Analysis of breast cancer progression using principal component analysis and clustering

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

[1]*The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge MA, 02142, USA*
[2]*The Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540, USA*
[3]*Molecular Biology, Cell Biology and Biochemistry Program, Boston University, Boston, MA 02215, USA*
[4]*Cancer Institute of New Jersey, 195 Little Albany Street, New Brunswick, NJ 08903, USA*
[5]*Center for Advanced Genomic Technology, Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA*
[6]*BioMaPS Institute and Department of Biomedical Engineering, Rutgers University, Piscataway, NJ 08854, USA*

**Joint first authors*
***Corresponding authors (Email, gyanbhanot@gmail.com; delisi@bu.edu)*

We develop a new technique to analyse microarray data which uses a combination of principal components analysis and consensus ensemble *k*-clustering to find robust clusters and gene markers in the data. We apply our method to a public microarray breast cancer dataset which has expression levels of genes in normal samples as well as in three pathological stages of disease; namely, atypical ductal hyperplasia or ADH, ductal carcinoma *in situ* or DCIS and invasive ductal carcinoma or IDC. Our method averages over clustering techniques and data perturbation to find stable, robust clusters and gene markers. We identify the clusters and their pathways with distinct subtypes of breast cancer (Luminal, Basal and Her2+). We confirm that the cancer phenotype develops early (in early hyperplasia or ADH stage) and find from our analysis that each subtype progresses from ADH to DCIS to IDC along its own specific pathway, as if each was a distinct disease.

[Alexe G, Dalgin G S, Ganesan S, DeLisi C and Bhanot G 2007 Analysis of breast cancer progression using principal component analysis and clustering; *J. Biosci.* **32 1**027–1039]

1. Introduction

Microarrays have the potential to identify pathways that are altered in disease. This promise has resulted in this technology being aggressively pursued by researchers, hospitals and pharmas because of its potential for an improved understanding of the disease process, better diagnostic protocols, new drugs, and new treatment regimens. A major application of microarrays has been to the analysis of cancer. The focus has been on identifying genes that are altered in the initiation, progression and metastasis of cancer. Such analysis is confounded by the fact that most cancers are highly heterogeneous, microarray signals are noisy and there is a large variation in the "normal" levels of most genes. Identification of the signals that are characteristic for the disease phenotype and its progression requires the use of robust techniques.

In this paper we develop a new robust method which first uses principal component analysis (PCA) (*see* Wall *et al* 2003; Evritt and Dunn 2001) to identify the overall structure of clusters in the data and select the subset of genes that distinguish the clusters. We use this set of genes and a new consensus ensemble *k*-clustering technique, which averages over several clustering methods and many data perturbations, to identify strong, stable clusters. We also define a simple criterion to find the optimum number of

Keywords. Breast cancer subtypes; clustering; metastatic risk; microarray

Abbreviations used: ADH, Atypical ductal hyperplasia; DCIS, ductal carcinoma *in situ;* FDR, false-discover-rate; IDC, invasive ductal carcinoma; PCA , principal component analysis; SNR, signal to noise ratio; WV, weighted voting clusters and a method to identify robust markers for disease progression within each cluster.

Our method results in stable lists of genes and pathways that distinguish high and low grade tumours and other sets which mark progression of disease from ductal carcinoma *in situ* to invasive ductal carcinoma. The clustering paints a portrait of the disease at varying levels of granularity. When the data is divided into two clusters, the normal samples form one cluster and the disease samples form another. At the next level of clustering, the low grade and high grade samples separate. The optimal number of clusters is seven, corresponding to a split of the low grade cluster into two and the high grade cluster into four sub-clusters. The sub-clusters are well separated by a strong set of markers which are able to distinguish them with sensitivity and specificity in the 80-100% range. We identify the genes and pathways that mark disease progression in each sub-cluster. A major result of our analysis is that each sub-cluster contains samples from non-invasive and invasive tumours from the same patient. This suggests that within each grade of breast cancer, different groups of patients progress to the same final phenotype along different pathways. This prediction needs to be validated by a study on a chip with more genes from the pathways we identify. If verified, it would have significant implications for disease identification and treatment.

2. Materials and methods

2.1 *Description of microarray data*

We used public data from Ma *et al* (2003) (*www.geneexpression_ma.org*) which consists of a cohort of 36 breast cancer patients of whom 31 were diagnosed with at least two out of three pathological stages of disease: atypical ductal hyperplasia or ADH, ductal carcinoma *in situ* or DCIS and invasive ductal carcinoma or IDC respectively. The remaining 5 patients were diagnosed to have pre-invasive disease (ADH) only.

As described in the original study of Ma *et al* (2003), normal as well as disease samples were harvested from each patient in as many different disease stages (ADH, DCIS, IDC) as possible by laser capture micro-dissection (Arcturus, CA, USA) in triplicate. Care was taken to avoid contamination between cells of different stages taken from the same patient. There were a total of 300 samples in the Ma *et al* (2003) study, each of which was analysed in duplicate with a 12,000 gene cDNA microarray. It was determined that the "normal cells" from cancer patients were highly similar to the normal epithelium of three non-cancer patients. This suggested that the "normal cells" from cancer patients could be used as a baseline to determine disease state and progression.

The data Ma *et al* (2003) consists of the expression values of 1940 genes across the 93 samples and 32 of these were from disease free or normal patients, 8 were ADH samples, 30 were DCIS samples and 23 were IDC samples. The 1940 (out of the 12,000) genes were selected in Ma *et al* (2003) by their ability to distinguish "normal cells" and each of the disease stages ADH, DCIS and IDC using a linear discriminant function. The patients were further classified by pathological analysis into 3 categories based on the tumour grade: grade I (18 patients), grade II (22 patients) and grade III (19 patients).

2.2 *Overview of analysis technique*

The flow chart of our analysis method is presented in figure 1. First the dataset was normalized and missing entries imputed robustly. Next, we used PCA to identify the genes which account for most of the variation in the data. The optimal number of clusters k_{opt} in the data was estimated using gap statistics (Tibshirani *et al* 2001) and silhouette scores (Kaufmann and Rousseeuw, 1990).

Consensus ensemble clustering (Monti *et al* 2003; Strehl and Ghosh 2002) was applied to the projection of the data on these genes to identify k=2, 3,…, k_{opt} data and method perturbation independent (robust) clusters. To ensure sensitivity to subtle signals that may be present, we used the full set of genes on the samples after each k level clustering to find the best pool of genes that distinguish disease classes within and between clusters. This non-stringent selection was motivated by the expectation that the key genes altered in disease pathways most likely change their expression levels in subtle ways, and may not necessarily be the same genes that are best to distinguish the clusters. On this larger set of genes for each k, we identified two sub-classes. The first set distinguished each cluster from its complement. The second set defined progression from non-invasive to invasive disease. Finally, we used annotated databases to identify the functional pathways that are most representative of the clusters identified. Each of these steps is described in detail below.

2.3 *Data normalization and imputation*

The genes were normalized by first applying a robust nonlinear local regression method as described in Ma *et al* (2003) and then by applying a global normalization procedure which consists of subtracting the median of each gene across the arrays. Thirteen genes had missing values in 13-15% of the samples and were discarded and 105 genes had missing entries for up to 5% of the samples. These missing entries were imputed using a dynamical kNN approach (Alexe *et al* 2006).

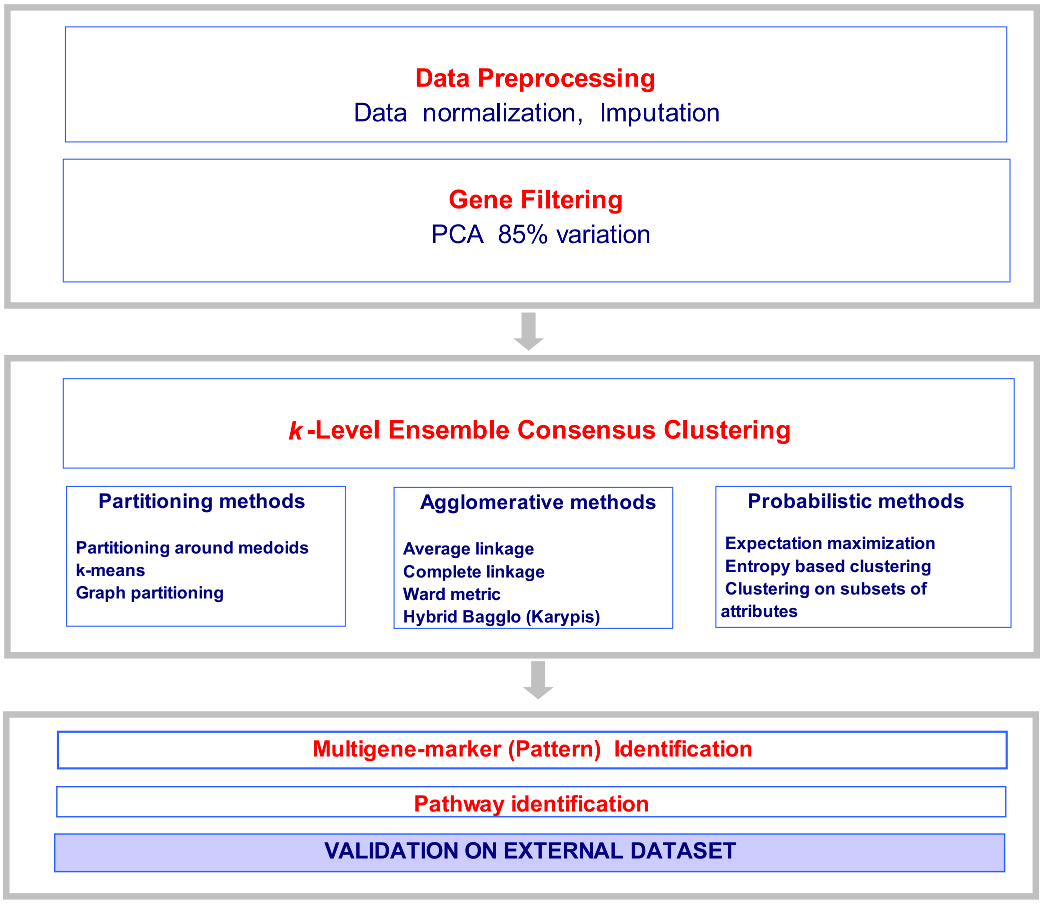

Figure 1. The flow chart of our method.

2.4 *Principal component analysis*

Principal component analysis or PCA (Kaufmann and Rousseeuw 1990; Everitt and Dunn 2001; Wall *et al* 2003) was used to retain those genes in the dataset that contribute most to its variance. PCA was applied to the expression matrix E_{ij} whose the rows were the 93 samples and whose columns were the 1927 genes. The analysis was done by a singular value decomposition of this matrix after it was centered and scaled to mean 0 and variance 1 per column. From the eigenvectors of the largest eigenvalues that accounted for 85% of the variation in the data we selected the subset of genes with coefficients in the top 25% in absolute value in these eigenvectors. This collection of genes was further used to find robust clusters in the data.

2.5 *Ensemble consensus k-clustering*

Using the genes from PCA, we first identified the optimal number of clusters using gap statistics (Tibshirani *et al* 2001) and silhouette scores (Kaufmann and Rousseeuw 1990). Next, we applied an ensemble consensus *k*-clustering approach (initiated by Monti *et al* 2003 and Strehl and Ghosh 2002) to group the samples into the optimum number of clusters. The ensemble consensus clustering integrates the results of various clustering techniques across sample data perturbations into a pairwise agreement matrix which is used to partition the samples into the optimum number of clusters.

The overall technique has two distinct parts: (i) a method which generates a collection of clustering solutions using different methods applied to many perturbations of the data, and (ii) a consensus function that combines the clusters found to produce a single output clustering of the data. The approach used in our paper is summarized below.

Step 1: 150 datasets were created from the imputed data restricted to the 207 significant genes identified by PCA. Fifty datasets came from bootstrapping the samples, 50 from bootstrapping genes and 50 by first projecting the data on bootstrapped genes and then by further bootstrapping on samples.

Step 2: The optimal number of clusters k_{opt} was inferred (*a priori*) using the gap statistic and silhouette scores.

Step 3: k=2,…, k_{opt} clusters were created using representative methods from the three major classes:

(i) *Partitioning*: partition around medoids (PAM) (Kaufmann and Rousseeuw 1990), *k*-means (Hartigan 1975) and graph partitioning (Zhao and Karypis 2003).

(ii) *Agglomerative*: hierarchical clustering based on average linkage, complete linkage and Ward metric (Kaufmann and Rousseeuw 1990) as well as bagglo, which is a hybrid agglomerative method developed in Zhao and Karypis (2003).

(iii) *Probabilistic*: expectation maximization (EM) method (Dempster *et al* 1977), entropy-based-clustering (ENCLUST) (Cheng *et al* 1999), clustering on subsets of attributes (COSA) (Friedman and Meulman 2004).

Step 4: Each clustering method was applied 50 times with different parameter initialization on the full dataset, and once on each of the 150 datasets from step 1. From the 200 resulting clusters, we constructed an agreement matrix of size N_{sample} x N_{sample} for each method, whose entries m_{ij} represented the fraction of times a pair of samples (i,j) occurred in the same cluster out of the number of times the pair was selected in the 200 datasets. Here N_{sample} denotes the number of samples in the dataset.

Step 5: For each *k*, the agreement results of step 4 were averaged across the clustering techniques. The samples were then sorted such that those with the highest pairwise agreement appeared along the diagonal of the agreement matrix in *k* blocks. We applied simulated annealing to find the *k* optimal clusters for which the average internal similarity (within each cluster) minus the average pairwise similarity (between all pairs of clusters) has a local maximum value.

2.6 *Identification of gene markers within clusters*

We now used the full collection of genes on each of the clusters identified at each *k* by consensus ensemble clustering. The markers were chosen to discriminate between two classes: class 1 = the group of interest (i.e. the entire cluster), class 0 = the samples not included in the group of interest (i.e. the complement of the cluster). The best markers were identified in two steps.

Step 1: A large pool of genes which distinguished the two labelled classes was selected based on a variant of the *t*-test statistic called the signal to noise ratio (SNR) (Golub *et al* 1999) with a permutation *P*-value of 0.1 and a false discovery rate (FDR) (Benjamini and Hochberg 1995) of 0.5. The SNR statistic computes the difference of the means in each of two classes scaled by the sum of the standard deviations: SNR = $(\mu_0 - \mu_1)/(\sigma_0 + \sigma_1)$, where μ_0 is the mean of class 0 and σ_0 is the standard deviation of class 0 and so on. 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 penalizes features that have higher variance in each class more than those features that have a high variance in one class and a low variance in another. This bias is particularly useful in distinguishing genes which are altered in normal/disease or stage/grade progression. For example, in the normal/disease case, the pathway in which the gene is involved is working correctly in one class, and hence is regulated strictly (has low variance) while in the other class, the pathway is compromised and the gene is less well regulated (has high variation).

Step 2: From the larger pool of genes from step 1, we identified the best genes correlated with the class label using stringent criteria which combined (i) a permutation *P*-value of 0.05, (ii) stability to sample perturbation through bootstrapping and (iii) stability to leave-one-out experiments in top 25% genes selected by weighted voting and kNN classifiers which distinguish the two classes with specificity and sensitivity above 0.75. This analysis was done using the software GenePattern from the Broad Institute (*http://www.broad.mit.edu/cancer/software/genepattern/*).

2.7 *Identification of pathways and biological/functional categories*

We used the bioinformatics public resources DAVID (Dennis *et al* 2003), iHOP (Hoffnamm and Valencia 2004), and MatchMiner (Bussey *et al* 2003). We also used 14 functional annotation sources including KEGG and GO annotations, Biocarta pathways, linked to DAVID as well as the functional classification tool implemented in DAVID. The Functional classification tool groups genes based on functional similarity. It uses Kappa statistics (Dennis *et al* 2003) which is an index that compares the agreement against the possibility that it appeared by chance. Thus,

$$\kappa = \frac{\text{Observed agreement} - \text{Chance agreement}}{1 - \text{Chance agreement}}.$$

The Kappa statistic can be thought of as the chance-corrected proportional agreement, and possible values range from +1 (perfect agreement) to 0 (no agreement above chance) to –1 (complete disagreement). The algorithm first generates a gene-to-gene similarity matrix (genes in rows and functional terms in columns) based on shared functional annotation. The matrix is made from binary entries. If a gene is annotated in a term, the term entry is 1, if not then the entry is 0. The algorithm adopts the kappa statistic to quantitatively measure the degree to which genes share similar annotation terms. The higher the value of κ, the stronger the agreement. The fuzzy heuristic partition algorithm (Dennis *et al* 2003), which allows a gene to participate in more than one cluster, was used to classify highly related genes into functionally related groups.

3. Results

3.1 *Principal component analysis*

We found that 50% of the variation in the data was represented by the first 5 PCs and 85% by the first 32 PCs. We identified 207 genes as those with highest absolute value (top 1st quartile) in the coefficients of the first 32 eigenvectors as representative of most of the data variation. These thresholds were estimated through a calibration step whose aim was to optimize the overall cluster membership assignment for the optimal number of clusters identified in the data restricted to the selected genes.

3.2 *Consensus ensemble* k*-clustering*

The gap statistic and the silhouette scores applied to the partition around medoids method identified k=7 as the optimal value for the number of clusters in the data, although the gap statistic output oscillated between 6 and 7. The data was divided into k=2, 3,…,7 clusters by using the 207 genes identified by PCA and by applying consensus ensemble k-clustering. The results are shown schematically in figure 2.

Note that even though the clusters at each level were determined independently, at clustering level k+1, two clusters were always obtained from splits of a parent cluster at level k, while the remaining k-1 clusters were inherited unchanged from the previous level k. This shows that the data inherently supports a clustering into a hierarchy of subtypes. Moreover, the separation of samples into "normal" and "disease" at k=2 and of the "disease" samples into "low" and "high" grades at k=3 and so on, strongly suggests that disease progression is a hierarchical process

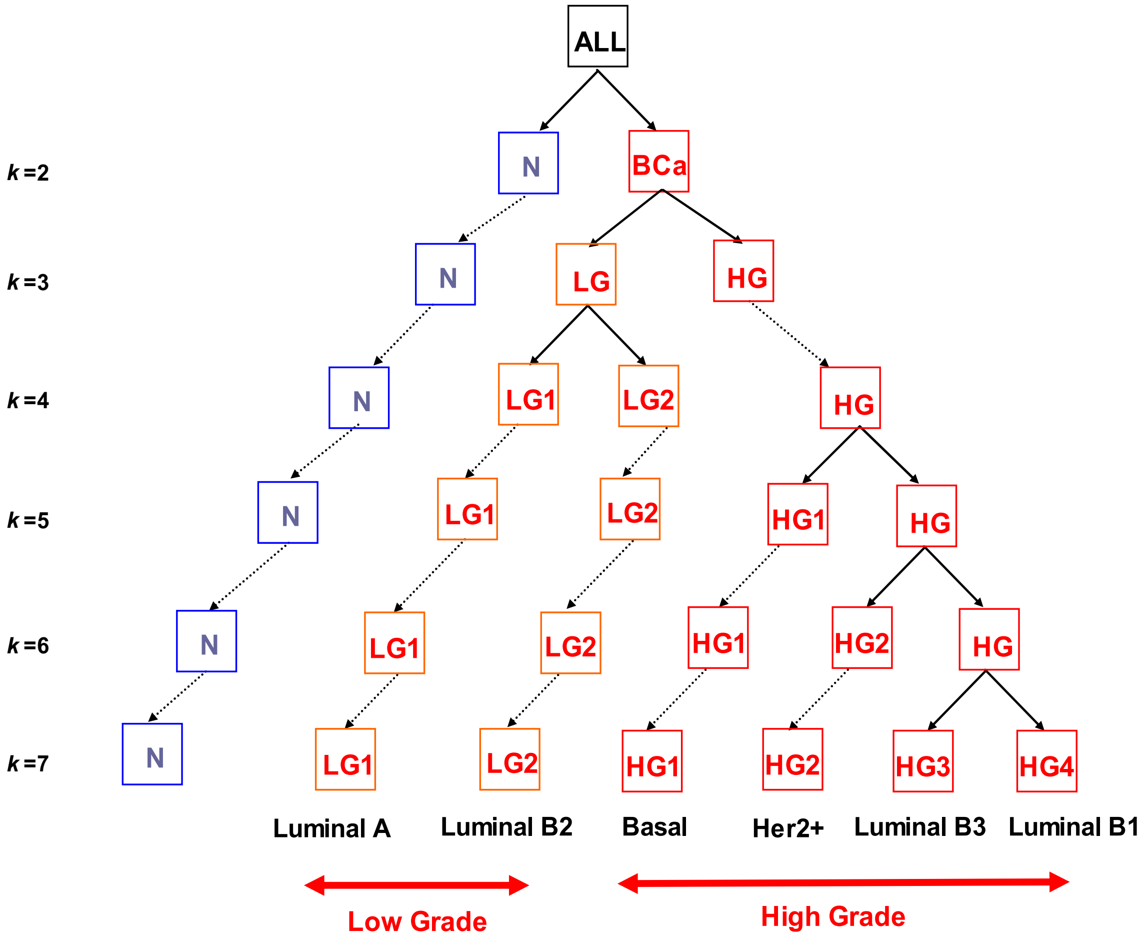

Figure 2. Consensus ensemble k-clustering tree reveals the recursive splitting of breast cancer subtypes

and is readily and robustly identifiable by our clustering procedure.

At *k*=2, the samples separated into a "normal" (N) group, which contained all the normal samples and one ADH sample (from patient id 210), and a "breast cancer" (BCA) group, which contained all the remaining breast cancer samples.

At *k*=3*,* the normal group was unaltered but the BCA group split into a low grade (LG) tumour group containing 18 samples labelled grade 1 and 9 samples labelled grade 2, and a high grade (HG) tumour group containing 13 samples labelled grade 2 and 19 samples labelled grade 3.

As *k* increased progressively from 4 through 7, the LG group split into 2 distinct subgroups (labelled LG1 and LG2 in figure 2) and the HG group split into 4 distinct subgroups (labelled HG1-HG4).

Table 1 shows the characteristics of the samples in these groups with respect to stage, ER, PR, Her2, lymph node and grade status for k = 2, 3 and 7. These subgroups of LG and HG are strongly dissimilar with respect to the cluster agreement matrix, which is shown in figure 3. We noticed that the HG1 subgroup is particularly different from the other HG subgroups (as is also evident in figure 2). All samples in it are ER-, PR- and mostly Her2-. The HG2 subgroup has a mixed ER signature, and the HG3 and HG4 subgroups consist mostly of ER positive samples. Based on these and other signatures (*see* below), we identify LG1 as Luminal A; LG2, HG3, HG4 as Luminal B; HG1 as Basal and HG2 as Her2+.

At k=7, each of the six BCA clusters always contained samples in both DCIS and IDC stages from the same patient. This suggests that breast cancer is composed of distinct disease subtypes that develop early and progress along different pathways because progression within a subtype is less distinct than the subtypes themselves. This strong heterogeneity in the genetic signature of subtypes also suggests that treatment decisions may benefit by taking account of the subtype as well as ER, PR and HER2 status. Note that the ensemble consensus clustering is absolutely critical to distinguish the subtypes. PCA by itself could identify a collection of useful markers, but could not identify the rich stratification discovered by consensus ensemble *k*-clustering.

3.1 *Identification of significant characteristic markers for the LG and HG subgroups*

3.3.1 *Markers for the LG and the HG groups:* Using a non-stringent SNR test (permutation *P*-value *P* = 0.10) we

Table 1. Clinical characteristics of *k* = 2, 3, 7 clusters

Cluster level *k*	Group	Size	Stage				ER			PR			Her2			Node		Grade		
			ADH	DCIS	IDC	N	Pos	Neg	ND	Pos	Neg	ND	Pos	Neg	ND	Pos	Neg	1	2	3
2	N	33	1			32														
	BCA	60	7	30	23		47	10	3	42	15	3	10	37	13	44	14	18	22	19
3	LG	28	7	13	8		26		2	21	5	2	4	18	6	20	8	18	9	
	HG	32		17	15		21	10	1	21	10	1	6	19	3	24	6		13	19
7	LG1	11	4	5	2		11			8	3		1	10		7	4	9	2	
	LG2	17	3	8	6		15		2	13	2	2	3	8	6	13	4	9	7	
	HG1	5		2	3			5			5		1	4		3				5
	HG2	10		7	3		7	3		5	5		3	4	1	9	1		2	8
	HG3	13		6	7		10	2	1	12		1	2	7	2	10	3		7	6
	HG4	4		2	2		4			4				4		2	2		4	

ND stands for "not determined". The "Node" and "Grade" status of some samples was not provided in the data. At *k*=2, the clustering splits the data into normal samples and disease samples (BCA), except for one ADH which is classified with the normals. At *k*=3, the BCA samples split into high grade (grade 2 or 3) and low grade (grade 1 or 2) categories. At *k*=7, the low grade samples split into two clusters LG1, LG2 and the high grade into four: HG1 – HG4. The HG1 samples are all ER-, PR- and mostly Her2-. The HG3 and HG4 clusters are mostly ER+, PR+, Her2-. The HG2 cluster has mixed ER, PR and Her2 signatures. Based on this and other gene signatures, using the Sørlie *et al* (2003) classification, we identify HG1 as the Basal subtype; LG1 as Luminal A; LG2, HG3 and HG4 as Luminal B and HG2 as the Her2+ subtype.

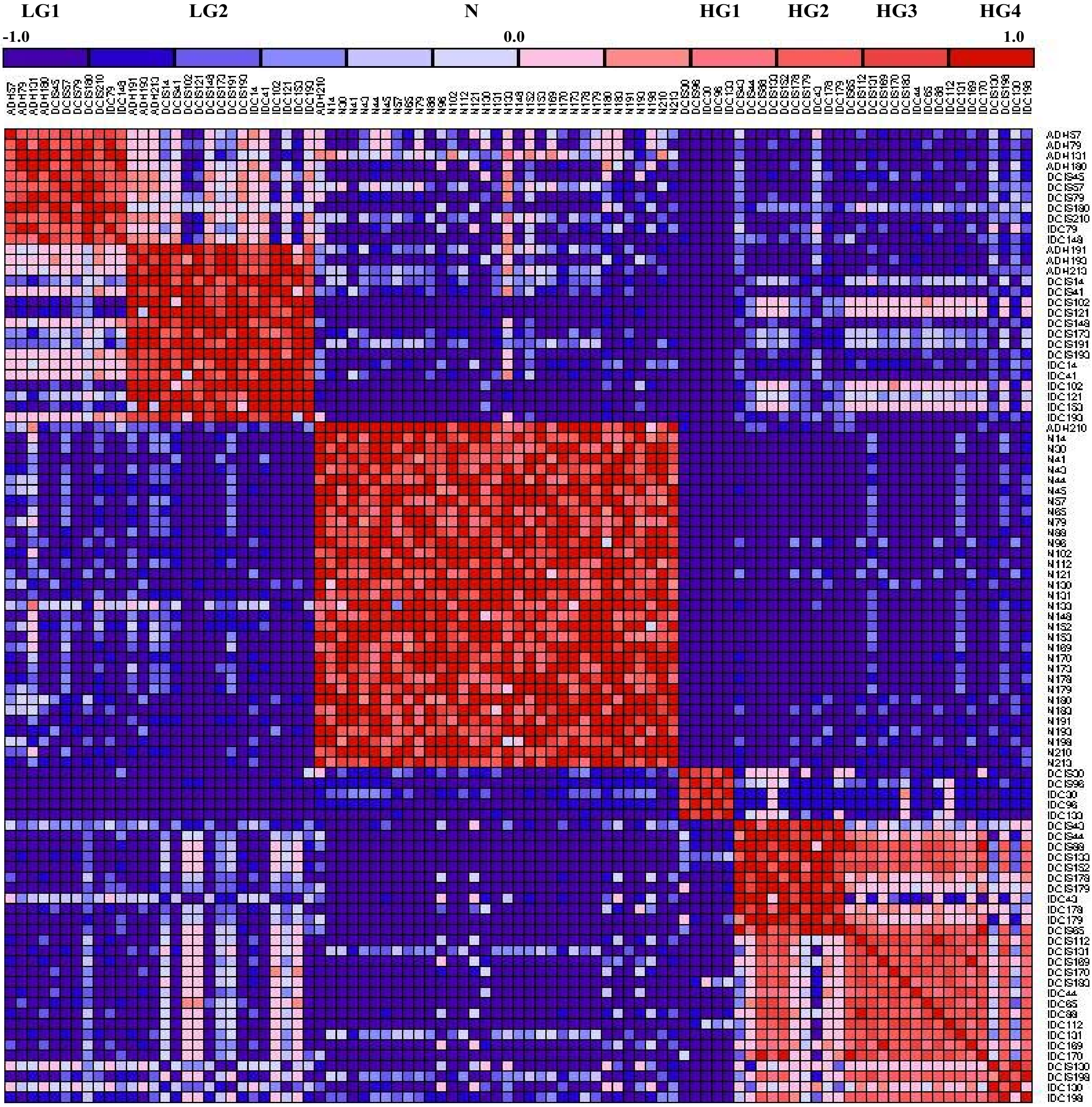

Figure 3. Heatmap of the agreement matrix for $k = 7$ clusters. Red/green represent strong/weak agreement across clustering methods and data perturbations. The normals and the LG1 and LG2 are cleanly separated while the HG1, HG2, HG3 and HG4 separation is weaker. We find that the optimum number of clusters using gap-statistics oscillates between 6 and 7 with the HG3 and HG4 clusters merging at $k = 6$.

found 223 gene markers which distinguish the group LG from HG. A subset of 10 markers was selected based on their performance on leave-one-out cross-validation experiments for weighted voting (WV) and k-Nearest Neighbors (kNN) classification models. The models trained on these 10 markers produced only 1 FP error (DCIS #79) and 1 FN error (DCIS #183) in leave-one-out experiments.

3.3.2 *Characteristic markers for LG:* Again, using the SNR test and leave-one-out experiments for the WV and kNN models, we identified 10 markers which distinguish the LG samples from all others (HG and N) with 90% accuracy. We find that RBSK, *Homo sapiens* cDNA FLJ12924 fis, clone NT2RP2004709 and CRIP1 are up-regulated in the LG group, and EYA2, ANXA1, RUNX3, DKFZp762A227, GPRC5B are down-regulated in the LG group.

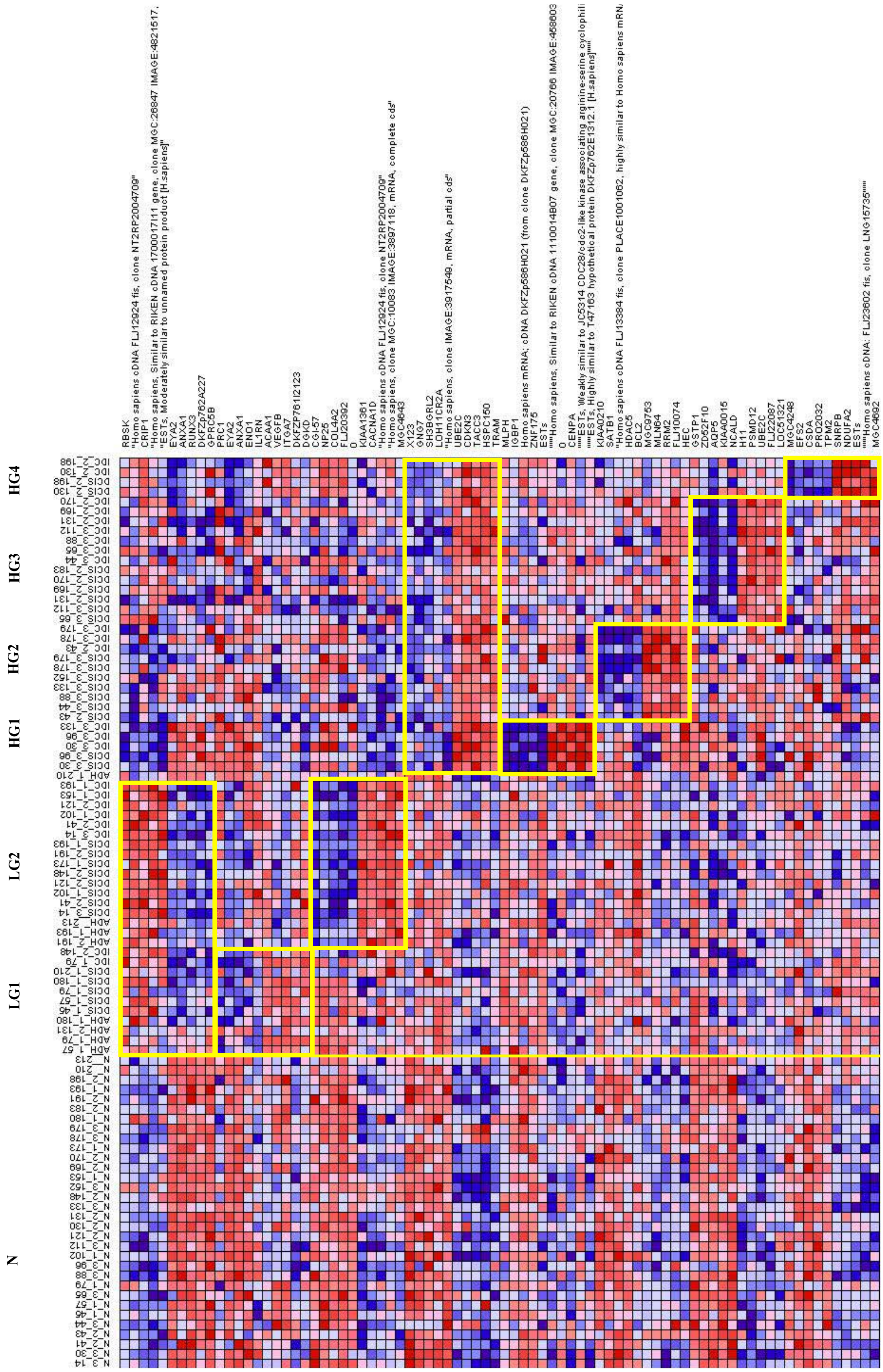

Figure 4. Heatmap of the expression levels of the top 10 markers for each subtype identified in the data. Red/green represent up/down regulation relative to black. Each subgroup is in a framed box to identify its samples and distinguish gene markers. The signatures of the genes specific to each subtype stand out distinctly compared to all other subtypes.

Table 2. Accuracy of weighted voting classifiers in distinguishing samples in a given subtype from all other samples

Group	Sensitivity	Specificity	Accuracy
LG	89.29	90.77	90.32
LG1	81.82	91.46	90.32
LG2	100.00	90.79	92.47
HG	96.88	95.08	95.70
HG1	100.00	96.59	96.77
HG2	100.00	96.39	96.77
HG3	84.62	92.50	91.40
HG4	100.00	94.38	94.62

The accuracy scores were computed using leave-one-out experiments. The gene markers for classification were selected as the top genes based on their collective power to accurately discriminate between a group and its complement.

3.3.3 *Characteristic markers for HG:* Here the classification accuracy was 97% for the markers shown in the table with 3 FP and no FN errors. The top markers up-regulated in HG are TRAM, HSPC150, TACC3, CDKN3, UBE2C, and top markers down-regulated in HG are X123, GNG7, SH3BGRL2, LOH11CR2A and *Homo sapiens*, clone IMAGE:3917549 mRNA, partial cds.

3.3.4 *Low grade substructure:* Table 1 shows that both LG1 and LG2 are ER+, PR+ and HER2-, which explains their pathological classification as low grade. We note that LG2 has a greater fraction of grade II samples compared to LG1 which identifies LG2 as the more aggressive subtype. The genes that discriminate LG1 from other low and high grade subgroups include the down-regulated BIRC5 (survivin) gene, which inhibits apoptosis and is suggested as a marker of poor prognosis in different cancer types (*see* Fangusaro *et al* 2005; Lee *et al* 2005). Two others are ACAA1 and ACOX1 enzymes, which are involved in fatty acid metabolism. LG2 markers include 190 genes, among which are many oncogenes [BCL2 (down, breast cancer poor prognosis marker), RAD51 (up), EGFR (up), RUNX3 (up), BCL9 (down) and VAV3 (down)] and tumour suppressor gene NME1 (up). The ER and Her2neu status suggest that both LG1 and LG2 are Luminals in the standard nomenclature (Perou *et al* 2000), with LG2 presenting more aggressive features than LG1.

3.3.5 *High grade substructure:* As seen in table 1, all the samples in the HG1 subgroup were ER and PR negative while those in the HG3 and HG4 subgroups were mostly ER and PR positive. The HG2 samples had mixed ER and PR signatures. The HG1 subgroup, which is the worst prognosis group based on clinical characteristics, has discriminatory markers of oncogenes BCL2 (up), RAD51 (down); GSTP1

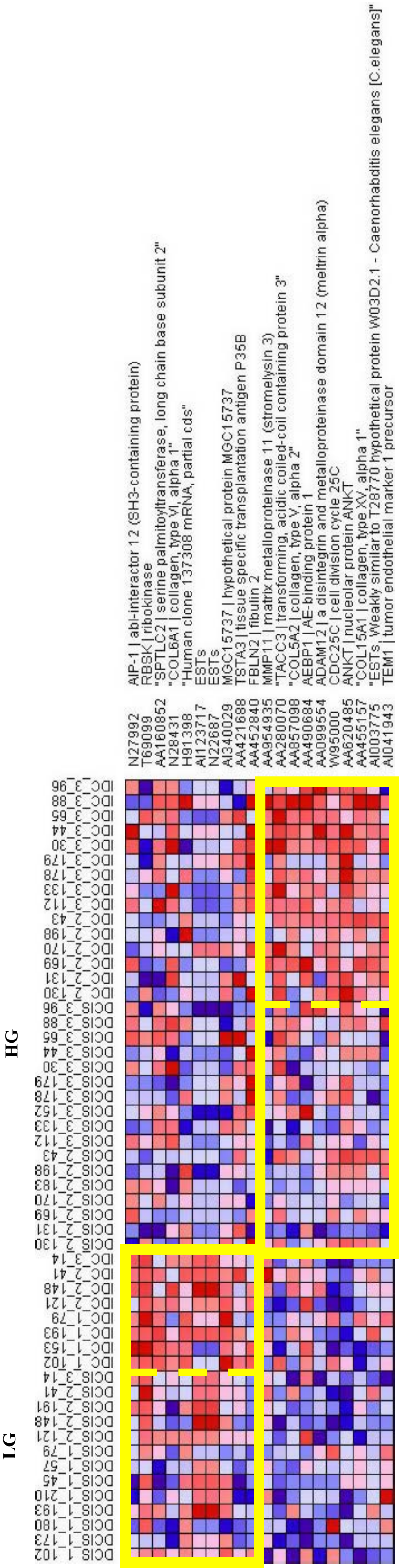

Figure 5. Heatmap of the expression levels of the top markers for progression from DCIS to IDC in the low grade and high grade tumour subgroups. The first 10 genes are top markers for DCIS/IDC progression in both LG and HG. Next 10 genes are top markers for progression in LG, and last 10 genes are top markers for progression in HG. Note that whereas the distinction between grades is not obvious by eye, it is readily identified by the consensus ensemble protocol described in the paper.

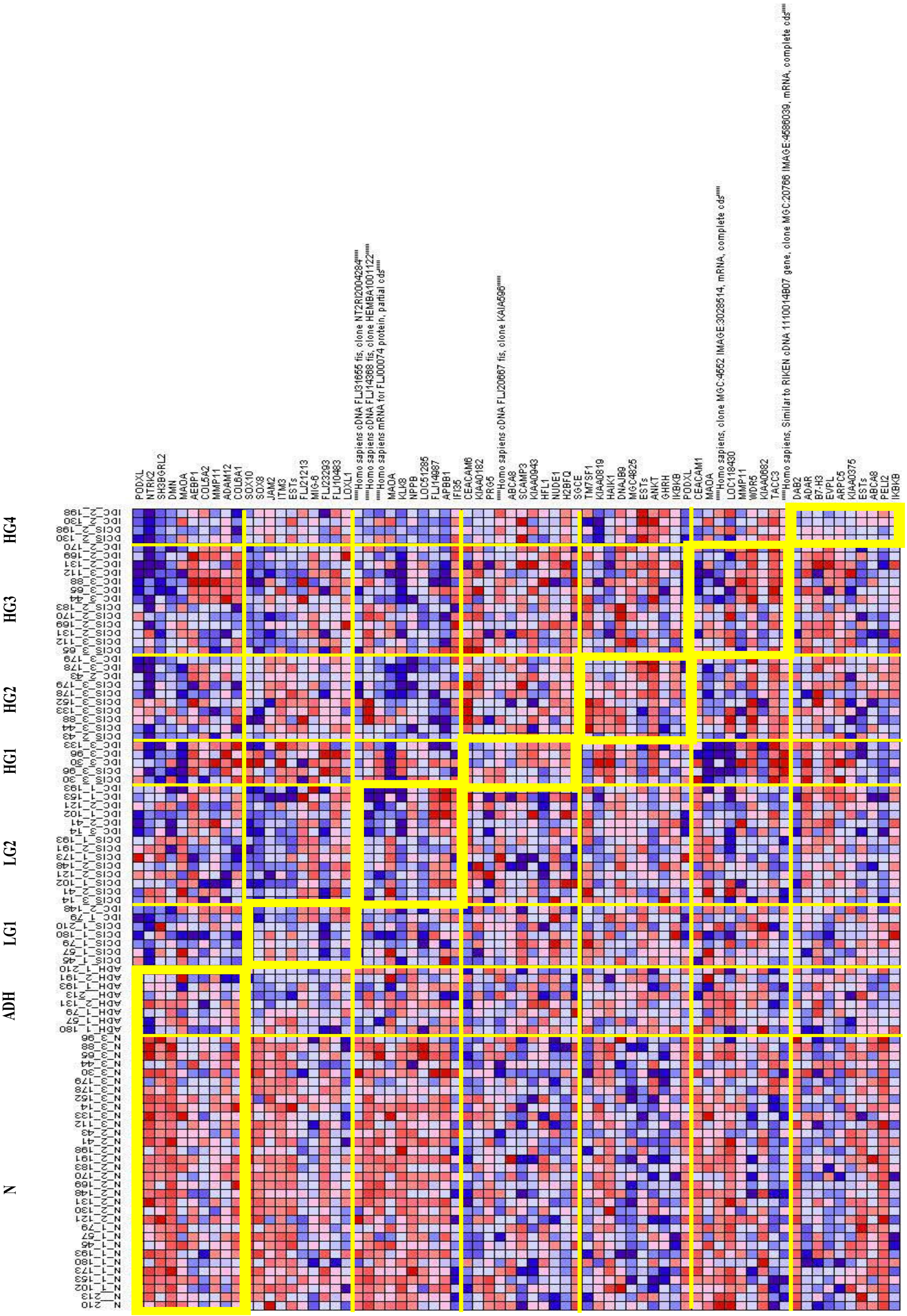

Figure 6. Heatmap of the expression level of top 10 markers for progression from DCIS to IDC for each subtype. Each subgroup is in a framed box to identify its samples and distinguish gene markers. Note that whereas the distinction between subtypes is not obvious by eye, it is readily identified by the consensus ensemble protocol described in the paper.

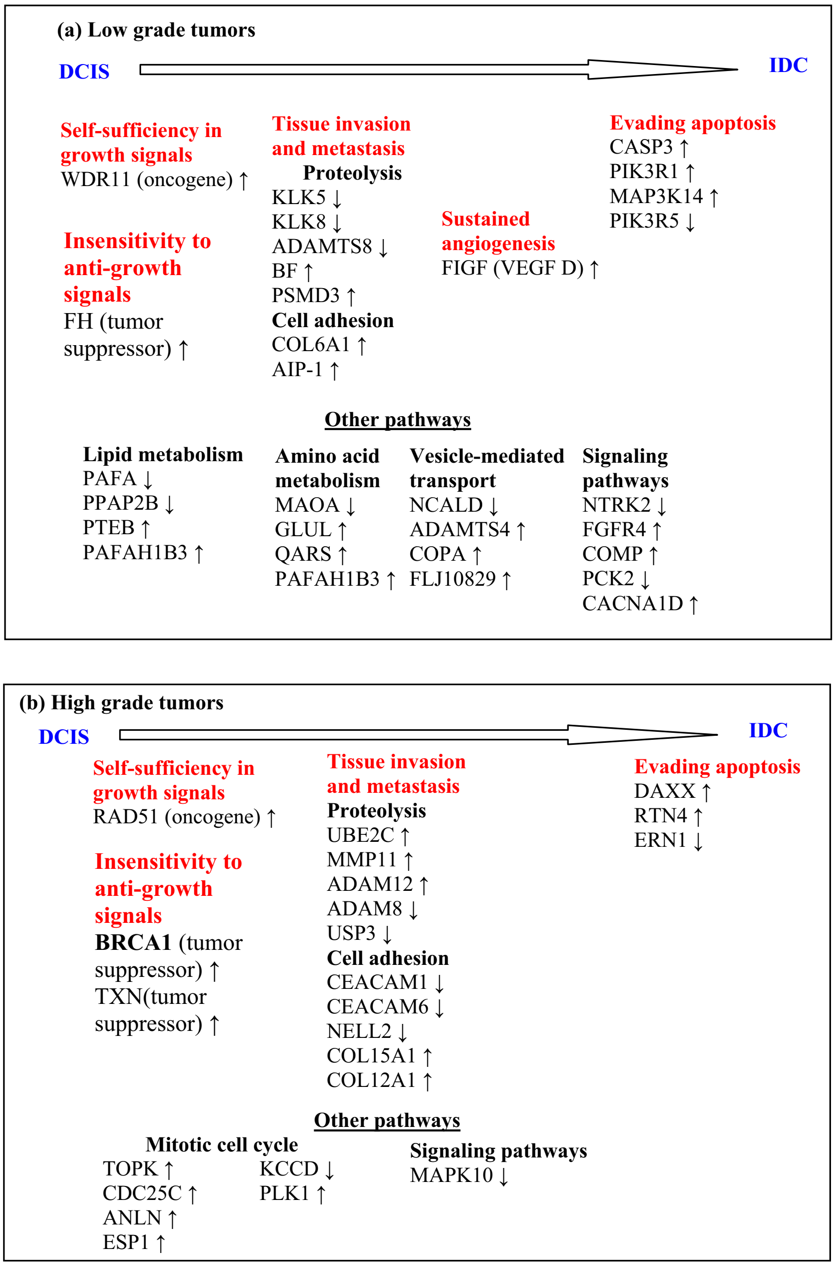

Figure 7. Progression models for low and high grade tumors. Marker genes were placed into Weinberg categories which are indicated in red. Although the exact order of these steps is not known, it has been suggested from other cancers that activation of oncogenes and loss of tumor suppressor genes are usually early events, and induction of angiogenesis is an early to mid-stage event.

(down) gene which leads to cancer susceptibility due to hyper-methylation or polymorphisms, and RRM2 (down). HG2 markers also include up-regulated BCL2 (1.7 fold less up-regulated than in HG1) and down-regulated RRM2. The HG3 markers, include a group of down-regulated genes in chromosomal region 17q23-25 which harbors the ERBB2 amplicon 17q 22.24. These genes are KPNA2 (17q23.1-q23.3), amplified in breast cancer 1 (AIBC1, 17q23.2), Bcl-2 inhibitor of transcription (BIT1, 17q23.2), hypothetical protein TANC2 (17q23.3), and two proteosome protein PSMC5 (17q23-Q25) and PSMD12 (17q24.2). This suggests the possibility that patients in the HG3 subgroup might have a re-arrangement or deletion of genes around the Her2/neu gene leading to loss of regulation or function for these genes which might explain why only 15% of HG3 patients are HER2+, while 53% are HER2 - and 15% are undetermined.

Table 3. Enriched pathways in low and high tumour groups (cf. DAVID *http://david.abcc.ncifcrf.gov/home.jsp*, *P*-value <0.01)

Group	Enriched pathways
LG	Lipid metabolism transcriptional regulation, vesicle-mediated transport, amino acid and derivative metabolism
LG1	Small GTPase, mediated signal transduction, intracellular trafficking and vasicular transport
LG2	Proteolysis collagens mRNA processing
HG	Mitotic cell cycle, ATM signalling pathway, role of BRCA1 BRCA2 and ATR in cancer susceptibility, cell cycle: G2/M checkpoint
HG1	Ion transport
HG2	Cell cycle proteolysis
HG3	Collagens proteolysis
HG4	Proteolysis

The most notable HG4 marker is down-regulated transforming growth factor, beta receptor II (TGFBR2). Mutations in this gene have been associated with the development of various types of tumours. The over-expression of this gene was found to be associated with poor prognosis breast tumours. Overall, gene markers and clinical parameters lead to the conclusion that among the high grade subgroups HG4 is probably the best prognosis group composed of grade II tumours that are all ER+ and PR+.

Based on these observations, we identify HG1 as Basal, HG2 as Her2+, and HG3 and HG4 as additional subtypes of Luminal (Perou *et al* 2000).

Figure 4 presents a heatmap of the top 10 gene markers characteristic of each significant phenotype identified in the data. At each k level, each set of markers distinguishes the subtype from all the other subtypes with an accuracy above 90% in leave-one-out experiments for WV and kNN classification models (*see* table 2). The signatures of the subgroups LG1-HG4 stand out clearly. Table 2 presents sensitivity and specificity scores on leave-one-out cross-validation experiments for WV models. Note that the specificity ranges from 92-97%, and the sensitivity from 82-100%.

3.4 *Identification of significant markers for breast cancer progression*

We have identified the progression markers at different levels of granularity in the data and found that the progression of the disease from non-invasive to invasive status occurs along different pathways. We have also identified the top 10 progression markers within each significant subgroup. Figures 5 and 6 presents the heatmaps of the expression levels of these markers.

Figure 7 summarizes the pathways and the genes for disease progression in low/high grade using an analysis motivated by Hanahan and Weinberg's "Hallmarks of Cancer" (Hanahan and Weinberg, 2000) (and *see*, Hanahan and Folkman, 1996). Progression in the low-grade groups seems to correlate with changes in metabolic and transportation pathways, while in the high grade groups it is related to alterations in cell-cycle and signaling pathways, with distinct subsets of genes involved in each.

Table 3 presents a summary of the significant pathways involved in the low-and high grade subgroups. We find that the differences between the levels in the DCIS and IDC groups are quite subtle and the accuracy of leave-one-out experiments of simple WV models trained to distinguish between DCIS and IDC in each group ranges between 60-70%.

4. Discussion

The main observation of the original paper of Ma *et al* was that the molecular signature of breast cancer is already present in the early (ADH) stage of the disease. The genes that distinguish ADH from normal progressively change their levels away from normal as the disease progresses to DCIS and IDC. They also noticed that that breast cancer progression is defined by distinct markers for low and high grade tumours.

Using a new technique, we refined these observations into a stratification of the molecular signature of breast cancer progression. Using the small gene set provided in the data, we identified at least six different subtypes of breast cancer with distinct patterns of progression. Four of these subtypes (LG1, LG2, HG3, HG4) have a Luminal signature (predominantly ER+, PR+, Her2–); one subtype (HG1) had the triple negative (ER-, PR-, Her2–) characteristic of the Basal subtype, and one subtype (HG2) had a predominantly Her2+ signature (mixed ER, mostly Her2+). The validation of these subtypes on a larger dataset with more genes is currently underway.

References

Alexe G, Dalgin G S, Ramaswamy R, DeLisi C and Bhanot G 2006 Data perturbation independent diagnosis and validation of breast cancer subtypes using clustering and patterns; *Cancer Informatics* **2** 243–274

Benjamini Y and Hochberg Y 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing; *J. R. Stat. Soc. Series B* **57** 289–300

Bussey K J, Kane D, Sunshine M, Narasimhan S, Nishizuka S, Reinhold W C, Zeeberg B, Ajay W and Weinstein J N 2004 MatchMiner: a tool for batch navigation among gene and gene product identifiers; *Genome Biol*. **4** R27

Cheng C-H, Fu A W and Zhang Y 1999 Entropy-based subspace clustering for mining numerical data; in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (San Diego, California, United States ACM Press)

Dempster A, Laird N and Rubin D 1977 Maximum likelihood from incomplete data via the EM algorithm; *J. R. Stat. Soc. Series B* **39** 1–38

Dennis G, Sherman B T, Hosack D A, Yang J, Gao W, Lane H C and Lempicki R A 2003 DAVID: Database for annotation, visualization, and integrated discovery; *Genome Biol.* **4** R60

Everitt B S and Dunn G 2001 *Applied multivariate data analysis* (Arnold and Oxford University Press)

Fangusaro J R, Jiang Y, Holloway M P, Caldas H, Singh V, Boue D R, Hayes J and Altura R A 2005 Survivin, Survivin-2B, and Survivin-deItaEx3 expression in medulloblastoma: biologic markers of tumour morphology and clinical outcome; *Br. J. Cancer* **92** 359–365

Friedman J H and Meulman J J 2004 Clustering objects on subsets of attributes; *J. R. Stat. Soc. Series B* **66** 815–850

Golub T R, Slonim D K, Tamayo P, Huard C, Gaasenbeek M, Mesirov J P, Coller H, Loh M L, Downing J R and Caligiuri M A 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring; *Science* **286** 531–537

Hanahan D and Folkman J 1996 Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis; *Cell* **86** 353–364

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

Hartigan J A 1975 *Clustering algorithms* (New York: John Wiley)

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

Kaufmann L and Rousseeuw P J 1990 *Finding groups in data: An introduction to cluster analysis* First edition (John Wiley)

Lee J P, Chang K H, Han J H and Ryu H S 2005 Survivin, a novel anti-apoptosis inhibitor, expression in uterine cervical cancer and relationship with prognostic factors; *Int. J. Gynecol. Cancer* **15** 113–119

Ma X J, Salunga R, Tuggle J T, Gaudet J, Enright E, McQuary P, Payette T, Pistone M, Stecker K, Zhang B M *et al* 2003 Gene expression profiles of human breast cancer progression; *Proc. Natl. Acad. Sci. USA* **100** 5974–5979

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

Perou C M, Sorlie T, Eisen M B, van de Rijn M, Jeffrey S S, Rees C A, Pollack J R, Ross D T, Johnsen H and Akslen L A 2000 Molecular portraits of human breast tumours; *Nature (London)* **406** 747–752

Sørlie T, Tibshirani R, Parker J, Hastie T, Marron J S, Nobel A, Deng S, Johnsen H *et al* 2003 Repeated observation of breast tumor subtypes in independent gene expression data sets; *Proc. Natl. Acad. Sci. USA* **100** 8418–8423

Strehl A and Ghosh J 2002 Cluster ensembles: a knowledge reuse framework for combining partitionings; in *Eighteenth National Conference on Artificial Intelligence,* July 28-August 01, 2002 (Edmonton, Alberta, Canada) pp 93–98

Tibshirani R, Walther G and Hastie T 2001 Estimating the number of clusters in a dataset via the Gap statistic; *J. R. Stat. Soc. Series B* 411–423

Wall M E, Rechtsteiner A and Rocha L M 2003 Singular value decomposition and principal component analysis; in *A practical approach to microarray data analysis* (eds) D P Berrar, W Dubitzky, M Granzow and M A Norwell (Kluwer) pp 91–109

Zhao Y and Karypis G 2003 Clustering in life sciences (Humana Press)

*e*Publication: 7 July 2007