Motivation Biomarker discovery from high-dimensional data is a crucial problem with enormous applications in biology and medicine. It is also extremely challenging from a statistical viewpoint, but surprisingly few studies have investigated the relative strengths and weaknesses of the plethora of existing feature selection methods.

Methods We compare 32 feature selection methods on 4 public gene expression datasets for breast cancer prognosis, in terms of predictive performance, stability and functional interpretability of the signatures they produce.

Results We observe that the feature selection method has a significant influence on the accuracy, stability and interpretability of signatures. Simple filter methods generally outperform more complex embedded or wrapper methods, and ensemble feature selection has generally no positive effect. Overall a simple Student's t-test seems to provide the best results.

Availability Code and data are publicly available at http://cbio.ensmp.fr/∼ahaury/.

1 Introduction

Biomarker discovery from high-dimensional data, such as transcriptomic or SNP profiles, is a crucial problem with enormous applications in biology and medicine, such as diagnosis, prognosis, patient stratification in clinical trials or prediction of the response to a given treatment. Numerous studies have for example investigated so-called *molecular signatures*, i.e., predictive models based on the expression of a small number of genes, for the stratification of early breast cancer patients into low-risk or high-risk of relapse, in order to guide the need for adjuvant therapy (

,).

While predictive models could be based on the expression of more than a few tens of genes, several reasons motivate the search for short lists of predictive genes. First, from a statistical

*To whom correspondance should be addressed: 35, rue Saint Honoré, F-77300 Fontainebleau, France.

and machine learning perspective, restricting the number of variables is often a way to reduce over-fitting when we learn in high dimension from few samples and can thus lead to better predictions on new samples. Second, from a biological viewpoint, inspecting the genes selected in the signature may shed light on biological processes involved in the disease and suggest novel targets. Third, and to a lesser extent, a small list of predictive genes allows the design of cheap dedicated prognostic chips.

Published signatures share, however, very few genes in common, raising questions about their biological significance (Ioannidis, 2005). Independently of differences in cohorts or technologies, Ein-Dor *et al.* (2005) and Michiels *et al.* (2005) demonstrate that a major cause for the lack of overlap between signatures is that many different signatures lead to similar predictive accuracies, and that the process of estimating a signature is very sensitive to the samples used in the phase of gene selection. Specifically, Ein-Dor *et al.* (2006) suggest that many more samples than currently available would be required to reach a descent level of signature stability, meaning in particular that no biological insight should be expected from the analysis of current signatures. On the positive side, some authors noticed that the biological functions captured by different signatures are similar, in spite of the little overlap between them at the gene level (Shen *et al.*, 2008; Reyal *et al.*, 2008; Wirapati *et al.*, 2008).

From a machine learning point of view, estimating a signature from a set of expression data is a problem of *feature selection*, an active field of research in particular in the high-dimensional setting (Guyon and Elisseeff, 2003). While the limits of some basic methods for feature selection have been highlighted in the context of molecular signatures, such as gene selection by Pearson correlation with the output (Ein-Dor *et al.*, 2006), there are surprisingly very few and only partial investigations that focus on the *influence of the feature selection method* on the performance and stability of the signature. Lai *et al.* (2006) compared various feature selection methods in terms of predictive performance only, and Abeel *et al.* (2010) suggest that ensemble feature selection improves both stability and accuracy of SVM recursive feature elimination (RFE), without comparing it with other methods. However, it remains largely unclear how "modern" feature selection methods such as the elastic net (Zou and Hastie, 2005), SVM RFE or stability selection (Meinshausen and Bühlmann, 2010) behave in these regards and how they compare to more basic univariate techniques.

Here we propose an empirical comparison of a panel of feature selection techniques in terms of accuracy and stability, both at the gene and at the functional level. Using four breast cancer datasets, we observe significant differences between the methods. Surprisingly, we find that ensemble feature selection, i.e., combining multiple signatures estimated on random subsamples, has generally no positive impact, and that simple filters can outperform more complex wrapper or embedded methods.

2 Methods

2.1 Feature selection methods

We compare eight common feature selection methods to estimate molecular signatures. All methods take as input a matrix of gene expression data for a set of samples from two categories (good and bad prognosis in our case), and return a set of genes of a user-defined size s. These genes can then be used to estimate a classifier to predict the class of any sample from the expression values of these genes only. Feature selection methods are usually classified into three categories (Kohavi and John, 1997; Guyon and Elisseeff, 2003): *filter methods* select subsets of variables as a pre-processing step, independently of the chosen predictor; *wrapper methods* utilize the learning machine of interest as a black box to score subsets of variable according to

their predictive power; finally, *embedded methods* perform variable selection in the process of training and are usually specific to given learning machines. We have selected popular methods representing these three classes, as described below.

2.1.1 Filter methods

Univariate filter methods rank all variables in terms of relevance, as measured by a score which depends on the method. They are simple to implement and fast to run. To obtain a signature of size s, one simply takes the top s genes according to the score. We consider the following four scoring functions to rank the genes: the *Student's t-test* and *Wilcoxon sum-rank test*, which evaluate if each feature is differentially expressed between the two classes; and the *Bhattacharyya distance* and *relative entropy* to calculate a distance between the distributions of the two groups. We used the MATLAB Bioinformatics toolbox to compute these scoring functions.

2.1.2 Wrapper methods

Wrapper methods attempt to select jointly sets of variables with good predictive power for a predictor. Since testing all combinations or variables is computationally impossible, wrapper methods usually perform a greedy search in the space of sets of features. We test *SVM recursive feature elimination (RFE)* (Guyon *et al.*, 2002), which starts with all variables and iteratively removes the variables which contribute least to a linear SVM classifier trained on the current set of variables. We remove 20% of features at each iteration until s remain, and then remove them one by one in order to rigourously rank the first s. Following (Abeel *et al.*, 2010), we set the SVM parameter C to 1, and checked afterwards that other values of C did not have a significant influence on the results. Alternatively, we test a *Greedy Forward Selection (GFS)* strategy for least squares regression also termed Orthogonal Matching Pursuit, where we start from no variable and add them one by one by selecting each time the one which minimizes the sum of squares, in a 3-fold internal cross-validation setting. This algorithm was implemented in the SPAMS toolbox for Matlab initially published along with Mairal *et al.* (2010).

2.1.3 Embedded methods

Embedded methods are learning algorithms which perform feature selection in the process of training. We test the popular *Lasso* regression (Tibshirani, 1996), where a sparse linear predictor $\beta \in \mathbb{R}^p$ is estimated by minimizing the objective function $R(\beta) + \lambda \|\beta\|_1$, where $R(\beta)$ is the mean square error on the training set (considering the two categories as ± 1 values) and $\|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|$. λ controls the degree of sparsity of the solution, i.e., the number of features selected. We fix λ as the smallest value which gives a signature of the desired size s. Alternatively, we tested the elastic net (Zou and Hastie, 2005), which is similar to the Lasso but where we replace the ℓ_1 norm of β by a combination of the ℓ_1 and ℓ_2 norms, i.e., we minimize $R(\beta) + \lambda \|\beta\|_1 + \lambda/2 \|\beta\|_2^2$ and $\|\beta\|_2^2 = \sum_{i=1}^{p} \beta_i^2$. By allowing the selection of correlated predictive variables, the elastic net is supposed to be more robust than the Lasso while still selecting predictive variables. Again, we tune λ to achieve a user-defined level of sparsity. For both algorithms, we used the code implemented in the SPAMS toolbox.

2.2 Ensemble feature selection

Many feature selection methods are known to be sensitive to small perturbations of the training data, resulting in unstable signatures. In order to "stabilize" variable selection, several authors have proposed to use ensemble feature selection on bootstrap samples: the variable

selection method is run on several random subsamples of the training data, and the different lists of variables selected are merged into a hopefully more stable subset (Bi *et al.*, 2003; Meinshausen and Bühlmann, 2010; Abeel *et al.*, 2010).

For each feature selection method described above, we tested in addition the following three aggregation strategies for ensemble feature selection. We first bootstrap the training samples $B = 50$ times (i.e., draw a sample of size n from the data with replacement B times) to get B rankings $(r^1 ... r^B)$ of all features by applying the feature selection method on each sample. For filter methods, the ranking of features is naturally obtained by decreasing score. For RFE and GFS, the ranking is the order in which the features are added or removed in the iterative process. For Lasso and elastic net, the ranking is the order in which the variables become selected when λ decreases. We then aggregate the B lists by computing a score $S_j = 1/B \sum_{b=1}^{B} f(r_j^b)$ for each gene j as an average function of its rank r_j^b in the b-th bootstrap experiment. We test the following functions of the rank for aggregation:

- *Ensemble-mean (Abeel* et al.*, 2010)*: we simply average the rank of a gene over the bootstrap experiments, i.e., we take $f(r) = r$.
- *Ensemble-stability selection (Meinshausen and Bühlmann, 2010)*: we measure the percentage of bootstrap samples fro which the gene ranks in the top s, i.e., $f(r) = 1$ if $r \leq s$, 0 otherwise.
- *Ensemble-exponential*: we propose a soft version of stability selection, where we average an exponentially decreasing function of the rank, namely $f(r) = \exp -r/s$.

Finally, for each rank aggregation strategy, the aggregated list is the set of s genes with the largest score.

2.3 Accuracy of a signature

In order to measure the predictive accuracy of a feature selection method, we assess the performance of various supervised classification algorithms trained on the data restricted to the selected signature. More precisely, we tested 5 classification algorithms: nearest centroids (NC), k-nearest neighbors (KNN) with $k = 9$, linear SVM with $C = 1$, linear discriminant analysis (LDA) and naive Bayes (BAYES). The parameters of the KNN and SVM methods were fixed to arbitrary default values, and we checked that no significantly better results could be obtained with other parameters by testing a few other parameters. We assess the performance of a classifier by the area under the ROC curve (AUC), in two different settings. First, on each dataset, we perform a 10-fold cross-validation (CV) experiment, where both feature selection and training of the classifier are performed on 90% of the data, and the AUC is computed on the remaining 10% of the data. This is a classical way to assess the relevance of feature selection of a given dataset. Second, to assess the performance of the signature across datasets, we estimate a signature on one dataset, and assess its accuracy on other datasets by again running a 10-fold CV experiment where only the classifier (restricted to the genes in the signature) is retrained on each training set. In both cases, we report the mean AUC across the folds and datasets, and assess the significance of differences between methods with a two-sided paired t-test.

2.4 Stability of a signature

To assess the stability of feature selection methods, we compare signatures estimated on different samples in various settings. First, to evaluate stability with respect to small perturbation of the training set, we randomly subsample each dataset into pairs of subsets with 80% of sample

overlap, estimate a signature on each subset, and compute the overlap between two signatures in a pair as the fraction of shared genes, i.e., $|S_1 \cap S_2|/s$. Note that this corresponds to the figure of merit defined by Ein-Dor *et al.* (2006). The random sampling of subsets is repeated 20 times on each dataset, and the stability values are averaged over all samples. We will refer to this procedure the *soft-perturbation* setting in the remaining. Second, to assess stability with respect to strong perturbation within a dataset, we repeat the same procedure but this time with no overlap between two subsets of samples. In practice, we can only sample subsets of size $N/2$, where N is the number of samples in a dataset, to ensure that they have no overlap. Again, we measure the overlap between the signatures estimated on training sets with no sample in common. We call this procedure the *hard-perturbation* setting. Finally, to assess the stability across datasets, we estimate signatures on each dataset independently, using all samples on each dataset, and measure their overlap. We call this procedure the *between-datasets setting* below.

2.5 Functional interpretability and stability of a signature

To interpret a signature in terms of biological functions, we perform functional enrichment analysis by inspecting the signature for over-represented Gene Ontology (GO) terms. This may hint at biological hypothesis underlying the classification (Shen *et al.*, 2008; Reyal *et al.*, 2008). We performed a hypergeometrical test on each of the 5830 GO biological process (BP) terms that were associated to at least one gene in our dataset, and corrected the resulting p-values for multiple testing through the procedure of Benjamini and Hochberg (1995). To assess the *interpretability* of a signature, i.e., how easily one can extract a biological interpretation, we computed the number of GO terms over-represented at 5% FDR. To compare two signatures in functional terms, we first extracted from each signature the list of 10 GO terms with the smallest p-values, and compared the two lists of GO terms by the similarity measure of Wang *et al.* (2007) which takes into account not only the overlap between the lists but also the relationships between GO BP. Finally, to assess the *functional stability* of a selection method, we followed a procedure similar to the one presented in Section 2.4 and measured the mean functional similarity of signatures in the soft-perturbation, hard-perturbation and between-datasets settings.

3 Data

We collected 4 breast cancer datasets from Gene Expression Omnibus (Barrett *et al.*, 2009), as described in Table 1. The four datasets address the same problem of predicting metastatic relapse in breast cancer on different cohorts, and were obtained with the Affymetrix HG-U133A technology. We used a custom CDF file with EntrezGene ids as identifiers (Dai *et al.*, 2005) to estimate expression levels for 12,065 genes on each array, and normalized all arrays with the Robust Multi-array Average procedure (Irizarry *et al.*, 2003).

Dataset name	♯ examples	♯ positives	source
GSE1456	159	40	Pawitan *et al.* (2005)
GSE2034	286	107	Wang *et al.* (2005)
GSE2990	125	49	Sotiriou *et al.* (2006)
GSE4922	249	89	Ivshina *et al.* (2006)

Table 1: The four breast cancer datasets used in this study.

4 Results

4.1 Accuracy

We first assess the accuracy of signatures obtained by different feature selection methods. Intuitively, the accuracy refers to the performance that a classifier trained on the genes in the signature can reach in prediction. Although some feature selection methods (wrapper and embedded) jointly estimate a predictor, we dissociate here the process of selecting a set of genes and training a predictor on these genes, in order to perform a fair comparison common to all feature selection methods. We tested the accuracy of 100-gene signatures obtained by each feature selection method, combined with 5 classifiers to build a predictor as explained in Section 2.3. Table 2 shows the mean accuracies (in AUC) over the datasets reached by the different combinations in 10-fold cross-validation.

Globally, we observe only limited differences between the feature selection methods, for a given classification method. In particular the selection of a random signature reaches a baseline AUC comparable to that of other methods, confirming results already observed by Ein-Dor *et al.* (2005). Second, we observe that, among all classification algorithms, the simple NC classifier consistently gives good results compared to other classifiers. We therefore choose it as a default classification algorithm for further assessment of the performance of the signatures below. Figure 1 depicts graphically the AUC reached by each feature selection method with NC as a classifier, reproducing the first three lines of Table 2. In the single-run framework, the t-test performs significantly better than most methods ($p < 0.001$ against random, $p < 0.01$ against entropy, Bhattacharyya, Wilcoxon and GFS). Lasso and Elastic Net perform similarly and show an AUC significantly higher than GFS and Entropy ($p < 0.05$). Except for the t-test, random feature selection is not significantly worse than any other algorithm. Finally, we observe that ensemble methods for feature selection do not bring any improvement in accuracy in general since only Bhattacharyya and GFS benefit from ensemble-mean (resp. $p < 0.05$ and $p < 0.1$) and no significant improvement is obtained from the use of the two remaining ensemble aggregation methods.

Class.	Type	Random	t-test	Entropy	Bhatt.	Wilcoxon	SVM RFE	GFS	Lasso	Elastic Net
NC	S	0.62(0.17)	0.66(0.14)	0.58(0.15)	0.60(0.15)	0.62(0.15)	0.62(0.15)	0.58(0.15)	0.63(0.15)	0.63(0.15)
	E-M	0.62(0.15)	0.65(0.14)	0.59(0.15)	0.63(0.15)	0.62(0.15)	0.63(0.14)	0.62(0.13)	0.61(0.16)	0.63(0.15)
	E-E	0.61(0.15)	0.65(0.14)	0.59(0.15)	0.61(0.16)	0.62(0.15)	0.61(0.15)	0.58(0.13)	0.63(0.13)	0.63(0.14)
	E-S	0.63(0.14)	0.65(0.14)	0.58(0.15)	0.61(0.15)	0.62(0.15)	0.63(0.15)	0.59(0.12)	0.63(0.13)	0.63(0.14)
KNN	S	0.59(0.16)	0.61(0.15)	0.58(0.11)	0.57(0.13)	0.63(0.15)	0.60(0.15)	0.59(0.13)	0.60(0.17)	0.60(0.17)
	E-M	0.61(0.14)	0.62(0.15)	0.57(0.15)	0.60(0.15)	0.64(0.16)	0.62(0.15)	0.61(0.12)	0.61(0.15)	0.60(0.12)
	E-E	0.55(0.13)	0.63(0.15)	0.53(0.10)	0.54(0.10)	0.63(0.16)	0.60(0.17)	0.54(0.16)	0.61(0.14)	0.60(0.17)
	E-S	0.60(0.13)	0.63(0.15)	0.54(0.11)	0.54(0.12)	0.62(0.16)	0.58(0.14)	0.55(0.14)	0.62(0.14)	0.60(0.14)
LDA	S	0.54(0.12)	0.56(0.12)	0.51(0.14)	0.55(0.13)	0.52(0.12)	0.56(0.12)	0.50(0.13)	0.58(0.14)	0.57(0.14)
	E-M	0.53(0.10)	0.55(0.13)	0.55(0.13)	0.58(0.12)	0.56(0.13)	0.60(0.15)	0.52(0.14)	0.59(0.14)	0.60(0.13)
	E-E	0.54(0.13)	0.53(0.15)	0.52(0.15)	0.53(0.11)	0.53(0.14)	0.57(0.13)	0.53(0.15)	0.59(0.12)	0.58(0.13)
	E-S	0.54(0.13)	0.52(0.13)	0.54(0.13)	0.55(0.12)	0.52(0.14)	0.57(0.16)	0.54(0.15)	0.59(0.15)	0.60(0.13)
NB	S	0.57(0.14)	0.60(0.13)	0.58(0.11)	0.58(0.14)	0.57(0.13)	0.56(0.14)	0.54(0.11)	0.59(0.15)	0.59(0.15)
	E-M	0.59(0.13)	0.59(0.14)	0.57(0.14)	0.59(0.13)	0.57(0.13)	0.56(0.13)	0.59(0.12)	0.57(0.15)	0.57(0.14)
	E-E	0.55(0.15)	0.60(0.14)	0.58(0.12)	0.57(0.13)	0.58(0.13)	0.57(0.14)	0.58(0.11)	0.58(0.12)	0.58(0.13)
	E-S	0.58(0.14)	0.60(0.14)	0.57(0.13)	0.57(0.13)	0.58(0.13)	0.56(0.14)	0.58(0.10)	0.58(0.11)	0.58(0.13)
SVM	S	0.56(0.18)	0.56(0.15)	0.55(0.11)	0.55(0.12)	0.54(0.15)	0.62(0.14)	0.51(0.16)	0.62(0.15)	0.62(0.15)
	E-M	0.51(0.15)	0.55(0.14)	0.59(0.16)	0.60(0.13)	0.56(0.13)	0.62(0.15)	0.55(0.16)	0.61(0.16)	0.61(0.16)
	E-E	0.54(0.16)	0.54(0.15)	0.54(0.13)	0.54(0.12)	0.55(0.15)	0.61(0.17)	0.56(0.17)	0.63(0.13)	0.62(0.16)
	E-S	0.54(0.17)	0.55(0.18)	0.56(0.12)	0.56(0.12)	0.54(0.14)	0.61(0.16)	0.55(0.17)	0.63(0.14)	0.62(0.16)

Table 2: AUC obtained for each combination of feature selection and classification method, in 10-fold cross validation and averaged over the datasets. Standard error is shown within parentheses. For each selection algorithm, we highlighted the setting in which it obtained the best performance. The *Type* column refers to the use of feature selection run a single time (S) or through ensemble feature selection, either with the mean (E-M), exponential (E-E) or stability selection (E-S) procedure to aggregate lists.

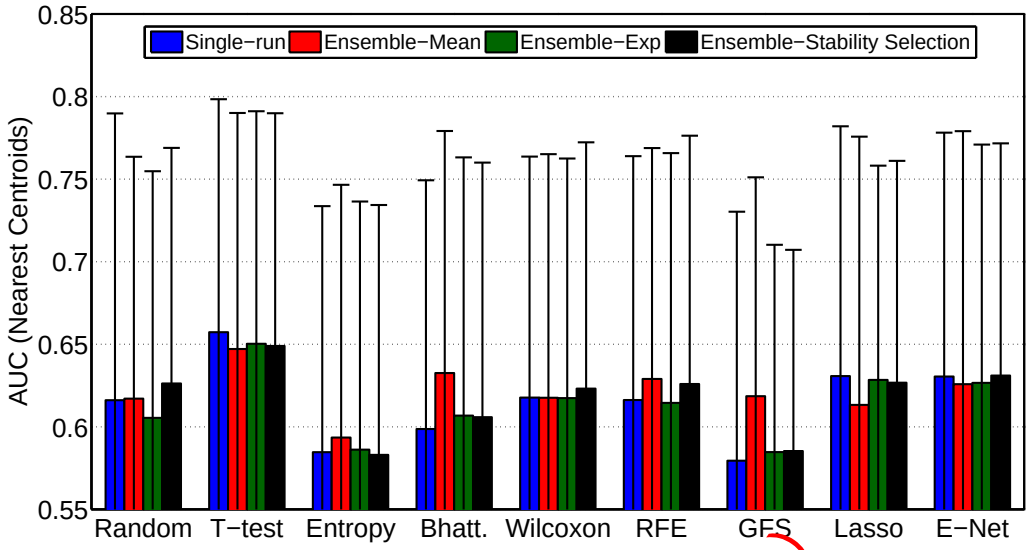


Figure 1: Area under the ROC curve for a signature of size 100 in a 10-fold CV setting and averaged over the four datasets

In order to check how these results depend on the size of the signature, we plot in Figure 2 the AUC of the 9 feature selection methods, with or without ensemble averaging, combined with a NC classifier, as a function of the size of the signature. Interestingly, we observe that in some cases the AUC seems to increase early, implying that fewer than 100 genes may be sufficient to obtain the maximal performance. Indeed, while it is significant that 100-gene signatures perform better than a list of fewer than 10 features ($p < 0.05$ regardless of the method or the setting), signatures of size 50 do not lead to significantly worse performances in general. It is worth noting that some algorithms have an increasing AUC curve in this range of sizes, and we observe no overfitting that may lead to a decreasing AUC when the number of features increases. Random selection was previously shown to give an AUC equivalent to other methods for a large signature, but as we observe on this picture, the fewer genes the larger the gap in AUC.

Finally, we estimate the predictive performance of a signature across datasets (Supplementary Table 1). Entropy is significantly less accurate than all other methods. T-test significantly outperforms other filter methods, and elastic net and Lasso also perform significantly better than Wilcoxon and SVM RFE. T-test and SVM RFE benefit from ensemble-mean, but no method significantly benefits from ensemble-exponential or ensemble-stability selection.

4.2 Stability of gene lists

We now assess the stability of signatures created by different feature selection methods at the gene level. Figure 3 compares the stability of 100-gene signatures estimated by all feature selection methods tested in this benchmark, in the three experimental settings: soft-perturbation, hard-perturbation and between-datasets settings. The results are averaged over the bootstrap replicates and the four datasets. It appears very clearly and significantly that filter methods provide more stable lists than wrappers and embedded methods. It also seems that ensemble-exponential and ensemble-stability selection yield much more stable signatures than ensemble-average. It is worth noting that a significant gain in robustness through bootstrap is only observable for relative entropy and Bhattacharyya distance. Interestingly, SVM-RFE seems to

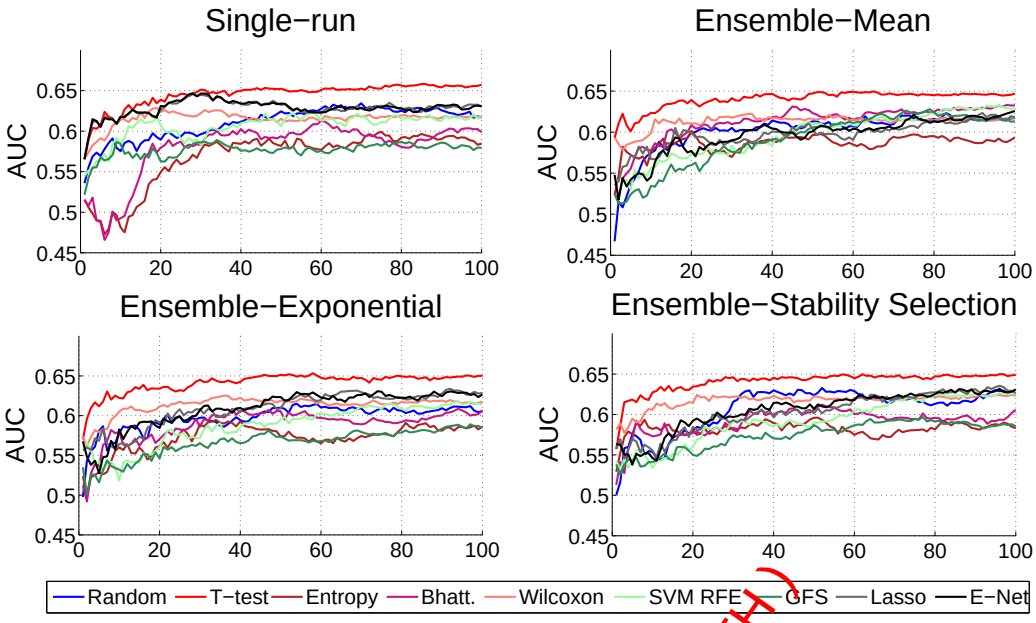


Figure 2: AUC of a NC classifier trained as a function of the size of the signature, for different feature selection methods, in a 10-fold CV setting averaged over the four datasets

benefit from ensemble aggregation in the soft-perturbation setting, as observed by Abeel *et al.* (2010), but this effect seems to vanish in the more relevant hard-perturbation and between-dataset settings. We also observed that the relative stability of the different methods does not depend on the size of the signature over a wide range of values, confirming that the differences observed for signatures of size 100 reveal robust differences between the methods (Supplementary Figure 1).

Obviously, Figures 3b and 3c are very much alike while Figure 3a stands aside. They confirm that the hard-perturbation setting is the best way to estimate the behavior of the algorithms between different studies. The larger stability observed in the between-datasets setting compared to the hard perturbation setting for some methods (e.g., t-test) is essentially due to the fact that signatures are trained on more samples in the between-dataset setting, since no split is required within a dataset (Supplementary Figure 2). This suggests that, as predicted by Ein-Dor *et al.* (2006), the main reason for signature instability for a given technology is really the sample size issue, and not differences in cohorts or experimental protocols.

4.3 Interpretability and functional stability

Even when different signatures share no or little overlap in terms of genes, it is possible that they encode the same biological processes and be useful if we can extract information about these processes from the gene lists in a robust manner. In the case of breast cancer prognostic signatures, for example, several recent studies have shown that functional analysis of the signatures can highlight coherent biological processes (Fan *et al.*, 2006; Reyal *et al.*, 2008; Shen *et al.*, 2008; Abraham *et al.*, 2010; Shi *et al.*, 2010). Just like stability at the gene level, it is therefore important to assess the stability of biological interpretation that one can extract from signatures.

First, we evaluate the *interpretability* of signatures of size 100, i.e., the ability of functional analysis to bring out a biological interpretation for a signature.

As shown on Figure 4a, the four filter methods appear to be much more interpretable than wrappers/embedded methods. However, it should be pointed out that the number of significant

GO terms is often zero regardless of the algorithm, leading to large error bars. Ensemble methods do not seem to enhance the interpretability of signatures.

Second, we assess on Figure 4b the *stability* of the functional interpretation in the between-dataset setting (the soft- and hard-perturbation settings are shown in Supplementary Figure 3). Stability results at the functional level are overall very similar to the results at the gene level, namely, we observe that univariate filters are overall the most stable methods, and that the hard-perturbation setting returns a trustworthy estimate of the inter-datasets stability. In particular, we note that in the single-run settings, only signatures obtained from filters are significantly more stable than random at ($p < 0.05$). We also note that Ensemble-mean never improves the functional stability and that Ensemble-exponential/Ensemble stability selection return more stable signatures than single-run for Entropy and Bhattacharyya ($p < 10^{-22}$) as well as for GFS ($p < 10^{-6}$) and Lasso ($p = 0.029$) although less significantly.

5 Discussion

We compared a panel of 32 feature selection methods in light of two important criteria: accuracy and stability, both at the gene and at the functional level. Figure 5 summarizes the relative performance of all methods, and deserves several comments.

Taking random feature selection as a baseline, we first notice the strange behaviour of gene selection by Batthacharyya distance and relative entropy: they are both more stable but less accurate than random selection. A careful investigation of the genes they select allowed us to identify that they tend to select genes with low expression levels, independently of the sample labels, as explained in Supplementary Figures 4 and 5. This unwanted behaviour can easily be fixed by pre-filtering genes with small variations, but it highlights the danger of blindly trusting a feature selection method, which in this case gives very stable and interpretable signatures.

Second, we observe that among the other methods, only elastic net, Lasso and t-test clearly outperform random in terms of accuracy, and only t-test outperforms it in terms of stability. Overall, t-test gives both the best performance and the best stability. The fact that the Lasso is not stable is not surprising since, like most multivariate methods, it tries to avoid redundant genes in a signature and should therefore not be stable in data where typically many genes encode for functionally related proteins. What was less expected is that neither the elastic net, which was designed exactly to fight this detrimental property of Lasso by allowing the selection of groups of correlated genes, nor stability selection, which is supposed to stabilize the features selected by Lasso, were significantly more stable than the Lasso. In addition, we also found very unstable behaviours at the functional level. This raises questions about the relevance of these methods for gene expression data. Similarly, the behavior of wrapper methods was overall disappointing. SVM RFE and Greedy Forward Selection are neither more accurate, nor more stable or interpretable than other methods, while their computational cost is much higher. Although we observed like Abeel *et al.* (2010) that SVM RFE can benefit from ensemble feature selection, it remains below the t-test both in accuracy and stability.

Overall we observed that ensemble method which select features by aggregating signatures estimated on different bootstrap samples increased the stability of some methods in some cases, but did not clearly improve the best methods. Regarding the aggregation step itself, we advise against the use of ensemble-average, i.e. averaging the ranks of each gene over the bootstrapped lists, regardless of the selection method. Ensemble-stability selection or ensemble-exponential gave consistently better results. The superiority of the latter two can be explained by the high instability of the rankings, as discussed in Iwamoto and Pusztai (2010).

Regarding the choice of method to train a classifier once features are selected, we observed that the best accuracy was achieved by the simplest one, namely the *nearest centroids* classifier,

used e.g. by Lai *et al.* (2006); Abraham *et al.* (2010). An advantage of this classifier is that it does not require any parameter tuning, making the computations fast and less prone to overfitting.

We noticed that evaluating the stability and the interpretability in a soft-perturbation setting may lead to untrustworthy results. The best estimation seems to be obtained in the hard-perturbation setting experiments. The lack of stability between datasets has been explained by four arguments. First data may come from different technological platforms, which is not the case here. Second and third, there are differences in experimental protocols and in patient cohorts, which is indeed the case between datasets; fourth, the small number of sample leads statistical instability. We however obtained very similar stability in the *hard-perturbation* setting (within each dataset) and in the *inter-datasets* results. This suggests that the main source of instability is not the difference in cohorts or experimental protocols, but really the statistical issue of working in high dimension with few samples.

References

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., and Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, **26**(3), 392–398.

Abraham, G., Kowalczyk, A., Loi, S., Haviv, I., and Zobel, J. (2010). Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. *BMC Bioinformatics*, **11**(1), 277.

Barrett, T., Troup, D., Wilhite, S., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I., Soboleva, A., Tomashevsky, M., Marshall, K., *et al.* (2009). NCBI GEO: archive for high-throughput functional genomic data. *Nucleic acids research*, **37**(Database issue), D885.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Bi, J., Bennett, K., Embrechts, M., Breneman, C., and Song, M. (2003). Dimensionality reduction via sparse support vector machines. *J. Mach. Learn. Res.*, **3**, 1229–1243.

Dai, M., Wang, P., Boyd, A., Kostov, G., Athey, B., Jones, E., Bunney, W., Myers, R., Speed, T., Akil, H., *et al.* (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic acids research*, **33**(20), e175.

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**(2), 171–178.

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA*, **103**(15), 5923–5928.

Fan, C., Oh, D., Wessels, L., Weigelt, B., Nuyten, D., Nobel, A., van't Veer, L., and Perou, C. (2006). Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med.*, **355**(6), 560.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**(1/3), 389–422.

Ioannidis, J. P. A. (2005). Microarrays and molecular research: noise discovery? *Lancet*, **365**(9458), 454.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level datas. *Biostatistics*, **4**(2), 249–264.

Ivshina, A., George, J., Senko, O., Mow, B., Putti, T., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., *et al.* (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research*, **66**(21), 10292.

Iwamoto, T. and Pusztai, L. (2010). Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data? *Genome Medicine*, **2**(11), 81.

Kohavi, R. and John, G. (1997). Wrappers for feature selection. *Artificial Intelligence*, **97**(1-2), 273–324.

Lai, C., Reinders, M., Van't Veer, L., Wessels, L., *et al.* (2006). A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC bioinformatics*, **7**(1), 235.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, **11**, 19–60.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B*, **72**(4), 417–473.

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**(9458), 488–492.

Pawitan, Y., Bjoehle, J., Amler, L., Borg, A., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., *et al.* (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research*, **7**(6), R953–R964.

Reyal, F., Van Vliet, M., Armstrong, N., Horlings, H., De Visser, K., Kok, M., Teschendorff, A., Mook, S., Van't Veer, L., Caldas, C., *et al.* (2008). A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast Cancer Res*, **10**(6), R93.

Shen, R., Chinnaiyan, A., and Ghosh, D. (2008). Pathway analysis reveals functional convergence of gene expression profiles in breast cancer. *BMC Medical Genomics*, **1**(1), 28.

Shi, W., Bessarabova, M., Dosymbekov, D., Dezso, Z., Nikolskaya, T., Dudoladova, M., Serebryiskaya, T., Bugrim, A., Guryanov, A., Brennan, R. J., Shah, R., Dopazo, J., Chen, M., Deng, Y., Shi, T., Jurman, G., Furlanello, C., Thomas, R. S., Corton, J. C., Tong, W., Shi, L., and Nikolsky, Y. (2010). Functional analysis of multiple genomic signatures demonstrates that classification algorithms choose phenotype-related genes. *The pharmacogenomics journal*, **10**, 310–23.

Sotiriou, C. and Pusztai, L. (2009). Gene-expression signatures in breast cancer. *N. Engl. J. Med.*, **360**(8), 790–800.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., *et al.* (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *JNCI Cancer Spectrum*, **98**(4), 262.

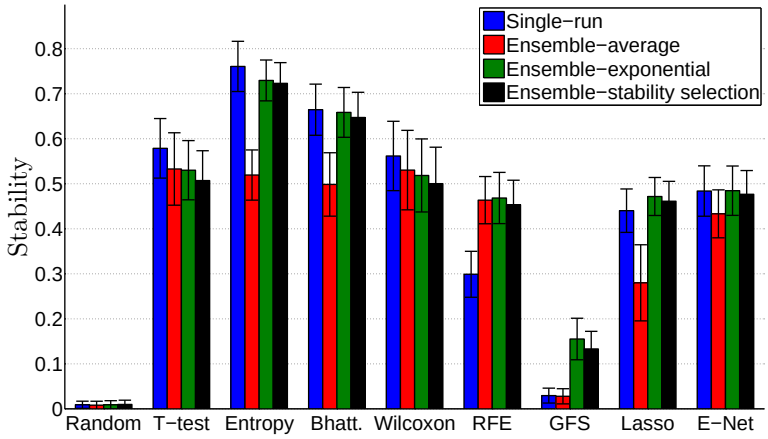
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**(1), 267–288.

Wang, J., Du, Z., Payattakool, R., Yu, P., and Chen, C. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**(10), 1274.

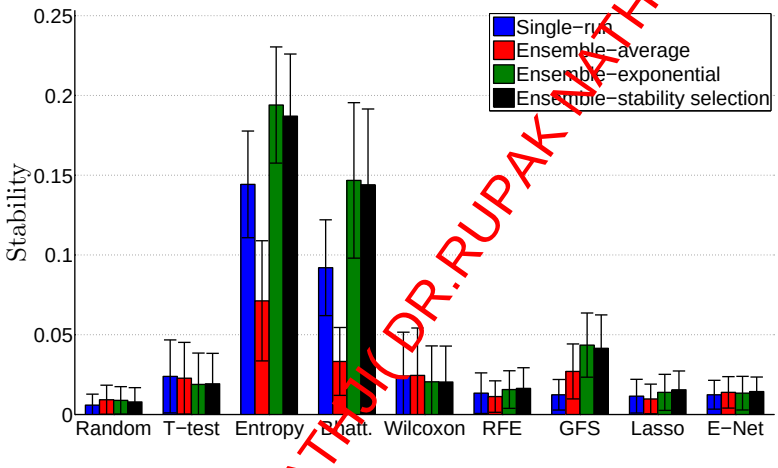
Wang, Y., Klijn, J., Zhang, Y., Sieuwerts, A., Look, M., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M., Yu, J., Jatkoe, T., Berns, E., Atkins, D., and Foekens, J. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancers. *Lancet*, **365**(9460), 671–679.

Wirapati, P., Sotiriou, C., Kunkel, S., Farmer, P., Pradervand, S., Haibe-Kains, B., Desmedt, C., Ignatiadis, M., Sengstag, T., Schutz, F., *et al.* (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res*, **10**(4), R65.

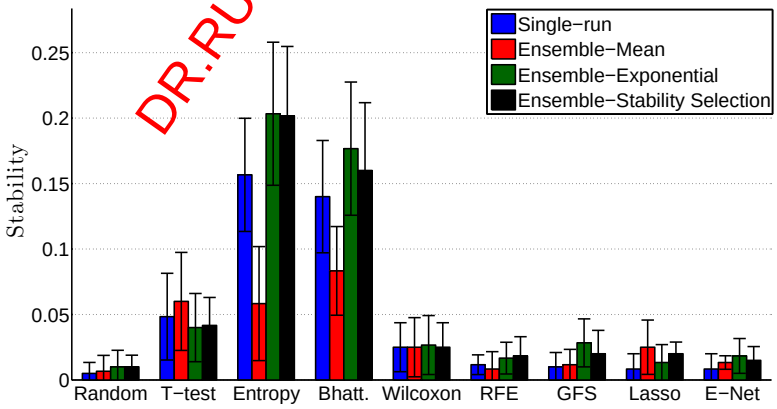
Zou, H. and Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. Ser. B*, **67**, 301–320.



(a)

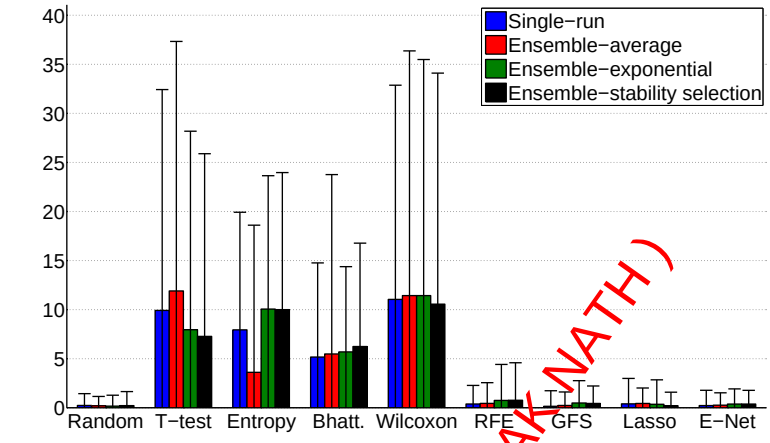


(b)

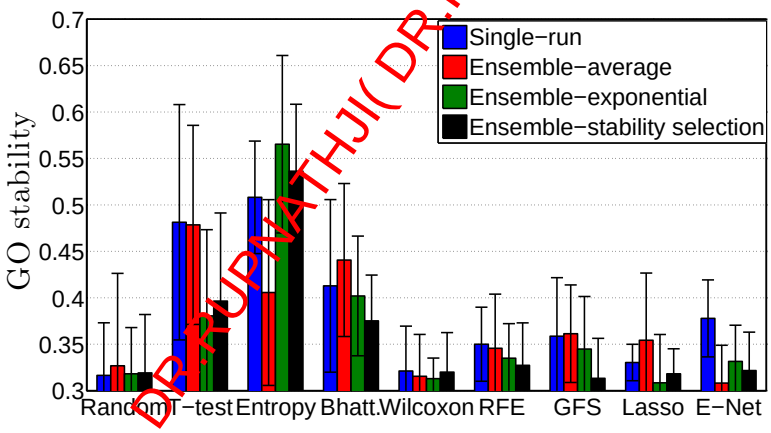


(c)

Figure 3: Stability for a signature of size 100. Average and standard errors are obtained over the four datasets. a) Soft-perturbation setting. b) Hard-perturbation setting. c) Between-datasets setting.



(a)



(b)

Figure 4: GO interpretability and functional stability of for a signature of size 100. a) Average number of GO BP terms significantly over-represented. b) Functional stability in the between-datasets setting.

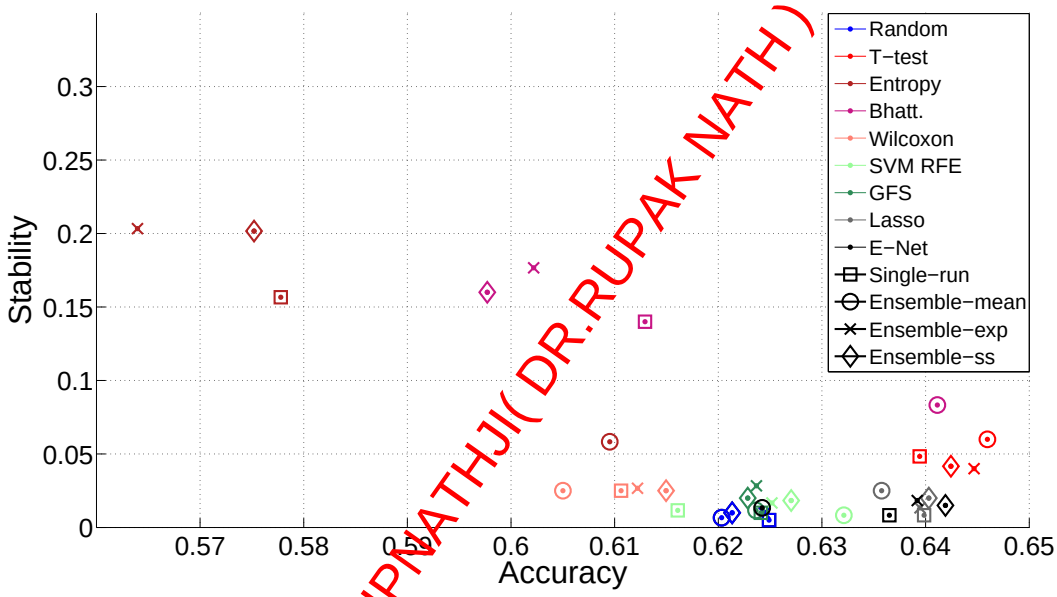


Figure 5: Accuracy versus stability for each method in the between-datasets setting. We show here the average results over the four datasets.