Statistically Weighted Voting Analysis of Microarrays for Molecular Pattern Selection and Discovery Cancer Genotypes

Vladimir A. Kuznetsov[1] , Oleg V. Senko[2],, Lance D. Miller[1] and Anna V. Ivshina[1]

Genome Institute of Singapore, 60 Biopolis Str, Singapore ; Computer Center of Russian Acad. of Sciences, Moscow, Russia.

Summary
We developed a methodological approach to genetic class discovery using gene expression microarray data, which is based a on statistically-oriented class-prediction method called Statistically Weighted Voting (SWV) analysis integrating with clinical risk factor and survival analyses, and statistics of Gene Ontology annotation terms which we use to validate candidate biomarker selection. Our approach provides a "voting" class prediction function constructed using the most informative and robust discrete segments (sub-regions) of all covariate ranges and their gradated pairs, which thus allows to model the interactions of variables (genes). We show here that the SWV-based methodology can be adapted for microarray data and profitably used to biomarker selection and discovered two genetic classes associated with essentially improvement of classical histological grade II of human breast cancer. Our findings show that small and reliable genetic grade signatures could improve an individual prognosis for patients with histologic grated II and, thus after further biomedical validation, be used in therapeutic planning for breast cancer patients.

Key words:
Voting Algorithms, Biomarker Selection, Prediction, Microarray, Histologic Grades, Cancer Classification.

Introduction

Gene expression microarrays are assays for quantitative studying of transcript abundance profile of large proportion of genes in a multi-cell sample. To date, global gene expression patterns have been used to classify human cancers into genetic classes related to different clinical outcomes [1, 2, 3, 4]. In these studies, different unsupervised methods such as hierarchical cluster analysis have been used [2, 3]. However, such methods, based on heuristic models, are quite sensitive to the number of samples, population bias in a sample set, missing values, model of distance measure, and different sources of technical noise. It is therefore no surprise that the microarray predictions of biologically and clinically significant tumor classes, as discovered by the different research groups using unsupervised methods, often exhibit poor reproducibility. Therefore, there is a serious concern regarding the ability of unsupervised methods to predict meaningful biologically and clinically significant tumor classes; these classifications, generated by cluster analysis, still remain extremely unstable [1,4,5]. There exist many class prediction approaches that, when applied to a given expression dataset, could result in a range of classification accuracies and gene numbers that comprise the classifier (a subset of high-informative and robust predictors selected by a supervised method). Supervised learning algorithms could provide more accurate statistically-oriented results than unsupervised methods, but, usually, they are used to identify novel markers in class prediction, not in class discovery tasks. A number of authors have underlined critical issues in gene selection bias, error estimation, fragility of gene signatures, and overoptimistic performance estimation due to model overfit [6,7]. Thus, there does not exist one “correct” method. This has motivated us to develop a more suitable and better validated methodological approach for inference of unknown classes from microarray data.

Our approach to genetic class discovery, which is based on supervised learning, uses a statistically-oriented class-prediction method called Statistically Weighted Syndromes (SWS) [8, 9]. Briefly, SWS provides a "voting" class prediction function constructed using the *most informative* and *robust* discrete segments (sub-regions) of all covariate ranges, which are thus discretized. The variants of the SWS have been successful in accurate predicting therapeutic outcome in bladder cancer patients using limited clinical data [8,10]. In clinical trials, it is important to minimize the cost of the trial and the total number of patients. In the case of small numbers of patients, SWS methodology has demonstrated higher robustness and predictive power than logistic

Manuscript received December 5, 2006.
Manuscript revised December 25, 2006.

regression-based analyses and classification and regression tree (CART) methods [10,11].

Breast cancer is most common malignancy among women. Histological grading of breast cancer provides clinically important prognostic information and defines morphological subtypes informative of patient risk. Approximately 50% of all breast cancers are classified as grade II [1,3], which is less informative for clinical decisions due to biological heterogeneity and intermediate risk of cancer recurrence. To discover the molecular basis of histologic grade II, we analyzed genome-wide expression profiles of 315 primary invasive breast tumors. In this work, using advised and computationally intensive version of SWS methodology, though not previously applied to large-dimension (microarray) data, combining with survival analysis, multivariate correlation analysis and gene ontology analysis we identified several small subsets of highly significant grade-associated markers, which could accurately classify tumors of grade I (G1) and grade III (G3) histology, and dichotomize G2 tumors into two highly discriminant classes (termed *G2a* and *G2b genetic grades*) with patient survival outcomes highly similar to those with G1 and G3 histology, respectively.

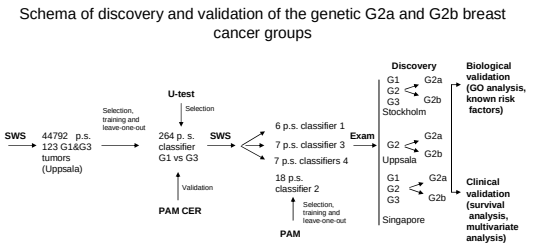
2. Methods, Algorithms and Equations

2.1. Breast cancer data and microarrays

To study the relationship between gene expression and histologic grade, we analyzed the expression patterns of approximately 23,000 gene transcripts (representing by 44,928 probesets (p.s.) on Affymetrix U133A and U133B arrays) in 315 primary breast tumors (NCBI Gene Expression Omnibus (GEO) data sets GSE4922 and GSE1456). The tumor samples were derived from three independent population-based cohorts: Uppsala (249 samples), Singapore (40 samples) and Stockholm (58 samples) (Figure 1) enabling the robust identification and cross-cohort validation of highly significant and predictive grade-associated genes. Details on patients, clinical information, tumor samples, microarrays see in [12].

2.2. A basis of SWV algorithm

In simplified terms, the statistically weighted voting (SWV) analysis of microarrys class prediction process can be described as follows. A *training set* consisting of samples of known classes (e.g., histologic grade I (G1) and histologic grade III (G3) tumors) is used to select the *variables* (i.e., gene expression measurements; *probesets or predictors*), that allow the most accurate discrimination (or prediction) of the samples in the training set. Once the SWV is *trained* on the optimal set of variables,



SWS: Statistically Weighted Syndromes method; PAM: Prediction Analysis for microarray method; CER: Class Error Rate plot; p.s. probe set:G1: grade 1; G3: grade 2; G3: grade 3; G2a: grade 2a; G2b: grade 2b; GO: gene ontology.

Figure 1. Schema of discovery and validation of genetically different subgroups within breast cancer patients with histologic grade II.

it is then applied to an independent *exam set* (i.e., a new set of samples not used in training) to validate its prediction accuracy. More details are given below.

Briefly, for constructing the class prediction function, the SWV uses the training set $\tilde{S}_0$ (comprised of G1 and G3 tumor samples) to evaluate statistically the weight of the graduated "informative" variables (predictors), and all possible pairs of these predictors. The predictors are automatically selected by SWV from n (n=44,500) probe sets (i.e., gene expression measurements) on U133A and U133B Affymetrix Genechips. The description of each patient includes n (potential) prognostic variables $X_1,\ldots,X_n$ (signals from probe sets of the U133A and U133B microarrays) and information about class to which a patient belongs. In particular, the predictors might be able to discriminate G1 and G3 tumors with minimum "a posteriori probability". Reliability of the SWV class prediction function is based on the standard "leave-one-out procedure" and on an additional *exam* of the class prediction ability on one or more independent sample populations (i.e., patient cohorts). In this application the G2 tumor samples from the Uppsala, Singapore and Stockholm cohorts have been used as exam datasets to test the SWV class prediction function.

Let us consider the available n-dimension domain of the variables $X_1,\ldots,X_n$ as prognostic variable space. The SWV algorithm is based on calculating

the posteriori probabilities of the tumors belonging to one of two classes using a weighted voting scheme involving the sets of so called "syndromes". A syndrome is the sub-region of prognostic variable space. Within the syndrome, one class of samples (for instance, G3 tumors) must be significantly highly represented than another class (for instance, G1), and in other sub-region(s) the inverse relationship should be observed. In the present version of the SWV method, one-dimensional and two-dimensional sub-regions (syndromes) are used.

Let b'_i and b''_i denote the boundaries of the sub-region for the variable X_i (the *i-th* probe set); $b'_i \geq X_i > b''_i$. One-dimensional syndrome for the variable X_i is defined as the set of points in variable space for which inequalities $b'_i \geq X_i > b''_i$ are satisfied. Two-dimensional syndrome for variables $X_{i'}$ and $X_{i''}$ is defined as a set of points in variable space for which inequalities $b'_{i'} \geq X_{i'} > b''_{i'}$ and $b'_{i''} \geq X_{i''} > b''_{i''}$ are satisfied. The syndromes are constructed at the initial stage of training using the optimal partitioning (OP) algorithm described below.

SWV training algorithm is based on several steps:
1) optimal recoding (partitioning) of the given variables (signal intensity values) to obtain discrete-valued variables with two or more gradations;
2) selection of the most informative and robust discrete-valued variables and their paired combinations (termed syndromes) that together best characterize the classes of interest;
3) tallying the statistically weighted votes of these syndromes to allow us to compute the value of the outcome prediction function.

In this study we present an advanced procedure of SWS method based on permutation statistics and high-intensive computational estimates of significant cut-off values providing an effective procedure of predictor selection.

2.2.1. Optimal partitioning (OP)

OP method is used for constructing the optimal syndromes for each class (G1 and G3) using the training set $\tilde{S}_0$. The OP is based on the optimal partitioning of some potential prognostic variable X_i the range that allows the best separation of the samples belonging to different classes. To evaluate the separating ability of partition *R* (see below) in the training set $\tilde{S}_0$ the chi-2 functional is used [9]. The optimal partitions are searched inside observed variable domain which contains partitions with critical values not greater than a fixed threshold (defined below). The "informative" partition with the maximal value of the chi-2 functional is considered optimal for the given variable.

2.2.2. Stability of partitioning

Another important characteristic that allows evaluation the prognostic ability of partitioning model for specific variables is the index of *boundary instability*. Let $R_o, R_1, \ldots, R_m$ be optimal partitions of variable X_i ranges that is calculated by training set $\tilde{S}_0, \tilde{S}_1, \ldots, \tilde{S}_m$, where $\tilde{S}_k$ is the training set without description of the k^{th} sample. Let K_j denote the different classes (*j=1,2*). Let $b_1^k, \ldots, b_{r-1}^k$ be boundary points of optimal partition R_k found by training set $\tilde{S}_k$; D_i is the variance of variable X_i. The boundary instability index $\kappa(\tilde{S}_0, K_j, r)$ for partitioning with r elements is calculated as the ratio:

$$\kappa(\tilde{S}_0, K_j, r) = \frac{1}{D_i(r-1)}[\sum_{k=1}^{m} \sum_{l=1}^{r-1} (b_l^k - b_l^0)^2].$$

2.2.3. Selecting of optimal variables set

The OP can be used at the initial stage of training for reducing the dimension of the prognostic variables set. Selection of the optimal set of prognostic variables depends on a sufficiently high partition value determined by the Chi-2 function. The threshold for selection of informative variables is estimated based on p-value of Chi-2 function estimated based on a permutation procedure.

The additional criterion of selection of prognostic variables is the instability index $\kappa(\tilde{S}_0, K_j, r)$. The variable is used if value $\kappa(\tilde{S}_0, K_j, r)$ is less than threshold κ_0 , defined *a priori* by the user. When the partition of the given variable is instable ($\kappa(\tilde{S}_0, K_j, r) < \kappa_0$), the variable is removed from the final optimal set of prognostic variables. Finally, the optimal set of prognostic variables is defined if both selection criteria are fulfilled.

2.2.4. The weighted voting procedure

Let $\tilde{Q}_j^0$ denote the set of constructed syndromes for class K_j. Let $\mathbf{x}^*$ denote the point of parametric space. The SWV estimates a posteriori probability $P_j^{sv}(\mathbf{x}^*)$ of the class K_j at the point $\mathbf{x}^*$ that belongs to the intersection of syndromes $q_1,\ldots,q_r$ from $\tilde{Q}_j^0$ as follows:

$$P_j^{sv}(\mathbf{x}^*) = \frac{\sum_{i=1}^{r} w_i^j \nu_i^j}{\sum_{i=1}^{r} w_i^j}, \quad (1)$$

where ν_i^j is the fraction of class K_j among objects with prognostic variables vectors belonging to syndrome q_i, w_i is the so-called "weight" of syndrome q_i. The weight w_i is calculated by the formula,

$$w_i = \frac{m_i}{m_i+1}\frac{1}{\hat{d}_i},$$

where $\hat{d}_i = (1-\nu_i^i)\nu_i^i + \frac{1}{m_i}(1-\nu_0^j)\nu_0^j$. The estimate of fraction ν_i^j variance has the second term $\frac{1}{m_i}(1-\nu_0^j)\nu_0^j$, which is used to avoid a value $\hat{d}_i$ equal to zero in cases when the given syndrome is associated only with objects of one class from the training set.

The results of testing applied and simulated tasks have demonstrated that formula (1) gives too low of estimates of conditional probabilities for classes that are of smaller fraction in the training set. So, in this study, the additional correction of estimates in (1) has been implemented. The final estimates of conditional probability at point $\mathbf{x}^*$ are calculated as

$$P_j^{SWV}(\mathbf{x}^*) = P_j^{sv}(\mathbf{x}^*)\,\chi(\tilde{S}_0, K_j),$$

where

$$\chi(\tilde{S}_0, K_j) = 1/(\sum P^{sv}(\mathbf{x}_k))$$

and where $\mathbf{x}_k$ is the vector of prognostic variables for the *k-th* samples from the training set.

2.3. Statistical analysis of Gene Ontology (GO) terms

GO analysis was facilitated by PANTHER software (https://panther.appliedbiosystems.com/). Selected gene lists were statistically compared (Mann–Whitney) with a reference list (ie, NCBI Build 35) comprised of all genes represented on the microarray to identify significantly over- and under-represented GO terms.

2.4. Survival analysis

The Kaplan Meier estimate was used to compute survival curves, and the p-value of the likelihood-ratio test was used to assess the statistical significance of the resultant hazard ratios. Disease-free survival (DFS) in the Uppsala, Stockholm cohorts was defined as the time interval from surgery until the first recurrence (local, regional, or distant) or last date of follow-up. Survival statistics were performed in the R survival package.

2.5. Descriptive statistics

For inter-group comparisons using the clinico-pathological measurements, Mann–Whitney U-test statistics were used for continuous variables and one-sided Fisher's exact test used for categorical variables (Statistica-6 and StatXact-6 software).

3. Results

3.1. SWV as the discovery method of novel classes of tumors

Our methodology is based on the schema presented in Fig 1. Beginning with the Uppsala dataset comprised of 68 G1 and 55 G3 tumors, we used SWV optimal partitioning (OP) at the initial stage of training to reduce the dimension of the prognostic set of variables. SWV rank orders the set of probes according to specific algorithmic criteria for assessing differential expression between classes. Based on this two-criteria selection algorithm, we used SWV chi-2 values more than 24.38 (at p-value less then 0.00001); in combination with low boundary instability index criteria (κ_0 <0.1 for 90% of the selected informative variables and κ_0 <0.4 for 10% of the other informative variables). This procedure provides optimal (robust) partitioning of the informative variables and leads to selection of relatively small sets of the potential gene predictors. We also used the U-test with critical value p=0.05 (with Bonferroni correction). Based on these criteria, we selected 264 probesets (see Supplementary Material in [12]). Table 1 shows 25 (of 264) top-level selected probesets exhibiting the highest SWV chi-2 values and the significant cut-off p-values of survival

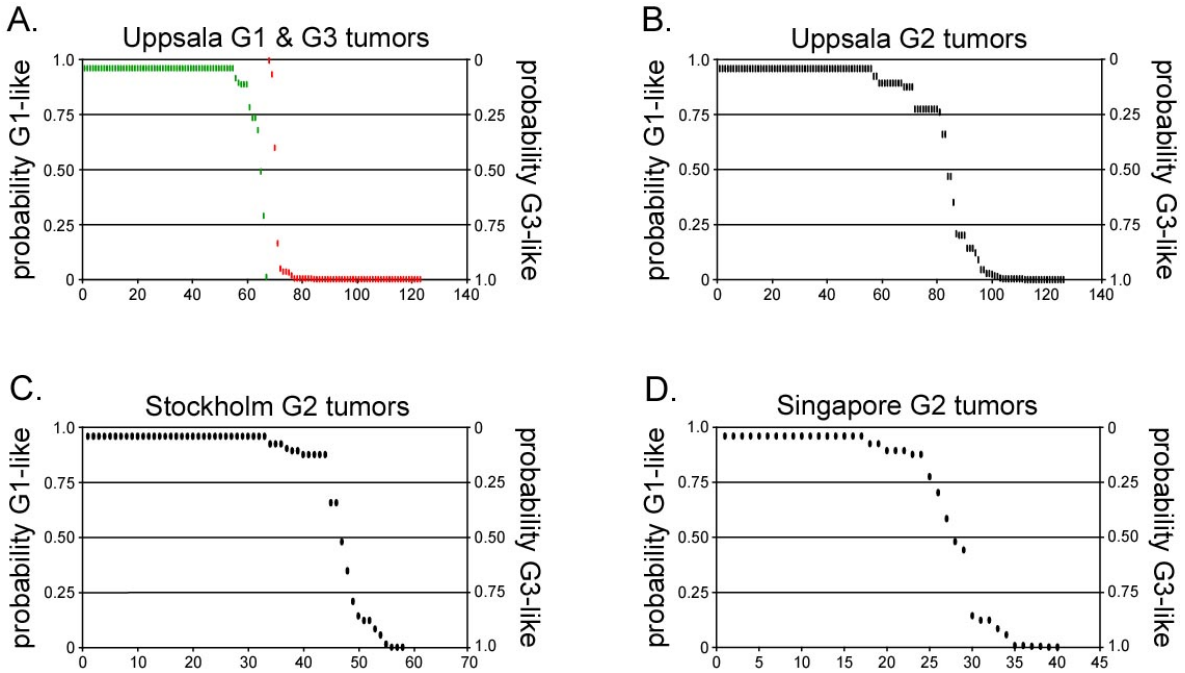


Figure 2. (A-D) Probability (Pr) scores from the SWV genetic grade classifier. Pr scores (0-1) generated by the class prediction algorithm are shown on the y-axes. Number of tumors per classification exercise is shown on the x-axis. For training set on Panel A: Green dot denotes G1 tumor; red dot denotes G3 tumors. Panels B,C and D show the results of predictions for three independent cohorts of patients with grade 2 tumors. In all these cohorts only few patients (0.25<Pr<0.75) might be considered as true Grade 2 tumor patients.

statistics (see below). Using 264 probesets, SWV provided small class error rate (CER) (4.5% for G1, and 5.5% for G3, respectively) when the leave-one-out cross-validation procedure is used. A posterior probability for G1 and G3 was also estimated by PAM [13] for each tumor sample by the leave-one-out cross-validation procedure with resulting CER of 5% for G1, and 6% for G3, respectively.

To extract the smallest possible genetic grade classifier from the 264 p.s., we varied the initial parameters of the SWV algorithm to minimize the number of predictors in training set providing the maximum correlation coefficient between posteriori probabilities and true class indicators (specifically, "0" was the indicator of G1 tumors, and "1" was the indicator of G3 tumors in the G1-G3 comparison) (Figure 2A). The smallest robust genetic grade signature contains only 6 gene probesets (A.212949_at; B.228273_at; B226936_at; A.208079_s_at; A.204825_at; A.204092_s_at) representing 5 genes: BRRN1, PRR11, C6orf173, STK6, MELK. CER was 4.4% for class G1 and 5.5% for class G3 (Figure 2.A). By PAM, for the G1-G3 comparisons, maximal prediction accuracies were obtained with 18 probesets (A.212949_at; A.221520_s_at; A.201710_at; B.228273_at;A.202768_at; B.226936_at; A.208079_s_at;B.222608_s_at;A.205046_at;A.204822_at ;A.219197_s_a;A.209189_at;A.210052_s_at; B.235572_at; A.202580_x_at;A.204825_at;B.224753_at;A.221436_s_at). Both SWV and PAM correctly classified ~96% (65/68) of the G1s and ~ 95% (52/55) of the G3s (by leave one-out method). The smaller number of probes sets required by SWV (6 probesets) compared to PAM (18 probesets) reflects an ability of SWV to use synergetic effect (co-expression patterns) during variable selection (see Methods).

Based on consistency between SWV and U-tests and PAM, we further considered the classification results using the 264 variables. In two-group comparisons high CERs were observed in the G1-G2 and G2-G3 predictions (data not shown), while in the G1-G3 CER was low (<5% errors). It suggests that G2 tumors could be not molecularly distinct from either low or high aggressive tumors.

3.2. Dichotomy of G2 tumors and disease prognosis

We next applied our grade G1-G3 predictors directly to the 126 G2 tumors of the Uppsala cohort to ask if these genetic determinants of low and high grade might resolve moderately differentiated G2 tumors into separable classes. To do that we estimated the posterior probability (Pr) as the likelihood that a sample from the exam group of tumors belongs to one class (termed "G1-like") or the other (i.e., "G3-like").

Using the 264 p.s. classifier, we found that the G2 tumors could be separated into G1-like (n=83) and G3-like (n=43) classes. We found 96% of the G2 tumors were assigned to either the G1-like or G3-like classes, indicating that almost all G2 tumors can be well separated into distinct low- and high- grade-like classes (henceforth referred to as "G2a" and "G2b" genetic grades). Only few G2 tumors exhibit intermediate Pr scores (<0.75). Using U-test and t-test, we confirmed a high separation ability of each of the SWV 264 predictors ($p<0.01$) We also evaluated a prognostic ability of our 264 p.s. classifier which we estimated using disease free survival (DFS) time data sets. Kaplan-Meier

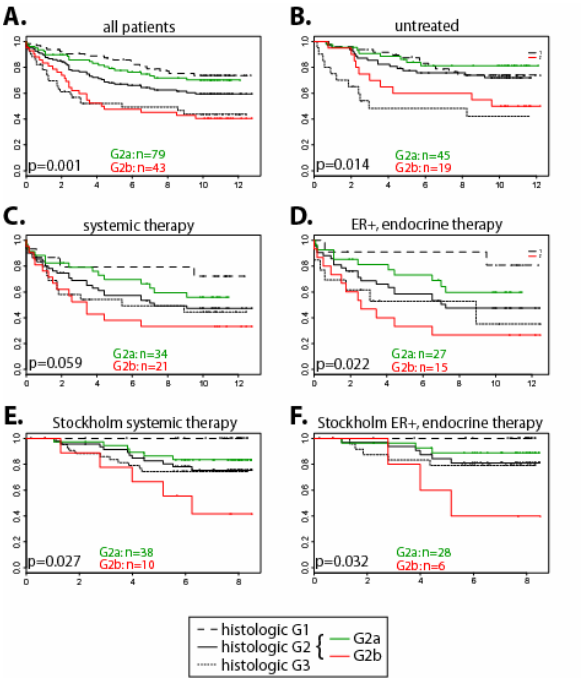


Figure 3. Survival differences between G2a and G2b genetic grade subtypes. (A) Expression profiles of the Uppsala and Stockholm tumors segregated by the SWV (5-gene) genetic grade signature are shown. Green and red vertical bars (top panel) denote histologic G1 and G3 tumors, respectively. (B-F) Kaplan-Meier survival curves for G2a (green) and G2b (red) subtypes are shown alone, or superimposed on survival curves of histologic grades 1, 2, and 3. Uppsala cohort survival curves are shown for (B) all patients, (C) patients who received no systemic therapy, and (D) patients positive for ER who received endocrine therapy only. Stockholm cohort survival curves are shown for (E) patients treated with systemic therapy and (F) those with ER positive cancer treated with endocrine therapy only. The likelihood ratio test p-value reflects the significance of the hazard ratios.

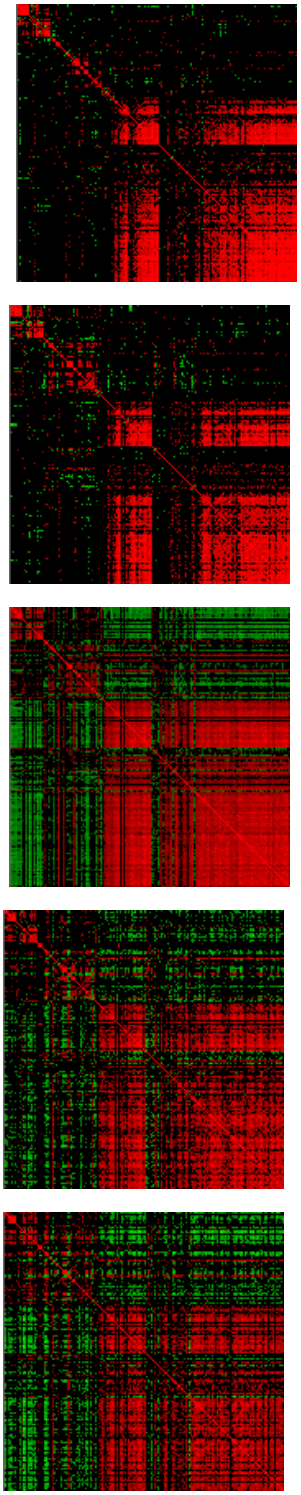


Figure 4. Correlation portraits of histological and genetic grades

Figure 2B shows that 96% of the G2 tumors (Uppsala cohort) were assigned by the classifier to either the G1-like or G3-like classes. The result was successfully verified using Stockholm and Singapore cohorts (Figure 2C,D, respectively).

survival analysis demonstrated a highly significant difference between survival curves of the G2a and G2b patients. (Cox proportional model likelihood ratio test=7.2, p=0.0071). The same classification results we obtained using the SWV 5-genetic grade classifier (Figure 3).

Survival analysis of G2a and G2b tumor subtypes based on the 5-genetic grade classifier, showed significant difference between survival curves of the G2a and G2b patients (Figure 3A). Notably, neither the G2a and G1 curves, nor the G2b and G3 curves were significantly different from each other, respectively. The G2a-G2b survival difference was further observed in specific therapeutic contexts including patients who received no systemic therapy (p=0.014; Figure 3B), with systemic therapy (Figure 3C), and those with ER positive tumors who received endocrine therapy only (p=0.022; Fig 3D). In a similar fashion, the genetic grade classifier was also predictive of recurrence in the Stockholm (G2) patients who received systemic therapy (i.e., chemotherapy, endocrine therapy or both) (p=0.027; Figure 3E) and those with ER positive disease who received only endocrine treatment (p=0.032; Figure 3F).

Affiymetrix ID	Chi-2 (G1vs G3)	Survival p value	Gene
A.212949_at	92.6	0.009	BRRN1
B.226936_at	92.6	0.004	
A.204822_at	82.2	0.002	TTK
B.228559_at	82.1	0.023	
A.218009_s_at	79.2	0.049	PRC1
A.204033_at	79.0	0.023	TRIP13
A.218726_at	75.5	0.039	DKFZ p762E1312
A.205024_s_at	73.4	0.004	RAD51
A.202870_s_at	73.0	0.051	CDC20
B.226473_at	72.6	0.004	
A.204444_at	69.3	0.034	KIF11
A.209773_s_at	67.4	0.022	RRM2
B.235572_at	67.3	0.008	Spc24
A.222077_s_at	66.5	0.020	RACGAP1
A.219990_at	66.3	0.011	FLJ23311
A.218755_at	66.3	0.016	KIF20A
A.219000_s_at	66.3	0.045	MGC5528
A.218662_s_at	66.2	0.034	HCAP-G
A.204146_at	65.3	0.027	
A.203438_at	64.0	0.032	STC2
A.209189_at	63.9	0.036	FOS
A.214039_s_at	63.3	0.017	LAPTM4B
A.205898_at	63.2	0.044	CX3CR1
A.222039_at	62.2	0.019	LOC146909
A.214710_s_at	60.8	0.021	CCNB1

Table 1. 25 (of 264) top level informative p.s. which were also significant in survival analysis (p<0.021).

3.3. The 264 p.s. provide multiple robust genetic grade signatures to discriminate G2a and G2b tumors

Due to high informatively and stability of variables of the 264-p.s. predictor, we hypothesized that there are at least several small alternative gene sub-sets (prognostic signatures) that could be used to classify low and high aggressive breast tumors with high accuracy (and therefore could provide individual classification of patients according to prognostic probability of G1 and G3).

To find such small-dimension predictors, we excluded the 6 probesets, representing the 5-genetic grade classifier, from the 264 probsets, and randomly selected two non-overlapping subsets (each of 40 probesets) from the remaining 258 probesets and applied the SWV algorithm to each selected probe subsets. In this way, we selected two additional small-dimension sets of genetic grade classifiers containing (A.221520_s_at; A.205046_at; A.211056_s_at;A.203929_s_at;B.222848_at;B.240112_at; A.221870_at)and(A.210052_s_at;A.218009_s_at;A.205794_s_at;A.203438_at;B.225191_at;A.218002_s_at;A.219197_s_at). For Uppsala, Stockholm and Singapore cohorts, each of these predictors provides similar high accuracy of classification in G1-G3 comparisons (>94% correct predictions), reproducible levels of separation of G2a and G2b subtypes for different cohorts (>94% patients assignment with G2a and G2b) and highly significant differences in G2a-G2b comparison based on survival analysis.

3.4 Comparison of performance of SWV with traditional pattern recognition algorithms

We compare performance of SWV with several traditional class prediction algorithms including Fisher Discriminant Analysis (FDA), Q-nearest neighbors (QNN), Support vector Machine (SVM). Our analysis shows that SWV can provide similar as other methods the high accuracy in leave-one-out analysis when 5-gene SWV signature or 232-gene SWV signature was used. Table 2 shows the results of such predictions base on the 5-gene SWV signature. To evaluate predictive power of the methods, after training step, we used the methods to predict G1 and G3 tumors of Stockholm and Singapore cohorts. In these two tests SWV provides better accuracy of the prediction G1 and G3 than the other methods. However, the most pronounce differences between methods we found when histologic G2 tumors from independent cohorts were tested. Table 3 demonstrates that only SWV and PAM

provide strong separation of the histologic grade 2 tumors (Uppsala cohort) on the G1-like and G3-like sets. Other methods provide diverse and poor discrimination ability of the G2 on these sets.

Method **Test 1: Uppsala DB**	**Prediction base on the 6 best SWV selected p.s.**	
	G1	**G3**
SWV	**65(95.6%)**	**52(94.5%)**
LFD	65(95.6%)	51(92.7%)
QNN	66(97.1%)	50(90.9%)
SVM	66(97.1%)	50(90.9%)
Test 2: Stockholm DB	**G1**	**G3**
SWV	**27(96.4%)**	**46(75.4%)**
LFD	27(96.4%)	40(65.6%)
QNN	27(96.4%)	45(73.8%)
SVM	27(96.4%)	41(67.2%)
Test 3 Singapore DB	**G1**	**G3**
SWV	**10(91%)**	**34(72.3%)**
LFD	10(91%)	27(57.4%)
QNN	10(91%)	29(61.7%)
SVM	10(91%)	29(61.7%)

Table 2. Evaluation of performance of pattern recognition methods.

Method	**G1-like (Pr $\leq$ 0.25)**	**True G2 (0.25< Pr <0.75)**	**G3-like (Pr $\geq$ 0.75)**
SWS (6 p.s.)	38(30%)	4(4%)	83(66%)
LFD (6 p.s.)	6(5%)	81(64%)	39(31%)
QNN (6 p.s.)	25(20%)	22(17%)	79(63%)
SVM (6 p.s.)	17(13%)	40(32%)	69(55%)
PAM (18 p.s.)	37(29%)	6(5%)	83(66%)

Table 3. Discrimination of Uppsala histologic G2 tumors base on 5-gene SWS signature using SWS, LFD, QNN, SVM and base on 17-gene signature using PAM.

3.5. Co-expression analysis of 264 gene predictors support genetic grade 2 re-classification

We found that the 264 gene predictors can be grouped base on their co-regulation patterns, which are represented on Figure 4 using Kendal tau correlation coefficient matrix for these predictors. Figure shows the images of the matrix of correlation coefficients clustered with respect values of paired correlation coefficients between probe sets into several separated groups of genes. The probes are ordered by using Gene Cluster software and then visualized using TreeView program (http:/www.lbl.gov/EisenSoftware.htm). To avoid possible bias in the images we selected randomly by 34 patents from G1, G2a, G2, G2b and G3 tumor sets. Only statistically significant correlations ($p<0.01$ after Bonferroni correction) were presented. Increasingly positive significant correlations are represented with reds of increasing intensity, and increasingly negative significant correlations are represented with greens of increasing intensity. Non-significant correlations are in black. The order of gene on all matrixes is the same. Figure 4 demonstrates the pronounced differences in expression co-regulation patterns of the genes differentially expressed in the G1, G2 and G3 groups. However, expression gene correlation matrix for G1 and G2a pair are very similar to the each other. The same phenomenon was found when we compared correlation matrixes in the pair G2a and G3 tumors. These findings support the view that low and high

	G1 vs G2a	**G2a vs G2b**	**G2b vs G3**
	p-value	p-value	p-value
	Biological process		
1	6.20E-06	5.70E-28	2.50E-06
2	1.30E-02	2.50E-02	--
3	2.70E-02	6.80E-15	1.10E-03
4	--	4.40E-03	4.90E-03
5	1.60E-02	5.50E-04	5.50E-03
6	--	3.60E-02	4.40E-02
7	--	5.00E-03	
	Molecular Function		
8	1.10E-03	7.20E-06	--
9	3.50E-03	5.00E-02	--
10	1.30E-02	--	--
11	--	7.60E-07	4.20E-04
12	--	--	7.50E-03
13	--	7.80E-04	--
14	--	1.90E-02	--
	Pathway		
15	4.90E-02	--	--
16	--	--	4.90E-02
17		3.00E-02	

Table 4. Gene ontology analysis of 264 p.s. grade classifier. Selected terms are shown with corresponding p-values that reflect significance of term enrichment. (by Panther software http://www.pantherdb.org/panther/). 1:Cell cycle; 2: Chromatin packaging and remodeling; 3: Mitosis; 4: Inhibition of apoptosis; 5: Oncogenesis; 6: Cell motility; 7: Stress response; 8: Kinase activator; 9: Histone; 10:Nucleic acid binding; 11: Microtubule family cytoskeletal protein; 12: Chemokine; 13: Non-receptor serine/threonine protein kinase; 14: Extracellular matrix linker

protein; 15: Insulin/IGF pathway-MAPKK/MAPK cascade; 16: Apoptosis signaling pathway; 17: Ubiquitin proteasome pathway. Genetic grade diseases (G1+G2a and G2b+G3) could be represented by different cancer cell precursors and Figure 4 reflects specific pathobiological pathways associated with intrinsic biological networks of these two tumor cell types.

3.6. Statistical Analysis of GO terms

A separation into the G2a and G2b is strongly supported by statistical analysis of enrichment of specific gene ontology (GO) categories of 237 RefSeq annotated gene names represented by 264 predictors in comparison to enrichment of the same GO categories in the human genome (NCBI Build 35.1). Table 4 displays a selected set of significantly enriched GO categories which includes cell cycle, inhibition of apoptosis, cell motility, stress response, kinase activators, microtubule family cytoskeletal proteins, ubiquitin proteasome pathway, suggesting essential differences in genetic programs and pathways of the G2a- and G2b-type tumor cells. Interestingly, GO comparison G1 vs G2a and G2b vs G3 also demonstrate some significant biological differences, however, these differences less different and multiple than in G2a vs G2b.

3.7. Many genetic grade features are significantly associated with cell cycle, mitosis and patient survival time

Interestingly, among patients separated by a median of DFS time in survival analysis, a large proportion probesets (58 of the 264) can significantly discriminate (at $p<0.05$) the patients on the poor and good responders (This result presents for the 25 top-level significant probesets in Table 1). GO analysis of the list of the gene assigned by these 58 probesets strongly indicates that the associated genes are essentially involved in cell cycle, mitosis including microtubule-based process, mitotic chromosome condensation, mitotic spindle organization and biogenesis. These biological processes are well-known as the essential in cancer outcome.

4. Discussion

We initially investigated several distinct class prediction/pattern recognition algorithms, including the classical Fisher discriminant analysis, *k*-nearest neighbors method, and Support Vector Machines (SVM) method. Empirically, these machine learning algorithms provide approximately similar discrimination ability on the training sets. However we found that SWV and PAM [13] had the greatest discriminative ability and were most robust regarding individual predictions when prediction rules on independent cohorts were used.

As we have shown, SWV method allows the selection of a smaller number of genes (only 5 genes representing by 6 probesets) compared to PAM (18 probesets) while the classification accuracies remain identical. SWV and PAM were used side-by-side in this study to allow a performance comparison between two robust but mathematically distinct class prediction algorithms in terms of classification accuracies and total number of genes required for maximum accuracy. PAM is a widely used statistical method for class prediction in large datasets. However, a limitation of PAM is that it is prone to over-parameterization (i.e., the selection of non-independent variables (genes) with redundant characteristics) because it does not take into account interactions between genes. The SWV method relies on a different statistical approach which involves a "voting" class prediction function using only the most informative and robust variables. SWV is strongly oriented towards the selection of a relatively small number of genes (which is more amenable to PCR-based diagnostic applications than large gene sets), even if the number of patients is limited. This is because SWV takes into account interactions between variables, thus minimizing the number of predictors needed and reducing the risk of over-parameterization. We have utilized both approaches to allow a simple performance comparison in terms of classification accuracies and total number of genes required for high-accuracy classification.

Ma et. al. (2003) were the first to report a histologic grade genetic signature capable of distinguishing low and high grade breast cancer. Using ~12K cDNA microarrays to analyze from 10 G1, 11G2, and 10 G3 microdissected tumor samples, they identified 200 genes differentially expressed between G1 and G3 tumors [3]. Using these genes for tumor clustering, they observed that the majority of G2 tumors possessed a hybrid signature intermediate to G1 and G3, with few exceptions (Figure 3 of their original report). Notably, this finding is in contrast with our discovery that the majority of G2 tumors do not display hybrid signatures, but rather possess clear G1-like or G3-like genetic features. According to our classifier, only a small percentage (~6% or less) of the tumors in our study had intermediate genetic grade measurements (ie, Pr scores <0.75 for G1-like and G3-like). To address this discrepancy, we cross-compared the 200 grade-associated genes in their list to our expanded set of 232 genes, and observed a statistically significant overlap of 35 genes ($p<1.0\text{x}10^{-7}$; Monte Carlo simulation). This overlap, however, represents only a small percentage of either gene list, indicating that the discrepant observations are most likely explained by fundamentally different signature

compositions. It is also possible that differences in sample selection and preparation, sample size (we have much larger samples and used 3 independent cohorts), RNA purification, quality of microarray analysis, and data normalization could have contributed to the variable results.

Ma et al. results inconsistent to our data (see also [12]) and Sotiriou et. al. data [14]. Sotiriou et al. published their findings of "the 97-gene expression grade index" associated with histologic grade and correlated with relapse-free survival in ER-positive breast cancer [14]. Their grade index, like our grade signatures, could dichotomize the vast majority of G2 tumors into two groups with expression profiles and survival characteristics resembling those of G1 and G3 tumors. Comparison of our gene classifiers and the 97-gene classifier we revealed that 3 of our 5-gene grade signature genes, and 68 of our larger 232-gene set, overlapped with their 97-gene index. This high degree of overlap suggests that the 232-gene set and 97-gene set may utilize the fragments of the same fundamental transcriptional programs/pathways for predicting patient outcomes. For instant, in the both studies cell cycle genes were essentially enriched in the classifiers. Whether the two predictors are collinear with respect to patient survival will be an important question moving forward. Nevertheless, our studies and [14] converge on similar findings reinforces the view that gene expression-based measurements of histologic grade can substantially contribute to patient prognosis.

We could consider the 264 probe sets and its smaller reliable subsets which we discussed in this study as genetic predictors of the G1+G2a (G1-like) and G2b (G3-like)+G3 tumor types. This finding shows that extensive molecular heterogeneity exists within the G2 tumor population, and this heterogeneity is robustly defined by the major determinants of G1 and G3 cancer. It also demonstrates that a much larger and pervasive transcriptional program underlies the genetic grade predictions of the several SWV signatures – despite its composition of the mere 5 – 17 genes. Based on SWV, PAM and multivariate analyses, a minor fraction (~6%) of grade II breast cancers is still unclassified and might be considered as the "mixture" cancers [3] or as the "technical noise".

Our findings show that genetic grade signatures could after additional biological validation improve a prognosis for patients with histologic grated II and, thus, be used in therapeutic planning for breast cancer patients.

Our results support the view that low and high grade disease (G1+G2a and G2b+G3), as re-defined genetically, reflect genetically stable independent pathobiological entities rather than a continuum of progression, which could be associated with distinct breast epithelium stem cell types [see [12] for references] .

This study demonstrates that a system approach to microarray data analysis combining with data mining, multivariate analysis, GO analysis, histopatological information from tissue samples and survival data of large patient cohorts can provide insight in molecular classification of cancers and other diseases. Such approach allows for the identification of coordinately expressed genes with essential biological and clinical associations.

References

[1]. van 't Veer, L.J., Dai, H., van de Vijver et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530-536.

[2]. Sotiriou, T., Perou, C.M., Tibshirani, R. et al. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 98:10869-10874.

[3]. Ma, X.J., Salunga, R., Tuggle, J.T., Gaudet, J. et al. 2003. Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci USA* 100:5974-5979.

[4]. Miller, L.D., Smeds, J., George, J. et al. 2005. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci USA*. 102:13550-13555.

[5] Loi, S., Sotiriou, C., Buyse, M., Rutgers, E., Van't Veer, L., , Piccart, M., Cardoso, F. 2006. Molecular forecasting of breast cancer: Time to move forward with clinical testing. *J Clin Oncol* 24: 721-722.

[6] Ransohoff D.F. 2004.Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* 4:309-314.

[7] Brenton, J.D., Carey, L.A., Ahmed, A.A., Caldas C. 2005. Molecular classification and molecular forecasting of breast cancer: ready for clinical application? *J Clin Oncol*. 23:7350-60.

[8]. Kuznetsov, V.A., Ivshina, A.V., Sen'ko, O.V., Kuznetsova, A.V. 1996.Syndrome approach for computer recognition of fuzzy systems and its application to immunological diagnostics and prognosis of human cancer. *Math. Comput. Modeling* 23:92-112.

[9]. Kuznetsov, V.A., Knott, G.D., Ivshina, A.V. 1998. Artificial immune system based on syndromes-response approach: Theory and their application to recognition of the patterns of immune response and prognosis of therapy outcome. In *Proc. of IEEE Intern. Conf. on Systems, Man, and Cybernetics*. San Diego, CA, USA. 3804-3809.

[10]. Jackson, A.M., Ivshina, A.V., Senko, O., Kuznetsova, A., Sundan, A., O'Donnell, M.A., Clinton, S., Alexandroff, A.B., Selby, P.J., James, K., Kuznetsov, V.A. 1998. Prognosis of intravesical bacillus Calmette-Guerin therapy for superficial bladder cancer by immunological urinary measurements: statistically weighted syndromes analysis. *J Urol* 159:1054-1063.

[11]. Mueller, B.U., Zeichner, S.L., Kuznetsov, V.A., Heath-Chiozzi, M., Pizzo P.A., and Dimitrov, D.S. 1998. Individual prognoses of long-term responses to antiretroviral treatment based on virological, immunological and pharmacological parameters measured during the first week under therapy. *AIDS*, 13: f191-f196.

[12] Ivshina, A.V., George, J., Senko, O.V., Mow, B., Putti, T.C., Smeds, J., Nordgren, H., Bergh, J., Liu, E. T-B., Kuznetsov, V.A., Miller, L.D. 2006. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.*, 66: 10292-10301.

[13] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA*, 99:6567-6572.

[14] Sotiriou, C., Wirapati, P., Loi, S. et al. 2006. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst, 98:262-272.

Acknowledgments

Grant support: Singapore Agency for Science Technology and Research.

Vladimir A. Kuznetsov received the Ph.D in biophysics in 1984 from Moscow State University and Dr. Sci. in math.-physics in 1992 from Science & Technical Union of the Russian Academy of Sciences, respectively. During 1982-1998 he was researcher, senior researcher and head of the laboratory of Mathematical Immunobiophysics at the Institute of Chemical Physics of Russian Academy of Sciences, Moscow, to study mathematical models of immune system and tumor-immune system interaction network. He also developed computational data mining tools, pattern recognition methods and its applications in clinical trials. During 1996-2004, he worked at FDA USA, and NCI/NIH, NICHD/NIH (Bethesda, MD, USA) as a senior researcher. He was involved in NIH Cancer Anatomy Genome Project and other genomics projects focusing on study of cancer and infectious diseases. From 2004 he is a senior group leader and head of the Laboratory Computational Genomics and Systems Biology at Genome Institute of Singapore.

Oleg Senko received the PhD in mathematics in 1990 from Computer Center of USSR Academy of Sciences. His interest is in development of novel statistical and combinatorial methods of pattern recognition, forecasting, data mining and its applications. He is a senior researcher at Dorodnicyn Computing Center of the Russian Academy of Science, Moscow, Russia.

Lance D. Miller received the PhD in 2001 in Genetics and Molecular Biology from University North Carolina at Chapel Hill Chapel Hill, NC, USA.
From 2001 at present a senior group leader and head of the Laboratory of Microarrays and Expression Genomics at Genome Institute of Singapore. His primary interest is the application of genomic technologies towards solving problems in human disease. His current research is focused primarily on the molecular characterization of human breast cancer and the microarray-based detection and characterization of human pathogens.

Anna Ivshina received the MD in 1979 from the Fist Moscow Medical Institute, USSR and PhD in 1986 from All Union Cancer Center of Medical Academy of USSR. During 1979-1996 she was a researcher at Clinical Institute of Russian Cancer Center of Medical Academy. At the Laboratory of Clinical Immunology of this institute she studied immune response against cancer and developed clinical immunodiagnostic methods. For several years she worked as a scientist at FDA USA and she developed methods of diagnostics and control of infectious diseases. She is now a scientist of Laboratory of Microarrays and Expression Genomics at Genome Institute of Singapore; research interest focuses on microarray expression data analysis and clinical risk factors aiming to reveal novel genetic markers for reliable classification of tumor sub-types and prediction of clinical outcome of cancer diseases.