TABLE 4
Performance Comparison of MAP and WMAP Using a 14-Class GCM Data Set

		1R	11	SR
Training	MAP	1 (0)	0.996 (0.004)	1 (0)
	WMAP	1 (0)	1 (0)	1 (0)
Test	MAP	**0.770** (0.036)	0.674 (0.055)	0.764 (0.031)
	WMAP	0.770 (0.034)	0.674 (0.062)	**0.795** (0.068)

	MS-SVM(WW)	MC-SVM(CS)	NSC ($\Delta = 0.45$)
Training	1 (0)	1 (0)	0.689 (0.029)
Test	0.754 (0.069)	0.753 (0.052)	0.605 (0.060)

intractable, because the number of targets $\#B_M^{AA}$ in the AA coding increases exponentially with respect to M:

$$\#B_M^{AA} = \sum_{i=1}^{\lceil M/2 \rceil} \sum_{j=1}^{M-i} \frac{1}{1+\delta_{ij}} \binom{M}{i+j} \binom{i+j}{j}, \tag{9}$$

where δ_{ij} is the Kronecker's delta.

In this section, we apply our code optimization based on the WMAP framework to a 14-class problem [3] of a global cancer map (GCM) data set, where the number of targets in the initial AA coding becomes as high as 2,375,101. Although the original GCM data set consists of 16,063 genes' expression profiles for 144 training samples and 54 test samples of 14 common cancer types, we merged the training and test samples to construct a data set of 190 samples (eight test samples of another cancer type were removed) in this experiment.

Although WMAP would be able to find an optimal and possibly sparse (or graded) subset by starting from the whole AA code matrix in principle, it is in practice difficult partly because of the lack of a sufficient number of data and partly because of the computational intractability on a set whose element number becomes exponential. By applying a sparse random coding (SR) to this 14-class problem, we obtained a set of 58 targets, where the number of targets was determined as a rough standard $\lceil 15 \log_2 M \rceil$, as proposed in [15], and the 58 targets were selected from 10,000 random sets of 58 targets, so that any two code words were the furthest apart from each other in the sense of Hamming distance. The details are described in Appendix D. The objective of our WMAP method here is to obtain an appropriate graded code matrix starting from the initial set SR and hence to optimize the weight values for this "random" but "rigid" code matrix.

We evaluated the six methods, namely, the combinations of three designs of code matrix, 1R, 11, and SR, and the two procedures, MAP and WMAP, for the 14-class problem by the means of a fivefold cross validation. Each binary classifier was implemented as a linear kernel SVM employing all genes. Multiclass SVMs implemented as MC-SVM(CS) and MC-SVM(WW), and the NSC method were also compared within the same conditions as Experiment 2. Table 4 shows the results.

A simple voting by linear kernel SVMs showed better performance with MAP-1R than with MAP-11. MAP-SR was also better than MAP-11 but did not exceed MAP-1R; these results were consistent with those by [3]. However, the results were different when the weight optimization was performed; WMAP-SR became better than MAP-1R. The performances of WMAP-1R and WMAP-11 were not improved by the weight optimization, probably because the training accuracy was saturated, suggesting this 14-class classification problem contains rather little data in comparison to the complexity of the multiclass problem. Even in such a saturated condition and when the initial code matrix includes a lot of targets, our weight optimization method works well, as confirmed by the improvement of WMAP-SR over MAP-SR. The other three state-of-the-art alternatives did not exceed the results of the WMAP-SR. From this experiment, we can see that an appropriate "rigid" code matrix may improve performance by appropriately decoding from it based on the MAP method, over the simple voting heuristics by 1R or 11, and introducing the weight optimization to seek the optimal "graded" set from the rigid one can further improve the performance.

6 Discussion and Conclusions

The statistical model of the MAP decoder is an expanded version of the pairwise coupling method of binary probability estimates [9], which used only B^{11}, and is conceptually similar to the method in [10], which was also an expansion in [9], while the MAP method incorporates an additional term that naturally represents prior knowledge of class distribution. In [10], B^{1R}, B^{11}, and random targets from B^{AA} were dealt with in combinations, but B^{AA} itself was not considered, and the optimization of code matrices (optimal coding problem) was unsolved. Our weight optimization method used in WMAP successfully obtained a graded code matrix starting from any code matrix without any prior knowledge about the data, and this method could be one answer to the optimal coding problem. This feature is essential for practical tumor classification problems using gene expression profiling, because we often do not have much information on the data.

When B^{AA}, containing all possible targets was used as the code matrix to be weighted, the WMAP method often showed the best performance. Especially when B^{AA} could not be used as in the 14-class GCM problem, various sets of targets can be considered by reducing B^{AA} but determining which code matrix shows the best performance among them requires some heuristics. In the current study, we used the SR method [15], but still, the weight estimation method worked well when applied to the reduced code matrix, implying that our method could solve, at least to some extent, the optimal coding problem by searching the analog coding space restricted within the initial binary coding. The current results suggest that the larger the code matrix to be optimized, the better the performance becomes, though the optimization of large code matrices requires heavy computation. Although it is important, in practice, to seek a better configuration of the initial code matrix than the exhaustive coding B^{AA} or the SR B^{SR} when the class number is not small, this problem is not the target of our current study but a future one, because our code optimization technique can in principle employ any initial setting of the code matrix.

In our code optimization method, overtraining of weights may occur, especially when the number of targets is large. This problem could be avoided by using large training data sets, but in many actual gene expression analyses, handling small data sets is required. One possible way of dealing with this problem is to use various parameter optimization techniques such as the leave-two-out (LTO) method [26],

which is a hierarchical cross-validation approach. When the training accuracy by the MAP decoder reaches a higher limit, i.e., saturation occurs, the room for taking advantage of the WMAP method is restricted. This is an another aspect of overtraining, and this tendency is more apparent when the binary classification method is strong enough. One way of solving this overtraining problem is to split the training data set into two or more subsets and to train the binary classifiers and adjust the weights individually using different data subsets.

The linear-kernel SVMs used in this study have been preferred in many bioinformatics studies [27], [28]. On the other hand, we can also tune parameters of the kernel and/or use other kernels in the SVM or use more sophisticated binary predictors such as AdaBoost [29]. The flexibility of our framework that can employ any binary classification algorithms could yield further performance and work well especially when a relatively large amount of data are available. For example, when we use a quadratic-kernel SVM as a unit classifier (the kernel function is $K(x, x') = (1 + x^T x')^2$), the resultant cross-validation accuracy for the thyroid data set was 1 in the training and 0.780 in the test. Note that this improvement was due to the change in the unit classifier rather than the weight optimization. If we employ highly adjustable binary classifiers and an appropriate code matrix, the classifiers are well adapted to the given data set, and the decoder can easily integrate them. In such a case, there is little room to achieve improvement by the WMAP's code optimization because of the saturation of the gain function. Accordingly, the performance improvement by WMAP is dependent on both the data set (and the class structure underlying the data) and the choice of binary classifiers. Still, however, it is important that our code optimization technique can be employed with any choice of binary classifiers.

From the view point of bioinformatics, our WMAP method may also be valuable for feature extraction and existing classifiers such as the NSC method [25]. The optimized weights can be interpreted as a numerical feature vector in the binary classifier space, each of whose elements characterizes the degree of the corresponding binary classifier's contribution to multiclass classification results, i.e., disease (medical or phenotypic) information in cancer classification problems. In other words, the code matrix optimized by WMAP would enable us to observe the interclass relationships in a multiclass classification problem from the perspective of several binary classification problems. For example, the code optimized by our WMAP method can provide information on the geometrical relationship of multiple classes, which can be seen in the experiment using a synthesized data set (Experiment 1). On the other hand, the centroids of NSC can be construed as characteristic pattern vectors in the gene expression space, each of which represents the corresponding class label and each of whose elements indicates gene's responsibility to each class label. While these are class specific patterns, they do not represent interclass relationship directly. Consequently, WMAP and NSC extract some characteristic features from data sets, but they contain different types of information. As stated above, the WMAP method itself does not select informative genes for cancer classification tasks. However, it is possible to obtain some evidences about the gene contribution by using binary classification algorithms such as the weighted voting method [1] and SVM-RFE [30] as a unit classifier of WMAP, which incorporates gene selection processes: The higher ranked or survived genes in a binary classifier that has a larger optimized weight are supposed to have substantial influence on multiclass classification results. To do this, we must not only optimize the code matrix but also tune gene selection parameters of binary classifiers, but the parameter tuning is an independent issue of our approach. However, we expect that our approach will elucidate biological meaning through linkages between the optimized code words and the class labels in gene expression analyses.

The novel approaches introduced in this study show promise as the means to differentiate similar tumor types of the same origin, as are thyroid and esophageal cancers, for example. Before the final determination of their efficacy, a number of confirmatory experiments are necessary. Nevertheless, we believe that our algorithms based on ECOC coding/decoding will contribute to providing advanced tools in the pathological diagnosis of cancer in the near future.

APPENDIX A

PROBABILITY ESTIMATION FROM DECISION VALUES OF BINARY CLASSIFIERS

In order to convert a discriminant function value from a binary predictor (in this study, a binary SVM) into the probability estimate, we employed a regression-based method proposed by Platt [17]. Let $d_j(x) \in \mathcal{R}$ be a discriminant function value from the jth predictor constructed based on the partial training data $L_j = \{x^{(n)}, i^{(n)}\}_{n \in n_j}$, where n_j is the index set of samples used to make this predictor. The logistic regression model assumes that the probabilistic guess $q_j^{(n)} = Pr(i \in 1_j | x^{(n)}, i \in 1_j \cup 0_j)$ is given by the parametric sigmoidal function of $d_j(x)$:

$$q_j^{(n)} = \frac{1}{1 + \exp(A_j d_j(x^{(n)}) + B_j)},$$

where A_j and B_j are the model parameters specific to the jth target. These parameters were estimated by maximizing the log likelihood on the transformed training data $L_j' = \{d_j(x^{(n)}), i^{(n)}\}_{n \in n_j}$:

$$\max \sum_{n \in n_j} \left\{ r_j^{(n)} \log q_j^{(n)} - (1 - r_j^{(n)}) \log(1 - q_j^{(n)}) \right\}. \quad (10)$$

$r_j^{(n)}$ is the target probability defined as

$$r_j^{(n)} = \begin{cases} r_{1_j} & \text{if } i^{(n)} \in 1_j, \\ r_{0_j} & \text{if } i^{(n)} \in 0_j, \end{cases}$$

where r_{1_j} and r_{0_j} are explained below. We used a gradient descent method to maximize (10) with respect to A_j and B_j.

Platt's method incorporated the following two techniques to avoid overfitting to the training data. First, the estimated target probabilities, $r_{1_j} = (N_{1_j} + 1)/(N_{1_j} + 2)$ and $r_{0_j} = 1/(N_{0_j} + 2)$, were used instead of typical choices, $r_{1_j} = 1$ and $r_{0_j} = 0$, where N_{1_j} and N_{0_j} are the numbers of samples belonging to 1_j and to 0_j, respectively. This setting is effective especially in dealing with unbalanced training data sets. Second, cross validation was used for generating

unbiased training data of the sigmoidal fitting. It should be noted that a naively made transformed data set L'_j could be biased because of the effects from the optimization of the jth predictor. We used fivefold cross validation in this study. Namely, the training data set L_j was divided into five blocks, five binary predictors were trained by using four out of the five blocks, and then, $d_j(x)$ was evaluated on the remaining block for each of the five predictors. The concatenation of the evaluation over the five disjoint blocks was used as the unbiased transformed training data set L'_j.

APPENDIX B
PSEUDOCODES FOR MAP AND WMAP

To provide the procedures of MAP and WMAP in step-by-step manner, we present their pseudocodes here. Algorithm 1 is a pseudocode for MAP that estimates the multiclass membership probability from the set of binary membership probabilities.

Algorithm 1. MAP class membership probability estimation

1: **procedure** ESTIMATEP(q, p^0, w, C, B)
2: **if** p^0 is NULL **then**
3: **for all** $i \in C$ **do** ▷ Initialize class membership probability if p^0 is not specified
4: $p_i^0 \leftarrow 1/\#C$
5: **end for**
6: **end if**
7: $V^0 \leftarrow V(p^0|q, w, B)$ ▷ Calculate the initial objective function value according to (4)
8: $T \leftarrow 0$
9: **repeat**
10: $T \leftarrow T + 1$
11: # Update p by the steepest descent method
12: **for all** $i \in C$ **do**
13: $\log p_i^T \leftarrow \log p_i^{T-1} + \alpha p_i^{T-1} \partial V(p|q, w, B)/\partial p_i|_{p=p^{T-1}}$ ▷ Step size α is determined by line search algorithm
14: **end for**
15: Normalize p^T so as to $\sum_{i \in C} p_i^T = 1$
16: $V^T \leftarrow V(p^T|q, w, B)$ ▷ Update the objective function value
17: **until** $T = MaxIter_p$ or $V^T - V^{T-1} < Threshold_p$ ▷ $MaxIter_p$ and $Threshold_p$ are arbitrary constants
18: $\tilde{p} \leftarrow p^T$
19: **return** $\tilde{p}$
20: **end procedure**

In this code, procedure ESTIMATEP takes a set of binary membership probability $q = \{q_j\}_{j \in B}$, an arbitrary initial value of multiclass membership $p^0 = \{p_i^0\}_{i \in C}$, a fixed weight vector $w = \{w_j\}_{j \in B}$, the set of class labels $C = \{1, \ldots, M\}$, and the code matrix B as inputs, and outputs the estimated multiclass membership $\tilde{p}$ according to the steepest descent method. Since $p_i > 0$, we used the gradient along $\log p$, which stabilizes the optimization (Algorithm 1, line 12). If we do not specify p^0, elements of p^0 are automatically set by the uniform probability $1/\#C$ (Algorithm 1, lines 2-6).

Algorithm 2 is a pseudocode for WMAP, which optimizes the weights for the given code matrix B.

Algorithm 2. WMAP weight optimization

1: **procedure** TRAINW ($Q = \{q^{(n)}\}_{n=1,\ldots,N}, T = \{t^{(n)}\}_{n=1,\ldots,N}, C, B$)
2: **for all** $j \in B$ **do** ▷ Initialize weight $w = \{w_j\}_{j \in B}$
3: $w_j^0 \leftarrow 1/\#B$
4: **end for**
5: $P^0 \leftarrow$ ESTIMATEPALL(Q, NULL, w^0, C, B)
6: $U^0 \leftarrow U(P^0, T)$ ▷ Calculate the initial objective function value according to (5)
7: $T \leftarrow 0$
8: **repeat**
9: $T \leftarrow T + 1$
10: $w^T \leftarrow \arg\max U(P^{T-1}, T)$ under $\sum_{j \in B} w_i^T = 1$ # Update w by a gradient ascent method based on gradient (12)
11: $P^T \leftarrow$ ESTIMATEPALL(Q, P^{T-1}, w^T, C, B) # Update P with new weights
12: $U^T \leftarrow U(P^T, T)$ ▷ Update the objective function value
13: **until** $T > MaxIter_w$ or $U^T - U^{T-1} < Threshold_w$ ▷ $MaxIter_w$ and $Threshold_w$ are arbitrary constants
14: $\tilde{w} \leftarrow w^T$
15: **return** $\tilde{w}$
16: **end procedure**

17: **procedure** ESTIMATEPALL ($\{q^{(n)}\}_{n=1,\ldots,N}, \{p^{(n)}\}_{n=1,\ldots,N}, w, C, B$)
18: **for** $n = 1$ to N
19: $p'^{(n)} \leftarrow$ ESTIMATEP($q^{(n)}, p^{(n)}, w, C, B$)
20: **end for**
21: **return** $\{p'^{(n)}\}_{i=1,\ldots,N}$
22: **end procedure**

This code consists of two procedures: a main procedure TRAINW and an auxiliary procedure ESTIMATEPALL, which is a wrapper of ESTIMATEP of Algorithm 1. TRAINW takes a set of binary membership $Q = \{q^{(n)}\}_{n=1,\ldots,N}$ of N samples, the corresponding true class label vectors $T = \{t^{(n)}\}_{n=1,\ldots,N}$, C, and B, and outputs the optimized weight vector $\tilde{w}$. ESTIMATEPALL is used for the inner optimization given by (8), which updates the multiclass membership probabilities of all samples $P = \{p^{(n)}\}_{n=1,\ldots,N}$ for the current weight estimate w^T (Algorithm 2, lines 8-13). TRAINW performs the outer optimization given by (7).

APPENDIX C
DERIVATION OF WMAP METHOD

The optimization of (8) can be simply executed by the MAP method, but we need a technique to optimize (7) because U depends on w indirectly through $\tilde{p} \equiv \{\tilde{p}^{(n)}\}$. We define a function $f(w, p)$ of w and $p \equiv \{p^{(n)}\}$:

$$f(w, p) \equiv \frac{\partial}{\partial p} \tilde{V}(p|w),$$

where $\tilde{V}$ is the sum of V and the Lagrange multiplier term. The stationary condition of $\tilde{V}$ with respect to p:

$$f(w, \tilde{p}) = 0$$

provides the $\tilde{p}(w)$ that satisfies (8). Then,

$$\begin{aligned} & f(w+dw, \tilde{p}+dp) \\ &= f(w,\tilde{p}) + \frac{\partial}{\partial w} f(w+\theta dw, \tilde{p}+\theta dp) dw \\ &\quad + \frac{\partial}{\partial \tilde{p}} f(w+\theta dw, \tilde{p}+\theta dp) dp \\ &= \sum_{j\in B} \frac{\partial f}{\partial w_j} dw_j + \sum_{i\in C} \sum_{n=1}^{N} \frac{\partial f}{\partial p_i^{(n)}} dp_i^{(n)} \\ &= 0 \end{aligned} \tag{11}$$

gives an another solution of (8), $\tilde{p}+dp$, when w has an infinitesimal change dw, where $0<\theta<1$.

For description simplicity, we introduce matrix $A=\{a(j,\mu)\}$, where indices j and μ correspond to a target $j\in B$ and an element $p_i^{(n)}$ of p, respectively. Each element of matrix A is defined by

$$a(j,(i,n)) = \frac{\partial^2 \tilde{V}}{\partial w_j \partial p_i^{(n)}}.$$

We also introduce a square matrix $H=\{h(\mu,\mu')\}$ whose element is defined by

$$h((i,n),(i',n')) = \frac{\partial^2 \tilde{V}}{\partial p_i^{(n)} \partial p_{i'}^{(n')}},$$

where μ and μ' index an element $p_i^{(n)}$ of p. By using these notations, solution condition (11) is expressed in an implicit function as

$$A dw + H dp = 0,$$

and when $dw \to \mathbf{0}$,

$$\left\{ \frac{dp_i^{(n)}}{dw_j} \right\} \equiv \frac{dp}{dw} = -H^{-1}A.$$

Using this derivative,

$$\frac{\partial U}{\partial w} = \frac{\partial \tilde{p}}{\partial w}\frac{\partial U}{\partial \tilde{p}} = -H^{-1}A\frac{\partial U}{\partial p}. \tag{12}$$

Each element of $\partial U/\partial p$ is written as

$$\frac{\partial U}{\partial p_i^{(n)}} = \left(1 - \frac{\exp(\beta p_i^{(n)})}{\sum_{i'\in C}\exp(\beta p_{i'}^{(n)})}\right) \frac{\exp(\beta p_i^{(n)})}{\sum_{i'\in C}\exp(\beta p_{i'}^{(n)})} \beta t_i^{(n)}, \tag{13}$$

and then, (7) can be optimized by a gradient ascent method based on gradient (12).

Appendix D

Sparse Random Coding

The SR, which enables us to design efficient initial code matrices for large-class problems, was proposed by Allwein et al. [15]. Code matrices of SR for M-class problems consist of nonoverlapping $l = \lceil 15 \log_2 M \rceil$ targets (row vectors). Each element was assigned a value from {"1," "0," "$*$"} with certain probabilities; "1" or "0" with 1/4 or "$*$" with 1/2. To obtain good error correcting properties, the minimum and averaged distance between each pair of code words (rows in the code matrix) should be large. For calculating distance of a pair of code words $u, v \in \{"1," "0," "*"\}^{l\times 1}$, we used a generalized Hamming distance:

$$\rho = \sum_{i=1}^{l} r_i, \quad \text{where } r_i = \begin{cases} 0 & \text{if } u_i = v_i \wedge u_i \neq 0 \wedge v_i \neq 0, \\ 1 & \text{if } u_i \neq v_i \wedge u_i \neq 0 \wedge v_i \neq 0, \\ 0.5 & \text{if } u_i = 0 \vee v_i = 0. \end{cases}$$

After generating 10,000 code matrices according to the process above, we selected the optimal code matrix whose minimum ρ value was the maximal among them, checking that no column or row contained only "$*$."

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. This work was partly supported by the 21st Century COE Research Program: Exploiting New Frontiers in Bioscience and by a Grant-in-Aid for Scientific Research on Priority Areas: Deepening and Expansion of Statistical Mechanical Informatics, both from the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

References

[1] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, no. 5439, pp. 531-537, Oct. 1999.

[2] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P.S. Meltzer, "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," *Nature Medicine*, vol. 7, no. 6, pp. 673-679, June 2001.

[3] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander, and T.R. Golub, "Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures," *Proc. Nat'l Academy of Sciences*, vol. 98, no. 26, pp. 15149-15154, Dec. 2001.

[4] I. Hedenfalk, M. Ringner, A. Ben-Dor, Z. Yakhini, Y. Chen, G. Chebil, R. Ach, N. Loman, H. Olsson, P. Meltzer, A. Borg, and J. Trent, "Molecular Classification of Familial non-BRCA1/BRCA2 Breast Cancer," *Proc. Nat'l Academy of Sciences*, vol. 100, no. 5, pp. 2532-2537, Mar. 2003.

[5] B. Schoelkopf, C. Burges, and V. Vapnik, "Extracting Support Data for a Given Task," *Proc. First Int'l Conf. Knowledge Discovery and Data Mining*, pp. 252-257, 1995.

[6] B. Schoelkopf, C. Burges, and A. Smola, *Advances in Kernel Methods Support Vector Learning*. MIT Press, 1999.

[7] T.G. Dietterich and G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes," *J. Artificial Intelligence Research*, vol. 2, pp. 263-286, 1995.

[8] E.L. Allwein, R.E. Schapire, and Y. Singer, "Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers," *Proc. 17th Int'l Conf. Machine Learning*, pp. 9-16, 2000.

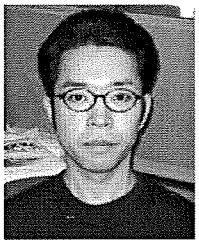
[9] T. Hastie and R. Tibshirani, "Classification by Pairwise Coupling," *Advances in Neural Information Processing Systems*, vol. 10, pp. 507-513, 1998.

[10] B. Zadrozny, "Reducing Multiclass to Binary by Coupling Probability Estimates," *Advances in Neural Information Processing Systems*, vol. 14, pp. 1041-1048, 2001.

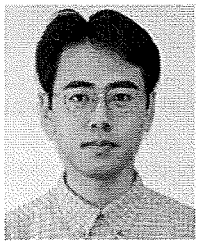
[11] T. Li, C. Zhang, and M. Ogihara, "A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Based on Gene Expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429-2437, Oct. 2004.

[12] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631-643, 2005.

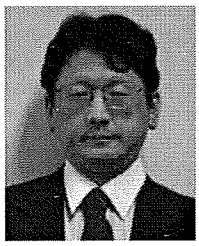
[13] J. Weston and C. Watkins, "Multi-Class Support Vector Machine," technical report, Univ. of London, 1998.
[14] K. Crammer and Y. Singer, "On the Algorithmic Implementation of Multiclass Kernel-Based Vector Machines," *J. Machine Learning Research*, vol. 2, pp. 265-292, 2001.
[15] E.L. Allwein, R.E. Schapire, and Y. Singer, "Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers," *J. Machine Learning Research*, vol. 1, pp. 113-141, 2001.
[16] L. Shen and E.C. Tan, "Reducing Multiclass Cancer Classification to Binary by Output Coding and SVM," *Computational Biology and Chemistry*, vol. 30, no. 1, pp. 63-71, Feb. 2006.
[17] J. Platt, "Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods," *Advances in Large Margin Classifiers*, A.J. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, eds., pp. 61-74, 2000.
[18] K. Kato, "Adaptor-Tagged Competitive PCR: A Novel Method for Measuring Relative Gene Expression," *Nucleic Acids Research*, vol. 25, no. 22, pp. 4694-4696, Nov. 1997.
[19] E. Saxen, K. Franssila, O. Bjarnason, T. Normann, and N. Ringertz, "Observer Variation in Histologic Classification of Thyroid Cancer," *Acta Pathologica et Microbiologica Scandinavica A*, vol. 86A, no. 6, pp. 483-486, Nov. 1978.
[20] A.S. Fassina, M.C. Montesco, V. Ninfo, P. Denti, and G. Masarotto, "Histological Evaluation of Thyroid Carcinomas: Reproducibility of the WHO Classification," *Tumori*, vol. 79, no. 5, pp. 314-320, Oct. 1993.
[21] Z.W. Baloch, S. Fleisher, V.A. LiVolsi, and P.K. Gupta, "Diagnosis of Follicular Neoplasm: A Gray Zone in Thyroid Fine-Needle Aspiration Cytology," *Diagnostic Cytophathology*, vol. 26, no. 1, pp. 41-44, Jan. 2002.
[22] K. Kato, R. Yamashita, R. Matoba, M. Monden, S. Noguchi, T. Takagi, and K. Nakai, "Cancer Gene Expression Database (CGED): A Database for Gene Expression Profiling and Accompanying Clinical Information of Human Cancer Tissues," *Nucleic Acids Research*, vol. 33, pp. D533-D536, 2005.
[23] K. Taniguchi, T. Takano, A. Miyauchi, K. Koizumi, Y. Ito, Y. Takamura, M. Ishitobi, Y. Miyoshi, T. Taguchi, Y. Tamaki, K. Kato, and S. Noguchi, "Differentiation of Follicular Thyroid Adenoma from Carcinoma by Gene Expression Profiling with Adapter-Tagged Competitive Polymerase Chain Reaction," *Oncology*, vol. 69, pp. 428-435, 2005.
[24] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, and S.J. Korsmeyer, "MLL Translocations Specify a Distinct Gene Expression Profile that Distinguishes a Unique Leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41-47, Jan. 2002.
[25] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proc. Nat'l Academy of Sciences*, vol. 99, no. 10, pp. 6567-6572, May 2002.
[26] M. Ohira, S. Oba, Y. Nakamura, E. Isogai, S. Kaneko, A. Nakagawa, T. Hirata, H. Kubo, T. Goto, S. Yamada, Y. Yoshida, M. Fuchioka, S. Ishii, and A. Nakagawara, "Expression Profiling Using a Tumor-Specific cDNA Microarray Predicts the Prognosis of Intermediate Risk Neuroblastomas," *Cancer Cell*, vol. 7, no. 4, pp. 337-350, Apr. 2005.
[27] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data," *Bioinformatics*, vol. 16, no. 10, pp. 906-914, evaluation studies, Oct. 2000.
[28] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *J. Am. Statistical Assoc.*, vol. 97, pp. 77-87, 2002.
[29] Y. Freund and R. Schapire, "Experiments with a New Boosting Algorithm," *Proc. Int'l Conf. Machine Learning (ICML '96)*, pp. 148-156, 1996.
[30] I. Guyon, J. Weston, S.M.D. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, vol. 46, pp. 389-422, 2002.



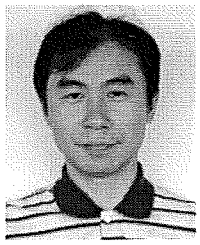
Naoto Yukinawa received the BS degree in bioscience from the Tokyo Institute of Technology in 2001 and the PhD degree from the Nara Institute of Science and Technology in 2006, where he studied the statistical machine learning approaches for system identification and classification of gene expression profiles. Since 2006, he has been working as a researcher in the group of Dr. Shin Ishii at the Nara Institute of Science and Technology. His current research interests include machine learning and their applications to transcriptomic and proteomic analyses.



Shigeyuki Oba received the MS degree in geophysics from Kyoto University in 1998 and the MS and PhD degrees in information science from the Graduate School of Information Science, Nara Institute of Science and Technology in 2001 and 2002, respectively. He has been in the Nara Institute of Science Technology since 2002 as a researcher and since 2003 as an assistant professor. His current research interests are machine learning and their application to bioinformatics.



Kikuya Kato received the MD degree and the PhD degree in molecular genetics and biochemistry from Osaka University Medical School in 1980 and 1984, respectively. He is currently the director of the Research Institute, Osaka Medical Center for Cancer and Cardiovascular Diseases. His current interest is molecular genetics and transcriptome analysis of human cancer tissues.



Shin Ishii received the BE, ME, and PhD degrees from the University of Tokyo in 1986, 1988, and 1997, respectively. He is currently a professor at the Graduate School of Informatics, Kyoto University, after a 10-year career at the Graduate School of Information Science, Nara Institute of Science and Technology. He has been interested in statistical bioinformatics and systems neurobiology and has approached these areas from both basic and practical viewpoints.

Int J Clin Exp Pathol (2009) 2, 154-162
www.ijcep.com/IJCEP806007

Original Article

Putative Precursor Cancer Cells in Human Colorectal Cancer Tissue

Teodora E Goranova[1], Masayuki Ohue[2] and Kikuya Kato[1]

[1]Research Institute and [2]Department of Surgery, Osaka Medical Center for Cancer and Cardiovascular Diseases, Higashinari-ku, Osaka, Japan

Received 11 June 2008; Accepted 1 July 2008; Available online 14 July 2008

Abstract: Multistage carcinogenesis is an important concept in cancer biology. Each new stage is triggered by the acquisition of an additional genetic aberration, leading to clonal expansion of the cancer cell. The resulting tumor mass consists of cancer cells with all genetic aberrations, but may include precursor cells at some point of carcinogenesis. We analyzed six colorectal cancer tissues with *APC*, *K-ras*, and *p53* mutations. From each sample, 40–50 areas (100×100×40μm) consisting only of cancer cells were microdissected, and genomic DNA was purified. Ratios of mutated and normal alleles were quantitated by the SNaPshot assay, a primer extension assay. In five tumor tissues, we identified cancer cell subpopulations corresponding to putative precursors, i.e., cells with mutations in one or two of the three genes. All samples were likely to be of monoclonal origin, and temporal sequences of the mutations could be deduced from the mutation patterns of putative precursors. The orders of mutation events were variable. However, the two carcinoma tissues accompanying adenoma regions started with the APC mutation, not contradicting the previous studies. The analysis also revealed considerable heterogeneity in allele ratios of one or two of the chromosomes. The current findings are promising to uncover the process of carcinogenesis directly from the tumor tissue of the patient.

Key Words: carcinogenesis, somatic mutation, intratumor heterogeneity, chromosome copy number variation, cancer stem cell

Introduction

Carcinogenesis is a multistage process in which an initial population of slightly abnormal cells, descendants of a single mutant ancestor, evolves through successive cycles of mutation and natural selection [1]. The model of Vogelstein and colleagues on the temporal sequence of genetic events in colorectal cancer is known as a typical example [2]. According to this model, each new stage is triggered by the acquisition of an additional genetic aberration which brings a growth or survival advantage to the cell. Eventually, this leads to clonal expansion of the cell and overgrowth over the other cells. This process may yield a tumor mass including cancer cells at some point of carcinogenesis as the minor population, leading to intratumoral heterogeneity in mutation patterns. There have been a considerable number of studies demonstrating intratumor genetic heterogeneity in human colorectal [3-8], prostate [9], breast [10], ovarian [11], and cervical [12] cancers. These studies excised sections from different areas of the tumor mass and compared mutation or chromosomal aberration patterns. One of the studies correlated the degree of heterogeneity with the evolutional process of cancer [7]. However, the origin of the heterogeneity is not clear from these studies.

In the present study, we focused on mutations in three genes, i.e., *APC*, *K-ras* and *p53*, and examined whether the tumor tissues comprised cancer cells at some point of carcinogenesis. To improve the resolution of the assay, we reduced the size of the excised sections from the reported sizes of previous studies by more than two orders of magnitude. In addition, applying the principle of competitive PCR, we quantitated relative abundance of each allele. The analysis revealed that five out of six tissues had cancer cells with one or two of the mutated genes,

corresponding to the precursors of cancer cells with three mutations. The identification of precursor cancer cells, defined here as cancer cells lacking a part of carcinogenic mutations, will enable us to trace the history of carcinogenesis.

Materials and Methods

Samples

Seventy-nine colorectal carcinoma tissues from our tumor tissue bank were screened for mutations in the coding regions of *APC*, *K-ras* and *p53* genes. DNA was extracted from frozen bulk tumor tissues by QIAamp DNA Micro kit (Qiagen). Coding regions of *APC*, *K-ras* and *p53* genes were screened for mutations by High-Resolution Melting on LightScanner (Idaho Technology Inc.). Samples with aberrant melting curves were analyzed by direct sequencing with BigDye Terminator Cycle Sequencing Kit (version 3.1, Applied Biosystems, USA) on ABI PRISM 3730 (Applied Biosystems, USA). Eleven out of 79 samples had mutations in all three genes. We selected six colorectal carcinoma samples with the largest size for the next experiment. The study was approved by the ethical committee of Osaka Medical Center for Cancer and Cardiovascular Diseases. Informed consents were obtained from all patients.

Laser Microdissection and DNA Extraction

Sections (40μm thick) from frozen cancer tissues of six patients were prepared on Leica CM1900 cryostat (Leica Microsystems). After mounting on a film-covered glass slide, the sections were stained with Mayer's hematoxylin (Wako). Microdissection was performed using Leica AS LMD system (Leica Microsystems). Genomic DNA was extracted by prepGEM kit (ZyGEM) according to the protocol; 20μl DNA mixture was prepared from each sample.

PCR Amplification and SNaPshot Assay

DNA fragments containing mutations were amplified by multiplex PCR on GeneAmp PCR System 9700 (Applied Biosystems). PCR mixture included 5μl DNA (250pg), 1xPCR buffer (Applied Biosystems), 2mM $MgCl_2$, 200μM each dNTP, Primer mix (0.2μM each primer, Supplemental Table S1) and 1U AmpliTaqGold polymerase (Applied Biosystems) in a 10 μl reaction. Cycling conditions were as follows: denaturing at 94°C for 5 min; 40 cycles of denaturing at 94°C for 30 s, annealing at 54-56°C for 30 s and synthesis at 72°C for 40 s; and final synthesis at 72°C for 5 min. SNaPshot assay is a primer extension assay: each primer is designed to bind to a complementary template right in front of the mutation site. Reaction is carried out in the presence of fluorescently labeled ddNTPs and DNA polymerase extends the primer by one nucleotide, adding a single ddNTP to its 3' end. Fluorescent dyes used for dideoxynucleotides are as follows: A, dR6G; C, dTAMRA; G, dR110; and T, dROX.

PCR fragments for primer extension were prepared by incubating 7.5 μl PCR product with 0.5 U shrimp alkaline phosphatase (TaKaRa) and 1 U exonuclease I (TaKaRa) for 40 min at 37°C in a final volume of 10 μl, followed by the inactivation of the enzymes for 20 min at 80°C. Primer extension was carried out in 5 μl containing 2 μl of treated PCR product, 2.5 μl ABI Prism SNaPshot Multiplex kit (Applied Biosystems) and 0.5 μl extension primers mix (0.2μM each primer). Primer sequences are shown in Supplemental Table S1. Cycling conditions were according to the manufacturer's protocol: 25 cycles of 10 s at 96°C denaturation, 5 s at 50°C annealing, and 30 s at 60°C extension. To remove unincorporated ddNTPs, 5 μl of SNaPshot products were incubated for 40 min at 37°C with 0.5 U shrimp alkaline phosphatase (TaKaRa) in a final volume of 6 μl, and the enzyme was deactivated as described above. A total of 1 μl of treated SNaPshot reaction was denatured in 9 μl of distilled water (in the presence of standard-LIZ 120) for 5 min at 95°C, and was analyzed on ABI PRISM 3100 Genetic Analyzer (Applied Biosystems). The fragment analysis was performed with Peak Scanner Software v1.0 (Applied Biosystems).

The mutated allele ratio, M/(M+N), was calculated where M is mutant allele peak height and N is normal peak height. Reproducibility of the amplification and the SNaPshot assay was checked with two series of experiments. All data are supplied as Supplemental Table S2. We plotted the results on 3D graphs using Grafis software (ver.2.9.22, Kylebank Software Ltd.). In most cases of allelic loss, the corresponding peak

Table 1 Sequence alterations identified in colorectal cancer tissues of the patients

Sample No.	*APC*		*K-RAS*		*P53*	
	Exon	Locus	Exon	Locus	Exon	Locus
1	16	c.2932 C>T Q978*	2	c.35 G>A G12D	7	c.767 C>A T256K
3	16	c.3956 delC P1319fs*2	2	c.34 G>A G12S	8	c.818 G>A R273H
33	16	c.2755 A>T R919*	2	c.38 G>A G13D	8	c.824 G>T C275F
41	16	c.4044 insA	3	c.204 G>C R68S	6	c.659 A>G Y220C
65	14	c.1690 C>T R564*	2	c.35 G>A G12D	6	c.659 A>G Y220C
74	16	c.2626 C>T R876*	2	c.35 G>T G12V	4	c.374 C>T T125M

*stop codon

height was zero. However, there were several areas with residual peaks, where we set the threshold as 0.05.

LOH Analysis

Amplification of microsatellite markers on chromosome 5q (D5S107, D5S82 and D5S346) and those on chromosome 17p (D17S796 and D17S786) was performed in two separate multiplex PCR reactions. PCR mixture included 5μl DNA, 1xPCR buffer (Applied Biosystems), 2mM $MgCl_2$, 200μM each dNTP, Primer mix (**Supplemental Table S1**) and 1U AmpliTaqGold polymerase (Applied Biosystems) in a 10 μl reaction. Cycling conditions were as follows: denaturing at 94°C for 5 min; 40 cycles of denaturing at 94°C for 30 s, annealing at 55°C for 30 s and synthesis at 72°C for 30 s; final synthesis at 72°C for 30 min. The fluorescent products were analyzed on ABI PRISM 3100 Genetic Analyzer (Applied Biosystems). The fragment analysis was performed with Peak Scanner Software v1.0 (Applied Biosystems).

Results

Outline of the Method

From our tumor tissue bank, we selected six colorectal cancer tissues carrying mutations in *APC*, *K-ras*, and *p53*, identified by analysis of bulk tissues. The details of the mutations are listed in **Table 1**. Although there may be additional mutations not detected by the bulk tissue analysis, we focused on these mutations for detailed analysis. For the analysis of genetic heterogeneity, we excised 40-50 small areas containing only cancer cells from frozen tissue sections, and purified genomic DNA. The sampling was random, but we avoided repeated sampling from the same cryptic region. After simultaneous amplification of three genes with multiplex PCR, mutation status was quantitatively determined with the SNaPshot assay, a primer extension assay. The illustration of the method is shown in **Figure 1.**

In our previous study on lung cancer [13], we determined the smallest amount of tissue section that enabled stable and unbiased PCR amplification. We set the size of the section as 100×100×40 μm, a slight increase in thickness from 35 μm. This scale was less than 1/100 of those used in previous studies: for example, 5×5×5 mm^3 in [12]. We used one-fourth of the genomic DNA purified from the excised section for a single multiplex PCR reaction.

To confirm the quantitative recovery of PCR products in our protocol, we performed the following experiment. We purified genomic DNA from the colorectal tumor section (0.01 mm^2 each) from two other patients: one with homozygous T for SNP in *TGFR2* (rs2228048) and one with homozygous C. Samples with various amount ratios of the two genomic DNA (1:0, 3:1, 1:1, 1:3, 0:1) were prepared, maintaining the total amount of DNA as that used for the single multiplex PCR reaction. To

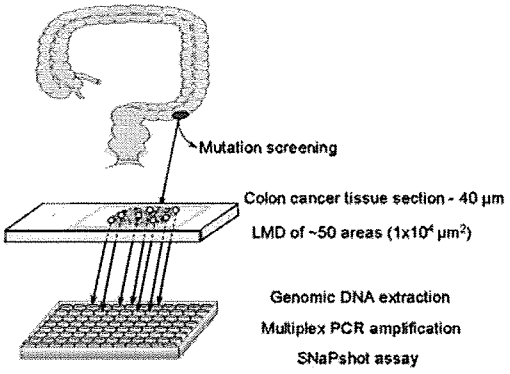


Figure 1 Graphic representation of the method.

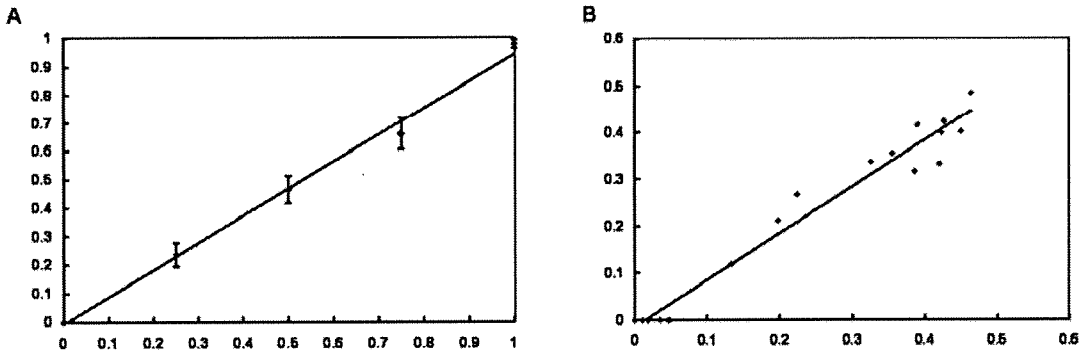


Figure 2 A. Quantitative accuracy of the SNaPshot assay of amplified products. Horizontal axis indicates the ratio of the mutant allele (T at rs2228048) of *TGFR2* in the template; vertical axis indicates the ratio of the mutant allele of *TGFR2* in the amplified product, i.e., M/(M+N), where M is mutant peak height, and N is normal peak height. The error bars correspond to the standard deviations of ten experiments. **B.** Contamination of normal cells in the excised section areas. Data obtained from areas including various fractions of normal cells among cancer cells were plotted. Horizontal axis indicates mutant allele ratios (M/(M+N)) of *APC*; vertical axis indicates that of *K-ras*.

simulate multiplex PCR reaction, we simultaneously amplified *TGFR2*, *K-ras* and *APC*. Then, allele ratios of amplified *TGFR2* were determined with the SNaPshot assay. There was a good correlation between the amount ratio in the templates and the ratio in amplified products (**Figure 2A**), assuring unbiased amplification.

Although we set the size of the excised section as small as possible, it consisted of 180-200 cells. One may argue that possible contamination of normal cells may lead to erroneous conclusions. We performed the following experiment: from a colorectal cancer tissue, we excised twenty areas (0.01 mm^2 each), which included various numbers of

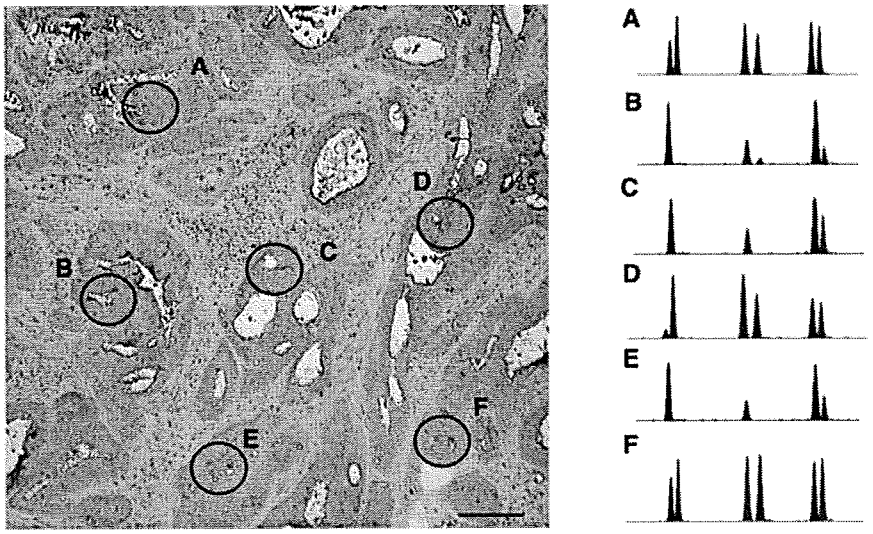


Figure 3 An example of intratumor genetic heterogeneity (Sample 1). A microscopic view of a colorectal cancer tissue section (left) with black circles indicating microdissected areas; (right), electropherograms of the SNaPshot assay. The first two peaks represent the mutation status of *p53* (C>A); third and fourth - that of *K-ras* (G>A); and the last two peaks - that of *APC* (C>T). Black peak, fragment amplified with ddC; green peak, fragment amplified with ddA; blue peak, fragment amplified with ddG; red peak, fragment amplified with ddT. Red bar, 100μm.

normal cells. This colorectal cancer tissue had mutations in all of the three genes. After isolation of genomic DNAs, we amplified corresponding exons of *APC*, *K-ras* and *p53*, and quantitated allele ratios with the SNaPshot assay. The relative allele ratios of *APC* and *K-ras* were plotted (**Figure 2B**). In the case of normal cell contamination, ratio of the normal allele is proportional to the fraction of normal cells included in the microdissected region. Thus, ratios of normal/mutant alleles of the two genes are proportional as shown in **Figure 2B**. The same correlation was observed with *APC–p53* and *K-ras–p53* (data not shown). In general, a data point from a mixture of mature cancer cells and normal cells would lie around a line connecting data points derived from each of them. We can exclude the possibility that an aberrant mutation pattern is due to the normal cell contamination by this plotting. As shown below, no areas had this characteristic.

Intratumor Heterogeneity of APC, K-ras and p53 Mutations

We examined six colorectal cancer tissues and found that five tissues contained cancer cells with mutation types different from that of the major population. **Figure 3** shows an example (sample 1). Six areas are presented. Electropherograms of three areas (A, D, F) revealed mutations in three genes, i.e., *APC*, *p53*, and *K-ras*. One area had mutations in *APC* and *K-ras* (B), whereas the other two areas (C, E) had mutations only in *APC*.

In our experimental system, the relative ratios of mutated and normal alleles were measured (**Supplemental Table S2**). Thus, we plotted areas in three-dimensional spaces created by mutated allele ratios of *APC*, *K-ras* and *p53* (**Figure 4**, top graphs). We also presented two-dimensional plots with color graduation for APC allele ratio (**Figure 4**, bottom graphs). QuickTime movies of the three-dimensional plots will be posted on our web site at: http://genome.mc.pref.osaka.jp/data_download.html.

The major population of sample 1 (30/41 areas) carries all three mutations (**Figure 4A**). Chromosomal status is stable with *APC* and *K-ras*, but there is a distinct heterogeneity with *p53*. Besides this major population, there are two subpopulations: with *APC* and *K-ras* mutations (4/41), and with *APC* mutation alone (7/41).

Most of the excised areas from sample 3 carry mutations in all three genes – 46 out of 48 areas (**Figure 4B**). However, we found two areas with mutations in both *K-ras* and *p53* but not in the *APC* gene.

The major population of sample 33 has all three mutations, but *p53* and *K-ras* alleles have marked heterogeneity (**Figure 4C**). Besides, there is one subpopulation with *APC* and *K-ras* mutations (6/48) and another with *APC* mutation alone (18/48).

In sample 41, there are two minor populations: one with *p53* mutation alone and one with *K-ras* and *p53* mutations (**Figure 4D**). The former lost the normal *p53* allele, and the latter lost the normal *K-ras* allele, whereas the major population always maintained the *K-ras* normal allele, and often the *p53* normal allele as well. Subsequent loss of normal alleles after divergence of the major population is suspected.

There are three subpopulations of cancer cells in sample 65 (**Figure 4E**). *APC* alleles are highly homogeneous – we detected the *APC* mutation in all 49 areas. A subpopulation with only *APC* mutation (2/49) and another carrying both *APC* and *p53* aberrations (2/49) were identified.

Although we identified minor subpopulations in the above cases, sample 74 showed no heterogeneity in the mutation pattern (**Figure 4F**).

Order of Genetic Events

In somatic mutations in *APC*, *K-ras*, and *p53*, mutations in various loci evoke similar biological effects onto cells. Because the chance that the second mutation is introduced into the same locus is very low, these mutations should have been introduced by single events. Therefore, subpopulations of each sample should have been derived from a single ancestor cell, and the order of mutation events can be deduced from mutation patterns of subpopulations. For example, sample 1 had a subpopulation with *APC* and *K-ras* mutations, and another with *APC* mutation alone. The order of mutation events is

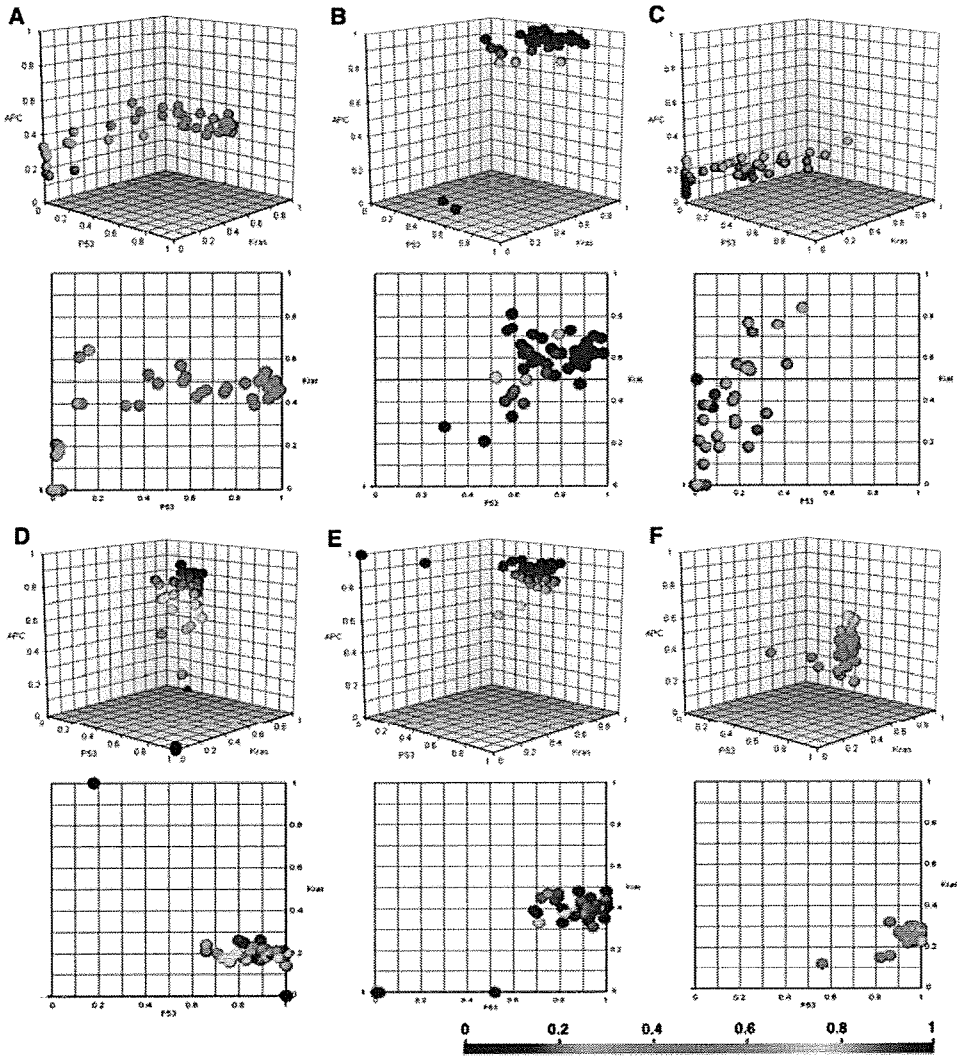


Figure 4 Genetic heterogeneity in six colorectal cancer tissues. The mutant allele ratios (M/(M+N)) of *p53*, *K-ras* and *APC* genes are plotted in the x, y and z axis, respectively. Each sphere represents a single area; the color indicates the mutant allele ratio of APC gene. The top graphs are 3D graphs; the bottom graphs show *p53* and *K-ras* only (the color is the same as in the 3D graph). **A.** Sample 1; **B.** Sample 3; **C.** Sample 33; **D.** Sample 41; **E.** Sample 65; **F.** Sample 74.

Table 2 Summary of the order of genetic events in the samples

Sample No.	Precursor genotype*	Order of genetic events	Accompanying adenoma
1	$APC^{M}K\text{-}RAS^{N}P53^{N}$ (7/41) $APC^{M}K\text{-}RAS^{M}P53^{N}$ (4/41)	*APC→K-RAS→P53*	Yes
3	$APC^{N}K\text{-}RAS^{M}P53^{M}$ (2/48)	*P53/K-RAS→APC*	No
33	$APC^{M}K\text{-}RAS^{N}P53^{N}$(18/48) $APC^{M}K\text{-}RAS^{M}P53^{N}$(6/48)	*APC→K-RAS→P53*	No
41	$APC^{N}K\text{-}RAS^{N}P53^{M}$(7/40) $APC^{N}K\text{-}RAS^{M}P53^{M}$(1/40)	*P53→K-RAS→APC*	No
65	$APC^{M}K\text{-}RAS^{N}P53^{N}$(2/49) $APC^{M}K\text{-}RAS^{N}P53^{M}$(2/49)	*APC→P53→K-RAS*	Yes
74	-	-	No

*N, normal; M, mutation; (areas/total)

therefore APC→K-ras→p53. The deduced order of mutation events is shown in **Table 2.**

A major concern is the possibility that the absence of mutation is due to loss of the mutated allele. Loss of the mutant allele was reported in the mismatch repair region [14], although it is not known with *APC*, *K-ras* and *p53*. We performed LOH analysis of 5q and 17p with relevant areas using microsatellite markers, and detected both alleles in all cases, excluding the possible loss of the mutated allele (**Supplemental Table S3**). During a review of archival HE sections, we found that in two samples adenocarcinoma was accompanied by adenoma regions (**Table 2**).

Discussion

The study of human carcinogenesis has been a difficult task due to unavailability of tumor samples of different stages, especially early stages, from the same patient. Thus, most models have been based on indirect evidence obtained by studies with patient populations [2, 15]. In the present study, we demonstrated that the primary colorectal cancer tissues contained cancer cells with part of the mutations found in the major population, which are likely to be precursors of the major population, and named them precursor cancer cells. In addition, from their mutation patterns, we deduced the order of genetic events. It is likely to be a common feature of colorectal cancer, because five out of six cases had such minor populations. The heterogeneity of somatic mutation patterns previously reported in colorectal carcinoma [3, 5, 7] is probably due to the precursor cancer cell. In colorectal adenoma, genetic heterogeneity of carcinogenic mutations was well established in the context of tumor evolution [16, 17]. The current findings imply that the heterogeneity found in early adenoma still remains in the later carcinoma stage, reducing its level. It should be noted that mutation patterns are not necessarily consistent in adjacent adenoma and carcinoma regions: *K-ras* mutations in the adenoma region were not found in the adjacent carcinoma in 24% of the cases [18]. This suggests that even adjacent carcinoma and adenoma arose from different ancestors. Also taking into account *de novo* carcinoma, it is important to collect information directly from carcinoma tissues.

Another important discovery is heterogeneity in allelic imbalance. It is interesting to note that the heterogeneity was usually restricted to some chromosomes. In sample 1, heterogeneity was found only with *p53*. In this context, it would be cautious to interpret results of comparative genomic hybridization (CGH) or array-CGH. CGH cannot discriminate cases with and without heterogeneity, presenting an averaged view for chromosome aberration with heterogeneity.

Because PCR from a small amount of DNA may lead to biased amplification, we carefully designed the whole experiment. In our previous study [13], we determined the minimum amount of tissue sections enabling unbiased quantitative amplification. Using the determined amount of section, we demonstrated quantitative recovery of amplified products under the condition used in this study. This excludes the possibility that the observed loss of mutant/normal allele is due to stochastic PCR reaction. In addition, we previously calculated the chance of erroneous identification of allelic loss by stochastic PCR: it was 3.5×10^{-5} and 6.04×10^{-7} in two different loci [13]. Except for one pattern of sample 41, all mutation patterns of precursors were identified with more than two areas, demonstrating reproducibility of the patterns.

There appears to be no rule in the deduced order of mutation events. It should be noted that *de novo* colorectal cancer is more frequent in Japanese than in other racial populations [19]. Investigators believe that this neoplasm develops through a pathway different from the polypoid one [19-21]. There were two cases (sample 1 and 65) in which studied carcinoma was accompanied by adenoma regions, and the *APC* mutation was the first event in both cases. Because the original Vogelstein model is for cancers developed from adenoma [2], our results do not contradict the previous studies based on patient populations. One sample (sample 74) was homogeneous – only one type of cells was found in it. According to the widely accepted theory, cancer tissue includes a dominant clone, which overgrows the other cells [1]. Although advanced cancer growth could be an explanation, it is not clear whether there was any other type of cells in this cancer tissue. There may have been a heterogeneous pattern in another part of the section or the tissue.

This approach for temporal sequence may be applied to other cancers, if precursor cancer cells exist. It should be noted that many genetic aberrations such as LOH occur frequently, and it is important to determine whether the aberration is from a single event or more than one event. Mutations in *APC* and *p53* do not need such attention, because the chance of having mutations at the same base is very low. For chromosomal aberrations, probably use of multiple markers would solve the problem.

Most models of carcinogenesis assume that tumors are monoclonal in origin. This conclusion is based largely on studies using X chromosome-linked markers in females [22]. However, a recent study demonstrated relatively large sizes of X-inactivation patches in normal tissues, confounding assessment of early studies [23]. Our results strongly suggest the monoclonal origin: the tumor tissues contained cancer cells with several different genetic types, of which mutation patterns indicated monoclonal origin. However, there is still possibility of other clones not detected by the above number of sampling.

Our results suggest that putative precursor cancer cells exist in the tumor mass surgically dissected, and we can perform molecular analysis through their purification and subsequent culture. One intriguing question is whether they are the same as cells that differentiated into cancer cells with the three mutations in the past. As they still exist as a minor population, their growth rate is smaller than the major population, suggesting no additional mutation for growth advantage. The smaller number of replication cycles indicates less number of new mutations. Thus, we suspect that their biological properties would not change. Another question is whether the precursor cancer cells are malignant or benign. Most adenoma cells accompany *APC* mutations, and precursor cancer cells with *APC* mutations might be reminiscent of adenoma cells.

Detailed comparison of mutation patterns in the primary tumor and metastasis revealed that a considerable number of cases had different mutation patterns [24, 25]. Unmatched mutation patterns indicate that metastatic lesions would be derived from a minor population in the primary tumor, suggesting possible involvement of precursor cancer cells. Additional experiments with primary tumor and its paired metastases are required to elucidate this issue.

We believe that finding putative precursor cancer cells is important for understanding carcinogenesis as well as for future drug discovery. If they have distinct molecular characteristics, it would be beneficial to develop anti-cancer drugs targeting them. In particular, due to possible involvement in metastasis, such drugs are of interest for their possible anti-metastasis activities.

Acknowledgements

The authors thank Dr. Kazuya Taniguchi for his advice on the experiments. This work was partly supported by the Knowledge Cluster Initiative (the Keihanna Science City area) of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Please address all correspondences to Kikuya Kato, MD, PhD, Professor and Director, Research Institute, Osaka Medical Center for Cancer and Cardiovascular Diseases, 1-3-2 Nakamichi, Higashinari-ku, Osaka, 537-8511, Japan. Tel:+81-6-6972-1181 (ext. 4105, 4304); Fax: +81-6-6973-1209; Email: katou-ki@mc.pref.osaka.jp

References

[1] Alberts B, Johnson A, Lewis J, Raff M, Roberts K and Walter P. Molecular Biology of the Cell. Fourth Edition; New York and London, Garland Science, 2002, available online at http://www.ncbi.nlm.nih.gov

[2] Fearon ER and Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* 1990;61: 759-767.

[3] Baisse B, Bouzourene H, Saraga EP, Bosman FT and Benhattar J. Intratumor genetic heterogeneity in advanced human colorectal adenocarcinoma. *Int J Cancer* 2001;93:346-352.

[4] Barnetson R, Jass J, Tse R, Eckstein R, Robinson B and Schnitzler M. Mutations associated with microsatellite unstable colorectal carcinomas exhibit widespread intratumoral heterogeneity. *Genes Chromosomes Cancer* 2000;29:130-136.

[5] Giaretti W, Rapallo A, Sciutto A, Macciocu B, Geido E, Hermsen MA, Postma C, Baak JP, Williams RA and Meijer GA. Intratumor heterogeneity of k-ras and p53 mutations among human colorectal adenomas containing early cancer. *Anal Cell Pathol* 2000;21:49-57.

[6] Gonzalez-Garcia I, Sole RV and Costa J. Metapopulation dynamics and spatial heterogeneity in cancer. *Proc Natl Acad Sci USA* 2002;99:13085-13089.
[7] Losi L, Baisse B, Bouzourene H and Benhattar J. Evolution of intratumoral genetic heterogeneity during colorectal cancer progression. *Carcinogenesis* 2005;26:916-922.
[8] Samowitz WS and Slattery ML. Regional reproducibility of microsatellite instability in sporadic colorectal cancer. *Genes Chromosomes Cancer* 1999;26:106-114.
[9] Konishi N, Hiasa Y, Matsuda H, Tao M, Tsuzuki T, Hayashi I, Kitahori Y, Shiraishi T, Yatani R, Shimazaki J and Lin JC. Intratumor cellular heterogeneity and alterations in ras oncogene and p53 tumor suppressor gene in human prostate carcinoma. *Am J Pathol* 1995;147:1112-1122.
[10] Wild P, Knuechel R, Dietmaier W, Hofstaedter F and Hartmann A. Laser microdissection and microsatellite analyses of breast cancer reveal a high degree of tumor heterogeneity. *Pathobiology* 2000;68:180-190.
[11] Takeshima Y, Amatya VJ, Daimaru Y, Nakayori F, Nakano T and Inai K. Heterogeneous genetic alterations in ovarian mucinous tumors: application and usefulness of laser capture microdissection. *Hum Pathol* 2001;32:1203-1208.
[12] Lyng H, Beigi M, Svendsrud DH, Brustugun OT, Stokke T, Kristensen GB, Sundfor K, Skjonsberg A and De Angelis PM. Intratumor chromosomal heterogeneity in advanced carcinomas of the uterine cervix. *Int J Cancer* 2004;111:358-366.
[13] Taniguchi K, Okami J, Kodama K, Higashiyama M and Kato K. Intratumor heterogeneity of epidermal growth factor receptor mutations in lung cancer and its correlation to the response to gefitinib. *Cancer Sci* 2008;99:929-935.
[14] Sanchez de Abajo A, de la Hoya M, van Puijenbroek M, Godino J, Diaz-Rubio E, Morreau H and Caldes T. Dual role of LOH at MMR loci in hereditary non-polyposis colorectal cancer? *Oncogene* 2006;25:2124-2130.
[15] Diep CB, Kleivi K, Ribeiro FR, Teixeira MR, Lindgjaerde OC and Lothe RA. The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes. *Genes Chromosomes Cancer* 2006;45:31-41.
[16] Shibata D, Schaeffer J, Li ZH, Capella G and Perucho M. Genetic heterogeneity of the c-K-ras locus in colorectal adenomas but not in adenocarcinomas. *J Natl Cancer Inst* 1993;85:1058-1063.
[17] Shih IM, Wang TL, Traverso G, Romans K, Hamilton SR, Ben-Sasson S, Kinzler KW and Vogelstein B. Top-down morphogenesis of colorectal tumors. *Proc Natl Acad Sci USA* 2001;98:2640-2645.
[18] Zauber NP, Sabbath-Solitare M, Marotta SP and Bishop DT. K-ras mutation and loss of heterozygosity of the adenomatous polyposis coli gene in patients with colorectal adenomas with in situ carcinoma. *Cancer* 1999;86:31-36.
[19] Soetikno R, Friedland S, Kaltenbach T, Chayama K and Tanaka S. Nonpolypoid (flat and depressed) colorectal neoplasms. *Gastroenterology* 2006;130:566-576.
[20] Speake D, Biyani D, Frizelle FA and Watson AJ. Flat adenomas. *ANZ J Surg* 2007;77:4-8.
[21] Uronis JM, Herfarth HH, Rubinas TC, Bissahoyo AC, Hanlon K and Threadgill DW. Flat colorectal cancers are genetically determined and progress to invasion without going through a polypoid stage. *Cancer Res* 2007;67:11594-11600.
[22] Vogelstein B, Fearon ER, Hamilton SR and Feinberg AP. Use of restriction fragment length polymorphisms to determine the clonal origin of human tumors. *Science* 1985;227:642-645.
[23] Novelli M, Cossu A, Oukrif D, Quaglia A, Lakhani S, Poulsom R, Sasieni P, Carta P, Contini M, Pasca A, Palmieri G, Bodmer W, Tanda F and Wright N. X-inactivation patch size in human female tissue confounds the assessment of tumor clonality. *Proc Natl Acad Sci USA* 2003;100:3311-3314.
[24] Albanese I, Scibetta AG, Migliavacca M, Russo A, Bazan V, Tomasino RM, Colomba P, Tagliavia M and La Farina M. Heterogeneity within and between primary colorectal carcinomas and matched metastases as revealed by analysis of Ki-ras and p53 mutations. *Biochem Biophys Res Commun* 2004;325:784-791.
[25] Zhang JS, Caplin S, Bosman FT and Benhattar J. Genetic diversity at the p53 locus between primary human colorectal adenocarcinomas and their lymph-node metastases. *Int J Cancer* 1997;70:674-678.

Int J Clin Exp Med (2009) 2, 193-202
www.ijcem.com/IJCEM906002

Review Article

Impact of the next generation DNA sequencers

Kikuya Kato

Research Institute, Osaka Medical Center for Cancer and Cardiovascular Diseases, 1-3-3 Nakamichi, Higashinari-ku, Osaka, 537-8511, Japan.

Received June 25, 2009; accepted June 31, 2009; available online July 8, 2009

Abstract: New generation sequencers have been developed with a strong impact on genomics. These sequencers are based on a principle different from the Sanger method, and can sequence one to several million templates in a single run, albeit read length is relatively small. The current large-scale efforts are: 1) complete genome sequencing of 1,000 individuals, the primary objective of which is identification of rare SNP variants, not identified by the international HapMap project; 2) large-scale sequencing of cancer genomes to construct a complete catalog of genomic changes. These sequencers are also being applied in the identification of new infectious agents. Steady increase in data production capacity and decrease of cost will definitely make the sequencers a powerful diagnostic tool, especially for screening of all genetic diseases. On the contrary, statistical problems inherent to large data sets need to be solved before application to specific problems in medical science.

Key words: Massive parallel analysis, the 1000 genomes project, the cancer genome atlas

Introduction

In the field of genomics, the next generation DNA sequencer is currently the hottest topic. These new sequencers can produce over 100 times more data compared to the most sophisticated capillary sequencers based on the Sanger method. The rapid developments of machines and bioinformatics are making the goal a "1,000 dollar genome sequence", i.e., sequencing individual human genomes at a cost of $ 1, 000 each. The entire scene of biomedical science may change when the goal has been reached.

In this review, I summarize the principle of the next generation sequencers, current applications, and their future prospects in medical science. The first generation sequencers refer to those based on the Sanger method, the second generation sequencers are those based on massive parallel analysis, and the third generation sequencers are those based on single molecule sequencing in addition to massive parallel analysis. Because the current excitement comes from the second-generation sequencers, I will show their basic principle first.

Principle of the second generation sequencer

Three second generation sequencers are commercially available: Roche FLX [1], Illumina Genome Analyzer (GA) [2], and Lifetechnologies' SOLiD [3, 4]. Those machines are widely distributed, and their performance has been well characterized. All sequencers are based on a similar principle.

1. Use PCR products from single molecules as templates. With FLX and SOLiD, PCR amplification is performed on microbeads using emulsion PCR so that PCR products from a single molecule are attached to a single bead. With FLX, each bead is located in a picoliter well. With GA, PCR amplification is performed on a slide glass, making "clusters" of PCR products derived from single molecules [5]. Cluster formation is more sophisticated because theoretically a higher density of templates can be achieved.

2. Sequence by repetitive reaction. Information

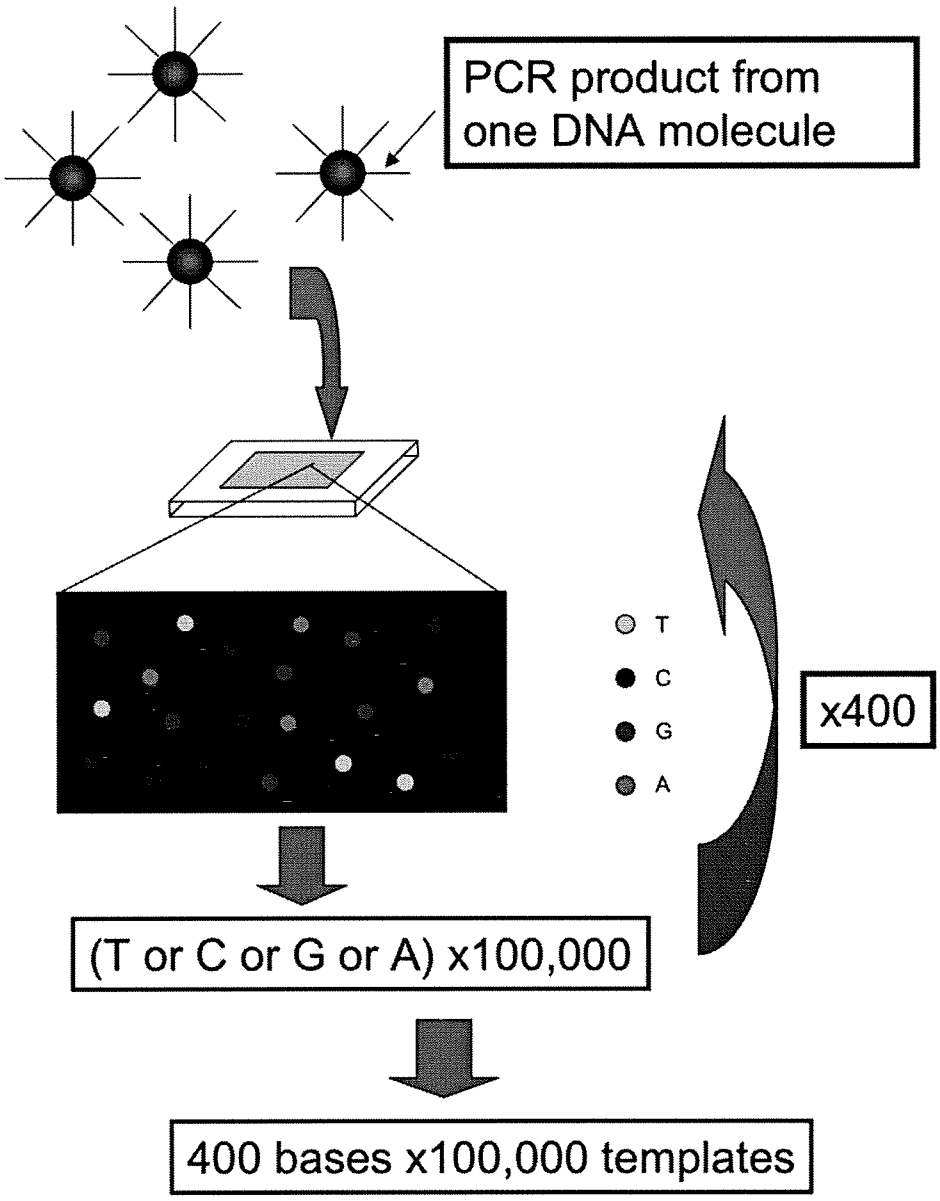


Figure 1. Schematic presentation of the principle of the second-generation sequencer. This scheme is based on Roche FLX.

of 1-2 bases from a large number of templates is obtained by a single reaction, where the bases are discriminated by a fluorescent dye. Each time, a fluorescent image of the entire field, i.e., all templates, is captured with a CCD camera so that all analyzed bases are recorded. After clearing out the dyes, the same cycle is continued until no further base information can be obtained.

A schematic representation of the representative sequence principle is shown in **Figure 1**. Each sequencer employs different principles of reaction including:

(1). Pyrosequencing [6] (FLX). When an extension reaction occurs, one dNTP is added, and pyrophosphate (PPi) is released. ATP sulfurylase quantitatively converts PPi to ATP in the presence of adenosine 5´ phosphosulfate. This ATP acts as fuel to the luciferase-mediated conversion of luciferin to

Table 1. Comparison of sequencers (January, 2009). It should be noted that the throughput of each sequencer is improving rapidly

	ABI 3730xl	Roche GS FLX	Illumina GA	ABI SOLiD
Bases / template	~1100	~400	~75	50
Templates / run	96	1,000,000	40,000,000	85,000,000
Data production /day	1 MB/day	400MB/run/7.5hr	3,000MB/run/6.5 days	4,000MB/run/6 days
Maximum samples	96	16 regions/plate	8 channels/flowcell	16chambers/2 slides
Sequence reaction	Sanger method	pyrosequencing	Reverse terminator	ligation sequencing

oxyluciferin that generates visible light in amounts that are proportional to the amount of ATP. The light produced in the luciferase-catalyzed reaction is detected by a camera and is analyzed in a program.

(2). Reversible terminator (GA). Using a fluorescent dye-labeled terminator, the single base extension reaction is performed. Then, the fluorescent dye and the blocking group are chemically removed, and the next extension reaction is performed. The terminators are similar to those reported in [7].

(3). Sequencing by ligation (SOLiD). This reaction utilizes the base discrimination ability of DNA ligase. Two bases adjacent to the ligation point are used for sequencing. One cycle consists of ligation of oligonucleotides, and cleavage and removal of the extended product. The cycles are repeated until no detectable fluorescent signals are obtained. One of the earliest examples of sequencing by ligation is described in [8].

The current benchmarks of the sequencers are summarized in **Table 1.** In brief, FLX produces long reads (~400 bases), but the number of templates per run is moderate (~1,000,000). GA and SOLiD produce short read (50~75 bases), but are characterized by the large number of templates per run (100,000,000~85,000,000). Their performance is increasing rapidly.

The third generation sequencer–single molecule sequencing

Pacific Bioscience Inc. is developing a sequencer based on a new principle, which should be categorized as third generation. This DNA sequencer uses single DNA molecules as templates. The main characteristic of this sequencer is real-time monitoring of nucleotide incorporation with DNA polymerase. The major drawback of the second-generation sequencers from the Sanger method is short read of templates. Not like the second-generation sequencer, this sequencer can obtain reads of several kilobases from a single template. This sequencer is based on the following three technical components.

1. Zero mode waveguide [9]. A slide glass is coated with a thin aluminum layer. The aluminum layer has many small holes, with a diameter d ~50 nm. Because the light, whose wave length is greater than 1.7 x d, is evanescent, the illuminating light exists only in the entrance of the hole. Because no propagation mode exists, these guides are referred to as "zero mode wave guide." To enable real-time monitoring of DNA polymerase, the concentration of substrates (deoxynucleotide triphosphate, dNTP) should be more than micromole. However, other technologies require a much lower concentration for detection of fluorescence. Zero mode waveguide is the first technique solving this problem.

2. Passivation of aluminum surface using polyphosphonate chemistry [10]. Aluminum surface is protected with polyvinylphosphonic chemistry from attachment of DNA polymerase. Thus, DNA polymerase molecules only attached to the silica surface, i.e., at the bottom of the holes, eliminating possible background fluorescent light.

3. Use of dNTPs whose terminal phosphate moieties are conjugated with fluorophores [11]. These fluorescence-labeled dNTPs release fluorescence when incorporated into DNA, and then lose the fluorophores. Thus,

these dNTPs enable real-time monitoring of incorporation of nucleotides.

Characteristics of sequence data generated by the second-generation sequencers

Because the third generation sequencer has yet to be commercialized, this review further focuses on the second-generation sequencers. It should be noted that there is plenty of room for improvement in throughput of GA and SOLID. With these systems, templates can be accumulated at a much higher density. On the contrary, Roche FLX has limitations. Because the fluorescent dye, i.e., oxyluciferin, diffuses into the reaction solution, each template bead must be separated in an individual well. This feature limits the template density

The surrounding situation of the second-generation sequencers is different from that of the first-generation sequencers. The most important factor is the completion of the human genome project. As shown above, a major drawback of the next-generation sequencers from the previous sequencers is the short read length: 350 bases (FLX) and 50-75 bases (GA, SOLiD), compared to > 800 bases with first-generation sequencers. The short read length is a considerable disadvantage for *de novo* sequencing. In *de novo* sequencing, it is necessary to construct a complete sequence from a large number of short sequence pieces. If the one read length is short, the short pieces make only small overlaps, making it difficult to construct contigs. Thus, the second-generation sequencers, especially GA and SOLiD, are not intended for *de novo* sequencing. However, in the human genome, the short pieces may be assembled into large sequences, being matched with the reference human genome sequence. In this way, the second-generation sequencers can produce complete genome sequences of individuals. The major genome centers now challenge two targets, i.e., the genomes of individuals and cancer genomes.

Sequencing genomes of individuals

For several years, single nucleotide polymerphism (SNP) and its application to human genetics has been the most intensive area in genomics. SNP was at first intensively collected using sequences obtained during the human genome project. These SNPs (roughly 100 million) were organized by haplotypes identified by the international HapMap project [12]. Consequently, about 50,000 tag SNPs representing haplotypes, were obtained. Genetic loci associated with a number of common diseases have been identified using the above tag SNP set through genome-wide association studies (GWAS). Accumulating results, however, show that GWAS generally failed to identify most of the genetic background of common diseases. A series of articles has been recently published to review the results from various viewpoints [13-15]. There are now a number of discussions to determine the research direction, i.e., continuation of GWAS or turning the research direction to complete sequencing of individual human genomes. Because the SNP markers used in GWAS are based on the international HapMap project, they detect allele variants whose frequencies are over 5 %. Therefore, rare variants (0.1 - 5 %) cannot be detected in GWAS. Proponents of the genome sequencing argue that genetic association may be found with rare variants, not detected by the current tag SNPs, and the complete genome sequences of a large number of individuals will uncover the more detailed view of variations. Currently, the "1,000 genomes" project (http://www.1000genomes.org), an international project to sequence genomes of 1,000 individuals, is ongoing. The outcome of the projects will be an important resource for human genome variation, but the direct objective is identification of rare variants to extend current GWAS.

It is important to confirm whether the second-generation sequencers can identify SNP equally as well as the Sanger method. Two Caucasian individual genomes have been determined before the "1,000 genomes" project. One that was obtained by the Sanger method [16], identified 2.8 million known SNPs and about 0.74 million novel SNPs. The other that was sequenced with GS20, a previous model of FLX [17], identified 2.72 million known and 0.61 million novel SNPs. Pilot experiments of the 1,000 genome project determined genomes of two individuals with GA [18, 19]. The sequence of a male Yoruba identified 3.8-4.1 million SNPs, 73.6% of which were in dbSNP [18]. The sequence of an Asian individual identified 3 million SNPs, 73.5 % of which were in dbSNP [19]. Recently, a new study compared the second-generation sequencers and a Sanger sequencer from the view point of GWAS [20]. In general, the

second-generation sequencers had very high sensitivity, i.e., identification of SNPs, but relatively low specificity. This tendency was more prominent with GA and SOLiD, because of short sequence reads: errors were more common in repeated sequence regions, probably due to errors during sequence assembly. The other obstacle is biases in representation among genomic regions. To obtain complete coverage of a genomic region, it is necessary to obtain more reads. These results suggest that the next-generation sequencers are useful for SNP studies, if enough reads are obtained.

Still the complete human genome sequencing is expensive. In addition, a huge computational load is required. Instead, sequencing of all protein coding regions, named "exome", is regarded as a cost-effective approach [21]. SNPs or mutations in coding regions are more informative and likely to be linked to diseases than those in non-coding regions. One of the examples is a study on pancreatic cancer described below [22].

Sequencing of cancer genomes

The objective of projects, such as The Cancer Genome Atlas (http://cancergenome.nih.gov), to sequence cancer genomes is a complete list of genomic changes contributing to carcinogenesis. These projects hypothesize that there would be undiscovered genes contributing to carcinogenesis, and they will accompany genomic changes such as mutations, copy number variations and translocations. Epigenetic events have also been known to contribute to carcinogenesis, and may be incorporated into the projects. Unbiased exploration of such events would substantially contribute to understanding of cancer, and lead to identification of new target molecules.

Several pilot experiments using the first generation sequencers have been performed. Due to limited throughput of the first generation sequencers, several early studies focused on specific gene families, such as tyrosine kinases, which were often activated by somatic mutations. An organized study was performed at the Welcome Trust Sanger Institute [23]. In that study, somatic mutations were classified into "driver" and "passenger" mutations. "Driver" mutations are defined as that conferring growth advantage, and "passenger" mutations are defined as those without any biological effects. Overall selection pressure by all the substitute mutations was calculated: 1.29 (95% confidence interval, 1.10-1.51; P=0.0013).197The other study examined the majority of the transcribed genes (18,191 genes) with eleven breast and eleven colorectal cancer tissues [24]. This study revealed that there were a large number of mutations with rare incidence, in addition to a small number of genes with mutations of high incidence. Both studies suggested that known somatic mutations were only a small fraction of mutations in cancer genomes, and more systematic analysis of the cancer genome, i.e., complete genome sequencing of a large number of cancer tissues, is necessary. These studies were followed by two studies on glioma [25, 26]. Both studies accompanied measurements of copy number variation by genome arrays and gene expression profiling [25] by microarrays or SAGE [26]. One of the studies found recurrent mutations at the active site of isocitrate dehydrogenase 1 (*IDH1*) in 12% of glioblastoma patients [26]. This result suggests that there would be additional important mutations not discovered so far.

Comparison of a cancer genome with the corresponding germline genome is very informative. One study analyzed the whole genome of malignant cells and normal cells from a single acute myelogenous leukemia (AML) patient [27]. The whole genome analysis revealed that the AML genome had only eight heterozygous, non-synonymous somatic mutations, all of which were novel. Another study to sequence all coding regions on a genome of familial pancreatic cancer identified that mutations in *PALB2* was responsible for the disease, validated with 96 additional samples [22]. Both studies could pinpoint out a small number of candidate genes, demonstrating the accuracy and thoroughness of the whole-genome approach.

The above early studies strongly suggest that the large-scale cancer genome projects would definitely contribute to our understanding of genetic changes in cancer. However, contribution to medicine is a different problem. The rationale to justify the large investments for these projects is identification of molecular targets and subsequent developments of anti-cancer drugs. The proponents of the projects argue that newly identified mutations will be

effective targets for anti-cancer drug development. This reflects the current trend of anti-cancer drug development: a large number of molecular target drugs are now being developed or during clinical trials with expectations to improve cancer therapy. However, when the above cancer genome projects were finished, the current trend and enthusiasm might be finished. Already, there is controversy among scientists on the future prediction of molecular target drugs [28, 29]. So far, all molecular target drugs except imatinib extend overall survival only several months. Molecular target therapy might turn out to be not attractive as it is: pharmaceutical companies might lose interest. In any case, the resulting data will be valuable as a resource for cancer research.

Discovery of new infectious agents

The third important application of the second-generation sequencers is identification of infectious agents. RNA or DNA of human tissues or cells infected by a specific infectious agent such as a virus, bacterium, contain the human genome sequences as well as sequences of the infectious agent. Sequencing a large number of RNA or DNA pieces from an infected sample, the resulting sequences contain those derived from the infectious agent as well as from the human genome. Now that the complete human genome sequence has been obtained, subtraction of the human genome sequence should theoretically yield sequences of the infectious agent. This idea is not new. In 2002, a computational experiment was performed, by searching the human genome sequences for expressed tag sequences (EST) of human origin using data in the public database [30]. Among sequences not matching the human genome, more than 50 sequences matching virus genomes were identified. The same group performed a model experiment with tissues of post-transplant lymphoproliferative disorder (PTLD), and successfully recovered Epstein-Bar virus sequences, the known agent of PTLD [31]. These studies suggested the plausibility of the above experimental strategy.

In spite of the potential strength of the strategy, the high cost of DNA sequencing has prevented real application. Due to the decreased cost of sequencing by the second-generation sequencer, two studies using FLX appeared in 2008. One study focused on patients who died of febrile illness after visceral organ transplantation [32]. Unbiased transcript sequencing from liver and kidney, and subsequent data analysis revealed infection of a new arena virus. The other study focused on Merkel cell carcinoma, a rare type of skin cancer [33]. Sequencing of nearly 400,000 transcripts identified sequences similar to known polyoma viruses, Further analysis revealed a new polyoma virus sequence named Merkel cell polyoma virus.

Application to gene expression profiling

The sequencers can be applied to gene expression profiling, i.e., a genome-scale analysis of gene expression. Sequencing a large number of transcripts purified from a tissue or cell, and subsequently matching them to the human reference genome reveals the identity of each transcript. The expression level of the gene can be determined from the number of times each gene sequence appeared. This approach of gene expression profiling has been named digital gene expression profiling, and was originally initiated in the early stage of the human genome project [34]. Later, a new technique named serial analysis of gene expression (SAGE) [35], appeared. In SAGE, a small tag (SAGE tag), with a size of 9 to 21 bases, is obtained from each transcript, and tens of tags are concatemerized, and read with a sequencer. With SAGE, from a single read, frequency information of tens of transcripts can be obtained. Even still with SAGE, it was not practical to process a large number of samples due to low throughput of the sequencers based on the Sanger method. With the next-generation sequencers, digital expression profiling has finally become a plausible method comparable to microarrays. Its major advantage over microarray is straightforward standardization of the data. In digital expression profiling, data is just molecular counts. In contrast, the data obtained by microarray analysis is expression level against some standard, and it is difficult to compare data from different experimental series. However, for laboratory use, i.e., comparison of global gene expression among samples of interest, digital expression profiling does not have clear advantage over microarrays.

Discussion for future applications