the classification performance of the proposed system. The cross-validation (CV) technique is used widely for predicting true classification error rate in samples that are not included in either the training or the test sample sets. Among the CV methods, the leave-one-out cross-validation technique (LOO) is often used because of its small bias.[18] These studies, however, paid little attention to the variances of estimated classification error rates.

The estimated variances in the assessment process are important for practical applications. Even if a classifier has a sufficiently low error rate accompanied instead by large variance in prediction, it suffers from a high risk of having a large actual error rate when applied to unknown test samples.[21] The LOO criterion sometimes selects a classifier involving a very small number of genes, or even a single gene. Although the single-gene classifier fits the 'as few genes as possible' criterion, classifiers involving redundant genes tend to exhibit lower noise and provide better prognosis.[9] Several recent methods consider the estimated error rate variances,[21–24] and unsupervised methods[25,26] also minimize the variance of the model by focusing on the stability of the signatures instead of on the supervised class labels. However, there has been no discussion from the viewpoint of mini-chip design, namely, to explore a reliable predictor based on as few genes as possible.

In the present study, we consider both the bias and the variance of performance estimation so as to achieve a reliable predictor. We applied a bootstrap sampling method to estimate the distribution of possible error rates, with bias and variance, and propose a min-max criterion to obtain a stable classifier. We conducted a simulation study and found that the min-max criterion tends to select better candidate predictors than the LOO criterion, especially when the number of samples is small. We then compared two supervised analysis procedures, T-WV and R-SVM, and showed that T-WV achieves reliable predictors with a small number of genes, indicating that T-WV with the min-max criterion is desirable for our purpose of obtaining a reliable predictor with as few genes as possible.

2. Methods

2.1. Notations

Let $x_i = (x_{i1}, ..., x_{iM})$ be a vector of the M-dimensional gene expression profile of the i-th sample, and y_i a binary class label $y_i \in \{-1,1\}$ representing the binary status of the i-th sample, for example, tumor or non-tumor. The numbers of samples in the negative ($y_i = -1$) and positive ($y_i = 1$) classes are denoted as n_n and n_p, respectively. Suppose that we have a dataset $D = \{d_i \mid i = 1, ..., N\}$ including N samples, where $d_i = (x_i, y_i)$ is a pair of input (expression) and output (class label) of the i-th sample. By applying a supervised machine learning method to the dataset D, we construct a discriminant function $h(x \mid D)$ such that we predict a label $\hat{y}(x')$ for a new input x' by

$$\hat{y}(x') = \begin{cases} 1 & \text{if } h(x' \mid D) \geq 0 \\ -1 & \text{if } h(x' \mid D) < 0 \end{cases} \quad (1)$$

2.2. T-WV method

The WV method is a typical supervised machine learning method that employs the top k significant genes. Since the significance of the j-th gene is defined according to the following t-score, the entire procedure is referred to as the T-WV method,

$$t_j = \frac{\bar{x}_{pj} - \bar{x}_{nj}}{\sqrt{1/n_p + 1/n_n}\, S_j}, \quad (2)$$

where $\bar{x}_{pj}$ and $\bar{x}_{nj}$ are the average expression levels of the j-th gene over the training samples labeled 1 and −1, respectively, and S_j^2 is the pooled within-class variance of the j-th gene,

$$S_j^2 = \frac{\sum_{i:y_i=-1}(x_{ij} - \bar{x}_{nj})^2 + \sum_{i:y_i=1}(x_{ij} - \bar{x}_{pj})^2}{n_n + n_p - 2}. \quad (3)$$

The genes are ranked according to the absolute value of $|t_j|$, and the top-ranked k genes are selected as significant genes so that the set of these genes is denoted as C_k. The discriminant function obtained by the T-WV method is then constructed as

$$h_k(x \mid D) = \frac{1}{k} \sum_{j \in C_k} t_j (x_j - \bar{x}_j), \quad (4)$$

where $\bar{x}_j \equiv \frac{1}{N} \Sigma_j^N x_{ij}$ is the average expression level of the j-th gene in the training samples.

In the discriminant function h_k, the difference between the j-th gene expression and its average is weighted by its significance, i.e. the t-score. Note that

the function h_k depends on the number k of significant genes, and thus we need to specify k appropriately.

2.3. R-SVM method

R-SVM is another typical supervised machine learning method, which was developed to select important genes for SVM classification.[15] An R code package is publicly available at http://www.hsph.harvard.edu/bioinfocore/R-SVM.html. The discriminant function of a linear SVM is defined as

$$h_k(x' \mid D) = (w \cdot x') + b = \sum_{i=1}^{N} \alpha_i y_i (x_i \cdot x') + b, \quad (5)$$

where x' is a new input expression vector and x_i is the i-th sample expression vector in the training dataset. α_i and b are parameters to be determined so that training data points with different class labels are classified with the largest margin. $x \cdot x' = \sum_{j=1}^{M} x_j x'_j$ denotes the inner product. Each element of w, w_j, is defined as

$$w_j = \sum_{i=1}^{n} \alpha_i y_i x_{ij}, \quad (6)$$

the absolute value $|w_j|$ of which represents the significance weight of the jth gene in the current discriminant function.

As in the T-WV method, the classification performance of SVM also depends on gene subset selection. R-SVM applies a recursive feature elimination (RFE) procedure.[27] In RFE, less significant genes in the current discriminant function are recursively eliminated, and the next discriminant function is constructed based on the new, smaller set of genes. Consequently, a sequence of discriminant functions with decreasing numbers of genes is constructed. Thus, the prediction performance of each discriminant function h_k depends on the number k of significant genes, which causes the same problem as in T-WV, i.e. setting an appropriate number k. In the following section, we describe a common way to set the number of genes in both T-WV and R-SVM.

2.4. LOO model selection

T-WV and R-SVM, both produce many candidate classifiers, from which we should select the best one by an assessment process. Although the true performance of a classifier is measured as classification accuracy on an unknown dataset given in the future, we should instead estimate the performance using the dataset obtained in the assessment process. Note that we refer to each candidate in the assessment process as a *model*, to clarify that we are assessing all procedures used to construct a classifier rather than assessing solely the classifier. In T-WV and R-SVM, a model is characterized by the number of significant genes that it includes.

The LOO procedure has been widely used to estimate, or predict, the future performance of a classifier. In LOO, a classifier h is built using each leave-one-out dataset D^{-i}, $i = 1, ..., N$; that is, the i-th sample d_i is excluded in the training procedure from the dataset D, and becomes a validation sample. The classification performance of h is assessed using the validation sample. After the assessments for $d_1, ..., d_N$, the LOO error rate of the classifier h, $G_{\mathrm{LOO}}(h \mid D)$, is calculated as the averaged error rate

$$G_{\mathrm{LOO}}(h \mid D) = \frac{1}{N} \sum_{i=1}^{N} I(y_i h(x_i \mid D^{-i}) < 0), \quad (7)$$

where $I(R)$ denotes the indicator function that takes a value of one if condition R holds, and is otherwise zero. When we select the number k of significant genes by

$$h_k^{\mathrm{LOO}} = \operatorname{argmin}_k G_{\mathrm{LOO}}(h_k \mid D), \quad (8)$$

this model selection is said to be based on the LOO criterion.

2.5. Resampling bootstrap method

It is known that the error rates used to estimate the LOO procedure are nearly unbiased. Molinaro et al[18] compared estimated generalization error rates between different resampling methods and showed that LOO had the smallest bias for a simulation dataset and a real microarray dataset. However, LOO has a tendency to include large variance, despite its small bias,[28] because classifiers constructed based on the leave-one-out datasets, D^{-i}, are quite similar to each other, whereas the data points used for validation vary widely. The large variance of the error rate estimation leads to a high risk of selecting a classifier whose 'true' performance is poor, and this risk becomes

higher as the number of candidates becomes larger. When we assess the performance of many candidate classifiers with large variances, some of the candidates often exhibit remarkably low errors, even if their true performance is poor. This is the same problem as overfitting, which was originally found in parametric learning especially when there are many parameters to be learnt. Therefore, it is important to reduce the estimation variance to obtain a robust classifier.

We applied a bootstrap method to simulate possible variation of the given dataset and to obtain the distribution of LOO error rates over the range of that variation. We generated bootstrap datasets $\{D^{*b} \mid b = 1, ..., B\}$, in which each bootstrap dataset is defined as

$$D^{*b} = \{d_r^{*b} = (x_r^{*b}, y_r^{*b}) \mid r = 1, ..., N-1\}, \quad (9)$$

where d_r^{*b} is randomly sampled with replacements from the LOO dataset D^{-i}. The single validation sample d_i is evaluated by the classifiers that were trained by different datasets D^{*b}, leading to a set of LOO error rates: $G_{\text{LOO}}(h_k^{*1} \mid D^{*1}), G_{\text{LOO}}(h_k^{*2} \mid D^{*2}), ..., G_{\text{LOO}}(h_k^{*B} \mid D^{*B})$. h_k^{*b}, $b = 1, ..., B$, is given by Eq. (4) after replacing the dataset D with the bootstrap dataset D^{*b}. This set of LOO error rates is considered to be a distribution of G_{LOO} and provides a guideline to determine the number of genes used in the T-WV classifier.

2.6. Min-max model selection

Using the simulated distribution of LOO error rates, $\{G_{\text{LOO}}(h_k^{*b} \mid D^{*b})\}_{b=1}^{B}$, we defined a risk score called a min-max criterion,

$$G_{\text{BOOT}}(h_k \mid D) = \text{Per95}\left(\{G_{\text{LOO}}(h_k^{*b} \mid D^{*b})\}_{b=1}^{B}\right), \quad (10)$$

where 'Per95' denotes the 95th percentile of the set of values. Based on this risk score, an appropriate model (i.e. the number of genes, k) is selected as

$$h_k^{\text{BOOT}} = \underset{k}{\text{argmin}}\{G_{\text{BOOT}}(h_k \mid D)\}. \quad (11)$$

We considered the 95th percentile with the number of bootstrap $B = 100$ as the representative of possible high error rates for each model with different numbers of genes. The 95th percentile is a robust criterion to estimate the risk of selecting a bad model against the possibly asymmetric nature of the error rate distribution.

Our approach is referred to as the "min-max" selection criterion because we minimized the risk of selecting a model for which the expected prediction error rate was almost the maximum in the distribution of possibilities. This min-max model selection is likely to refuse classifiers for which the estimated error rates are distributed with a large variance, even if LOO shows the lowest error rate from a single dataset. Therefore, the min-max criterion reduces the instability stemming from the variation of possible future datasets that could be simulated by random sampling from a large pool of samples.

In other words, the min-max criterion assumes an underlying game between an analyzer and nature. A dataset is given by nature, and a model is selected by an analyzer. For the analyzer to achieve stability, one good idea is to minimize the risk (Eq. (11)), which stems from the possibility that nature could provide a bad situation (and hence the classifier has been over-trained) (Eq. 10).

The number 95 of the percentile and number of bootstrap $B = 100$ were determined arbitrarily by considering trade-offs between computation time, estimation variance of the percentile point, and appropriateness as a representative of high error rates:

- The computation time is proportional to the number of bootstrappings.
- Estimation variance is a monotonic function of both the percentile number and the number of bootstrappings. Namely, the variance becomes large as the percentile number diverges from 50 and as the number of bootstrappings is small.
- The criterion should evaluate possible high error rates even when the distribution of bootstrap samples is asymmetric.

We did not select the 50th percentile, i.e. the median, because of the third reason above; we attempted to obtain a safe classifier rather than to show good average performance. Although the 99th percentile could be another representative of possible high error rates, we rejected it, because it relies on 1% of bootstrap samples, and will therefore lead to high variance especially with small B. The estimation

variance of each percentile of the bootstrap error rate can be evaluated in terms of the standard deviation of the corresponding order statistic if the distribution of error rates is known. Table 1 shows the standard deviations (SDs) of several percentiles when the distribution of error rates is a standard normal distribution. These SDs are proportional to the SD of the distribution of error rates, implying that the SDs of the percentiles can represent their variation well even for non-normal distributions.

3. Results

3.1. Results for real datasets

We evaluated our method using four published real gene expression profile datasets:

- Breast cancer
  van't Veer et al[3] obtained gene expression microarray data for approximately 5,000 genes for 78 + 19 breast cancer tissue samples. The samples were classified into favorable and unfavorable samples: patients with recurrence-free survival in five years and those with recurrence in five years, respectively. The authors trained supervised classifiers using 78 samples (34 favorable and 44 unfavorable samples), which we call the training dataset, and tested using 19 independent samples (7 favorable and 12 unfavorable samples), which we call the test dataset. The same group also provided a larger dataset consisting of 295 samples.[29] Among the 295 samples, 32 samples were also included in the former dataset[3] and 10 samples were censored in five years; hence, we used the remaining 253 (192 favorable and 61 unfavorable) samples for the second test dataset.
- Colon cancer
  The colon cancer dataset[30] contains microarray expression data for 2,000 genes for 62 colon tissues. Among the 62 tissue samples, 40 and 22 were labeled as "tumor" and "normal," respectively, and these were used as the labels to be predicted.
- Neuroblastoma (NBL)
  The NBL dataset[5] consists of microarray expression data for 5,180 genes for 136 patients. Among the 136 samples, 25 and 102 were labeled as "favorable" and "unfavorable" patients, respectively, according to their status at 24 months after diagnosis, and these were used as the labels to be predicted. The remaining nine samples of unknown status at 24 months after diagnosis were omitted.
- Breast cancer Affymetrix (Affymetrix)
  Wang et al[31] analyzed 286 breast cancer patients with an Affymetrix chip harboring 22,283 genes. Among the 286 patients, 183 and 93 were labeled as favorable and unfavorable, respectively, and these were used as the labels to be predicted. We omitted 10 samples which were censored in five years. Although this dataset concerned breast cancer, we did not consider relationship between this set and the breast cancer datasets at the top of this list because these two datasets were assembled by entirely different systems and hence had fairly different characters in distribution. Considering different systems of microarrays together may be an important issue, but is beyond the scope of the current study.

Table 1. Estimated standard deviations of bootstrap percentiles. Bold type marks the setting which we used in the current study.

	$B = 100$	$B = 500$	$B = 1000$
99th	0.315	0.171	0.120
95th	**0.216**	0.095	0.067
90th	0.172	0.077	0.054
50th	0.125	0.056	0.040

For each of the above four datasets, we trained T-WV and R-SVM classifiers with various numbers of genes using the training samples, and assessed their classification errors in terms of LOO, 3-, 5- and 10-fold-CV, and min-max criteria. In the case of the breast cancer dataset with large numbers of test samples,[3,29] we also assessed their classification errors in the test datasets.

Figure 1 shows the results for the breast cancer dataset. The results with the T-WV classifier (left panel), indicated characteristic behaviors of the three criteria to assess the classification error rate, LOO (dashed line), 3-fold-CV (dotted line), and the proposed min-max criterion (solid line at the top of the blue area). The 90% interval of LOO error rates (blue area), which was estimated by the resampling bootstrap method, describes the estimation variance of error rates. The LOO error rate profile showed the

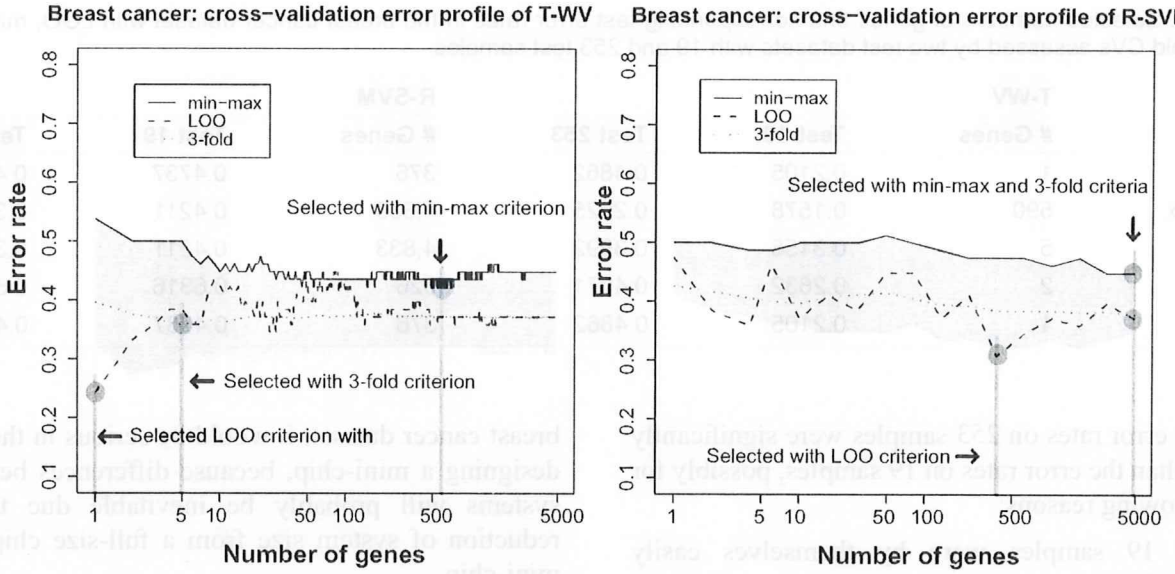

Figure 1. Estimated classification errors in the breast cancer dataset. The left and right panels show the results obtained with the T-WV and R-SVM methods, respectively. The vertical and horizontal axes denote classification error rates estimated by various criteria and the number of genes included in each classifier, respectively. The 90% interval of resampling bootstrap of the estimated classification errors at each number of genes is denoted by blue areas. The classification errors estimated by the three criteria, min-max criterion (solid line on the top of blue area), LOO error rate (dashed line), and 3-fold-CV error rate (dotted line), are plotted against different numbers of genes. Vertical lines indicate the numbers of genes selected by the three criteria.

lowest value with a small number of genes, $k = 1$, so that $k = 1$ was selected as the best number of genes by the LOO criterion. On the other hand, the 90% interval of the bootstrap distribution at $k = 1$ exhibited a large width in the error rate, and the 95th percentile error rate was above the chance level 0.5, suggesting large risk of the $k = 1$ classifier falling into a poor predictor around the chance level. Also, the LOO error rate at $k = 1$ was below both the 5th percentile and the 3-fold-CV error rate, indicating that the low LOO error rate at $k = 1$ could have been obtained by chance. The 3-fold-CV showed a smoother profile than those obtained by the LOO, and stayed in the midst of the 90% interval. The 3-fold-CV criterion selected a classifier with $k = 5$ where the 90% interval was narrower than that at $k = 1$. We also calculated 5- and 10-fold-CVs and observed similar curves to that of the 3-fold-CV. The proposed min-max criterion, i.e. the 95th percentile, selected a larger number of genes, $k = 590$. The LOO and 3-fold-CV error rates at $k = 590$ were higher than those at $k = 1$ and $k = 5$; however, we expected that the classifier of $k = 590$ would have a lower risk of being a poor predictor than those at $k = 1$ and $k = 5$.

In the right panel of Figure 1, a similar comparison is shown between LOO, 3-fold-CV, and the min-max criteria with the R-SVM classifier. The LOO criterion showed an instability similar to that of T-WV, so that the lowest LOO error rate at $k = 376$ seems to have been obtained by chance. All criteria selected larger numbers of genes than in the cases of T-WV classifiers.

In Table 2, test error rates of the selected predictors were assessed using two test datasets with 19 and 253 samples, where five criteria (LOO, min-max, and 3-, 5- and 10-fold-CVs) with two classifiers (T-WV and R-SVM) are compared. The min-max criterion outperformed the other criteria, LOO and k-fold-CVs, on both test sets. The LOO exhibited poor performance with 19 test samples and worse with 253 test samples whose test error rate was around the chance level. Intuitively, this result pointed out a defect of the LOO criterion in terms of the risk of taking a poor classifier, which has already been suggested by the 90% interval shown in Figure 1. The 3-, 5- and 10-fold-CVs achieved better performance in test error rates than LOO, but worse than the min-max criterion. T-WV tended to exhibit lower error rates than R-SVM with smaller numbers of genes, although we cannot conclude the general superiority of T-WV based on this single example.

Table 2. Selected numbers of genes and corresponding test error rates in the breast cancer dataset with LOO, min-max, and k-fold CVs assessed by two test datasets with 19 and 253 test samples.

	T-WV			R-SVM		
	# Genes	Test 19	Test 253	# Genes	Test 19	Test 253
LOO	1	0.2105	0.4862	376	0.4737	0.4664
min-max	590	0.1578	0.2925	4,833	0.4211	0.3992
3-fold	5	0.3158	0.3992	4,833	0.4211	0.3992
5-fold	2	0.2632	0.4071	626	0.6316	0.5217
10-fold	1	0.2105	0.4862	376	0.4737	0.4664

Test error rates on 253 samples were significantly worse than the error rates on 19 samples, possibly for the following reasons:

- The 19 samples were by themselves easily classified.
- The number of samples (19) was too small to reproduce the error rate with low variance.
- The test data of 253 samples were gathered from different populations from those for the training data of 78 samples and the other test data of 19 samples.
- The microarray measurement system differed between the two sets of data.

The considerations above will be important when designing mini-chips based on training datasets. Although the last reason, difference in microarray systems, may not be very serious in the case of this breast cancer dataset, it would be serious in the case designing a mini-chip, because differences between systems will probably be inevitable due to the reduction of system size from a full-size chip to a mini-chip.

We compared three criteria, LOO, min-max, and 3-fold-CV, with the two classifiers T-WV and R-SVM on the other three datasets (NBL, colon cancer and breast cancer Affymetrix) in Figures 2, 3 and 4, respectively. From the total comparisons over Figures 1–4, we observed the following tendencies:

- Although the error rates estimated by LOO fluctuate as the number of genes increases, they stay mostly within the 90% interval. This suggests that the LOO estimation of the tuned number of genes includes a large variance and the character

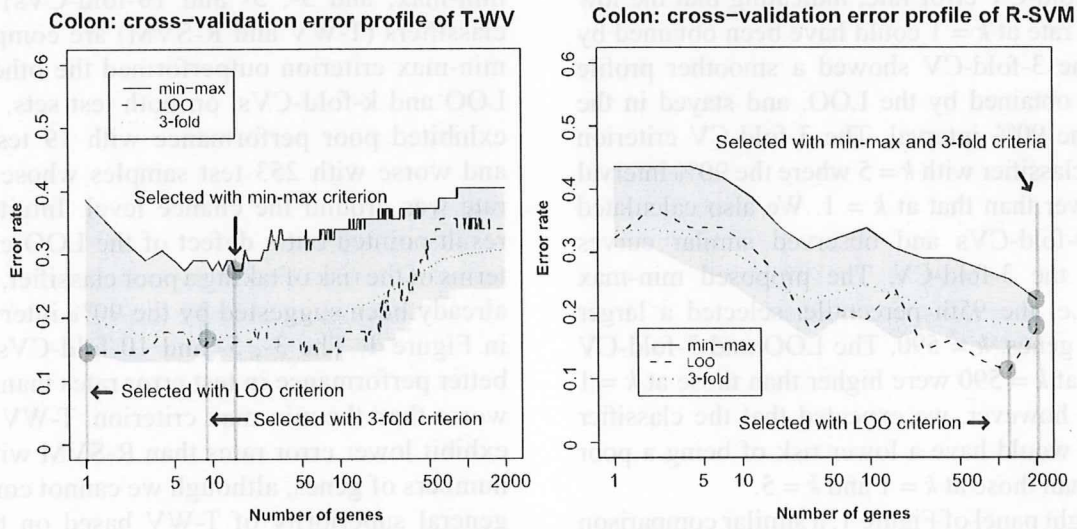

Figure 2. Estimated classification errors in the colon cancer dataset. See Figure 1 legend for details.

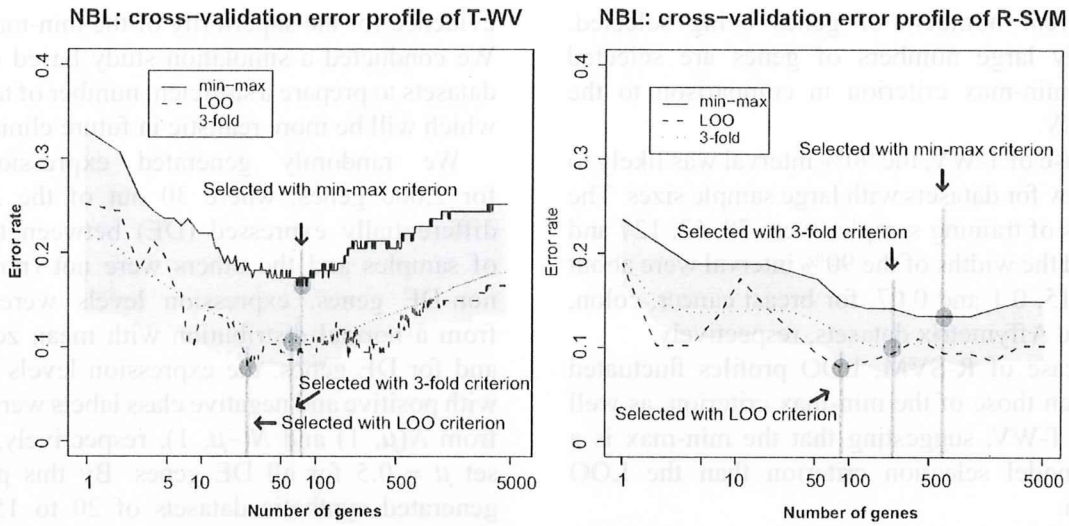

Figure 3. Estimated classification errors in the NBL dataset. See Figure 1 legend for details.

of the variance is well captured by the estimated 90% interval.

- In contrast to the fluctuating profile of LOO error rates, the profiles of the 3-fold-CV and the 95th percentile (G_{BOOT}) exhibit smoother curves. This suggests a more stable character for the 3-fold-CV and the min-max criterion than the LOO criterion.
- With T-WV, the 90% confidence interval was likely to be wide when the number of genes was small, $k < 10$, indicating that prediction based on too few genes is risky; we occasionally get a model with poor performance. The 95th percentile is likely to show a higher error rate for a smaller number of genes, e.g. $k < 10$, than for a large number of genes. Thus, the min-max criterion based on the 95th percentile can avoid risky prediction so that a smaller error rate is achieved on average.
- The 3-fold-CV profile stayed almost in the middle of the 90% interval and showed a similar curve to the 95th percentile. However, there was difference between the 3-fold-CV and the 95th percentile in the range of 90% interval, which was prominent in T-WV with small numbers of genes, $k < 10$. The 3-fold-CV and the min-max criterion lead

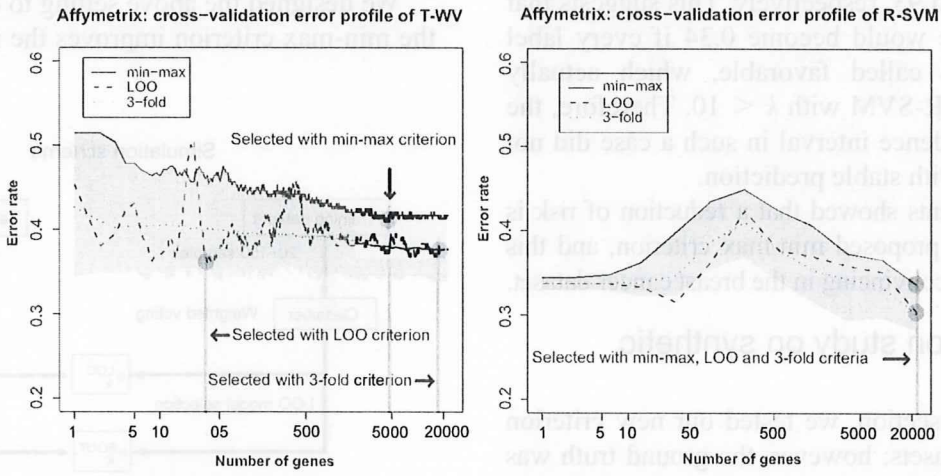

Figure 4. Estimated classification errors in the Affymetrix dataset. See Figure 1 legend for details. Note that the errors of R-SVM with $k < 10$ are not reliable because of unbalanced numbers of labeled samples.

to different numbers of genes being selected; relatively large numbers of genes are selected by the min-max criterion in comparison to the 3-fold-CV.

- In the case of T-WV, the 90% interval was likely to be narrow for datasets with large sample sizes. The numbers of training samples were 78, 62, 127 and 276, and the widths of the 90% interval were about 0.15, 0.15, 0.1 and 0.07, for breast cancer, colon, NBL and Affymetrix datasets, respectively.
- In the case of R-SVM, LOO profiles fluctuated more than those of the min-max criterion, as well as with T-WV, suggesting that the min-max is a better model selection criterion than the LOO criterion.
- Whereas the best performance was comparable between R-SVM and T-WV, a larger number of genes was required to achieve the best performance by R-SVM than by T-WV. Thus, T-WV employing a relatively small number of genes is more suitable for practical clinical applications, which is consistent with a previous finding.[12]
- The confidence intervals for R-SVM were likely to be narrower than those for T-WV, implying that SVM, as a large margin classifier, is more stable against observation noise than T-WV. Even though we are not interested here in classifiers with a large number of genes, say $k > 1{,}000$, this finding may be important for applications other than mini-chip construction.
- The Affymetrix data set was unbalanced, with the numbers of favorable and unfavorable samples being 183 and 93, respectively. This suggests that the error rate would become 0.34 if every label prediction is called favorable, which actually occurred for R-SVM with $k < 10$. Therefore, the narrow confidence interval in such a case did not correspond with stable prediction.

The experiments showed that a reduction of risk is achieved by the proposed min-max criterion, and this was particularly convincing in the breast cancer dataset.

3.2. Simulation study on synthetic datasets

In the previous section, we tested our new criterion on four real datasets; however, the ground truth was unknown and the number of samples was limited in many cases, which prevented us from obtaining strong evidence for the superiority of the min-max criterion. We conducted a simulation study based on artificial datasets to prepare a sufficient number of test samples, which will be more realistic in future clinical studies.

We randomly generated expression profiles for 2,000 genes, where 30 out of the 2,000 were differentially expressed (DE) between two classes of samples and the others were not (non-DE). For non-DE genes, expression levels were generated from a normal distribution with mean zero, $N(0,1)$, and for DE genes, the expression levels of samples with positive and negative class labels were generated from $N(\mu, 1)$ and $N(-\mu, 1)$, respectively, where we set $\mu = 0.5$ for all DE genes. By this process, we generated synthetic datasets of 20 to 150 samples for training, and 1,000 samples for testing, where the numbers of samples with the two class labels were set to be equal.

The proposed simulation scheme is illustrated in Figure 5. For each training dataset, the candidate classifiers involving various numbers of genes were trained and assessed, and the best numbers of genes were selected by the LOO and the min-max criteria, where the number B of the bootstrap in the min-max procedure was set at 100. The performance of the finally selected classifier was then assessed by a test dataset with 1,000 samples. We repeated this process with a randomly generated training dataset and assessed the corresponding test error rates by using a test dataset of 1,000 samples. The distributions of the test error rates were compared between different conditions.

We designed the above setting to clarify how well the min-max criterion improves the model selection.

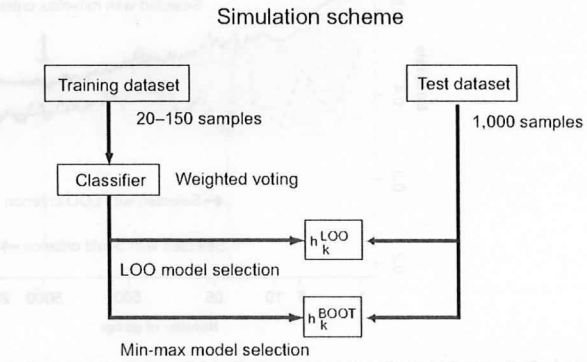

Figure 5. Setting of the simulation experiment.

The number of test datasets was set sufficiently large, and is commonly used in various settings of the other features to reduce the variance of error rates that stems from random sampling of the test dataset. The number of DE genes (30) and the strength of differential expression ($\mu = 0.5$) were determined to examine typical situations that arise in realistic cases. We omitted other realistic features of datasets that may arise such as variation in the number of DE genes, strength μ, and the proportion of numbers of positive and negative samples, because they had shown no significant effect in our preliminary experiments. We also omitted correlations of gene expression patterns between DE genes because such correlations would not affect either T-WV or R-SVM.

Figure 6 shows the distributions of test error rates of the T-WV classifiers selected by LOO and min-max, with 20, 50, 100 and 150 training samples. We found that there were certain levels of variance for both criteria, and the variance was larger for smaller numbers of samples. LOO sometimes showed much worse results than min-max, as indicated by the points in the bottom-right area of each panel in Figure 6. Note that the number of test samples, 1,000, was so large that there was no significant increase in sampling variance. Table 3 shows the means and

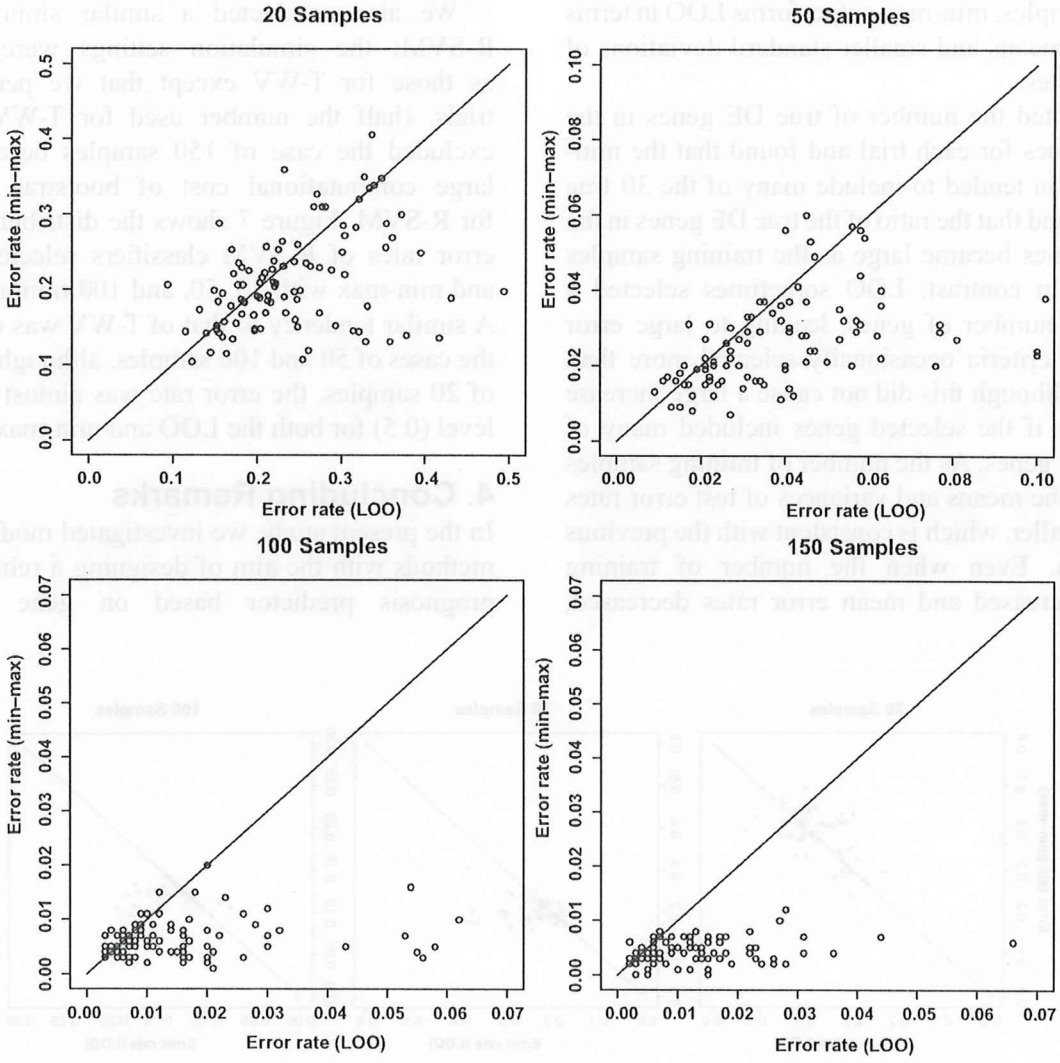

Figure 6. Distribution of test error rates of T-WV. The vertical and horizontal axes denote the test error rates of classifiers selected by the min-max and LOO criteria, respectively. The results from 100 trials of random sampling of 20, 50, 100 and 150 samples are shown in the four panels.

Table 3. Test error rate of simulation dataset.

Number of training samples	Selection criterion	Mean	Standard deviation
20	LOO	0.241	0.077
	min-max	0.210	0.064
50	LOO	0.042	0.024
	min-max	0.026	0.012
100	LOO	0.015	0.013
	min-max	0.006	0.003
150	LOO	0.012	0.010
	min-max	0.004	0.002

standard deviations of test error rates of the classifiers selected by LOO and min-max. Through 20–150 training samples, min-max outperforms LOO in terms of smaller means and smaller standard deviations of test error rates.

We counted the number of true DE genes in the selected genes for each trial and found that the min-max criterion tended to include many of the 30 true DE genes, and that the ratio of the true DE genes in the selected genes became large as the training samples increased. In contrast, LOO sometimes selected a very small number of genes, leading to large error rates. Both criteria occasionally selected more than 30 genes, although this did not cause a large increase in the error if the selected genes included many of the true DE genes. As the number of training samples increased, the means and variances of test error rates became smaller, which is consistent with the previous observation. Even when the number of training samples increased and mean error rates decreased, however, the test error rates of LOO still showed larger variance than those of min-max.

We also conducted a similar simulation with R-SVM; the simulation settings were the same as those for T-WV except that we performed 50 trials, (half the number used for T-WV), and we excluded the case of 150 samples because of the large computational cost of bootstrap simulation for R-SVM. Figure 7 shows the distributions of test error rates of R-SVM classifiers selected by LOO and min-max with 20, 50, and 100 training samples. A similar tendency to that of T-WV was observed in the cases of 50 and 100 samples, although in the case of 20 samples, the error rate was almost the chance level (0.5) for both the LOO and min-max criteria.

4. Concluding Remarks

In the present study, we investigated model selection methods with the aim of designing a reliable cancer prognosis predictor based on gene expression

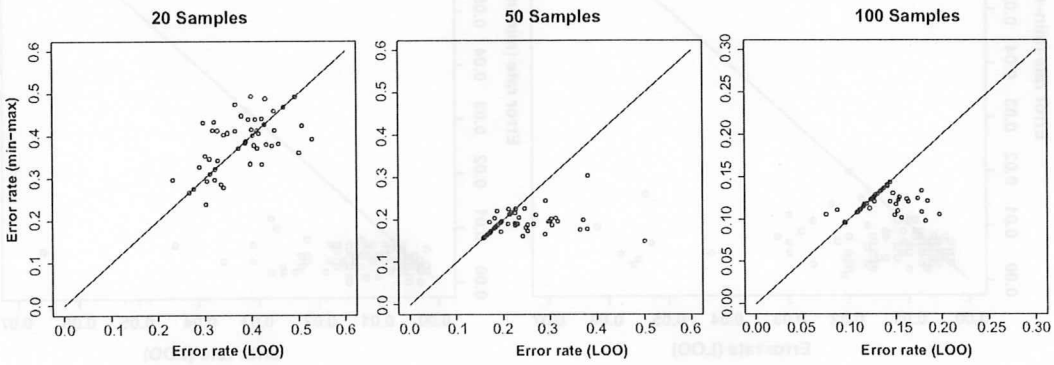

Figure 7. Distribution of test error rates of R-SVM. The vertical and horizontal axes denote the test error rates of classifiers selected by the min-max and LOO criteria, respectively. The results from 50 trials of random sampling of 20, 50 and 100 samples are shown in the three panels.

microarrays involving as small a number of genes as possible. We assessed possible variation in prediction error rate of each microarray-based predictor by simulating a distribution of classification error rates via a resampling bootstrap method. Accordingly, we proposed a novel min-max criterion to select a predictor from multiple candidates. In numerical comparisons that used real and synthetic datasets, we showed that the conventional LOO estimation of their error rates resulted in large variances; consequently, the LOO criterion had a large risk of choosing inappropriate classifiers that would exhibit extremely poor prediction performance. In contrast, we showed the stability of the min-max criterion relative to well-established statistical criteria including the LOO. We also compared two different supervised analysis procedures, T-WV and R-SVM, and found that, in general, T-WV performed the best when it involved a small or moderate number of genes in contrast to that R-SVM performed the best when it involved almost all genes, although the mean and variance of the best possible performances were not always significantly different between those achieved with T-WV and R-SVM. Thus, overall, we concluded to recommend T-WV with the min-max criterion, which satisfied our demand; the most reliable predictor involving as small a number of genes.

It should be important to note that, we proposed our procedure to select a set of genes for designing a good predictor of cancer prognosis, rather than for determining a set of genes which have statistically significant relationship to the prognosis; these purposes are different from each other in general. In other words, the 'robust' model selection is meant to lower the risk to select an extremely poor predictor, rather than to select a stable set of genes. In fact, different research groups reported prognosis prediction systems with different sets of genes based on different sets of microarray data for the same type of cancer.[6] The microarray-based predictors for breast cancer, were designed with 70 and 76 genes by two different research groups,[3,31] respectively, and these gene sets had only three genes in common. Namely, the selected sets of genes were not stable at all, however, the 70 gene-based diagnosis system of breast cancer have been verified by increasingly large number of new patients and authorized by Food and Drug Administration in USA.[6] In our own numerical experiments, we also observed that number of common genes tended to be small between any gene sets that were selected based on different datasets generated by resampling bootstrap (data not shown), although we achieved good predictors in vast amount of the cases as we had shown. Thus, it should be emphasized that such an instable selection of gene subsets did not necessarily cause a poor predictor as long as the predictor was selected by a robust model selection method.

Once a prediction system based on a small number of genes is developed, the system can be transfered not only to mini-chip microarrays but also to other easy accessible devises such as quantitative real-time polymerase-chain-reaction (RT-PCR) analysis,[32] which would be tractable if only tens of genes were targeted. Robust model selection methods, like the proposed one, will be needed especially when we consider such a transfer work between different measurement devises because large bias is often expected between different devises. In general, when a procedure is designed to be robust against measurement variance, such a method is also robust against an unknown bias which would appear like in the above transfer; thus, our min-max criterion will be used for this purpose.

In order to design a practical tool for real scenes in clinical cancer therapy, new demands in informatics can always arise. As we had seen in this study, although past efforts in informatics tended to pursue good performances in average, minimizing risk to catch poor predictor against possible variability in cancer diagnosis systems becomes a next issue. There are few methods to directly seek such risk minimization as long as we know. Reducing cost by selecting relevant genes based on high-dimensional gene expression profile is a relatively well-investigated field of research. However, the combination of the cost and reliability is not investigated well. Thus, there must be room to develop a novel supervised classification method that satisfies these demands for designing mini-chip systems, and future studies in cancer informatics should proceed to such directions.

Authors Contributions

IS performed the experiments and wrote the manuscript. IS and SO proposed the main idea that the variance influenced the performance of the classifiers. TT contributed to the construction of the simulation scheme and the development of the

variance estimation methods. SI provided advice on the min-max strategy and supervised the present study. MO provided several topics concerning real cancer therapy and future directions. All five authors participated in the preparation of the final manuscript.

Acknowledgement

We are grateful to Dr. Ian Smith for his careful proofreading of our manuscript. Thanks are also due to anonymous reviewers for their critical reviews and useful suggestions.

Disclosure

The authors report no conflicts of interest.

References

1. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*. 1999;21:33–7.
2. Alizadeh AA, Eisen MB, Davis EE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403:503–11.
3. van't Veer L, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415:530–6.
4. Iizuka N, Oka M, Yamada-Okabe H, et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet*. 2003;361:923–9.
5. Ohira M, Oba S, Nakamura Y, et al. Expression profiling using a tumor-specific cDNA microarray predicts the prognosis of intermediate risk neuroblastomas. *Cancer Cell*. 2005;7:337–50.
6. van't Veer L, Bernards R. Enabling personalized cancer medicine though analysis of gene-expression patterns. *Nature*. 2008;415:564–70.
7. Draghici S, Khatri P, Eklund AC, et al. Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics*. 2006;22:101–9.
8. Glas AM, Floore A, Delahaye LJMJ, et al. Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics*. 2006;7:278.
9. Jaeger J, Spang R. Selecting normalization genes for small diagnostic microarrays. *BMC Bioinformatics*. 2006;7:388.
10. Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*. 2002;97:77–87.
11. Li T, Zhang C, Ogihara MA comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*. 2004;20:2429–37.
12. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*. 2001;98:15149–54.
13. Golub T, Slonim D, Tamayo P, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*. 1999;286:531–7.
14. Vapnik VN. *The nature of statistical learning theory*. Springer, New York, 2000.
15. Zhang X, Lu X, Shi Q, et al. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*. 2006;7:197.
16. Friedman JH. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining Knowledge Discovery*. 1997;1:55–77.
17. Borgne YL. Bias variance trade-off characterization in a classification. what differences with regression? Technical report, ULB, 2005.
18. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*. 2005;21:3301–7.
19. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91.
20. Wood IA, Visscher PM, Mengersen KL. Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics*. 2007;23:1363–70.
21. Braga-Neto UM, Dougherty E. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*. 2004;20:374–80.
22. Fu WJ, Carroll RJ, Wang S. Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics*. 2005;21:1979–86.
23. Xu Q, Hua J, Braga-Neto U, et al. Confidence intervals for the true classification error conditioned on the estimated error. *Technology in Cancer Research and Treatment*. 2006;5:579–89.
24. Jiang W, Varma S, Simon R. Calculating confidence intervals for prediction error in microarray classification using resampling. *Statistical Applications in Genetics and Molecular Biology*. 2008;7.
25. Davis CA, Gerick F, Hintermair V, et al. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*. 2006;22:2356–63.
26. Haibe-Kains B, Desmedt C, Loi S, et al. *Computational Intelligence in Clinical Oncology: Lessons Learned from an Analysis of a Clinical Study*, volume 122. Springer-Verlag Berlin/Heidelberg, 2008.
27. Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using support vector machines. *Machine Learning*. 2002;46:389–422.
28. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning*. Springer, New York, 2001.
29. van de Vijver MJ, He YD, van't Veer L, et al. A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*. 2002;347:1999–2009.
30. Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*. 1999;96:6745–50.
31. Wang Y, Klijn J, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005;365:671–9.
32. Lossos IS, Czerwinski DK, Alizadeh AA, et al. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *New England Journal of Medicine*. 2004;350:1828–37.

Supplementary Figures for "Robust Model Selection for Classification of Microarrays"

Ikumi Suzuki, Takashi Takenouchi, Miki Ohira, Shigeyuki Oba, and Shin Ishii

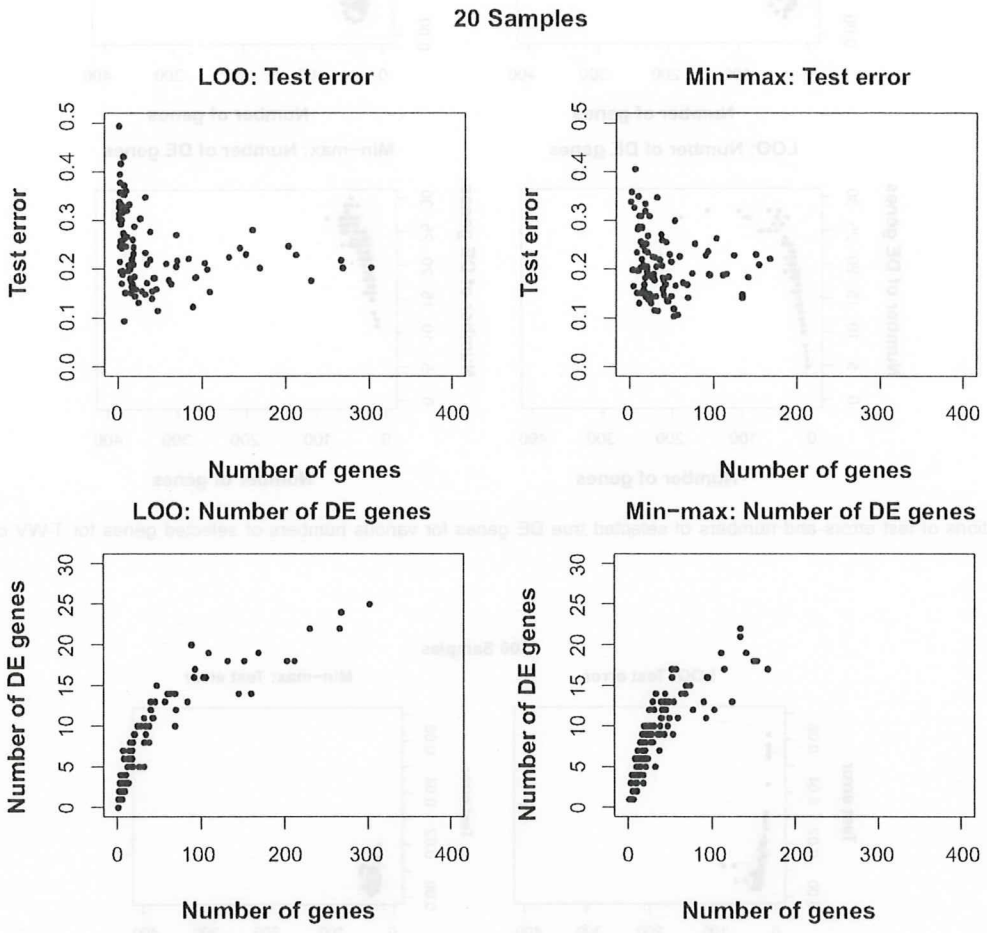

Figure S1. Distributions of test errors and numbers of selected true DE genes for various numbers of selected genes for T-WV classifiers based on 20 artificial samples. Each point denotes one of 100 trials in each setting. Horizontal axes denote the number of genes selected by either LOO or min-max criterion. The vertical axes in the top two panels and the bottom two panels denote the test error estimated by 1000 test samples and the number of true DE genes in the selected set of genes, respectively.

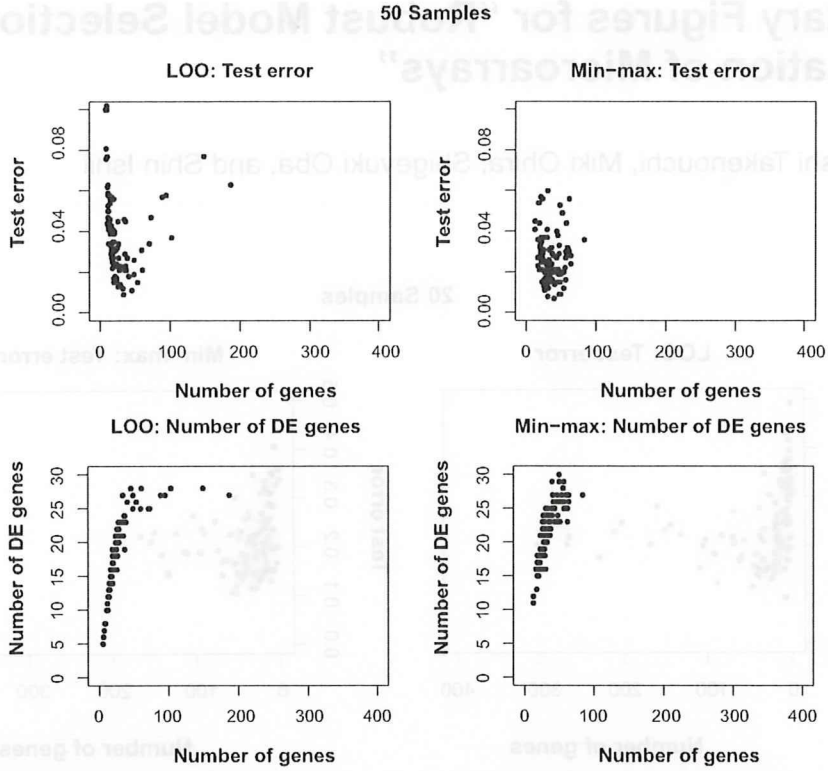

Figure S2. Distributions of test errors and numbers of selected true DE genes for various numbers of selected genes for T-WV classifiers based on 50 artificial samples

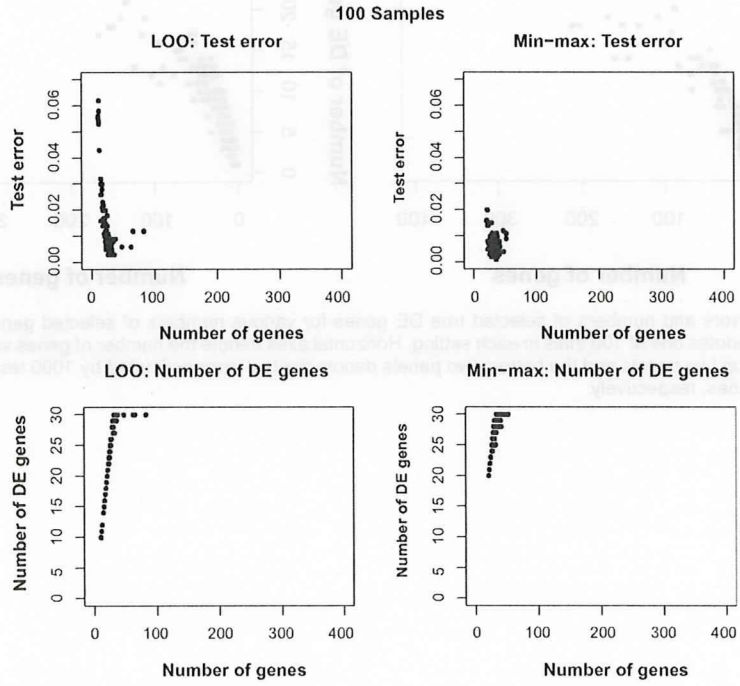

Figure S3. Distributions of test errors and number of selected true DE genes for various numbers of selected genes for T-WV classifiers based on 100 artificial samples.

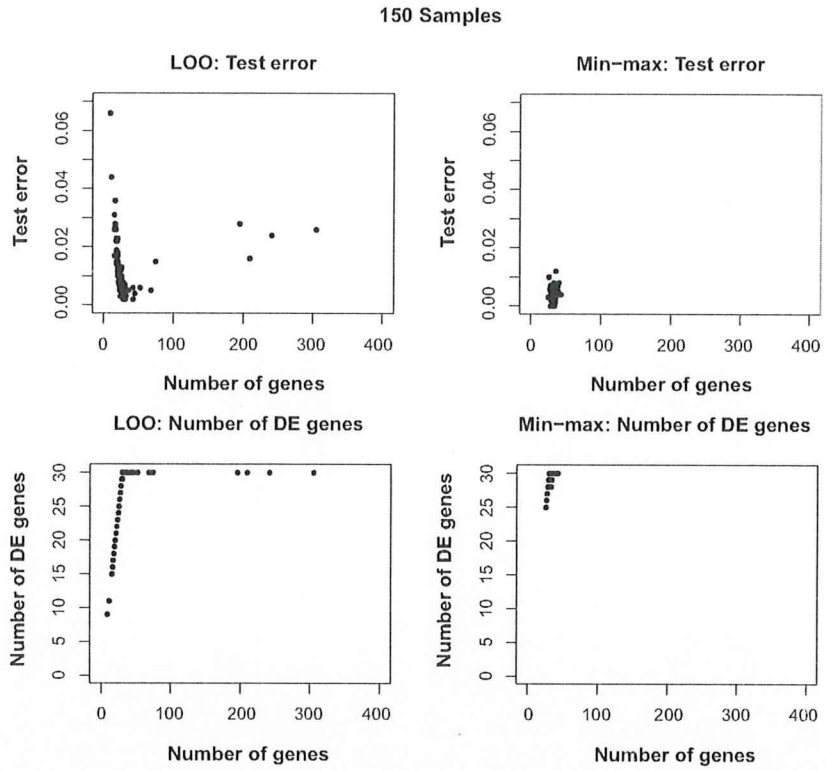

Figure S4. Distributions of test errors and number of selected true DE genes for various numbers of selected genes for T-WV classifiers based on 150 artificial samples.

High expression of *ncRAN*, a novel non-coding RNA mapped to chromosome 17q25.1, is associated with poor prognosis in neuroblastoma

MENG YU[1,2], MIKI OHIRA[1], YUANYUAN LI[1], HIDETAKA NIIZUMA[1], MYAT LIN OO[1], YUYAN ZHU[1,3,4], TOSHINORI OZAKI[1], ERIKO ISOGAI[1], YOHKO NAKAMURA[1], TADAYUKI KODA[3], SHIGEYUKI OBA[5], BINGZHI YU[2] and AKIRA NAKAGAWARA[1]

[1]Division of Biochemistry and Innovative Cancer Therapeutics, Chiba Cancer Center Research Institute, Chiba 260-8717, Japan; [2]Department of Biochemistry and Molecular Biology, China Medical University, Shenyang 110001, P.R. China; [3]Research Center for Functional Genomics, Hisamitsu Pharmaceutical Co., Inc., Chiba 260-8717, Japan; [4]Department of Urology, No. 1 Hospital, China Medical University, Shenyang 110001, P.R. China; [5]Integrated Systems Biology Laboratory, Department of Systems Science, Graduate School of Informatics, Kyoto University, Kyoto 611-0011, Japan

Received October 14, 2008; Accepted December 16, 2008

DOI: 10.3892/ijo_00000219

Abstract. Neuroblastoma shows complex patterns of genetic aberrations including *MYCN* amplification, deletion of chromosome 1p or 11q, and gain of chromosome 17q. The 17q gain is frequently observed in high-risk neuroblastomas, however, the candidate genes still remain elusive. In the present study, we integrated the data of comparative genomic hybridization of 236 tumors by BAC array and expression profiling of 136 tumors by using the in-house cDNA microarray carrying 5,340 genes derived from primary neuroblastomas. A novel candidate gene mapped to chromosome 17q25.1 with two splicing variants, *Nbla10727* and *Nbla12061*, was identified. The transcript size appeared to be 2.3 kb by Northern blot, however, the cDNA sequences had no obvious open reading frame. The protein product was undetectable by both *in vivo* and *in vitro* translation assays, suggesting that the transcript might not encode any protein product. Therefore, we named it as *ncRAN* (non-coding RNA expressed in aggressive neuroblastoma). In analysis of 70 patients with sporadic neuroblastoma, the high levels of *ncRAN* mRNA expression were significantly associated with poor outcome of the patients ($p<0.001$). The multivariate analysis showed that expression of *ncRAN* mRNA was an independent prognostic factor among age, stage, origin and *MYCN* expression. Ectopic expression of *ncRAN* induced transformation of NIH3T3 cells in soft agar, while knockdown of endogenous *ncRAN* with RNA interference significantly inhibited cell growth in SH-SY5Y cells. Collectively, our results suggest that *ncRAN* may be a novel non-coding RNA mapped to the region of 17q gain and act like an oncogene in aggressive neuroblastomas.

Correspondence to: Dr Akira Nakagawara, Division of Biochemistry and Innovative Cancer Therapeutics, Chiba Cancer Center Research Institute, 666-2 Nitona, Chuo-ku, Chiba, Chiba 260-8717, Japan
E-mail: akiranak@chiba-cc.jp

Key words: neuroblastoma, non-coding RNA, ncRAN, prognosis

Introduction

Neuroblastoma is one of the most common pediatric solid tumors in children and originates from sympathoadrenal lineage of the neural crest. Its clinical behavior is heterogeneous because the tumors often regress spontaneously when developed in patients under one year of age, while they grow rapidly and cause very poor clinical outcome when occurring in patients over one year of age (1). Recent cytogenetic analyses have revealed that given subsets of neuroblastoma with unfavorable prognosis often have *MYCN* amplification, gains of chromosome 1q, 2p, and 17q as well as allelic losses of 1p, 3p, and 11q (1). However, the precise molecular mechanisms underlying pathogenesis and progression of neuroblastoma still remain unclear.

Accumulating evidence shows that gain of chromosome 17 or 17q is the most frequent genetic abnormality in neuroblastoma (1-4). We have previously conducted microarray-based comparative genomic hybridization (array-CGH) with a DNA chip carrying 2,464 BAC clones to examine genomic aberrations in 236 primary neuroblastomas (5). Our array-CGH analysis demonstrated three major groups of genomic aberrations in sporadic neuroblastomas (n=112) that can well define the prognoses of neuroblastomas: a genetic group of silent chromosomal aberration (GGS, 5-year cumulative survival rate: 68%), a genetic group of partial chromosomal gains and/or losses (GGP, 43%), and a genetic group of whole chromosomal gains and/or losses (GGW, 80%). The classification of three genetic groups corresponded well with the pattern of chromosome 17 abnormalities, namely, no gain of either chromosome 17 or 17q, gain of chromosome

17q, and gain of whole chromosome 17, respectively (5). Thus, 17q gain has been implicated in close correlation with aggressiveness of neuroblastoma (5-7). The region has been narrowed down to 17q21-qter, in which several important candidate genes such as *Survivin* and *PPM1D* were suggested to be involved in acquiring aggressiveness of neuroblastoma (4,7,8).

In the present study, by combining with our previous array-CGH data, we searched for the candidate 17q gain gene(s) by applying the results of our gene-expression profiling obtained from the analysis of 136 neuroblastoma samples using an in-house cDNA microarray carrying 5,340 genes isolated from primary neuroblastomas (9,10). This approach has led us to identify a novel non-coding RNA as the candidate mapped to the region of chromosome 17q gain. Its high expression is significantly associated with aggressiveness of primary neuroblastomas.

Materials and methods

Patients. Tumor specimens were collected from the patients with neuroblastoma who had undergone biopsy or surgery at various institutions in Japan. Two hundred and thirty-six and 136 tumor samples were used for array-CGH and expression profiling, respectively (5,10). Among them, sporadic cases were 112 and 70, respectively. The clinical stage of tumor was classified according to the INSS criteria (11). Expression data for the latter 70 sporadic neuroblastomas, which were composed of 15 stage 1, 8 stage 2, 17 stage 3, 25 stage 4, and 5 stage 4s tumors, were used for the Kaplan-Meier analysis. The status of *MYCN* amplification in each tumor had been determined as described previously (8). Patients were treated according to previously described protocols (12,13). The procedure of this study was approved by the Institutional Review Board of the Chiba Cancer Center (CCC19-9).

Microarray-based comparative genomic hybridization (array-CGH) and gene expression profiling. Array-based CGH experiments for 236 neuroblastomas by using a chip carrying 2,464 BAC clones which covers the whole human genome at ~1.2-Mb resolution were performed as described previously (5). For the gene expression profiling of 136 neuroblastomas, we employed an in-house cDNA microarray, carrying 5,340 cDNAs obtained from the oligo-capping cDNA libraries generated from anonymous neuroblastoma tissues (10,14-16). The array-CGH and gene expression profile data are available at NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) with accession numbers GSE 5784 and GSE 5779, respectively.

Cells, culture and transfection. NIH3T3, COS7 and human neuroblastoma cell lines were cultured in Dulbecco's modified Eagle's medium (DMEM) or RPMI-1640 medium containing 10% (vol/vol) heat-inactivated fetal bovine serum (FBS) and antibiotics. Cultures were maintained in a humidified atmosphere containing 5% CO_2 at 37°C. COS7 and NIH3T3 cell lines were transiently transfected using Lipofectamine 2000 reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's protocol.

Construction of expression plasmid. The full-length cDNAs of *Nbla10727* and *Nbla12061* were cloned from the established full length-enriched cDNA libraries which we made from the primary neuroblastomas as described (14-16). The full-length cDNAs were then inserted into pcDNA3 or pcDNA3-FLAG plasmids.

In vitro transcription and translation assay. In vitro translation was carried out in the presence of [^{35}S]-methionine using TNT T7 Quick coupled transcription/translation system (Promega, Madison, WI, USA) according to the manufacturer's instructions. The products were resolved by SDS-PAGE and detected by autoradiography.

In vivo [^{35}S]-labeling experiment. COS7 cells were transfected with the FLAG-tagged *ncRAN* expression vectors or the HA-tagged MEL1 expression plasmid. After 24 h, cells were rinsed with 1X PBS 3 times and recultured in fresh growth medium without methionine and antibiotics. Two hours later, [^{35}S]-methionine (GE Healthcare, Tokyo, Japan) was added to the medium to a final concentration of 0.1 mCi/ml, and cells were further incubated. Cells were harvested and whole cell lysates were subjected to immunoprecipitation using a monoclonal anti-Flag antibody or a polyclonal anti-HA antibody. Immunoprecipitates were resolved by SDS-PAGE and detected by autoradiograph.

RNA isolation and semi-quantitative reverse transcription-PCR (RT-PCR). Total RNA was isolated from frozen tumor tissues by an AGPC method (8). Total RNA (5 μg) was employed to synthesize the first-strand cDNA by means of random primers and SuperScript II reverse transcriptase (Invitrogen) following the manufacturer's protocol. We prepared appropriate dilutions of each single stranded cDNA for subsequent PCR by monitoring an amount of glyceraldehyde-3-phosphate dehydrogenase *(GAPDH)* as a quantitative control. The PCR amplification was carried out for 28 cycles (preheat at 95°C for 2 min, denature at 95°C for 15 sec, annealing at 55°C 15 sec, and extension at 72°C 20 sec) for *ncRAN* (*Nbla10727* and *Nbla12061*). The primers used were: *ncRAN* (*Nbla10727*) 5'-CAGTCAGCCTCAGTTTC CAA-3' (forward); 5'-AGGCAGGGCTGTGCTGAT-3' (reverse), *ncRAN* (*Nbla12061*) 5'-ATGTTAGCTCCCA GCGATGC-3' (forward); 5'-CTAACTGCCAAAAGGTTT TCC-3' (reverse).

Northern blot analysis. Total RNA (20 μg) was subjected to electrophoresis and Northern blotting. The cDNA insert (*Nbla10727*) was labeled with [α-^{32}P]-dCTP (GE Healthcare) by the BcaBEST™ labeling kit (Takara, Tokyo, Japan) and used for the hybridization probe.

Soft agar assay. NIH3T3 cells were transfected with FLAG-*Nbla10727*, FLAG-*Nbla12061* or empty vector, and resuspended in 0.33% agar (wt/vol) in DMEM with 10% FBS at a density of 500 cells/plate. Cell suspensions were poured on the top of the base layer (0.5% agar (wt/vol) in fresh medium, and grew in a 5% CO_2 incubator for 14 days. Colonies >100 μm were counted under an Olympus microscope.

Table I. The comparison of *ncRAN/Nbla10727/Nbla12061* expression level among three major groups of genomic aberrations in neuroblastomas.

Genetic group	n	*ncRAN* expression Mean ± SD (log2 ratio)	p-value	
ncRAN-long/Nbla10727				
GGS (silent)	n=10	-1.12±0.39	p=0.004	
GGP (partial 17q+)	n=26	-0.60±0.48		p=0.952
GGW (whole 17+)	n=35	-1.11±0.48	p<0.001	
ncRAN-short/Nbla12061				
GGS (silent)	n=10	-1.60±0.33	p=0.070	
GGP (partial 17q+)	n=26	-1.23±0.59		p=0.163
GGW (whole 17+)	n=35	-1.81±0.43	p<0.001	

n, number of samples; GGS, Genetic group silent (normal 17); GGP, Genetic group partial gains/losses (17q gain); GGW, Genetic group whole gains/losses (17 gain); *ncRAN* expression levels are shown as normalized log2 ratio of microarray data. p-values were calculated based on statistical t-test.

RNA interference. Oligonucleotides for knocking down the *ncRAN* with *Sac*I and *Xho*I extension were inserted into pMuni vector. The oligonucleotides used were: 5'-CCC CATCCTCTAGTAGCCACGGTTTCAAGAGAACCGT GGCTACTAGAGGATTTTTTGGAAAC-3' and 5'-TCG AGTTTCCAAAAAATCCTCTAGTAGCCACGGTTCTCT TGAAACCGTGGCTACTAGAGGATGGGGAGCT-3'. The plasmids containing the oligonucleotide sequence were transfected into SH-SY5Y cells by using Lipofectamine 2000 reagent (Invitrogen) according to the manufacturer's protocol.

Statistical analysis. The Student's t-tests were used to explore possible associations between *ncRAN* expression and other factors, such as age. Kaplan-Meier curves were calculated and survival distributions were compared using the log-rank test. Univariate and multivariate analyses were made according to the Cox hazard models. q-value was also calculated because *ncRAN* expression was measured with 5340 genes in the microarray (17). Statistical significance was set at $p<0.05$.

Results

Identification of a novel Nbla10727/12061 gene mapped to chromosome 17q25.1 upregulated in advanced neuroblastomas with gain of chromosome 17q. To explore the candidate genes for therapeutic target against aggressive neuroblastomas, the genomic and molecular characteristics specific to high-risk tumors were surveyed. We previously conducted array-CGH analysis with a microarray carrying 2,464 BAC clones to examine genomic aberrations in 236 primary neuroblastomas and found that the gain of chromosome 17q was most strongly correlated with the patient's prognosis (5). The genetic group of 'silent chromosomal aberrations' (GGS) could be defined as the tumor group without apparent abnormalities in chromosome 17, and the genetic group of 'whole chromosomal gains and/or losses' (GGW) as that with gain of whole chromosome 17 (5-year cumulative survival rate in 112 sporadic neuroblastomas: 68 and 80%, respectively, according to ref. 5). On the other hand, the genetic group of 'partial chromosomal gains and/or losses' (GGP) with gain of chromosome 17q showed poor prognosis (43%).

According to the different grade of aggressiveness among the genetic groups, we hypothesized that the GGP tumors may have higher levels of expression of the activated 17q candidate gene(s) that is (are) involved in defining the grade of malignancy of neuroblastoma than the GGS or GGW tumors. We then used our data set of gene expression profile in 136 neuroblastomas to subtract the genes mapped to the commonly gained region of chromosome 17q and differentially expressed in the GGP tumors at high levels and the GGS or GGW tumors at low levels. Consequently, we found two cDNA clones *Nbla10727* and *Nbla12061* (Fig. 1A) on our in-house microarray carrying 5,340 cDNAs obtained from oligo-capping cDNA libraries generated from different subsets of primary neuroblastomas (10,14-16), both of which were splicing variants of the same gene mapped to chromosome 17q25.1 (Table I and Fig. 1B, expression in GGP more than that in GGS or GGW). Database searching showed that both 2,087-bp and 2,186-bp insert sequences (Genbank/DDBJ accession numbers: AB447886 and AB447887) did not exhibit significant similarity to any previously known genes. As the size of mRNA was ~2.3 kb by Northern blot (Fig. 1C), the clones *Nbla10727* and *Nbla12061* appeared to be almost full-length cDNAs. Therefore, *Nbla10727*/12061 appeared to be the gene activated for its expression in neuroblastomas with partial gain of chromosome 17q, but not activated in those with diploid or triploid pattern of whole chromosome 17.

The *Nbla10727*/12061 gene was expressed in multiple human tissues with preferential expression in heart, kidney, lung, spleen, mammary gland, prostate and liver, but with low expression in neuronal tissues such as brain and cerebellum, fetal brain and adrenal gland (Fig. 1D).

High expression of Nbla10727/12061 is associated with poor prognosis of neuroblastoma. The analysis by semi-quantitative RT-PCR in a panel of cDNAs obtained from 8 favorable (stage 1, <1-year-old, single copy of *MYCN* and high expression

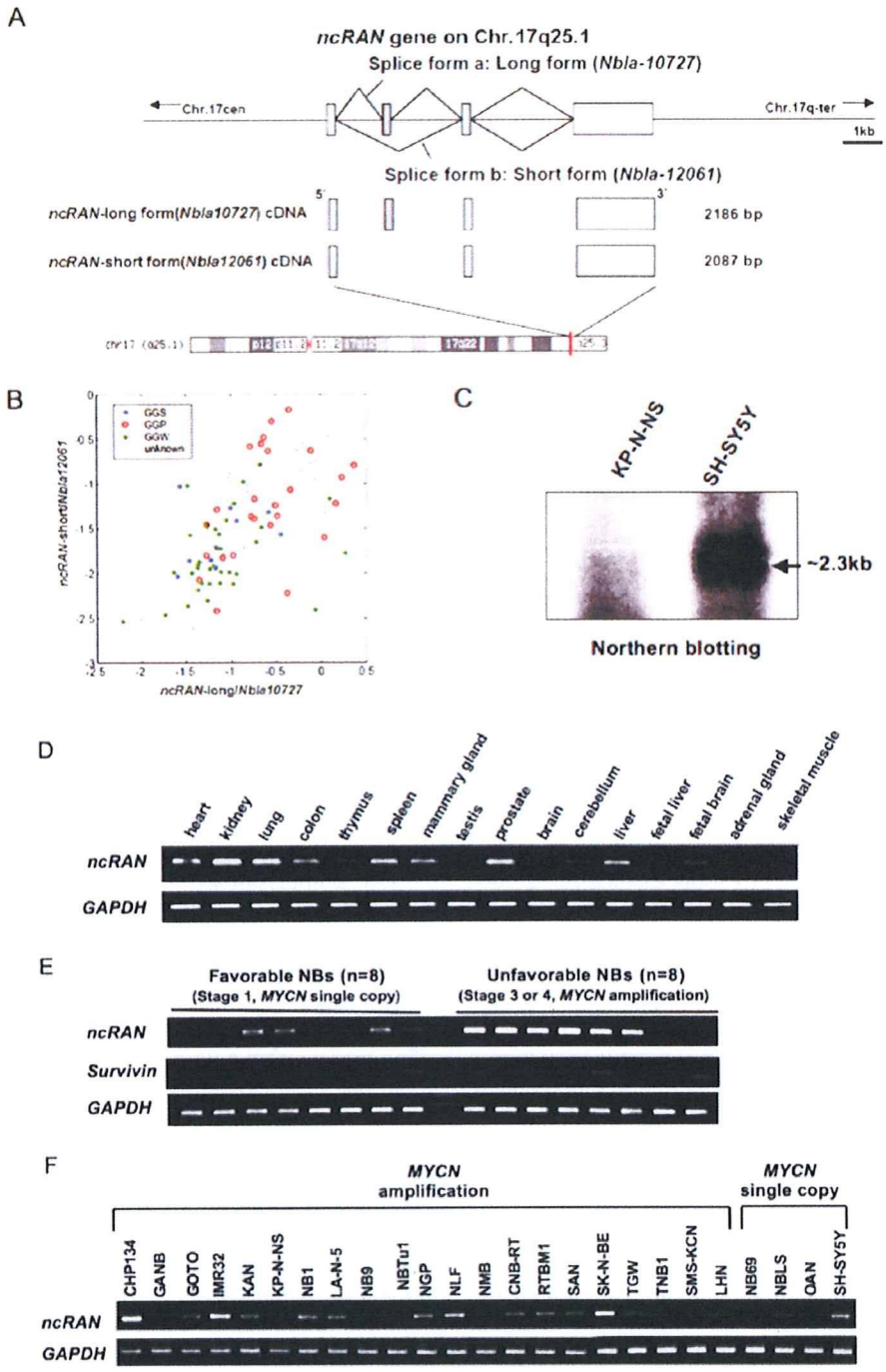

Figure 1. *ncRAN* is mapped to the 17q gain region. A, Genomic structure of *ncRAN* region on chromosome 17q25.1. Splicing variants, whose sequences were contained in cDNAs as *ncRAN-long/Nbla10727* and *ncRAN-short/Nbla12061*, are schematically shown. These are transcribed from a single gene, *ncRAN* (see text). B, High expression of *ncRAN* is associated with high malignant subset of neuroblastoma. Scatter plot of the expression levels of the *ncRAN-long/Nbla10727* and *ncRAN-short/Nbla12061* in 71 primary neuroblastomas with both accompanying expression and aCGH data. Blue, red, green, and black spots denote GGS, GGP, GGW and unknown genomic group samples, respectively. As shown in Table I, the expression levels of the *ncRAN* were significantly higher in GGP tumors (+17q gain) than in GGS (no 17 gain) or GGW (+ whole 17 gain) tumors ($p=0.004$ and $p<0.001$ for *ncRAN-long/Nbla10727*, and $p=0.070$ and $p<0.001$ for *ncRAN-short/Nbla12061*, respectively), whereas their expression levels in GGS and GGW tumors were comparable ($p=0.952$ for *ncRAN-long/Nbla10727*, and $p=0.163$ for *ncRAN-short/Nbla12061*, see also Table I), suggesting that the acquired allele(s) at 17q might be silenced at least for the *ncRAN* expression in GGW tumors, and that high expression of *ncRAN* is associated with high malignant subset of neuroblastoma. C, Northern blot analysis of *ncRAN*. Total RNA (20 μg) prepared from neuroblastoma cell lines, SH-SY5Y and KP-N-NS were used. A 2.3-kb band was visible in only SH-SY5Y cells. The cDNA insert (*Nbla10727*) was labeled with [α-^{32}P]-dCTP and used for the hybridization probe. D, Semiquantitative RT-PCR of *ncRAN* in multiple human tissues and neuroblastoma cell lines. Total RNA of 25 adult tissues and two fetal tissues were purchased from Clontech Co. Ltd. The expression of *GAPDH* is also shown as a control. E, Semi-quantitative RT-PCR of *ncRAN* in favorable and unfavorable subsets of primary neuroblastomas. The mRNA expression patterns for *ncRAN* and *Survivn*, a known oncogene identified at 17q, were detected by semi-quantitative RT-PCR procedure in eight favorable (lanes: 1-8, stage 1, with a single copy of *MYCN*) and eight unfavorable (lanes: 9-16, stage 3 or 4, with *MYCN* amplification) neuroblastomas. F, Semiquantitative RT-PCR of *ncRAN* in neuroblastoma cell lines. Twenty-one neuroblastoma cell lines with *MYCN* amplification and 4 cell lines with a single copy of *MYCN* were used for this study as templates.

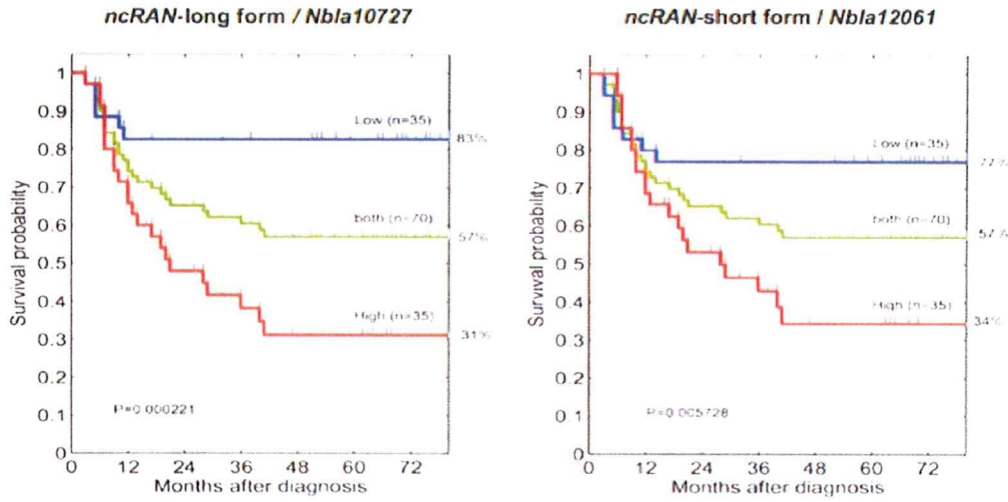

Figure 2. The high expression of *ncRAN/Nbla10727/12061* mRNA is a prognostic indicator of unfavorable neuroblastomas. The Kaplan-Meier survival curves were drawn from the results of the cDNA microarray data of 70 sporadic neuroblastomas (log-rank test, p=0.000221 and p=0.005728, respectively).

of *TrkA*) and 8 unfavorable (stage 3 or 4, >1-year-old, amplified *MYCN* and low expression of *TrkA*) primary neuroblastomas confirmed that this novel gene was expressed at significantly high levels in the latter compared to the former (Fig. 1E), such as *Survivin* which we have previously reported as one of the candidate genes mapped at the region of 17q gain (9). Among neuroblastoma cell lines, high or moderate levels of expression of *Nbla10727/12061* was observed in cell lines with *MYCN* amplification most of which had 17q gain, whereas it was relatively low in those with a single copy of *MYCN* and without the 17q gain (Fig. 1F).

As shown in Fig. 2, our microarray data of 70 sporadic neuroblastomas showed that the high levels of *Nbla10727/12061* expression were significantly associated with poor prognosis (log-rank test, p=0.000221 and p=0.005728, respectively). The multivariate analysis using Cox proportional hazard model demonstrated that expression of *Nbla10727/12061* was an independent prognostic factor among age at diagnosis, disease stage, tumor origin and *MYCN* expression (Table II). Thus, the expression level of *Nbla10727/12061* is a novel prognostic factor of neuroblastoma that is closely associated with gain of chromosome 17q.

Nbla10727/12061 is involved in inducing enhancement of cell growth in neuroblastoma cells and transformation of NIH3T3 cells. To investigate function of *Nbla10727/12061*, we transfected SH-SY5Y neuroblastoma cells with the siRNA, since SH-SY5Y cells have 17q gain in their genome as well as higher mRNA expression of *Nbla10727/12061*. As shown in Fig. 3A, suppression of endogenous levels of *Nbla10727/12061* transcripts significantly inhibited cell growth in SH-SY5Y neuroblastoma cells as compared with the control cells. On the other hand, the soft agar colony formation assay showed that the enforced expression of *Nbla10727/12061* significantly enhanced the anchorage-independent growth of NIH3T3 mouse fibroblast cells (Fig. 3B). These results suggested that *Nbla10727/12061* was a novel candidate gene of the region of 17q gain with an oncogenic function.

ncRAN-Nbla10727/12061 is a large non-coding RNA. Several lines of evidence from the gene structure analysis as well as the comparative genomic analysis described below further suggested that *Nbla10727/12061* is a non protein-coding but functional RNA. We therefore tentatively named this gene as *ncRAN* (non-coding RNA expressed in aggressive neuroblastoma).

First, the full-length cDNA sequences of *ncRAN*, which are suggested to be relevant to both *Nbla10727* and *Nbla12061* cDNAs by Northern blot analysis (Fig. 1C), did not contain any long-enough open reading frames (>200 bp). Bioinformatic analysis indicated that there were no ESTs longer than those two cDNAs at the genomic locus, and that the CpG island was located at the 5' region of the cDNA sequences.

Second, no protein product was translated both *in vivo* and *in vitro* from the *ncRAN* transcripts (Fig. 4). Though only the possible open reading frames (>150 bp) within the *ncRAN* cDNA were from n.t. 190 to 354 (55 amino acids) and from 293 to 469 (59 amino acids) in *Nbla10727*, none of the putative translation start sites contains the Kozak consensus sequence. In addition, these predicted protein products of 55 and 59 amino acids did not exhibit significant similarity to any other known protein or protein domain. Furthermore, *in vivo* transcription and translation of the full-length *ncRAN* did not lead to the synthesis of any peptide or protein (Fig. 4B), though endogenously and ectopically expressed *ncRAN* were easily detectable at mRNA level (Fig. 4A). Coincident with the above observation, the *ncRAN* protein product could not be detected using [^{35}S]-methionine-labeling system *in vitro* (Fig. 4C).

Third, we performed sequence comparison of the *ncRAN* gene with genome sequences of other species and found it has high similarity (>90% identity in nucleotides) with primates including orangutan, chimpanzee and rhesus, but not those with mice and rat (Fig. 5). We also searched for the possible long open reading frames of *ncRAN* homologs in these highly similar species, resulting in failure. The highly conserved sequence similarity only with primates may