Supervised classification of combined copy number and gene expression data

S. Riccadonna[1,2], G. Jurman[1], S. Merler[1], S. Paoli[1,2], A. Quattrone[3], C. Furlanello[*1]

[1]FBK-irst, via Sommarive 18, I-38100 Povo (Trento), Italy
http://mpa.itc.it

[2]DIT, University of Trento, via Sommarive 14, I-38100 Povo (Trento), Italy
http://dit.unitn.it

[3]CIBIO and DISI, University of Trento, via Sommarive 14, I-38100 Povo (Trento), Italy
http://www.unitn.it

Summary

In this paper we apply a predictive profiling method to genome copy number aberrations (CNA) in combination with gene expression and clinical data to identify molecular patterns of cancer pathophysiology. Predictive models and optimal feature lists for the platforms are developed by a complete validation SVM-based machine learning system. Ranked list of genome CNA sites (assessed by comparative genomic hybridization arrays – aCGH) and of differentially expressed genes (assessed by microarray profiling with Affy HG-U133A chips) are computed and combined on a breast cancer dataset for the discrimination of Luminal/ER+ (Lum/ER+) and Basal-like/ER- classes. Different encodings are developed and applied to the CNA data, and predictive variable selection is discussed. We analyze the combination of profiling information between the platforms, also considering the pathophysiological data. A specific subset of patients is identified that has a different response to classification by chromosomal gains and losses and by differentially expressed genes, corroborating the idea that genomic CNA can represent an independent source for tumor classification.

1 Introduction

Both transcriptome profiling, by gene expression microarray, and genomic copy number aberrations (CNA) detection, by comparative genomic hybridization arrays (aCGH), have been used to produce molecular portraits of breast cancer specimens. These are used to derive signatures of prognostic value for patients by means of unsupervised hierarchical clustering. Microarray-derived data allowed the identification of five breast cancer subtypes (basal-like, luminal A, luminal B, ERBB2, normal breast-like) two of which (basal-like and ERBB2) have been associated with strongly reduced survival [14, 15, 18, 19]. aCGH also consistently detected characteristic CNA in breast cancer, which allows the classification of tumor samples on the basis of their pattern of chromosomal gains and losses [10, 11, 16]. Based on unsupervised clustering on aCGH data, the breast tumor genomes fall into one of three categories (called 1q/16q or simple, complex and amplifier) [5]. The emergence of features distinguishing breast cancer subgroups

*Corresponding author

on the basis of either their genomic and transcriptomic blueprints raises the possibility combining them to produce compound signatures potentially endowed with extended predictive power. A recent extensive study [4] explored the value of the combination of microarray-derived and aCGH-derived data obtained from the same collection of tumor samples, coming from patients undergoing aggressive adjuvant chemotherapy and endowed with adequate clinical description and follow-up. The conclusion was that the classes obtained by expression profiling have different recurrent CNA, and that the parallel use of these data can improve patient stratification according to the outcome. These results suggest that, for breast cancer, integration of genomic and transcriptomic abnormalities could provide a potential enrichment in predictive power, justifying further attempts to unravel the structure of the relationship between these two levels of observation of cancer genomic instability.

We introduce in this context a set of machine learning methods to investigate common and different structures within a possible integrative space of high-throughput features. In terms of machine learning, our main task is the predictive classification and feature ranking of gene expression and genome copy number with respect to the Luminal/ER+ and Basal/ER- classes. We first develop predictive models and optimal feature lists for the two platforms separately. Different encodings are applied to the CNA data. Then we filter data and variables and consider the combination of profiling information between the platforms, also considering the pathophysiological data. Finally, we analyze in detail a specific subset of patients in which differences are found by BAC aCGH and gene expression. We show that SVM-based classification of microarray-derived and aCGH-derived data on a common subset of 63 samples detects a subset of 8 cases with specific pathophysiological characteristics.

To our knowledge, this is one of the first studies applying supervised classification in a complete validation context to aCGH data and combining them with gene expression data with the same machine learning methodology. The structure of the rest of the paper is as follows: the dataset and preprocessing methods are discussed in Sec. 2, and Sec. 3 details the machine learning framework, whose application is described in Sec. 4. Results are finally discussed in Sec. 5.

2 Data description

This paper is based on the aCGH and gene expression data first presented in [4]. In this section we summarize the main information on the original data and provide details on different preprocessing methods we applied for supervised classification tasks.

The array-CGH data were obtained from the Scanning and OncoBAC methodologies as described in [4]. The data consists of 2149 BAC describing 149 samples. This dataset includes missing values (NAs). In this study, we consider the data either as BAC or by different encodings and we impute the missing values through kNN as in [20], we eliminate samples and/or genes with too many NAs, or we obtain imputation directly from another type of preprocessing which encodes BAC into segments. The segmented data are obtained by processing the original aCGH data as in [12, 21]: circular binary segmentation is applied to all samples to obtain, for each chromosome of each sample, a piecewise constant function with the `DNAcopy` R/Bioconductor package. By intersecting all those values (Fig. S1 in Supplementary Material), we build a sparse matrix of BioDCV inputs. Note that NA imputation is automatically performed within the segmentation algorithm.

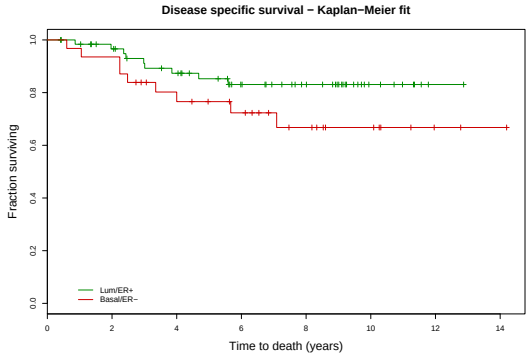


Figure 1: Kaplan-Meyer disease-specific survival stratified by Lum/ER+ vs. Basal/ER- classes. The graph includes 94 samples (63 Lum/ER+, 31 Basal/ER-) in [4] having complete ER status and survival information.

Expression data for 118 samples described by 22215 probes were collected by using Affymetrix HG-U133A arrays and preprocessed as detailed in [4]. Tumor subtype classes can be assigned on these data by clustering with the 70-genes molecular signature from [19]. The samples were labelled as belonging to one out of the five classes (basal-like, ERBB2, luminal A, luminal B and normal-like) according to proximity to the molecular-signature cluster computed both by the Euclidean distance and by Pearson's correlation (in case of ambiguity we consider the latter) [4]. Although an alternative classification in subtypes of biological relevance have been recently proposed [9], in this study we follow the original five class partition.

Clinical data are available and consist of 136 fields for 174 patients. In particular, the ER status and the disease-specific survival information are provided. The number of samples that are available for integrative genomic study (i.e. with both Affy U133A and aCGH, as well as the clinical information) is 89. Their partition into subtypes, according to Pearson's correlation, is the following: 35 luminal A , 11 luminal B, 23 basal-like, 10 ERBB2, and 10 normal-like. From all the 89 samples we extracted the 42 samples belonging to the luminal subtype (aggregation of luminal A and luminal B) with positive ER status (discarding 4 luminal samples with negative ER status) and the 23 samples of the basal-like subtype (all with negative ER status), for a total of 65 cases. The Kaplan-Meyer disease-specific survival graph for the two classes luminal/ER+ (Lum/ER+) and basal/ER- (94 samples, having complete ER status and disease specific survival information) is shown in Fig. 1. The resulting p-value is 0.131.

3 Methods

Our approach towards supervised classification relies on a set of algorithms aimed at achieving predictive classification together with stable molecular profiles, avoiding overfitting risks due to selection bias effects [1]. This method has been used so far with satisfying results in different functional genomics and proteomics tasks (e.g. profiling microarrays, time series microarray, integration with clinical data, and mass spectrometry data) and also by using grid computing infrastructures [2, 3, 6, 7, 8, 13].

The method's core is a complete validation procedure [8] realizing repeated training/test splits of the original dataset. A feature ranking algorithm is applied to the training portion, and classification models with increasing number of best ranked features are computed on the test part.

Accuracy performance is assessed by averaging classification errors on all the test splits (ATE, Average Test Error) both globally over all samples and on a samplewise basis (sample-tracking analysis). Given the population of ranked lists produced at each training/test split, a global ranked list is then computed by means of an aggregating algorithm (Borda count method). As a rule of thumb, the more the splits, the smoother the obtained results. Class label randomization is also applied to validate the entire method. The whole procedure is implemented in BioDCV, a Python/C environment freely available at `http://biodcv.itc.it`.

In this paper, Support Vector Machines (SVM) are used as the classification algorithm. They are well-known both in machine learning as well as in bioinformatics literature for their good performance on high-throughput data, characterized by the huge unbalance $p >> n$ between the number n of samples and the number p of features [17]. Regularization parameter and other kernel-specific parameters (such as bandwith for Gaussian kernels) are preliminarly tuned by training error minimization in bootstrap experiments. Finally, each feature in the data matrix is usually standardized across samples to zero mean and one standard deviation to avoid unwanted effects due to strong dishomogeneities in the features' ranges.

Entropy-based Recursive Feature Elimination (E–RFE) is used as the ranking algorithm [7]. RFE-like algorithms are wrapping methods that recursively discard a bunch of the features which are less contributing to the classifier until all features have been eliminated. Different choices in the discarding rule lead to variants of the original RFE algorithm, which eliminates just the least important feature at each loop (stepwise backward selection). The E–RFE algorithm has been shown to achieve performances comparable with the original RFE with a consistent improvement in terms of computing time. Nevertheless, performing the complete validation procedure for a large number of splits (e.g. 400 as in the experiments discussed in this paper) is still computationally heavy: reasonable computing times can thus be obtained by distributing loads on High Performance facilities such as clusters or grid infrastructures [3].

4 Results

In this study, we initially performed a 10-NN imputation of missing values on the aCGH data. BAC with more than 10% missing values ($\geq$ 7 cases) across samples were discarded, leaving a total of 1590 BAC features. We also performed a circular binary segmentation on the same 65 samples, obtaining 1674 features. The BioDCV-based predictive profiling was applied to these datasets, following the steps detailed in Methods. The experiment on the gene expression data (BioDCV, linear SVM, E–RFE) reaches near perfect classification with very few genes (ATE $<$ 1% with 3 genes, on average). Results are summarized in Tab. 1 (see also Fig. S2 and S3): predictive average test errors (ATE) and their standard deviations are listed for increasing feature set sizes, globally and separately for the classes. This result, which extends the original study [4], is however predictable since tumor subtype labeling was set by clustering of the gene expression data and these two tumor subtypes are considered as well separated also in known breast cancer signatures [19]. In particular, the most discriminating feature, for all the 400 test sets, is the probe 205225_at (ESR1).

Classification of the aCGH data as BAC (also with BioDCV, linear SVM and E–RFE) gives an ATE $<$ 12% for all models with less than 300 features, reaching 9.8% as the minimum value with 900 features. Two samples (s0004 and s0138) are however consistently misclassified in all

tests according to the BioDCV procedure. Note that the subtype class of s0004 is not univocally assigned by the Sorlie signature: the label is basal-like by Pearson correlation and ERRB2 by Euclidean distance.

We thus applied the shaving procedure detailed in [13] and removed the two samples from both the BAC and the expression data. The shaving lowers the no-information rate from the original $\frac{23}{23+42} = 35.4\%$ to $\frac{21}{21+42} = 33.3\%$. The classification exercise was then repeated on both the datasets. For the gene expression data, the shaving procedure did not significantly affect the performances of the experiment (ATE $<$1% with four features). As above, the best discriminating gene was ESR1, typically the gene more directly associated with the ER status, introducing a first order effect that masks other potential descriptors of the underlying pathophysiology. We thus removed the ESR1 gene from the feature set and repeated the classification. We did not remove other ER-related probes that are poorly discriminating on this dataset. For gene expression, performances were only slightly affected (ATE $<$2% with 3 features and ATE $<$1% with 15 features), confirming the effect of other genes in this classification task. Moreover the ordering of features in the ranked list showed only minor rearrangements (discussed below). In summary, classification by gene expression is not modified by the removal of the two samples (problematic for the task on BAC) and the ESR1 probe.

Classification of the aCGH BAC data is instead improved in this setting. As shown in Tab. 2 and displayed in Fig. 2, now ATE is less than 10% for models with at least 15 features, with a minimum ATE 6.9% reached with 50 BAC. The shaving procedure has thus effectively reduced noise from the analysis. Note that performance with Gaussian kernel does not improve with respect to the simpler linear SVM model, also after parameter tuning by bootstrap-based procedures, while the resulting optimal lists are similar. We then analyzed the problem of combining genomic and trascriptomic information as derived from profiling. Alignment of BAC and corresponding HG-U133A probes for the the best features is detailed in Tab. 3 (from probes to BAC) and Tab.4 (from BAC to probes). The correspondances were seeked by using GenomeMap (NCBI: `http://www.ncbi.nlm.nih.gov/mapview`), NetAffix (`https://www.affymetrix.com`), UCSC Genome Browser (`http://genome.ucsc.edu`). Given the first 15 probes (ATE $<$ 0.9% on the shaved dataset without ESR1), 11 BAC were mapped (3 corresponding to one single probe). However, we were able to map only 2 BAC of the 1590 conserved after imputation. In the other direction (Tab. 4), when we considered the first 15 ranked BAC, 22 Affy probes were found for 5 different BAC, of which 11 for the HG-U133A platform used in [4]. The differential expressions for these 11 probes are shown

# feat	$M_{\text{Lum/ER+}}$	$SD_{\text{Lum/ER+}}$	$M_{\text{Basal/ER-}}$	$SD_{\text{Basal/ER-}}$	M	SD
1	2.08	4.69	3.05	9.79	2.42	6.50
2	0.53	2.74	2.15	6.96	1.10	4.23
3	0.18	1.80	1.50	5.64	0.64	3.16
4	0.13	1.11	1.30	5.33	0.54	2.60
5	0.13	1.11	1.15	4.87	0.49	2.44
10	0.03	0.50	1.50	5.82	0.55	2.38
25	0.00	0.00	1.40	5.30	0.50	1.88
50	0.00	0.00	1.15	5.27	0.41	1.86
100	0.00	0.00	0.75	4.53	0.27	1.60
500	0.00	0.00	0.40	2.80	0.14	0.99
1000	0.00	0.00	0.10	1.41	0.04	0.50
22215	0.00	0.00	0.15	1.73	0.05	0.61

Table 1: Average Test Error (M) on the gene expression data for the Lum/ER+ vs. Basal/ER- task (global and classwise), with standard deviation (SD).

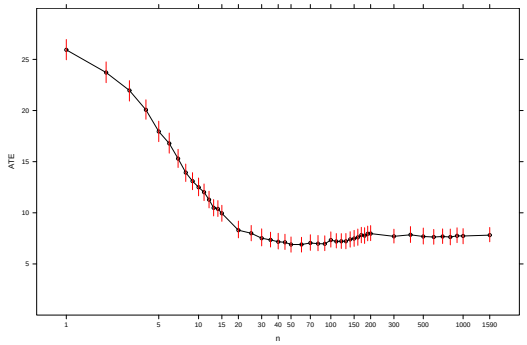


Figure 2: Average Test Error on the copy number data (BAC) for the Lum/ER+ and Basal/ER- task, with 95% student's bootstrap confidence intervals.

in Fig. 3: best separation is shown for probe 201805_at, which is the best ranked for gene expression classifiers. It is worth noting that in both cases the mappings found low ranked features, suggesting independence between the optimal lists produced by the two experiments on the different datasets for the admissible mappings with current information. Note that greater insights may be gained by using higher resolution arrays such as SNP.

As an alternative way of comparing the best ranked lists from the two platforms, we first explored a binning technique discussed in [4], in which every chromosome was subdivided into non-overlapping sections (bins) of length 20Mb. Locations of top-20 probesets and BAC are listed in Tab. 5. Only 5 common bins were detected, supporting a possible independence between the feature sets. None of the top 20 BAC and genes lie on chromosomes 7, 8, 9, 11, 17. As an intermediate feature encoding between BAC and coarse binning, we performed the predictive profiling study on the aCGH data preprocessed by circular binary segmentation (see Methods). Classification with the segmented data gave poor performances: as shown in Tab. S5 and Fig. S6, ATE always remains above 13%, and above 20% with less than 70 segments used as features in the SVM classifiers. In order to enhance possible effects due to CNA high-amplification, the classification exercise was repeated with both BAC and segments without standardizing data, after imputation. This variant worsened the performance with BAC (minimum ATE at 7.8% for 90 features), while a negligible higher accuracy (minimum ATE 12.8% at 600 features and 19.4% at 70 features) was obtained for the segmented data. In summary, seg-

# feat	$M_{\text{Lum/ER+}}$	$SD_{\text{Lum/ER+}}$	$M_{\text{Basal/ER-}}$	$SD_{\text{Basal/ER-}}$	M	SD
1	13.75	12.32	50.30	23.61	25.93	16.08
2	15.23	11.48	40.70	22.33	23.72	15.10
3	14.63	12.24	36.65	21.28	21.97	15.25
4	12.78	11.57	34.65	21.94	20.07	15.03
5	11.25	11.08	31.35	21.85	17.95	14.67
10	8.33	8.49	20.85	20.87	12.50	12.62
15	6.60	7.69	16.65	18.94	9.95	11.44
20	5.78	7.11	13.35	17.65	8.30	10.62
25	5.95	7.47	12.10	17.33	8.00	10.75
50	5.58	7.20	9.55	15.70	6.90	10.03
100	5.73	7.49	10.55	16.56	7.33	10.52
500	4.63	6.67	13.80	17.39	7.68	10.24
1000	4.73	6.64	13.75	18.02	7.73	10.43
1590	4.80	6.64	13.85	17.78	7.82	10.35

Table 2: Predictive errors (M) with standard deviation (SD) on the copy number data (BAC) for the Lum/ER+ and Basal/ER- task (separately on the two classes and for all data), after shaving two samples.

# [ER]	# [no ER]	Probeset ID	Location	Gene symbol	Mapped BAC	BAC cytoband (FISH)	BAC ranking
1	(shaved)	205225_at	6q25.1	ESR1	(RP1-63IE)	6q25.1-6q26	
2	2	204623_at	21q22.3	TFF3			
3	6	219497_s_at	2p16.1	BCL11A			
4	11	215867_x_at	16q23	AP1G1			
5	(18)	221880_s_at	15q26.1	LOC400451			
6	1	214164_x_at	16q23	AP1G1			
7	14	212692_s_at	4q31.3	LRBA	(RP11-29P18)	4q32-4q33	
8	4	203963_at	15q22.2	CA12	RP11-100N8	15q21.3	97
8	4	203963_at	15q22.2	CA12	(RP11-209D15)	15q22	
8	4	203963_at	15q22.2	CA12	(RP11-91E13)	15q22	
9	9	214404_x_at	6p21.31	SPDEF	(RP11-375E1)	6q21.31	
10	15	209623_at	5q13.2	MCCC2	(RP11-88J2)	5q13.2	
11	12	204667_at	14q21.1	FOXA1			
12	10	209871_s_at	15q13.1	APBA2	(RP11-165M18)	15q23.1-15q23.33	
13	5	204198_s_at	1p36.11	RUNX3	(RP3-398I9)	1p34.3-36.13	
14	(17)	203929_s_at	17q21.31	MAPT	(RP11-669E14)		
15	3	212956_at	4q31.21	TBC1D9	(RP11-5K16)	4q31.1	
(27)	7	214053_at	2q33.3-q34	ERBB4	(RP11-300D24)	2q34	
(27)	7	214053_at	2q33.3-q34	ERBB4	CTD-2067J6	2q34	619
(27)	7	214053_at	2q33.3-q34	ERBB4	(CTD-2204E9)	2q34	
(25)	8	212442_s_at	2q24.3	LASS6			
(21)	13	206373_at	3q24	ZIC1			

Table 3: Mapping of the top-15 ranked probesets on the available BAC. Probeset ranking position for gene expression classification is listed for ESR1 probeset included (col. *# [ER]*) or excluded *# [noER]*). The BAC clones in parentheses are not present (missing or discarded by imputation).

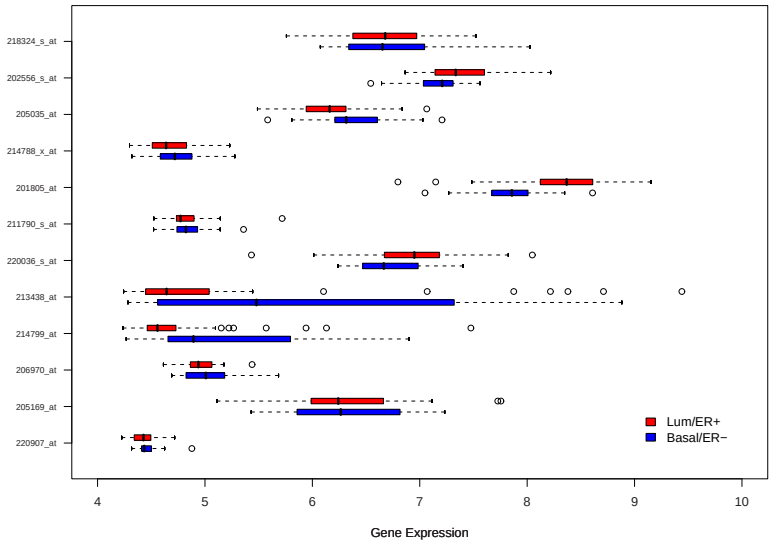


Figure 3: Expression level of the affy probesets mapped from top-ranked BAC.

mentation and non-standardized data gave worse results than standardized BAC on this dataset. Because differences were found by comparing and mapping the profiling results on the two different platforms, we consider the individual ATE error curves computed for each sample (i.e. sample-tracking curves, see Methods and [8]). The sample-tracking curves are displayed in Fig. 4. As shown in the figure, eight aCGH samples (4 Lum/ER+ and 4 Basal/ER-) are mostly misclassified. As they are altogether perfectly classified by gene expression, it is suggestive to analyze this subset in some detail. Some of their clinical features (SBR grade and survival time) are differently distributed w.r.t. to the rest of samples: SBRgrade=3 for 6/8 cases vs SBR-grade=3 for 22/55; median recurrence 3.3 years vs 6.9 years, suggesting a worse prognosis for

#	BAC	Cytoband (FISH-mapped)	Mapped Gene Symbol	Associated Affy Probeset	Gene Ranking
1	RP11-23B22	1			
2	RP11-277N18	5q21	none		
3	RP11-25K5	12p12-13	SPATS2	218324_s_at	9184
3	RP11-25K5	12p12-13	SPATS2	(222593_s_at)	
3	RP11-25K5	12p12-13	SPATS2	(222594_s_at)	
3	RP11-25K5	12p12-13	KCNH3	(223726_at)	
3	RP11-25K5	12p12-13	MCRS1	202556_s_at	1762
3	RP11-25K5	12p12-13	C12orf25	(224039_at)	
4	RP11-176I4	4p15.1		no probe	
5	CTD-2048F17	18q23	FLJ25715	no probe	
5	CTD-2048F17	18q23	CTDP1	205035_at	8110
6	CTD-2271B13	4			
7	RP11-40N8	5q13.1-5q12	none		
8	RP11-253O10	16q23	LOC645799	no probe	
9	CTD-2008N7	12q13	DDN	214788_x_at	21293
9	CTD-2008N7	12q13	PRKAG1	201805_at	252
9	CTD-2008N7	12q13	MLL2	211790_s_at	21071
9	CTD-2008N7	12q13	RHEBL1	no probe	
9	CTD-2008N7	12q13	DHH	no probe	
9	CTD-2008N7	12q13	STX6 LMBR1L	220036_s_at	14078
9	CTD-2008N7	12q13	MLL2	(227527_at)	
9	CTD-2008N7	12q13	MLL2	(227528_s_at)	
9	CTD-2008N7	12q13	MLL2	(231974_at)	
10	RP11-109L8	12q15	HELB	no probe	
11	RP11-243M13	1q32.1	NFASC	213438_at	8856
11	RP11-243M13	1q32.1	NFASC	214799_at	8484
11	RP11-243M13	1q32.1	NFASC	(230242_at)	
11	RP11-243M13	1q32.1	CNTN2	206970_at	6392
11	RP11-243M13	1q32.1	TMTM81 (LOC388730)	no probe	
11	RP11-243M13	1q32.1	CNTN2	(230045_at)	
12	RMC05P007	5			
13	RP11-39C2	6p12	GPR110	220907_at	13553
13	RP11-39C2	6p12	GPR110	(235988_at)	
13	RP11-39C2	6p12	GPR110	(238689_at)	
14	RP11-22L20	5q11.2	none		
15	RP11-141N1	12q21.3-22	none		

Table 4: Top-15 ranked BAC with the included genes. The Affy probesets in parentheses belongs to the Affymetrix U133B platform, and thus they are not present in the Chin06 dataset [4].

Probe Set ID	chrom.	portion	BAC	chrom.	portion
204198_s_at	1	2	RP11-219O7	1	2
219497_s_at	2	4	RP11-243M13	1	11
212442_s_at	2	9	RP11-176I4	4	2
214053_at	2	11	CTD-2271B13	4	3
206373_at	3	8	RP11-22L20	5	3
202341_s_at	4	8	RP11-40N8	5	4
212692_s_at	4	8	RP11-277N18	5	6
212956_at	4	8	RMC05P007	5	7
209623_at	5	4	CTD-2118F18	6	2
214404_x_at	6	2	RP11-39C2	6	3
201984_s_at	7	3	RP11-72C6	10	1
212771_at	10	1	RP11-70B16	10	2
204667_at	14	2	CTD-2008N7	12	3
209871_s_at	15	2	RP11-25K5	12	3
203963_at	15	4	RP11-109L8	12	4
221880_s_at	15	5	RP11-141N1	12	5
214164_x_at	16	4	RP11-106L3	16	3
215867_x_at	16	4	RP11-253O10	16	4
203929_s_at	17	3	CTD-2048F17	18	1
204623_at	21	3	RP11-23B22	20	2

Table 5: Chromosome portions of the top-20 affy probesets and top-20 BAC.

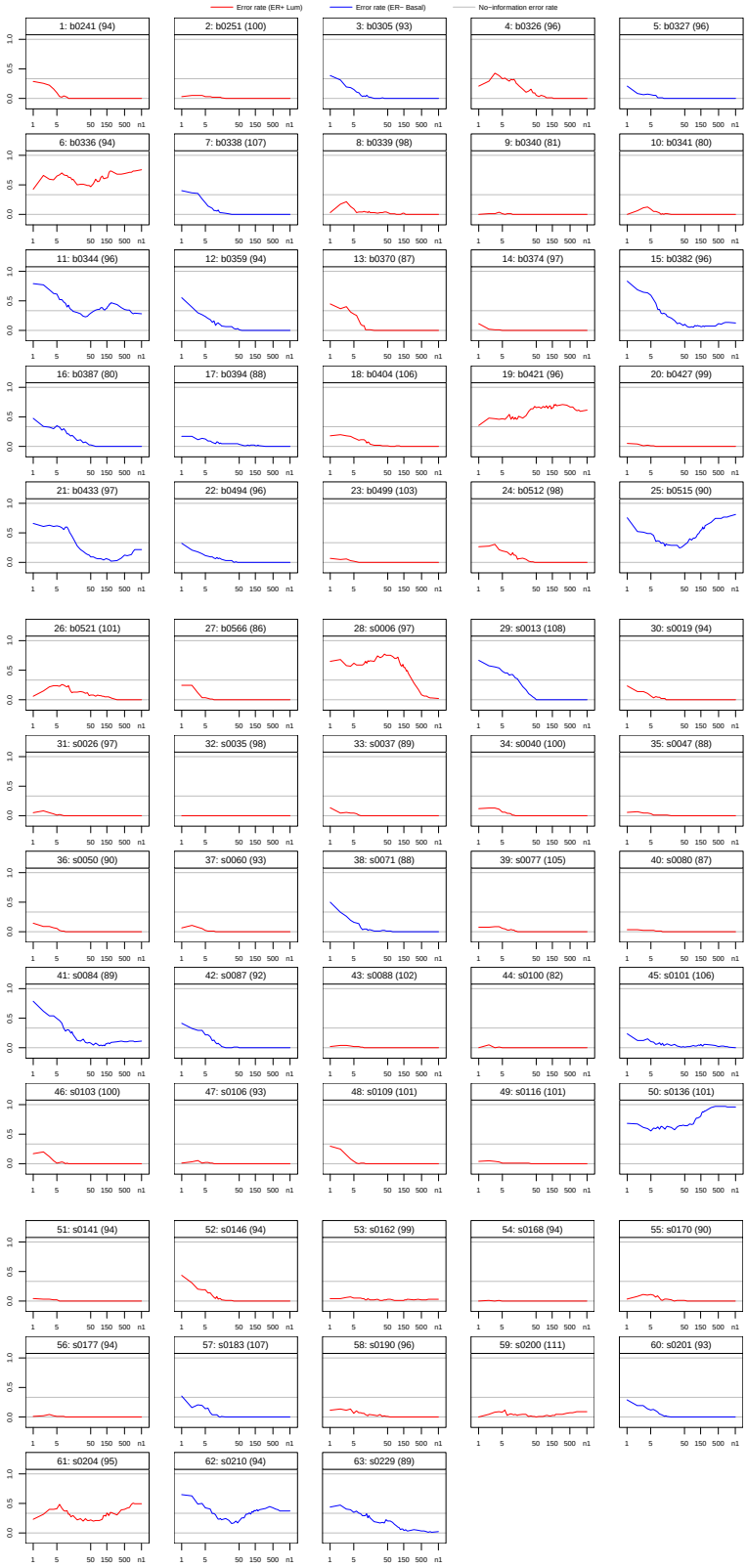


Figure 4: Sample-tracking error analysis of Lum/ER+ and Basal/ER- task on the copy number data (BAC features). For each sample, the plot indicates percent error for BioDCV runs (indicated in parenthes) in which the sample is in test, averaged on models of the same feature set size. The horizontal grey line indicates the no-information error rate for baseline comparison.

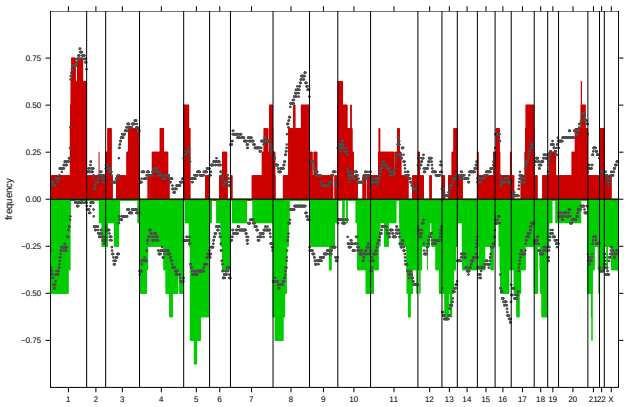


Figure 5: Gains (upper red) and losses (lower green) frequencies of the eight misclassified samples are compared to the frequencies for the remaining 55 samples (gray dots).

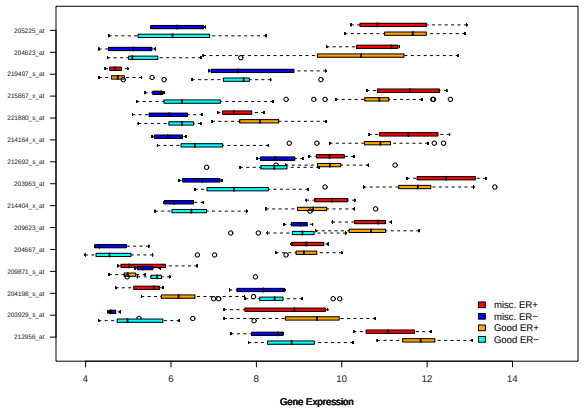


Figure 6: Comparison of gene expression on optimal gene expression signature of the aCGH misclassified samples with the remaining data. All data are perfectly classified by the selected Affy probes. However, probes such as 212692_s_at discriminate within the subset better than within the rest of the data.

those eight samples. By comparing the frequency of gains and losses (as in [4]) of those eight samples vs. the remaining 55 samples, significant differences emerge for many chromosomes (see Fig. 5). Further differences can also be dectected by the box plots in Fig. 6 and Fig. S6.

5 Conclusions

In this paper we propose a combined approach to supervised classification of breast cancer specimens based on genomic lesion detection from aCGH and transcriptome analysis by microarray profiling. The CNA signature we derived discriminates the Lum/ER+ and Basal/ER- subgroups with a 12.5% predictive error rate (with 10 BAC features); this suggests that aCGH data could be used in classification tasks where transcriptome profiling data are unavailable or cannot be obtained for the high degree of sample degradation. Moreover, the absence of co-occurence between the genes located in the top ranked BAC and the location of the genes detected by the top ranked Affy probesets suggests a segregation of predictive power at the genome and the transcriptome level in these samples, further corroborating the idea that genomic CNA can

represent an independent source for tumor classification. The 8 samples misclassified by the supervised aCGH-based analysis are characterized by a common overall worse prognosis and by statistically significant differences in the CNA profile, with substantial changes in the number of gains and losses in the 5, 6, 7, 10, 12 and 14 chromosomes. Therefore, this tumor subset could be endowed with other features emerging only on the basis of the CNA profile, and having no effect on the discriminating ability of the transcriptome profiles. This finding again reinforces the interest in aCGH-derived information for breast tumor supervised classification.

Acknowledgments Research partially funded by the AIRC grants "IFOM-Breast Cancer Complexity Reduction by Molecular Subtyping associated with Clinical Outcome" and "IFOM-Breast Cancer Complexity Reduction by Molecular Subtyping associated with Clinical Outcome".

References

[1] C. Ambroise and G.J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA*, 99(10):6562–6566, 2002.

[2] A. Barla, B. Irler, S. Merler, G. Jurman, S. Paoli, and C. Furlanello. Proteome profiling without selection bias. In *Proc. CBMS 2006*, pages 941–946. IEEE, 2006.

[3] M. Cannataro, A. Barla, R. Flor, G. Jurman, S. Merler, G. Paoli, S. Tradigo, P. Veltri, and C. Furlanello. A grid environment for high-throughput proteomics. *IEEE Trans. Nanobiosciences*, 6(2):117–123, 2007.

[4] K. Chin, S. DeVries, J. Fridlyand, P. T. Spellman, R. Roydasgupta, W. L. Kuo, A. Lapuk, R. M. Neve, Z. Qian, T. Ryder, F. Chen, H. Feiler, T. Tokuyasu, C. Kingsley, S. Dairkee, Z. Meng, K. Chew, D. Pinkel, A. Jain, B. M. Ljung, L. Esserman, D. G. Albertson, F. M. Waldman, and J. W. Gray. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541, 2006.

[5] J. Fridlyand, A.M. Snijders, B. Ylstra, H. Li, A. Olshen, R. Segraves, S. Dairkee, T. Tokuyasu, B.M. Ljung, and A.N. et al. Jain. Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer*, 6:96, 2006.

[6] C. Furlanello, S. Merler, and G. Jurman. Combining feature selection and DTW for time-varying functional genomics. *IEEE Trans. on Signal Processing*, 54(6):2436–2443, 2006.

[7] C. Furlanello, M. Serafini, S. Merler, and G. Jurman. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics*, 4:54, 2003.

[8] C. Furlanello, M. Serafini, S. Merler, and G. Jurman. Semisupervised learning for molecular profiling. *IEEE-ACM Trans. Comp. Biology and Bioinf.*, 2(2):110–118, 2005.

[9] A.V. Kapp, S.S. Jeffrey, A. Langerød, A.L. Børresen-Dale, W. Han, D.Y. Noh, I.R.K. Bukholm, M. Nicolau, P.O. Brown, and R. Tibshirani. Discovery and validation of breast cancer subtypes. *BMC Genomics*, 7:231, 2006.

[10] L.W. Loo, D.I. Grove, E.M. Williams, C.L. Neal, L.A. Cousens, E.L. Schubert, I.N. Holcomb, H.F. Massa, J. Glogovac, and C.I. et al. Li. Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. *Cancer Res.*, 64(8541–8549), 2004.

[11] T.L. Naylor, J. Greshock, Y. Wang, T. Colligon, Q.C. Yu, V. Clemmer, T.Z. Zaks, , and B.L. Weber. High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. *Breast Cancer Res.*, 7:R1186–R1198, 2005.

[12] A.B. Ohlsen, E.S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5:557–572, 2004.

[13] S. Paoli, G. Jurman, D. Albanese, S. Merler, and C. Furlanello. Integrating gene expression profiling and clinical data. Int. J Approx. Reas., in press, 2007.

[14] C.M. Perou, S.S. Jeffrey, M. van de Rijn, C.A. Rees, M.B. Eisen, D.T. Ross, A. Pergamenschikov, C.F. Williams, S.X. Zhu, and J.C. et al. Lee. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA*, 96:9212–9217, 1999.

[15] C.M. Perou, T. Sorlie, M.B. Eisen, M. van de Rijn, S.S. Jeffrey, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, and L.A. et al. Akslen. Molecular portraits of human breast tumours. *Nature*, 406:747–752, 2000.

[16] J.R. Pollack, T. Sorlie, C.M. Perou, C.A. Rees, S.S. Jeffrey, P.E. Lonning, R. Tibshirani, D. Botstein, A.L. Borresen-Dale, , and P.O. Brown. Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. USA*, 99:12963–12968, 2002.

[17] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, and T.R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 2002.

[18] T. Sorlie, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M.B. Eisen, M. van de Rijn, and S.S. et al. Jeffrey. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA*, 98:10869–10874, 2001.

[19] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lonning, P. O. Brown, A. L. Borresen-Dale, and D. Botstein. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*, 100(14):8418–8423, 2003.

[20] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[21] E.S. Venkatraman and A.B. Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657–663, 2007.

Supervised classification of combined copy number and gene expression data

S. Riccadonna[1,2], G. Jurman[1], S. Merler[1], S. Paoli[1,2], A. Quattrone[2], C. Furlanello[*1]

[1]FBK-irst, via Sommarive 18, I-38100 Povo (Trento), Italy
http://mpa.itc.it

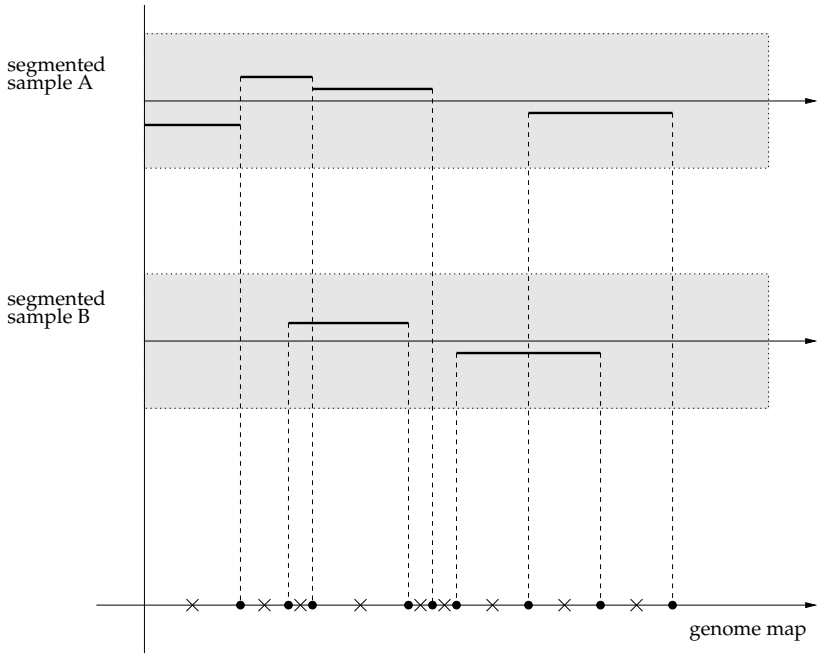
[2]DIT, University of Trento, via Sommarive 14, I-38100 Povo (Trento), Italy
http://dit.unitn.it

SUPPLEMENTARY MATERIAL

[*]Corresponding author

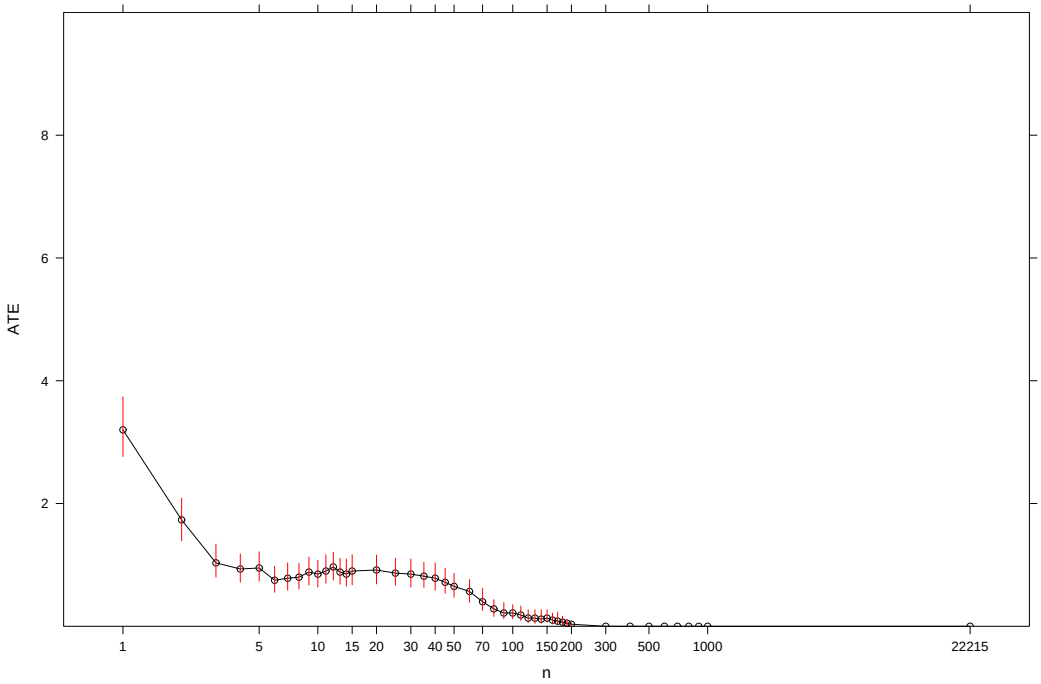
1 aCGH preprocessing

S1 Building the BioDCV matrix from segmented aCGH data: after identifying all the change-points (the solid dots on the bottom line) across samples, features are labeled by using the midpoint of each segment (the crosses on the bottom line). For each sample, the log2ratio value of the associated segment is assigned to each feature.



2 Gene expression classifiers

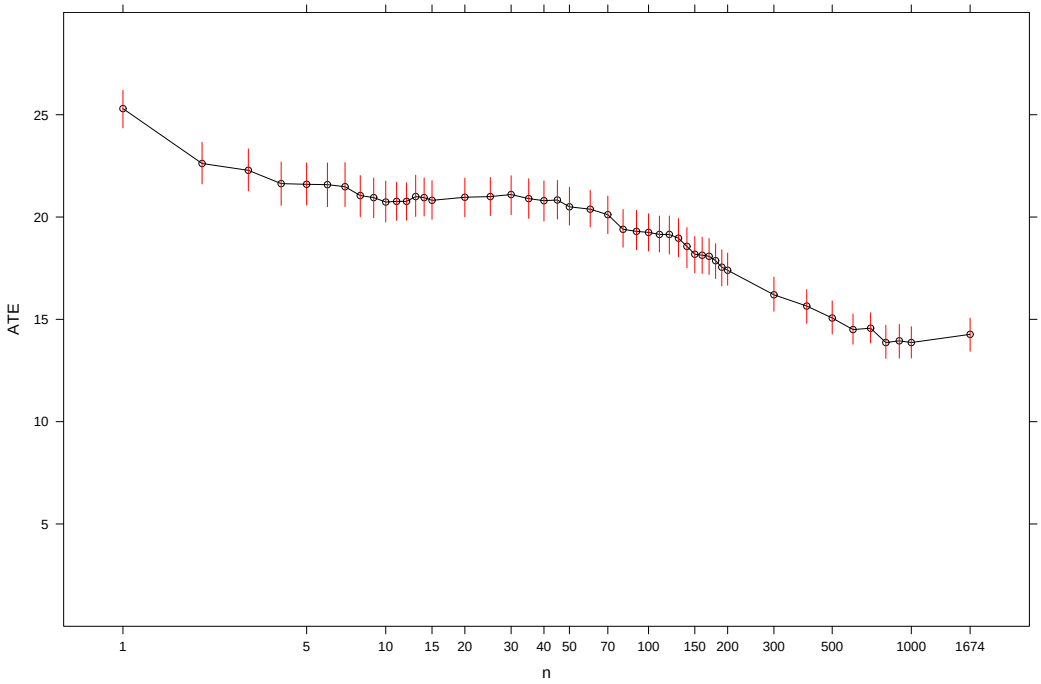
S2 The Average Test Error is displayed for gene expression data and the Lum/ER+ vs. Basal/ER- task, with 95% student's bootstrap confidence intervals.



S3 Sample-tracking analysis of the Lum/ER+ vs. Basal/ER- task on the expression data. For each sample, the plot indicates percent error for BioDCV runs (indicated in parenthes) in which the sample is in test, averaged on models of the same feature set size. The horizontal grey line indicates the no-information error rate for baseline comparison.

3 Segmented aCGH classifiers

S4 The Average Test Error is displayed for the segmented copy number data and the Lum/ER+ and Basal/ER- task, with 95% student's bootstrap confidence intervals.



S5 Average Test Error (M) on the segmented copy number data for the Lum/ER+ and Basal/ER- task (global and classwise), with standard deviations (SD).

# feat	$M_{\text{Lum/ER+}}$	$SD_{\text{Lum/ER+}}$	$M_{\text{Basal/ER-}}$	$SD_{\text{Basal/ER-}}$	M	SD
1	10.15	10.06	55.60	22.76	25.30	14.29
2	12.03	11.68	43.80	21.51	22.62	14.96
3	13.80	12.27	39.25	20.63	22.28	15.05
4	14.35	12.75	36.20	21.88	21.63	15.79
5	14.93	12.38	34.95	22.06	21.60	15.61
10	14.63	11.95	32.95	21.46	20.73	15.12
15	14.93	11.85	32.60	21.08	20.82	14.92
20	14.88	11.97	33.15	20.86	20.97	14.94
25	15.03	12.08	32.95	20.75	21.00	14.97
50	15.58	12.11	30.35	20.42	20.50	14.88
100	15.20	11.67	27.35	20.00	19.25	14.45
200	14.78	11.26	22.65	19.82	17.40	14.11
500	13.98	11.26	17.25	17.73	15.07	13.41
1000	11.95	10.75	17.70	18.48	13.87	13.32
1674	12.38	10.63	18.05	18.60	14.27	13.28

4 Subclass analysis

S6 Comparison of gene expression on optimal gene expression signature of the 8 aCGH misclassified samples with the remaining 55 samples.