M-step:

$$\hat{p}(g_j \mid z_h) \propto \pi_{cg} \sum_{i'} \frac{N_{i,j}}{N_{CG}} p(z_h \mid c_i, g_j) + \pi_{CC} \sum_{i'} \frac{L_{i,i'}}{N_{CC}} p(z_h \mid c_i, c_{i'})$$

$$\hat{p}(g_j \mid z_h) \propto \pi_{CG} \sum_{i} \frac{N_{i,j}}{N_{CG}} p(z_h \mid c_i, g_j) + \pi_{GG} \sum_{j'} \frac{M_{j,j'}}{N_{GG}} p(z_h \mid g_j, g_{j'})$$

$$\hat{p}(z_c) \propto \pi_{CG} \sum_{i,j} \frac{N_{i,j}}{N_{CG}} p(z_h \mid c_i, g_j) + \pi_{GG} \sum_{j',j''} \frac{M_{j',j''}}{N_{GG}} p(z_h \mid g_{j'}, g_{j''}) + \pi_{CC} \sum_{i',i''} \frac{L_{i',i''}}{N_{CC}} p(z_h \mid c_{i'}, c_{i''})$$

Parameter Settings in Our Experiments

We set the number of latent clusters, H, at 128 and used a uniform distribution for the weights (ie π) of both 2MAM and 3MAM in all cases. We iterated the EM algorithm until the improvement of the observed log-likelihoods between two successive iterations is less than 0.001.

Data

Cancer-Gene and Cancer-Cancer Co-occurrences

OMIM (Online Mendelian in Man) is a human-curated database, containing the comprehensive and authoritative information on human genes and genetic disorders. Our focus is placed on genes which are related with cancers, and we used a software tool CGMIM, which extracts the description section of OMIM records to obtain cancers and associated genes. The CGMIM builds a synonym list from International Classification of Disease for Oncology (ICD-O) [14]. The list maps genetic disorders into 21 different types of cancers, which are defined by the National Cancer Institute of Canada. They are bladder, brain, breast, cervix, colorectal, esophagus, kidney, larynx, leukemia, lung, lymphoma, melanoma, myeloma, oral, ovary, pancreas, prostate, stomach, testis, thyroid and body-of-uterus. We obtained the two types of co-occurrence datasets from the OMIM database downloaded in Oct 2005. Our datasets are altogether 2,017 genes associated to cancers, 3,743 cancer-gene pairs and 206 cancer-cancer pairs.

Gene-Gene Co-occurrences

Since gene-gene co-occurrences are not available in OMIM, we obtained this kind of co-occurrences from the Medline database. We used Locuslink [34], ie a human curated database, to avoid errors that may occur in identifying gene names in Medline. The Locuslink has a list of links, each of which connects a Locus ID with a PubMed ID, meaning that we can see whether a gene (specified by a Locus ID) is in an abstract (specified by a PubMed ID) or not.

We used a file available at the following ftp site, and the file we used was generated at Dec 2004:

ftp://ftp.ncbi.nih.gov/refseq/LocusLink

From this list, we selected Medline records containing one or more human genes, focusing on "human" genes only. We then generated gene-gene co-occurrences from the selected Medline records. That is, if two genes are in a same Medline record, we can say that these two genes co-occur.

We found some Medline records have a large number of genes. For example, a record with PubMed ID 12477932 contains more than 9,000 human genes by showing all genes in a microarray experiment. Thus, we removed the record, each of which has more than 10 genes. We note that this is a normal procedure in dealing with Medline records. For example, Wilkinson et al also put this kind of restriction to filtering Medline records for finding communities of related genes [46].

Our focus is on cancer associated genes, and a gene-gene co-occurrence pair was removed unless both genes of the pair are in the 2,017 genes of our cancer-gene co-occurrence dataset. Finally we obtained 3,118 gene-gene pairs from Medline. Table 1 shows a summary of the data information.

Table 1: The size of co-occurrence datasets.

Item	Size
gene type	2,017
gene-gene	3,118
cancer type	21
cancer-cancer	206
cancer-gene	3,743

Preliminary Verification on Gene-Gene Co-occurrence Dataset

Focusing on genes in cancer-gene co-occurrence pairs from OMIM, we attempted to confirm that two genes in each gene-gene pair from Medline are associated to a same cancer with high probability. When both two genes in a gene-gene pair are associated with at least one same cancer, we call such a gene-gene pair a *positive pair,* and we computed the ratio of positive pairs to all gene-gene pairs, which we call the *positive ratio.*

We found that among total 3,118 gene-gene co-occurrence pairs, 1,804 (57.86%) are positive pairs. We then reduced the size of gene-gene pairs by the number of co-occurrences and checked the positive ratio. Table 2 summarizes the obtained results.

As shown in the table, with increasing the co-occurrence number of gene-gene pairs, the positive ratio increased. For example, when the number of co-occurrences is set at more than one, 490 (64.64%) out of 758 gene-gene pairs are positive pairs. Furthermore, as a baseline, we checked the positive ratio of randomly generated pairs. That is, we randomly generated 3,118 gene-gene pairs 1,000 times using our 2,017 cancer associated genes and checked the average positive ratio for them. The average positive ratio was only 26.65%, with minimum 24.05%, maximum 29.76% and standard deviation 0.0083, which is far less than those obtained by our gene-gene co-occurrence dataset. These results clearly indicate that the motivation of adding gene-gene co-occurrence data in Medline to the cancer-gene and cancer-cancer data from OMIM would be reasonable.

Experimental Results

Predictive Performance of Mixture Aspect Model

Evaluation Procedure

We evaluated the performance of MAM by cross-validation on predicting associated cancer-gene pairs. We examined four types of MAM (including AM). That is, we first built AM using only the cancer-gene co-occurrence dataset. We then tested two different 2MAM by adding cancer-cancer or gene-gene pairs to the cancer-gene pairs, which correspond to 2MAM (CG+CC) or 2MAM (CG+GG), respectively. Finally 3MAM was examined by using all these three types of co-occurrence datasets.

To examine the effect of the training data size on the performance of our models, we checked three different data-size ratios of training to test datasets, 3:1, 1:1 and 1:3, in our cross-validation experiment. For example, in the 1:1 case, we randomly divided the original cancer-gene dataset into two subsets of roughly equal size, and then alternately selected one subset as a test set and the other as a training set. We carried out 50 rounds of the cross-validation to reduce the possible biases caused by random partitioning. In each round, to compare the performance of different models, we kept the testing dataset unchanged while adding another type of co-occurrence dataset. In this way, we made predictions on the same test dataset. We note that AM cannot compute the likelihood for a cancer gene pair in the test dataset unless a gene of this pair appears in the training data. So we removed all the pairs which are not in the training data but in the test dataset. We then used all remaining pairs as positive test examples. Please note that this experimental setting is advantageous to AM and not to MAM. Negative examples, which were used for evaluation only, were randomly generated to be included in neither the training dataset nor the positive test dataset. The size of negative test dataset was set as the same as that of positive test dataset.

Evaluation Measures

1) Area Under the ROC Curve (AUC)

The performance of each probabilistic model is evaluated by the ability to discriminate positive examples from negative examples in test data of our cross-validation. We used AUC (Area Under the ROC curve) to evaluate the discriminative performance of a model. The AUC is computed from an ROC (Receiver Operator Characteristic) curve. The ROC curve is drawn by plotting "sensitivity" against "false positive rate", using the ranked cancer-gene pairs. The sensitivity

Table 2: The ratio of positive pairs in gene-gene co-occurrence dataset.

# co-occurrences	- (random)	> = 1	>1	>2	>3	>4	>5	>6
Dataset size	3,118	3,118	758	379	276	152	122	99
Positive ratio (%)	26.65	57.86	64.64	68.34	69.91	70.2	72.13	76.77

(or true positive rate) is the proportion of the number of correctly predicted positive examples to the total number of positive examples. The false positive rate is the proportion of the number of false positive examples to the total number of negative examples. More concretely, once we estimated the parameters of a probabilistic model from training data, we computed the likelihood of each cancer-gene pair in test data and ranked them according to their likelihoods. We then set a cut-off value to separate positive examples from negatives and computed the sensitivity and the false positive rate by changing the cut-off value from the highest likelihood to the lowest. We finally plotted all obtained values of the sensitivity and the false positive rate to draw an ROC curve.

The AUC, a popular metric for measuring the performance of different models [5, 18], can be computed as the area under this ROC curve. We can see that the larger the AUC, the better the performance of the model. We further used the paired sample two-tailed *t*-test to statistically evaluate the performance difference between 3MAM and another model. Since we run crossvalidation 50 times, we have at least 100 values in each of the three different ratios, and so if the *t*-value is greater than 3.50 (2.36) then the difference is more than 99.9% (98%) statistically significant.

2) Log-likelihood Distribution on Positive Test
All these four probabilistic models are trained in an unsupervised manner and the maximum likelihood setting, meaning that they are trained to provide the maximum likelihoods to given training data. In addition, conveniently enough, they have the same (common) set[1] of parameters, ie $p(c_i|z_h)$, $p(g_j|z_h)$ and $p(z_h)$. Thus, we can compare the four models each other by the distribution of the likelihoods for positive test examples, given by each of the models. If a model provides positive examples with higher likelihoods than those of another, we can say that this model is better than the other.

Results

1) AUC
Table 3 shows the AUC for each of the four models at different data settings and the *t*-value (in parenthesis) between the AUC of 3MAM and that of another model.

Table 3: AUCs and *t*-values (in parenthesis) obtained by 50 rounds of cross-validation on cancer-gene pairs.

Model	Ratio of training to test data		
	3:1	1:1	1:3
3MAM (CG+CC+GG)	**76.1**	**74.6**	**73.2**
2MAM (CG+CC)	75.8 **(2.56)**	74.2 **(2.44)**	71.8 **(12.9)**
2MAM (CG+GG)	73.9 **(17.2)**	71.4 **(22.5)**	68.3 **(38.0)**
AM (CG)	74.1 **(14.7)**	70.5 **(26.3)**	64.9 **(55.1)**

This table clearly shows that 3MAM outperformed the other three models, and the second best model is 2MAM (CG+CC). We can easily see that, compared with AM, the 3MAM improved around 2 to 9% in the discriminative accuracy. Furthermore, the *t*-values showed that 3MAM outperformed all other models by a statistically significant factor in all cases. These results indicate that incorporating cancer-cancer and gene-gene pairs from diverse sources improved the predictive performance obtained by cancer-gene pairs only.

In addition, we note the following two points on these results: First, interestingly, 2MAM (CG+GG) outperformed AM in 1:1 and especially 1:3 cases, but not 3:1 case. This is probably because gene-gene co-occurrence data comes from the different source, Medline, which can supplement original data, when it is scarce, and can achieve better performance. Second, since we have only 21 type of cancers and 2,017 genes, some putative negative test examples must be positive. This means that the performance of our model may be under-estimated.

2) Log-likelihood Distribution on Positive Test
When the probability parameter has a uniform distribution, a randomly generated cancer-gene pair has the following log-likelihood:

$$\log\left(\frac{1}{21}\times\frac{1}{2,017}\right) = -4.63$$

In our unsupervised setting, the log-likelihood of a positive example should be larger than the above value. In other words, when positive (test) examples are given, a better trained probabilistic model would provide a larger number of examples whose log-likelihoods are larger than the above value.

[1] We note that trained models have different parameter values because the training algorithms are different.

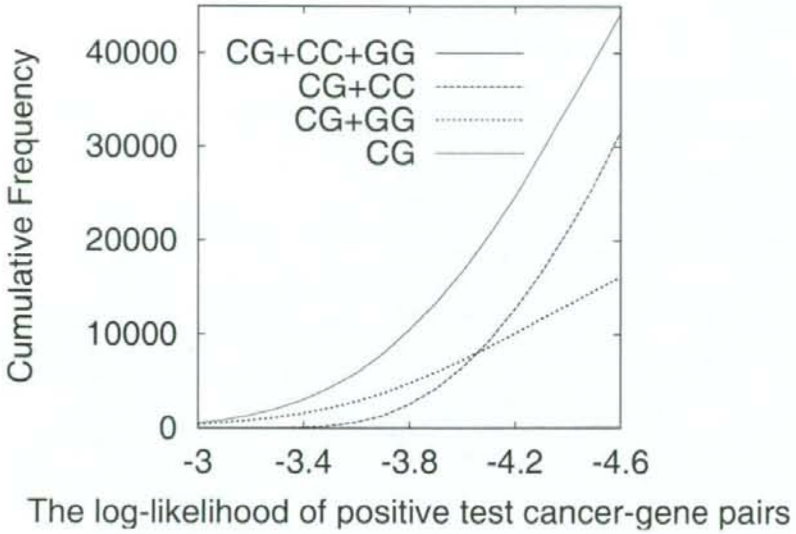

Figure 1: Cumulative number of positive examples with higher log-likelihoods.

Thus, given a cut-off value, we checked the number of positive test examples having log-likelihoods larger than the given cut-off value. Figure 1 shows the counted cumulative number of positive test pairs with higher likelihoods against a given cut-off value. This figure is drawn from the average over the 50 rounds of our cross-validation at the 3:1 ratio of training to test data. We found that 3MAM is clearly the best among the four models, always keeping the largest number of examples whose likelihoods higher than a given cut-off value. These results also confirmed the performance advantage of 3MAM over other models and showed adding cancer-cancer and cancer-gene datasets is effective. Another empirical finding in this analysis is that 2MAM (CG+GG) outperformed 2MAM (CG+CC) in the range of larger than – 4, while 2MAM (CG+CC) outperformed 2MAM (CG+GG) in the range between – 4.6 and – 4.

Mining and Analyzing Unknown Cancer Associated Genes

Mining New Cancer-Gene Co-occurrences

We trained 3MAM using all three types of co-occurrence data and tried to find new associated cancer gene pairs which are unknown in the current literature. The procedure is as follows: We first trained 3MAM using all the three types of co-occurrence data and then computed the log-likelihoods of all cancer-gene paris that are not in the current cancer-gene co-occurrence data. We repeated this procedure 100 times and ranked the new pairs according to the average log-likelihoods over 100 times. Table 4 shows the list of top 20 pairs with their log-likelihoods, and a more

Table 4: 20 Cancer-gene pairs with highest log-likelihoods that are not in our training dataset.

Cancer Type	Gene Name	Log-likelihood
OVARY	TP53	– 3.078
COLORECTAL	BCL2	– 3.085
STOMACH	TP53	– 3.113
LEUKEMIA	CDKN1A	– 3.176
LYMPHOMA	BAX	– 3.191
PANCREAS	TP53	– 3.199
BREAST	NFKB1	– 3.222
THYROID	TP53	– 3.234
LYMPHOMA	TNF	– 3.235
LUNG	BCL2	– 3.244
BREAST	BCL2	– 3.266
KIDNEY	TP53	– 3.269
BREAST	TNF	– 3.293
LEUKEMIA	TNF	– 3.300
COLORECTAL	TNF	– 3.312
LYMPHOMA NF	NFKB1	– 3.316
LUNG	TNF	– 3.323
COLORECTAL	CASP8	– 3.330
LEUKEMIA	NFKB1	– 3.336
BRAIN	BCL2	– 3.340

detailed list of top 1,000 pairs is given in Table 1 of the on-line supplementary information. The first, second, third and fourh columns of the on-line information show cancer names, HUGO IDs [43], genes and log-likelihoods, respectively.

As shown in Table 4, the top 20 list has some famous oncogenes such as TP53, BCL2 and TNF. This result implies that our prediction worked well, because these popular genes must be related with a lot of different types of cancers. So we can expect that these relations must exist, even if the cancer-gene co-occurrences in Table 4 are not in OMIM. In other words, we may say that these relations are easily expected. Thus in the next section, we focused on genes which are specific to some cancer but unknown and tried to analyze how the found genes are related with the corresponding cancer.

Mining New Genes Specific to Cancer

We computed the following score for all cancer-gene pairs by using the probability parameters of 3MAM, which was trained by using all three types of training data.

$$R(g_j,c_i)=\frac{p(g_j \mid c_i)}{\Sigma_i p(g_j \mid c_i)}$$

where

$$p(g_j \mid c_i)=\frac{\Sigma_h p(c_i \mid z_h)p(g_j \mid z_h)p(z_h)}{\Sigma_{j',h'} p(c_i \mid z_{h'})p(g_{j'} \mid z_{h'})p(z_{h'})}.$$

The $p(g_j|c_i)$ is the conditional probability that given a cancer type c_i, g_j is related with the c_i. Thus the score $R(g_j, c_i)$ is the ratio that a gene g_j is related with c_i, comparing to all the other cancer types. That is, it is the probability over cancer types and shows to what extent gene g_j is specific to cancer c_i. Once we computed the score for each pair, we sorted the values for each cancer and selected the top 20 genes which are not in the cancer-gene pairs in the training data. Table 2 of the on-line supplementary information shows the list of top 20 genes of each cancer. The first, second, third and fourh columns of this file show cancer names, HUGO IDs, genes and parameter values, respectively.

These pairs are unknown pairs in OMIM and Medline, but our method suggested that each of them has a strong relationship between a cancer and a gene. In fact, we can see a biological relationship for each pair from the literature. Below we briefly describe the biological, medical and genetic relationships on each pair of the list, for only the top gene of seven cancers out of all 21 cancers, owing to the space limitations.

Brain:

The top is MMP17. According to Puente et al [36], they revealed that MMP17 is expressed mainly in the brain, leukocytes, colon, ovary and testis, using northern blot analysis of polyadenylated RNAs isolated from a variety of human tissues. This implies MMP17 can be related with brain cancer.

Breast:

The top is ZAP70, a member of the Syk tyrosine kinase family. Recently, Gatalica and Bing [15] pointed out that the loss of Syk tyrosine kinase expression characterises a subset of breast carcinomas. This implies a relationship between ZAP70 and breast cancer.

Colorectal:

The top is CYP1A1. Hou et al [21] recently reported the relationship between the CYP1A1 polymorphism and the risk for colorectal adenoma. Their summary is that the joint carriage of CYP1A1 and NQO1 polymorphisms, particularly in smokers, was related to colorectal adenoma risk, with a propensity for formation of multiple lesions. This would be an evidence for the relationship between CYP1A1 and colorectal cancer. The second is MAD2. The expression profile of MAD2 in colorectal cancer was investigated by Li et al [26]. Their result shows that the defect of spindle checkpoint gene MAD2 is involved mainly in colorectal carcinogenesis. So this clearly indicates the relationship between MAD2 and colorectal cancer.

Lymphoma:

The top is LMO1. In the recent study of leukemogenesis, Lin et al [27] found that almost 60% of transgenic mice that overexpressed both OLIG2 and LMO1 developed pre-T LBL with large thymic tumor masses. This reveals the association between LMO1 and lymphoma cancer.

Pancreas:

The top is NR5A2. NR5A2, a member of a nuclear receptor subfamily, is a liver recepter homolog1 (LRH-1). Fayard et al [12] showed that LRH-1 is abundantly expressed in pancreas. Furthermore, their in situ hybridization and gene expression studies demonstrated that both LRH and carboxyl ester lipase (CEL) are co-expressed and confined to the exocrine pancreas.

Prostate:

The top is KLK10, ie kallikrein 10. Bharaj et al [3] showed the association between single nucleotide polymorphisms in the human KLK10 and prostate cancer. Petraki et al [31] studied the localization of human KLK10 in benigh and malignant prostatic tissues and the correlation between the expression of KLK10 and prostate cancer (PC) prognosis. They pointed out that kallikreins may function as tumor suppressors or are down-regulated during cancer progression. These results imply the relationship between KLK10 and prostate cancer.

Testis:

GAGEB1 is the top. Chen et al [9] isolated GAGEB1 by differential display PCR. They found that GAGEB1 expression was restricted to testes and placenta on human multiple tissue Northern blots. This shows some relationship GAGEB1 and testis cancer.

Concluding Remarks

We have applied a new probabilistic model MAM, which was proposed by us in our research on mining implicit chemical compound-gene relationship, to the problem of finding new cancer associated genes from OMIM and Medline. MAM can integrate different types of co-occurrence datasets effectively, and we found that MAM performed very well even when co-occurrence datasets are gathered from heterogeneous sources.

In this work, we used a uniform distribution for the component weights (π) of our mixture model to allow users additional control. Interesting future work would adjust the weights to achieve the maximum predictive performance. On the other hand, the gene-gene co-occurrence data can come from a different source other than Medline. Since microarray expression data can reveal the biological relationship of genes, it would be very interesting to integrate gene-gene co-occurrence data from microarray expressions.

Acknowledgement

This work is supported in part by Bioinformatics Education Program "Education and Research Organization for Genome Information Science" and Kyoto University 21st Century COE Program "Knowledge Information Infrastructure for Genome Science" with support from MEXT (Ministry of Education, Culture, Sports, Science and Technology), Japan.

References

[1] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol*. 215(3): 403–410.

[2] Bajdik CD, Kuo B, Rusaw S, Jones S and Brooks-Wilson A. (2005). CGMIM: Automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and Candidate genes, *BMC Bioinformatics*, 6:78–84.

[3] Bharaj BB, Luo LY, Jung K, Stephan C, Diamandis EP (2002) Identification of single nucleotide polymorphisms in the human kallikrein 10 (KLK10) gene and their association with prostate, breast, testicular, and ovarian cancers. *Prostate*, 51(1):35–41.

[4] Boguski MS, Lowe TM, Tolstoshev CM. (1993) dbEST–database for "expressed sequence tags" *Nat Genet.* 4(4):332–3.

[5] Bradley A. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, 30:1145–1159.

[6] Brancolini V and Devoto M. (1996) Genetic linkage studies for the identification of cancer-related genes. *Ann Ist Super Sanita*. 32(1):173–180.

[7] Cardon LR and Bell JI. (2001) Association study designs for complex diseases. *Nat Rev Genet.* 2(2):91–99.

[8] Chang JT and Altman RB. (2004) Extracting and characterizing gene-drug relationships from the literature, *Pharmacogenetics*, 14:577–586.

[9] Chen ME, Lin SH, Chung LW, Sikes RA. (1998) Isolation and characterization of PAGE-1 and GAGE-7. New genes expressed in the LNCaP prostate cancer progression model that share homology with melanoma-associated antigens. *J. Biol. Chem.*, 273(28):17618–17625.

[10] Dempster A, Laird N and Rubin D. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39:1-38.

[11] Forozan F, Karhu R, Kononen J, Kallioniemi A and Kallioniemi OP.(1997) Genome screening by comparative genomic hybridization. *Trends Genet.* 13(10):405–409.

[12] Fayard E, Schoonjans K, Annicotte JS and Auwerx J. (2003) Liver receptor homolog 1 controls the expression of carboxyl ester lipase. *J. Biol. Chem.* 278(37):35725–35731.

[13] Freudenberg J and Propping P. (2002), A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18, Suppl. 2:S110–S115.

[14] Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin DM and Whelan S. *International Classification of Diseases for Oncology* Third edition. World Health Organization; 2000.

[15] Gatalica Z and Bing Z., Syk tyrosine kinase expression during multistep mammary carcinogenesis. *Croat Med J.*, 46(3):372–376.

[16] Guo QM, DNA microarray and cancer. (2003) *Curr Opin Oncol*, 15:36–43.

[17] Hamosh A, Scott AF, Amberger JS, Bocchini CA and McKusick VA. (2005) Online Meddelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33:D514–D517.

[18] Hand DJ and Till RJ. (2001) A simple generalization of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45:171–186.

[19] Hofmann T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196.

[20] Hofmann T. (2004) Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22: 89–115.

[21] Hou L, Chatterjee N, Huang WY, Baccarelli A, Yadavalli S, Yeager M, Bresalier RS, Chanock SJ, Caporaso NE, Ji BT,Weissfeld JL and Hayes RB. (2005) CYP1A1 Val462 and NQO1 Ser187 polymorphisms, cigarette use, and risk for colorectal adenoma. Carcinogenesis, 26(6):1122–1128.

[22] Jenssen T, Laegreid A, Komorowski J and Hovig E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28:21–28.

[23] Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F and Pinkel D. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258(5083):818–821.

[24] Kim JM, Sohn HY, Yoon SY, Oh JH, Yang JO, Kim JH, Song KS, Rho SM, Yoo HS, Kim YS, Kim JG and Kim NS. (2005) Identification of gastric cancer-related genes using a cDNA microarray containing novel expressed sequence tags expressed in gastric cancer cells. *Clinical Cancer Research* 11:473–482.

[25] Kinzler KW and Vogelstein B, (2002) *The genetic basis of human cancer* edn 2, Toronto, McGraw-Hill.

[26] Li GQ, Li H and Zhang HF (2003) Mad2 and p53 expression profiles in colorectal cancer and its clinical significance. *World J Gastroenterol.*, 9(9):1972–1975.

[27] Lin YW, Deveney R, Barbara M, Iscove NN, Nimer SD, Slape C and Aplan PD (2005) OLIG2 (BHLHB1), a bHLH transcription factor, contributes to leukemogenesis in concert with LMO1. *Cancer Research,* 65(16):7151–7158.

[28] Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, West JA, Rostan S, Nguyen KC, Powers S, Ye KQ, Olshen A, Venkatraman E, Norton L and Wigler M. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.* 13(10):2291–2305.

[29] McKusick VA (1998) Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders, 12th edn. Johns Hopkins University Press, Baltimore, MD.

[30] Perez-Iratxeta C, Bork P and Andrade MA (2002). Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* 31:316–319.

[31] Petraki CD, Gregorakis AK, Papanastasiou PA, Karavana VN, Luo LY and Diamandis EP. (2003) Immunohistochemical localization of human kallikreins 6, 10 and 13 in benign and malignant prostatic tissues. *Prostate Cancer Prostatic Dis.* 6(3):223–227.

[32] Pinkel D, Segraves R, Sudar D, et al. (1998) High resolution analysis of DNA copy-number variation using comparative genomic hybridization to microarray. *Nat. Genet,* 20:207–211.

[33] Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D and Brown PO. Genome-wide analysis of DNA copy-number changes using cDNA microarrays.*Nat Genet.* 23(1):41–46.

[34] Pruitt K and Maglott D. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.,* 29:137–140.

[35] Pearson WR and Lipman DJ.(1988) Improved tools for biological sequence comparison. *PNAS,* 85(8):2444–2448.

[36] Puente XS, Pendas AM, Llano E, Velasco G and Lopez-Otin C. (1996) Molecular cloning of a novel membrane-type matrix metalloproteinase from a human breast carcinoma. *Cancer Research,* 56(5): 944–949.

[37] Qiu P, Wang L, Kostich M, Ding W, Simon JS and Greene JR.(2004) Genome wide in silico SNP-tumor association analysis. *BMC Cancer.* 4:4.

[38] Roylance R,(2002) Methods of molecular analysis: assessing losses and gains in tumors. *Mol Pathol* 55:25–28

[39] Shen D, He J and Chang HR. In silico identification of breast cancer genes by combined multiple hight throughput analyses. *Int J Mol Med,* 15(2):205–212.

[40] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM and Sirotkin K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29(1):308–311.

[41] Thorisson GA, Smith AV, Krishnan L and Stein LD. (2005) The International HapMap Project web site. *Genome Research,* 15: 1592–1593.

[42] Velculescu VE, Zhang L, Vogelstein B and Kinzler KW. (1995) Serial analysis of gene expression. *Science,* 270:484–487.

[43] Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW and Povey S. (2002) Guidelines for human gene nomenclature. *Genomics.* 79(4):464–470.

[44] Wang TL, Maierhofer C, Speicher MR, Lengauer C, Vogelstein B, Kinzler KW and Velculescu VE. (2002). Digital karyotyping. *PNAS.* 99(25):16156–16161.

[45] Wheeler D, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S and Helmberg W et al (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 33: D39–D45.

[46] Wilkinson DM and Huberman BA. (2004). A method for finding communities of related genes. *PNAS:* 101, 5241–5248.

[47] Yandell MD and Majoros WH. (2002) Genomics and natural language processing. *Nat. Rev. Genet.,* 3: 601–610.

[48] Zhu S, Okuno Y, Tsujimoto G, and Mamitsuka H. (2005). A probabilistic model for mining implicit "Chemical compound-gene" relations from literature. *Proc. of ECCB2005 (Bioinformatics 21 Supplement 2):* ii245–ii251.

Available online at www.sciencedirect.com

ScienceDirect

Toxicology Letters 167 (2006) 106–113

Toxicology Letters

www.elsevier.com/locate/toxlet

Altered gene expression of transcriptional regulatory factors in tumor marker-positive cells during chemically induced hepatocarcinogenesis

Shigehiro Osada[a,b,*,1], Ayako Naganawa[a,1], Masashi Misonou[a], Soken Tsuchiya[c,d], Shigero Tamba[c], Yasushi Okuno[d,e], Jun-ichi Nishikawa[a], Kimihiko Satoh[f], Masayoshi Imagawa[b], Gozoh Tsujimoto[e], Yukihiko Sugimoto[c], Tsutomu Nishihara[a]

[a] *Laboratory of Environmental Biochemistry, Graduate School of Pharmaceutical Sciences, Osaka University, 1-6 Yamada-Oka, Suita, Osaka 565-0871, Japan*

[b] *Department of Molecular Biology, Graduate School of Pharmaceutical Sciences, Nagoya City University, 3-1 Tanabe-dori, Mizuho-ku, Nagoya, Aichi 467-8603, Japan*

[c] *Department of Physiological Chemistry, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan*

[d] *Department of Pharmacoinformatics, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan*

[e] *Department of Genomic Drug Discovery Science, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan*

[f] *Department of Organic Function, Hirosaki University, School of Health Science, Hon-Cho 66-1, Hirosaki 036-8564, Japan*

Received 10 August 2006; received in revised form 29 August 2006; accepted 29 August 2006
Available online 25 September 2006

Abstract

Glutathione-*S*-transferase placental form (GST-P) is markedly and specifically inducible in rat chemical hepatocarcinogenesis and is a reliable marker protein for pre-neoplasia. To gain insights into the molecular mechanisms at the early stage of hepatocarcinogenesis and hepatotoxicity, we investigated the gene expression profile by DNA microarray analysis. We prepared RNA from GST-P-positive foci in three individual rats and compared with normal liver sections from three individual rats, and labeled RNA was individually hybridized onto Affymetrix GeneChip Rat Expression Array 230A. DNA microarray analysis showed distinctly different profiles of dysregulated gene expression and supported the previous finding that some enzymes involved in metabolism and detoxification are overexpressed and suppressed. Here we discovered that several DNA-binding transcription factors and cofactors, including sterol-regulatory-element binding protein 1 (SREBP1) and Wilms' tumour 1 (WT1)-interacting protein, and their target genes were dysregulated in GST-P-positive foci. Moreover, genes involved in chromatin components, histone modification enzymes, and centrosome duplication were highly expressed. These genes were not previously known to be up-regulated during chemically induced hepatocarcinogenesis. DNA microarray analysis using RNA prepared from tumor marker-positive foci and control tissues provided a candidate gene link to the early stage of carcinogenesis and hepatotoxicity.

Keywords: Chemical hepatocarcinogenesis; Tumor marker; Gene expression; Transcription factor; Chromatin; Histone

[*] Corresponding author at: Department of Molecular Biology, Graduate School of Pharmaceutical Sciences, Nagoya City University, 3-1 Tanabe-dori, Mizuho-ku, Nagoya, Aichi 467-8603, Japan. Tel.: +81 52 836 3456; fax: +81 52 836 3456.
E-mail address: osada@phar.nagoya-cu.ac.jp (S. Osada).

[1] These authors contributed equally to this work.

0378-4274/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.toxlet.2006.08.014

1. Introduction

Rat glutathione-S-transferase placental form (GST-P) is a phase II detoxification enzyme and its expression is completely repressed in normal liver. GST-P is also a well-known tumor marker that is specifically induced during chemical hepatocarcinogenesis in rats (Sato, 1989; Satoh et al., 1985). GST-P expressed single cells are detected in the liver after treatment of diethylnitrosamine (DEN) and might be precursors of preneoplastic foci and nodules (Satoh et al., 1989; Satoh et al., 2005). Hepatocyte nodules in six models of liver carcinogenesis were analyzed and the amount of GST-P was elevated in all types of nodules (Eriksson et al., 1983). Measurement of GST-P-positive foci is rapid detection of carcinogenic agents in the medium-term rat liver bioassay (Ito test), which is considered to be a reliable tool for prediction of promoting or reducing activity of chemicals on hepatocarcinogenesis. Over 300 chemicals have already observed and this test was recommended as an alternative to long-term carcinogenicity testing at the International Conference on Harmonization (ICH) (Ito et al., 2003). GST-P-positive foci are induced by not only DEN but many chemicals. Gamma-glutamyltranspeptidase is also expressed in GST-P-positive foci derived from precursor cells and GST-P-positive foci are important for detoxification for carcinogen (Satoh et al., 2005). Carcinogenic activity of nitroso compounds, including DEN, is well studied, but the mechanisms of hepatotoxicity of these compounds are poorly understood. Analysis of GST-P-positive foci is valuable for the understanding of the molecular mechanisms of hepatocarcinogenesis, detoxification and hepatotoxicity.

Transgenic rats using the regulatory element of the GST-P gene revealed that a gene involved in liver cell transformation is not physically linked with the GST-P gene, but the expression is regulated by common transcription factors (Morimura et al., 1993). Further, we identified the enhancer element responsible for tumor-specific expression of the GST-P gene (Sakai and Muramatsu, 2005; Suzuki et al., 1995). This indicates that analysis of the expression profile in GST-P-positive foci leads to the identification of the responsible gene for liver cancer and understanding the mechanism of hepatocarcinogenesis.

Carcinogenesis is a genetic and epigenetic disease arising from multiple molecular changes and these events lead to changes in gene expressions. Recently, it was reported that specific differences in the gene expression profile revealed by cDNA microarray analysis of GST-P-positive foci and the surrounding tissue, and metabolic enzymes were found as up- and down-regulated genes (Suzuki et al., 2004). Direct comparisons of gene expressions between normal liver and chemically induced preneoplastic foci provide more useful information related to the molecular mechanisms of carcinogenesis. Although genes involved in transcriptional regulation are one of the most important factors in carcinogenesis, their expression levels are generally lower than those of metabolic enzymes and are hard to evaluate by DNA microarray analysis.

In this study, we conducted microarray analysis of mRNA from GST-P-positive foci in three individual rats and compared with normal liver sections from three individual rats. Labeled RNA was individually hybridized onto GeneChip Rat Expression Array 230A, and several differentially expressed genes were found to be involved in transcriptional regulation, which were not previously known to be regulated during chemically induced hepatocarcinogenesis.

2. Materials and methods

2.1. Chemical hepatocarcinogenesis of rats

Carcinogenic experiments were done according to the Solt–Farber protocol (Solt and Farber, 1976). Experiments were initiated by intraperitoneal injection of DEN (200 mg/kg) (Wako Pure Chemical Industries, Ltd., Osaka, Japan) into 5-week-old Sprague–Dawley rats. After the animals had been fed basal diets for 2 weeks, they were changed to basal diets containing 0.02% 2-acetylaminofluorene (Nacalai Tesque, Kyoto, Japan). Three weeks after DEN injection, partial hepatectomy was performed and livers were extirpated 8 weeks after DEN injection. Control rats were injected with saline and fed basal diets. All animal care and handling procedures were approved by the animal care and use committee of Osaka University.

2.2. Preparation of RNA from rat liver

To map the exact location of GST-P-positive foci, one of the serial frozen sections (10 μm) from the liver was treated with rabbit anti-GST-P antibody and immunohistochemical staining was performed with the DAKO ENVISION System (DAKO Co., Tokyo, Japan). RNA was prepared from the area corresponding to GST-P-positive foci in hyperplastic nodules induced in three individual rats or sections from three control rats by RNeasy Mini Kit (QIAGEN, Hilden, Germany).

2.3. Oligonucleotide microarray and data analysis

Target RNA amplification and labeling with biotinylated nucleotides were carried out using MEGAscript T7 Kit (Ambion, Austin, TX) and Enzo BioArray High Yield RNA Transcript Labeling Kit (Enzo Diagnostics, Farmingdale, NY) as specified by the manufacturer. The quality and size distribution of the targets were determined using the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA). Labeled and fragmented RNA of individual rats was hybridized onto GeneChip Rat Expression Array 230A (Affymetrix, Santa Clara, CA) using standard methods. We calculated the background correction and normalization of the array data using Robust Multi-Array (RMA) method in the R package. Statistics of differential expression between genes was estimated using the linear modeling features of the limma library of the R. Limma computes p-values of moderated t-statistics by emprical Bayes shrinkage of the standard error toward a common value.

3. Results and discussion

3.1. Detection of GST-P-positive foci

To examine the expression profile of GST-P-positive foci during hepatocarcinogenesis, hyperplastic nodule-induced rats were prepared according to the Solt–Farber procedure (Solt and Farber, 1976). Eight weeks after DEN treatment, the livers, which had a large number of foci and nodules, were excised and immunohistochemical experiments indicated that an approximately 70–80% region contained GST-P-positive foci (data not shown).

3.2. DNA microarray analysis of gene expression

We prepared biotinylated target RNA from GST-P-positive foci and normal liver sections in three hyperplastic nodule-induced rats and three control rats, respectively. Each target was individually hybridized with the Rat 230A Array containing the primary probe sets against well-annotated full-length genes. The scatter plot of the gene expression pattern between three independent control rats showed excellent reproducibility of results with an average correlation coefficient ± S.D. (0.93 ± 0.035). In the case of GST-P-positive foci, good reproducibility was also obtained (average of correlation coefficient ± S.D., 0.95 ± 0.0068). On the other hand, the average of the correlation coefficient ± S.D. derived from comparisons of control versus GST-P-positive foci was 0.74 ± 0.026. These results indicate that expression profiles in the same groups were indistinguishable, but dysregulation in many genes was observed during hepatocarcinogenesis.

3.3. Expression profile of enzymes involved in metabolism and detoxification

Genes were examined in which the expression was enhanced or reduced in GST-P-positive foci compared with control liver. Significantly changed transcripts were selected by moderated t-statistics. There were 15,923 probes on the chip, and 375 and 199 genes were significantly up- and down-regulated, respectively, with log ratio values outside of 1 to -1 ($p < 0.05$). Of these, the twenty most up- and down-regulated genes are shown in Tables 1 and 2 together with p-values for statistical significance. Significant up-regulation of the GST-P gene (*Gstp1/Gstp2*) expression was observed in GST-P-positive foci (Table 1). It is known that enzymes involved in metabolism and detoxification are induced or repressed during chemical hepatocarcinogenesis (Sato, 1989; Suzuki et al., 2004). Overexpression of metabolic enzymes, which were reported to demonstrate increased expression in hyperplastic nodules, including aldehyde dehydrogenase, aflatoxin B1 aldehyde reductase, NAD(P)H dehydrogenase and glutathione peroxidase 2, and the suppression of carbonic anhydrase 3, were detected by DNA microarray analysis (Tables 1 and 2). Semi-quantitative reverse transcriptase-coupled PCR experiments were performed on several selected genes and it was confirmed that the expression patterns were similar to those observed with microarray (data not shown). These results indicate that our study would be suitable for discovering new genes to provide new information on hepatocarcinogenesis, detoxification, and hepatotoxicity.

3.4. Expression profile of transcripts involved in transcription

Probes on the chip were divided into various categories based on Gene Ontology (Ashburner et al., 2000). Observation and analysis of the expression profile for genes involved in transcription, one of the categories, provides valuable information to understand the mechanism of carcinogenesis. Transcripts categorized as transcription with significantly changed expression with log ratio values outside 1 to -1 are listed in Tables 3 and 4, and most have not previously been found to be differentially expressed during chemically induced hepatocarcinogenesis. For example, Pawr was overexpressed in GST-P-positive foci. Pawr also termed par-4, which interacts with Wilms' tumor 1 (WT1) and modulates functions of WT1 (Johnstone et al., 1996). WT1 is a sequence-specific DNA-binding protein and functions as both a tumor suppressor and an oncogenic factor (Loeb and Sukumar, 2002). The WT1 gene exerts an oncogenic function rather than a tumor-suppressor gene function in solid tumors as well as leukemias (Sugiyama, 2001). In prostate cancer cell line, ectopic expression PAWR repressed Bcl-2 expression through WT1 (Cheema et al., 2003). However, Loeb revealed that

Table 1
A list of the twenty genes most highly induced in GST-P-positive foci

Gene symbol	Gene title	Log ratio	*p*-value	GenBank accession no.
Akr1b8	Aldo-keto reductase family 1, member B8	6.65	8.40E−07	NM_173136
Yc2	Glutathione-*S*-transferase Yc2 subunit	5.38	1.32E−06	NM_001009920
Gstp1/Gstp2	Glutathione-*S*-transferase, pi 1/2	5.14	4.31E−06	NM_012577 NM_138974
Aldh1a1	Aldehyde dehydrogenase family 1, member A1	4.42	9.18E−05	NM_022407
Akr7a3	Aflatoxin B1 aldehyde reductase	4.17	4.17E−05	NM_013215
Aldh3a1	Aldehyde dehydrogenase family 3, member A1	3.89	1.07E−04	NM_031972
Nqo1	NAD(P)H dehydrogenase, quinone 1	3.72	9.79E−05	NM_017000
Serpinb 1 a_predicted	Serine (or cysteine) proteinase inhibitor, clade B, member 1a (predicted)	3.72	3.22E−02	NM_001031642
LOC294067	Similar to ww domain binding protein 5	3.71	1.92E−05	XM_215278
RDG:621458	Neurofilament, light polypeptide	3.64	8.68E−04	NM_031783
RGD1310542_predicted	Similar to RIKEN cDNA 4930457P18 (predicted)	3.59	3.13E−04	NM_001014154
–	Brain expressed X-linked 1	3.56	6.18E−04	NM_001037365
Rnf30_predicted	RING finger protein 30 (predicted)	3.54	2.56E−04	NM_001013217
Anxa2	Annexin A2	3.53	6.18E−04	NM_019905
Ca2	Carbonic anhydrase 2	3.51	4.17E−05	NM_019291
Ddit4l	DNA-damage-inducible transcript 4-like	3.45	5.30E−06	NM_080399
Gpx2	Glutathione peroxidase 2	3.41	8.23E−05	NM_183403
RGD:1303152	Ectodermal-neural cortex 1	3.31	1.36E−04	NM_001003401
Dscr1l1	Down syndrome critical region gene 1 -like 1	3.27	3.17E−05	NM_175578
LOC500040	Similar to testis-derived transcript	3.23	2.40E−04	XM_575396

Log ratio indicates a logarithm of the fold-change vs. the expression level of the control rats. Statistics of differential expression between genes was estimated using the linear modeling features of the limma library of the R. Limma computes *p*-values of moderated *t*-statistics by emprical Bayes shrinkage of the standard error toward a common value.

Table 2
A list of the twenty genes most highly repressed in GST-P-positive foci

Gene symbol	Gene title	Log ratio	*p*-value	GenBank accession no.
Pgcl4	Alpha-2u globulin PGCL4	−4.75	1.88E−02	NM_147215
Pgcl3/5/1/4	Alpha-2u globulin PGCL3/5/1/4	−4.71	1.79E−02	NM_147212 NM_147213 NM_147214 NM_147215
Pgcl4	Alpha-2u globulin PGCL4	−4.48	8.86E−03	NM_147215
Ca3	Carbonic anhydrase 3	−3.72	2.20E−02	NM_019292
Apoa4	Apolipoprotein A-IV	−3.63	1.01E−04	NM_012737
Cyp3a13	Cytochrome P450, family 3, subfamily a, polypeptide 13	−3.53	6.18E−04	NP_671739.1
Cyp2c	Cytochrome P450, subfamily IIC (mephenytoin 4-hydroxylase)	−3.50	8.20E−03	NM_019184
Ca3	Carbonic anhydrase 3	−3.45	6.62E−03	NM_019292
LOC368066	Similar to thioether S-methyltransferase	−3.25	2.21E−03	XM_347233
Fasn	Fatty acid synthase	−3.04	2.66E−03	NM_017332
Cyp2a2	Cytochrome P450, subfamily 2A, polypeptide 1	−3.03	2.28E−02	NM_012693
Sult1a2	Sulfotransferase family 1 A, member 2	−2.83	2.52E−03	NM_031732
Ust5r	integral membrane transport protein UST5r	−2.75	2.21E−02	NM_134380
Apoa2	Apolipoprotein A-II	−2.74	1.83E−03	NM_013112
Slc27a5	bile acid CoA ligase	−2.69	2.10E−03	NM_024143
–	Ab2-060	−2.68	1.36E−04	AI411138
Avpr1a	Arginine vasopressin receptor 1A	−2.53	7.61E−03	NM_053019
–	Malic enzyme 3, NADP(+)-dependent, mitochondrial (predicted)	−2.51	6.18E−04	AA964869
Thrsp	Thyroid hormone responsive protein	−2.46	2.54E−04	NM_012703
Slc21a10	Solute carrier family 21, member 10	−2.41	3.05E−02	NM_031650

Log ratio and *p*-value are described in Table 1.

Table 3
A list of genes involved in transcription induced in GST-P-positive foci

Gene Symbol	Gene title	Log ratio	p-value	GenBank accession no.
Rnf30_predicted	Ring finger protein 30 (predicted)	3.54	2.56E−04	NM_001013217
Copeb	Core promoter element binding protein	1.92	2.21E−02	NM_031642
Basp1	Brain acidic membrane protein	1.69	1.50E−02	NM_022300
Copeb	Core promoter element binding protein	1.66	1.70E−03	NM_031642
Htatip2_predicted	HIV-1 Tat interactive protein 2 (predicted)	1.61	6.54E−04	XM_214927
L3mbtl2_predicted	l(3)mbt-like 2 (Drosophila) (predicted)	1.59	1.29E−03	NM_001033695
Ppp2ca	Protein phosphatase 2a, catalytic subunit, alpha isoform	1.58	3.77E−03	NM_017039
Ppp2ca	Protein phosphatase 2a, catalytic subunit, alpha isoform	1.56	2.53E−03	AI009467
Maged1	Melanoma antigen, family D, 1	1.41	9.41E−03	NM_053409
Als2cr3	Amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 3 homolog (human)	1.41	4.83E−02	NM_133560
Hmgb2	High mobility group box 2	1.35	4.16E−02	XM_573272
Npm1	Nucleophosmin 1	1.34	4.18E−02	NM_012992
Pdlim1	PDZ and LIM domain 1	1.27	1.22E−02	NM_017365
Mdm2_predicted	Transformed mouse 3T3 cell double minute 2 (predicted)	1.24	2.64E−02	XM_235169
Sox4_predicted	SRY-box containing gene 4 (predicted)	1.19	2.36E−02	XM_344594
Npm1	Nucleophosmin 1	1.18	3.92E−03	NM_012992
Carm1_predicted	Coactivator-associated arginine methyltrnsferase1 (predicted)	1.14	2.07E−02	NM_001030041
Tgif_predicted	TG interacting factor predicted	1.14	3.74E−03	NM_001015020
Ivns 1 abp_predicted	Influenza virus NS1A binding protein (predicted)	1.09	1.09E−02	XM_213898
Pawr	PRKC, apoptosis, WT1, regulator	1.07	4.28E−03	NM_033485
RGD1304726_predicted	Similar to RIKEN cDNA 6330509G02 (predicted)	1.06	2.49E−02	NM_001024993
Ets2	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)	1.03	1.24E−02	XM_239510
Rbbp7	Retinoblastoma binding protein 7	1.02	4.15E−03	NM_031816

Log ratio and p-value are described in Table 1.

WT1 transcriptionally up-regulates anti-apoptotic genes such as Bcl-2 in rhavdoid cell line (Loeb, 2006; Mayo et al., 1999). In GST-P-positive foci, we found mRNA overexpression of Bcl-2 (log ratio, 0.776; $p = 0.0330$) by microarray. The different regulation mechanism of Bcl-2 expression is caused by cell lineage and isoform-specific differences in WT1 function (Loeb, 2006; Mayo et al., 1999). Further characterization of pawr would lead the

Table 4
A list of genes involved in transcription repressed in GST-P-positive foci

Gene symbol	Gene title	Log ratio	p-value	GenBank accession no.
Thrsp	Thyroid hormone responsive protein	−2.46	2.54E−04	NM_012703
Thrsp	Thyroid hormone responsive protein	−1.82	2.14E−02	NM_012703
Thrsp	Thyroid hormone responsive protein	−1.68	6.14E−03	NM_012703
Srebf1	Sterol regulatory element binding factor 1	−1.57	4.24E−03	XM_213329
Atf5	Activating transcription factor 5	−1.39	8.82E−03	NM_172336
Sec 14l2	SEC14-like 2 (S. cerevisiae)	−1.36	5.06E−03	NM_053801
	Protocadherin 1 (cadherin-like 1) (predicted)	−1.27	8.40E−03	XM_225997
Gls2	Liver mitochondrial glutaminase	−1.22	1.40E−02	NM_138904
Per2	Period homolog 2	−1.14	7.45E−03	NM_031678
Idb4	Inhibitor of DNA binding 4	−1.12	9.70E−03	NM_175582
Clp1	Cardiac lineage protein 1	−1.11	3.12E−02	NM_001025136
Tgfb1i4	Transforming growth factor beta 1 induced transcript 4	−1.10	5.86E−03	L25785
Rxra	Retinoid X receptor alpha	−1.02	1.24E−03	NM_012805
Hes6_predicted	Hairy and enhancer of split 6 (Drosophila) (predicted)	−1.01	5.41E−03	NM_001013179

Log ratio and p-value are described in Table 1.

understanding of oncogenic or tumor suppressor gene function of WT1 during hepatocarcinogenesis.

On the other hand, several sequence-specific DNA-binding transcription factors were repressed during hepatocarcinogenesis (e.g. Sterol-regulatory-element binding factor 1 (Srebf1)/Sterol-regulatory-element binding protein 1 (Srebp1) and retinoid X receptor alpha (RXRalpha)) (Table 4). SREBPs have been established as lipid synthetic transcription factors for cholesterol and fatty acid synthesis (Eberle et al., 2004). The expression of fatty acid synthase and apolipoprotein A-II are mainly regulated by SREBP1, and these genes were suppressed in GST-P-positive foci (Table 2). Further, SREBP1 is required for the induction of thyroid hormone-responsive protein (THRSP) in hepatocytes (Martel et al., 2006). Brown et al. (1997) reported that exposure of Thrsp antisense oligonucleotide inhibited the expression of mRNAs encoding lipogenic (fatty acid synthase, ATP citrate lyase and malic enzyme) and glycolytic (pyruvate kinase) enzymes. The log ratios of these genes were -3.04 ($p=0.00266$), -1.51 ($p=0.00105$), -0.508 ($p=0.0221$) and -2.05 ($p=0.000531$), respectively. These observations suggest that the aberrant decrease of lipogenic and glycolytic enzymes may be caused by the suppression of SREBP1. This raises the possibility that hepatotoxicity induced by nitroso compounds would be cased by the down regulation of SREBP1.

One of nuclear receptors, retinoid X receptor alpha (RXRalpha) was also decreased in GST-P-positive foci. RXRalpha dimerizes with constitutive androstane receptor (CAR), pregnane X receptor (PXR) and peroxisome proliferator-activated receptor (PPARalpha). Hepatocyte RXRalpha-deficient mice revealed that hepatocyte RXRalpha is required for induction of metabolic enzymes by the ligands of CAR, PXR, and PPARalpha, and is essential for xenobiotic metabolism in vivo (Cai et al., 2002). Hepatotoxicity may be caused by decrease of RXRalpha expression in GST-P-positive foci.

3.5. Expression of transcripts coding chromatin modification enzymes

Sequence-specific transcription factors require cofactors for transcription from the chromatin context and chromatin components affect gene expression (Sterner and Berger, 2000). The expression of cofactors and chromatin components during hepatocarcinogenesis has not been studied well. Recent studies demonstrated that cofactors possess histone modification activities, which are required for the change of chromatin conformation and the regulation of gene function. Generally, histone acetylation promotes transcription, although histone methylation both positively and negatively regulates gene expression dependent on the position of the lysine residue on histone. An epigenetic program including histone and DNA modifications is important for the maintenance of inheritable information and the disturbance of epigenetic balances may lead to alterations in gene expression, resulting in cellular transformation and malignant growth (Lund and van Lohuizen, 2004). Microarray analysis revealed that coactivator-associated arginine methyltransferase (Carm1) and Rbbp7, also termed retinoblastoma suppressor-associated protein 46 (RbAp46), were induced in GST-P-positive foci. CARM1 catalyzes the methylation of histone H3 at Arg17 and can also function as a coactivator for transcription factor NF-E2-related factor 2 (Nrf2), which regulates the induction of Phase II detoxifying enzymes, including GST-P, through its transactivation domain (Lin et al., 2006; Miao et al., 2006). Although increased Nrf2 was detected in hyperplasic nodules, the extremely high level of GST-P expression during hepatocarcinogenesis was difficult to explain by the slight induction of Nrf2 alone (Ikeda et al., 2004). Here we found the overexpression of Carm1 in GST-P-positive foci. Increase of both Nrf2 and Carm1 expression and the cooperative regulation of gene expression may lead to the induction of GST-P expression during hepatocarcinogenesis.

We also found the induction of RbAp46, which contributed to the regulation of gene expression as a subunit of histone acetyltransferase, histone deacetylase and chromatin remodeling complexes NURD (Zhang et al., 1999). Li et al., 2003 reported that the expression of RbAp46 suppressed colony formation in soft agar, and inhibited tumor formation in nude mice. They also showed that high levels of RbAp46 expression promoted apoptotic cell death, resulting in the inhibition of tumorigenicity of neoplastigenic breast epithelial cells. These results suggest that overexpressed RbAp46 in GST-P-positive cells may function as a suppressor of tumorigenicity in the early stage of hepatocarcinogenesis.

3.6. Expression of transcripts coding chromatin components and related factors

High mobility group box 2 (Hmgb2), a member of HMGB family proteins, was up-regulated in GST-P-positive cells. HMGB proteins are abundant nonhistone nuclear proteins that have been found in association with chromatin. HMGB family proteins contain two DNA-binding HMG-box domains and bind to DNA without sequence specificity, but play important architectural roles in the assembly of nucleoprotein complexes in a variety of biological processes including the initiation of transcription and DNA repair (Thomas, 2001). Further, HMGB2, while showing no coactivator activity on its own, can promote transcription activity together with histone acetyltransferase. HMGB2 acts mainly at the level of elongation and is a coactivator for transcription from chromatin templates (Guermah et al., 2006). Hmgb2 is frequently overexpressed in malignant gastrointestinal stromal tumors and ovarian cancer (Koon et al., 2004; Ouellet et al., 2006). Overexpression of Hmgb2 may be common feature of carcinogenesis. HMGB2 binds with high affinity to DNA modified with the cancer chemotherapeutic drug cisplatin and enhancement of cisplatin sensitivity in Hmgb2 transfected human lung cancer cells (Arioka et al., 1999; Farid et al., 1996). Cisplatin-induced hepatotoxicity may be promoted by overexpressed Hmgb2.

Nucleophosmin was identified as a positively regulated gene in GST-P-positive cells. Nucleophosmin is a key regulator for centrosome duplication, the maintenance of genomic integrity, and ribosome assembly. At the steady state, nucleophosmin localizes mainly in the nucleolus, whereas aberrant cytoplasmic localization of nucleophosmin is observed in acute myeloid leukemias (Mariano et al., 2006). Observation of localization of nucleophosmin would be important for the understanding of roles of overexpressed nucleophosmin in GST-P-positive foci. Recent studies suggest that nucleophosmin may be a Ran-Crm1 substrate that controls centrosome duplication and utilizes a conserved Crm1-dependent nuclear export sequence in its amino terminus to enable shuttling between the nucleolus/nucleus and cytoplasm (Wang et al., 2005; Yu et al., 2006). Further, purification of nucleophosmin binding protein revealed that nucleophosmin directly interacted with ribosomal protein L5. This interaction mediated the colocalization of nucleophosmin with both maturing nuclear 60S ribosomal subunits and newly exported and assembled 80S ribosomes (Yu et al., 2006). Interestingly, Crm1 (log ratio, 0.908; $p = 0.0427$) and ribosomal protein L5 (log ratio, 0.830; $p = 0.00429$) were also up-regulated in GST-P-positive foci. Overexpression of these genes may disturb multiple processes involved in nucleophosmin, which accelerate oncogenesis.

DNA microarray analysis in this study uncovered several genes, which expression was induced or repressed during hepatocarcinogenesis, and some of these genes possess anti-oncogenic as well as oncogenic activities and may be involved in regulation of GST-P expression. Our study provided a candidate gene link to the early stage of carcinogenesis and hepatotoxicity. To elucidate the mechanisms of the early stage of hepatocarcinogenesis mediated by these genes, further characterization of aberrantly expressed genes in GST-P-positive cells is necessary. We proceed to observe the effect of overexpression of these genes up-regulated during hepatocarcinogenesis, especially epigenetics regulatory factors, on transformation and the induction of GST-P expression.

Acknowledgements

This research was supported in part by grants from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, Japan Society for the Promotion of Science (JSPS), a Grant-in-Aid for Scientific Research from the Ministry of Health and Labor of Japan, Long-range Research Initiative (LRI) by Japan Chemical Industry Association (JCIA) and Sankyo Foundation of Life Science.

References

Arioka, H., Nishio, K., Ishida, T., Fukumoto, H., Fukuoka, K., Nomoto, T., Kurokawa, H., Yokote, H., Abe, S., Saijo, N., 1999. Enhancement of cisplatin sensitivity in high mobility group 2 cDNA-transfected human lung cancer cells. Jpn. J. Cancer Res. 90, 108–115.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.

Brown, S.B., Maloney, M., Kinlaw, W.B., 1997. "Spot 14" protein functions at the pretranslational level in the regulation of hepatic metabolism by thyroid hormone and glucose. J. Biol. Chem. 272, 2163–2166.

Cai, Y., Konishi, T., Han, G., Campwala, K.H., French, S.W., Wan, Y.J., 2002. The role of hepatocyte RXR alpha in xenobiotic-sensing nuclear receptor-mediated pathways. Eur. J. Pharm. Sci. 15, 89–96.

Cheema, S.K., Mishra, S.K., Rangnekar, V.M., Tari, A.M., Kumar, R., Lopez-Berestein, G., 2003. Par-4 Transcriptionally Regulates Bcl-2 through a WT1-binding Site on the bcl-2 Promoter. J. Biol. Chem. 278, 19995–20005.

Eberle, D., Hegarty, B., Bossard, P., Ferre, P., Foufelle, F., 2004. SREBP transcription factors: master regulators of lipid homeostasis. Biochimie 86, 839–848.

Eriksson, L.C., Sharma, R.N., Roomi, M.W., Ho, R.K., Farber, E., Murray, R.K., 1983. A characteristic electrophoretic pattern of cytosolic polypeptides from hepatocyte nodules generated during liver carcinogenesis in several models. Biochem. Biophys. Res. Commun. 117, 740–745.

Farid, R.S., Bianchi, M.E., Falciola, L., Engelsberg, B.N., Billings, P.C., 1996. Differential binding of HMG1, HMG2, and a single HMG box to cisplatin-damaged DNA. Toxicol. Appl. Pharmacol. 141, 532–539.

Guermah, M., Palhan, V.B., Tackett, A.J., Chait, B.T., Roeder, R.G., 2006. Synergistic functions of SII and p300 in productive activator-dependent transcription of chromatin templates. Cell 125, 275–286.

Ikeda, H., Nishi, S., Sakai, M., 2004. Transcription factor Nrf2/MafK regulates rat placental glutathione-S-transferase gene during hepatocarcinogenesis. Biochem. J. 380, 515–521.

Ito, N., Tamano, S., Shirai, T., 2003. A medium-term rat liver bioassay for rapid in vivo detection of carcinogenic potential of chemicals. Cancer Sci. 94, 3–8.

Johnstone, R.W., See, R.H., Sells, S.F., Wang, J., Muthukkumar, S., Englert, C., Haber, D.A., Licht, J.D., Sugrue, S.P., Roberts, T., Rangnekar, V.M., Shi, Y., 1996. A novel repressor, par-4, modulates transcription and growth suppression functions of the Wilms' tumor suppressor WT1. Mol. Cell. Biol. 16, 6945–6956.

Koon, N., Schneider-Stock, R., Sarlomo-Rikala, M., Lasota, J., Smolkin, M., Petroni, G., Zaika, A., Boltze, C., Meyer, F., Andersson, L., Knuutila, S., Miettinen, M., El-Rifai, W., 2004. Molecular targets for tumour progression in gastrointestinal stromal tumours. Gut 53, 235–240.

Li, G.C., Guan, L.S., Wang, Z.Y., 2003. Overexpression of RbAp46 facilitates stress-induced apoptosis and suppresses tumorigenicity of neoplastigenic breast epithelial cells. Int. J. Cancer 105, 762–768.

Lin, W., Shen, G., Yuan, X., Jain, M.R., Yu, S., Zhang, A., Chen, J.D., Kong, A.N., 2006. Regulation of Nrf2 transactivation domain activity by p160 RAC3/SRC3 and other nuclear co-regulators. J. Biochem. Mol. Biol. 39, 304–310.

Loeb, D.M., 2006. WT1 Influences apoptosis through transcriptional regulation of Bcl-2 family members. Cell Cycle 5, 1249–1253.

Loeb, D.M., Sukumar, S., 2002. The role of WT1 in oncogenesis: tumor suppressor or oncogene? Int. J. Hematol. 76, 117–126.

Lund, A.H., van Lohuizen, M., 2004. Epigenetics and cancer. Genes Dev. 18, 2315–2335.

Mariano, A.R., Colombo, E., Luzi, L., Martinelli, P., Volorio, S., Bernard, L., Meani, N., Bergomas, R., Alcalay, M., Pelicci, P.G., 2006. Cytoplasmic localization of NPM in myeloid leukemias is dictated by gain-of-function mutations that create a functional nuclear export signal. Oncogene 25, 4376–4380.

Martel, P.M., Bingham, C.M., McGraw, C.J., Baker, C.L., Morganelli, P.M., Meng, M.L., Armstrong, J.M., Moncur, J.T., Kinlaw, W.B., 2006. S14 protein in breast cancer cells: direct evidence of regulation by SREBP-1c, superinduction with progestin, and effects on cell growth. Exp. Cell Res. 312, 278–288.

Mayo, M.W., Wang, C.Y., Drouin, S.S., Madrid, L.V., Marshall, A.F., Reed, J.C., Weissman, B.E., Baldwin, A.S., 1999. WT1 modulates apoptosis by transcriptionally upregulating the bcl-2 proto-oncogene. EMBO J. 18, 3990–4003.

Miao, F., Li, S., Chavez, V., Lanting, L., Natarajan, R., 2006. Coactivator-associated arginine methyltransferase-1 enhances nuclear factor-kappaB-mediated gene transcription through methylation of histone H3 at arginine 17. Mol. Endocrinol. 20, 1562–1573.

Morimura, S., Suzuki, T., Hochi, S., Yuki, A., Nomura, K., Kitagawa, T., Nagatsu, I., Imagawa, M., Muramatsu, M., 1993. Trans-activation of glutathione transferase P gene during chemical hepatocarcinogenesis of the rat. Proc. Natl. Acad. Sci. U.S.A. 90, 2065–2068.

Ouellet, V., Page, C.L., Guyot, M.C., Lussier, C., Tonin, P.N., Provencher, D.M., Mes-Masson, A.M., 2006. SET complex in serous epithelial ovarian cancer. Int. J. Cancer 119, 2119–2126.

Sakai, M., Muramatsu, M., 2005. Regulation of GST-P gene expression during hepatocarcinogenesis. Methods Enzymol. 401, 42–61.

Sato, K., 1989. Glutathione transferase as markers of preneoplasia and neoplasia. Adv. Cancer Res. 52, 205–255.

Satoh, K., Hatayama, I., Tateoka, N., Tamai, K., Shimizu, T., Tatematsu, M., Ito, N., Sato, K., 1989. Transient induction of single GST-P positive hepatocytes by DEN. Carcinogenesis 10, 2107–2111.

Satoh, K., Kitahara, A., Soma, Y., Inaba, Y., Hatayama, I., Sato, K., 1985. Purification, induction, and distribution of placental glutathione transferase: a new marker enzyme for preneoplastic cells in the rat chemical hepatocarcinogenesis. Proc. Natl. Acad. Sci. U.S.A. 82, 3964–3968.

Satoh, K., Takahashi, G., Miura, T., Hayakari, M., Hatayama, I., 2005. Enzymatic detection of precursor cell populations of preneoplastic foci positive for gamma-glutamyltranspeptidase in rat liver. Int. J. Cancer 115, 711–716.

Solt, D., Farber, E., 1976. New principle for the analysis of chemical carcinogenesis. Nature 263, 701–703.

Sterner, D.E., Berger, S.L., 2000. Acetylation of histones and transcription-related factors. Microbiol. Mol. Biol. Rev. 64, 435–459.

Sugiyama, H., 2001. Wilms' tumor gene WT1: its oncogenic function and clinical application. Int. J. Hematol. 73, 177–187.

Suzuki, S., Asamoto, M., Tsujimura, K., Shirai, T., 2004. Specific differences in gene expression profile revealed by cDNA microarray analysis of glutathione-S-transferase placental form (GST-P) immunohistochemically positive rat liver foci and surrounding tissue. Carcinogenesis 25, 439–443.

Suzuki, T., Imagawa, M., Hirabayashi, M., Yuki, A., Hisatake, K., Nomura, K., Kitagawa, T., Muramatsu, M., 1995. Identification of an enhancer responsible for tumor marker gene expression by means of transgenic rats. Cancer Res. 55, 2651–2655.

Thomas, J.O., 2001. HMG1 and 2: architectural DNA-binding proteins. Biochem. Soc. Trans. 29, 395–401.

Wang, W., Budhu, A., Forgues, M., Wang, X.W., 2005. Temporal and spatial control of nucleophosmin by the Ran-Crm1 complex in centrosome duplication. Nat. Cell Biol. 7, 823–830.

Yu, Y., Maggi Jr., L.B., Brady, S.N., Apicelli, A.J., Dai, M.S., Lu, H., Weber, J.D., 2006. Nucleophosmin is essential for ribosomal protein L5 nuclear export. Mol. Cell. Biol. 26, 3798–3809.

Zhang, Y., Ng, H.H., Erdjument-Bromage, H., Tempst, P., Bird, A., Reinberg, D., 1999. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. Genes Dev. 13, 1924–1935.

Journal of Pharmacological Sciences
©2006 The Japanese Pharmacological Society

Current Perspective

MicroRNA: Biogenetic and Functional Mechanisms and Involvements in Cell Differentiation and Cancer

Soken Tsuchiya[1], Yasushi Okuno[1], and Gozoh Tsujimoto[1,*]

[1] *Department of Genomic Drug Discovery Science, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan*

Received June 2, 2006

Abstract. MicroRNAs (miRNAs) are endogenous small noncoding RNAs (20 – 23 nucleotides) that negatively regulate the gene expressions at the posttranscriptional level by base pairing to the 3' untranslated region of target messenger RNAs. Hundreds of miRNAs have been identified in humans and evolutionarily conserved from plants to animals. It is revealed that miRNAs regulate various physiological and pathological pathways such as cell differentiation, cell proliferation, and tumoriogenesis. By the computational analysis, it is predicted that 30% of protein-encoding genes are regulated by miRNAs. In this review, we discuss recent remarkable advances in the miRNA biogenetic and functional mechanisms and the involvements of miRNAs in cell differentiation, especially in hematopoietic lineages, and cancer. These evidences offer the possibility that miRNAs would be potentially useful for drug discovery.

Keywords: microRNA, RNA cleavage, translational repression, target mRNA, base pairing

Introduction

MicroRNAs (miRNAs) are endogenous short noncoding RNA molecules (20 – 23 nucleotides) that regulate cell differentiation, cell proliferation, and apoptosis through post-transcriptional suppression of gene expression by binding to the complementary sequence in the 3' untranslated region (3'UTR) of target messenger RNAs (mRNAs) (1). Hundreds of miRNAs have been identified in humans and they are evolutionarily conserved (1, 2). In addition, the presence of up to 1000 miRNAs is estimated by computational analysis (3). Strikingly, 30% of protein-encoding genes in humans are predicted to be regulated by miRNAs (4). Recently, it has been revealed that altered expression of specific miRNA genes contributes to the initiation and progression of diseases such as cancer (5 – 10). This review focuses on the biogenetic and functional mechanisms and the involvements in cell differentiation and cancer in mammalian miRNAs and the utility of miRNAs in drug discovery.

Mechanisms of biogenesis and function

Most miRNA genes are located in the introns of host genes or outside genes. Unlike *Drosophila*, most of the human miRNA genes individually exist, although some human miRNAs are found in polycistronic clusters (5, 8, 9).

The miRNAs are synthesized through multiple steps (Fig. 1). Initially, the miRNAs are transcribed as long RNA precursors (pri-miRNAs) (11). As pri-miRNAs usually contain the cap structure and the poly(A) tail, it is suggested that the transcription of miRNAs is carried out by RNA polymerase II (12). The pri-miRNAs are processed into the precursors of approximately 70 nucleotides (pre-miRNAs) with a stem-loop structure and a two nucleotide 3' overhang by the RNase III enzyme Drosha and the double-stranded-RNA-binding protein DGCR8/Pasha (13, 14), and pre-miRNAs are exported from the nucleus to the cytoplasm by Exportin-5 in a Ran guanosine triphosphate-dependent manner (15). Pre-miRNAs exported in the cytoplasm are processed by another RNase III enzyme, Dicer, and only one strand (guide strand) as a mature

*Corresponding author. gtsuji@pharm.kyoto-u.ac.jp
Published online in J-STAGE
doi: 10.1254/jphs.CPJ06013X

Invited article

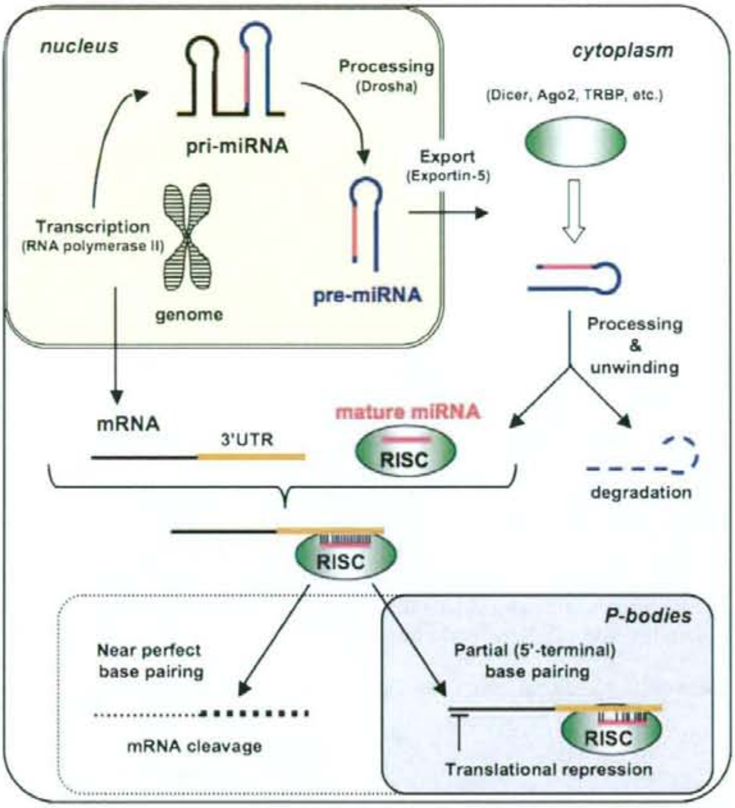

Fig. 1. Diagram of the miRNA biogenetic and functional mechanisms. Whether the target mRNA cleavage by RISC occurs in the cytoplasm or P-bodies remains unknown.

miRNA is incorporated into a RNA-induced silencing complex (RISC) that mediates either target RNA cleavage or translational inhibition, while the another strand (passenger strand) is excluded. Which strand is incorporated in RISC is determined by the stability of the base pairs at the 5' end of the duplex (16, 17). The incorporated guide strand guides the RISC to the complementary sequence in the 3'UTR of target mRNA. When the guide strand shares perfect or near perfect base pairing with the 3'UTR of target mRNA, the target mRNA is degraded by Argonaute2 (Ago2), a component of RISC (18). On the contrary, when the guide strand shares partial base pairing, translation is target-specifically repressed without the target mRNA degradation (19). Recent studies have revealed that RISC is at least composed of Dicer, Ago2, and the double-strand RNA binding protein TRBP, and RISC efficiently processes pre-miRNAs to mature miRNAs (20). Furthermore, RISC more efficiently cleaves target mRNAs by using the pre-miRNAs than the duplex miRNAs that do not have the stem-loop. These results suggest that miRNA processing by Dicer, assembly of the mature miRNA into RISC, and target RNA cleavage by Ago2 are coupled. Compared to the RNA cleavage mechanism by Ago2, the translational repression mechanism by miRNAs had been poorly understood. Recently, it was revealed that the target mRNAs binding to RISC through partial base pairing are accumulated in the cytoplasmic foci referred to as processing bodies (P-bodies) (21, 22). P-bodies, in which the mRNAs are stored or degraded by the decapping enzymes and exonucleases, do not contain the translational machinery (23). Furthermore, the disruption of P-bodies by the silencing of GW182, a key protein in P-body, inhibits translational silencing in not only partial base pairing but also perfect base pairing (24), although the localization of target mRNA with perfect base pairing is not detected in P-bodies (21). These results suggest that, at least in part, translational repression appears to be caused by the recruitment of target mRNAs to P-bodies. However, whether localization of the RISC-target mRNA complex in P-bodies is a cause or a result of the translational repression and whether the target mRNA cleavage by RISC occurs in the cytoplasm or P-bodies remain controversial issues.

Cell differentiation

Increasing evidence indicates that miRNAs have distinct expression patterns among tissues and cells in different differentiation stage (25). It is reported that overexpression of miR-124, which is preferentially expressed in brain, shifted the gene expression profile of HeLa cells towards that of the brain. Similarly, overexpression of miR-1 shifted the expression profile towards that of the muscle in that miR-1 is preferentially expressed (25). These results indicate that miRNAs play important roles in cell differentiation and characterization.

Recently, it was revealed that miRNAs also played critical roles in the differentiation of mammalian hematopoietic lineage. For example, miR-181 is preferentially expressed in the thymus and B-lymphoid cells of mouse bone marrow and promotes B cell differentiation by overexpression in hemapoietic stem/progenitor cells (26). Conversely, overexpression of the miR-181a, one member of the miR-181 family, was reported to repress megakaryoblast differentiation in humans (27). By the induction of megakaryoblast differentiation, the expression of endogenous miR-181a is downregulated through the acetylcholinesterase, protein kinase (PK) C, and PKA cascade. The expression of miR-130a is also downregulated by the induction of megakaryoblast differentiation (28). miR-130a targets the transcriptional factor MAFB that is a transcriptional activator of GPIIB, an important protein for platelet physiology. Furthermore, miR-223 is upregulated by the retinoic acid-induced replacement of NFI-A with CCAAT/Enhancer binding protein (C/EBP) α, and promotes human granulopoiesis (29). As miR-223 repressed NFI-A translation, the upregulation of miR-223 by C/EBPα and granulopoiesis further accelerated through positive feedback by miR-223.

Cancer

It has been revealed that the change of miRNA expressions contributes to the initiation and progression of cancer. More than 50% of miRNAs are located in cancer-associated genomic regions or in fragile sites (5). The expression of miR-15a and miR-16, which locate as a cistronic cluster at 13q14, is deleted or decreased in most cases (approx. 68%) of B cell chronic lymphocytic leukemia (B-CLL) (6). Both these miRNAs negatively regulate the expression of B cell lymphoma 2 (Bcl2), that is reported to be expressed in many types of cancer including leukemias, and inhibit cell death (7). Overexpression of miR-15 and miR-16 in the MEG-01 cell line actually induces the apoptosis. Inversely, one cluster of miRNAs, miR-17 – 92 polycistron, was found to increase in the cancers such as B-CLL (8). The expression of six miRNAs in this cluster is upregulated by c-myc, whose expression and/or function are one of the most common abnormalities in human cancers, and miR-17-5p and miR-20a included in this cluster negatively regulate the expression of transcriptional factor E2F1 (9). Furthermore, mice reconstituted with hemotopoietic stem cells overexpressing miR-17 – 19b exhibit accelerated c-myc-induced lymphomagenesis (8). Furthermore, it was revealed that miRNA expression profiles enable researchers to successfully classify poorly characterized human tumors that can not be accurately classified by mRNA expression profiles (10). These results show the possibility that miRNAs have clinical benefits as not only therapeutic targets but also a tool for cancer diagnosis.

Drug discovery

miRNAs are expected to be potential targets of therapeutic strategies applied to drug discovery for a number of reasons. Firstly, in addition to the initiation and progression of tumor, miRNAs play critical roles in various biological pathways such as differentiation of adipocyte and insulin secretion and diseases such as diabetes and hepatitis. Therefore, the possibility that various human diseases are caused by abnormalities in miRNAs is indicated. Actually, miR-15 and miR-16 have been deleted or decreased in most cases of B-CLL and are identified as tumor suppressor genes (6, 7). Secondly, miRNA expression profiles are correlated with clinical severity of cancer malignancy, and because of this, miRNAs are expected to be powerful tools for cancer diagnosis (10). Thirdly, miRNAs are applicable in gene therapy. The expression of miRNAs can be introduced in vivo by using viral vectors and chemical modifications. Finally, antisense oligonucleotides are potent inhibitors of miRNA, and they can be applied to gene therapy. Actually, it was reported that introduction of 2'-*O*-methoxyethyl phosphorothioate antisense oligonucleotide of miR-122, which is abundant in the liver and regulates cholesterol and fatty-acid metabolism, decreases plasma cholesterol levels and improves liver steatosis in mice with diet-induced obesity (30). These findings indicate that miRNAs and the antisense oligonucleotides are potential targets for drug discovery.

Perspective

It has been established that miRNAs play critical roles in cell differentiation, proliferation, and apoptosis, and the abnormalities of specific miRNA expression contribute to the initiation and progression of tumor. However, identification of target mRNAs negatively regulated by miRNAs remain largely to be explored. Although up to hundreds of target genes toward a single miRNA were predicted by bioinformatics approaches (4), there is no comprehensive assay to biologically validate the prediction algorithm. Therefore, establishment of a method to comprehensively and rapidly identify target mRNAs for the miRNA is necessary for understanding biological and functional mechanisms of miRNA.

References

1 Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116:281–297.

2 Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature. 2000;408:86–89.

3 Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, Cuppen E. Phylogenetic shadowing and computational identification of human microRNA genes. Cell. 2005;120:21–24.

4 Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120:15–20.

5 Calin GA, Sevignani C, Dumitru CD, Hyslop T, Noch E, Yendamuri S, et al. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proc Natl Acad Sci U S A. 2004;101:2999–3004.

6 Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, et al. Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A. 2002;99:15524–15529.

7 Cimmino A, Calin GA, Fabbri M, Iorio MV, Ferracin M, Shimizu M, et al. miR-15 and miR-16 induce apoptosis by targeting BCL2. Proc Natl Acad Sci U S A. 2005;102:13944–13949. Erratum in: Proc Natl Acad Sci U S A. 2006;103:2464–2565.

8 He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, et al. A microRNA polycistron as a potential human oncogene. Nature. 2005;435:828–833.

9 O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT. c-Myc-regulated microRNAs modulate E2F1 expression. Nature. 2005;435:839–843.

10 Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature. 2005;435:834–838.

11 Lee Y, Jeon K, Lee JT, Kim S, Kim VN. MicroRNA maturation: stepwise processing and subcellular localization. EMBO J. 2002;21:4663–4670.

12 Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, et al. MicroRNA genes are transcribed by RNA polymerase II. EMBO J. 2004;23:4051–4060.

13 Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, et al. The nuclear RNase III Drosha initiates microRNA processing. Nature. 2003;425:415–419.

14 Gregory RI, Yan KP, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, et al. The Microprocessor complex mediates the genesis of microRNAs. Nature. 2004;432:235–240.

15 Yi R, Qin Y, Macara IG, Cullen BR. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev. 2003;17:3011–3016.

16 Khvorova A, Reynolds A, Jayasena SD. Functional siRNAs and miRNAs exhibit strand bias. Cell. 2003;115:209–216. Erratum in: Cell. 2003;115:505.

17 Schwarz DS, Hutvágner G, Du T, Xu Z, Aronin N, Zamore PD. Asymmetry in the assembly of the RNAi enzyme complex. Cell. 2003;115:199–208.

18 Meister G, Landthaler M, Patkaniowska A, Dorsett Y, Teng G, Tuschl T. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol Cell. 2004;15:185–197.

19 Hutvágner G, Zamore PD. A microRNA in a multiple-turnover RNAi enzyme complex. Science. 2002;297:2056–2060.

20 Gregory RI, Chendrimada TP, Cooch N, Shiekhattar R. Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. Cell. 2005;123:631–640.

21 Pillai RS, Bhattacharyya SN, Artus CG, Zoller T, Cougot N, Basyuk E, et al. Inhibition of translational initiation by Let-7 MicroRNA in human cells. Science. 2005;309:1573–1576.

22 Liu J, Valencia-Sanchez MA, Hannon GJ, Parker R. MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. Nat Cell Biol. 2005;7:719–723.

23 Andrei MA, Ingelfinger D, Heintzmann R, Achsel T, Rivera-Pomar R, Lu hrmann R. A role for eIF4E and eIF4E-transporter in targeting mRNPs to mammalian processing bodies. RNA. 2005;11:717–727.

24 Liu J, Rivas FV, Wohlschlegel J, Yates JR 3rd, Parker R, Hannon GJ. A role for the P-body component GW182 in microRNA function. Nat Cell Biol. 2005;7:1261–1266.

25 Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature. 2005; 433:769–773.

26 Chen CZ, Li L, Lodish HF, Bartel DP. MicroRNAs modulate hematopoietic lineage differentiation. Science. 2004;303:83–86.

27 Guimaraes-Sternberg C, Meerson A, Shaked I, Soreq H. MicroRNA modulation of megakaryoblast fate involves cholinergic signaling. Leuk Res. 2006;30:583–595.

28 Garzon R, Pichiorri F, Palumbo T, Iuliano R, Cimmino A, Aqeilan R, et al. MicroRNA fingerprints during human megakaryocytopoiesis. Proc Natl Acad Sci U S A. 2006;103: 5078–5083.

29 Fazi F, Rosa A, Fatica A, Gelmetti V, De Marchis ML, Nervi C, et al. A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis. Cell. 2005;123:819–831.

30 Esau C, Davis S, Murray SF, Yu XX, Pandey SK, Pear M, et al. miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting. Cell Metab. 2006;3:87–98.

Nucleic Acids Research, 2006, Vol. 34, Database issue ***D673–D677***
doi:10.1093/nar/gkj028

GLIDA: GPCR-ligand database for chemical genomic drug discovery

Yasushi Okuno*, Jiyoon Yang, Kei Taneishi, Hiroaki Yabuuchi and Gozoh Tsujimoto

Department of Genomic Drug Discovery Science, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida-Shimo-Adachi-cho, Sakyo-ku, Kyoto 606-8501, Japan

Received August 15, 2005; Revised and Accepted September 22, 2005

ABSTRACT

G-protein coupled receptors (GPCRs) represent one of the most important families of drug targets in pharmaceutical development. GPCR-LIgand DAtabase (GLIDA) is a novel public GPCR-related chemical genomic database that is primarily focused on the correlation of information between GPCRs and their ligands. It provides correlation data between GPCRs and their ligands, along with chemical information on the ligands, as well as access information to the various web databases regarding GPCRs. These data are connected with each other in a relational database, allowing users in the field of GPCR-related drug discovery to easily retrieve such information from either biological or chemical starting points. GLIDA includes structure similarity search functions for the GPCRs and for their ligands. Thus, GLIDA can provide correlation maps linking the searched homologous GPCRs (or ligands) with their ligands (or GPCRs). By analyzing the correlation patterns between GPCRs and ligands, we can gain more detailed knowledge about their interactions and improve drug design efforts by focusing on inferred candidates for GPCR-specific drugs. GLIDA is publicly available at http://gdds.pharm.kyoto-u.ac.jp:8081/glida. We hope that it will prove very useful for chemical genomic research and GPCR-related drug discovery.

INTRODUCTION

The superfamily of G-protein coupled receptors (GPCRs) forms the largest class of cell surface receptors. These molecules regulate various cellular functions responsible for physiological responses (1). GPCRs represent one of the most important families of drug targets in pharmaceutical development (2). A large majority of human-derived GPCRs still remain 'orphans' with no identified natural ligands or functions, and thus a key goal of GPCR research related to drug design is to identify new ligands for such orphan GPCRs.

With the unprecedented accumulation of the genomic information, databases and bioinformatics have become essential tools to guide GPCR research. The GPCRDB (http://www.gpcr.org/7tm/) (2) and IUPHAR (http://iuphar-db.org/iuphar-rd/index.html) (3) receptor databases are representatives of widely used public databases covering GPCRs. These databases, which provide substantial data on the GPCR proteins and pharmacological information on receptor proteins containing GPCRs, are mainly focused on biological aspects of the gene products or proteins. In spite of the significance of ligand compounds as drug leads, the relationships between GPCRs and their ligands and/or chemical information on the ligands themselves are not yet fully covered.

On the other hand, there is increasing interest in collecting and applying chemical information in the post-genome era. This new trend is called 'chemical genomics', in which biological information and chemical information are integrated on the genome scale (4,5). PubChem (http://pubchem.ncbi.nlm.nih.gov/) (6), KEGG/LIGAND (http://www.genome.jp/kegg/ligand.html) (7) and ChEBI (http://www.ebi.ac.uk/chebi/) (8) have been developed as databases related to chemical genomics. KEGG/LIGAND and ChEBI contain primarily biochemical information on reported enzymatic reactions. Recently, NIH (the National Institutes of Health) opened PubChem, a public database providing information on the chemical structures of small molecules. However, one cannot retrieve direct information relating these chemical structures to gene or protein entries. Although chemical genomic approaches have thrown new light on relationships between receptor sequences and compounds that interact with particular receptors, the GPCR-ligand information is not well represented in these large-scale databases for chemical genomics.

There are still very few publicly available databases or tools for GPCR-specialized drug discovery from the viewpoint of chemical genomics. Herein, we have developed a novel relational database, GLIDA (GPCR-LIgand DAtabase) (9).

*To whom correspondence should be addressed. Tel: +81 75 753 9264; Fax: +81 75 753 4544; Email: okuno@pharm.kyoto-u.ac.jp