# BREAST CANCER STRATIFICATION FROM ANALYSIS OF MICRO-ARRAY DATA OF MICRO-DISSECTED SPECIMENS

GABRIELA ALEXE[1,2,*] GUL S. DALGIN[3,*] DANIEL SCANFELD[1] PABLO TAMAYO[1]
galexe@broad.mit.edu sdalgin@bu.edu scanfeld@broad.mit.edu tamayo@broad.mit.edu

JILL P MESIROV[1] SHRIDAR GANESAN[4] CHARLES DELISI[3]
mesirov@broad.mit.edu ganesash@umdnj.edu delisi@bu.edu

GYAN BHANOT[2,4,5]
gyanbhanot@gmail.com

[1] *The Broad Institute of MIT and Harvard, Cambridge MA, 02142, USA*
[2] *Institute for Advanced Study, Princeton, NJ, 08540, USA*
[3] *Boston University, Boston, MA, 02215, USA*
[4] *Cancer Institute of New Jersey, New Brunswick, NJ, 08903, USA*
[5] *Rutgers University, Piscataway, NJ 08854, USA*
[*] *Joint first authors*

We describe a new method based on principal component analysis and robust consensus ensemble clustering to identify and elucidate the subtypes of breast cancer disease. The method was applied to microarray gene expression data using micro-dissection of samples from 36 breast cancer patients with at least two of three pathological stages of disease. Controls were normal breast epithelial cells from 3 disease free patients. Our method identified an optimum set of genes and strong, stable clusters which correlated well with clinical classification into Luminal, Basal and Her2+ subtypes based on ER, PR and Her2 status. It also revealed a hierarchical portrait of disease progression through various grades and stages and identified genes and functional pathways for each stage, grade and disease subtype. We found that gene expression heterogeneity across subtypes is much greater than the heterogeneity of progression from DCIS to IDC within a subtype, suggesting that the disease subtypes are distinct disease processes. The averaging over data perturbations and clustering methods is critical in the robust identification of subtypes and gene markers for grade and progression.

*Keywords*: breast cancer; microarray analysis; disease subtypes; progression; principal component analysis; consensus ensemble clustering.

## 1. Introduction

The probability for women in the US to get breast cancer in their lifetime is ~ 10-13%. Standard treatment includes surgery, radiation, and hormonal, chemo and/or biological therapy. 60-80% of tumors are positive for the estrogen receptor (ER+) and respond to treatment with hormonal agents such as Tamoxifen [10,18]. 20-40% have amplification of the Her2 gene [20], a marker of higher recurrence and poor prognosis. The outcome of Her2+ tumors can be improved using humanized anti-Her2 antibody trastuzumab (Herceptin). 10-15% of tumors neither express the estrogen receptor nor have Her2 amplification [25]. These tumors, called Basal-like [21,23], are high grade with poor prognosis and no known targeted therapy. Overall, therapy for breast cancer is often

confounded by the fact that tumors with similar histopathology have divergent course and varied response to therapy [24].

Microarrays should be able to identify the genes and pathways altered in cancer initiation, progression and metastasis. However, their success has been limited by practical considerations, the most important of which are sensitivity to noise and method of analysis [22]. Such limitations have often resulted in publications with ambiguous and contradictory results and biologically non-intuitive genes and pathways for stratification, making it difficult to move analysis results from bench to bedside.

In this paper, we develop a robust method which addresses these issues. We first use Principal Component Analysis (PCA) [15] to identify the overall structure of clusters in the data and to select the subset of genes that distinguish the clusters. We use this gene set and a new consensus ensemble clustering technique, which averages over several clustering methods and many data perturbations, to identify strong, stable clusters. We use simple criteria to find the optimum number of clusters and describe methods to identify robust markers for subtype/grade separation and disease progression.

Applied to a public breast cancer microarray data set [17], our method results in stable lists of genes and pathways that distinguish high and low grade tumors and progression. A hierarchy of clusters paints a portrait of the disease at varying levels of granularity. With two clusters, the normal samples separate from the disease samples. At the next level of clustering we get a separation of low and high grade samples. The optimal number of clusters identified is seven and correspond to two low grade (LG1 and LG2) and four high grade (HG1-HG4) sub-clusters with strong markers which can distinguish them with sensitivity/specificity in the 80-100% range. Using the gene markers and measured ER, PR, Her2+ levels, we match the sub-clusters to standard clinical classification of breast cancer as in [26]. The low grade clusters correspond to one Luminal A subtype and one Luminal B subtype. The high grade samples correspond to two additional Luminal B subtypes, one Her2+ subtype and one Basal subtype. A major overall observation is that *each sub-cluster contains samples from non-invasive and invasive tumors from the same patient,* which suggests that the sub-clusters are distinct diseases.

## 2. Results

Data provided in [17] (www.geneexpression_ma.org) consisted of microarrays from micro-dissected samples from 36 breast cancer patients. 31 patients had at least two out of three stages of disease: atypical ductal hyperplasia or ADH, ductal carcinoma *in situ* or DCIS and invasive ductal carcinoma or IDC respectively. Five patients had pre-invasive disease (ADH) only. Samples were also collected from normal breast epithelial tissue from three healthy women during mammoplasty. Multiple samples were analyzed for each stage using a 12,000 gene cDNA microarray. The samples were pathologically classified into grades I, II and III. The expression levels of “normal cells” from cancer patients matched those of normal epithelial cells from disease free patients; which meant

that normal samples from cancer patients could serve as controls. The data provided in [17] was limited to 1940 genes across 93 microarrays; 32 from disease free patients, 8 ADH, 30 DCIS and 23 IDC samples.

Fig 1**.** shows the flow chart for our method. The data was normalized and missing entries imputed robustly. PCA showed that 85% of the data variation was due to the first 32 PCs and 207 genes, which were identified from high values of their coefficients in the corresponding eigenvectors. The optimal number of clusters $k_{opt}$ was estimated using gap statistics [29] and silhouette scores [16]. A variety of clustering techniques and data perturbations were then averaged to identify $k = 2,3,\ldots k_{opt}$ clusters using an agreement matrix.

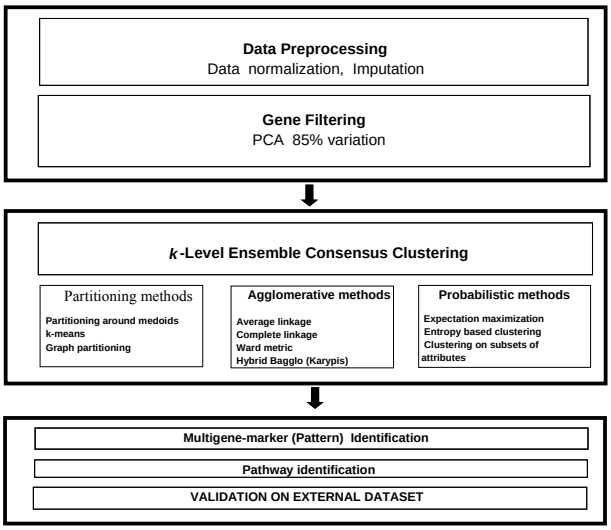


Fig 1: Flow chart of the analysis method.

The clustering results are shown in Fig 2. At $k$=2, the samples separated into the Normal (N) group with all normal samples and one ADH sample, and a Cancer (BCA) group, with only cancer samples. At $k$=3, the normal group was unaltered and the BCA group split into Low grade (LG: 18 samples labeled grade I and 9 samples labeled grade II) and High grade (HG: 13 samples labeled grade II and 19 samples labeled grade III). As $k$ increased from 4 through 7, the LG group split into 2 subgroups (LG1, LG2) and the HG group into 4 subgroups (HG1-HG4). Fig 2. suggests that disease progression is a hierarchical process which is readily and robustly identified by our clustering procedure. Table 1 shows the number of samples in each subtype broken down into the clinical characteristics of stage, ER, PR, Her2, node and grade status in the ADH, DCIS and IDC category for $k$ = 2,3 and 7. Based on this and gene signatures (see below), we identify LG1 as Luminal A; LG2, HG3, HG4 as Luminal B; HG1 as Basal and HG2 as Her2+.

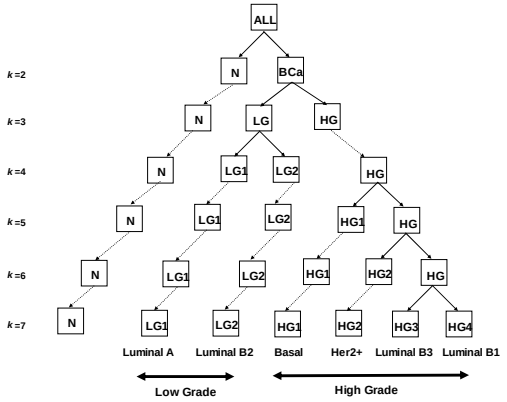


Fig 2: Consensus ensemble clustering tree showing the recursive splitting of samples into six cancer subtypes. The normal cluster separates at the k=2 split and remains distinct.

Table 1: The number of samples in various subtypes as a function of their clinical characteristics.

| Cluster level | Group | Size | Stage | | | | ER | | | PR | | | Her2 | | | Node | | Grade | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *k* | | | 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 | 9 | 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 | |

The ER and Her2 signatures suggest that both LG1 and LG2 are Luminals in the standard nomenclature [21], with LG2 presenting more aggressive features than LG1. In the nomenclature of [27] we identify LG1 as the Luminal A subtype and LG2 as one Luminal B subtype. All 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 have mixed ER and PR signatures. The HG1 subgroup had the worst prognosis based on clinical markers. HG3 markers included a group of down-regulated genes in chromosomal region 17q23-25 which harbors the ERBB2 amplicon 17q 22.24. This identifies the clinical signature of HG3 as HER2-. Based on these observations, we identify HG1 as Basal [21,23], HG2 as Her2+, and HG3 and HG4 as additional subtypes of Luminal B [21].

Table 2: Characteristic and progression markers for grade and subtype.

| Characteristic markers | | | | | | | |
|---|---|---|---|---|---|---|---|
| LG | | | | HG | | | |
| marker | score | pval | FDR | marker | score | pval | FDR |
| EYA2 | -0.89 | 0.01 | 0.06 | X123 | -2.02 | 0.01 | 0.03 |
| ANXA1 | -0.73 | 0.01 | 0.06 | GNG7 | -1.90 | 0.00 | 0.03 |
| RUNX3 | -0.73 | 0.01 | 0.06 | SH3BGRL2 | -1.89 | 0.01 | 0.03 |
| DKFZp762A227 | -0.61 | 0.02 | 0.06 | LOH11CR2A | -1.86 | 0.01 | 0.03 |
| GPRC5B | -0.61 | 0.02 | 0.06 | IMAGE3917549 | -1.85 | 0.02 | 0.03 |
| RBSK | 2.83 | 0.02 | 0.02 | UBE2C | 1.28 | 0.22 | 0.02 |
| FLJ12924 fis, c | 2.85 | 0.02 | 0.02 | CDKN3 | 1.28 | 0.21 | 0.02 |
| CRIP1 | 2.86 | 0.01 | 0.02 | TACC3 | 1.28 | 0.21 | 0.02 |
| sim RIKEN cD | 3.07 | 0.01 | 0.02 | HSPC150 | 1.30 | 0.21 | 0.02 |
| 503671 | 3.12 | 0.01 | 0.02 | TRAM | 1.31 | 0.19 | 0.02 |

| Progression markers | | | | | | | |
|---|---|---|---|---|---|---|---|
| LG | | | | HG | | | |
| marker | score | pval | FDR | marker | score | pval | FDR |
| AIP-1 | 0.89 | 0.01 | 0.06 | MMP11 | 0.80 | 0.00 | 0.02 |
| RBSK | 0.73 | 0.01 | 0.06 | TACC3 | 0.58 | 0.00 | 0.02 |
| STLC2 | 0.67 | 0.01 | 0.06 | COL5A2 | 0.56 | 0.01 | 0.02 |
| COL6A1 | 0.65 | 0.02 | 0.06 | AEBP1 | 0.55 | 0.01 | 0.02 |
| "clone137308" | 0.63 | 0.02 | 0.21 | ADAM12 | 0.50 | 0.01 | 0.04 |
| AI123717 | 0.61 | 0.02 | 0.21 | CDC25C | 0.44 | 0.02 | 0.05 |
| N22687 | 0.57 | 0.02 | 0.45 | ANKT | 0.39 | 0.05 | 0.07 |
| MGC15737 | 0.56 | 0.02 | 0.57 | COL15A1 | 0.37 | 0.05 | 0.07 |
| TSTA3 | 0.53 | 0.04 | 0.57 | AI003775 | 0.35 | 0.07 | 0.20 |
| FBLN2 | 0.42 | 0.09 | 0.57 | TEM1 | 0.33 | 0.07 | 0.20 |

| Characteristic markers | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LG1 | | | | LG2 | | | | HG1 | | | | HG2 | | | | HG3 | | | | HG4 | | | |
| marker | score | pval | FDR | marker | score | pval | FDR | marker | score | pval | FDR | marker | score | pval | FDR | marker | score | pval | FDR | marker | score | pval | FDR |
| PRC1 | -0.20 | 0.02 | 0.04 | CGI-57 | -2.30 | 0.00 | 0.05 | MLPH | -7.45 | 0.00 | 0.04 | KIAA0210 | -1.34 | 0.01 | 0.00 | GSTP1 | -1.37 | 0.00 | 0.09 | MGC4248 | -0.84 | 0.00 | 0.03 |
| EYA2 | -0.20 | 0.02 | 0.04 | NP25 | -1.63 | 0.00 | 0.05 | IGBP1 | -5.12 | 0.00 | 0.04 | SATB1 | -1.31 | 0.00 | 0.00 | ZD52F10 | -1.17 | 0.00 | 0.09 | EFS2 | -0.57 | 0.00 | 0.03 |
| ANXA1 | -0.20 | 0.02 | 0.04 | COL4A2 | -1.33 | 0.00 | 0.05 | DKFZp586H0 | -2.80 | 0.00 | 0.04 | FLJ13384 fis, clon | -1.11 | 0.02 | 0.01 | AQP5 | -1.12 | 0.00 | 0.09 | CSDA | -0.43 | 0.00 | 0.03 |
| ENO1 | -0.20 | 0.02 | 0.04 | FLJ20392 | -1.27 | 0.01 | 0.05 | ZNF175 | -2.63 | 0.00 | 0.04 | HDAC5 | -1.03 | 0.02 | 0.01 | KIAA0015 | -1.04 | 0.01 | 0.09 | PRO2032 | -0.42 | 0.00 | 0.03 |
| IL1RN | -0.21 | 0.02 | 0.04 | BE222894 | -1.11 | 0.00 | 0.05 | T68510 | -2.12 | 0.00 | 0.04 | BCL2 | -1.01 | 0.03 | 0.02 | NCALD | -1.03 | 0.00 | 0.09 | TPM2 | -0.40 | 0.00 | 0.03 |
| ACAA1 | 2.81 | 0.06 | 0.03 | KIAA1361 | 2.39 | 0.02 | 0.02 | RIKEN cDNA | 2.78 | 0.20 | 0.01 | MGC9753 | 0.98 | 0.09 | 0.05 | H11 | 0.87 | 0.01 | 0.09 | SNRPB | 0.43 | 0.01 | 0.03 |
| VEGFB | 2.80 | 0.01 | 0.03 | CACNA1D | 2.47 | 0.03 | 0.02 | AI284295 | 2.85 | 0.20 | 0.01 | MLN64 | 1.02 | 0.03 | 0.02 | PSMD12 | 0.89 | 0.01 | 0.08 | NDUFA2 | 0.55 | 0.01 | 0.03 |
| ITGA7 | 2.67 | 0.03 | 0.03 | FLJ12924 fis, clon | 2.55 | 0.01 | 0.02 | CENPA | 3.34 | 0.20 | 0.01 | RRM2 | 1.04 | 0.02 | 0.01 | UBE2C | 0.91 | 0.01 | 0.08 | AA151092 | 0.63 | 0.01 | 0.03 |
| DKFZP761I212 | 2.63 | 0.02 | 0.03 | MGC:10083 IMAG | 2.80 | 0.01 | 0.02 | JC5314 CDC2 | 3.49 | 0.20 | 0.01 | FLJ10074 | 1.11 | 0.02 | 0.01 | FLJ22087 | 0.93 | 0.01 | 0.08 | FLJ23602 fis, | 0.86 | 0.01 | 0.03 |
| DGKD | 2.59 | 0.02 | 0.03 | MGC4643 | 3.13 | 0.01 | 0.02 | T47163 hypoth | 5.45 | 0.20 | 0.01 | HEC | 1.12 | 0.02 | 0.01 | LOC51321 | 1.15 | 0.00 | 0.08 | MGC4692 | 1.28 | 0.01 | 0.03 |

| Progression markers | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LG1 | | | | LG2 | | | | HG1 | | | | HG2 | | | | HG3 | | | | HG4 | | | |
| marker | score | pval | FDR | marker | score | pval | FDR | marker | score | pval | FDR | marker | score | pval | FDR | marker | score | pval | FDR | marker | score | pval | FDR |
| MIG-6 | 2.88 | 0.00 | 0.02 | NPPB | 0.88 | 0.00 | 0.14 | H2BFQ | 1.83 | 0.00 | 0.02 | AA992253 | 1.31 | 0.00 | 0.14 | MMP11 | 1.37 | 0.00 | 0.05 | 283748 | 3.29 | 0.00 | 0.01 |
| FLJ21213 | 2.86 | 0.00 | 0.02 | ST5 | 0.92 | 0.01 | 0.14 | SCAMP3 | 2.00 | 0.00 | 0.02 | MGC4825 | 1.34 | 0.01 | 0.21 | WDR5 | 1.17 | 0.00 | 0.05 | IDH2 | 3.36 | 0.00 | 0.01 |
| FLJ23293 | 2.84 | 0.00 | 0.02 | FLJ14987 | 0.95 | 0.01 | 0.14 | KIAA0943 | 5.12 | 0.00 | 0.02 | GHRH | 1.11 | 0.02 | 0.35 | KIAA0682 | 1.12 | 0.00 | 0.05 | NS1-BP | 3.19 | 0.00 | 0.01 |
| FLJ10483 | 2.40 | 0.00 | 0.02 | RAD50 | 0.96 | 0.01 | 0.14 | HFL1 | 2.80 | 0.00 | 0.02 | PRELP | 0.98 | 0.02 | 0.35 | TACC3 | 1.15 | 0.01 | 0.05 | KLK6 | 2.99 | 0.00 | 0.01 |
| RAE1 | 2.24 | 0.00 | 0.02 | APBB1 | 0.79 | 0.02 | 0.16 | NUDE1 | 2.12 | 0.00 | 0.02 | MGC12936 | 1.02 | 0.03 | 0.48 | COL5A2 | 0.89 | 0.01 | 0.05 | CA5B | 2.90 | 0.00 | 0.01 |
| LOXL1 | 2.20 | 0.11 | 0.44 | "cl 137308" | 0.86 | 0.02 | 0.17 | CEACAM6 | 5.45 | 0.20 | 0.78 | CCNE1 | 0.88 | 0.08 | 0.61 | TEM1 | 1.04 | 0.01 | 0.05 | SLC35A3 | 2.64 | 0.37 | 0.02 |
| ZMPSTE24 | 2.07 | 0.11 | 0.44 | AKAP12 | 0.72 | 0.03 | 0.17 | KIAA0182 | 2.85 | 0.20 | 0.78 | RRAS2 | 1.11 | 0.09 | 0.61 | AA628867 | 0.72 | 0.02 | 0.06 | KIAA0440 | 3.75 | 0.37 | 0.02 |
| DHCR24 | 1.96 | 0.11 | 0.44 | COL6A1 | 0.63 | 0.04 | 0.19 | PRG5 | 3.49 | 0.20 | 0.78 | SAA1 | 0.92 | 0.09 | 0.61 | MAPK1 | 0.80 | 0.03 | 0.06 | KIAA0375 | 3.99 | 0.37 | 0.02 |
| STAT5A | 1.69 | 0.11 | 0.44 | TSTA3 | 0.64 | 0.04 | 0.19 | FLJ20667 fis, | 2.78 | 0.20 | 0.78 | R09201 | 0.90 | 0.10 | 0.61 | ANKT1 | 0.58 | 0.07 | 0.06 | IKBKB | 2.94 | 0.37 | 0.02 |
| KIAA0523 | 1.45 | 0.11 | 0.44 | AIP-1 | 0.65 | 0.05 | 0.21 | ABCA8 | 2.17 | 0.20 | 0.78 | IKBKB | 0.75 | 0.18 | 0.68 | AI123617 | 0.58 | 0.08 | 0.06 | N55063 | 2.40 | 0.37 | 0.02 |

Using the Signal-to-Noise-Ratio (SNR) test and leave-one-out (LOO) experiments for Weighted Voting (WV) and kNN models, we identified the top 10 markers which distinguish the LG samples from the rest (HG and N) with 90% accuracy. HG samples could be distinguished from the rest (LG and N) with an accuracy of 97%. Table 2 presents the top ten genes for classification of grade and subtype. Each marker set distinguished a subtype from all others with accuracy exceeding 90% in LOO experiments for WV and kNNs. Progression markers were identified for LG/HG and within subtypes as the top up-regulated genes by SNR which distinguish DCIS from IDCs. In LOO experiments for WV and kNN, the average accuracy for predicting progression was 76% in HG/LG and 68-73% within subtypes. We believe that these low accuracies derive from the limited sample sizes and genes in this data. The p values were obtained using permutation experiments and the FDR rates inferred from these. Fig 3. maps the genes identified for progression in different grades into pathways for disease progression using the classification of Hanahan and Weinberg [11,12].

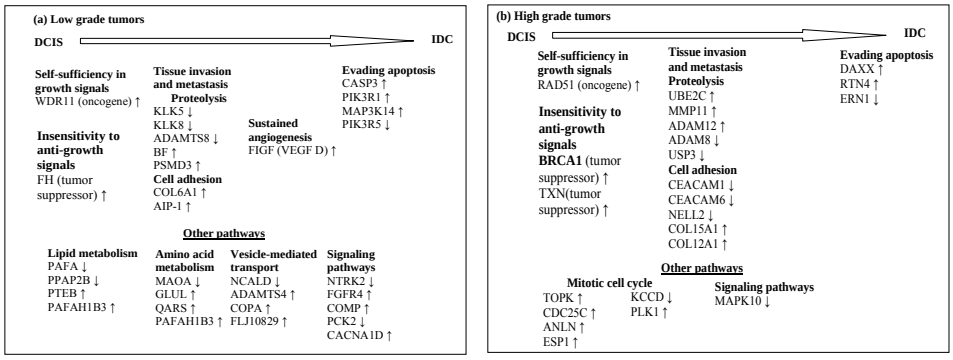


Fig 3: Pathways involved in the progression of low and high grade tumors.

## 3. Discussion

The use of ensemble consensus clustering was critical to identifying the subtypes. PCA by itself could identify useful markers, but could not find the rich stratification discovered by consensus ensemble *k*-clustering. Hierarchical clustering by itself was sensitive to bootstrap, indicating that its clusters are unstable to data perturbation. Robustness of clustering was obtained only after the averaging over clustering techniques and data perturbations.

The development of cancers is accompanied by alterations in cell physiology [12] involving disturbances in many regulatory mechanisms: environment independent growth, insensitivity to antigrowth factors, evasion of apoptosis, limitless replicative potential, sustained angiogenesis and tissue invasion and metastasis. We created progression model for each group based on the pathways (KEGG and GO biological process) and disease associations (OMIM, Genetic Association Database) of genes. The analysis included genes known to be involved in cancer (tumor suppressors and oncogenes) as well as genes not directly associated with cancer but which have a

function critical to tumor development (proteolysis and cell adhesion markers linked to tissue invasion and metastasis). We found that progression from DCIS to IDC occurred along different pathways in low and high grade. In low grade tumors, progression correlated with alteration in lipid metabolism, transcriptional regulation, vesicle-mediated transport, amino-acid and derivative metabolism. High grade tumor progression correlated with alterations in genes in the cell cycle, ATM signaling pathway, BRCA1, BRCA2. Table 3 presents a summary of the enriched pathways in low-and high grade subtypes.

Table 3: Pathways enriched in various grades and subtypes.

| Group | Enriched pathway |
|---|---|
| LG | lipid metabolism, transcriptional regulation, vesicle-mediated transport, amino-acid and derivative metabolism |
| LG1 | small GTPase, mediated signal transduction, intracellular trafficking and vesicular transport |
| LG2 | proteolysis collagens mRNA processing |
| HG | mitotic cell cycle, ATM signaling 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 progression models inferred for each subgroup suggest group-specific genes that are activated/repressed which contribute to the six key steps towards tumor progression as well as specific pathways altered in each subgroup. The number of available genes (1940) limits the identification of progression markers. Nevertheless, the evaluation of these markers in the context of key steps necessary for tumor progression would be valuable in the analysis of the subtypes as distinct diseases.

The main observation of the original paper of Ma et al [17] 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 and high from low grade progressively change their levels away from Normal as the disease progresses to DCIS and IDC. Our results, particularly the hierarchy we see when the data is grouped into $k$=2,3,…7 clusters (Fig 2.) agree with this observation.

Our methods identified six different subtypes of breast cancer with distinct patterns of progression. From histopathology, four subtypes (LG1, LG2, HG3, HG4) had a strong Luminal signature (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.

At $k$=7, each of the six BCA clusters always contained samples in both DCIS and IDC stages *from the same patient*. This strong heterogeneity in the genetic signature of subtypes (progression within a subtype is less distinct than the subtypes themselves) suggests that breast cancer decides its progression path early and progression happens along different pathways in each subtype.

## 4. Methods

After normalization [17] and imputation of genes with < 5% missing entries using a dynamical *k*NN approach [1] we were left with 1927 genes for 93 samples. PCA [7,16,30] was done using singular value decomposition on this matrix and the eigenvectors of the largest eigenvalues that accounted for 85% of the variation in the data were used to find the subset of genes with coefficients in the top 25% in absolute value in these eigenvectors. This collection of genes was then used to find robust clusters in the data.

The optimal number of clusters was identified using gap statistics [29] and silhouette scores [16]. Next, ensemble consensus clustering [19,28] was used to divide the data into $k$=2, 3,…, $k_{opt}$ clusters. The technique has two parts: [1] a method to generate clustering solutions using different methods applied to many perturbations of the data, and [2] a consensus function to combines the clusters into a single output clustering.

150 datasets were created by bootstrapping samples and genes and each was partitioned into $k$=2,…, $k_{opt}$ clusters using representative methods from the three major clustering methods: *Partitioning*: partition around medoids (PAM) [16], *k*-means [13] and graph partitioning [31]; *Agglomerative*: hierarchical clustering based on average linkage, complete linkage and Ward metric [16], including bagglo [31]; *Probabilistic*: expectation maximization (EM) method [5], entropy-based-clustering (ENCLUST) [4], clustering on subsets of attributes (COSA) [8]. Each clustering method was applied 50 times with different parameter initialization on the full dataset, and once on each of the 150 datasets. The 200 resulting clusters were combined into 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. For each $k$, the agreement matrices were averaged across the clustering techniques. The samples were then sorted using simulated annealing to create a $k$-block diagonal structure in the combined agreement matrix.

For optimum sensitivity, we used the full collection of genes to identify a large pool of markers to distinguish the group of interest from its complement. This was done using SNR [9] with a permutation p-value of 0.1 and a False Discovery Rate (FDR) [2]of 0.5. We then identified a smaller subset of these genes by using stringent criteria which combined (a) a permutation p-value of 0.05 (b) stability to sample perturbation through bootstrapping (c) stability to leave-one-out experiments in top 25% genes selected by WV and kNN classifiers which distinguish the two classes with specificity and sensitivity above 0.75. The analysis used GenePattern from the Broad Institute: (http://www.broad.mit.edu/cancer/software/genepattern/).

We used the bioinformatics public resources DAVID [6], iHOP [14], and MatchMiner [3] for functional and pathway annotation. 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 using Kappa Statistics.

Acknowledgments:
We thank Dr. Xiao-Jun Ma (Arcturus, CA) for providing the gene expression breast cancer data and Dr. Stefano Monti (The Broad Institute of MIT and Harvard) for discussions on estimating the number of clusters.

References

[1] Alexe, G., Dalgin, G.S., Ramaswamy, R., DeLisi, C., and Bhanot, G., Data Perturbation Independent Diagnosis and Validation of Breast Cancer Subtypes Using Clustering and Patterns, *Cancer Informatics* 2 (online): 2006.

[2] Benjamini, Y. and Hochberg, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistics Society (Series B),* 57:289-300, 1995.

[3] Bussey, K.J., Kane, D., Sunshine, M., Narasimhan, S., Nishizuka, S., Reinhold, W.C., Zeeberg, B., Ajay, W., and Weinstein, J.N., MatchMiner: a tool for batch navigation among gene and gene product identifiers, *Genome Biol.,* 4(4):R27, 2003.

[4] Cheng, C-H., Fu, A.W., and Zhang, Y., 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, 1999.

[5] Dempster, A., Laird, N., and Rubin, D., Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society ( Series B),* 39:11-38, 1977.

[6] Dennis, G., Sherman, B.T., Hosack, D.A,, Yang, J., Gao, W., Lane,H.C., and Lempicki, R.A., DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.,* 4:R60, 2003.

[7] Everitt, B.S. and Dunn, G., *Applied Multivariate Data Analysis*, London: Arnold, 2001.

[8] Friedman, J.H. and Meulman, J.J., Clustering objects on subsets of attributes, *Journal of the Royal Statistical Society (Series B),* 66:815-850, 2004.

[9] Golub, T.R., Slonim, D.K,, Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science,* 286(5439):531-537, 1999.

[10] Gruvberger, S., Ringner, M., Chen, Y., Panavally, S., Saal, L.H., Borg, A., Ferno, M., Peterson, C., and Meltzer, P.S., Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns, *Cancer Res.,* 61(16):5979-5984, 2001.

[11] Hanahan, D. and Folkman, J., Patterns and emerging mechanisms of the angiogenic switch during tumorigenesis, *Cell,* 86(3):353-364, 1996.

[12] Hanahan, D. and Weinberg, R.A., The hallmarks of cancer, *Cell,* 100(1):57-70, 2000.

[13] Hartigan, J.A., *Clustering algorithms*, New York: John Wiley & Sons, 1975.

[14] Hoffmann, R. and Valencia, A., A gene network for navigating the literature, *Nat. Genet.,* 36(7):664, 2004.

[15] Jolliffe, I.T., *Principal Component Analysis*, Springer, 2002.

[16] Kaufmann, L. and Rousseeuw, P.J., *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, 1990.

[17] Ma, X.J,, Salunga, R., Tuggle, J.T,, Gaudet, J., Enright, E., McQuary, P., Payette, T., Pistone, M., Stecker, K., Zhang, B.M,, Zhou, Y.X,, Varnholt, H., Smith, B., Gadd, M., Chatfield, E., Kessler, J., Baer, T.M,, Erlander, M.G., and Sgroi, D.C., Gene expression profiles of human breast cancer progression, *Proc. Natl. Acad. Sci. USA,* 100(1):5974-5979, 2003.

[18] Mauriac, L., Aromatase inhibitors: effective endocrine therapy in the early adjuvant setting for postmenopausal women with hormone-responsive breast cancer, *Best Pract Res. Clin. Endocrinol Metab,* 20(Suppl 1):S15-29, 2006.

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

[20] Morris, S.R. and Carey, L.A., Molecular profiling in breast cancer, *Rev. Endocr Metab Disord,* 2007.

[21] 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., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, SX., Lonning, P.E., Borresen-Dale, AL., Brown, P.O., and Botstein, D., Molecular portraits of human breast tumours, *Nature,* 406(6797):747-752, 2000.

[22] Quackenbush, J., Microarray analysis and tumor classification, *N. Engl. J. Med.,*354(23): 2463-2472, 2006.

[23] Rakha, E.A., El-Rehim, D.A., Paish, C., Green, A.R., Lee, A.H., Robertson, J.F., Blamey, R.W., Macmillan, D., and Ellis, I.O., Basal phenotype identifies a poor prognostic subgroup of breast cancer of clinical importance, *Eur. J. Cancer,* 42(18): 3149-3156, 2006.

[24] Sorlie, T., Perou, C.M., Fan, C., Geisler, S., Aas, T., Nobel, A., Anker, G., Akslen, L.A., Botstein, D., Borresen-Dale, A.L., and Lonning, P.E., Gene expression profiles do not consistently predict the clinical treatment response in locally advanced breast cancer, *Mol. Cancer Ther.,* 5(11):2914-2918, 2006.

[25] Sorlie, T., Wang, Y., Xiao, C., Johnsen, H., Naume, B., Samaha, R.R., and Borresen-Dale, A.L., Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms, *BMC Genomics,* 7:127, 2006.

[26] Sorlie, T., Wang, Y., Xiao, C., Johnsen, H., Naume, B., Samaha, R.R., and Borresen-Dale, A.L., Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms, *BMC Genomics,* 7:127, 2006.

[27] Sorlie, T., Wang, Y., Xiao, C., Johnsen, H., Naume, B., Samaha, R.R., and Borresen-Dale, A.L., Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms, *BMC Genomics,* 7:127, 2006.

[28] Strehl, A. and Ghosh, J., Cluster ensembles: a knowledge reuse framework for combining partitionings, In: *Eighteenth national conference on Artificial intelligence*, Edmonton, Alberta, Canada , 2002.

[29] Tibshirani, R., Walther, G., and Hastie, T., Estimating the number of clusters in a dataset via the Gap statistic, *Journal of the Royal Statistics Society (Series B),* 411-423, 2001.

[30] Wall, M.E., Rechtsteiner, A., and Rocha, L.M., *A Practical Approach to Microarray Data Analysis*, Norwell, MA: Kluwer, 2003.

[31] Zhao, Y. and Karypis, G., *Clustering in Life Sciences*, Humana Press, 2003.