Data Perturbation Independent Diagnosis and Validation of Breast Cancer Subtypes Using Clustering and Patterns

G. Alexe[1,2,*], G.S. Dalgin[3,*], R. Ramaswamy [2,4], C. DeLisi[5] and G. Bhanot[1,2,5,6]

[1]Computational Biology Center, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A. [2]The Simons Center for Systems Biology, Institute for Advanced Study, Princeton NJ 08540, U.S.A. [3]Molecular Biology, Cell Biology and Biochemistry Program, Boston University, 2 Cummington Street, Boston, MA 02215, U.S.A. [4]School of Information Technology, Jawaharlal Nehru University, New Delhi 110 067, India. [5]Biomedical Engineering, Boston University, 44 Cummington Street, Boston, MA 02215, U.S.A. [6]Department of Biomedical Engineering and BioMaPS Institute, Rutgers University, Piscataway, NJ 08854.
[*]Jont First Authors.

Abstract: Molecular stratification of disease based on expression levels of sets of genes can help guide therapeutic decisions if such classifications can be shown to be stable against variations in sample source and data perturbation. Classifications inferred from one set of samples in one lab should be able to consistently stratify a different set of samples in another lab. We present a method for assessing such stability and apply it to the breast cancer (BCA) datasets of Sorlie et al. 2003 and Ma et al. 2003. We find that within the now commonly accepted BCA categories identified by Sorlie et al. Luminal A and Basal are robust, but Luminal B and ERBB2+ are not. In particular, 36% of the samples identified as Luminal B and 55% identified as ERBB2+ cannot be assigned an accurate category because the classification is sensitive to data perturbation. We identify a "core cluster" of samples for each category, and from these we determine "patterns" of gene expression that distinguish the core clusters from each other. We find that the best markers for Luminal A and Basal are (ESR1, LIV1, GATA-3) and (CCNE1, LAD1, KRT5), respectively. Pathways enriched in the patterns regulate apoptosis, tissue remodeling and the immune response. We use a different dataset (Ma et al. 2003) to test the accuracy with which samples can be allocated to the four disease subtypes. We find, as expected, that the classification of samples identified as Luminal A and Basal is robust but classification into the other two subtypes is not.

Keywords: Breast cancer, Clusters, Patterns, Multi-gene Biomarkers, Diagnosis.

Introduction

Breast cancer (BCA) is a common and heterogeneous disease affecting women of all ages. Its occurrence is correlated with levels of estrogen (ER), progesterone (PR) and Her2neu (ERBB2) (Gruvberger et al. 2001; Lacroix and Leclercq 2005). Clinically, BCA is classified into two major subtypes: ER+ and ER-. These groups are sometimes stratified further by ERBB2 and/or PR levels. Across all treatments, ER+ and/or PR+ patients have a better prognosis than ER- and/or PR- tumors (Anim et al. 2005) and are also more likely to respond to hormone therapy (e.g. tamoxifen). Over-expression of ERBB2, seen in 25–30% of cases, is often a marker of aggressive disease, poor prognosis and mixed treatment results (Diermeier et al. 2005).

In spite of sustained research and medical and pharmaceutical effort, the incidence and death rate of BCA remains high. In 2005, more than 1.2 million new cases were diagnosed world wide and more than 20% of these will die from the disease (http://imaginis.com/breasthealth/). A major cause of treatment failure is that tumors with similar histopathology have divergent clinical courses and prognoses. The goal of the present study is the same as that of many others (Bieche et al. 1995; West et al. 2001; van't Veer et al. 2002; Honig et al. 2004; Ahnstrom et al. 2005; Sharma et al. 2005; Osipo et al. 2005), that molecular profiling of BCA will clarify molecular correlates of disease, and this in turn will improve choice of therapy, and provide leads to new and more effective therapeutics.

In a series of papers on analysis of cDNA data of BCA tissue samples (Sorlie et al. 2001; Perou et al. 2000, 2001) the samples were uniquely assigned to one of four distinct categories: Luminal A, Luminal B, ERBB2+ (or Her2+) and Basal-like. These subtypes were later validated by Sotiriou et al. 2003, Loi et al. 2005 and Kristensen et al. 2005. The first two categories were mostly ER+ and the latter two mostly ER– negative. In the original analysis of Perou et al. 2000, Basal tumors were characterized by high levels of keratins 5 and 17, laminin, and fatty acid binding protein 7 genes (see also Charafe-

Correspondence: Gyan Bhanot. Email: gyanbhanot@gmail.com; Fax: 609-951-4438.

Jauffret et al. 2005), whereas ERBB2+ was characterized by high levels of several genes in the ERBB2 amplicon at 17q12.21 including ERBB2 and GRB7. Other studies identified different markers (Abd El-Rehim et al. 2005; Bertucci et al. 2005; Farmer et al. 2005; Hu et al. 2006; Sorlie et al. 2006) and a consensus set of markers for all BCA patients is not currently available.

Luminal and Basal-like tumors arise in distinct breast tissue cell types (Perou et al. 2000) and have very different disease course (Sorlie et al. 2001, 2003) and response to therapeutics (Troester et al. 2004; Bertucci et al. 2005). The Luminal A subtype has the best overall prognosis followed by Luminal B while the other two subtypes are more aggressive and difficult to treat. The nomenclature of these subtypes has found its way into the language and culture of clinical practice and affects treatment options offered to patients. This makes it important to validate the stability of the original classification of Sorlie et al. This is the main goal of the present paper.

The original analysis used simple hierarchical clustering (Eisen et al. 1998) which is known to be sensitive to data perturbation (Monti et al. 2003; van der Kloot et al. 2005). We re-analyzed the data using a robust averaging procedure to access the stability of imposing five clusters (4 disease subtypes + Normal) on the data. The goal was to identify a "core" set of samples in each subtype which were stable under data perturbations, and to use these cores to determine "patterns" of gene expression for each core. We found stable core clusters for samples in the Luminal A, Basal and Normal clusters of the original analysis. However, the "Luminal B" and "ERBB2+" clusters of Sorlie et al. were unstable, with only a subset of the samples from the previous assignment remaining in stable core clusters under data perturbation. Instead, the originally assigned samples scattered over two or more clusters. This suggests that the Luminal B and ERBB2+ clusters (and their markers) as identified in Sorlie et al. 2003, are unstable to data perturbation and need further analysis.

For the Luminal A and Basal categories, we find a robust set of gene markers and patterns. If we combine the Sorlie et al. dataset with a new dataset from Ma et al. and cluster the combined data using these robust gene markers and patterns, then in the new data, we can assign a robust subtype label for Luminal A and Basal but not for the other two disease phenotypes.

Materials and Methods

Datasets

Data 1: The cDNA dataset of (Sorlie et al. 2003) was obtained from http://genome-www.stanford.edu/breast_cancer/robustness/data/SupplText.html. The data had expression levels of $N = 552$ genes for $M = 122$ samples of which 112 were from BCA patients and 10 controls. The 552 genes were selected by Sorlie et al. to have small variation in tissue samples from the same patient and a high variation in tissue samples from different patients.

Data 2: The Ma et al. dataset was downloaded from www.geneexpression_ma.org. It consisted of expression levels of 1940 genes for 93 samples micro-dissected from 36 BCA patients and 3 normals. The samples were from three stages of disease: atypical ductal hyperplasia or ADH, ductal carcinoma *in situ* or DCIS and invasive ductal carcinoma or IDC respectively. The genes made available in the data were chosen by linear discriminant analysis as markers for breast cancer progression. ER, PR and HER2neu levels measured through immunohistochemistry were available.

Preprocessing and Imputation for Data 1

The matrix of samples (columns) and genes (rows) was normalized to mean 0 and variance 1 first across columns and then across rows, ignoring missing entries. The matrix had 5,027 missing entries. We first eliminated genes and samples with more than 20% missing entries. This reduced the data to $N = 530$ genes and $M = 118$ samples. We imputed the missing entries using a simple generalization of the *k*NN method of Troyanskaya et al. 2001 as follows:

We identified the k nearest neighbor entries for missing entry x_{ij} using the Euclidean metric,

$$d(i,i') = \left(\sum_j (x_{ij} - x_{i'j})^2\right)^{1/2}$$

with the requirement that the genes chosen as nearest neighbors have at least t% filled entries. Twenty imputations were done at each x_{ij} using the range $10 \le k \le 14$ for k and varying t from 50% to 80% in increments of 10. Let $\{x_1, x_2, \ldots, x_k\}$ be the k-nearest neighbor entries in increasing order of distance and R be a uniform random number in (0,1). Then the imputed value

y is given by $y = x_j$, which satisfies

$$\sum_{i=1}^{j-1} \frac{x_i}{X} < R \leq \sum_{i=1}^{j} \frac{x_i}{X},$$

$$\text{where } X = \sum_{i=1}^{k} x_i .$$

Twenty datasets were generated in this way, one for each (k,t) value. The clustering was averaged over these twenty datasets in order to create a set of clusters insensitive to parameter choice in data imputation. This averaging is an improvement over the kNN method because it is stable to both variation in k and variation in how the neighbors are chosen (as measured by t). Multiple clones in the data were eliminated by averaging after discarding outliers outside a 95% confidence interval. This process left 523 genes with no missing entries or clones. The final data is given in Supplementary Table 1.

Results

Identifying "Core" Clusters

We use the letters A, B, C, D, E to denote the five phenotypes: Luminal A, Luminal B, ERBB2+, Basal, and Normal respectively. The clusters were identified using the consensus hierarchical clustering technique of Monti et al. 2003 implemented in GenePattern (http://www.broad.mit.edu/cancer/software/genepattern/). This method assesses the stability of hierarchical clustering across multiple perturbations of the data. We generated 100 copies of the dataset by randomly selecting 80% of the samples. Each copy was hierarchically clustered using a Euclidean distance metric and the top 5 clusters were selected. For each distinct sample pair (i, j) in the data, we computed the frequency F_{ij} with which the pair clustered together over the 100 copies of the datasets. The matrix of F_{ij} values is called the "agreement matrix." Repeating this for all 20 data imputations and averaging gave the final "consensus agreement matrix" which is shown in Supplementary Table 2.

The five core clusters were identified as bicliques (Alexe et al. 2004) using the agreement matrix entries as a measure of similarity. We used the criterion that two samples have the same phenotype and belong to the same core cluster if they have a consensus agreement matrix score greater than P. For the Luminal A and Basal subtypes, the value $P = 90\%$ was sufficient to get an exact match between the core cluster identified by us and the assignment in Perou et al. 2000 and Sorlie et al. 2003. However, for samples assigned to Luminal B and ERBB2+ by the earlier study, these thresholds needed to be lowered to 50% and 25% respectively to get agreement with the previous assignments, suggesting that these categories are considerably less stable to data perturbation. The five core clusters contained 60 out of the 118 samples.

From the F_{ij} values, we define the average agreement score between a sample i and other samples j in a given cluster C as

$$F_{i,C} = \frac{\sum_{j=1}^{n} F_{ij}}{n},$$

where $j = 1, \ldots, n$, and n is the number of samples in the cluster C. $F_{i,C}$ was calculated for each of the five clusters. The results are shown in Figures 1 a–e. For each phenotype, we used a cutoff criterion on $F_{i,C}$ to assign it to the corresponding core cluster and these samples are shown in color. Many samples earlier identified as Luminal B also have a high score in our Basal core cluster (Figure 1b and 1d). This suggests that the Luminal B identification is problematic. Figure 1e also shows that some samples identified earlier as Luminal A are placed in our "Normal" core cluster, suggesting that these patients may have minimal disease. Overall, our analysis shows that Luminal A, Basal and Normal phenotypes are robustly classifiable into homogeneous clusters but Luminal B and ERBB2+ do not cluster well. We find that 36% of the samples previously placed in the Luminal B category and 55% of samples previously classified as ERBB2+ are in fact ambiguous; i.e., their assignments are highly sensitive to data perturbation and they should be reanalyzed or classified as ambiguous. The scores of some unclassified samples in Sorlie et al. 2003 are shown in Figure 1f. For the samples where these scores are higher than the cutoff in one core cluster but not in any other, the corresponding sample can be assigned a category label by our clustering.

Table 1 compares the original assignments of Sorlie et al. with our core clusters of Figure 1 and shows the sample id's from the original study.

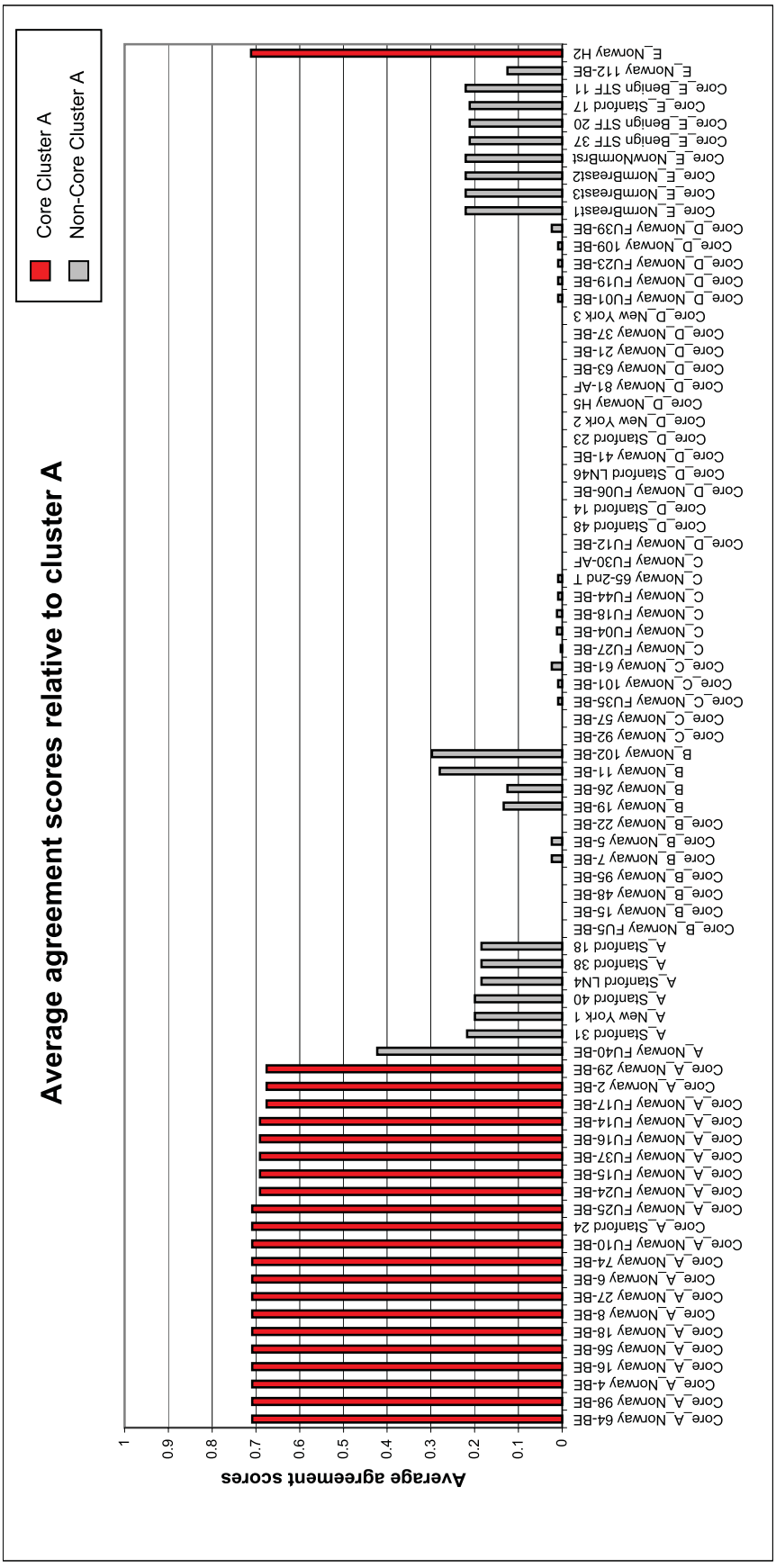

Figure 1a. Average agreement scores relative to cluster A.

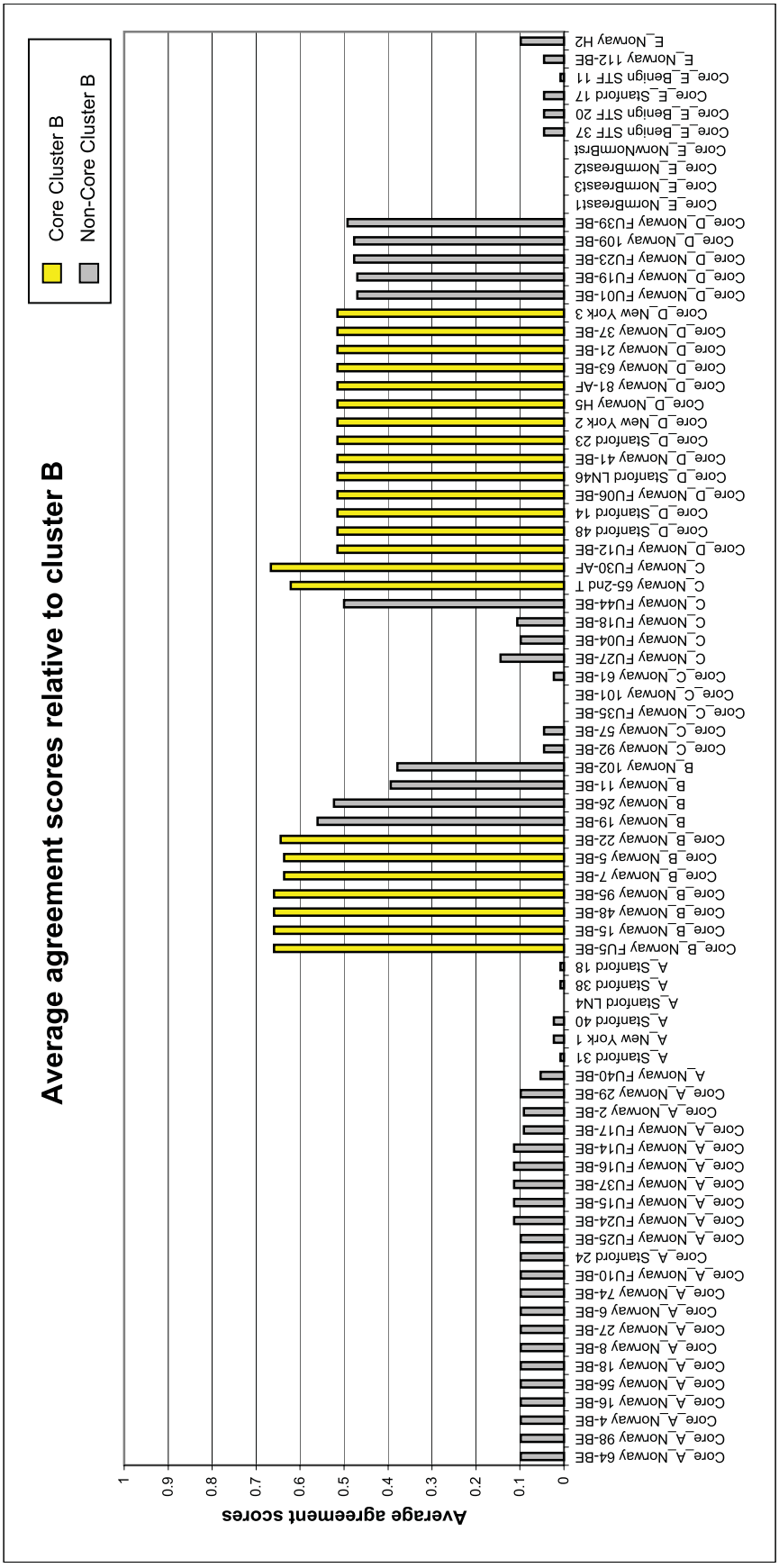

Figure 1b. Average cluster agreement scores relative to cluster B.

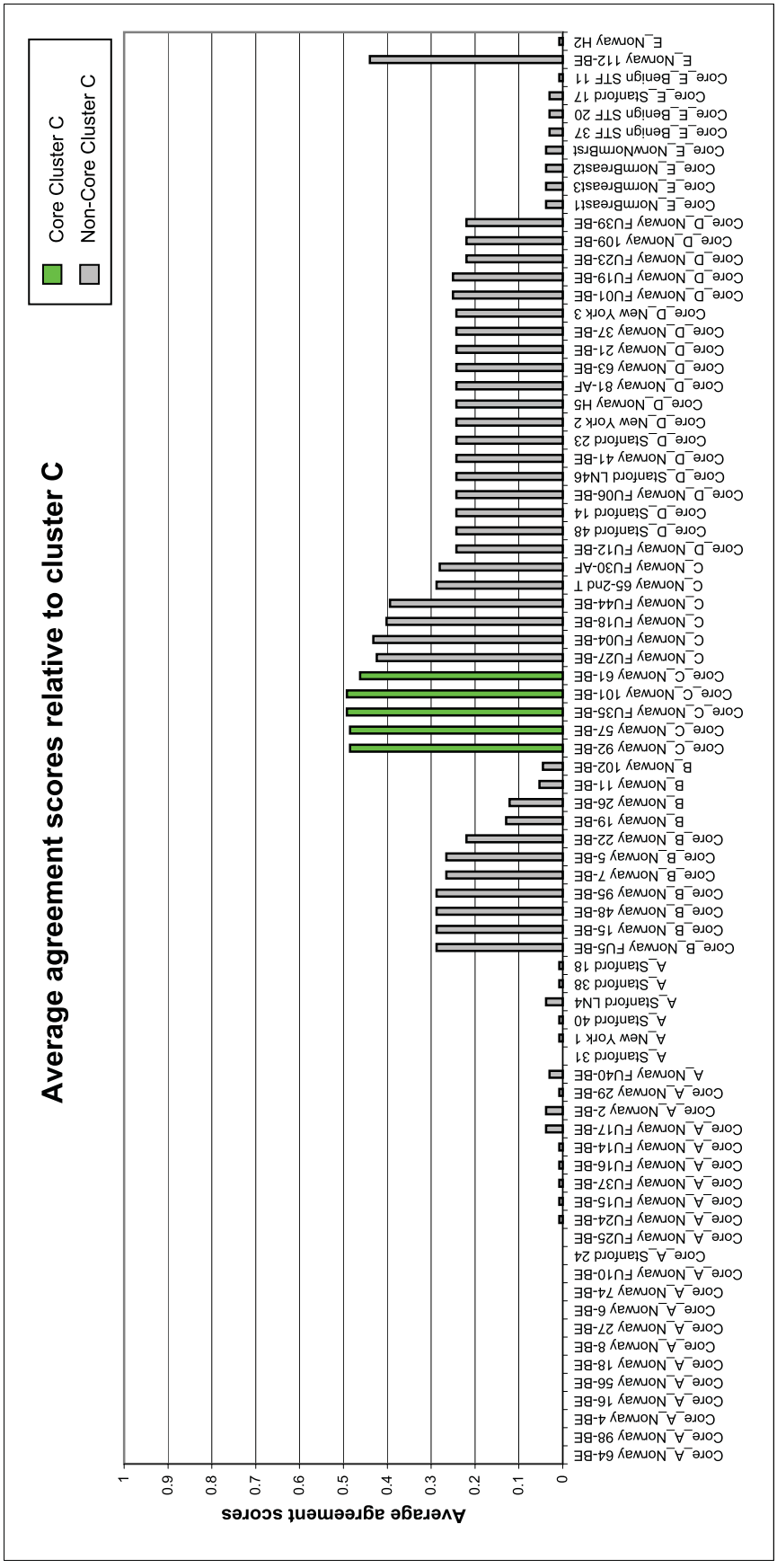

Figure 1c. Average cluster agreement scores relative to cluster C.

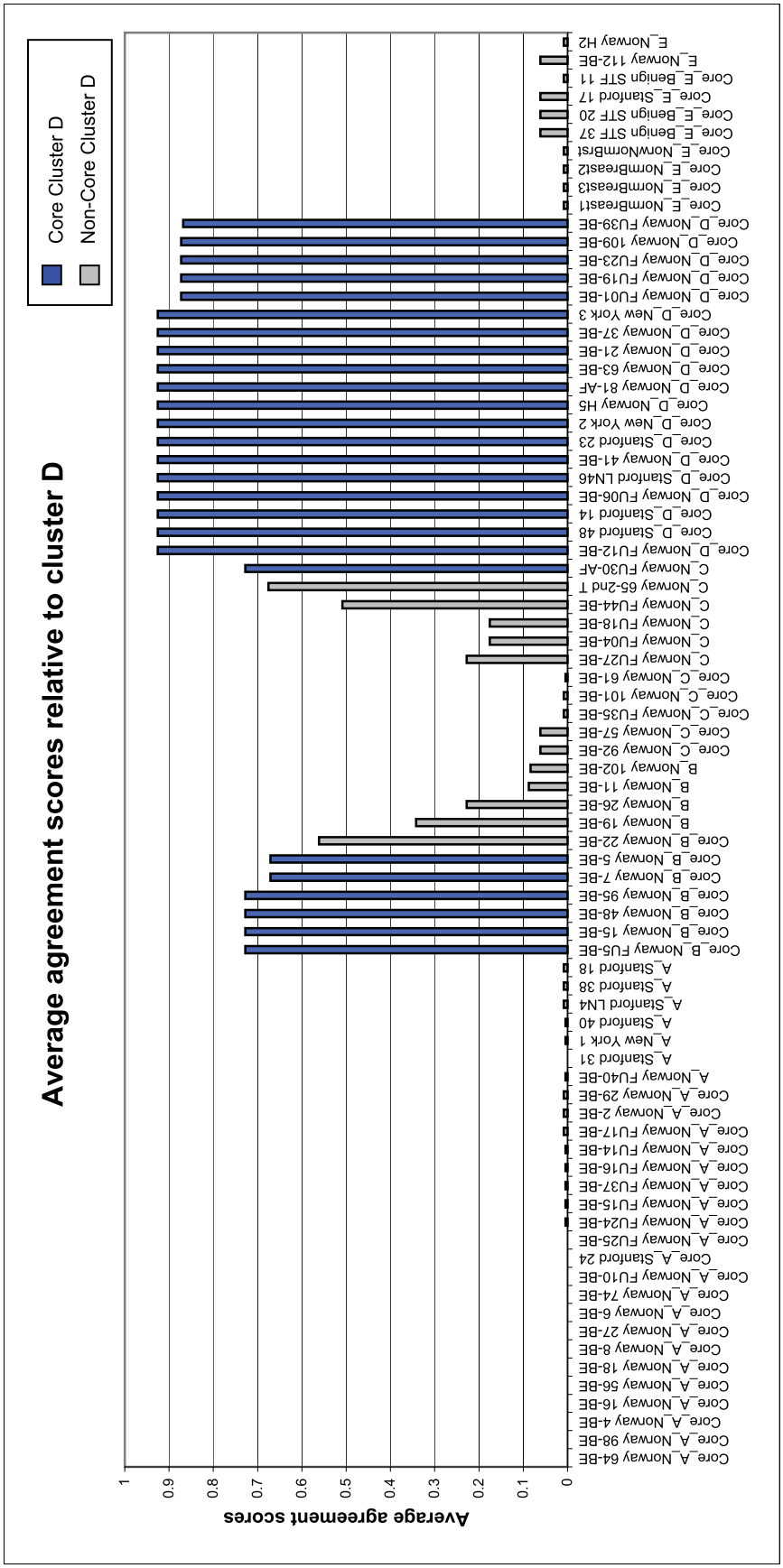

Figure 1d. Average cluster agreement scores relative to cluster D.

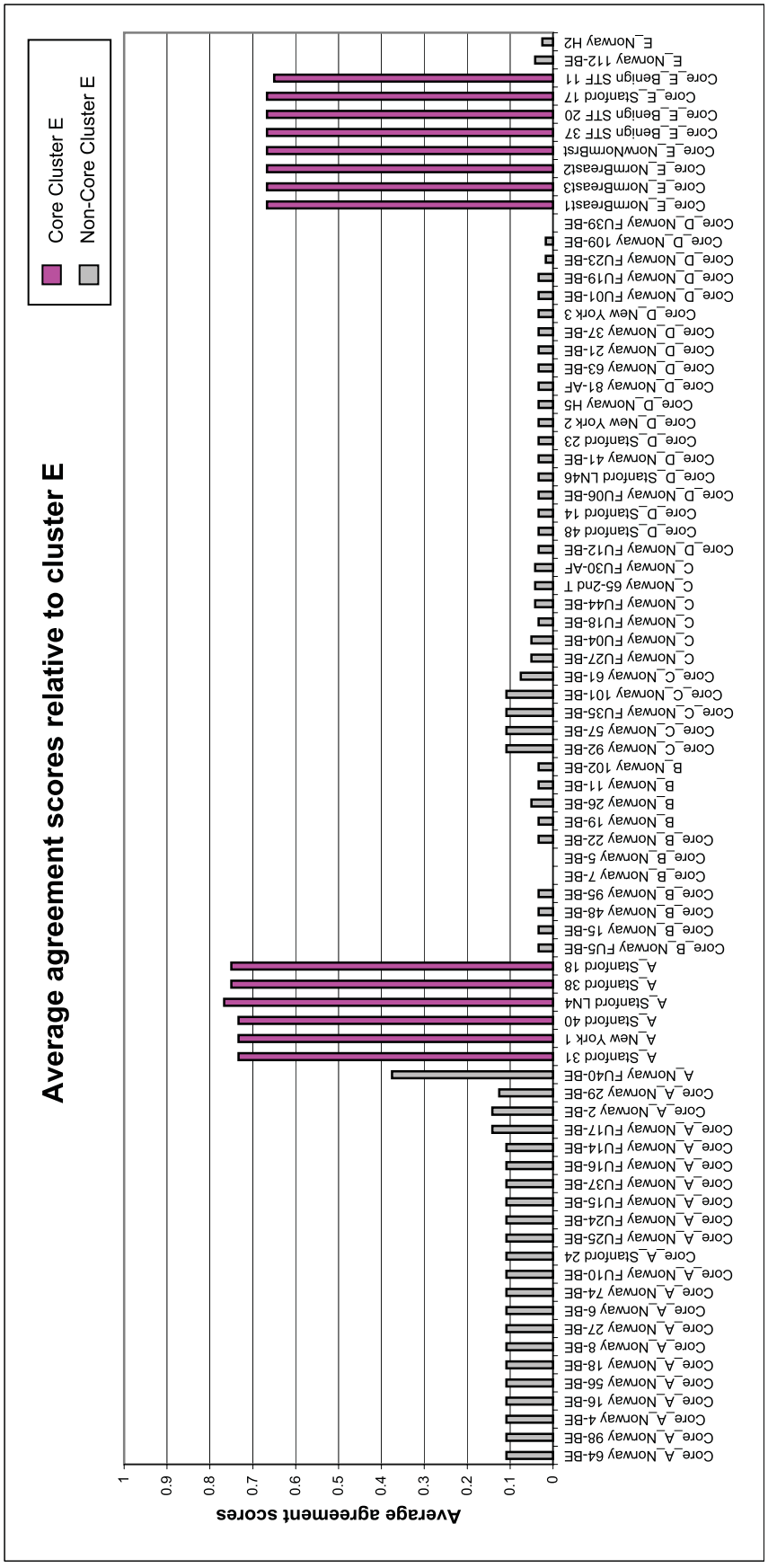

Figure 1e. Average cluster agreement scores relative to cluster E.

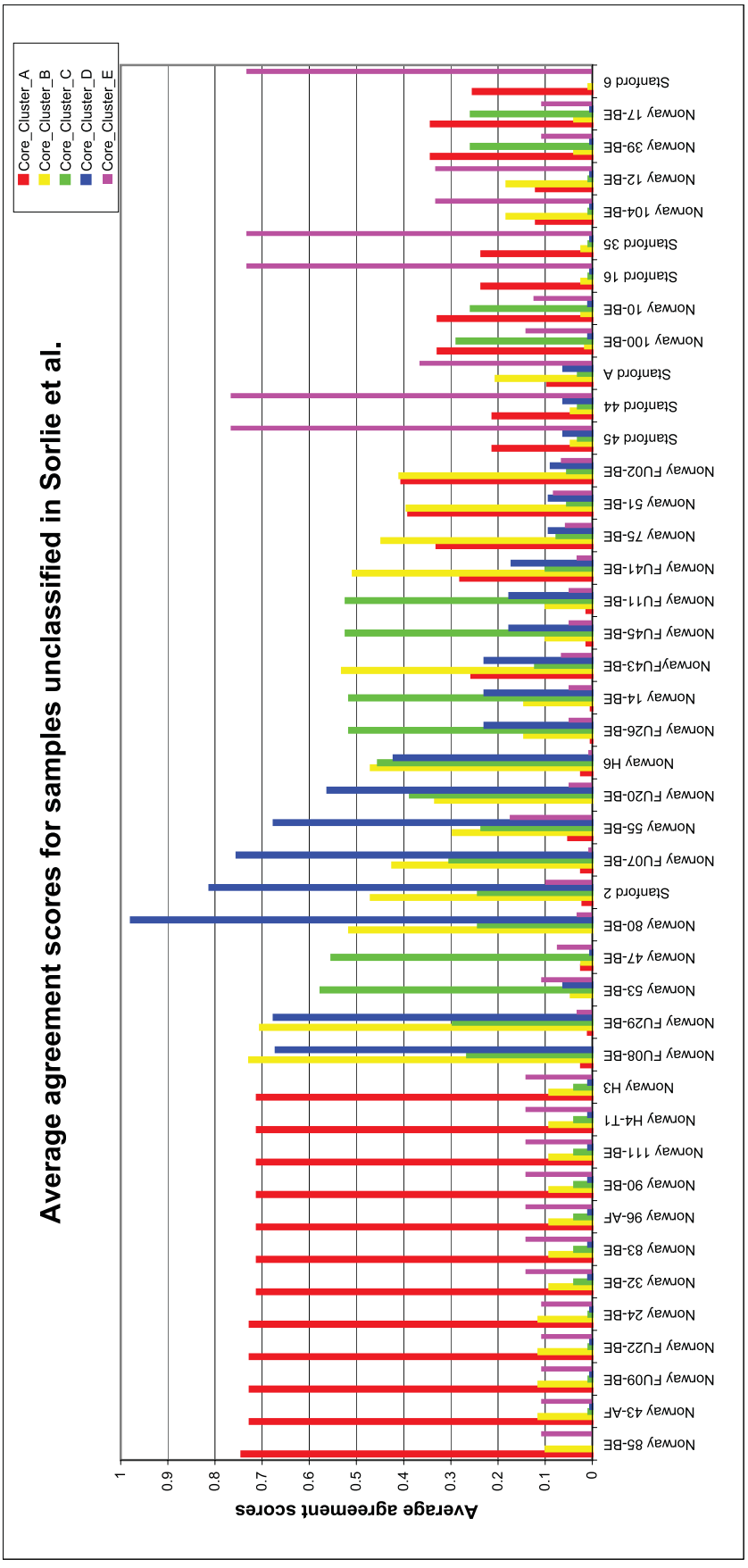

Figure 1f. Agreement scores for the unclassified samples in Sorlie et al.

Table 1. Summary of the classifications of tumor samples in the core samples (present study) and previous work. Sample identification numbers refer to the original data of Sorlie et al. 2003. The numbers of samples assigned to each phenotype by the original classification and by our clustering are shown in columns 5 and 6. We see that a larger fraction of assignments into the phenotypes Normal, Luminal A and Basal are correct. The silhouette scores are given in columns 7 and 8.

Tumor phenotype	Core cluster label	Tumor samples by id in phenotype		# samples assigned to phenotype		Quality scores (silhouette width) of clusters	
		Sorlie et al. 2003	This study core cluster	Sorlie et al. 2003	This study core cluster	Sorlie et al. 2003	This study core cluster
Luminal A	A	13-40	13-24,26,31-35,38-40	27	21	0.06	0.07
Luminal B	B	54-64	54-62	11	7	0.09	0.10
ERBB2+	C	79-89	79-85, 87, 89	11	5	0.08	0.16
Basal	D	93-111	93-111	19	19	0.14	0.13
Normal	E	112-121	113-120	10	8	0.12	0.17

The agreement fraction between the original assignment and our assignments is highest for the Normal, Luminal A and Basal categories and lower in the other two phenotypes.

For each sample i in a core cluster, we also calculated the silhouette score (Rousseeuw,1987) defined by

$$s(i) = s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

where $a(i)$ is the average dissimilarity between i and all other samples in the cluster, and $b(i)$ is the minimum average dissimilarity of i to all samples in other clusters. If $s(i)$ values in a cluster are close to unity, the cluster is well defined. An $s(i)$ value near zero indicates that the sample is between two clusters. Negative values of $s(i)$ mean that the sample is in the wrong cluster. The "silhouette width" of a cluster is the average of the $s(i)$ scores of all samples in that cluster. The silhouette widths for our core clusters as well as for the Sorlie et al. clusters are given in Table 1. The low values of the average silhouette scores are worrisome. They suggest either that the stratification into these phenotypes is problematic or that a better choices of genes is necessary to separate the phenotypes more reliably.

Identifying Robust Gene Markers

Microarray datasets suffer from an overabundance of genes, most of which do not contribute to the signal. Identifying differentially expressed genes for a given set of phenotypes is a difficult problem for which many methods have been proposed. These can be divided into two major groups (Guyon and Ellisseeff, 2003, Inza et al. 2004, Lai et al. 2006, Jeffery et al. 2006) for supervised learning:

(*i*) *Filtering or Variable Ranking methods:* These select features based on quality scores. They include the fold change test (e.g. Mutch et al. 2002; Breitling and Herzyk, 2005), the t-test (Gossett, 1908, Tusher et al. 2001), the Wilcoxon-Mann-Whitney test (Bradley, 1968; Lehman, 1975), the Signal-to-Noise Ratio (SNR) test (Golub et al. 1999), the J5 test (Patel and Lyons-Weiler, 2004), the D1 test (Patel and Lyons-Weiler, 2004) etc. Another set of methods measure the "separability" of data into different phenotype classes. These include simple separability (Patel and

Lyons-Weiler, 2004), weighted separability (Patel and Lyons-Weiler, 2004), envelope eccentricity (Alexe et al. 2006), separation measure (Alexe et al. 2006b) etc. A third class uses information-theoretic methods such as the entropy criterion (e.g. Furlanello et al. 2003; Liu et al. 2005), mutual information (e.g. Tourassi et al. 2001), information gain (Liu, 2004) etc. Finally, there are the statistical impurity measures (Su et al. 2003) which include the two-ing rule, the Gini index, max-minority, sum-minority, sum-of-variances etc.

(*ii*) *Feature Subset Selection Methods:* One such method selects those features which are useful for classification for a given machine learning algorithm (e.g. SVM (Vapnik, 1998), ANN (Bishop, 1995), kNN (Ripley, 1996) etc). More sophisticated approaches are embedded methods which include the selection of features as part of the training process for the classifier. These methods are computationally intensive and require efficient search strategies or a preliminary filtering of the non-reliable genes to reduce the dimensionality of the problem.

The existence of such a variety of feature selection methods poses a challenge in microarray data analysis. There have been recent attempts to combine various approaches into a meta selection procedure based on "majority-voting" using ranking by predictive content across many data perturbations and machine learning methods (e.g. Bhanot et al. 2005; Alexe et al. 2005a). Several studies (Guyon and Ellisseeff, 2001; Alexe et al. 2005b) have shown that variables which are only weakly correlated with phenotype are very useful when used in combinations. This principle has lead to the development and study of combinatorial markers or patterns (Crama et al. 1988; Bhanot et al. 2005; Alexe et al. 2006b).

In the present study, we have chosen to use a single feature selection method (namely the SNR test, Golub et al. 1999) which has been shown (Alexe et al. 2006b) to have good performance on genomic and proteomic data. However, we cannot guarantee that it is the best method, particularly because of the need to impute the missing data in the dataset of Sorlie et al. As an added check on the feature selection, we also use the combinatorial "pattern" method and averaging over data perturbations to reduce the errors from potentially "less than optimum" choice of features.

We identified a large pool of uni-gene markers for each core that distinguish it from the others using the signal-to-noise statistic. For gene i, if $\mu_1(i)$ and $\mu_2(i)$ be the average gene expression levels for the core and its complement and $\sigma_1(i)$ and $\sigma_2(i)$ the corresponding standard deviations, the signal-to-noise ratio (SNR) is defined as $SNR = (\mu_0 - \mu_1)/(\sigma_0 + \sigma_1)$. The t-test statistic is the same as the SNR except that the denominator is $(\sigma_0^2 + \sigma_1^2)^{1/2}$. Since $(\sigma_0 + \sigma_1) > (\sigma_0^2 + \sigma_1^2)^{1/2}$ SNR is a more conservative criterion than the t-test.

The SNR statistic is preferred over the t-test in situations when the sample size in a class is small (less than 30) because it does not assume a Gaussian distribution for the underlying variables; an assumption which is implicit in the t-test. When combined with a permutation test for measuring p-values, the SNR statistic is a powerful and widely used technique for feature selection and class discrimination (e.g. Golub et al. 1999; Ramaswamy et al. 2001; Shipp et al. 2002; Sun et al. 2004; Goh and Kasabov 2005; Monti et al. 2005) and is implemented in several software packages (e.g. GenePattern and Gene Set Enrichment Analysis (GSEA), http://www.broad.mit.edu/tools/software.html).

The signal-to-noise (SNR) was computed for each gene for each of the 20 imputed datasets and for each of the 60 leave-one-out sample perturbation experiments for the core samples. The selected genes were those whose *p*-value for the SNR was below 0.01 and the significance of the SNR for false discovery rate (FDR) (Benjamini and Hochberg, 1995) was above 0.95 in each experiment.

This procedure identified 391 robust uni-gene markers (given in Supplementary Table 3) for the five core clusters. They consisted of overlapping sets of genes, 238 for Luminal A, 234 for Basal, 66 genes for Luminal B, 35 genes for ERBB2+ and 118 genes for Normals. These included many genes identified in previous studies (Perou et al. 2000; Sorlie et al. 2003; Loi et al. 2005). For example, the Luminal A set included the known estrogen pathway genes (ESR1, LIV1, GATA-3) and the Basal set the known genes CCNE1, LAD1, and KRT5.

We further reduced this pool to 148 genes using the more stringent criteria which used the significance of the SNR for several metrics: the false discovery rate, the Q value (Storey and Tibshirani, 2003), FWER (Dudoit et al. 2002), Bonferroni correction (Bonferroni, 1935). More details about the multiple testing metrics we used are given in Supplementary Information I. These 148 genes included 79 genes for Luminal A and 60 for Basal with an overlap of 31 genes. The other phenotypes

Table 2a. Collection of uni-gene markers for the Luminal A phenotype. The markers are sorted in decreasing order with respect to to the signal-to-noise ratio.

Group	Gene index	Gene Description	GeneBank Acc	SNR	FDR
Core A	437	GATA3 GATA binding protein 3	H72474	1.46	0.00
	431	NAT1 N-acetyltransferase 1 (arylamine N-acetyltransferase)	T67128	1.38	0.00
	438	ESR1 estrogen receptor 1	AA291702	1.38	0.00
	432	LIV-1 LIV-1 protein, estrogen regulated	H29315	1.37	0.00
	420	FLJ11280 **hypothetical protein FLJ11280	N54608	1.21	0.00
	416	TCEAL1 transcription elongation factor A (SII)-like 1	AA451969	1.17	0.00
	434	HNF3A hepatocyte nuclear factor 3, alpha	T74639	1.09	0.00
	436	FLT1 fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)	AA058828	1.04	0.00
	421	Homo sapiens mRNA; cDNA DKFZp313L231 (from clone DKFZp313L231)	AA029948	0.95	0.00
	444	FBP1 fructose-1,6-bisphosphatase 1	AA699427	0.95	0.00
	440	PTP4A2 protein tyrosine phosphatase type IVA, member 2	AA504327	0.91	0.00
	439	RAB5EP rabaptin-5	AA428477	0.90	0.00
	455	KIAA0239 KIAA0239 protein	AA454740	0.87	0.00
	378	BECN1 beclin 1 (coiled-coil, myosin-like BCL2 interacting protein)	AA427367	0.83	0.00
	448	KIAA1025 KIAA1025 protein	T72613	0.81	0.00
	445	MGC27171 hypothetical protein MGC27171	R23619	0.81	0.00
	435	XBP1 X-box binding protein 1	W90128	0.80	0.00
	453	NPEPPS aminopeptidase puromycin sensitive	R24894	0.79	0.00
	425	LOC51313 **AD021 protein	N95180	0.79	0.00
	454	HIS1 HMBA-inducible	N21081	0.78	0.00
	446	HSD17B4 hydroxysteroid (17-beta) dehydrogenase 4	AA487914	0.78	0.00
	495	CYB5 cytochrome b-5	R91950	0.78	0.00
	429	FLJ10980 hypothetical protein FLJ10980	N45467	0.78	0.00
	442	CEGP1 CEGP1 protein	W74079	0.77	0.00
	443	ACADSB acyl-Coenzyme A dehydrogenase, short/branched chain	H95792	0.76	0.00
	426	Homo sapiens mRNA; cDNA DKFZp434E033 (from clone DKFZp434E033)	N63001	0.76	0.00
	491	MGST2 microsomal glutathione S-transferase 2	W73474	0.75	0.00
	380	IGBP1 immunoglobulin (CD79A) binding protein 1	AA463498	0.74	0.00
	418	POLYDOM likely ortholog of mouse polydom	R33004	0.73	0.00
	496	ALCAM activated leukocyte cell adhesion molecule	R13558	0.73	0.00
	414	ASAH1 N-acylsphingosine amidohydrolase (acid ceramidase) 1	AA664155	0.71	0.00
	399	GRLF1 glucocorticoid receptor DNA binding factor 1	N72276	0.71	0.00
	402	BF B-factor, properdin	H80257	0.70	0.00
	39	GLUD1 glutamate dehydrogenase 1	AA017175	0.69	0.00
	428	KIAA0876 KIAA0876 protein	AA431721	0.69	0.00
	398	FLJ11730 hypothetical protein FLJ11730	AA427401	0.69	0.00
	493	D5S346 DNA segment, single copy probe LNS-CAI/LNS-CAII (deleted in polyposis)	H99681	0.68	0.00
	403	FMO5 flavin containing monooxygenase 5	H52001	0.67	0.00
	411	Homo sapiens cDNA FLJ40901 fis, clone UTERU2003704	AA418564	0.67	0.00
	430	TLE3 transducin-like enhancer of split 3 (E(sp1) homolog, Drosophila)	AA057737	0.67	0.00
	433	Homo sapiens, clone MGC:22588 IMAGE:4696566, mRNA, complete cds	N74131	0.66	0.00
	494	C4B complement component 4B	AA664406	0.65	0.00
	371	QDPR quinoid dihydropteridine reductase	R38198	0.65	0.00
	262	ACTG2 actin, gamma 2, smooth muscle, enteric	T60048	-0.65	0.00
	256	ESTs	AA074677	-0.66	0.00
	69	KIT v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	N20798	-0.68	0.00
	211	TMSB10 thymosin, beta 10	AA486085	-0.68	0.00
	238	FLJ10697 hypothetical protein FLJ10697	H80748	-0.69	0.00
	243	SLPI secretory leukocyte protease inhibitor (antileukoproteinase)	AA026192	-0.70	0.00
	180	GSTP1 glutathione S-transferase pi	R33642	-0.70	0.00
	254	Homo sapiens cDNA FLJ11796 fis, clone HEMBA1006158, highly similar to Homo sapiens transcription factor forkhead-like 7 (FKHL7) gene	N22552	-0.71	0.00
	186	NRG1 neuregulin 1	R72075	-0.72	0.00
	261	FLJ22678 **hypothetical protein FLJ22678	N90109	-0.73	0.00
	244	TONDU TONDU	AA700322	-0.74	0.00
	232	PTPRK protein tyrosine phosphatase, receptor type, K	R78776	-0.76	0.00
	248	TRIM29 tripartite motif-containing 29	AA055485	-0.76	0.00
	185	PTK7 **PTK7 protein tyrosine kinase 7	AA453789	-0.77	0.00
	177	CRABP1 cellular retinoic acid binding protein 1	AA454702	-0.77	0.00
	241	CXCL1 chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)	W42723	-0.77	0.00
	240	Homo sapiens cDNA FLJ14761 fis, clone NT2RP3003302	W93120	-0.77	0.00
	187	PREP prolyl endopeptidase	AA664056	-0.77	0.00
	257	FLJ14525 hypothetical protein FLJ14525	AA464028	-0.78	0.00
	234	KIP2 DNA-dependent protein kinase catalytic subunit-interacting protein 2	N79761	-0.81	0.00
	199	BTG3 BTG family, member 3	N52496	-0.82	0.00
	97	ID4 inhibitor of DNA binding 4, dominant negative helix-loop-helix protein	AA453341	-0.82	0.00
	245	GABRP gamma-aminobutyric acid (GABA) A receptor, pi	AA101225	-0.84	0.00
	259	SLC5A6 solute carrier family 5 (sodium-dependent vitamin transporter), member 6	AA186605	-0.84	0.00
	68	CSDA cold shock domain protein A	AA455300	-0.85	0.00
	242	CDH3 cadherin 3, type 1, P-cadherin (placental)	AA425217	-0.86	0.00
	258	B3GNT5 UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 5	AA043551	-0.86	0.00
	222	LAD1 ladinin 1	T97710	-0.87	0.00
	213	KIAA1691 KIAA1691 protein	N58487	-0.87	0.00
	96	ITM3 integral membrane protein 3	AA034213	-0.88	0.00
	204	NSEP1 nuclease sensitive element binding protein 1	AA599175	-0.90	0.00
	250	MFGE8 milk fat globule-EGF factor 8 protein	AA054753	-0.93	0.00
	183	PLOD procollagen-lysine, 2-oxoglutarate 5-dioxygenase (lysine hydroxylase, Ehlers-Danlos syndrome type VI)	AA476240	-0.96	0.00
	251	ZDHHC5 **zinc finger, DHHC domain containing 5	AA448941	-1.08	0.00
	210	FLJ12442 hypothetical protein FLJ12442	R17469	-1.10	0.00
	252	CX3CL1 chemokine (C-X3-C motif) ligand 1	R66139	-1.18	0.00

Table 2b. Collection of uni-gene markers for the Luminal B phenotype. The markers are sorted in decreasing order with respect to to the signal-to-noise ratio.

Group	Gene index	Gene Description	GeneBank Acc	SNR	FDR
Core B	192	SDHA succinate dehydrogenase complex, subunit A, flavoprotein (Fp)	T70043	1.14	0.00
	138	ADRM1 adhesion regulating molecule 1	T46897	1.10	0.00
	219	SQLE squalene epoxidase	R01118	1.07	0.00
	205	GGH gamma-glutamyl hydrolase (conjugase, folylpolygammaglutamyl hydrolase)	AA455800	1.03	0.00
	206	LC27 putative integral membrane transporter	AA600214	0.88	0.00
	137	MGC2477 hypothetical protein MGC2477	T49801	0.84	0.00
	195	MDS029 uncharacterized hematopoietic stem/progenitor cells protein MDS029	AA431199	0.82	0.00
	280	KCNK1 potassium channel, subfamily K, member 1	N62620	0.80	0.00
	426	Homo sapiens mRNA; cDNA DKFZp434E033 (from clone DKFZp434E033)	N63001	-0.82	0.00
	442	CEGP1 CEGP1 protein	W74079	-0.83	0.00
	477	FLJ10948 hypothetical protein FLJ10948	T71152	-0.83	0.00
	351	PON3 paraoxonase 3	R95740	-0.88	0.00
	266	LAMC2 laminin, gamma 2	AA677534	-0.88	0.00
	324	PAM peptidylglycine alpha-amidating monooxygenase	R66309	-0.89	0.00
	332	Homo sapiens cDNA FLJ37284 fis, clone BRAMY2013590	N89738	-0.93	0.00

Table 2c. Collection of uni-gene markers for the ERBB2+ phenotype. The markers are sorted in decreasing order with respect to to the signal-to-noise ratio.

Group	Gene index	Gene Description	GeneBank Acc	SNR	FDR
Core C	7	ERBB2 v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	AA480116	2.27	0.00
	9	**Homo sapiens mRNA; cDNA DKFZp761B0319 (from clone DKFZp761B0319)	AA504615	1.84	0.00
	10	TBPL1 TBP-like 1	AA448001	1.58	0.00
	5	TRAP100 thyroid hormone receptor-associated protein (100 kDa)	N54470	1.43	0.00
	213	KIAA1691 KIAA1691 protein	N58487	1.39	0.00
	8	GRB7 growth factor receptor-bound protein 7	H53702	1.21	0.00
	235	KIAA1971 **similar to junction-mediating and regulatory protein p300 JMY	N71692	1.12	0.00
	304	LOX lysyl oxidase	AA037732	1.08	0.00
	306	OSF-2 osteoblast specific factor 2 (fasciclin I-like)	AA598653	1.02	0.00
	485	FLNB filamin B, beta (actin binding protein 278)	AA486238	-1.01	0.00
	270	CABC1 chaperone, ABC1 activity of bc1 complex like (S. pombe)	H67202	-1.06	0.00
	104	H2BFQ H2B histone family, member Q	AA010223	-1.06	0.00
	300	CDC42EP4 CDC42 effector protein (Rho GTPase binding) 4	W32509	-1.08	0.00
	111	FLJ10509 hypothetical protein FLJ10509	R18902	-1.11	0.00

Table 2d. Collection of uni-gene markers for the Basal phenotype. The markers are sorted in decreasing order with respect to to the signal-to-noise ratio.

Group	Gene index	Gene Description	GeneBank Acc	SNR	FDR
Core D	258	B3GNT5 UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 5	AA043551	1.10	0.00
	254	Homo sapiens cDNA FLJ11796 fis, clone HEMBA1006158, highly similar to Homo sapiens transcription factor forkhead-like 7 (FKHL7) gene	N22552	1.07	0.00
	256	ESTs	AA074677	1.03	0.00
	183	PLOD procollagen-lysine, 2-oxoglutarate 5-dioxygenase (lysine hydroxylase, Ehlers-Danlos syndrome type VI)	AA476240	0.92	0.00
	257	FLJ14525 hypothetical protein FLJ14525	AA464028	0.85	0.00
	255	CHI3L2 chitinase 3-like 2	AA668821	0.85	0.00
	175	SIAT4C sialyltransferase 4C (beta-galactosidase alpha-2,3-sialytransferase)	AA453813	0.84	0.00
	215	CCNE1 cyclin E1	T54121	0.84	0.00
	172	STK38 serine/threonine kinase 38	AA521346	0.84	0.00
	199	BTG3 BTG family, member 3	N52496	0.84	0.00
	273	DGUOK deoxyguanosine kinase	R07506	0.83	0.00
	228	TP53BP2 tumor protein p53 binding protein, 2	H69077	0.81	0.00
	204	NSEP1 nuclease sensitive element binding protein 1	AA599175	0.79	0.00
	238	FLJ10697 hypothetical protein FLJ10697	H80748	0.76	0.00
	272	PRAME preferentially expressed antigen in melanoma	AA598817	0.76	0.00
	243	SLPI secretory leukocyte protease inhibitor (antileukoproteinase)	AA026192	0.75	0.00
	268	CP ceruloplasmin (ferroxidase)	H86554	0.75	0.00
	259	SLC5A6 solute carrier family 5 (sodium-dependent vitamin transporter), member 6	AA186605	0.72	0.00
	231	CDK2AP1 CDK2-associated protein 1	R78607	0.72	0.00
	224	MAFG v-maf musculoaponeurotic fibrosarcoma oncogene homolog G (avian)	AA045436	0.71	0.00
	269	RCL putative c-Myc-responsive	AA132086	0.70	0.00
	226	TMSNB thymosin, beta, identified in neuroblastoma cells	N91887	0.70	0.00
	217	LANP-L lecuine-rich acidic protein-like protein	AA130595	0.70	0.00
	245	GABRP gamma-aminobutyric acid (GABA) A receptor, pi	AA101225	0.69	0.00
	233	S100A11 S100 calcium binding protein A11 (calgizzarin)	AA464731	0.68	0.00
	185	PTK7 **PTK7 protein tyrosine kinase 7	AA453789	0.68	0.00
	173	DKFZP434L0718 hypothetical protein DKFZp434L0718	AA437140	0.67	0.00
	239	Homo sapiens cDNA FLJ31360 fis, clone MESAN2000572	AA031989	0.67	0.00
	222	LAD1 ladinin 1	T97710	0.66	0.00
	506	CRAT carnitine acetyltransferase	AA621218	-0.65	0.00
	394	RGS5 regulator of G-protein signalling 5	AA668470	-0.66	0.00
	428	KIAA0876 KIAA0876 protein	AA431721	-0.66	0.00
	431	NAT1 N-acetyltransferase 1 (arylamine N-acetyltransferase)	T67128	-0.66	0.00
	443	ACADSB acyl-Coenzyme A dehydrogenase, short/branched chain	H95792	-0.67	0.00
	458	FMOD fibromodulin	AA485748	-0.68	0.00
	442	CEGP1 CEGP1 protein	W74079	-0.70	0.00
	454	HIS1 HMBA-inducible	N21081	-0.70	0.00
	498	ESTs	N73949	-0.70	0.00
	488	ECE1 endothelin converting enzyme 1	H18427	-0.71	0.00
	457	RNASE4 ribonuclease, RNase A family, 4	T60163	-0.71	0.00
	452	PLAT plasminogen activator, tissue	AA447797	-0.73	0.00
	421	Homo sapiens mRNA; cDNA DKFZp313L231 (from clone DKFZp313L231)	AA029948	-0.76	0.00
	346	HRASLS3 HRAS-like suppressor 3	AA476438	-0.77	0.00
	425	LOC51313 **AD021 protein	N95180	-0.78	0.00
	501	MRPS14 **mitochondrial ribosomal protein S14	T51290	-0.79	0.00
	387	PRO1489 hypothetical protein PRO1489	AA131299	-0.80	0.00
	500	SLC11A3 solute carrier family 11 (proton-coupled divalent metal ion transporters), member 3	AA056733	-0.80	0.00
	495	CYB5 cytochrome b-5	R91950	-0.82	0.00
	429	FLJ10980 hypothetical protein FLJ10980	N45467	-0.83	0.00
	444	FBP1 fructose-1,6-bisphosphatase 1	AA699427	-0.83	0.00
	440	PTP4A2 protein tyrosine phosphatase type IVA, member 2	AA504327	-0.85	0.00
	439	RAB5EP rabaptin-5	AA428477	-0.88	0.00
	436	FLT1 fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)	AA058828	-0.91	0.00
	502	DKFZp586H0623 putative UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase T9	T51229	-0.94	0.00
	447	**Homo sapiens cDNA FLJ11796 fis, clone HEMBA1006158, highly similar to Homo sapiens transcription factor forkhead-like 7 (FKHL7) gene	AA495790	-1.11	0.00
	445	MGC27171 hypothetical protein MGC27171	R23619	-1.14	0.00
	434	HNF3A hepatocyte nuclear factor 3, alpha	T74639	-1.18	0.00
	435	XBP1 X-box binding protein 1	W90128	-1.20	0.00
	437	GATA3 GATA binding protein 3	H72474	-1.27	0.00
	433	Homo sapiens, clone MGC:22588 IMAGE:4696566, mRNA, complete cds	N74131	-1.46	0.00

Table 2e. Collection of uni-gene markers for the Normal phenotype. The markers are sorted in decreasing order with respect to to the signal-to-noise ratio.

Group	Gene index	Gene Description	GeneBank Acc	SNR	FDR
Core E	252	CX3CL1 chemokine (C-X3-C motif) ligand 1	R66139	1.38	0.00
	477	FLJ10948 hypothetical protein FLJ10948	T71152	1.25	0.00
	488	ECE1 endothelin converting enzyme 1	H18427	1.18	0.00
	317	EPAC Rap1 guanine-nucleotide-exchange factor directly activated by cAMP	AA453497	1.18	0.00
	249	KRT17 keratin 17	AA026100	1.16	0.00
	478	Homo sapiens cDNA: FLJ22566 fis, clone HSI01980	AA054715	1.15	0.00
	457	RNASE4 ribonuclease, RNase A family, 4	T60163	1.08	0.00
	384	GSTM1 glutathione S-transferase M1	AA290737	1.07	0.00
	329	Ells1 hypothetical protein Ells1	N35592	1.02	0.00
	248	TRIM29 tripartite motif-containing 29	AA055485	1.00	0.00
	474	ACADVL acyl-Coenzyme A dehydrogenase, very long chain	AA464163	0.97	0.00
	517	APOD apolipoprotein D	AA456975	0.95	0.00
	219	SQLE squalene epoxidase	R01118	-0.96	0.00
	148	LOC55829 AD-015 protein	W69583	-1.00	0.00
	156	no_name_3	AA598508	-1.01	0.00
	50	UNG uracil-DNA glycosylase	H15111	-1.02	0.00
	91	TAP1 transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)	AA487429	-1.05	0.00
	203	EBNA1BP2 EBNA1 binding protein 2	R45255	-1.06	0.00
	207	PRDX4 peroxiredoxin 4	AA459663	-1.18	0.00
	41	ARPC5 actin related protein 2/3 complex, subunit 5, 16kDa	W55964	-1.35	0.00

(Luminal B, ERBB2+ and Normal) had far fewer gene markers (15 for Luminal B, 14 for ERBB2+ and 20 for Normal core clusters). These genes are listed in Tables 2 a–d and those also identified in Sorlie et al. (2003) are marked with a*. A heat map of the core clusters using these 148 genes is shown in Figure 2.

Patterns (Multi-gene Markers) for the Core Clusters

The complexity of BCA makes it unlikely that single genes can predict phenotype. Instead, one expects combinations of genes to be better at identifying phenotype. Consequently, we used " patterns" (as defined in Crama et al. 1988; Alexe and Hammer, 2005; Bhanot et al. 2005) to distinguish the core clusters. A pattern is a set of linear constraints on the expression levels of a group of genes satisfied by many samples in a particular cluster and by few samples in other clusters. For example, the pattern $\boldsymbol{P_A}$ below is satisfied by all samples in the "Luminal A" cluster and by none of the non-Luminal A samples:

$$\boldsymbol{P_A} = [\text{Expression of } \textit{GATA3} \geq 0.49\ ].\text{AND}.\ [\text{Expression of } \textit{Liv-1} \geq -0.25]$$

For illustration, Figure 3 shows two patterns $\boldsymbol{P_A}$ and $\boldsymbol{N_A}$, in the 2-d expression plane for *GATA3* and *Liv*-1.

A pattern is characterized by its degree, prevalence, and homogeneity. The *degree* is the number of genes appearing in its defining conditions. The *prevalence* of a pattern is the percent of positive (negative) cases which satisfy the pattern. The *homogeneity* of a pattern is the percentage of positive (negative) cases covered by it. In general, patterns useful for classification have low degree and high prevalence and homogeneity.

We identified all patterns for the 60 core samples over the selected 148 genes by applying the combinatorial algorithm described in (Alexe and Hammer, 2005). Briefly, each sample from a core cluster was placed in a box by defining cuts in gene expression space which distinguish it from the samples belonging to other core clusters. The boxes were then merged by extending them along

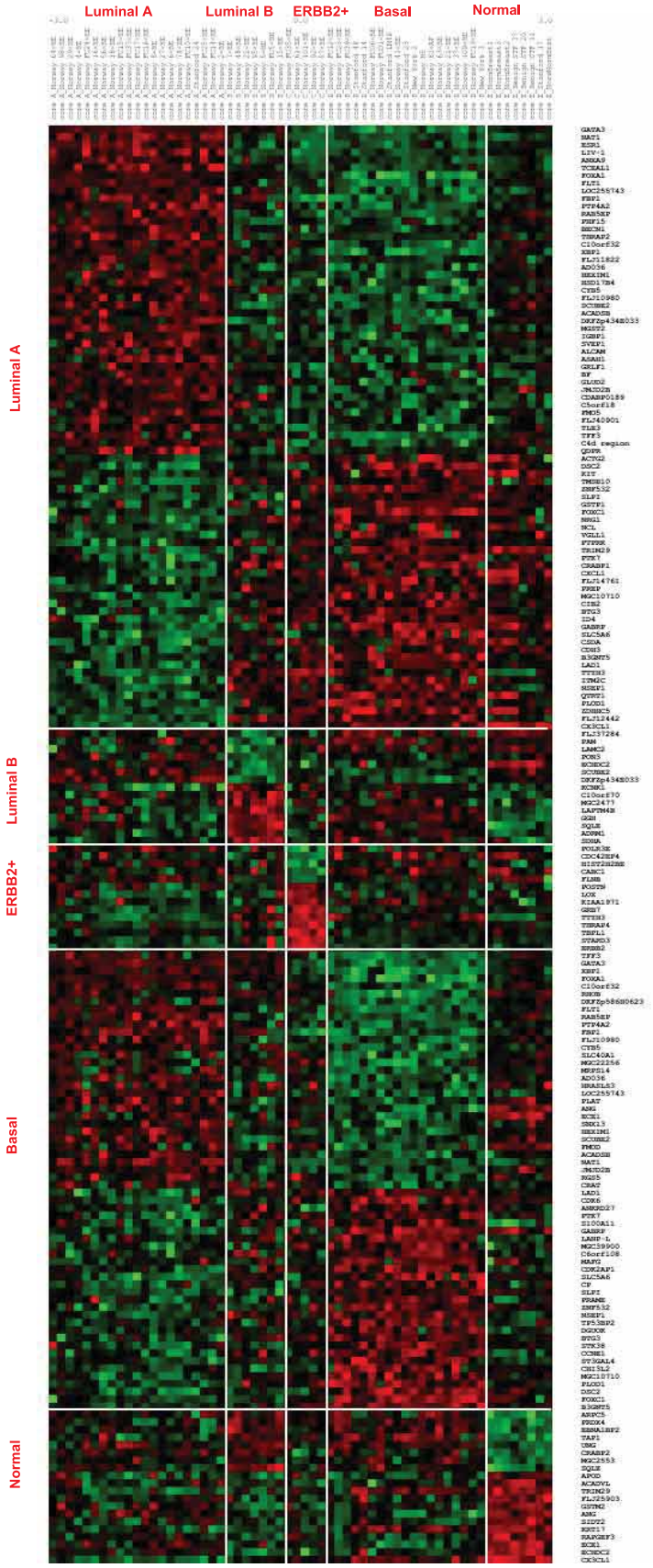

Figure 2. Heatmap of 148 uni-genes for the samples in core categories.

all possible dimensions without allowing any member of the opposite class to be included in the box. The maximal boxes so obtained defined the patterns.

The pattern parameters (degree, prevalence, and homogeneity) were determined by estimating the classification accuracy of a weighted-voting model constructed on pattern data through 10-fold cross-validation experiments. Pattern-based weighted voting is a meta-classification scheme in which individual patterns are "voters" for a phenotype. The performance of a multi-pattern meta-classification system is better than the performance of single patterns if the patterns are uncorrelated (Merz, 1998). Uncorrelated patterns were selected by requiring the patterns to be defined on non-overlapping subsets of features. To avoid over-fitting, the patterns were required to use no more than five genes each.

We found many patterns of degree 2 and 3 for each phenotype, each of which was common to more than 90% of the samples in the cores. Table 3 presents some of these patterns. The striking feature of Table 3 is that simple conditions on a few genes are able to generate a very clean classification in the cores. *Several genes occurred frequently in the patterns, suggesting an active association with disease*. For example, KIAA1691, PREP, CX3CL1, LIV-1, PLOD, GATA-3 occur in 20% of patterns for Luminal A, while PRAME, PLAT, CCNE1, FKHL7, clone MGC:22588 IMAGE:4696566, occur in 15% of the patterns for the Basal group. There are also several genes which are good uni-gene markers but are not found in patterns.

Consistency of Core Assignments Using Either Patterns or Clustering

A positive pattern is a set of conditions satisfied by a sample that belongs to a core cluster. A negative pattern is a set of conditions satisfied by a sample that belongs to the complement of the core cluster. For each unlabeled sample we counted the number of positive minus the number of negative patterns satisfied by it for each core cluster. The sample was assigned to the core cluster for which the ratio obtained by dividing this number to the total number of patterns for the core cluster, was positive and maximum. If the maximum ratio was negative or if it was assigned to multiple core clusters then the sample remained unclassified (Alexe et al. 2005c). The classification of samples to cores was validated using leave-one-out experiments on patterns. Over the sixty samples in the cores, in each such experiment, the entire procedure (gene selection, pattern extraction and sample classification) was repeated sixty times, once for each omitted sample.

A comparison of our clustering and pattern assignments with the original classification is presented in Table 4. The color scheme is that if the sample is robustly assigned to a phenotype, its entry is the color of that phenotype. Samples whose classification is either poor or ambiguous are in black or left blank respectively. When the

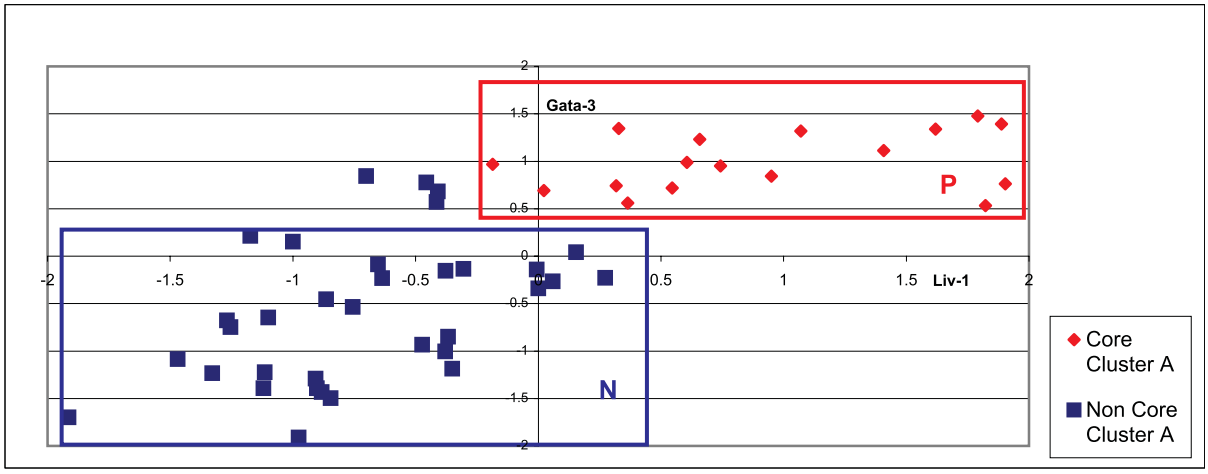

Figure 3. An example of a pattern (pattern P_A) characteristic of the Luminal A core cluster (Cluster A) and an example of a pattern (pattern N_A) characteristic of the non-Luminal A cases. Notice that ***P*** is satisfied by all the samples in the Luminal A group, while ***N*** is satisfied by 88% of the non-Luminal A cases. Both patterns ***P*** and ***N*** are expressed as bounding constraints on the expressions of genes Liv-1 and Gata-3.

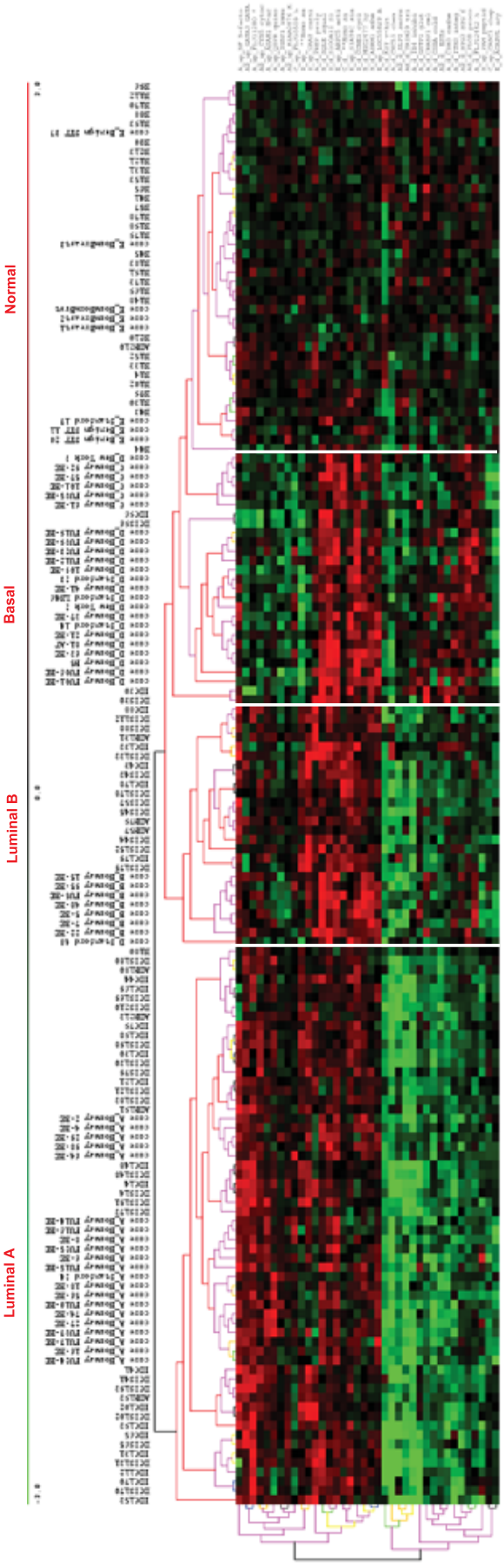

Figure 4. Heatmap of combined Ma et al. and Sorlie et al. data using the 38 genes identified in the latter data. There are four distinct clusters which are separtaed by vertical lines in the plot. The Normals, Luminal A and Basal core samples from Sorlie et al. cluster well enough with samples in the Ma et al. data to make a phenotype identification possible for the latter data. The B core cluster (Luminal B) looks similar to the Luminal A core cluster with some genes over expressed. Core cluster C (ERBB2+) is most similar to Core D (Basal) presumably because the discriminator gene ERBB2 gene is not on the Ma et al. chip set . The sample labels in the Ma et al. data indicate stages of disease (ADH, DCIS or IDC) and the index number of the patient. Notice that samples from the same patient, even if in different stages of BCA, cluster together.

Table 3. Collections of patterns for the breast cancer phenotypes.

Patterns core A	Gene description	GLUD1 glutamate dehydrogenase 1	CRABP1 cellular retinoic acid binding protein 1	KIAA1691 KIAA1691 protein	CDH3 cadherin 3, type 1, P-cadherin (placental)	TRIM29 tripartite motif-containing 29	MFGE8 milk fat globule-EGF factor 8 protein	FLJ11280 **hypothetical protein FLJ11280	LIV-1 LIV-1 protein, estrogen regulated	ESR1 estrogen receptor 1
	GeneBank acc	AA017175	AA454702	N58487	AA425217	AA055485	AA054753	N54608	H29315	AA291702
	Prevalence (%)									
P1	100		≤0.99		≤0.58			>−0.29		
P2	100	>−0.68		≤0.55			≤0.2			
P3	100					≤1.13			>−0.47	>−0.24

Patterns core B	Gene description	SDHA succinate dehydrogenase complex, subunit A, flavoprotein (Fp)	LC27 putative integral membrane transporter	SQLE squalene epoxidase	LAMC2 laminin, gamma 2	KCNK1 potassium channel, subfamily K, member 1	PAM peptidylglycine alpha-amidating monooxygenase	Homo sapiens cDNA FLJ37284 fis, clone BRAMY2013590	PON3 paraoxonase 3	FLJ10948 hypothetical protein FLJ10948
	GeneBank acc	T70043	AA600214	R01118	AA677534	N62620	R66309	N89738	R95740	T71152
	Prevalence (%)									
P1	100	>0.2			≤−0.21	>−0.44				
P2	100						≤0.04	≤0.03		≤0.1
P3	100		>0.46	>0.83					≤0	

Patterns core C	Gene description	ERBB2 v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	TRAP100 thyroid hormone receptor-associated protein (100 kDa)	KIAA1691 KIAA1691 protein	KIAA1971 **similar to junction-mediating and regulatory protein p300 JMY	CABC1 chaperone, ABC1 activity of bc1 complex like (S. pombe)	CDC42EP4 CDC42 effector protein (Rho GTPase binding) 4	FLNB filamin B, beta (actin binding protein 278)
	GeneBank acc	AA480116	N54470	N58487	N71692	H67202	W32509	AA486238
	Prevalence (%)							
P1	100	>0.05	>−0.22		>0.67			
P2	100					≤−0.02	≤−0.55	≤−0.02
P3	100	>0.05		>0.9				

Patterns core D	Gene description	Homo sapiens, clone MGC:22588 IMAGE:4696566, mRNA, complete cds	RAB5EP rabaptin-5	S100A11 S100 calcium binding protein A11 (calgizzarin)	HRASLS3 HRAS-like suppressor 3	Homo sapiens mRNA; cDNA DKFZp313L231 (from clone DKFZp313L231)	DKFZp586H0623 putative UDP-GalNAc:polypeptide Nacetylgalactosaminyl transferase T9	CRAT carnitine acetyltransferase
	GeneBank acc	N74131	AA428477	AA464731	AA476438	AA029948	T51229	AA621218
	Prevalence (%)							
P1	100	≤0.08		>−0.9		≤0.53		
P2	100		≤−0.19				≤−0.22	≤0.33
P3	100	≤0.08	≤−0.19		≤0.59			

Patterns core E	Gene description	TAP1 transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)	KRT17 keratin 17	CX3CL1 chemokine (C-X3-C motif) ligand 1	ACADVL acyl-Coenzyme A dehydrogenase, very long chain	ECE1 endothelin converting enzyme 1	APOD apolipoprotein D
	GeneBank acc	H15111	AA055485	AA026100	T60163	AA054715	H18427
	Prevalence (%)						
P1	100	≤0.32		>1.15			
P2	100		>−0.41		>0	>0.2	
P3	100						>−0.3

Table 4. Phenotype classification of breast cancer based on core clusters and pattern scores.

Sample id	P1 P2 P3	P1 P2 P3	P1 P2 P3	P1 P2 P3	P1 P2 P3	Sorlie et al Clusters	Core clusters	Classification based on core cluster scores	Classification based on pattern models
Norway 64-BE						A	core A	A	A
Norway 98-BE						A	core A	A	A
Norway 29-BE						A	core A	A	A
Norway 4-BE						A	core A	A	A
Norway FU24-BE						A	core A	A	A
Norway 16-BE						A	core A	A	A
Norway 56-BE						A	core A	A	A
Norway 18-BE						A	core A	A	A
Norway FU15-BE						A	core A	A	A
Norway FU37-BE						A	core A	A	A
Norway FU17-BE						A	core A	A	A
Norway FU16-BE						A	core A	A	A
Norway 8-BE						A	core A	A	A
Norway 27-BE						A	core A	A	A
Norway 6-BE						A	core A	A	A
Norway 74-BE						A	core A	A	A
Norway FU10-BE						A	core A	A	A
Stanford 24						A	core A	A	A
Norway FU25-BE						A	core A	A	A
Norway FU14-BE						A	core A	A	A
Norway 2-BE						A	core A	A	A
Stanford LN4						A		E	A
New York 1						A		E	A
Stanford 38						A		E	
Stanford 31						A		E	
Stanford 18						A		E	
Stanford 40						A		E	
Norway FU40-BE						A			
Norway 7-BE						B	core B	B	B
Norway 48-BE						B	core B	B	B
Norway 22-BE						B	core B	B	B
Norway 95-BE						B	core B	B	B
Norway 5-BE						B	core B	B	B
Norway FU5-BE						B	core B	B	B
Norway 15-BE						B	core B	B	B
Norway 26-BE						B			
Norway 19-BE						B			
Norway 102-BE						B			A
Norway 11-BE						B			
Norway FU35-BE						C	core C	C	C
Norway 61-BE						C	core C	C	C
Norway 101-BE						C	core C	C	C
Norway 92-BE						C	core C	C	C
Norway 57-BE						C	core C	C	C
Norway FU27-BE						C			
Norway FU18-BE						C			E
Norway FU04-BE						C			D
Norway 65-2nd T						C			
Norway FU44-BE						C			
Norway FU30-AF						C			B
Norway FU12-BE						D	core D	D	D
Norway FU23-BE						D	core D	D	D
Norway FU39-BE						D	core D	D	D
Stanford 48						D	core D	D	D
Stanford 14						D	core D	D	D
Norway FU06-BE						D	core D	D	D
Norway FU01-BE						D	core D	D	D
Stanford LN46						D	core D	D	D
Norway 41-BE						D	core D	D	D
Stanford 23						D	core D	D	D
New York 2						D	core D	D	D
Norway H5						D	core D	D	D
Norway 81-AF						D	core D	D	D
Norway 63-BE						D	core D	D	D
Norway 21-BE						D	core D	D	D
Norway 37-BE						D	core D	D	D
Norway 109-BE						D	core D	D	D
Norway FU19-BE						D	core D	D	D
New York 3						D	core D	D	D
NormBreast1						E	core E	E	E
NormBreast3						E	core E	E	E
NormBreast2						E	core E	E	E
Benign STF 37						E	core E	E	E
Benign STF 20						E	core E	E	E
Benign STF 11						E	core E	E	E
Stanford 17						E	core E	E	E
NorwNormBrst						E	core E	E	E
Norway H2						E		A	
Norway 112-BE						E			

Table 5. Phenotype prediction for previously unassigned breast cancer samples.

Sample id	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3	Sorlie et al Clusters	Core clusters	Classification based on core cluster scores	Classification based on pattern models
Norway 51-BE																			A
Stanford 16																			
Norway 39-BE																			
Norway 17-BE																			
Norway 10-BE																			A
Norway 43-AF																		A	A
Norway 32-BE																		A	A
Norway 85-BE																		A	A
Norway FU09-BE																		A	A
Norway 83-BE																		A	
Norway FU22-BE																		A	
Stanford 6																			E
Stanford 35																			A
Norway 75-BE																			
NorwayFU43-BE																		E	
Norway 96-AF																		A	A
Norway 90-BE																		A	A
Norway 100-BE																			A
Norway 111-BE																		A	
Norway 24-BE																		A	A
Norway 104-BE																			
Norway H4-T1																		A	A
Norway H3																		A	
Norway FU41-BE																			
Norway FU29-BE																		B	B
Norway FU08-BE																		B	B
Stanford A																			A
Norway 80-BE																		D	D
Norway H6																		E	
Norway 53-BE																		C	
Norway 47-BE																		C	
Norway FU26-BE																		E	
Norway 14-BE																		E	D
Norway FU45-BE																			
Norway 55-BE																			E
Stanford 2																		D	D
Stanford 45																			
Stanford 44																			E
Norway FU07-BE																		D	

Table 6. Classification accuracy of pattern models through leave-one-out cross validation experiments.

Phenotype	Sensitivity (%)	Specificity (%)
Core A	100.00	97.44
Core B	71.42	98.11
Core C	80.00	98.18
Core D	100.00	100.00
Core E	87.50	96.15
Average	87.78	97.98
Std. Deviation	12.52	1.39
Confidence Interval (95%)	75.26 - 100.00	96.58 - 99.37

pattern and cluster classifiers agree, the assignment can be considered accurate. When they differ, no classification is possible. From a treatment perspective, the recommendation of such an inconclusive assignment would be retesting. The clustering and patterns classifiers for the unassigned samples in the Sorlie et al. paper are shown in Table 5. Some of these originally unassigned samples are assigned to a consistent phenotype by our methods.

Table 6 summarizes the sensitivity and specificity of the pattern based classifier showing once again the robustness of the classification into phenotypes Normal, Luminal A and Basal and the unreliability of the other two phenotype classifications.

Validation on an External Dataset Data 2

We used the markers identified in Data 1 to classify samples in Data 2. These two datasets had 93 genes in common. Of these, 79 were in our 391 uni-gene set and a subset of 38 of these were in the smaller subset of 148 genes. Of the latter, 23 were markers for Luminal A, 4 were markers for Luminal B, 3 were markers for ERBB2+ and 12 were markers for the Basal group. For each of the 38 genes, we normalized the data sets relative to each other by equating the average intensity of each gene for the normal samples in the two data sets. In each dataset, the expression level of each gene was replaced with its quartile value across all samples. We recomputed a pattern-based classifier trained on the known core clusters in the Sorlie et al. (2003) data and used it to predict the phenotype for Ma et al. 2003 samples.

Figure 4 shows a heat map of the 38 genes in common between the datasets. This plot includes all core samples from Data 1 and all samples from Data 2. The Normal samples from both sets cluster nicely showing that the global normalization was done correctly. The Luminal A cluster is easily identified because all Luminal A core samples from Data 1 cluster together with several samples from Data 2. There is also a distinct Basal cluster with most Data 1 Basal samples and a few Data 2 samples on its edges. Finally, there is another cluster with some Core B samples which looks quite similar to Luminal A. The core C samples are mixed in with the Basal cluster (as was already noticed in Figure 1c). We conclude that it is not possible to assign Luminal B or ERBB2+ phenotypes to samples in Data 2 based on Data 1 because a) There are very few genes in these categories (3/38 for ERBB2+ and 4/38 for Luminal B), b) the ERBB2 gene is missing in Data 2 and c) The quality of the patterns using the 38 genes for these two phenotypes is poor. Indeed, for core C, there are no patterns at all and for core B, the patterns are of poor statistical quality.

To further validate the consistency of our assignments, we trained a pattern-based classification model on quartile discretized Data 1 samples and used it to predict the phenotype for the samples in Data 2 using majority voting. When the prediction from patterns agreed with the prediction from clustering as in Figure 4, we felt confident of the diagnosis, otherwise not. Our predicted phenotypes for the Ma et al. data are given in Table 7.

Pathways for each Core

To identify processes/pathways that are common and particular to the different phenotypes, we used the bioinformatics public resources DAVID (Dennis et al. 2003), BioRag (Pandey et al. 2004),

Table 7. Predicted phenotype for samples in Ma et al. data using patterns from core clusters in Sorlie et al. 2003. We are confident of the phenotype assignment for those samples marked in color in columns 9 and 10.

Sample label	Case ID	Stages Microdissected	Age	ER	PR	HER2	Node[a]	Predicted phenotype	
								Cluster scores	Pattern model
DCIS14	14	N, DCIS (I), IDC (I)	44	Pos	Pos	ND	ND	A	A
IDC14	14	N, DCIS (I), IDC (I)	44	Pos	Pos	ND	ND	A	A
DCIS30	30	N, DCIS (III), IDC (III)	47	Neg	Neg	Neg	Neg	D	D
IDC30	30	N, DCIS (III), IDC (III)	47	Neg	Neg	Neg	Neg	D	D
DCIS41	41	N, DCIS (II), IDC (II)	55	Pos	Pos	ND	ND	A	A
IDC41	41	N, DCIS (II), IDC (II)	55	Pos	Pos	ND	ND	A	A
DCIS43	43	N, DCIS (II), IDC (II)	53	Pos	Neg	Neg	Neg	C	
IDC43	43	N, DCIS (II), IDC (II)	53	Pos	Neg	Neg	Neg	B	
DCIS44	44	N, DCIS (III), IDC (III)	28	Pos	Pos	Neg	Neg	C	
IDC44	44	N, DCIS (III), IDC (III)	28	Pos	Pos	Neg	Neg	A	
DCIS45	45	N, DCIS (I)	36	Pos	Neg	Neg	Neg	A	
ADH57	57	N, DCIS (I)	36	Pos	Neg	Neg	Neg	E	D
DCIS57	57	N, DCIS (I)	36	Pos	Neg	Neg	Neg	A	D
DCIS65	65	N, DCIS (III), IDC (III)	39	Pos	Pos	Neg	Neg	A	A
IDC65	65	N, DCIS (III), IDC (III)	39	Pos	Pos	Neg	Neg	A	A
ADH79	79	N, ADH, DCIS (I), IDC (I)	54	Pos	Pos	Neg	Neg	E	
DCIS79	79	N, ADH, DCIS (I), IDC (I)	54	Pos	Pos	Neg	Neg	A	D
IDC79	79	N, ADH, DCIS (I), IDC (I)	54	Pos	Pos	Neg	Neg	A	
DCIS88	88	N, DCIS (III), IDC (III)	35	Pos	Pos	ND	ND	E	
IDC88	88	N, DCIS (III), IDC (III)	35	Pos	Pos	ND	ND	A	
DCIS96	96	N, DCIS (III), IDC (III)	31	Neg	Neg	Neg	Neg	D	D
IDC96	96	N, DCIS (III), IDC (III)	31	Neg	Neg	Neg	Neg	D	D
DCIS102	102	N, DCIS (I), IDC (I)	55	Pos	Neg	Neg	Neg	A	A
IDC102	102	N, DCIS (I), IDC (I)	55	Pos	Neg	Neg	Neg	A	A
DCIS112	112	N, DCIS (III), IDC (III)	31	Neg	Pos	Neg	Neg	A	
IDC112	112	N, DCIS (III), IDC (III)	31	Neg	Pos	Neg	Neg	A	
DCIS121	121	N, DCIS (II), IDC (II)	45	Pos	Pos	Pos	Pos	A	
IDC121	121	N, DCIS (II), IDC (II)	45	Pos	Pos	Pos	Pos	A	A
DCIS130	130	N, DCIS (II), IDC (II)	54	Pos	Pos	Neg	Neg	A	
IDC130	130	N, DCIS (II), IDC (II)	54	Pos	Pos	Neg	Neg	A	
ADH131	131	N, ADH, DCIS (II), IDC (II)	37	Pos	Pos	Pos	Pos	E	D
DCIS131	131	N, ADH, DCIS (II), IDC (II)	37	Pos	Pos	Pos	Pos	A	A
IDC131	131	N, ADH, DCIS (II), IDC (II)	37	Pos	Pos	Pos	Pos	A	A
DCIS133	133	N, DCIS (III), IDC (III)	44	Neg	Neg	Pos	Pos	C	D
IDC133	133	N, DCIS (III), IDC (III)	44	Neg	Neg	Pos	Pos	D	D
DCIS148	148	N, DCIS (II), IDC (II)	42	Pos	Pos	Neg	Neg	A	A
IDC148	148	N, DCIS (II), IDC (II)	42	Pos	Pos	Neg	Neg	A	A
DCIS152	152	N, DCIS (II), IDC (II)	42	Pos	Pos	Neg	Neg	A	
IDC153	153	N, IDC (I)	46	Pos	Pos	Pos	Pos	A	A
DCIS169	169	N, DCIS (II), IDC (II)	34	Pos	Pos	Neg	Neg	A	A
IDC169	169	N, DCIS (II), IDC (II)	34	Pos	Pos	Neg	Neg	A	A
DCIS170	170	N, DCIS (II), IDC (II)	44	Pos	Pos	Pos-FISH	Pos-FISH	A	
IDC170	170	N, DCIS (II), IDC (II)	44	Pos	Pos	Pos-FISH	Pos-FISH	A	
DCIS173	173	N, DCIS (I), IDC (I)	52	Pos	Pos	Neg	Neg	A	A
DCIS178	178	N, DCIS (III), IDC (III)	43	Pos	Pos	Pos	Pos	A	
IDC178	178	N, DCIS (III), IDC (III)	43	Pos	Pos	Pos	Pos	A	
DCIS179	179	N, DCIS (III), IDC (III)	37	Neg	Neg	Pos-FISH	Pos-FISH	C	
IDC179	179	N, DCIS (III), IDC (III)	37	Neg	Neg	Pos-FISH	Pos-FISH	C	
ADH180	180	N, ADH, DCIS (I), IDC (I)	46	Pos	Pos	Neg	Neg	A	A
DCIS180	180	N, ADH, DCIS (I), IDC (I)	46	Pos	Pos	Neg	Neg	A	
DCIS183	183	N, DCIS (II)	46	ND	ND	ND	ND	A	D
ADH191	191	N, DCIS (II)	46	ND	ND	ND	ND	A	A
DCIS191	191	N, DCIS (II)	46	ND	ND	ND	ND	A	A
ADH193	193	N, ADH, DCIS (I), IDC (I)	45	Pos	Pos	Neg	Neg	A	A
DCIS193	193	N, ADH, DCIS (I), IDC (I)	45	Pos	Pos	Neg	Neg	A	A
IDC193	193	N, ADH, DCIS (I), IDC (I)	45	Pos	Pos	Neg	Neg	A	A
DCIS198	198	N, DCIS (II), IDC (II)	30	Pos	Pos	Neg	Neg	A	A
IDC198	198	N, DCIS (II), IDC (II)	30	Pos	Pos	Neg	Neg	A	
ADH210	210	N, DCIS (II), IDC (II)	30	Pos	Pos	Neg	Neg	E	D
DCIS210	210	N, DCIS (II), IDC (II)	30	Pos	Pos	Neg	Neg	A	
ADH213	213	N, DCIS (II), IDC (II)	30	Pos	Pos	Neg	Neg	A	D

Table 8. A complete listing of the associated pathways for the biomarkers available in different databases on the web (BIOCARTA, KEGG, GENMAPP).

Group	Gene description	GeneBank	Pathway	Related cancer type or pathway
core A	ESR1 estrogen receptor 1	AA291702	Nuclear_Receptors	Breast cancer related
	KIT v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	N20798	Regulation of BAD phosphorylation	Breast cancer related, Loss of c-kit expression has been reported in 80-90% of breast cancer specimens, suggesting a possible role in the development of tumors. Introduction of the c-kit gene leads to growth suppression of a breast cancer cell line, MCF-7 (Nishida et al., 1996)
	NRG1 neuregulin 1	R72075	Neuroregulin receptor degredation protein-1 Controls ErbB3 receptor recycling	Breast cancer related, direct ligand for ERBB3 and ERBB4. Indirect activator of ERBB2.
	NSEP1 nuclease sensitive element binding protein 1	AA599175	D4-GDI Signaling Pathway	Breast cancer related. Target of Akt phosphorylation. Disruption inhibits tumor growth (Sutherland et al., 2005)
	ID4 inhibitor of DNA binding 4, dominant negative helix-loop-helix protein	AA453341	TGF-beta signaling pathway	Cancer related. May contribute to rat mammary gland carcinogenesis by inhibiting mammary epithelial cell differentiation and stimulating mammary epithelial cell growth (Shan et al., 2003). Down-regulated in gastric adenocarcinoma and leukemia.
	GSTP1 glutathione S-transferase pi	R33642	Multi-Drug Resistance Factors,Glutathione metabolism	Cancer related. Lost in prostate cancer, lung cancer and squamous cell carcinoma.
	TFF3. Homo sapiens, clone MGC:22588 IMAGE:4696566, mRNA, complete cds	N74131	Trefoil Factors Initiate Mucosal Healing	Cancer related. TFF3, activates STAT3 (oncogene) signaling in human colonic cancers (Rivat et al., 2005).
	FLT1 fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)	AA058828	VEGF, Hypoxia, and Angiogenesis	Cancer related, angiogenesis.
	SLPI secretory leukocyte protease inhibitor (antileukoproteinase)	AA026192	Proepithelin Conversion to Epithelin and Wound Repair Control	Immune response related.
	BF B-factor, properdin	H80257	Complement and coagulation cascades	Immune response related.
	C4B complement component 4B	AA664406	Complement and coagulation cascades	Immune response related.
	ASAH1 N-acylsphingosine amidohydrolase (acid ceramidase) 1	AA664155	Glycosphingolipid metabolism	Anti-apoptotic. Metabolizes ceramide to sphingosine-1-phosphate (SPP), an inducer of proliferation.
	PLOD procollagen-lysine, 2-oxoglutarate 5-dioxygenase (lysine hydroxylase, Ehlers-Danlos syndrome type VI)	AA476240	Lysine degradation	tissue modelling
	ACTG2 actin, gamma 2, smooth muscle, enteric	T60048	Cholera - Infection	tissue modelling
	ACADSB acyl-Coenzyme A dehydrogenase, short/branched chain	H95792	Fatty_Acid_Synthesis,Bile acid biosynthesis	
	FBP1 fructose-1,6-bisphosphatase 1	AA699427	Glycolysis / Gluconeogenesis	
	HSD17B4 hydroxysteroid (17-beta) dehydrogenase 4	AA487914	Mechanism of Gene Regulation by Peroxisome Proliferators via PPARa(alpha), Androgen and estrogen metabolism	
	MGST2 microsomal glutathione S-transferase 2	W73474	Glutathione metabolism	
	QDPR quinoid dihydropteridine reductase	R38198	Folate biosynthesis	
	GLUD1 glutamate dehydrogenase 1	AA017175	Glutamate metabolism	
core B	GGH gamma-glutamyl hydrolase (conjugase, folylpolygammaglutamyl hydrolase)	AA455800	Folate biosynthesis	Cancer related. Identified as a biomarker for pulmonary neuroendocrine tumors (he et al., 2004)
	LAMC2 laminin, gamma 2	AA677534	Inflammatory_Response_Pathway	Cancer related. Involved in tumor invasion and metastases e.g. in pancreatic ductal adenocarcinoma (Takahashi et al., 2002) and endometrial adenocarcinomas (Maatta et al., 2004).
	SDHA succinate dehydrogenase complex, subunit A, flavoprotein (Fp)	T70043	Oxidative phosphorylation	
	PON3 paraoxonase 3	R95740	gamma-Hexachlorocyclohexane degradation	

(continued)

(continued)

Group	Gene description	GeneBank	Pathway	Related cancer type or pathway
core C	ERBB2 v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)	AA480116	Role of ERBB2 in Signal Transduction and Oncology	Breast cancer related
	F2R coagulation factor II (thrombin) receptor	AA455910	Thrombin signaling and protease-activated receptors	Breast cancer related, matrix metalloprotease-1 receptor that promotes invasion and tumorigenesis of breast cancer cells (Boire et al, 2005)
	PPARBP PPAR binding protein	T57034	CARM1 and Regulation of the Estrogen Receptor	Breast cancer related, ESR1 coactivator. Overexpressed in breast cancer. May play a role in mammary epthelial differentiation (Zhu et al., 1999)
	FLNB filamin B, beta (actin binding protein 278)	AA486238	MAPK signaling pathway	
core D	CDK6. Homo sapiens cDNA FLJ31360 fis, clone MESAN2000572	AA031989	Cyclins and Cell Cycle Regulation	Breast cancer related. CDK6 gene, inhibits proliferation of human mammary epithelial cells (Lucas et al., 2004)
	SIAT4C sialyltransferase 4C (beta-galactosidase alpha-2,3-sialytransferase)	AA453813	Steps in the Glycosylation of Mammalian N-linked Oligosaccarides	Cancer related. Down-regulated in RCC (Saito et al., 2002)
	Homo sapiens, clone MGC:22588 IMAGE:4696566, mRNA, complete cds	N74131	Trefoil Factors Initiate Mucosal Healing	Cancer related. TFF3, activates STAT3 (oncogene) signaling in human colonic cancers (Rivat et al., 2005).
	FLT1 fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)	AA058828	VEGF, Hypoxia, and Angiogenesis	Cancer related, angiogenesis.
	PLOD procollagen-lysine, 2-oxoglutarate 5-dioxygenase (lysine hydroxylase, Ehlers-Danlos syndrome type VI)	AA476240	Lysine degradation	Catalyzes the hydroxylation of lysyl residues in collagen-like peptides. The resultant hydroxylysyl groups are attachment sites for carbohydrates in collagen
	**Homo sapiens cDNA FLJ11796 fis, clone HEMBA1006158, highly similar to Homo sapiens transcription factor forkhead-like 7 (FKHL7) gene	AA495790	Integrin-mediated cell adhesion	Cancer related. RHOB protein, tumor suppressor and proapoptotic.
	SLPI secretory leukocyte protease inhibitor (antileukoproteinase)	AA026192	Proepithelin Conversion to Epithelin and Wound Repair Control	Immune response related.
	PLAT plasminogen activator, tissue	AA447797	Complement and coagulation cascades	Tissue remodelling
	FMOD fibromodulin	AA485748	Small Leucine-rich Proteoglycan (SLRP) molecules	Affects the rate of fibrils formation. May have a primary role in collagen fibrillogenesis
	DGUOK deoxyguanosine kinase	R07506	Purine metabolism	
	ACADSB acyl-Coenzyme A dehydrogenase, short/branched chain	H95792	Fatty_Acid_Synthesis,Bile acid biosynthesis	
	FBP1 fructose-1,6-bisphosphatase 1	AA699427	Glycolysis / Gluconeogenesis	
	MAFG v-maf musculoaponeurotic fibrosarcoma oncogene homolog G (avian)	AA045436	Oxidative Stress Induced Gene Expression Via Nrf2	
	CP ceruloplasmin (ferroxidase)	H86554	Porphyrin and chlorophyll metabolism	
core E	GSTM1 glutathione S-transferase M1	AA290737	Glutathione metabolism	
	ACADVL acyl-Coenzyme A dehydrogenase, very long chain	AA464163	Fatty_Acid_Synthesis,Bile acid biosynthesis	

iHOP (Hoffmann and Valencia, 2004) and BRB Tools (http://linus.nci.nih.gov/BRB-ArrayTools.html). The method used for GO functional class scoring is given in Supplementary Information II.

Table 8 is a detailed explanation of some of the 148 uni-gene biomarkers identified for each core (see also Tables 2a–d). Table 9 presents the GO categories enriched for the genes associated with the cores. The statistical significance of the enriched GO categories is computed as described in Supplementary Information II. The complete list of gene markers for the core phenotypes involved

in the enriched GO categories is available in Supplementary Table 4.

Whereas we discuss markers for each core subtype, we have strong confidence only in the markers for Luminal A and Basal.

In Luminal A, ESR1 is up-regulated, indicating that the estrogen receptor pathway is turned on.

The KIT gene was already known to be lost in breast cancer. Introduction of the c-kit gene leads to growth suppression of a breast cancer cell line, MCF-7 (Nishida et al. 1996). The Neuregulin 1 gene, which is up-regulated, is a direct ligand for ERBB3 and ERBB4, and an indirect activator of ERBB2, though the ERBB2+ subtype is identified with Cluster C. The nuclease sensitive element binding protein (NSEP1), which is also up-regulated, is known to inhibit p53 induced apoptosis (Zhang et al. 2003). It has also been recently shown to be a target of Akt phosphorylation, and that disruption of phosphorylation inhibits tumor growth (Sutherland et al. 2005). This gene is involved in D4-GDI signaling pathway, which may also be up-regulated.

A number of Luminal A markers were previously identified cancer related genes. The ID4 gene, which was also reported to be down-regulated in gastric adenocarcinoma and leukemia, may cause the alteration of the TGF-beta signaling pathway which regulates the growth and proliferation of cells, blocking the growth of many different cell types. The TGF-beta receptor includes Type I and Type II subunits that are serine-threonine kinases that signal through the Smad family of proteins. Another cancer related gene is GSTP1, which was reported to be lost in different types of cancers including prostate cancer, lung cancer and squamous cell carcinoma. Other cancer related genes include the TFF3 gene, which was shown to activate STAT3, (an oncogene) signaling in human colonic cancers (Rivat et al. 2005) and the VEGF receptor FLT1 gene.

Other Luminal A marker genes include up-regulated immune system related genes (SLPI , BF, and C4B), anti-apoptotic gene ASAH1; collagen related gene PLOD and actin gamma 2 gene. Other genes constitute mostly metabolic genes (with a significant enrichment, see Table 9), including fructose-1,6-bisphosphatase 1 (FBP1), glutamate dehydrogenase 1 (GLUD1) and acyl-Coenzyme A dehydrogenase (ACADSB).

Biomarkers for Cluster B (Luminal B) include fibroblast growth factor FGFR4 which might be from the fact that this family of genes is known to be overexpressed in cancers of the cervix and bladder, though their role in breast cancers is more controversial (Streit et al. 2004; Jezequel et al. 2004); two cancer related genes: Gamma-glutamyl hydrolase (GGH) gene, which was also identified as a biomarker for pulmonary neuroendocrine tumors (He et al. 2004), and laminin, gamma 2 (LAMC2) gene, which was reported to be involved in tumor invasion and metastases in pancreatic ductal adenocarcinoma (Takahashi et al. 2002) and endometrial adenocarcinomas (Maatta et al. 2004). The latter gene is down-regulated in the breast cancer data sets analyzed here.

Generally, Cluster C (ERBB2+ subtype) biomarkers appear to be mostly receptors, receptor binding proteins and signal transduction related proteins (Table 9). As expected, the most characteristic of these genes is the up-regulated ERBB2 gene. Other important genes include two breast cancer related genes, namely, the F2R gene, a matrix metalloprotease-1 receptor that promotes invasion and tumorigenesis of breast cancer cells (Boire et al. 2005); and PPAR binding protein, coactivator of ESR1 and overexpressed in breast cancer (Zhu et al. 1999). The down-regulation of FLNB filamin B alters the MAP Kinase pathway with implications in both growth control and development.

The marker genes for the Basal phenotype (Cluster D) are significantly involved in cell cycle, regulation of cell proliferation, endoplasmic reticulum as well as in various metabolic processes. Important cancer related genes identified for this phenotype are CDK6 gene, which inhibits proliferation of human mammary epithelial cells (Lucas et al. 2004); SIAT4C, which is down-regulated in RCC (Saito et al. 2002), RHOB, which is known to be a pro-apoptotic and tumor suppressor gene, and the FLT1 and TFF3 gene. Plasminogen activator gene (PLAT) is involved in tissue remodeling while fibromodulin (FMOD) gene has a primary role in collagen fibrillogenesis.

The last of the clusters is the control or normal group. Here we find that the genes identified as significant markers are involved in organelle organization and biogenesis, cytoskeleton organization and biogenesis, or in metabolic pathways (e.g. cofactor biosynthesis). These represent genes that are pathologically expressed in all tumor strata; consequently they are able to robustly stratify BCA samples from control (Normals).

Table 9. Enriched GO properties for the core phenotypes.

Group	GO category	GO description	Number of genes	LS Permutation p-value	KS Permutation p-value
core A	19752	carboxylic acid metabolism	19	0.002	0.000
	6519	amino acid and derivative metabolism	6	0.040	0.001
core B	6732	coenzyme metabolism	8	0.008	0.068
core C	16591	DNA-directed RNA polymerase II\, holoenzyme	5	0.000	0.126
	5102	receptor binding	25	0.000	0.005
	5654	nucleoplasm	9	0.000	0.156
	4872	receptor activity	40	0.001	0.025
	7165	signal transduction	96	0.004	0.214
	6366	transcription from RNA polymerase II promoter	22	0.005	0.088
core D	5783	endoplasmic reticulum	24	0.047	0.003
	74	regulation of progression through cell cycle	20	0.008	0.064
	19752	carboxylic acid metabolism	19	0.020	0.005
	4674	protein serine/threonine kinase activity	17	0.297	0.008
	42127	regulation of cell proliferation	12	0.011	0.008
core E	6996	organelle organization and biogenesis	24	0.006	0.029
	5200	structural constituent of cytoskeleton	9	0.001	0.009
	30036	actin cytoskeleton organization and biogenesis	7	0.008	0.073
	6928	cell motility	7	0.008	0.056
	51188	cofactor biosynthesis	5	0.006	0.032

Overall, the biomarkers notably constitute genes that participate in breast cancer related pathways (e.g. marker genes involved in estrogen receptor pathway) and genes that were previously implicated in other cancer types (e.g. GSTP1, FLT1, see Table 8). Moreover, the enriched categories in each phenotype are biologically plausible, having already been implicated in cancer transformation (e.g. cell cycle, cell motility, cytoskeleton organization) (Hanahan and Weinberg, 2000) or being potentially important in transformation (signal transduction pathways, metabolism).

Summary and Discussion

We have presented a robust clustering and pattern based analysis of the phenotypes identified by Sorlie et al. 2003. We find that the clusters for Luminal A, Basal and Normal subtypes are homogenous and have predictive content. However, the Luminal B and ERBB2+ assignments are sensitive to data perturbations. One reason for this is that the genes chosen for the classification are too few and not appropriate for these two categories. This is evidenced by the fact that the number of genes for Luminal B and ERBB2+ that pass our stringent robustness filters is small. Another reason is that hierarchical clustering is inappropriate to resolve the subtleties of the Luminal B and ERBB2+ categories. Finally, these subtypes are more heterogeneous than Luminal A and Basal and possibly have further substructure not classifiable with the genes in this dataset. A larger number of samples and better/more genes are necessary to test these conclusions.

Several samples previously unclassified in Sorlie et al. 2003 were classifiable by our techniques. We also found several samples which show a complex (multiple) phenotype signature. Given the treatment implications, the patients from whom

these samples were taken should undergo further analysis or different treatment.

We also describe a general method to deal with *sensitivity to noise* in gene array data, which often confounds the analysis. There are four principal sources of noise. The first, which we cannot do anything about, is the experiment itself: a) different samples handled differently in and experiment or between different labs; b) data improperly collected or improperly recorded/measured; c) microarray or cDNA readout with missing or unreliable entries. The second type of "noise" is stochastic noise; from statistical errors in the measurement of the signal or from normal variation within a phenotype in the sample population. We show how to partially account for this noise by data perturbations and consensus analysis. A third source of noise is the data analysis methods used. In particular, there are many different definitions of distance between gene expression vectors and many different clustering techniques. These often lead to different clusters depending on parameter choices, and to clusters that are unstable to perturbations. Our method robustly deals with this issue to get reliable predictions. A fourth source of noise derives from the genes selected as the basis for the analysis (Ein-Dor et al. 2005). This set results both from the initial choice of genes on the chip and the subset of genes that is used in the clustering. The choice of genes on chips will improve only if chip manufactures come up with better chips, possibly motivated by the biology of the underlying processes. However, given a gene set, this paper describes a procedure to select a data perturbation independent and predictive subset of the genes.

The fundamental requirement of any clustering analysis is the assignment of confidence levels to clusters. This is particularly important in gene expression analysis where a small sample set is clustered using a large set of noisy genes which makes the clustering results sensitive to noise and susceptible to over-fitting. Our methods use re-sampling and cross validation to simulate perturbations of the data, and this allows us assess the stability of the clustering with respect to sample variability.

In functional genomics, agglomerative hierarchical clustering (HC) has been widely adopted as the unsupervised analysis tool of choice, mainly because of its intuitive appeal and its visualization properties. By not committing to a specific number of clusters, HC provides for a multi-resolution view of the data that can be extremely useful in exploratory data analysis. However, the method does not provide for an "objective" criterion to establish the number of clusters and the clusters' boundaries. Furthermore, the resulting trees are known to be highly unstable to small perturbations of the data. The trees also tend to preserve sample joining errors made at earlier stages.

To correct for these problems, we recommend averaging over perturbations of the original data. The hierarchical clustering algorithm can then be applied to each of the perturbed data sets, and the agreement, or consensus, among the multiple runs can be assessed. This technique will measure the "stability" of the discovered clusters to sampling variability. The basic assumption of the method is intuitively simple: if the data represent a sample of items drawn from distinct sub-populations, and if we were to observe a different sample drawn from the same subpopulations, the induced cluster composition and number should not be radically different. Therefore, the more the attained clusters are robust to sampling variability, the more confident we can be that these clusters represent real structure. Overall, the procedures suggested here will be of use in examining any data in a way that makes the predictions insensitive to stochastic and systematic variation.

A frequent concern in gene-array data and analysis is whether the data is reproducible, and whether the inferences are consistent with current biological knowledge. In this paper we address the first issue by applying the results of our analysis on one data set to make predictions on another. For the phenotypes which cluster well, we can make definite predictions on the unseen data. In addition, we identify pathways via genes whose markers are predictive of phenotype. It is likely that these genes have only diagnostic value, i.e. they are downstream effects of an established disease process whose cause is outside the identified set of genes. This is a problem with most microarray data which is usually available only for cells which show established disease.

Acknowledgments

We thank Professor Arnold J. Levine and Dr. Gustavo Stolovitzky for discussions and Dr. Wentian Li for helpful comments on an early version of the manuscript. RR thanks the Institute for Advanced Study, for sabbatical support and GB thanks them for continuing visiting membership

status. We are grateful to Xia-Jun Ma for providing the raw data. The work of GA was supported by the New Jersey Commission on Cancer Research (CCR-703054-03) and the Institute for Advanced Study, through The David and Lucile Packard Foundation and The Shelby White and Leon Levy Initiative Fund.

References

Abd El-Rehim, D.M., Ball, G. and Pinder, S.E. et al. 2005. High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses. *Int. J. Cancer*, 116:340–50.

Ahnstrom, M., Nordenskjold, B. and Rutqvist, L.E. et al. 2005. Role of cyclin D1 in ErbB2-positive breast cancer and tamoxifen resistance. *Breast Cancer Res. Treat.*, 91:145–51.

Alexe, G., Alexe, S. and Crama, Y. et al. 2004. Consensus algorithms for the generation of all maximal bicliques. *Disc. Appl. Math.*, 145:11–21.

Alexe, G. and Hammer, P.L. 2005. Spanned patterns in logical analysis of data. *Discr. Appl. Math.*, 154:1039–49.

Alexe, G., Bhanot, G. and Venkataraghavan, B. et al. 2005a. A robust meta-classification strategy for cancer diagnosis from gene expression data. *Proc. IEEE Comput. Syst. Bioinform. Conf.*, 322–5.

Alexe, G., Alexe, S. and Axelrod, D.E. et al. 2006a. Breast cancer prognosis by combinatorial analysis of gene expression data. *Breast Cancer Res*., 8(4):R41.

Alexe, G., Alexe, S. and Axelrod, D.E. et al. 2005b. Logical analysis of diffuse large B-cell lymphomas. *Artif. Intell. Med.*, 34(3):235–67.

Alexe, G., Alexe, S. and Kogan, A. et al. 2005c. Comprehensive vs. comprehensible classifiers in Logical Analysis of Data. *Discr. Appl. Math.*, in press.

Alexe, G., Alexe, S. and Vizvari, B. et al. 2006b. Pattern-Based Feature Selection in Genomics and Proteomics. *Annals of Operations Research, Optimization in Medicine* (in press).

Anim, J.T., John, B. and Abdulsathar, S. SA. et al. 2005. Relationship between the expression of various markers and prognostic factors in breast cancer. *Acta. Histochem.*,107:87–93.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B.*, 57:289–300.

Bertucci, F., Finetti, P. and Rougemont, J. et al. 2005. Gene expression profiling identifies molecular subtypes of inflammatory breast cancer. *Cancer Res.*, 65:2170–8.

Bhanot, G., Alexe, G. and Levine, A.J. et al. 2005. Robust diagnosis of non-Hodgkin lymphoma phenotypes validated on gene expression data from different laboratories. *Genome Inform. Ser. Workshop Genome Inform.*, 16:233–44.

Bieche, I. and Lidereau, R. 1995. Genetic alterations in breast-cancer. *Genes Chromosomes Cancer*, 14:227–51.

Bishop, C.M. 1995. Neural Networks for Pattern Recognition. Oxford, Oxford University Press.

Boire, A., Covic, L. and Agarwal, A. et al. 2005. PAR1 is a matrix metalloprotease-1 receptor that promotes invasion and tumorigenesis of breast cancer cells. *Cell*, 120:303–13.

Bonferroni, C.E. 1935. "Il calcolo delle assicurazioni su gruppi di teste." In Studi in Onore del Professore Salvatore Ortu Carboni. Rome: Italy, p 13–60.

Bradley, J.V. 1968. Distribution Free Statistical Tests. Prentice Hall: Englewood Cliffs, NJ.

Breitling, R. and Herzyk, P. 2005. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *J. Bioinform. Comput. Biol.*, 3:1171–1189.

Charafe-Jauffret, E., Ginestier, C. and Monville, F. et al. 2005. Gene expression profiling of breast cell lines identifies potential new basal markers. *Oncogene*, 1–12.

Crama, Y., Hammer, P.L. and Ibaraki, T. 1988. Cause-effect relationships and partially defined Boolean functions. *Ann. Oper. Res.*, 16:299–326.

Dennis, G., Sherman, B.T. and Hosack, D.A. et al. 2003. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.*, 4:R60. URL: (http://david.niaid.nih.gov/david/)

Diermeier, S., Horvath, G. and Knuechel-Clarke, R. et al. 2005. Epidermal growth factor receptor coexpression modulates susceptibility to Herceptin in HER2/neu overexpressing breast cancer cells via specific erbB-receptor interaction and activation. *Exp. Cell Res.*, 304:604–19.

Dudoit, S., Popper Shaffer, J. and Boldrick, J.C. 2003. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103.

Ein-Dor, L., Kela, I. and Getz, G. et al. 2005. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics,* 21:171–8.

Eisen, M.B., Spellman, P.T. and Brown, P.O. et al.1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95:14863–68.

Farmer, P., Bonnefoi, H. and Becette, V. et al. 2005. Identification of molecular apocrine breast tumours by microarray analysis, *Oncogene*, 24:4460–71.

Furlanello, C., Serafini, M. and Merler, S. et al. 2003. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics*, 6:4:54.

Goh, L. and Kasabov, N. 2005. An integrated feature selection and classification method to select minimum number of variables on the case study of gene expression data. *J. Bioinform. Comput. Biol*., 3(5):1107–36.

Golub, T.R., Slonim, D.K. and Tamayo, P. et al. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 536:531–537.

Gosset, W.S. 1908. The probable error of a mean. *Biometrika*, 6:1–25.

Gruvberger, S., Ringner, M. and Chen, Y. et al. 2000. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression pattern. *Cancer Res.*, 61:5979–84.

Guyon, I. and Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3:1157–1182.

Hanahan, D. and Weinberg, R.A. 2000. The hallmarks of cancer. *Cell*, 100:57–70.

He, P., Varticovski, L. and Bowman, E.D. et al. 2004. Identification of carboxypeptidase E and gammaglutamyl hydrolase as biomarkers for pulmonary neuroendocrine tumors by cDNA microarray. *Hum Pathol*, 35:1196–209.

Hoffmann, R. and Valencia, A. 2004. A gene network for navigating the literature. *Nat. Genet.*, 36, 664.

Honig, A., Rieger, L. and Sutterlin, M. et al. 2004. Preoperative chemotherapy and endocrine therapy in patients with breast cancer. *Clin. Breast. Cancer*, 5:198–207

Hu, Z., Fan, C. and Oh, D.S. et al. 2006. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, 7:96.

Inza, I., Larranaga, P. and Blanco, R. et al. 2004. Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med.*, 31(2):91–103.

Jeffery, I.B., Higgins, D.G. and Culhane, A.C. 2006. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 7:359.

Jezequel, P., Campion, L. and Joalland, M.P. et al. 2004. G388R mutation of the FGFR4 gene is not relevant to breast cancer prognosis. *Br. J. Cancer*, 90:189–93.

van der Kloot, W.A., Spaans, A.M. and Heiser, W.J. 2005. Instability of hierarchical cluster analysis due to input order of the data. *Psychol Methods.*, 10(4):468–76.

Kristensen, V.N., Sorlie, T. and Geisler, J. et al. 2005. Gene expression profiling of breast cancer in relation to estrogen receptor status and estrogen-metabolizing enzymes: clinical implications. *Clin. Cancer Res.*, 11:878–83.

Lacroix, M. and Leclercq, G. 2005. The portrait of hereditary breast cancer. *Breast. Cancer Res. Treat.*, 89:297-304. URL: http://www.geocities.com/m.lacroix/intro1.htm.

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

Lehmann, E.L. 1975. Nonparametrics: Statistical Methods Based on Ranks. *San Francisco: Holden-Day, Inc*.,

Liu. X., Krishnan, A. and Mondry, A. 2005. An entropy-based gene selection method for cancer classification using microarray data. *BMC Bioinformatics*, 6:76.

Liu, Y. 2004. A Comparative Study on Feature Selection Methods for Drug Discovery. *J. Chem. Inf. Comput. Sci.,* 44 (5):1823–1828.

Loi, S., Desmedt, C. and Cardoso, F. et al. 2005. Breast cancer gene expression profiling: clinical trial and practice implications. *Pharmacogenomics*, 6:49–58.

Lucas, J.J., Domenico, J. and Gelfand, E.W. 2004. Cyclin-dependent kinase 6 inhibits proliferation of human mammary epithelial cells. *Mol. Cancer Res.*, 2:105–14.

Ma, X.J., Salunga, R. and Tuggle, J.T. et al. 2003. Gene expression profiles of human breast cancer progression. *Proc. Natl. Acad. Sci. U.S.A.*, 100:5974–9.

Maatta, M., Salo, S. and Tasanen, K. et al. 2004. Distribution of basement membrane anchoring molecules in normal and transformed endometrium: altered expression of laminin gamma2 chain and collagen type XVII in endometrial adenocarcinomas. *J. Mol. Histol.*, 35:715–22.

McShane, L.M., Radmacher, M.D. and Freidlin, B. et al. 2002. Methods of assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18:1462–79.

Merz, C. 1998. Classification and Regression by Combining Models. Dissertation, University of California at Irvine.

Monti, S., Tamayo, P. and Mesirov, J. et al. 2003. Consensus Clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52:91–118.

Monti, S., Savage, K.J. and Kutok, L. et al. 2005. Molecular profiling of diffuse large B-cell lymphoma reveals a novel disease subtype with brisk host inflammatory response and distinct genetic features. *Blood*, 105:1851–1861.

Mutch, D.M, Berger, A. and Mansourian, R. et al. 2002. The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics*, 3:17.

Nishida, K., Tsukamoto, T. and Uchida, K. et al. 1996. Introduction of the c-kit gene leads to growth suppression of a breast cancer cell line, MCF-7. *Anticancer Res.*, 16:3397–402.

Osipo, C., Gajdos, C. and Cheng, D. et al. 2005. Reversal of tamoxifen resistant breast cancer by low dose estrogen therapy. *J. Steroid Biochem. Mol. Biol.*, 93:249–56.

Pandey, R., Guru, R.K. and Mount, D.W. 2004. Pathway Miner: Extracting gene association networks from molecular pathways for predicting the biological significance of gene Expression microarray data. *Bioinformatics*, 20, 2156–8. URL: http://www.biorag.org/.

Patel, S. and Lyons-Weiler, J. 2004. caGEDA: a web application for the integrated analysis of global gene expression patterns in cancer. *Appl. Bioinformatics*, 3(1):49–62.

Pavlidis, P., Qin, J. and Arango, V. et al. 2004. Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res.*, 29:1213–22.

Perou, C.M., Jeffrey, S.S. and van de Rijn, M. et al. 2001. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. U.S.A.,* 96:9212–7.

Perou, C.M., Sorlie, T. and Eisen, M.B. et al., 2000. Molecular portraits of human breast tumours. *Nature*, 406:747–752.

Ramaswamy, S., Tamayo, P. and Rifkin, R. et al. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U.S.A.*, 98:15149–57.

Reich, M., Liefeld, T. and Gould, J. et al. 2006. GenePattern 2.0. *Nature Genetics* 38:500–501.

Ripley, B.D. 1996. Pattern Recognition and Neural Networks. Cambridge.

Rivat, C., Rodrigues, S. and Bruyneel, E. et al. 2005. Implication of STAT3 signaling in human colonic cancer cells during intestinal trefoil factor 3 (TFF3) -- and vascular endothelial growth factor-mediated cellular invasion and tumor growth. *Cancer Res.*, 65:195–202.

Rousseeuw, P.J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65.

Saito, S., Yamashita, S. and Endoh, M. et al. 2002. Clinical significance of ST3Gal IV expression in human renal cell carcinoma. *Oncol Rep.*, 9:1251–5.

Sharma, R., Beith, J. and Hamilton, A. 2005. Systematic review of LHRH agonists for the adjuvant treatment of early breast cancer. *Breast.*, 14:181–91.

Shipp, M.A., Ross, K.N. and Tamayo, P. et al. 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8:68–74.

Sorlie, T., Perou, C.M. and Tibshirani, R. et al. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci., U.S.A.*, 98:10869–74.

Sorlie, T., Tibshirani R., and Parker, J., et al. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U.S.A.*, 100:8418–23.

Sorlie, T., Wang Y., and Xiao, C., et al. 2006. Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. *BMC Genomics,* 7:127.

Sotiriou, C., Neo, S.Y., and McShane, L.M., et al. 2003. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. U.S.A.,* 100(18):10393–8.

Storey, J.D., and Tibshirani, R. 2003. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U.S.A.,* 100:9440–5.

Streit, S., Bange, J., and Fichtner, A., Ihrler, S., et al. 2004. Involvement of the FGFR4 Arg388 allele in head and neck squamous cell carcinoma. *Int. J. Cancer*, 111:213–7.

Su, Y., Murali, T.M., and Pavlovic, V., et al. 2003. RankGene: identification of diagnostic genes based on expression data. *Bioinformatics*, 19(12):1578–9.

Sun, Z., Yang, P. and Aubry, M.C., et al. 2004. Can gene expression profiling predict survival for patients with squmous cell carcinoma of the lung? *Molecular Cancer*, 3:35.

Sutherland, B.W., Kucab, J., and Wu, J., et al. 2005. Akt phosphorylates the Y-box binding protein 1 at Ser102 located in the cold shock domain and affects the anchorage-independent growth of breast cancer cells. *Oncogene*, 24:4281–92.

Takahashi, S., Hasebe, T., and Oda, T., et al. 2002. Cytoplasmic expression of laminin gamma2 chain correlates with postoperative hepatic metastasis and poor prognosis in patients with pancreatic ductal adenocarcinoma. *Cancer*, 94:1894–901.

Troester, M.A., and Hoadley, K.A., Sorlie, T., et al. 2004. Cell-type specific responses to chemotherapeutics in breast cancer. *Cancer Res.,* 64:4218:26.

Troyanskaya, O., Cantor, M., and Sherlock, G., et al. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–5.

Tourassi, G.D., Frederick, E.D., and Markey, M.K., et al. 2001. Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Med. Phys.*, 28(12):2394–402.

Tusher, V.G., Tibshirani, R., and Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U.S.A.,* 98:5116–5121.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley-Interscience.

van't Veer, L.J., Dai, H.Y., and van de Vijver, M.J., et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536.

Weinberg, R.A. 2006. Biology of cancer. Garland Science, 1st edition.

West, M., Blanchette, C., and Dressman, H., et al. 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. U.S.A.,* 98:11462–67.

Zhang, Y.F., Homer, C., and Edwards, S.J., et al. 2003. Nuclear localization of Y-box factor YB1 requires wild-type p53. *Oncogene,* 22:2782–94.

Zhu, Y., Qi. C., and Jain, S., et al. 1999. Amplification and overexpression of peroxisome proliferator-activated receptor binding protein (PBP/PPARBP) gene in breast cancer. *Proc. Natl. Acad. Sci. U.S.A.,* 96:10848–53.

Supplementary Information

Supplementary Information I: Multiple Testing Correction Metrics

The general multiple hypothesis testing analysis used in our paper results in the following matrix:

	# non–rejected hypotheses	#rejected hypotheses	
# true null hypotheses (non-diff. genes)	*U*	*V* ***Type I error***	M_0
# false null hypotheses (diff. genes)	*T* ***Type II error***	*S*	M_1

We use the following statistics to analyze this table.

False discovery rate (*FDR*). The FDR (Benjamini and Hochberg 1995) is the expected proportion of Type I errors among the rejected hypotheses: $FDR = E(Q)$; with $Q = V/R$ if $R > 0$ and $Q = 0$; if $R = 0$.

The *q-value* of a gene (Storey and Tibshirani, 2003) is defined as the minimal FDR at which it appears significant.

Family–wise error rate (FWER, Dudoit et al. 2003). The FWER is defined as the probability of at least one Type I error (false positive): $FWER = \Pr(V > 0)$

The Bonferroni correction (Bonferroni 1935) : Suppose we conduct a hypothesis test for each gene $g = 1,\ldots,N$, producing an observed test statistic: T_g , an unadjusted *p*–value: p_g. = the probability under the null hypothesis that the test statistic is at least as extreme as T_g. Under the null hypothesis, $\Pr(p_g < a) = a$.

Bonferroni adjusted p–values: $\mathbf{p}_g = \min(1, \mathrm{N}\, p_g.)$

References for Supplementary Information III

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B.*, 57:289–300.

Bonferroni, C.E. 1935. "Il calcolo delle assicurazioni su gruppi di teste." In Studi in Onore del Professore Salvatore Ortu Carboni. Rome: Italy, p 13–60.

Dudoit, S., Popper Shaffer, J. and Boldrick, J.C. 2003. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103.

Storey, J.D. and Tibshirani, R. 2003. Statistical significance for genome-wide studies. *Proc. Natl. Acad. Sci. U.S.A.*, 100:9440–5.

Supplementary Information II: Functional class scoring for GO categories

We computed the statistical significance of a GO category within a collection of *N* gene markers by following Pavlidis et al. 2004: A *p*-value was computed for each of the *N* marker genes in our collection. Next, the set of *p*-values was tested for enrichment in a GO category by using the Functional Class (*LS*) and the Kolmogorov-Smirnov (*KS*) statistics. For a set of *N* genes, these are defined as

$$LS = \sum_{i=1}^{N} (-\log p_i)/N$$

$$KS = \max_{i=1,\ldots,N} \frac{i}{N} - p_i$$

The statistical significance of a GO category with *N* genes was measured by computing the empirical distribution of *LS* and *KS* from 100,000 random selections of *N* genes in the complete pool of genes. The *LS*/*KS* permutation *p*-value was computed by comparing the *LS*/*KS* statistics in these experiments to the measured value of these statistics for the selected genes. A GO category was considered enriched if its corresponding LS or KS re-sampling *p*-value was below 0.005.

References for Supplementary Information II

Pavlidis, P., Qin, J., Arango, V., et al. 2004. Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochem. Res.*, 29:1213–22.