STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:1106–1120
Published online 6 August 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2997

Clustering of significant genes in prognostic studies with microarrays: Application to a clinical study for multiple myeloma

Shigeyuki Matsui[1,2,*,†], Takeharu Yamanaka[3], Bart Barlogie[4], John D. Shaughnessy Jr[4] and John Crowley[5]

[1]*Department of Pharmacoepidemiology, School of Public Health, Kyoto University, Yoshida Konoe-cho, Sakyo-ku, Kyoto, Japan*
[2]*Translational Research Informatics Center, Foundation for Biomedical Research and Innovation, Minatojima-minami-machi, Chuo-ku, Kobe, Japan*
[3]*Cancer Statistics Laboratory, Institute for Clinical Research, National Kyushu Cancer Center, Fukuoka, Japan*
[4]*The Myeloma Institute for Research and Therapy, University of Arkansas for Medical Science, Little Rock, AR, U.S.A.*
[5]*Cancer Research and Biostatistics, Southwest Oncology Group Statistical Center, Seattle, WA, U.S.A.*

SUMMARY

When a large number of genes are significant in correlating microarray gene expression data with patient prognosis, clustering of significant genes may be effective not only for further dimension reduction but also for identifying co-regulated genes that belong to the same molecular pathway related to disease biology and aggressiveness. Moreover, a reduced feature, such as the average expression across samples for a cluster of significant genes, can play an important role in reducing variance in prediction analysis. We propose a simple procedure to select gene clusters that have strong marginal association with survival outcome from a large pool of candidate hierarchical clusters of significant genes. Selected gene clusters can have better predictive capability than the other gene clusters and singleton genes. Application of such clustering to the data set from a clinical study for patients with multiple myeloma and associated microarrays is given. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: gene expression; microarrays; prognostic analysis; gene clustering; multiple myeloma

*Correspondence to: Shigeyuki Matsui, Department of Pharmacoepidemiology, School of Public Health, Kyoto University, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan.
†E-mail: matsui@pbh.med.kyoto-u.ac.jp

Received 6 November 2006
Accepted 30 May 2007

1. INTRODUCTION AND BACKGROUND

Multiple myeloma is a malignancy of the plasma cells of the bone marrow and is the second most common hematological malignancy. In spite of substantial progress in therapeutics, the outcome for patients requiring therapy is still highly variable. Although the median survival for patients with standard treatment is from 30 to 36 months, the range is from less than a few months to greater than 10 years. This variability derives from heterogeneity in both myeloma cell biology and multiple host factors. Knowledge of tumor and host factors associated with prognosis is critical for understanding disease biology, identifying risk groups, and optimizing patient treatment [1].

Although several staging systems based on clinical and laboratory tests have been developed for multiple myeloma [1, 2], standard prognostic factors, such as β2-microglobulin, albumin, and C-reactive protein, account for only 15–20 per cent of outcome heterogeneity [3]. Recently, conventional cytogenetics has emerged as a relevant prognostic factor in myeloma patients. Deletion of chromosome 13 is the most common and the most significant prognostic abnormality observed [1, 4, 5]. In spite of recent advances in molecular cytogenetics, many undefined genetic abnormalities may exist, owing to the genetic complexity in multiple myeloma.

The high-throughput DNA microarrays that allow the simultaneous measurement of the level of expression for thousands of genes, or even an entire genome, from human plasma cells are anticipated to reveal a deeper understanding of the molecular biology and clinical behavior of the disease. In an effort to identify genes linked to an aggressive clinical course, RNA from highly purified plasma cells derived from 351 newly diagnosed patients with multiple myeloma at the Myeloma Institute for Research and Therapy, University of Arkansas, was applied to Affymetrix (Santa Clara, CA) U133Plus2.0 microarrays [6, 7]. Shaughnessy *et al.* [7] correlated expression extremes of approximately 54 000 genes with disease-related and overall survival after high-dose chemotherapy supported by autologous stem cell transplantation. Specifically, for each gene, logrank tests were performed for quartile 1 *versus* quartiles 2–4 and quartile 4 *versus* quartiles 1–3 in the expression level, in order to identify under- and over-expressed prognostic genes, respectively.

As an alternative to the analysis by Shaughnessy *et al.* [7] for this data set, we tried to detect genes globally correlated with disease-related survival. To be robust to outliers in gene expression data when performing extremely large numbers of statistical tests, as many as the number of genes, we ranked expression levels across all 351 patients for each gene, and adopted a logrank test for this ranked expression data [8, 9] for each gene. The test statistic corresponds to the average of the (two-sample) logrank statistics with respect to all possible cut-off points (including quartiles 1 and 4 used in Shaughnessy *et al.* [7]) to divide the entire samples into two groups [9]. To obtain an estimate of the false discovery rate (FDR) [10] as the expected proportion of errors among the rejected hypotheses for a given P-value cut-off for the two-sided logrank tests, we adopted a multivariate permutation procedure that randomly matches the survival data with the gene expression data. Table I summarizes the estimated FDR for various P-value cut-offs. For the estimated FDR of 5 per cent (the P-value cut-off of 0.11 per cent), 1173 genes were significant: over-expression of 644 genes and under-expression of 529 genes were linked to short survival.

When confronted with over one thousand significant genes, clustering of significant genes would be useful not only for further dimension reduction but also possibly for identifying co-regulated genes that belong to the same molecular pathway related to disease biology and aggressiveness. Several authors have considered clustering of genes, by hierarchical clustering [11], k-means clustering [12], and model-based clustering [13, 14]. In subsequent prediction analyses, use of the

DOI: 10.1002/sim

can reflect poor structure near the top of the tree. However, the risk for clustering on noise is mitigated by the restriction to relatively small clusters in agglomerative clustering. In contrast, for divisive hierarchical clustering that operates by partitioning clusters starting with the complete set of data, small clusters may suffer from the risk more seriously. Partitioning clustering such as k-means clustering is another possible choice for a given number of clusters and initial partitions of genes. However, a limitation of partitioning clustering is that the initial inputs can greatly impact the final result. Such a property is likely to be detrimental when coping with a large number of relatively small clusters.

The clustering procedure requires specification of both a pair-wise dissimilarity or a distance metric between genes and a linkage. For a distance metric, we have chosen the Euclidean distance for ranked expression levels across samples so as to be robust to outliers in high-dimensional gene expression data. Note that this choice is essentially the same as choosing one minus Spearman correlation coefficient for the original expression data without the rank transformation. With respect to the linkage method, complete linkage tends to yield smaller clusters with higher correlations among the component genes compared with single linkage, but, because of the small group sizes, the variance reduction might not be as large as in the case of the single linkage. Average linkage as a compromise is thus suggested [15].

The selection procedure starts with the calculation of the ranked average gene expression within clusters for each of all hierarchical clusters of gene with size $\leqslant 30$. We perform the univariate logrank test of no association between the ranked average gene expression within-cluster and disease-related survival. We select the cluster with the greatest chi-square statistic (or smallest two-sided P-value) as the first cluster. Before the next selection, we delete all the clusters containing gene(s) in the first cluster (or delete all the clusters that have any overlap in the membership of genes with the first cluster) from the list of candidate clusters. We then select the cluster with the greatest chi-square statistic as the second cluster from the remaining clusters in the list of candidate clusters. Before the next selection, we delete all the clusters containing gene(s) in the second cluster from the list of candidates. The third and subsequent clusters can be similarly identified. As such, we identify disjoint clusters of genes (no overlapped genes across top clusters) that are strongly associated with the disease-related survival outcome.

The procedure provides candidate clusters of informative genes. Suppose that a gene cluster T in hierarchical clustering is truly informative in terms of disease biology and aggressiveness. When a larger gene cluster that involves T is 'erroneously' selected as one of the top clusters, the cluster T is eliminated from the remaining selection pool. However, this would not be so serious because all genes in T are detected *via* the selected cluster and there still remains the possibility that T is identified by further investigation of all genes in the selected cluster. Meanwhile, when a sub-cluster in T, say S, is erroneously selected at a selection round, T is eliminated from the selection pool. However, the sub-clusters in T that are composed of genes other than those in S still remain in the selection pool. If these sub-clusters are selected in subsequent selection rounds, the possibility that T is identified may be enhanced through inspecting the relationship among top clusters in hierarchical clustering.

3. THE CHOICE OF PARAMETERS: SCREENING OR PREDICTION?

The choice of the three parameters, the number of significant genes, the restriction of cluster size, and the number of top clusters, depends on the strategy for developing prognostic markers.

DOI: 10.1002/sim

to 0.96 (the median of correlations between all the 1173 significant genes was 0.00 with a range of −0.58 to 0.99), which indicates that genes within top clusters tended to be positively correlated, as one expects, but the strength of the correlations was generally not so high. There was no clear trend in association between the strength of within-cluster correlation and the ranking in top clusters. Again, cluster size is restricted to be equal to or less than 30. Because derivation of the general criterion for variance reduction in prediction seems to be intractable for censored survival outcome, we alternatively assessed the strength of association between average gene expressions within gene clusters and survival outcome, which was compared with that between gene expressions of singleton genes and survival outcome.

Figure 1(a) shows the distribution of a chi-square statistic of the univariate logrank test for top 30 clusters, for the other 986 ($=1016-30$) clusters not selected as top clusters, and for the 1173 significant singleton genes. Here, ranked average expression from gene clusters or ranked expression from singletons across samples was associated with the disease-related survival outcome. Under the univariate Cox regression using a single ranked covariate R with the structural component of proportional hazard of $\exp(\beta R)$ for a particular patient, where β is the log hazard ratio associated with the unit increase in R, the non-centrality parameter of the asymptotic chi-square logrank statistic may be expressed as $D\beta^2\sigma^2$, where D is the expected number of events and σ^2 is the variance of R across patients [19]. Because D and σ^2 are common for all gene clusters and singletons, the differences in the chi-square statistic among gene clusters and singletons represent those in β^2, a measure of strength of association with survival outcome, among gene clusters and singletons. As one expects from the property of the selection procedure, the chi-square statistics for the top 30 clusters tended to be greater than those for the other 986 clusters. We also note that the chi-square statistics for the 986 clusters tended to be greater than those for the 1173 singleton genes. Gene clusters had greater β^2 and probably greater predictive ability than singletons. This could be explained by the variance reduction by averaging expressions within gene cluster. See Section 5 for comparison of chi-squares among gene clusters and singletons for an independent test data set.

For some of the top clusters, we observed over-representations of particular chromosomes. Table II shows the chromosome distribution for genes in each of the top five clusters and in each of the 14th, 18th, 21st, 24th, 26th and 30th clusters, with the relative frequency for a particular chromosome being greater than 50 per cent. For example, over-representation of chromosome 1, which was reported by the previous analysis of the same data set by Shaughnessy *et al.* [7], was observed for the 1st, 14th, 24th, 26th, and 30th clusters. As a new finding, over-representation of other chromosomes was also observed — for example, over-representation of chromosome 5 for the 18th cluster and that of chromosome 13 for the 21st cluster. Note that the deletion of chromosome 13 is a prognostic cytogenetic abnormality in myeloma [1, 4, 5]. The average gene expression for the 21st cluster had a significant correlation with chromosome 13 deletion by fluorescence *in situ* hybridization (FISH), where the chromosome 13 deletion was considered positive when >80 per cent of clonal plasma cells had only one signal (the chi-square of a Wilcoxon two-sample test to compare the average gene expression within cluster between positive and negative cases was 126.3). The 21st cluster associated with both the disease-related survival outcome and chromosome 13 deletion provides some insight for understanding the biological mechanism related to progression of the disease. The average gene expression levels within the top 30 clusters also had significant links to seven molecular subgroups of myeloma defined by unsupervised hierarchical clustering of myeloma cases [20]. The identified gene clusters have the potential to reflect co-regulated genes, arising from alternation of specific signaling pathways and/or changes in DNA copy number,

DOI: 10.1002/sim

Table II. Chromosome distribution for top clusters: (a) the top 5 cluster and (b) the 14, 18, 21, 24, 26, and 30th cluster.

		Top clusters				
Chromosome	U133Plus2.0*	1st $\chi^2=62.7^{\dagger}$ Over‡	2nd $\chi^2=62.7$ Under	3rd $\chi^2=59.2$ Over	4th $\chi^2=57.4$ Over	5th $\chi^2=53.5$ Under
(a)						
1	5379 (10.0)	6 (23.1)	0	1 (8.3)	1 (6.3)	3 (10.7)
2	3958 (7.3)	0	0	1 (8.3)	0	0
3	3275 (6.1)	0	0	0	3 (18.8)	0
4	2314 (4.3)	1 (3.9)	3 (17.7)	1 (8.3)	0	0
5	2615 (4.8)	0	1 (5.9)	1 (8.3)	0	0
6	2956 (5.5)	3 (11.5)	0	0	0	2 (7.1)
7	2769 (5.1)	1 (3.9)	1 (5.9)	1 (8.3)	5 (31.3)	0
8	2014 (3.7)	0	0	0	1 (6.3)	3 (10.7)
9	2139 (4.0)	0	0	0	0	0
10	2192 (4.1)	0	1 (5.9)	0	1 (6.3)	2 (7.1)
11	2889 (5.4)	1 (3.9)	5 (29.4)	0	2 (12.5)	4 (14.3)
12	2739 (5.1)	0	0	1 (8.3)	0	1 (3.6)
13	1250 (2.3)	0	0	0	0	0
14	1793 (3.3)	1 (3.9)	0	0	0	0
15	1805 (3.3)	1 (3.9)	1 (5.9)	0	0	0
16	2084 (3.9)	1 (3.9)	1 (5.9)	0	0	2 (7.1)
17	2843 (5.3)	4 (15.4)	0	2 (16.7)	0	8 (28.6)
18	966 (1.8)	1 (3.9)	1 (5.9)	1 (8.3)	0	0
19	2839 (5.3)	1 (3.9)	0	0	1 (6.3)	2 (7.1)
20	1487 (2.8)	0	0	0	0	1 (3.6)
21	622 (1.2)	1 (3.9)	0	0	1 (6.3)	0
22	1225 (2.3)	0	0	0	1 (6.3)	0
X	1691 (3.1)	0	0	0	0	0
Y	107 (0.2)	0	0	0	0	0
Other	684	4 (15.4)	3 (17.7)	3 (25.0)	0	0
Total	54 675	26	17	12	16	28

		Top clusters					
Chromosome	U133Plus2.0*	14th $\chi^2=42.0^{\dagger}$ Over‡	18th $\chi^2=40.2$ Under	21st $\chi^2=38.2$ Under	24th $\chi^2=37.9$ Over	26th $\chi^2=36.7$ Over	30th $\chi^2=35.3$ Over
(b)							
1	5379 (10.0)	6 (85.7)	2 (6.7)	0	8 (100.0)	5 (100.0)	9 (81.8)
2	3958 (7.3)	1 (14.3)	0	1 (3.3)	0	0	0
3	3275 (6.1)	0	1 (3.3)	0	0	0	0
4	2314 (4.3)	0	0	0	0	0	0
5	2615 (4.8)	0	18 (60.0)	0	0	0	0
6	2956 (5.5)	0	0	0	0	0	0
7	2769 (5.1)	0	0	0	0	0	0
8	2014 (3.7)	0	0	0	0	0	0
9	2139 (4.0)	0	1 (3.3)	1 (3.3)	0	0	0
10	2192 (4.1)	0	0	0	0	0	1 (9.1)

DOI: 10.1002/sim

Table III. Reproducibility of the top clusters in Table II in the sensitivity analyses.

Top clusters	680 significant genes	2050 significant genes
1st (26 genes)	10,* 10 (=4+2+1+1)†	26, 12 (=5+5+2)
2nd (17 genes)	8, 8 (=6+2)	17, 11 (=5+4+1+1)
3rd (12 genes)	7, 7 (=4+1+1+1)	12, 10 (=4+2+2+2)
4th (16 genes)	8, 8 (=7+1)	16, 16 (=11+3+1+1)
5th (28 genes)	19, 19 (=6+4+3+3+2+1)	28, 25 (=8+7+5+3+2)
14th (7 genes)	6, 6 (=4+1+1)	7, 3 (=3)
18th (30 genes)	20, 19 (=14+2+1+1+1)	30, 0
21st (30 genes)	13, 13 (=10+3)	30, 2 (=2)
24th (8 genes)	7, 7 (=5+1+1)	8, 4 (=4)
26th (5 genes)	5, 5 (=5)	5, 5 (=5)
30th (11 genes)	6, 6 (=3+2+1)	11, 6 (=3+2+1)

*The number of genes that were included in the set of (680 or 2050) significant genes.

†The total number of genes that were included in the top (50 or 31) clusters, followed by the number parentheses that represent the number of genes included in respective top clusters. For example, when using the set of 680 significant genes, of the 30 genes in the 18th cluster in Table II, 20 were included in the set of significant genes. After performing the selection procedure, of the 20 genes, 19 genes were included in five of the top clusters. The respective clusters included 14, 2, 1, 1, and 1 gene.

the original top clusters in Table II. The 18th cluster was not selected at all, while the 26th cluster was entirely reproduced without deletion or addition of genes. As such, the sensitivity analyses provide information on how selected clusters are reproduced. One may place more confidence on well-reproduced clusters. For identifying clusters of informative genes, it is advisable to trace each gene in clustering results in the series of sensitivity analyses.

The impact by mitigating the restriction on cluster size was generally small. For the cluster size restriction of 60, instead of 30, we selected the top 21 clusters (472 genes) for the upper limit of 500 genes for further investigation. Out of the original top 30 clusters (484 genes) for the cluster size restriction of 30, 24 clusters (425 genes) were reselected and the other six clusters (59 genes), 22nd, 23rd, 24th, 26th, 28th, and 30th, were not reselected in the 21 top clusters. Of the 24 clusters, 14 clusters were entirely reselected as one of the 21 top clusters without deletion or addition of genes. Meanwhile, the other four clusters formed two clusters in the top 21 clusters, and the rest six clusters formed four clusters in the top 21 clusters with other gene clusters that were not selected when the cluster size restriction was 30.

5. PREDICTIVE PERFORMANCE OF SELECTED CLUSTERS

We attempted to assess how the top 30 gene clusters selected from the 1173 significant genes for all 351 patients are predictive, as given by secondary analysis for the gene selection analysis described in Section 4. We divided the 351 patients into 234 patients for training and 117 patients for validation (the division ratio = 2:1) by a block randomization. The three patients, P_1, P_2, and P_3, say, with the longest observed survival times (failure or censored cases) formed a block with a size of 3. The block (P_1, P_2, P_3) was assigned randomly to one of (1, 1, 2), (1, 2, 1), and (2, 1, 1), where 1 corresponds to assignment to the training set and 2 to the validation set. Another block of three patients with the next longest survival times was similarly assigned and so forth. For

DOI: 10.1002/sim

Table IV. Multivariate Cox regression analysis with the five prognostic factors and the expression-based risk classes.

Variable	Hazard ratio	95 per cent confidence interval	*P*-value
Beta-2-microglobulin ⩾4.0 mg/L	2.08	0.94–4.59	0.0693
Albumin <3.5 g/dL	0.96	0.37–2.49	0.9319
Creatinine ⩾2.0 mg/dL	2.70	1.05–6.94	0.0398
Cytogenetic abnormality	0.98	0.49–1.95	0.9551
FISH-defined chromosome 13 deletion	2.89	1.43–5.85	0.0032
Expression-based risk classes			
Intermediate *versus* low	3.67	1.38–9.79	0.0093
High *versus* low	3.66	1.42–9.44	0.0074

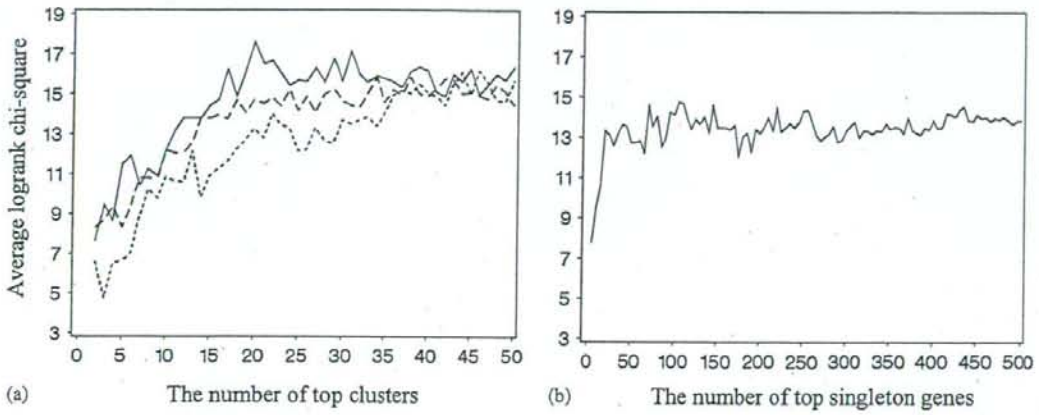

Figure 2. The average logrank chi-squares for the validation set over the 10 repetitions of prediction analysis. For cluster-based prediction labeled (a), the average chi-squares were calculated for each combination of the numbers of significant genes, 680 (solid line), 1173 (dashed line), or 2050 (dotted line), and the numbers of top clusters. For gene-based prediction labeled (b), the average chi-squares were calculated for each number of top singleton genes used for prediction.

Here, we considered the number of significant genes as 680, 1173, or 2050, which corresponds to an FDR of 2.5, 5.0 or 10.0 per cent, respectively, for all 351 patients. We changed the number of top clusters from 1 to 50. Figure 2(a) shows the average of the logrank chi-squares for association between the three risk classes and survival outcome (d.f. = 2) for the validation set over the 10 repetitions for each combination of the two parameters. First, we note that a high chi-square value such as 15 can be obtained for various combinations of the parameters. However, the chi-squares for the smallest number of significant genes, 680, were generally highest, which indicates the possibility that smaller numbers of significant genes than 680 are expected to yield higher chi-squares. For the number of significant genes of 680 and 1173, the chi-square reached a peak for the number of top clusters of around 20.

DOI: 10.1002/sim

regression [11] and application of the lasso [29] to a unique set of genes and cluster features of the average gene expression within clusters at each level of the hierarchy in hierarchical clustering [15]. These approaches select clusters (and singleton genes) through a prediction model building. However, forward stepwise regression is a greedy algorithm [30], and such greediness is likely to be detrimental to the predictive performance for high-dimensional microarray data [16, 31]. More importantly, subset selection can be extremely variable, which would be a drawback to gene screening. The lasso can also suffer from large variability in subset selection by changing the lasso bound [22]. On the other hand, subset selection in our approach will be more stable because it is based on the marginal association with the outcome variable. A point of concern for our selection procedure, pointed out by an anonymous reviewer, the procedure is greedy in a different sense that, each time a gene cluster is chosen and all the other overlapping clusters are eliminated, it narrows down the selection pool in a greedy way. In gene screening, a backup inspection of the relationship among top clusters in hierarchical clustering would be effective for detecting clusters of informative genes as described in Section 2. For prediction, while selected clusters were demonstrated to be good predictors in Section 5, modification to lessen the greediness for improving predictive accuracy is worthwhile and thus is a subject for future research.

REFERENCES

1. Greipp PR, San Miguel J, Durie BG, Crowley JJ, Barlogie B, Blade J, Boccadoro M, Child JA, Avet-Loiseau H, Kyle RA, Lahuerta JJ, Ludwig H, Morgan G, Powles R, Shimizu K, Shustik C, Sonneveld P, Tosi P, Turesson I, Westin J. International staging system for multiple myeloma. *Journal of Clinical Oncology* 2005; **23**:3412–3420.
2. Durie BG, Salmon SE. A clinical staging system for multiple myeloma. Correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival. *Cancer* 1975; **36**:842–854.
3. Harousseau JL, Shaughnessy JD, Richardson P. Multiple myeloma. *Hematology, American Society of Hematology Education Program Book* 2004; 237–256.
4. Shaughnessy JD, Barlogie B. Interpreting the molecular biology and clinical behavior of multiple myeloma in the context of global gene expression profiling. *Immunological Reviews* 2003; **194**:140–163.
5. Stewart AK, Fonseca R. Prognostic and therapeutic significance of myeloma genetics and gene expression profiling. *Journal of Clinical Oncology* 2005; **23**:6339–6344.
6. Shaughnessy JD, Barlogie B. Using genomics to identify high-risk myeloma after autologous stem cell transplantation. *Biology of Blood and Marrow Transplantation* 2006; **12**:77–80.
7. Shaughnessy JD, Zhan F, Burington BE, Huang Y, Colla S, Hanamura I, Stewart JP, Kordsmeier B, Randolph C, Williams DR, Xiao Y, Xu H, Epstein J, Anaissie E, Krishna SG, Cottler-Fox M, Hollmig K, Mohiuddin A, Pineda-Roman M, Tricot G, van Rhee F, Sawyer J, Alsayed Y, Walker R, Zangari M, Crowley J, Barlogie B. A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* 2007; **109**:2276–2284.
8. O'Quigley J, Prentice RL. Nonparametric tests of association between survival time and continuously measured covariates: the logit-rank and associated procedures. *Biometrics* 1991; **47**:114–127.
9. Jung SH, Owzar K, George SL. A multiple testing procedure to associate gene expression levels with survival. *Statistics in Medicine* 2005; **24**:3077–3088.
10. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 1995; **57**:259–300.
11. Hastie T, Tibshirani R, Botstein D, Brown P. Supervised harvesting of expression trees. *Genome Biology* 2001; **2**:0003.1–0003.12.
12. Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, Bild A, Iversen ES, Liao M, Chen CM, West M, Nevins JR, Huang AT. Gene expression predictors of breast cancer outcomes. *Lancet* 2003; **361**:1590–1596.
13. McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 2002; **18**:413–422.
14. McLachlan GJ, Do KA, Ambroise C. *Analyzing Microarray Gene Expression Data*. Wiley: Hoboken, NJ, 2004.
15. Park MY, Hastie T, Tibshirani R. Averaged gene expressions for regression. *Biostatistics* 2007; **8**:212–227.

BIOMETRICS 64, 217–226
March 2008

DOI: 10.1111/j.1541-0420.2007.00875.x

Sample Size Calculations Based on Ranking and Selection in Microarray Experiments

Shigeyuki Matsui,[1,2,*] Shu Zeng,[3] Takeharu Yamanaka,[4] and John Shaughnessy[5]

[1]Department of Pharmacoepidemiology, School of Public Health,
Kyoto University, Yoshidakonoe-cho, Sakyo-ku, Kyoto 606-8501, Japan
[2]Translational Research Informatics Center, Foundation for Biomedical Research and Innovation,
1-5-4 Minatojima-minami, Chuo-ku, Kobe 650-0047, Japan
[3]Insilico Technologies Incorporation, 22-30 Okuike-cho, Ashiya, 650-0003, Japan
[4]Cancer Biostatistics Laboratory, Institute for Clinical Research, National Kyushu Cancer Center,
3-1-1 Notame, Minami-ku, Fukuoka 811-1395, Japan
[5]The Myeloma Institute for Research and Therapy, University of Arkansas for Medical Sciences,
301 West Markham Street, Little Rock, Arkansas 72205, U.S.A.
*email: matsui@pbh.med.kyoto-u.ac.jp

SUMMARY. We develop formulae to calculate sample sizes for ranking and selection of differentially expressed genes among different clinical subtypes or prognostic classes of disease in genome-wide screening studies with microarrays. The formulae aim to control the probability that a selected subset of genes with fixed size contains enough truly top-ranking informative genes, which can be assessed on the basis of the distribution of ordered statistics from independent genes. We provide strategies for conservative designs to cope with issues of unknown number of informative genes and unknown correlation structure across genes. Application of the formulae to a clinical study for multiple myeloma is given.

KEY WORDS: Gene expression; Microarrays; Ranking; Sample size; Selection.

1. Introduction

The advent of microarray technology allows us to perform genome-wide screening of differentially expressed genes among different clinical subtypes or prognostic classes of disease. Such genome-wide screening studies generate large amounts of data and warrant careful planning, including determination of the number of samples. Most methods for sample size calculation recently published in the microarray literature are based on statistical significance of the difference in gene expression (Dobbin and Simon, 2005; Hu, Zou, and Wright, 2005; Jung, 2005; Jung, Bang, and Young, 2005; Li et al., 2005; Pawitan et al., 2005; Tsai et al., 2005). Although multiple statistical tests are frequently used at the analysis stage, the primary outcome relevant to screening purpose would be the rankings of genes with regards to a statistical measure of differential expression, such as the test statistic. Typically, top genes are selected for further investigation in subsequent studies, possibly using technically simpler, but more reliable assays such as polymerase chain reaction (PCR) (e.g., Lossos et al., 2004), where the number of selected genes may be prespecified owing to limited resources, irrespective of the number of significant genes in screening studies. For this ranking and selection task, it is warranted to ensure that the most informative, truly top-ranking genes have a high chance of being selected for further investigation. For disease screening, Pepe et al. (2003) proposed sample size calculations to satisfy the requirement on the probability such as that of truly top-ranking genes will rank in a selected subset of genes using the bootstrap from a real dataset.

In this article, we develop formulae for sample size calculations in the framework of ranking and selection (see Wetherill and Ofosu, 1974, for a review for normal populations). The formulae ensure that the most informative genes have a high chance of being selected, which can be assessed on the basis of the distribution of ordered statistics from independent genes. This criterion corresponds to that given in the fixed subset-size approach of Mahamunulu (1967).

The article is organized as follows. After providing the framework of ranking and selection in Section 2, we develop the formulae under the assumption of a single group of informative genes with a common effect size on a clinical phenotype variable in Section 3. We consider the extension to more plausible settings with multiple groups of informative genes with distinct effect sizes in Section 4. We provide strategies for conservative designs to cope with issues of unknown number of informative genes and unknown correlation structure across genes. Application to a clinical study for multiple myeloma is given in Section 5, followed by discussion in Section 6.

2. Gene Ranking and Selection

We suppose that a binary phenotype variable with class 1 or 0, e.g., success or failure for treatment, is observed for n samples. Let n_1 be the number of samples for class 1, so that the number of samples for class 0 is $n - n_1$. The m candidate genes are ranked on the basis of the strength of their association with the phenotype variable using normalized log ratios from two-color arrays or normalized log signals from oligonucleotide arrays. For gene j, let U_j be a standardized statistic to measure association between gene expression levels and the phenotype variable calculated from n samples ($j = 1, \ldots, m$). A commonly used statistic for an association for two classes, a difference in mean expression between two classes, is a Student's t-statistic (e.g., McLachlan, Do, and Ambroise, 2004). Without loss of generality, a negative (positive) value of U_j represents overexpression (underexpression) for class 1. When we are interested in genes overexpressed (underexpressed) for class 1, we select the top K genes with the greatest negative (positive) values for statistic U_j, where K is a prespecified number, which may represent an upper limit of genes that can be investigated in subsequent studies.

3. Sample Size Calculations

Sample size is calculated for one of two selection tasks, i.e., selection of overexpressed genes and selection of underexpressed genes for class 1. In what follows, we restrict the attention to selection of overexpressed genes for class 1, unless otherwise noted. Adaptation for selecting underexpressed genes for class 1 is obvious.

3.1 *Models*

We assume that, of m genes, m_1 genes are informative, i.e., associated with the phenotype variable, while the rest $m_0 (= m - m_1)$ are not informative. The m_1 informative genes are overexpressed for class 1, i.e., changing in the same direction. Underexpressed genes for class 1 are absorbed into noninformative genes. For noninformative genes, we assume that U_j's follow independently the same null distribution with the distribution function, F_0. Note that the assumption of the null distribution for underexpressed genes for class 1 may yield a conservative design because the probability of an underexpressed gene for class 1 being selected is less than the probability of a gene unrelated to the phenotype variable in selecting overexpressed genes for class 1. The impact of this assumption on sample size, however, should be ignorable because the number of underexpressed genes for class 1 would be much smaller than that of genes unrelated to the phenotype variable. For informative genes, we assume U_j's follow independently the same non-null distribution with the distribution function F_1. Let f_0 and f_1 be the corresponding density functions for F_0 and F_1, respectively. In sample size calculations, we have to specify these distributions. A model for the Student's t-statistic is normally distributed expression levels with a common standard deviation between two classes for each gene. For informative genes, the mean difference between two classes and the common standard deviation can vary over genes, but their ratio, i.e., the standardized mean difference, is assumed to be constant, Δ, across genes ($\Delta < 0$ for overexpressed genes for class 1). Under this model, F_0 would be a central t-distribution with degrees of freedom (df) of $n - 2$, while F_1 be a non-central t-distribution with the same df and the non-centrality parameter $\sqrt{\frac{n_1(n-n_1)}{n}}\Delta$.

3.2 *The Probability of Gene Selection*

For fixed n and n_1, we calculate the probability that, of the K selected genes, we have, at least, H informative genes, denoted by $Q_{K,H}$, where $\max(1, K + m_1 + 1 - m) \leq H \leq \min(m_1, K)$. This requirement corresponds to that given in the fixed subset-size approach of Mahamunulu (1967). Let $S_{(1)} < S_{(2)} < \cdots < S_{(m_1)}$ be the ordered statistics of U_j for m_1 informative genes. The distribution and density function of $S_{(r)}$ are given by

$$F_{1,(r)}(s) = \Pr\{S_{(r)} \leq s\} = \sum_{k=r}^{m_1} \binom{m_1}{k} F_1(s)^k \{1 - F_1(s)\}^{m_1 - k},$$

and

$$f_{1,(r)}(s) = \frac{m_1!}{(r-1)!(m_1 - r)!} F_1(s)^{r-1} \{1 - F_1(s)\}^{m_1 - r} f_1(s),$$

respectively (e.g., Cox and Hinkley, 1974, p. 466). Similarly, we consider the ordered statistics, $T_{(1)} < T_{(2)} < \cdots < T_{(m_0)}$, for m_0 noninformative genes. Let $F_{0,(r)}$ and $F_{0,(r)}$ be the distribution and density function of $T_{(r)}$, respectively. Then, $Q_{K,H}$ can be calculated as

$$\begin{aligned} Q_{K,H} &= \Pr\{S_{(H)} < T_{(J)}\} \\ &= \int_{-\infty}^{\infty} F_{1,(H)}(t) f_{0,(J)}(t)\, dt, \end{aligned}$$

based on the distribution function of $S_{(H)}$ where $J = K - H + 1$ (Mahamunulu, 1967). By setting $u = F_0(t)$ in the integration, $Q_{K,H}$ can be expressed as

$$\begin{aligned} Q_{K,H} = {} & \frac{m_0!}{(J-1)!(m_0 - J)!} \\ & \times \int_0^1 F_{1,(H)}\left\{F_0^{-1}(u)\right\} u^{J-1} (1-u)^{m_0 - J}\, du. \end{aligned} \tag{1}$$

Alternatively, $Q_{K,H}$ can be calculated as $Q_{K,H} = \Pr(T_{(J)} > S_{(H)})$ based on the survival function of $T_{(J)}$ (Mahamunulu, 1967). By setting $u = F_1(t)$, we have an alternative formula,

$$\begin{aligned} Q_{K,H} = {} & \frac{m_1!}{(H-1)!(m_1 - H)!} \\ & \times \int_0^1 R_{0,(J)}\left\{F_1^{-1}(u)\right\} u^{H-1} (1-u)^{m_1 - H}\, du, \end{aligned} \tag{2}$$

where

$$R_{0,(r)}(s) = 1 - F_{0,(r)}(s) = \sum_{k=0}^{r-1} \binom{m_0}{k} F_0(s)^k \{1 - F_0(s)\}^{m_0 - k}.$$

3.3 *Sample Size Calculations Based on the Probability of Gene Selection*

We consider hereinafter equal numbers of samples for two classes ($n = 2n_1$), unless otherwise noted (see Section 6 for unequal cases). For a given value for parameters, m, K, H, m_1, and Δ, we can search a minimal sample size that satisfies $Q_{K,H} \geq c$, where c is the acceptable lowest level for $Q_{K,H}$. One may specify a high value such as 0.8 or 0.9 for c, which

Table 1
Expected sample size to satisfy the requirement of $Q \geq c = 0.8$

		$m = 2000$				$m = 5000$			
		$K = 30$		$K = 50$		$K = 30$		$K = 50$	
Δ[a]	m_1	$f = 0.8$	$f = 0.9$	$f = 0.8$	$f = 0.9$	$f = 0.8$	$f = 0.9$	$f = 0.8$	$f = 0.9$
−0.58	10	129	162	110	139	157	192	137	170
	30	166	228	123	162	193	257	151	193
	50	97	123	150	211	118	145	177	241
	100	57	75	71	94	74	92	91	115
−1.00	10	46	57	39	49	56	68	49	60
	30	60	81	44	57	69	92	54	68
	50	36	45	54	75	43	53	63	86
	100	22	29	27	35	28	35	34	43
−2.00	10	14	17	12	14	17	20	15	17
	30	18	24	13	17	21	27	16	20
	50	12	15	16	22	15	17	19	25
	100	9	11	10	12	11	13	12	14

[a]The values of −0.58, −1.00, and −2.00 of Δ correspond to fold change in overexpression of 1.5, 2, and 4, respectively, for class 1 in untransformed data before log transformation with base 2.

is an analogue to the power requirement in designing clinical trials for development of therapeutics.

Determination of the value for the parameters m and K may largely relate to feasibility of studies. With respect to H, we introduce a ratio, f, so that $H = [f \times \min(m_1, K)]$, where the notation $[a]$ denotes the largest integer less than a. The range [0.8, 1.0] would be reasonable for f. Thus, our formulae is to determine sample size to achieve a high value for the ratio, H/m_1, when $m_1 \leq K$ or for the ratio, H/K, when $m_1 \geq K$. The former corresponds to sensitivity, while the latter to true positive rate in the framework of multiple testing (e.g., Tsai et al., 2005). Mahamunulu (1967) considered $f = 1.0$ as cases of special interest.

In determining the effect size Δ, one needs to consider the likely size of the difference between classes one is seeking to detect and the amount of variability between experimental units. A pilot study may be run to estimate these quantities, or else earlier experiments may be used. Also, plausible values should be specified for m_1. When relevant data are available from a pilot study or earlier experiments, we can estimate m_1 or m_0 (Schweder and Spjotvoll, 1982; Benjamini and Hochberg, 2000; Allison et al., 2002; Storey 2002; Hsueh, Chen, and Kodell, 2003; Storey and Tibshirani, 2003). Unfortunately, it is often difficult to obtain reliable estimates for m_1 from limited data. We will provide strategies for conservative designs in Section 3.6.

When both selections, i.e., selection of overexpressed genes and selection of underexpressed genes for class 1, are planned at the analysis stage, one may calculate sample size separately for each selection and take the maximum of the calculated sample sizes, which may generally yield a conservative design for one of the two selections. Note that, for the same specification of design parameters with the same $|\Delta|$, the sample size estimates will be identical for both selections.

3.4 *Illustration*

We supposed situations in which the top 30 or 50 genes were selected from 2000 or 5000 genes. Table 1 summarizes the expected total number of samples to satisfy the requirement of $Q_{K,H} \geq 0.8$ for each of several configurations of parameters. The impacts of Δ, f, and K were substantial. Also, the required sample size largely varied depending on the value of unknown parameter m_1 with the maximum at $m_1 = K$. Figure 1 plots the expected value of the ratio, $H/\min(m_1, K)$, for m_1 for a fixed sample size when $K = 30$. The expected ratio decreases as m_1 increases in the interval $[1, K]$, while it increases as m_1 increases in the interval $[K, \infty]$. For such values of m_1 that give lower values of the expected ratio, more additional samples may be needed to satisfy the requirement on the selection probability. At $m_1 = K$, the expected ratio is minimized, hence the required sample size is maximized.

3.5 *Simulations with Correlated Genes*

The formulae for sample size calculations assume statistical independence of genes and hence may not reflect real data. We

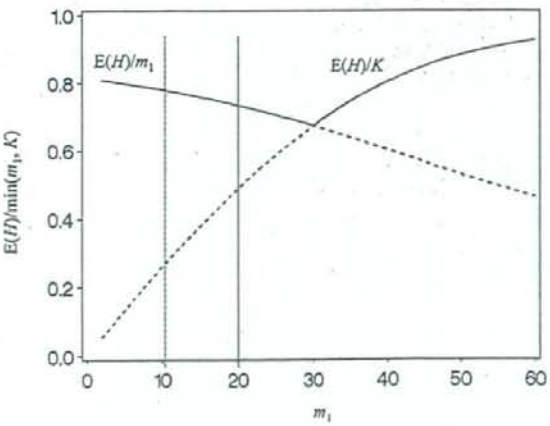

Figure 1. Expected value of $H/min(m_1, K)$ for m_1 for a given sample size: $K = 30$.

evaluated performance of our formulae for correlated genes via simulation. Positively correlated genes have the potential to be coregulated genes in the same molecular pathway, changing in the same direction. On the other hand, negative correlations are almost as likely as positive correlations because a group of genes may inhibit or switch off other genes. Many of such genes would change in the opposite direction ($\Delta > 0$ in the current setting), which could be detected by the selection of underexpressed genes for class 1 and this selection can be assured by another sample size calculation for detecting these genes. Hence, in sample size calculations for detecting informative genes (changing in the same direction), positive within-class correlations would be generally suggested for informative genes. In the myeloma dataset introduced in Section 5, the observed within-class correlations among the top 50 genes with the greatest negative t-statistics to compare the responders (class 1) and the non-responders (class 0) for a standard therapy ranged from -0.34 to 0.86 for class 0. While nearly 70% of absolute correlations were <0.2, nearly 25% of correlations were >0.2, and the rest 5% were <-0.2.

For informative genes, we simulated expression data from m_1-dimensional multivariate normal distributions with the mean of 0 for class 0 and Δ for class 1 and a common variance–covariate matrix for two classes. We set the variance equal to one. We considered the correlation matrix within class with all pairwise correlations being equal to $\rho = 0.0$, 0.4, or 0.8. For noninformative genes, we considered a normal distribution with mean 0 and variance–covariance equal to the identity matrix. Table 2 summarizes empirical selection probabilities obtained from 1000 simulations for some configurations with $m = 2000$, $K = 30$, $c = 0.8$, and $f = 0.9$. For $\rho > 0$, the empirical selection probability tended to be greater than the nominal value, 0.8, for $m_1 \leq K$, while less than that for $m_1 > K$.

As Pepe et al. (2003) pointed out, correlated genes will behave as a unit and reduce variability, compared with the setting where all genes are independent. The number of "independent units" for positively correlated informative genes may be less than the number of informative genes. The horizontal axis in Figure 1 could be regarded as the number of independent units for informative genes, instead of the number of informative genes. If the number of informative genes is 20, for example, the number of independent units for informative genes may be less than them, 10 say, due to correlation (see Figure 1). In our formulae, the number of independent units for informative genes is assumed to be identical with the number of informative genes, 20 in this example. Because the expected ratio is a decreasing function of m_1 in the interval $[1, K]$, we will have an overpowered design with our formulae for correlated genes when $m_1 \leq K$. By applying similar arguments, we can have an underpowered design with our formulae for correlated genes when $m_1 > K$.

Table 2
Empirical selection probability from 1000 *simulations:* $m = 2000$, $K = 30$, $f = 0.9$, *and* $c = 0.8$

Δ	m_1	n	$\rho = 0.0$	$\rho = 0.4$	$\rho = 0.8$	Negative	Myeloma
−0.58	10	162	0.81	0.81	0.84	0.80	0.80
	30	228	0.82	0.80	0.85	0.82	0.81
	50	123	0.80	0.64	0.60	0.80	0.80
−1.00	10	57	0.80	0.80	0.84	0.79	0.81
	30	81	0.78	0.81	0.86	0.80	0.78
	50	45	0.81	0.63	0.62	0.77	0.79
−2.00	10	17	0.85	0.80	0.88	0.84	0.84
	30	24	0.82	0.80	0.86	0.83	0.82
	50	15	0.82	0.66	0.63	0.81	0.81

For completeness, we assessed the impact of informative genes changing in the same direction, but having negative within-class correlations, although such informative genes would be generally in a minority (nearly 5% of correlations were <-0.2 among the top 50 genes in the myeloma dataset). We considered independent $m_1/2$ pairs of two informative genes ($m_1 = 10$, 30, or 50), where the two were negatively correlated with the underlying correlation coefficient of -0.8. Interestingly, the empirical selection probabilities (the column labeled "Negative" in Table 2) were almost equal to the nominal level, as in the case of independent informative genes ($\rho = 0.0$). This result indicates that the impact of negative within-class correlation is small. We also considered the simplest case of only two informative genes with the underlying correlation coefficient of γ and normally distributed U_j. Under $m = 1000$, $K = 2$, $f = 1.0$, $\Delta = -1.0$, and $n = 83$, the selection probabilities numerically calculated was the nominal level, 0.80, for independent informative genes ($\gamma = 0.0$). On the other hand, the selection probability was 0.79 for $\gamma = -0.8$ and 0.85 for $\gamma = 0.8$. Again, the result indicates that the impact of negative within-class correlation is small.

As a more real correlation structure for informative genes, we also used the observed correlations among the m_1 genes with the greatest negative t-statistics for class 0 in the myeloma dataset, which was a mixture of positive and negative within-class correlations. The empirical selection probability was nearly identical with the nominal level (see the column labeled "Myeloma" in Table 2), which indicates that our formulae under independent informative genes should be adequate for microarray data such as this.

The impact of correlations among noninformative genes was generally small. We assumed various correlations for 500 noninformative genes (approximately a quarter of m_0). For example, consider a group of 250 positively correlated genes and another group of 250 positively correlated genes. For the configuration with $K = 30$, $f = 0.9$, $c = 0.8$, $\Delta = -1.0$, $m_1 = 30$, and independent informative genes, the empirical selection probability was 0.82 when the two groups were positively correlated and 0.81 when the two were negatively correlated, where the magnitude of all the underlying correlations was set to be 0.4.

3.6 *Conservative Designs*

The arguments in the previous section may lead to strategies for conservative, overpowered designs to cope with issues of unknown number of informative genes and unknown correlation structure across genes. If we have a prior information from earlier studies that there may be an upper limit in the number of informative genes which is less than or equal to K, one may consider conservative designs in which the value of m_1 is set equal to this limit. For more general cases where

there is no such prior information about m_1, one may set $m_1 = K$. Note that these strategies are predicated on the assumption that informative genes are positively correlated within class. Again, under the configurations with $m_1 = 50$ and $\rho = 0.8$, empirical selection probabilities from 1000 simulations for $\Delta = -0.58$, -1.00, and -2.00 were 0.60, 0.62, and 0.63, respectively (Table 2). When the sample size determined by specifying $m_1 = K$ was used for these configurations (i.e., 228, 81, and 24 for the configurations with $\Delta = -0.58$, -1.00, and -2.00, respectively), the empirical selection probabilities for $\Delta = -0.58$, -1.00, and -2.00 increased to 0.95, 0.94, and 0.95, respectively. This strategy for conservative designs would satisfy the requirement on the selection probability for various values of m_1 and various correlation structures for informative genes under the assumption that informative genes are positively correlated.

3.7 *Impact of Normality Assumption*

The sample size calculations presented here are based on normality assumption on gene expression data. We tested the validity of our sample size calculations via simulation. The conditions of simulations were the same as those in Section 3.5, except that expression data are composed of two parts; r percent of m genes independently follow a central t-distribution with $df = 10$, which largely deviates from normality, and the rest $1 - r$ percent a central t-distribution with $df = 300$, which is approximately equal to the standard normal distribution (e.g., Gottardo et al., 2006, for using t-distribution for gene expression data). The two groups of genes were randomly determined. The parameter r represents the degree of deviation from normality for all the genes (m genes). We considered the value of 0, 10, 20, or 30 for r. A constant Δ was added to the expression data of class 1 for informative genes. Table 3 summarizes empirical selection probability from 1000 simulations. The empirical selection probability was nearly identical with the nominal level of 80% for $r = 0$, but slightly less than that for $r = 10$. For $r = 30$, the sample sizes could permitted the selection probability less than the nominal level of 80% by 10% or more.

Again, a pilot study may be run, or earlier experiments may be used to obtain information about the proportion of genes whose expression levels deviate from normality. For the myeloma dataset with the standardization of expression levels across patients within class for each gene, we assumed a central t-distribution and estimated df for each gene by a maximum likelihood method. Of 11,177 genes, 1876 (17%) genes had estimates of $df \leq 10$, 1372 (12%) had estimates ranging from 10 to 50, 430 (4%) had estimates ranging from 50 to 300, and the rest 7467 (67%) had estimates >300. The proportion of non-normality seems to be around the borderline for the myeloma dataset. Another approach to assess the impact of non-normality is a resampling exercise using a real dataset as illustrated in Section 5 for the myeloma dataset.

Table 3

Empirical selection probability when r percent of genes were independently followed by a t-distribution with $df = 10$ within class from 1000 simulations: $m = 2000$, $K = 30$, $f = 0.9$, and $c = 0.8$

Δ	m_1	n	$r = 0$	$r = 10$	$r = 20$	$r = 30$
-0.58	10	162	0.78	0.77	0.74	0.73
	30	228	0.78	0.74	0.70	0.65
	50	123	0.79	0.75	0.70	0.64
-1.00	10	57	0.79	0.78	0.76	0.70
	30	81	0.78	0.74	0.69	0.62
	50	45	0.77	0.75	0.72	0.67
-2.00	10	17	0.81	0.79	0.80	0.73
	30	24	0.81	0.74	0.72	0.66
	50	15	0.80	0.80	0.75	0.75

3.8 *Comparison with Sample Size Calculation Based on Multiple Testing*

Under the assumption that all of the informative genes have the same effect size Δ, our formulae are essentially to guarantee detection of informative genes, like those for sample size calculations based on multiple testing (Dobbin and Simon, 2005; Jung, 2005; Jung et al., 2005; Tsai et al., 2005), where m tests for individual genes are performed. Therefore, our formulae are expected to yield similar results with those based on multiple testing.

Recently, Tsai et al. (2005) developed general formulae for sample size calculations to achieve the desired fraction of the specified measure such as sensitivity and true discovery rate with a desired family-wise power for a given type I error rate. Here sensitivity and true discovery rate for m tests are expressed as Y/m_1 and Y/R, respectively, where Y is the number of true positives, out of R tests declared significant. The comparison-wise error rate α is specified by x/m_0, where x is the number of false positives admissible. Under the assumption that m tests are independent, the comparison-wise power $1 - \beta$ is specified so that the probability of at least y true discoveries computed by summing the binominal probabilities, $Q = \sum_{\ell=y}^{m_1} \frac{m_1!}{\ell!(m_1-\ell)!}(1-\beta)^{\ell}\beta^{m_1-\ell}$, is greater than a prespecified level c, where y is specified as $[f\,m_1]$ or $[fR]$ for fixed $f (0 < f \leq 1)$, which are to limit sensitivity or true positive rate to f, respectively. The required sample size can be obtained as such n that satisfies $n = 4[\{t_\alpha(n-2) + t_\beta(n-2)\}/\Delta]^2$, where $t_q(a)$ is the upper q point of the central t-distribution with $df = a$ (Note: Tsai et al. used formulae based on a normal approximation). If all the genes declared significant are subject to further investigation in subsequent studies, R corresponds to K, although R is a random variable, not a fixed number as in our framework of selection problems. For some settings, our formulae derived from the distribution of ordered statistics of U_j and the Tsai et al.'s formulae based on U_j may yield almost the same results. For example, suppose 30 genes are selected from 2000 genes with 10 informative genes ($K = 30$, $m = 2000$, $m_1 = 10$). Out of 30 selected genes, we need 10 true positive positives (sensitivity = 100%) ($f = 1.0$) and allow 20 ($v = K - m_1$) false positives. The required sample size from the formulae based on multiple testing was 78 for $\Delta = -1.0$ and 22 for $\Delta = -2.0$, which were the same as those from our formulae. On the other hand, settings such as $m_1 > K$ and $f = 1.0$ that require all the selected genes are informative, with no false positives, could not be handled with Tsai et al.'s or other formulae based on multiple testing.

4. Multiple Groups of Informative Genes

The assumption that all the informative genes have the same effect size Δ is obviously unrealistic. The effect size may vary by virtue of complex pathway and gene network relationships linked to the phenotype variable. In this case, one may specify the level of effect size one is seeking to detect. If all the informative genes (changing in the same direction) have effect sizes greater than this level, the previous formula with this level as a common effect size may yield a conservative design. However, it is likely to exist as informative genes whose effect sizes are smaller than this level. Because the probability that one of these informative genes is selected is greater than the probability that a noninformative gene is selected, an adaptation is needed to assure selection of informative genes whose effect sizes are greater than the specified level.

The formulae developed in the previous sections can be extended to multiple groups of informative genes with distinct effect sizes. For simplicity, we assume two groups of informative genes, one group comprised m_1 genes with effect size, Δ_1, which represents the size one is seeking to detect, and the other group comprised m_2 genes with effect size, Δ_2. We assume that the first group with m_1 genes is strongly associated with the phenotype variable, while the second group with m_2 genes is weakly associated with phenotype variable, i.e., $|\Delta_1| > |\Delta_2|$.

In sample size calculations, we assume statistical independence of genes as before. We calculate the probability that, of the K selected genes, we have, at least, H genes from the first group of informative genes, denoted by $Q_{K,H}$, where $\max(1, K + m_1 + 1 - m) \le H \le \min(m_1, K)$. Again, we introduce a ratio, f, so that $H = [f \times \min(m_1, K)]$. Note that f does not correspond to sensitivity or true positive rate in multiple testing for multiple groups of informative genes because f does relate to selection of the first informative gene group, a subset of informative genes, not selection of all the informative genes. Let $S_{(1)} < \cdots < S_{(m_1)}$ be the ordered statistics of U_j for the first informative gene group (m_1 genes) and $U_{(1)} < \cdots < U_{(m-m_1)}$ be those for a pooled group of the second informative gene group (m_2 genes) and the rest $m - m_1 - m_2$ for noninformative genes. Again, let $J = K - H + 1$. For detecting overexpressed genes for class 1, as an analogue to the expression (2), $Q_{K,H}(= \Pr\{U_{(H)} > S_{(J)}\})$ can be calculated as

$$Q_{K,H} = \frac{m_1!}{(H-1)!(m_1-H)!} \times \int_0^1 R_{2,0,(J)}\{F_1^{-1}(u)\}u^{H-1}(1-u)^{m_1-H}\,du,$$

where

$$R_{2,0,(r)}(s) = \sum_{k=0}^{r-1}\sum_{\ell=0}^{k}\binom{m_2}{\ell}F_2(s)^{\ell}\{1-F_2(s)\}^{m_2-\ell} \times \binom{m_0}{k-\ell}F_0(s)^{k-\ell}\{1-F_0(s)\}^{m_0-k+\ell},$$

and F_j is the distribution function of a non-central t-distribution with $df = n - 2$ and the non-centrality parameter $\sqrt{\frac{n_1(n-n_1)}{n}}\Delta_j (j = 1, 2)$.

The extension to three or more groups of informative genes with distinct effect sizes is straightforward. Since the impacts on sample size estimates by introducing additional informative gene groups to the first informative gene group can be substantial as indicated by Sections 4.1 and 5, sample size calculations with plausible specifications for additional informative gene groups are largely warranted. Again, a pilot study may be run, or earlier experiments may be used to estimate the distribution of effect sizes for informative genes. We will provide a simple estimation method for the myeloma dataset in Section 5.

4.1 *Illustration*

Table 4 summarizes the expected sample size to satisfy the requirement on the selection probability, $Q_{K,H} \ge 0.8$, for the settings with $\Delta_1 = -1.0$ and $\Delta_2 = \Delta_1/2 = -0.5$ for the same configurations as those in Table 1 for the other parameters. Note that, as seen for the formulae under the assumption of a single group of informative genes with a common effect size Δ for informative genes (Table 1) (or entries for $m_2 = 0$ in Table 4), the required sample size was maximized when $m_1 = K$, suggesting conservative designs with $m_1 = K$ for positively correlated informative genes. As one expects, the impact of m_2 could be substantial, and greater sample sizes were required for greater m_2.

4.2 *Simulations with Correlated Genes*

We assessed the performance of our formulae based on statistical independence of genes for correlated genes via simulation. The conditions of simulations were the same as those in Section 3.5. We considered the correlation matrix with the common pairwise correlation and that observed in the myeloma dataset, for $m_1 + m_2$ informative genes. We set $\Delta_1 = -1.0$ for the first informative gene group (m_1 genes) and $\Delta_2 = \Delta_1/2 = -0.5$ for the second informative gene group (m_2 genes). Table 5 summarizes empirical selection probabilities obtained from 1000 simulations. Again, the empirical selection probability was nearly identical with the nominal level of 80% when informative genes had the correlations observed in the myeloma dataset. For greater common correlation ρ, our formulae yielded overpowered designs when $m_1 \le K$ and underpowered designs for smaller m_2, e.g., $m_2 = 10$, when $m_1 > K$, as is the case for $m_2 = 0$ (see Table 2). When $m_1 > K$, it is interesting that we had overpowered designs for greater m_2, e.g., $m_2 = 50$. Again, the required sample size could be reduced for smaller numbers of independent units for the second informative gene group. For greater m_2, the effects of correlation to reduce the required sample size through decreasing the number of independent units for the second informative gene group can outperform the effects of correlation to enlarge the required sample size through decreasing the number of independent units for the first informative gene group when $m_1 > K$, resulting in overpowered designs. To have conservative designs, one may set a large value for m_2, possibly in addition to setting $m_1 = K$ under the assumption that informative genes are positively correlated.

5. Myeloma Example

We illustrate the formula for sample size calculations using one of the microarray studies at the Myeloma Institute for

Table 4

Expected sample size to satisfy the requirement of $Q \geq c = 0.8$ for the two groups of informative genes: $\Delta_1 = -1.0$, $\Delta_2 = \Delta_1/2 = -0.5$

		$m = 2000$				$m = 5000$			
		$K = 30$		$K = 50$		$K = 30$		$K = 50$	
m_1	m_2	$f = 0.8$	$f = 0.9$	$f = 0.8$	$f = 0.9$	$f = 0.8$	$f = 0.9$	$f = 0.8$	$f = 0.9$
10	0[a]	46	57	39	49	56	68	49	60
	10	47	59	40	50	57	70	50	61
	30	50	63	41	52	59	73	51	63
	50	53	69	43	55	62	77	52	65
30	0	60	81	44	57	69	92	54	68
	10	63	92	45	59	72	100	55	70
	30	73	126	47	63	80	128	57	73
	50	86	152	50	68	89	152	59	77
50	0	36	45	54	75	43	53	63	86
	10	37	48	56	82	44	55	65	91
	30	40	55	62	104	47	60	70	108
	50	43	63	69	127	49	66	75	128

[a]For $m_2 = 0$, the formula for a single group of informative genes with a common effect size in Section 3.2 was used.

Research and Therapy in Little Rock, Arkansas that were on CD138-enriched plasma cells from bone marrow aspirates taken from newly diagnosed multiple myeloma patients entering the National Cancer Institute-sponsored Phase III clinical trial (Barlogie et al., 2006). Specifically, we focus on a preliminary study of 221 patients with approximately 12,000 genes on the Affymetrix oligonucleotide chip (Harousseau, Shaughnessy, and Richardson, 2004). We supposed a situation where, following the results of this study, we designed a similar study. We considered comparison of gene expression data between responders (complete response) and non-responders for the standard therapy with high-dose chemotherapy supposed by autologous hematopoietic stem-cell transplantation.

We assessed the adequacy of our sample size calculations by a resampling exercise (Li et al., 2005) using the real dataset from the 221 patients, which provided an empirical selection probability for a given configuration of design parameters. This resampling exercise is to give another angle on adequacy of our sample size calculations with the assumptions of independent genes and normality of gene expression data using the real dataset. Here we provide results for selecting overexpressed genes for responders. Similar results were obtained for selecting underexpressed genes for responders.

Table 5

Empirical selection probability for the two groups of informative genes from 1000 simulations: $m = 2000$, $K = 30$, $c = 0.8$, $f = 0.9$, $\Delta_1 = -1.0$, $\Delta_2 = -0.5$

m_1	m_2	n	$\rho = 0.0$	$\rho = 0.4$	$\rho = 0.8$	Myeloma[a]
10	10	59	0.81	0.83	0.86	0.79
	30	63	0.79	0.86	0.89	0.80
	50	69	0.79	0.89	0.94	0.81
30	10	92	0.80	0.89	0.91	0.79
	30	126	0.79	0.99	0.99	0.80
	50	152	0.83	0.99	0.99	0.80
50	10	48	0.80	0.68	0.64	0.77
	30	55	0.82	0.79	0.73	0.80
	50	63	0.81	0.87	0.86	0.81

[a]The correlation matrix observed in the myeloma dataset was used.

We first removed the potential differences in expression levels between two response classes, responders and non-responders, not attributable to noise for each gene, by subtracting the class means from the original expression levels so that classes have the same mean levels. The top ranking genes with the greatest negative t-statistics in the original dataset were regarded as informative genes. For each of the informative genes, a difference Δs was added to the mean difference between classes in the modified dataset, where s is an estimate of a common standard deviation between classes for that gene, and the resultant modified dataset was regarded as the dataset of the population from which patients are resampled. In order to generate the expression levels for each class with a given sample size n^*, we repeatedly sampled from the modified expression levels of the whole patients within that class with replacement, where the total sample size was $n = 2n^*$. As such, the generated dataset is expected to preserve the properties of the original dataset in terms of the marginal distribution of expression levels for individual genes and the correlation structure among genes within class.

We tried to estimate the distribution of effect sizes using the original dataset. Table 6 shows the observed frequency of the estimated standardized mean difference between classes for all the genes ($m = 11{,}177$). We estimated the proportion of genes unrelated to classes using a simple procedure by Storey and Tibshirani (2003). Most of genes with small absolute standardized mean differences would be unrelated to classes. The proportion of genes whose absolute standardized mean differences <0.2 was 70.5% and the average of the corresponding proportion obtained from 1000 permutations of class label was 78.0%, suggesting the proportion of genes unrelated to classes of 0.9 ($= 70.5/78.0$). Note that almost the same estimates were obtained using other smaller cut-offs

Table 6
The observed and estimated frequencies of the standardized mean difference between the response classes in the myeloma data

Standardized mean difference	Observed frequency	Estimated frequency for genes unrelated to classes[a]	Estimated frequency for genes related to classes[b]
<-0.7	1 (0.0%)	0.1 (0.0%)	0.9
-0.7 to -0.6	10 (0.1%)	1.2 (0.0%)	8.8
-0.6 to -0.5	25 (0.2%)	9.9 (0.1%)	15.1
-0.5 to -0.4	145 (1.3%)	59.9 (0.6%)	85.1
-0.4 to -0.3	410 (3.7%)	258.9 (2.6%)	151.1
-0.3 to -0.2	1066 (9.5%)	774.6 (7.7%)	291.4
-0.2 to -0.1	1715 (15.3%)	1615.8 (16.1%)	0.0
-0.1 to 0.0	2353 (21.1%)	2316.8 (23.0%)	0.0
0.0 to 0.1	2199 (19.7%)	2312.0 (23.0%)	0.0
0.1 to 0.2	1605 (14.4%)	1598.4 (15.9%)	0.0
0.2 to 0.3	945 (8.5%)	771.8 (7.7%)	173.2
0.3 to 0.4	464 (4.2%)	263.2 (2.6%)	200.8
0.4 to 0.5	161 (1.4%)	63.8 (0.6%)	97.2
0.5 to 0.6	60 (0.5%)	11.4 (0.1%)	48.6
0.6 to 0.7	13 (0.1%)	1.4 (0.0%)	11.6
>0.7	5 (0.0%)	0.2 (0.0%)	4.8
Total	11,177	10,059.3	1088.6

[a]The estimated relative frequency (the entries in parentheses) was the average relative frequency obtained from 1000 permutations of class label using all the genes. The estimated frequency was then calculated as the estimated relative frequency multiplied by the estimated number of genes unrelated to classes, $10{,}059.3 (= 11{,}177 \times 0.9)$.

[b]Quantity obtained by subtracting the estimated frequency for genes unrelated to classes from the observed frequency for absolute standardized mean differences ≥ 0.2. The estimated frequency was set equal to zero for absolute standardized mean differences <0.2.

such as 0.1. Based on this estimate, we estimated the frequency distribution of standardized mean difference for genes unrelated to classes and for genes related to classes, respectively (see Table 6). We sought to detect genes whose standardized mean differences <-0.6 (the first informative gene group in Section 4). Table 6 indicated nearly 10 ($\approx 0.9 + 8.8$) such genes.

We calculated the sample size for selecting 50 genes ($K = 50$). We set the number of genes of the first informative gene group as 10 ($m_1 = 10$). We determined the sample size to satisfy the selection probability that, at least, eight of the selected K genes were from the first group of informative genes ($f = 0.8$) was greater than 0.8 ($c = 0.8$). For an ideal situation where only this single gene group (a common effect size) was informative and the other genes were noninformative, the required sample size was 58 for $\Delta_1 = -1.0$ and 154 for $\Delta_1 = -0.6$. The empirical selection probability from 1000 resampled datasets was 0.82 for the former and 0.80 for the latter configuration. For more realistic situations, we also calculated sample sizes reflecting the distribution of effect sizes indicated by Table 6. We assumed two additional groups of informative genes (the second and third informative gene groups). One is 100 genes ($\approx 15.1 + 85.1$) with effect sizes ranging from -0.6 to -0.4 and the other is 443 genes ($\approx 151.1 + 291.4$) with effect sizes ranging from -0.4 to -0.2. We set the effect size for the former group as a weighted average of effect sizes 0.42 ($= \{15.1 \times (-0.5) + 85.1 \times (-0.4)\}/(15.1 + 85.1)$). Similarly, we set the effect size for the latter group as 0.24. For this setting, the required sample size increased to 65 for $\Delta_1 = -1.0$ and 284 for $\Delta_1 = -0.6$. The empirical selection probability was 0.83 for the former and 0.82 for the latter configuration. The empirical selection probabilities were generally close to the nominal level, which indicates the adequacy of our sample size calculations for microarray data such as this. Lastly, for the latter configuration with 284 patients for $\Delta_1 = -0.6$, we also specified the estimated frequency distribution of effect sizes in the range of the standardized mean difference >-0.6 for genes related to classes in Table 6, when obtaining empirical selection probability to mimic the real dataset. The resultant empirical selection probability was 0.80 and still close to the nominal level, which indicates the adequacy of adopting the formulae with two additional groups of informative genes (two distinct effect sizes) for more realistic situations with various effect sizes for the myeloma dataset.

6. Discussion

In this article, we have developed formulae to calculate sample sizes for genome-wide screening with microarrays in the framework of ranking and selection problems. A pilot study or earlier experiments are essential for specifying plausible values of design parameters as well as for checking assumptions such as normality. Although we have derived strategies for conservative designs heuristically, it is advisable to find a feasible conservative design on a case-by-case basis, possibly through sensitivity analyses with several possible characteristics of gene expression of informative and noninformative genes indicated by a pilot study or earlier experiments. The formulae are also applicable to exploratory screening studies using other recently developed high-throughput technologies, for example, proteomics studies with protein mass spectrometry to select a set of informative proteins from among a large set of candidate proteins. SAS and R-codes to calculate sample sizes are available upon request.

Although we have considered equal numbers of samples within both classes in sample size calculations, the generalization to incorporate unequal numbers of samples is straightforward by introducing the parameter representing the proportion of samples for class 1, $p = n_1/n$. One can control this parameter to be close to 0.5, so that the required sample size is minimized in the sample size calculations based on the distribution of the two-sample t-statistic. This type of control may be possible when large specimen banks from large cohort studies or clinical trials are available for both classes. On the other hand, there is another situation in which this type of control can be difficult at the design stage. For example, in pharmacogenomic prospective studies with relatively small or moderate sample sizes to correlate pretreatment gene expression data with the response to treatment, the proportion p, which corresponds to the response rate, is unknown or a random variable at the design stage. In such studies, there is often no guarantee to obtain enough equal numbers of samples from both classes, which may oblige one to analyze unequal numbers of samples. A practical approach for this situation is to calculate sample sizes for a range of plausible values for p. Another, more formal approach is to assume a binomial distribution $B(n, \pi)$ for n_1 with the probability function q. Plausible values for π may be indicated by earlier clinical studies. As the selection probability, we use a marginalized version of $Q_{K,H}$ for $n_1, Q^*_{K,H} = \sum_{n_1=1}^{n-1} q(n_1) Q_{K,H}$, instead of $Q_{K,H}$. Generally, we need greater sample sizes than those from the calculation based on $Q_{K,H}$ with $p = \pi$.

A critical assumption of our formulae is the normality assumption on gene expression data. As noted in Section 3.7, assessment of the impact of nonnormality on the sample size estimates using available datasets is warranted. For the myeloma dataset, we can conclude that the impact of nonnormality is not large through the estimation of df for t-distributed data in Section 3.7 and the resampling exercise in Section 5. An approach to circumvent the normality assumption is to use a distribution-free statistic, such as Mann and Whitney test, for U_j. Sample size calculations based on an asymptotic normal distribution for the Mann and Whitney test statistic (e.g., Hettmansperger, 1984) is possible. Adequacy for using asymptotic distributions for distribution-free test statistics, including that for a log-rank score for right-censored survival outcomes (Hsieh and Lavori, 2000), in sample size calculations is under investigation and thus subject to future report.

Acknowledgements

We thank John Crowley of Cancer Research and Biostatistics for coordinating this collaborative study. We also thank the associate editor for helpful comments that substantially improved this article.

References

Allison, D. B., Gadbury, G. L., Heo, M., Fernandez, J. R., Lee, C.-K., Prolla, T. A., and Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis* **39**, 1–20.

Barlogie, B., Tricot, G., Anaissie, E., Shaughnessy, J., Rasmussen, E., van Rhee, F., Fassas, A., Zangari, M., Hollmig, K., Pineda-Roman, M., Lee, C., Talamo, G. et al. (2006). Thalidomide and hematopoietic-cell transplantation for multiple myeloma. *New England Journal of Medicine* **354**, 1021–1030.

Benjamini, Y. and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25**, 60–83.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Dobbin, K. and Simon, R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics* **6**, 27–38. Erratum in: *Biostatistics*, **6**, 348.

Gottardo, R., Raftery, A. E., Yeung, K. Y., and Bumgarner, R. E. (2006). Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics* **62**, 10–18.

Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks*. New York: John Wiley and Sons.

Harousseau, J. L., Shaughnessy, J. J., and Richardson, P. (2004). Multiple myeloma. *Hematology* (Am Soc Hematol Educ Program), 237–256.

Hsieh, F. Y. and Lavori, P. W. (2000). Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Controlled Clinical Trials* **21**, 552–560.

Hsueh, H.-M., Chen, J. J., and Kodell, R. L. (2003). Comparison of methods for estimating the number of true hypotheses in multiplicity testing. *Journal of Biopharmaceutical Statistics* **13**, 675–689.

Hu, J., Zou, F., and Wright, F. A. (2005). Practical FDR-based sample size calculations in microarray experiments. *Bioinformatics* **21**, 3264–3272.

Jung, S. H. (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics* **21**, 3097–3104.

Jung, S. H., Bang, H., and Young, S. (2005). Sample size calculation for multiple testing in microarray data analysis. *Biostatistics* **6**, 157–169.

Li, S. S., Bigler, J., Lampe, J. W., Potter, J. D., and Feng, Z. (2005). FDR-controlling testing procedures and sample size determination for microarrays. *Statistics in Medicine* **24**, 2267–2280.

Lossos, I. S., Czerwinski, D. K., Alizadeh, A. A., Wechser, M. A., Tibshirani, R., Botstein, D., and Levy, R. (2004). Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *New England Journal of Medicine* **350**, 1828–1837.

Mahamunulu, D. M. (1967). Some fixed-sample ranking and selection problems. *The Annals of Mathematical Statistics* **38**, 1079–1091.

McLachlan G. J., Do, K. -A., and Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data*. Hoboken, New Jersey: Wiley.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., and Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* **21**, 3017–3024.

Pepe, M. S., Longton, G., Anderson, G. L., and Schummer, M. (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics* **59**, 133–142.

Schweder, T. and Spjotvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B* **64**, 479–498.

Storey, J. D. and Tibshirani, R. (2003). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data: Methods and Software*, G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger (eds), 272–290. New York: Springer.

Tsai, C. A., Wang, S. J., Chen, D. T., and Chen, J. J. (2005). Sample size for gene expression microarray experiments. *Bioinformatics* **21**, 1502–1508.

Wetherill, G. B. and Ofosu, J. B. (1974). Selection of the best of k normal populations. *Applied Statistics* **23**, 253–277.

Received June 2006. *Revised April* 2007. *Accepted May* 2007.

Endoscopic submucosal dissection of recurrent or residual superficial esophageal cancer after chemoradiotherapy

Yutaka Saito, MD, PhD, Hajime Takisawa, MD, Haruhisa Suzuki, MD, Kouhei Takizawa, MD, Chizu Yokoi, MD, Satoru Nonaka, MD, Takahisa Matsuda, MD, Yukihiro Nakanishi, MD PhD, Ken Kato, MD

Tokyo, Japan

Background: Treatment of local recurrent or residual superficial esophageal squamous-cell carcinoma (SCC) with conventional EMR often results in a piecemeal resection that requires further intervention.

Objective: The aim of this study was to evaluate the efficacy of endoscopic submucosal dissection (ESD).

Design: A case series.

Patients: Between January 2006 and September 2006, 4 local recurrent or residual superficial esophageal SCCs were treated by ESD.

Interventions: ESD procedures were performed by using a bipolar needle knife and an insulation-tipped knife. After injection of glycerol into the submucosal (sm) layer, a circumferential incision was made, and an sm dissection was performed. All lesions were determined to be intramucosal or sm superficial, without lymph-node metastasis by EUS before treatment.

Main Outcome Measurements: Tumor size, en bloc resection rate, tumor-free lateral margin rates, and complications were recorded.

Results: All 4 ESD cases were successfully resected en bloc, and the tumor-free lateral margin rate was 75% (3/4) by histopathology examination. The mean tumor size of the resected specimens was 35 mm (range, 15-50 mm). There were no complications.

Limitations: The number of ESDs in our series was limited, and there are no long-term follow-up data.

Conclusions: ESD for recurrent or residual superficial esophageal tumors after chemoradiotherapy achieves the goal of an en bloc resection, with a low rate of incomplete treatment without any greater risk than the EMR technique.

Esophageal cancer is one of the most difficult GI cancers to detect at an early stage, even by endoscopy. Recently, a narrow-band imaging endoscope was developed and was shown to be advantageous for the early detection of squamous-cell carcinoma (SCC) in the esophagus and the pharynx, although it still is not widely in use.[1,2]

Some esophageal cancers have been detected as invasive tumors, and surgery has been the standard treatment for such lesions. However, higher mortality rate because of surgery has been reported (range 2.1% to 13.7%), as has poor patient quality-of-life after surgery.[3,4]

There is a current preference to treat esophageal SCC by primary chemoradiotherapy (CRT),[5,6] but 13% of patients treated for esophageal SCC with CRT have a recurrence or a residual tumor. Surgery after CRT is unsatisfactory,[7,8] and endoscopic treatment can be proposed when the tumor is superficial,[9-13] but a strip biopsy is difficult, because fibrosis and piecemeal resection frequently occur even for small lesions. A search of the literature confirmed that en bloc resection by endoscopic submucosal dissection (ESD) provides better results in the stomach.[14-17] ESD was recently reported to be useful in the treatment of superficial esophageal SCC[18-20]; however, the feasibility and safety of ESD for local recurrent or residual tumors is unclear. Previously, we reported on

Abbreviations: B-knife, bipolar needle-knife; CRT, chemoradiotherapy; ESD, endoscopic submucosal dissection; IT-knife, insulation-tipped-knife; NCCH, National Cancer Center Hospital; SCC, squamous-cell carcinoma; sm, submucosal.

0016-5107/$32.00
doi:10.1016/j.gie.2007.10.008

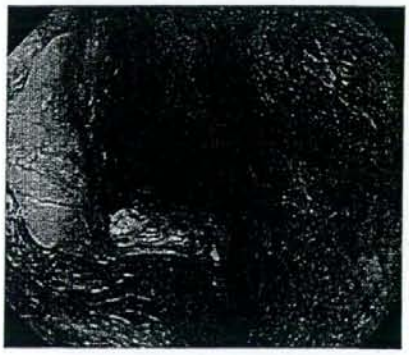

Figure 1. The primary tumor before CRT was diagnosed as a type 1 SCC, with a circumferential intraepithelial lesion, which had been located in the mid esophagus at a previous hospital.

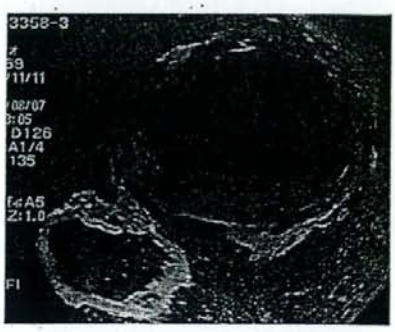

Figure 2. An endoscopy revealed a 0-IIC superficial residual lesion, 40 mm in diameter, located in the mid esophagus. After iodine staining, the lesion became more apparent and was larger than 50% in circumference.

the effectiveness and safety of ESD for colorectal tumors by using a bipolar needle-knife (B-knife) and an insulation-tipped knife (IT-knife), neither of which has any coagulation effect at the needle tip.[21-24] The aim of our study was to evaluate the efficacy and safety of ESD for local recurrent or residual esophageal tumors by using a B-knife and an IT-knife.

PATIENTS AND METHODS

Four patients with esophageal SCC, each of whom had developed a local recurrent or residual tumor (2 recurrent tumors and 2 residual tumors) after CRT, were included in this study, which was conducted between January 2006 and September 2006 at the National Cancer Center Hospital (NCCH) in Tokyo. Three of the ESD cases involved stage I lesions treated by CRT, and the other case was of a stage II lesion. The 4 ESDs were performed from 217 days to 1377 days after the initial CRT.

ESDs by using a B-knife and an IT-knife were performed on all 4 patients, with Glyceol (Chugai, Tokyo, Japan)[25] used in each case as the submucosal (sm) injection solution to maintain proper sm elevation. All of the local recurrent or residual tumors were confirmed as intramucosal or sm superficial, without lymph-node metastasis, by EUS and a CT before treatment.

Endoscopic operating system

ESD procedures were performed by using video endoscopes (GIF-Q240 or GIF-Q260; Olympus Optical Co, Ltd, Tokyo, Japan).

ESD procedure

A transparent disposable attachment (D-201-1074; Olympus) was fitted onto the tip of the endoscope to retract the sm layer and to facilitate dissection. Lesion margins were delineated before ESD by using 1.5% iodine staining (Figs. 1 and 2). After sm injection of Glyceol, a circumferential incision in the mucosa was made by using a B-knife and an IT-knife.[21-24] Additional Glyceol was then injected into the sm layer to lift the lesion, and the thickened sm layer was dissected by using an IT-knife (Figs. 3 and 4). The B-knife was mainly used for the dissection of fibrosis caused by CRT.[21-24] The operation time was recorded for all patients.

Sedation

Midazolam (3-5 mg intravenously) was administered in all cases. An additional 2 mg was given as necessary, whenever indicated, based on the individual endoscopist's judgment.

Histologic assessment

All specimens were evaluated after being cut into 2-mm slices; they were examined microscopically for histologic type, depth of invasion, lateral resection margin, and vertical resection margin.

Follow-up care

All patients who had an ESD at the NCCH were regularly observed, with annual endoscopic and EUS examinations and CTs. Complete follow-up care was available for all 4 patients in the ESD group.

Statistical analysis

All variables in this study were described as mean (SD). All statistical analyses were performed by using SAS version 8.0 (SAS Institute Inc, Cary, NC). The P value was 2 sided, and $P < .05$ was used to determine statistical significance.

Ethics

The ethics committee at the NCCH approved the study protocol, and written informed consent was obtained from all 4 patients in the ESD group before entering the study.