regression (Wold, 1966) is based on the simultaneous decomposition of X and Y into latent variables and associated loading vectors. The latent variables methods (*e.g.* PLS, Principal Component Regression) assume that the studied system is driven by a small number of n-dimensional vectors called latent variables. These latter may

correspond to some biological underlying phenomenons which are related to the study (Wold et al., 2004).
Like CCA, the PLS components (latent variables) are linear combinations of the predictor variables, but the objective function differs as it is based on the maximization of the covariance between each linear combination of the two sets of variables:

$$\arg \max_{a_h' a_h = 1, b_h' b_h = 1} cov(X a_h, Y b_h), \qquad h = 1 \dots H.$$

We denote $\xi_h = X a_h$ and $\omega_h = Y b_h$ the latent variables associated to each loading vector a_h and b_h, h being the chosen PLS dimension. On one hand, and in contrary to CCA, the loading vectors (ξ_h, ω_h) are interpretable and can give information about how the x^j and y^k variables combine to explain the relationships between X and Y. On the other hand, the PLS latent variables (a_h, b_h) indicate the similarities or dissimilarities between the individuals, related to the loading vectors.
Many PLS algorithms exist, not only for different shapes of data (SIMPLS, de Jong, 1993, PLS1 and PLS2, Wold, 1966, PLS-SVD, Lorber et al., 1987) but also for different aims (predictive, like PLS2 or modelling, like PLS-mode A, see Tenenhaus, 1998; Wegelin, 2000; Waaijenborg et al., 2008). In this study we especially focus on a modelling aim (canonical mode) between the two data sets, by deflating X and Y in a symmetric way (see appendix A).

1.3 Penalized Correlation Analysis with Elastic Net (CCA-EN)

Waaijenborg et al. (2008) proposed a sparse penalized variant of CCA using Elastic Net (Zou and Hastie, 2005; Zou et al., 2006) in a regression framework. To do so, the authors used the PLS-mode A formulation (Tenenhaus, 1998; Wegelin, 2000) to introduce penalties. Note that Elastic Net is well adapted in this particular framework. It combines the advantages of the ridge regression, that penalizes the covariance matrices XX' and YY' which become non singular, and the lasso (Tibshirani, 1996) that allows variable selection, in a one step procedure. However, when $p + q$ is very large, the resolution of the optimization problem requires intensive computations, and Zou and Hastie (2005); Waaijenborg et al. (2008) proposed instead to perform a univariate thresholding, that leaves only the lasso estimates to compute (see appendix C).

1.4 sparse PLS

Lê Cao et al. (2008) proposed a sparse PLS approach (sPLS) based on a PLS-SVD variant, so as to penalize both loading vectors a_h and b_h simultaneously.
For any matrix M$(p \times q)$ of rank r, the SVD of M is given by:

$$M = A \Delta B'$$

where the columns of A $(n \times r)$ and $B(r \times p)$ are orthonormal and contain the eigenvectors of MM' and $M'M$, Δ is a diagonal matrix of the squared eigenvalues of MM' or $M'M$.
If $M = X'Y$, then the column vectors of A (resp. B) correspond to the loading vectors of the PLS, and sparsity in both vectors can be introduced by iteratively penalizing a_h and b_h with a soft-thresholding penalization, as was proposed in a similar manner by Shen and Huang (2007) for a sparse PCA (see appendix B for more details). Both deflation modes, as referred

in section 1.2, were proposed. In this paper, we will focus on the canonical mode only. The regression mode has been already been discussed in Lê Cao et al. (2008) where a thorough biological interpretation was provided in this framework.

1.5 Co-Inertia Analysis

Co-Inertia analysis (CIA) was first introduced by Doledec and Chessel (1994) in the context of ecological data, before Culhane et al. (2003) applied it to high-throughput biological data. CIA is suitable for a canonical framework, as it is adapted for a symmetric analysis. It involves analyzing each data set separately either with principal component analyses, or with correspondence analyses, such that the covariance between the two new sets of projected scores vectors (that maximize either the projected variability or inertia) is maximal. This results in two sets of axes, where the first pair of axes are maximally co-variant, and are orthogonal to the next pair (Robert and Escoufier, 1976).
CIA does not propose a built-in variable selection, but we can perform instead a two-step procedure by ordering the weight vector (loadings) for each CIA dimension and select the top variables.

1.6 Differences between the approaches

The three canonical approaches that we want to compare (CCA-EN, sPLS, CIA) profoundly differ in their construction, and hence their aims.
CCA-EN looks for canonical variate pairs (Xa_h, Yb_h), such that a penalized version of the canonical correlation is maximized. This explains why a non monotonic decreasing trend in the canonical correlation can sometimes be obtained (Waaijenborg et al., 2008). On the other hand, sPLS (canonical mode) and CIA aim at maximizing the covariance between the scores vectors, so that there is a strong symmetric relationship between both sets. However, here CIA is based on the construction of two Correspondence Analyses, whereas sPLS is based on a PLS analysis.

1.7 Parameters tuning

In CCA-EN, the authors proposed to tune the penalty parameters for each dimension, such that the canonical correlation $cor(Xa_h, Yb_h)$ is maximized. In practice, they showed that the correlation did not change much when variables were added in the selection. Hence, an appropriate way of tuning the parameters would be to choose instead the degree of sparsity (*i.e.* the number of variables to select), as proposed by Zou et al. (2006) for their sparse PCA in the `elasticnet` `R` package, and rely on the biologists needs. Indeed, a too short gene selection may lack in information, as some of the functions or annotations may be missing. The same strategy will be used for sPLS. No other parameters than the number of selected variables is needed in CIA either.

1.8 Outputs

Graphical representations should be a an important issue to help biologists interpret the results. Hence we propose to homogenize all ouputs to get comparable results.
Samples will be represented with the scores or latent vectors, in a superimposed manner, as proposed in the `R` package `ade4` (Thioulouse et al., 1997), first to show how samples are

clustered based on their biological characteristics, and second to measure if both data sets strongly agree according to the applied approach.
Variables will be represented on correlation circles, as proposed by González et al. (2008b). Correlations between the original data sets and the loading vectors are computed so that highly correlated variables will cluster together in the resulting graphics. Only the selected variables in each dimension will be represented. This type of graphic will allow to identify interactions between the two types of variables and relate the variable clusters to their associated sample clusters.

2 Cross-platform study

2.1 Data sets and relevance for a canonical analysis

We compared the three canonical methods (CCA-EN, CIA and sPLS) for their ability to highlight the relationships between two gene expression data sets both obtained on a panel of 60 cell lines (NCI60) from the National Cancer Institute (NCI). This panel consists of human tumor cell lines derived from patients with leukemias (LE), melanomas (ME) and cancers of ovarian (OV), breast (BR), prostate (PR), lung (LU), renal (RE), colon (CO) and central nervous system (CNS) origin. The NCI60 is used by the Developmental Therapeutics Program (DTP) of the NCI to screen thousands of chemical compounds for growth inhibition activity and it has been extensively characterized at the DNA, mRNA, protein and functional levels. The data sets considered here have been generated using Affymetrix (Butte et al., 2000; Staunton et al., 2001) or spotted cDNA (Ross et al., 2000) platforms. These data sets are highly relevant to an analysis in a canonical framework since 1) there is some degree of overlap between the genes measured by the two platforms but also a large degree of complementarity through the screening of gene sets representing common pathways or biological functions (Culhane et al., 2003) and 2) they play fully symmetric roles as one data set cannot be explained by the other, it as would be done in a regression framework. Considering that the aim of the canonical methods is to capture the relationships between two data sets, each of which should be relevant to the problem under study (here, the characteristics of the gene expression profiles of tumor cell lines of different origins), we believe that these methods should primarily apply to pre-processed data sets, where data transformation, background correction and normalization steps were performed beforehand. These steps and the resulting data sets that were analyzed here are briefly described below.

2.2 The Ross Data set

Ross et al. (2000) used spotted cDNA microarrays containing 9,703 human cDNAs to profile each of the 60 cell line in the NCI60 panel (Ross et al., 2000). Here, we used a subset of 1,375 genes that has been selected using both non-specific and specific filters described in Scherf et al. (2000). In particular, genes with more than 15% of missing values were removed and the remaining missing values were imputed by k-nearest neighbours (Culhane et al., 2003). The pre-processed data set containing log ratio values is available in Culhane et al. (2003).

2.3 The Staunton Data set

Hu6800 Affymetrix microarrays containing 7,129 probe sets were used to screen each of the 60 cell lines in another study (Butte et al., 2000; Staunton et al., 2001). Pre-processing steps are described in Staunton et al. (2001) and Culhane et al. (2003). They include 1) replacing average difference values less than 100 by an expression value of 100, 2) eliminating genes whose expression was invariant across all 60 cell lines and 3) selecting the subset of genes displaying a minimum change in expression across all 60 cell lines of at least 500 average difference units. The final analyzed data set contained the average difference values for 1,517 probe sets, and is available in Culhane et al. (2003).

2.4 Application of the three canonical methods

We applied CCA-EN, CIA and sPLS to the Ross (X) and Staunton (Y) data sets. For each dimension h, $h = 1 \dots 3$, we performed variable selection of 100 genes from each data set. The number of dimensions was arbitrarily chosen, given that if $H \geq 4$, the interpretation of the results becomes more difficult due to the high number of graphical outputs, and the results were less relevant. The size of the selection (100) was judged small enough to allow for the identification of into individual relevant genes and large enough to reveal gene groups belonging to the same functional category or pathway.
The graphical representation of the individuals, as described in section 1.8, is displayed in a superimposed manner, where each sample will be indicated using an arrow. The start of the arrow will indicate the location of the sample in X in one plot, and the tip the location of the sample in Y in the other plot. Short arrows will therefore indicate if both data sets strongly agree and long arrows a disagreement between the two data sets.

3 Results and Discussion

We apply the three canonical approaches to the NCI60 data set and assess the results in two different ways. First we examine few statistical criteria, then we provide an interpretation of the results from each method, using graphical representations along with database mining.

3.1 How to assess the results ?

Canonical methods are statistically difficult to assess. Firstly because they do not fit into a regression/prediction framework, meaning that cross-validation cannot be computed to evaluate the quality of the model. Secondly because in many two-block biological studies, the number of samples n is very small compared to the number of variables $p+q$. This makes any statistical criteria difficult to compute. This is why graphical outputs are important to analyse the results (see for example Tenenhaus, 1998; Culhane et al., 2003).
When working with biological data, a new way of assessing the results should be to strongly rely on the biological interpretation. Indeed our aim is to show the applicability of each approach and to show if they answer the biological question. We hence propose to base most of our comparative study on the biological interpretation by using appropriate graphical representations of the samples and of the selected variables.

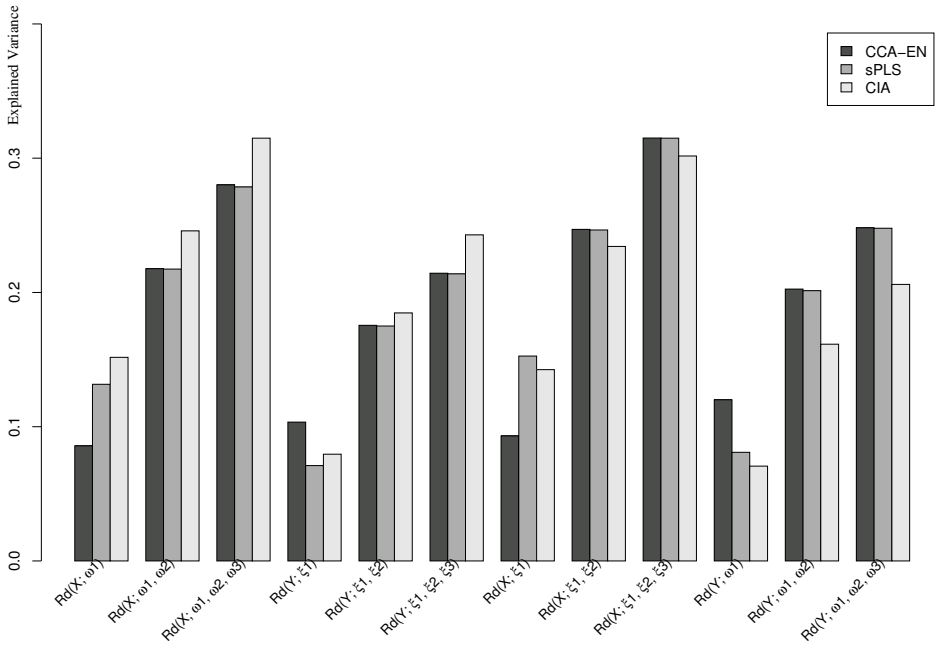

Figure 1: Cumulative explained variance of each data set in relation to its component score (CCA-EN, CIA) or latent variable (sPLS)

3.2 Link between two-block data set

Variance explained by each component. Tenenhaus (1998) proposed to estimate the variance explained in each data set X and Y in relation to the "opposite" component score or latent variables $(\omega_1, \ldots, \omega_H)$ and $(\xi_1, \ldots, \xi_H)$, where $\xi_h = Xa_h$ and $\omega_h = Yb_h$ in all approaches. The redundancy criterion Rd, or part of explained variance, is computed as follows:

$$Rd(X; \omega_1, \ldots, \omega_H) = \frac{1}{p} \sum_{h=1}^{H} \sum_{j=1}^{p} cor^2(x^j, \omega_h)$$

$$Rd(Y; \xi_1, \ldots, \xi_H) = \frac{1}{q} \sum_{h=1}^{H} \sum_{k=1}^{q} cor^2(y^k, \xi_h)$$

Similarly, one can compute the variance explained in each component in relation with its associated data set:

$$Rd(X; \xi_1, \ldots, \xi_H) = \frac{1}{p} \sum_{h=1}^{H} \sum_{j=1}^{p} cor^2(x^j, \xi_h)$$

$$Rd(Y; \omega_1, \ldots, \omega_H) = \frac{1}{q} \sum_{h=1}^{H} \sum_{k=1}^{q} cor^2(y^k, \omega_h)$$

Figure 1 displays the Rd criterion for $h = 1 \ldots 3$ for each set of components $(\xi_1, \ldots, \xi_)$, $(\omega_1, \ldots, \omega_H)$ and for each approach. While there seems to be a great difference in the first dimension between CCA and the other methods, the components in dimensions 2 and 3 explain the same amount of variance in both X and Y for CCA-EN and sPLS. This suggests

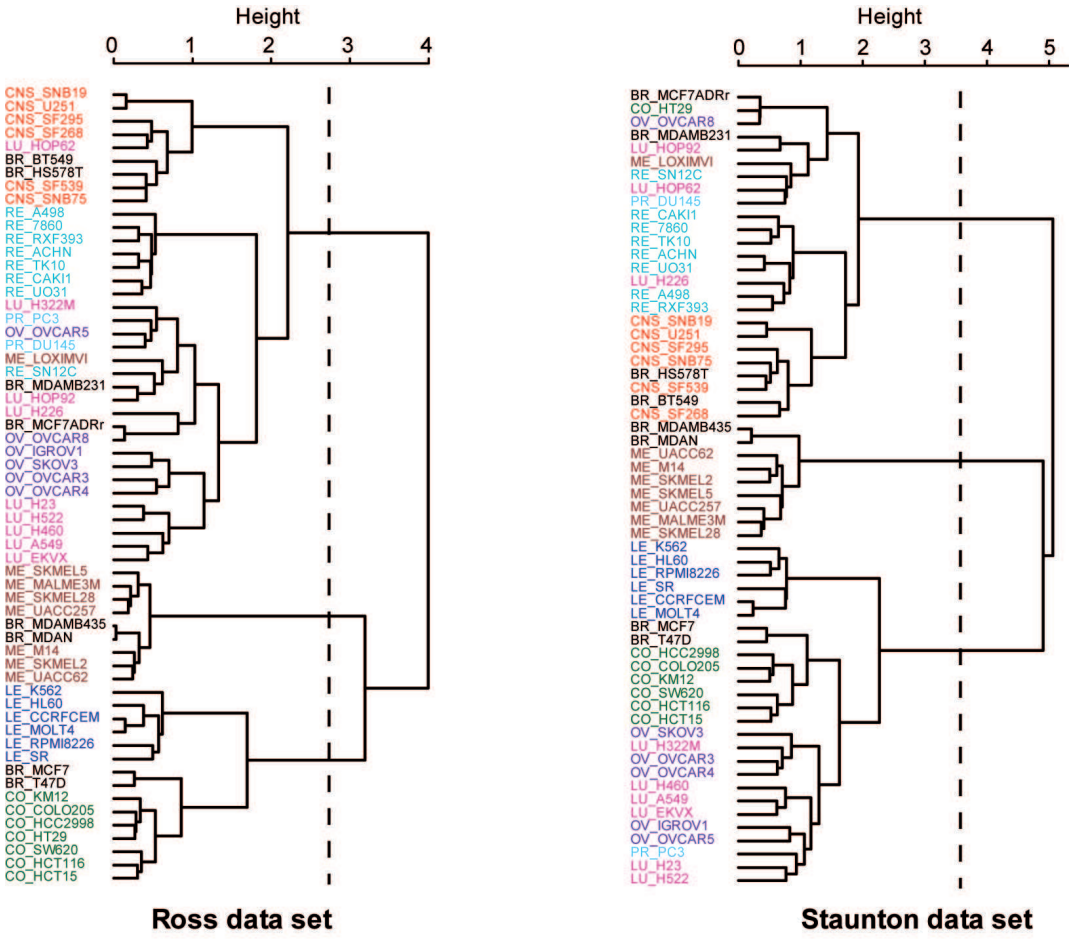

Figure 2: Hierarchical clustering of Ross and Staunton data sets expression profiles of the cell lines, which are coded as CO = Colon, ME = Melanoma, BR = Breast, CNS= Central Nervous System , OV= Ovarian, RE = Renal, PR = Prostate.

Table 1: Correlations of the score vectors/latent variables for each dimension.

	CCA-EN	CIA	sPLS
$cor(\xi_1, \omega_1)$	0.967	0.935	0.938
$cor(\xi_2, \omega_2)$	0.937	0.967	0.964
$cor(\xi_3, \omega_3)$	0.953	0.955	0.944

a strong similarity at this stage between these two approaches. On the other hand, CIA differs from these two methods. The components computed from the "opposite" set explain more variance than CCA/sPLS, and less in their respective set.
More generally, we can observe that more information seems to be present in the X rather than in the Y data set. Indeed, similarly to Culhane et al. (2003), we noticed that a hierarchical clustering of the samples using the distance $1-$correlation with the Ross data set allows a better clustering of the cell lines based on their tissue of origin than with the Staunton data set (Figure 2).

Correlations between each component. The canonical correlations between the pair of score vectors (or latent variables) were very high (>0.93) for any approach and in any dimension (see Table 1). This comfort our hypothesis regarding the canonical aim of each method.
The non monotonic decreasing trend in the canonical correlations in CCA-EN is not what can be expected from a CCA variant, but was also pointed out by Waaijenborg et al. (2008) as the optimization criterion differs from ordinary CCA. However, the computations of the *Rd* criterion (Figure 1) seem to indicate that the cumulative variance explained by the latent variables increases with h. In sPLS and CIA, which aim is to maximize the covariance, we can see that in fact they also highlight very strongly correlated components. This suggests that the associated loading vectors may also bring related information from both data sets.
The maximal canonical correlation ($\simeq 0.97$) is obtained on the first dimension for CCA-EN, and surprisingly only on the second dimension for CIA and sPLS. In the next sections, we will see that in fact CCA-EN and sPLS "swap" their components between the first and second dimensions.

3.3 Interpretation of the observed cell line clusters

Figures 3 and 4 display the graphical representations of the samples in dimension 1 and 2 **(a)**, or 1 and 3 **(b)** for CCA-EN (Fig. 3) and sPLS (Fig. 4), CIA showing patterns similar to sPLS and to those presented in Culhane et al. (2003).
All graphics show that both data sets are strongly related (short arrows), but depending on the applied approach, the components differ. In dimension 1, the pair (ξ_1, ω_1) tends to separate the melanoma cell lines from the other cell lines in CCA-EN (Fig. 3 **(a)**), whereas sPLS and CIA tend to separate the LE and CO cell lines on one side from the RE and CNS cell lines on the other side (Fig. 4 **(a)**). As previously proposed (Culhane et al., 2003), we interpreted this clustering of the cell lines along the first axes of sPLS and CIA as the separation of cell lines with *epithelial* characteristics (mainly LE and CO) from those with *mesenchymal* characteristics (in particular RE and CNS). Epihelial cell generally form layers

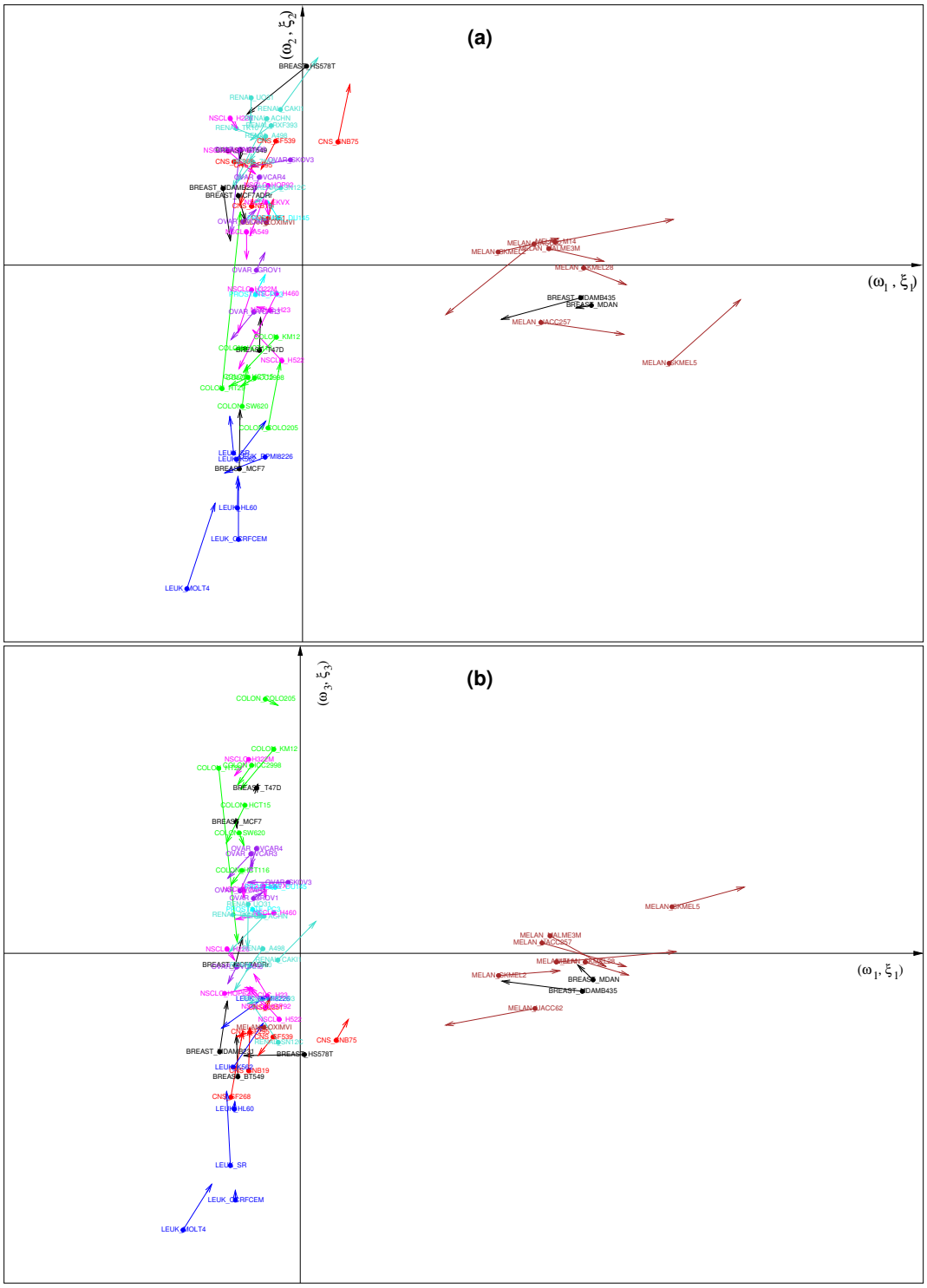

Figure 3: Graphical representations of the cell lines. CCA-EN component scores are displayed in a superimposed manner, where the start of the arrow show the location of the Ross samples, and the tip the Staunton samples. The first and second axis (first and third) are shown in **(a)** and **(b)**.

by making junctions between them and interacting with the extracellular matrix (ECM). On the other hand, mesenchymal cells are able to migrate through the ECM and are found in the connective tissues. We will see that the interpretation of the genes lists selected on the axes separating LE and CO versus RE and CNS strongly argue for such an interpretation of the individuals plot. In addition, it has been previously described that glioblastoma cell lines (CNS) do express mesenchymal stem-like properties at multiple levels, including gene expression (Tso et al., 2006). In dimension 2, we observe the opposite tendency: the pair (ξ_2, ω_2) separates the cell lines with epithelial characteristics from the cell lines with mesenchymal characteristics in CCA-EN while it separates the melanoma samples from the other samples in sPLS and CIA.

When performing hierarchical clustering of the 60 cell lines (with 1−correlation distance) separately on each data set (Figure 2), it appears that the three main clusters of samples largely correspond to the three groups that are separated by all three methods in dimensions 1 and 2 *i.e.* they correspond to 1) cell lines with epithelial characteristics (LE and CO for both data sets), 2) cell lines with mesenchymal characteristics (in particular RE and CNS) and 3) melanomas with which two breast cancer cell lines (MDA_N and MDA_MB435) are systematically clustered. Among these clusters, only the third one is strictly identical for the two data sets. This illustrates that CCA-EN primarily captures the sample characteristics in the clusters that are most conserved between the two data sets, even if these do not underlie the separation of the main clusters within each data set. The fact that, based on their gene expression profiles, ME samples form a relatively homogeneous and compact cluster along with two breast tumor cell lines (MDA_N and MDA_MB435 which are indeed melanoma metastases derived from a patient diagnosed with breast cancer), has been previously shown by other authors (Ross et al., 2000; Scherf et al., 2000; Culhane et al., 2003) and seems largely independent of the initial gene selections that were used here. We believe that the strongest canonical correlation can only be found when separating this specific set of cell lines (see Table 1). This explains why CCA-EN, that looks for maximal correlation, first focuses on this particular axis. On the other hand, sPLS and CIA first focus on the separation between cell lines with epithelial versus mesenchymal characteristics, a separation that is slightly more obvious in the dendrograms obtained from the two data sets, but where the cluster members substantially change between the two data sets. In particular, most OV and LU cell lines are clustered with LE and CO in the Staunton data set while they are clustered with RE and CNS cell lines in the Ross data set (Figure 2). To further evaluate this hypothesis, we permuted the labels from 1 to 4 (out of 7) melanoma cell lines with randomly selected cell lines in one of the data set, thus artificially reducing the consistency between the clustering of the melanoma cell lines in the two data sets. Resulting graphics in CCA-EN happened to be similar to those obtained for sPLS and CIA in the absence of permutation (Figure 3 **(a)**), hence separating epithelial-like versus mesenchymal-like cell lines on the first dimension. By contrast, sPLS and CIA graphics remained the same after the permutations.

3.4 Interpretation of the observed genes clusters

We computed the correlations between the original data sets and the scores vectors or latent variables (ξ_1, ξ_2, ξ_3) and $(\omega_1, \omega_2, \omega_3)$. Only the genes selected in each dimension are displayed. Figure 5 provides an illustrative example of these types of figures in the case of sPLS. These graphical outputs proposed by González et al. (2008a) improve the

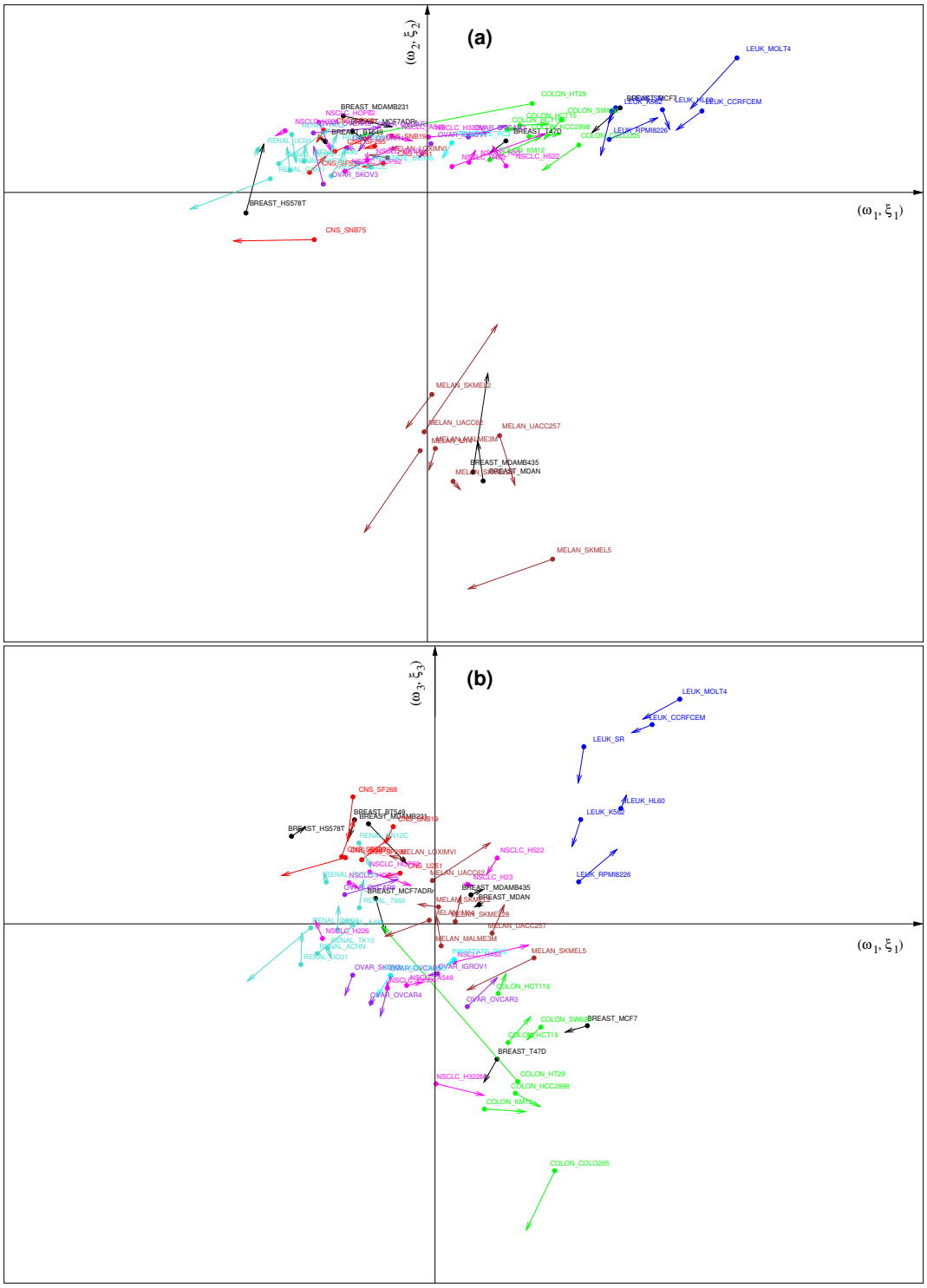

Figure 4: Graphical representations of the cell lines. sPLS latent variables are displayed in a superimposed manner, where the start of the arrow show the location of the Ross samples, and the tip the Staunton samples. The first and second axis (first and third) are shown in **(a)** and **(b)**.

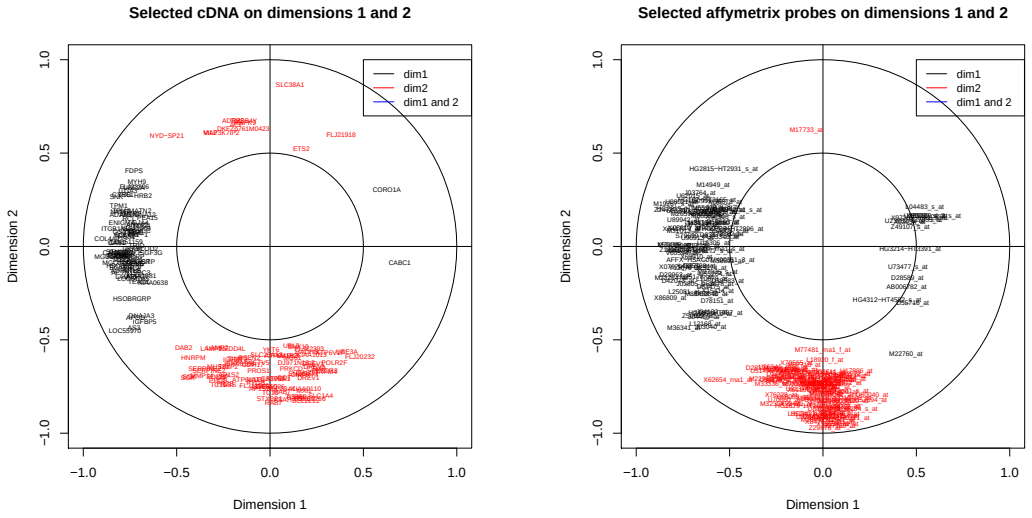

Figure 5: Graphical representation of the genes selected on the first two axes with sPLS. The coordinates of each gene are obtained by computing the correlation between (ξ_1, ξ_2) (resp. (ω_1, ω_2)) and the original Ross (resp. Staunton) data set. Selected cDNAs from the Ross data set (left) or selected Affymetrix probes from the Staunton data set (right) are displayed.

interpretability of the results in the following manner. First they allow for the identification of correlated gene subsets from each data set, which are either up or down regulated. Second they help revealing the correlation between gene subsets from both data sets (by superimposing both graphics). And third they help relating these correlated subsets to the associated tumor cell lines by combining the information contained in Fig. 5 and Fig. 4 **(a)**. For example, we can make the assumption that the genes which were selected on the second sPLS dimension for both data sets should help discriminating melanoma tumors from the other cell lines.
In our case, these types of graphics usually show that there is few overlap between the gene selections in dimensions 1, 2 or 3. This means that each selection focus on a specific aspect of the data set (a specific tumor), and that the loading vectors are orthogonal ($cor(a_s, a_r) = 0$, $cor(b_s, b_r) = 0$, $r < s$). This valuable property is still kept in the sparse methods (sPLS, CCA-EN), which is not often the case (see the sparse PCA approaches, Zou and Hastie, 2005; Jolliffe et al., 2003; Shen and Huang, 2007). This results in a very small overlap between each gene list from each CCA-EN or sPLS dimension (Table 2). In fact, only 0 to 2 genes are overlapping the dimensions 1-2 and 1-3 in X, and between 1 to 13 genes in Y for both approaches. On the other hand, there is no orthogonality between CIA loadings vectors, leading to a high number of genes that are overlapping.

Comparisons of the gene lists.

Based on the interpretation of the cell line clusters (paragraph 3.3), our analysis of gene clusters relied on three sets of gene lists (3 methods $\times$ 2 data sets = 6 lists of 100 genes per set):

- **Set 1:** the lists associated with the separation of cell lines with epithelial (mainly LE and CO) versus mesenchymal (mainly RE and CNS) characteristics (CCA-EN axis 2, CIA and sPLS axes 1),

Table 2: Number of genes commonly selected between all dimensions for each approach.

	X=Ross-cDNA				Y=Staunton-Affymetrix			
	dim 1-2	dim 1-3	dim 2-3	dim 1-2-3	dim 1-2	dim 1-3	dim 2-3	dim 1-2-3
CCA-EN	0	2	2	0	1	3	13	1
CIA	20	17	31	2	14	21	24	1
sPLS	0	0	2	0	0	8	1	0

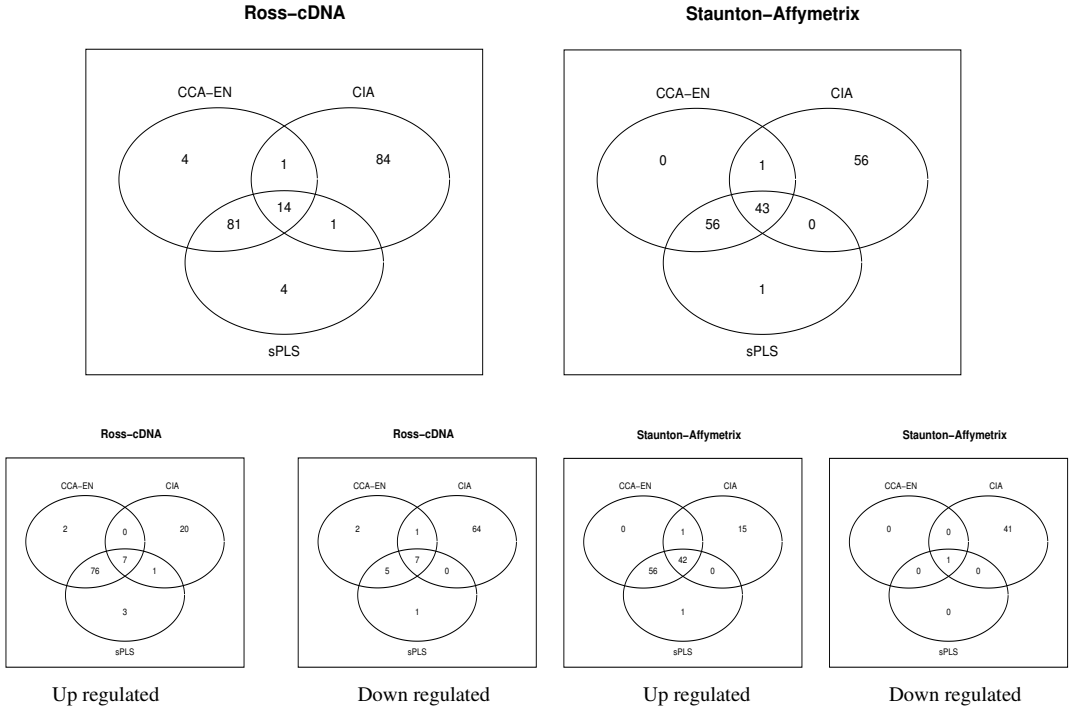

Figure 6: Venn diagrams for 100 selected genes associated to melanoma *vs.* the other cell lines for each data set (top) and by identifying up and down regulated genes in these lists (bottom).

- **Set 2:** the lists associated with the separation of the melanoma cell lines (ME, BR_MDAN and BR_MDAMB435) from the other cell lines (CCA-EN axis 1, CIA and sPLS axes 2),
- **Set 3:** the lists associated with the separation of the LE cell lines from the CO cell lines (axis 3 of each method).

First, we evaluated for each set of gene lists and for each data set the number of genes that were selected in common by the different methods. Figure 6 displays the Venn diagrams for the lists of genes associated with melanoma *vs.* the other cell lines in dimension 1 for CCA-EN and in dimension 2 for CIA and sPLS.
For each set of gene lists, the Venn diagrams revealed a very strong similarity between the gene selections obtained by CCA-EN and sPLS, whereas CIA seemed to select other genes linked to the cell lines. Note that the same trend was observed if more than 100 variables were selected on each dimension.

Second, we evaluated for each dimension from each method the degree of overlap between the two data sets. In fact, it could be expected from the canonical methods that they identify correlations between measurements obtained from the two platforms when these correspond to the same gene. To evaluate this aspect, the identifiers of the features from each platform were mapped to unique gene identifiers using Ingenuity Pathway Analysis Application (IPA, see Supplemental Table). For all three dimensions, CCA-EN and sPLS selected approximately 20 features from the Ross and Staunton data sets that corresponded to identical genes. On the other hand, CIA selected 15 to 17 genes that were common to the two data sets.

We obtained heatmaps (Sup. Fig. 7, 8 and 9) for each of the 18 gene lists. For all heatmaps, we used the clustering of the individuals obtained with the Ross and Staunton data sets, presented in Fig.2. These heatmaps illustrate well the general finding that CCA-EN and sPLS are most similar, and that CIA tends to select genes with a higher variance across all cell lines compared to CCA - EN and sPLS.
The 3 sets of gene lists were loaded into IPA along with their corresponding log ratios and we focused on 1) *biological functions* that were significantly over-represented (right-tailed Fisher's exact test) in the gene lists compared to the initial sets of genes (1,375 and 1,517 genes for the Ross and Staunton data sets respectively), 2) *canonical pathways* in which the genes from the lists were significantly over-represented compared to the genes in the initial sets and 3) the first *networks* generated by IPA from the gene lists. These networks are build by combining the genes into small (35 molecules maximum) networks that maximize their specific connectivity (Calvano et al., 2005) which result in highly-interconnected networks. The main results from these analyses are presented below for each set.

Set 1: Epithelial-Mesenchymal Transition (EMT). As previously described for a CIA analysis (Culhane et al., 2003), axes 1 (CIA and sPLS) or 2 (CCA-EN) of the 3 methods distinguished cell lines with epithelial characteristics (mainly CO and LE) from cell lines with stromal/mesenchymal characteristics (mainly RE and CNS). The epithelial to mesenchymal transition (EMT) is a key process underlying various tissue remodeling events during embryonic development. The EMT is thought to be also involved in establishing the metastatic potential of carcinoma cells (Yang and Weinberg, 2008). During the EMT, cells acquire

morphological and biochemical characteristics that enables them to limit their contacts with neighboring cells and to invade the extracellular matrix. Studying the events underlying this process is thus of primary importance to better understand tumor malignancy.
The most significant biological functions identified in common by the three methods ($p < 0.001$ for each method) were:

- Cellular movement, skeletal and muscular system development and function, tissue development, cell-to-cell signaling and interaction, cellular assembly and organization and cancer for the Ross data set
- Cell morphology, cellular movement, cell death, cancer, reproductive system disease, cell-to-cell signaling and interaction, connective tissue development and function, cellular function and maintenance, cardiovascular system development and function, renal and urological system development and function and cellular development for the Staunton data set

The lists of genes involved in these biological functions are available as Supplemental Table). First, this illustrates well the complementarity of the two data sets, which interrogate very different sets of genes (see Culhane et al., 2003 for such a comparison) and may thus identify complementary aspects of the same biological process. Second, most of the biological functions identified are highly relevant to the EMT transition which involves modifications of the connective tissue and of cell morphology, cell movement and cell-to-cell interactions in particular. Genes involved in skeletal and muscular system development were found to be more highly expressed in stromal/mesenchymal cell lines and is consistent with previous observations (Ross et al., 2000; Tso et al., 2006). Similarly, genes involved in the function "reproductive system disease" were mostly over expressed in stromal/mesenchymal cell lines and were mainly associated with breast cancer cell lines biological functions. This is consistent with the presence of most breast cancer cell lines on the stromal/mesenchymal side of the corresponding axes. Other biological functions were more specifically identified by CIA or CCA-EN/sPLS. Generally, the latter two methods identified the same biological functions, which is consistent with the similarity of their gene selections. However, CIA systematically identified (sometimes many) more highly significant biological functions compared to CCA-EN/sPLS (*e.g.* for the Ross data set, CCA-EN and sPLS identified 7 functions with $p < 0.001$ while CIA identified 21 functions using the same threshold). Since many of these functions were found significant for 2 to all 3 sets of gene lists, this likely reflects the redundancy in gene selections among the CIA axes. Thus, while some of these additional biological functions evidenced by CIA may be relevant, their interpretation may also be misled by less specific findings.
This hypothesis was strengthened when we focused on the canonical pathways identified by IPA analysis. CCA-EN and sPLS both found that the integrin and the actin cytoskeleton pathways contained a significantly higher number of genes that were over expressed in RE and CNS cell lines compared to LE and CO than could be expected by chance. This finding was consistent between the two data sets. These two central pathways in cell movement, which appear highly relevant to the EMT, displayed much higher p-values for the analysis of the gene lists selected by CIA. It is thus likely that less specific genes contained in the CIA gene selections limit the enrichment of a sufficient number of genes in a given pathway to yield low enough p-values.
Finally, the first networks identified by IPA for all three methods were highly connected and

were associated with cellular movement for both data sets and in addition with cell-to-cell signaling and interaction for the Ross data set. Interestingly, all six networks pointed to the ERK (extracellular-signal-regulated kinase) signaling pathway as a central player in the gene expression modulations that were selected, which is consistent with its known role in cell migration (Juliano et al., 2004). However, the CIA network for the Ross data set failed to identify the integrin pathway as an upstream regulator of ERK. Merging the first 3 networks from the 3 canonical methods for each data set yielded two highly similar networks (Supplemental Figures 10 and 11). However, only the network built from the Satunton data set highlighted the transforming growth factor$-\beta$ (TGF$-\beta$) pathway which is thought to be a primary inducer of the EMT (Yang and Weinberg, 2008). Despite this difference, the most connected nodes (including integrins α and β alpha-actinin, connective tissue growth factor, fibronectin 1, SERPINE1, plasminogen activator urokinase, Ras or ERK) were found in both networks. These likely represent central players in establishing the different phenotypes of LE and CO cell lines on one hand and of RE and CNS cell lines on the other hand.

Set 2: Melanoma cell lines. Axis 1 of CCA and axes 2 of CIA and PLS clearly separate all except one (LOXIMVI) melanoma cell lines, along with the melanoma metastasis BR_MDAN and BR_MDAMB435 from all other cell lines. The melanoma cell line LOXIMVI has previously been shown to lack melanine and several typical markers of melanoma cells (Stinson et al., 1992), which likely explains its absence in the cluster of ME cell lines. For these axes, the selections made by CCA-EN and sPLS are almost identical for the two data sets (only 1 and 5 genes specific to each method for the Staunton and the Ross data sets respectively).
For this cluster, less significant biological functions were identified compared to Set 1 and these differed substantially between CIA and the other two methods. The most significant biological functions ($p < 0.001$ for both methods) identified by CCA-EN/ sPLS were:

- Molecular transport, amino acid metabolism and small molecule biochemistry for the Ross data set
- Hair and skin development and function, amino acid metabolism, cellular development, small molecule biochemistry, cell morphology, dermatological diseases and conditions, nervous system development and function

On the other hand, CIA identified the following significant biological functions ($p < 0.001$):

- Cancer, reproductive system disease, cellular movement and cell morphology for the Ross data set
- Cellular growth and proliferation, cancer, hair and skin development and function, reproductive system disease, amino acid metabolism, cell morphology, cellular assembly and organization, ophthalmic disease and small molecule biochemistry for the Staunton data set

As for Set 1, CIA identified more biological functions than CCA-EN/sPLS but some, such as "cancer", appear less specific and are common to all three sets of gene lists. Overall, the biological functions identified by the three methods appear relevant to the characterization of melanoma cell lines, particularly those related to skin biology. The categories related to amino

acid metabolism (including small molecule biochemistry and molecular transport which contains many genes involved in amino acid transport and metabolism) are likely found because ME cell lines are characterized by melanin synthesis which involves the amino acids tyrosine and cysteine.
Similarly to Set 1, CCA-EN/ sPLS identified more significant canonical pathways compared to CIA which allowed a more precise understanding of the gene lists selected by these two methods. In particular, CCA-EN/sPLS identified glycosphingolipid biosynthesis pathways from both the Ross (ganglioside biosynthesis only) and the Staunton data sets (ganglioside and globosid biosynthesis pathways). Melanoma tumors are known to be rich in these glycosphingolipids (Portoukalian et al., 1979). Indeed, their presence at the cell membrane makes them interesting targets for immunotherapy and vaccination strategies (Fredman et al., 2003). Noticeably, the tyrosine metabolism pathway was identified by all three methods ($p < 0.05$) in the Staunton data set but only by CCA-EN/sPLS in the Ross data set ($p < 0.05$). Genes involved in this pathway included tyrosinase, tyrosinase related proteins 1 and 2 and dopachrome tautomerase which are all involved in melanin biosynthesis and were found over expressed in melanoma cell lines accordingly.
Finally, the first networks generated by IPA from the CCA-EN/sPLS and the CIA gene lists pointed to differential activities or expression of several components of signaling pathways including TGF-β, PDGF, TNF, Mek, Erk, Mapk, Ras, PKA, PKCδ, Jnk, AP1, PI3K or Akt in melanoma cell lines compared to the other cell lines. These networks, especially those obtained from the Staunton data set, also highlighted several markers used for the diagnosis of melanomas including the over expressed MITF, Vimentin, S-100A1, S-100B and Melan-A and the under expressed keratins 7, 8, 18 and 19.

Set 3: Leukemia cell lines compared to colon tumor cell lines. The axes 3 from each of the three canonical methods separated the LE from the CO cell lines highlighting that these two groups could also be distinguished through gene expression profiles of selected genes from both data sets.
For the Ross data set, CIA found only one significant biological function (tissue development) that had not been found significant at the 0.001 threshold for Sets 2 and 3. Most of the genes in this category were expressed at lower levels in LE compared to CO cell lines and were implicated in the adhesion of epithelial cells or tissue and in the formation and assembly of extracellular matrix. CCA-EN and sPLS identified the hematological and immunological disease categories as relevant biological functions that separate the LE from the CO cell lines for the Ross data set. In addition, they identified the cell death category that was also found for the Staunton gene lists of Set 1 but the genes implicated in this biological function were almost completely different between Set 1 and Set 3. For the Staunton data set, CCA-EN alone identified a set of genes implicated in embryonic development that were over expressed in CO cell lines compared to LE cell lines (except CXCR4 that was over expressed in LE compared to CO). Interestingly, all three methods identified a set of three genes implicated in severe combined immunodeficiency (CD3D, IL2RG and ZAP70) that were up regulated in LE compared to CO cells.
Surprisingly, CIA seemed to identify many more canonical pathways for the Ross data set compared to CCA-EN and sPLS. Indeed these were all specific metabolic pathways involving the same three isoforms of poorly specific aldehyde dehydrogenase. sPLS alone identified the tight junction signaling pathway which included in particular Claudin 4 (CLDN4) and

Zona occludens 1 (ZO1) that are strongly expressed in CO cell lines but not in LE cell lines and are key components of the tight junctions between epithelial cells. A similar bias in canonical pathway identification was observed for the Staunton data set for which CCA-EN and sPLS had selected two aldehyde dehydrogenases along with other enzymes involved in several metabolic pathways.
The first networks found by IPA for the Ross data set were mainly focused on genes involved in cell-to-cell signaling and interaction and in cellular movement, assembly and organization. In particular, most of these genes were components of the cytoskeleton, of the basement membrane or of cell-cell junctions. They were also involved in cell-cell contacts or in cell migration and adhesion. Most of them were expressed at much higher levels in CO versus LE cell lines, which is consistent with the typical epithelial characteristics of the colon tumor cell lines compared to the leukemia cell lines. For the Staunton data set, the first networks identified by IPA were also mainly focused on cell-to-cell signaling and interaction and on cellular movement. Overall, these results highlighted the fact that the CO cell lines are much more characteristic of an epithelium than the LE cell lines.

Conclusion

The analysis of the NCI60 data sets with CCA-EN, CIA and sPLS evidenced the main differences between these methods.

CIA. CIA does not propose a built-in variable selection procedure and requires a two-step analysis to perform variable selection. The main individual effects were identified. However, the loadings or weight vectors obtained were not orthogonal, in contrary to CCA-EN and sPLS. This resulted in some redundancy in the gene selections on the first three axes, which may be a limitation for the biological interpretation, as there may be less specific genes related to some cell lines types that were identified.
The gene selections obtained on each dimension generally led to interpretations that were overall similar to those obtained with CCA-EN and sPLS. However, the interpretations of the gene selections were clearly affected by genes selected on several axes, leading to less specific results.

CCA-EN. CCA-EN first captured the main robust effect on the individuals that is present in the two data sets. Consequently, it may hide strongest individual effects that are present in only one data set, but bring robust results.
We observed a strong similarity between CCA-EN and sPLS in the gene selections, except that the axes were permuted. In fact, we believe that CCA-EN can be considered as a sparse PLS variant with a canonical mode. Indeed, the elastic net is approximated with a univariate threshold, similar to a soft-thresholding penalization, and the whole algorithm uses PLS and not CCA computations, which explains why the canonical correlations do not monotonically decrease. The only difference that distinguishes sPLS canonical mode from CCA-EN is the initialization of the algorithm for each dimension. CCA-EN maximizes the correlation between the latent variables, whereas sPLS maximizes the covariance.

sPLS. We found that sPLS makes a good compromise between all these approaches. It includes variable selection and the loading vectors are orthogonal. Apart from the fact that

sPLS and CCA-EN do not order the axis in the same manner, both approaches were highly comparable, except for slight but significant differences when studying LE *vs.* CO (axes 3). In this particular case, the resulting gene lists clearly provided complementary information.

We believe that all approaches are easy to use and fast to compute. These approaches would benefit from the development of an `R` package that could harmonize their inputs and outputs to facilitate their use and their comparison. Based on the present study, we would primarily recommend the use of CCA-EN or sPLS when gene selection is an issue. Like CCA-EN, sPLS includes a built-in variable selection procedure but captured subtle individual effects. Therefore, the choice of one of these methods would take into consideration the fundamental difference between them in the building of the first axes.

References

Butte, A., Tamayo, P., Slonim, D., Golub, T., and Kohane, I. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proceedings of the National Academy of Sciences, page 220392197.

Bylesjö, M., Eriksson, D., Kusano, M., Moritz, T., and Trygg, J. (2007). Data integration in plant biology: the o2pls method for combined modeling of transcript and metabolite data. The Plant Journal, 52:1181–1191.

Calvano, S., Xiao, W., Richards, D., Felciano, R., Baker, H., Cho, R., Chen, R., Brownstein, B., Cobb, J., Tschoeke, S., et al. (2005). A network-based analysis of systemic inflammation in humans. NATURE-LONDON-, 437(7061):1032.

Combes, S., González, I., Déjean, S., Baccini, A., Jehl, N., Juin, H., Cauquil, L., and Batrice Gabinaud, Franois Lebas, C. L. (2008). Relationships between sensorial and physicochemical measurements in meat of rabbit from three different breeding systems using canonical correlation analysis. Meat Science, in press.

Culhane, A., Perriere, G., and Higgins, D. (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. BMC Bioinformatics, 4(1):59.

de Jong, S. (1993). Simpls: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 18:251–263.

Doledec, S. and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology, 31(3):277–294.

Fredman, P., Hedberg, K., and Brezicka, T. (2003). Gangliosides as Therapeutic Targets for Cancer. BioDrugs, 17(3):155.

Gittins, R. (1985). Canonical Analysis: A Review with Applications in Ecology. Springer-Verlag.

González, I., Déjean, S., Martin, P., Goncalves, O., Besse, P., and Baccini, A. (2008a). Highlighting Relationships Between Heteregeneous Biological Data Through Graphical Displays Based On Regularized Canonical Correlation Analysis. Technical report, Université de Toulouse.

González, I., Déjean, S., Martin, P. G. P., and Baccini, A. (2008b). Cca: An r package to extend canonical correlation analysis. Journal of Statistical Software, 23(12).

Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28:321–377.

Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A Modified Principal Component Technique Based on the LASSO. Journal of Computational & Graphical Statistics, 12(3):531–547.

Juliano, R., Reddig, P., Alahari, S., Edin, M., Howe, A., and Aplin, A. (2004). Integrin regulation of cell signalling and motility. Biochem Soc Trans, 32:443–446.

Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). Sparse PLS: Variable Selection when Integrating Omics data. Technical report, Université de Toulouse et Institut National de la Recherche Agronomique.

Lorber, A., Wangen, L., and Kowalski, B. (1987). A theoretical foundation for the PLS algorithm. Journal of Chemometrics, 1(19-31):13.

Portoukalian, J., Zwingelstein, G., and Dore, J. (1979). Lipid composition of human malignant melanoma tumors at various levels of malignant growth. FEBS Journal, 94(1):19.

Robert, P. and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. Applied Statistics, 25(3):257–265.

Ross, D., Scherf, U., Eisen, M., Perou, C., Rees, C., Spellman, P., Iyer, V., Jeffrey, S., Van de Rijn, M., Waltham, M., et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet, 24(3):227–35.

Scherf, U., Ross, D., Waltham, M., Smith, L., Lee, J., Tanabe, L., Kohn, K., Reinhold, W., Myers, T., Andrews, D., et al. (2000). A gene expression database for the molecular pharmacology of cancer. Nat Genet, 24(3):236–244.

Shen, H. and Huang, J. Z. (2007). Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis, to appear.

Staunton, J., Slonim, D., Coller, H., Tamayo, P., Angelo, M., Park, J., Scherf, U., Lee, J., Reinhold, W., Weinstein, J., et al. (2001). Chemosensitivity prediction by transcriptional profiling. Proceedings of the National Academy of Sciences, 98(19):10787.

Stinson, S., Alley, M., Kopp, W., Fiebig, H., Mullendore, L., Pittman, A., Kenney, S., Keller, J., and Boyd, M. (1992). Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. Anticancer Res, 12(4):1035–53.

Tenenhaus, M. (1998). La régression PLS: théorie et pratique. Editions Technip.

Thioulouse, J., Chessel, D., Dole´ dec, S., and Olivier, J. (1997). ADE-4: a multivariate analysis and graphical display software. Statistics and Computing, 7(1):75–83.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58(1):267–288.

Tso, C., Shintaku, P., Chen, J., Liu, Q., Liu, J., Chen, Z., Yoshimoto, K., Mischel, P., Cloughesy, T., Liau, L., et al. (2006). Primary Glioblastomas Express Mesenchymal Stem-Like Properties. Molecular Cancer Research, 4(9):607.

Vijayendran, C., Barsch, A., Friehs, K., Niehaus, K., Becker, A., and Flaschel, E. (2008). Perceiving molecular evolution processes in Escherichia coli by comprehensive metabolite and gene expression profiling. Genome Biology, 9(4):R72.

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. Journal of Econometrics, 4(2):147–166.

Waaijenborg, S., de Witt Hamer, V., Philip, C., and Zwinderman, A. (2008). Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. Statistical Applications in Genetics and Molecular Biology, 7(1):3.

Wegelin, J. (2000). A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical Report 371, Department of Statistics, University of Washington, Seattle.

Wold, H. (1966). Multivariate Analysis. Academic Press, New York, Wiley, krishnaiah, p.r. (ed.) edition.

Wold, S., Eriksson, L., Trygg, J., and Kettaneh, N. (2004). The PLS method–partial least squares projections to latent structures–and its applications in industrial RDP (research, development, and production). Technical report, Umea University.

Yang, J. and Weinberg, R. (2008). Epithelial-Mesenchymal Transition: At the Crossroads of Development and Tumor Metastasis. Developmental Cell, 14(6):818–829.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B, 67(2):301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. Journal of Computational and Graphical Statistics, 15(2):265–286.

Appendix

A PLS algorithm (canonical mode)

1. $X_0 = X$, $Y_0 = Y$
2. For $h = 1 \dots H$:
   (a) Initialize
   ξ_h = first column of X_{h-1} $\quad \omega_h$ = first column of Y_{h-1}
   (b) Until convergence of a_h:
      i. $a_h = X'_{h-1}\xi_h / \xi'_h\xi_h$, norm a_h
      ii. $\xi_h = X_{h-1}a_h$, norm ξ_h
      iii. $b_h = Y'_{h-1}\xi_h / \xi'_h\xi_h$, norm b_h
      iv. $\omega_h = Y_{h-1}b_h$, norm ω_h
   (c) $c_h = X'_{h-1}\xi_h \qquad e_h = Y'_{h-1}\omega_h$
   (d) $X_h = X_{h-1} - \xi_h c'_h \quad Y_h = Y_{h-1} - \omega_h e'_h$

Step (c) computes the regression coefficients of the matrices X_{h-1} and Y_{h-1} on the latent variables ξ_h and ω_h.
Step (d) computes the deflated (residual) matrices.

B sparse PLS algorithm (canonical mode)

Sparse PLS initializes step (a) in PLS by extracting the first pair of singular vectors (a_h, b_h) of the crossproduct $X'_{h-1}Y_{h-1}$, which includes variation in both X and Y and the correlation between the two sets.
The two loading vectors (a_h, b_h) are then computed with penalizations λ_a and λ_b in step (b), and the latent vectors (ξ_h, ω_h) are then computed, where $g_\lambda(y) = sign(y)(|y| - \lambda)_+$ is the soft-thresholding penalty function.

1. $X_0 = X \qquad Y_0 = Y$
2. For $h = 1 \dots H$:
   (a) Set $\tilde{M}_{h-1} = X'_{h-1}Y_{h-1}$, decompose $\tilde{M}_{h-1}$ and extract the first pair of singular vectors $a_{old} = a_h$ and $b_{old} = b_h$
   (b) Until convergence of a_{new} and b_{new}:
      i. $a_{new} = g_{\lambda_a}(\tilde{M}_{h-1}b_{old})$, norm a_{new}
      ii. $b_{new} = g_{\lambda_b}(\tilde{M}'_{h-1}a_{old})$, norm b_{new}
      iii. $a_{old} = a_{new}$, $b_{old} = b_{new}$
   (c) $\xi_h = X_{h-1}a_{new}$
   $\omega_h = Y_{h-1}b_{new}$
   (d) $c_h = X'_{h-1}\xi_h \quad e_h = Y'_{h-1}\omega_h$
   (e) $X_h = X_{h-1} - \xi_h c'_h \quad Y_h = Y_{h-1} - \omega_h e'_h$

C Canonical Correlation Analysis with Elastic Net penalization

CCA-EN initializes step (a) in PLS by setting $\xi_h = X^j_{h-1}$ and $\omega_h = Y^k_{h-1}$ such that $cor(X^j_{h-1}, Y^k_{h-1})$ is maximized, for $j = 1 \dots p$ and $k = 1 \dots q$. Hence, this algorithm aims at maximizing the correlation (rather than the covariance for PLS and sPLS).
The approximation on Elastic Net penalization finally consists in introducing soft-thresholding penalizations, as in sparse PLS, which makes both algorithms similar, except for the initilization step.

1. $X_0 = X \qquad Y_0 = Y$
2. For $h = 1 \dots H$:
   (a) Set $\xi_h = X^j_{h-1}$ and $\omega_h = Y^k_{h-1}$ such that $cor(X^j_{h-1}, Y^k_{h-1})$ is maximized
   $a_{new} = X'_{h-1}\xi_h / \xi'_h\xi_h \quad b_{new} = Y'_{h-1}\xi_h / \xi'_h\xi_h$, norm a_{new} and b_{new}
   (b) Until convergence of a_{new} and b_{new}:
      i. $a_{new} = g_{\lambda_a}(Y_{h-1}b_{old})$, norm a_{new}
      ii. $b_{new} = g_{\lambda_b}(X_{h-1}a_{old})$, norm b_{new}
      iii. $a_{old} = a_{new}$, $b_{old} = b_{new}$
   (c) $\xi_h = X_{h-1}a_{new}$, norm ξ_h
   $\omega_h = Y_{h-1}b_{new}$ norm ω_h
   (d) $c_h = X'_{h-1}\xi_h \quad e_h = Y'_{h-1}\omega_h$
   (e) $X_h = X_{h-1} - \xi_h c'_h \quad Y_h = Y_{h-1} - \omega_h e'_h$

D Supplemental figures

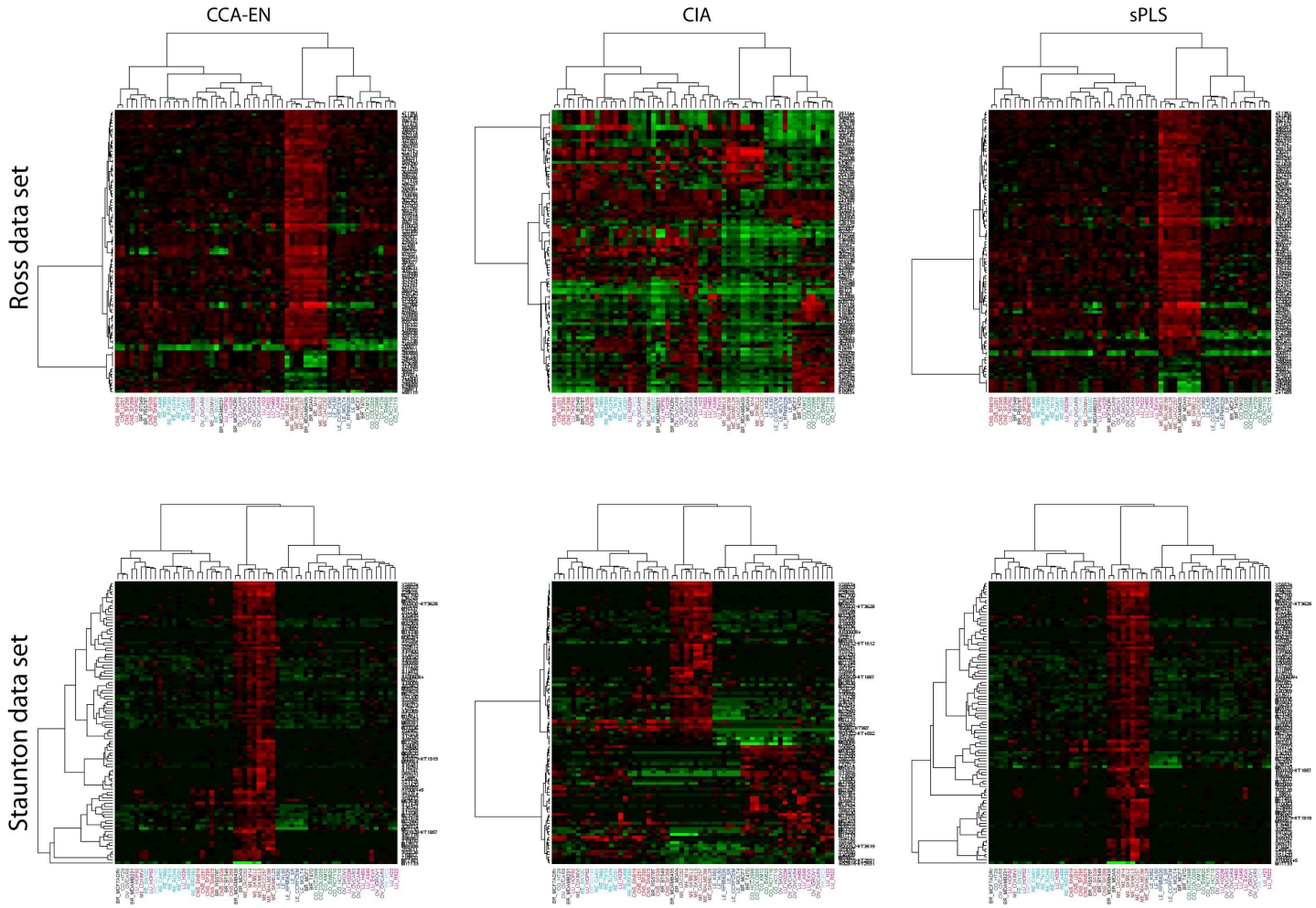

Figure 7: Heat map displays of hierarchical clustering results with the Ward method and correlation distance with genes in lines and cell lines in columns. The red (green) colour represents over-expressed (under-expressed) genes. Selection of 100 genes on dimension 2 (resp. 1) for CIA and sPLS (resp. CCA-EN) that separate melanoma *vs.* the other cell lines.

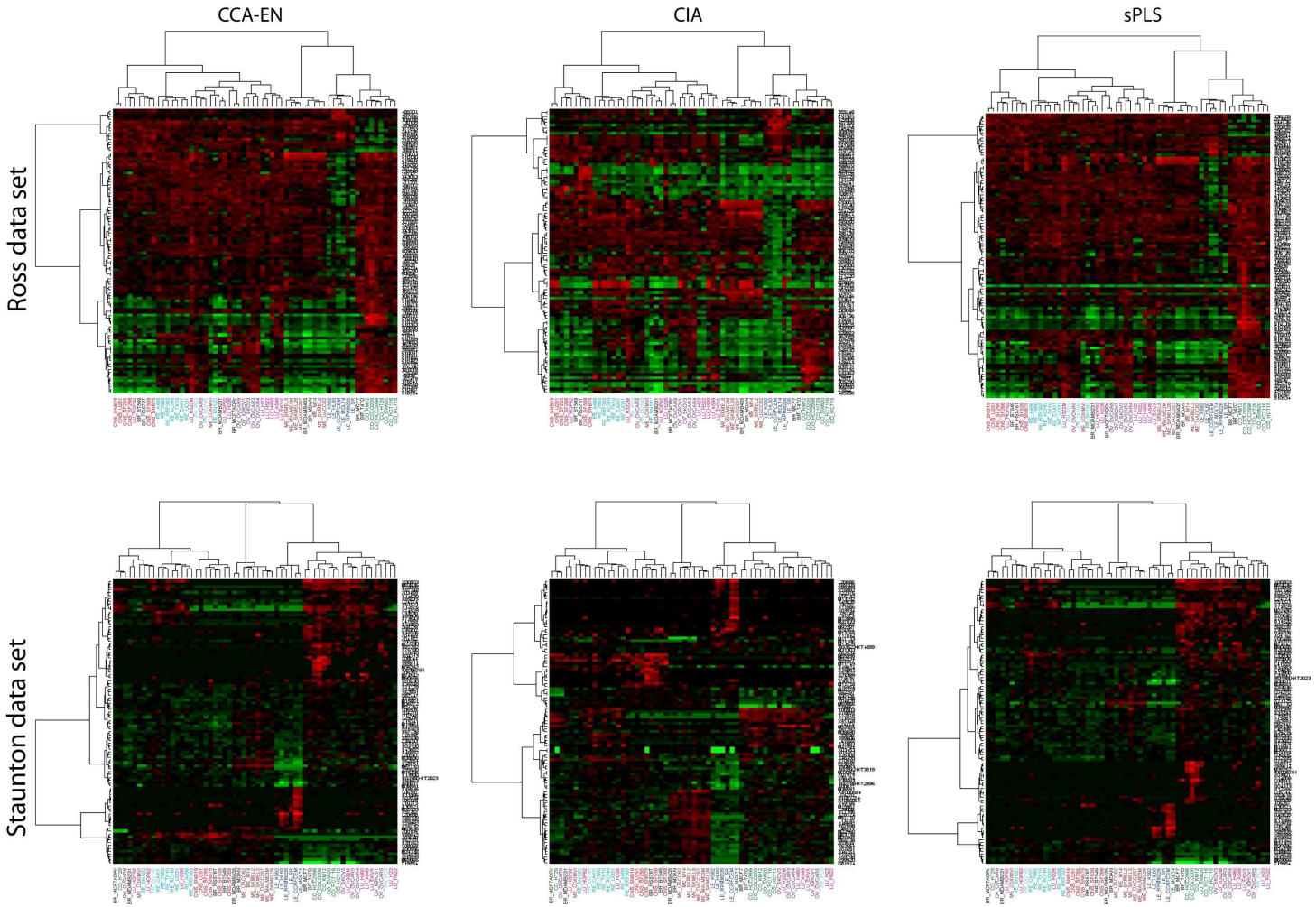

Figure 8: Heat map displays of hierarchical clustering results with the Ward method and correlation distance with genes in lines and cell lines in columns. The red (green) colour represents over-expressed (under-expressed) genes. Selection of 100 genes on dimension 3 (resp. 2) for CIA and sPLS (resp. CCA-EN) that separate LE *vs.* CO.

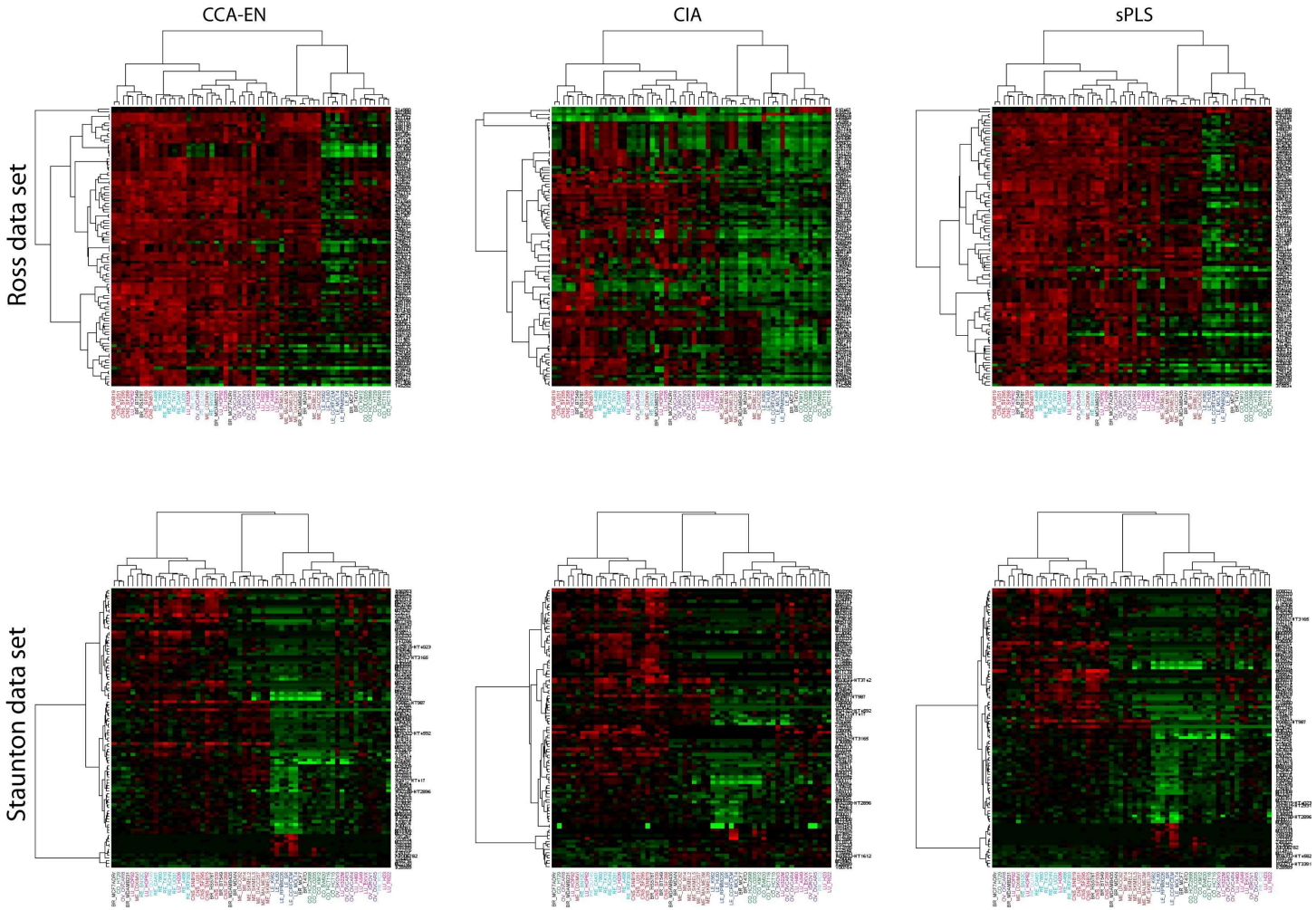

Figure 9: Heat map displays of hierarchical clustering results with the Ward method and correlation distance with genes in lines and cell lines in columns. The red (green) colour represents over-expressed (under-expressed) genes. Selection of 100 genes on dimension 1 (resp. 2) for CIA and sPLS (resp. CCA-EN) that separate epithelial-like *vs.* mesenchymal-like.

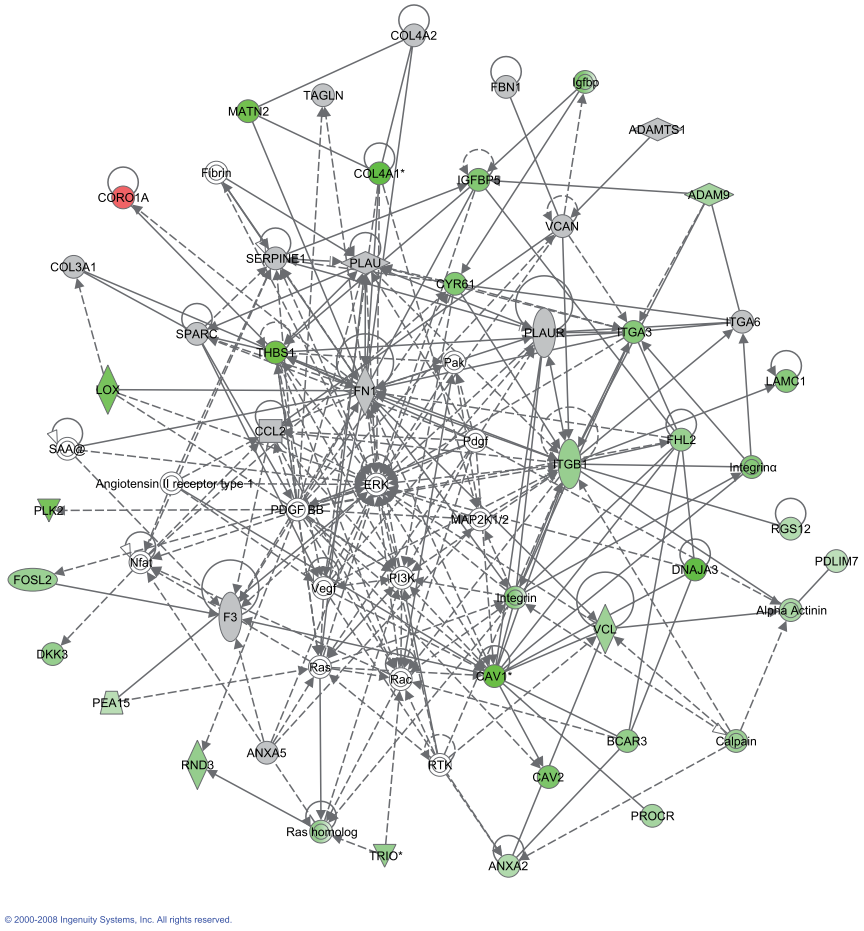

Figure 10: Molecular network obtained from the Ross gene lists from Set 1. For each canonical method (CCA-EN, CIA or sPLS), molecular networks were built from the Ross gene lists (focus genes) of Set 1 using Ingenuity Pathways Analysis (IPA, `www.ingenuity.com`). The first networks obtained from each method were merged into the presented network. Green and red colors indicate under- and over-expressions respectively in the LE/CO cell lines compared to the RE/CNS cell lines. Only the genes selected by sPLS have been colored in red or green. Genes colored in grey have been selected by CCA-EN or sPLS and all correspond to genes that are under-expressed in the LE/CO cell lines compared to the RE/CNS cell lines. Genes in white have been added by IPA based on their high connectivity with focus genes.

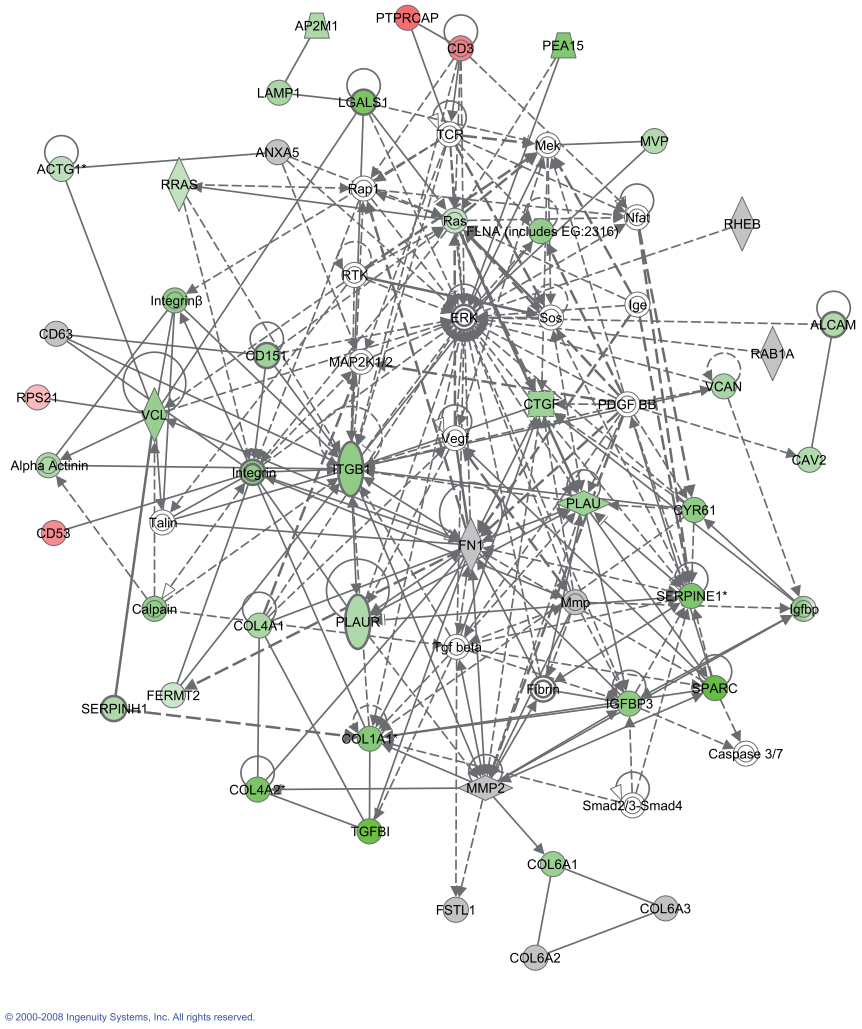

Figure 11: Molecular network obtained from the Staunton gene lists from Set 1. For each canonical method (CCA-EN, CIA or sPLS), molecular networks were built from the Staunton gene lists (focus genes) of Set 1 using Ingenuity Pathways Analysis (IPA, `www.ingenuity.com`). The first networks obtained from each method were merged into the presented network. Green and red colors indicate under- and over-expressions respectively in the LE/CO cell lines compared to the RE/CNS cell lines. Only the genes selected by sPLS have been colored in red or green. Genes colored in grey have been selected by CCA-EN or sPLS and all correspond to genes that are under-expressed in the LE/CO cell lines compared to the RE/CNS cell lines. Genes in white have been added by IPA based on their high connectivity with focus genes.