Table 1. Fold changes of mRNA levels by fatty acid treatment in HepG2 cells.

Gene	Accession	Average difference[a]	Fold change				Function
			OA	AA	EPA	DHA	
Interferon-gamma receptor alpha chain	U19247	226	−1.1	−2.3	−2.3	−2.2	antiviral activity
Mitochondrial NADH dehydrogenase	U65579	407	1.8	2.0	3.1	2.7	aspiratory chain
Heparan sulfate proteoglycan (HSPG2)	M85289	146	−1.5	1.3	5.3	1.3	cell adhesion
cdc25Hs	M34065	−26	2.7*	1.9*	1.4*	2.1*	cell differentiation
Interleukin 1 alpha (IL 1)	M28983	−51	2.6*	2.0*	2.7*	1.8*	cell differentiation
MAC30	L19183	1,769	1.0	−2.8	−2.1	−1.8	cell differentiation
Protein tyrosine phosphatase (PTP-PEST)	M93425	73	2.0*	1.9*	−1.0*	3.1*	cell differentiation
Small proline-rich protein 2 (SPRR2B)	L05188	−110	2.9*	2.8*	3.0*	1.8*	cell differentiation
SWI/SNF complex 155 kDa subunit (BAF155)	U66615	197	1.4	1.5*	2.3	2.2*	cell differentiation
Drosophila female sterile homeotic (FSH)	X62083	4	1.5	1.2	13.3*	2.9*	cell proliferation
Glial growth factor 2		394	−5.5*	−3.0	−5.3*	−5.6*	cell proliferation
Membrane-associated protein (HEM-1)	M58285	193	1.8	2.5*	2.9	2.4*	cell proliferation
Sec23A isoform	X97064	51	3.3*	1.1*	2.1*	2.6*	cell proliferation
Sec23B isoform	X97065	230	2.3	1.8	2.4	2.1	cell proliferation
S-lac lectin L-14-II (LGALS2)	M87860	−3	1.5*	2.0*	3.2*	4.5*	cell proliferation
Microsomal glutathione S-transferase (GST-II)	U77604	2,836	1.0	−1.1	1.2	−2.0	detoxification
FDXR gene (adrenodoxin reductase)	M58509	287	1.2	1.5	1.6	2.2	electron transport system
Uncoupling protein homolog (UCPH)	U94592	169	2.8	−2.7	2.2	−2.0	energy consumption
Fatty acid synthase	S80437	4,358	−1.0	−2.1	−2.1	−2.3	fatty acid synthesis
Stearoyl-CoA desaturase		1,416	1.1	−2.9	−2.9	−3.1	fatty acid synthesis
Liver fatty acid binding protein (FABP)	M10050	6,859	1.1	−2.0	−1.6	−1.5	fatty acid transport
Ceruloplasmin (ferroxidase)	M13699	309	1.2	−1.9	−2.9	−3.1	Fe oxidation
Galactokinase (GALK1)	L76927	120	2.5	2.8	−1.9*	3.1	glycogenesis/glycolysis
RASF-A PLA2	M22430	122	2.0	2.2	1.5	2.7	inflamation
S-lac lectin L-14-II (LGALS2)	M87860	−3	1.5*	2.0*	3.2*	4.5*	lectin
Deleted in split hand/split foot 1 (DSS1)	U41515	102	1.4	3.7	4.4	4.8	limb development
Urokinase-type plasminogen activator receptor	U09937	−19	1.6*	1.*	4.9*	2.4*	platelet coagulation
Metallothionein-IG (MT1G)	J03910	195	1.8	3.6	2.3	5.6	protection against heavy metal toxicity
Inter-alpha-trypsin inhibitor subunit 3	X16260	238	−1.1	−4.8*	−4.2*	−2.8	proteinase inhibitor
Vacuolar proton pump, 116-kDa subunit	U45285	41	2.0*	4.4*	4.3*	7.1*	proton pump
Prostasin	L41351	749	−1.2	−2.6	−8.3	−3.8	serine proteinase
Extracellular-superoxide dismutase (SOD3)	J02947	124	1.7	1.3	3.6*	2.2	superoxiside scavenger
Manganese superoxide dismutase (SOD2)	X65965	611	−1.2	−1.2	−2.0	−1.1	superoxiside scavenger
2-Oxoglutarate dehydrogenase	D10523	143	1.4	−1.0	2.0	1.4	TCA cycle
Isocitrate dehydrogenase	Z68129	202	1.5	1.9	2.5	2.0	TCA cycle
Succinate dehydrogenase (SDH)	L21936	496	1.9	1.4	2.1	2.5	TCA cycle
Succinyl-CoA synthetase	Z68204	6	2.1*	1.2*	2.1*	2.1*	TCA cycle
LXR-alpha	U22662	67	1.4*	−1.3*	2.1*	1.3*	transcription factor
NF-kappa-B p65 subunit	L19067	201	2.3	1.7	1.4	1.8	transcription factor
Nuclear factor I-X	L31881	94	1.4	−1.2	4.4	1.4	transcription factor
PPAR alpha	L02932	4	−1.4*	1.2*	1.5*	1.1*	transcription factor
PPAR gamma	L40904	99	2.0	−1.6	−1.1	1.0	transcription factor
Rad2		40	2.6	2.1*	3.5*	2.6*	transcription factor
SREBP-1	U00968	1,105	1.0	−1.7	−1.2	−1.8	transcription factor
SREBP-2	U02031	559	1.1	−1.5	−1.9	−1.7	transcription factor
KIAA0030	D21063	45	2.5*	2.2*	8.1*	1.9*	unknown
KIAA0092	D42054	325	−1.1	−1.3	−1.7	−3.8	unknown
KIAA0219	D86973	96	4.2	1.9	2.7	3.2	unknown
Inducible protein	L47738	−55	3.9*	3.5*	4.5*	3.5*	unknown

HepG2 cells were treated with 0.25 mM of oleic acid (OA), arachidonic acid (AA), eicosapentaenoic acid (EPA), or docosahexaenoic acid (DHA) for 24 h.

[a] Average differences were expressed the intensities of the mRNA levels in control HepG2 cells.

* The value of fold change was calculated using the noise, since the noise of either array was greater than the average differece of the transcript in both the control and the FA-treated groups.

Table 2. Changes of mRNA levels in genes related to cholesterol and lipoprotein metabolism by FA-treatment.

Gene	Accession	Average difference[a]	Fold change			
			OA	AA	EPA	DHA
Repressed						
HMG-CoA reductase	M11058	614	−1.5	−2.9	−2.2	−3.1
HMG-CoA synthase	L25798	226	−1.5	−2.9	−2.4*	−2.0
Mevalonate kinase	M88468	276	−1.2	−1.2	−2.7	−1.1
Mevalonate pyrophosphate decarboxylase	U49260	1,638	−1.3	−3.4	−1.9	−9.5
Squalene epoxidase	D78129	1,782	−1.0	−2.0	−1.2	−2.2
2,3-Oxidosqualene-lanosterol cyclase	U22526	200	−1.0	−2.5	−2.8*	−4.8*
LDL receptor	L00352	1,358	−1.1	−2.6	−2.1	−2.3
Lysosomal acid lipase	U04285	957	−1.1	−1.6	−2.2	−1.6
Induced						
Hepatic triglyceride lipase	M29194	−1	1.7*	1.2*	1.8*	2.6*
Apolipoprotein(a)	X06290	89	2.2	1.3	2.4	−1.3 *
ICAM-2	M32334	5	2.0*	1.3*	3.1*	−1.3*
No change						
Apolipoprotein AI regulatory protein (ARP-1)	M64497	40	1.2*	1.1*	−1.7*	1.1*
Ear-3		75	1.0	1.1*	−1.1*	−1.5*
Lectin-like oxidized LDL receptor	D89050	−21	1.1*	−2.0*	−1.1*	1.1*
Lipoprotein lipase	M15856	52	−1.4*	−1.0*	−1.3*	−1.3*
Scavenger receptor type I	D13264	−13	−1.2*	−1.3*	1.0*	1.2*
CLA-1 (SR-BI)	Z22555	0	0.0*	0.0*	0.0*	0.0*
CD36	Z32765	731	1.1	−1.3	1.3	1.5
HDL binding protein	M64098	942	1.2	1.2	1.2	1.4
CD6 ligand (ALCAM/HB2)	L38608	87	1.0	−1.8*	−1.5*	1.2
Cdc42 GTPase-activating protein	U02570	310	1.3	1.2	1.2	1.4
LCAT	M12625	741	1.0	1.1	−1.2	1.2
ACAT	L21934	−12	1.4*	1.1*	1.2*	1.2*
CETP	M30185	−140	−2.9 *	−1.2 *	1.3*	−1.9 *
Phospholipid transfer protein		245	1.4	−1.2	1.4	−1.0
MTP	X91148	0	0.0*	0.0*	0.0*	0.0*

HepG2 cells were treated with 0.25 mM of oleic acid (OA), arachidonic acid (AA), eicosapentaenoic acid (EPA), or docosahexaenoic acid (DHA) for 24 h.

[a] Average differences were expressed the intensities of the mRNA levels in control HepG2 cells.

* The value of fold change was calculated using the noise, since the noise of either array was greater than the average differece of the transcript in both the control or the FA-treated groups.

may be mediated through SREBPs. PUFA reduce the mRNA expression of SREBPs (*4, 19–22*), which regulate lipogenic gene transcription (SREBP-1) and control cholesterol metabolism (SREBP-2) (*12–14*). Sakakura et al. reported that SREBP regulates the gene expression of all of the enzymes involved in cholesterol synthesis including MPD (*23*). These results indicate that PUFA down-regulates the entire cholesterol synthetic pathway.

Yoshikawa et al. reported that the PUFA suppression of SREBP-1c expression is mediated through competition with liver X factor receptor (LXR) ligand during activation of the ligand-binding domain of LXR (*24*). On the other hand, Tobin et al. (*6*) reported that fatty acids induced the LXR alpha expressions that regulate the fatty acid and cholesterol metabolism. LXR alpha was not changed in our gene chip data. Although we need to analyze LXR further, Cyp7A1 was up-regulated by PUFA using RT-PCR (data not shown).

The PUFA response region is located in the promoter of the stearoyl-CoA desaturase 1 (SCD1) gene (*25*). SREBP may play an important role to regulate the SCD1 because its rate of down-regulation was similar to those of the genes related to cholesterol metabolism. However, Kim et al. (*26*) recently demonstrated that cholesterol overrides the PUFA-mediated repression of the SCD1 gene and regulates SCD1 gene expression through a mechanism independent of SREBP-1 maturation in vivo. The detailed mechanism of the down-regulation of SCD1 caused by PUFA has not been resolved. Furthermore, Matsuzaka et al. (*27*) reported that Δ6-desaturase and Δ5-desaturase expression is dually regulated by SREBP-1c and PPAR-α. At least, PUFA are thought to also autoregulate their biosynthesis through SREBP. In addition, CETP might interact with SREBP-1 (*28*) and is down-regulated by PUFA (*16*). As the expression level of CETP in HepG2 cells is very low, we could not evaluate the effect of PUFA on its expression using the oligonucleotide chip system (Table 1).

On the other hand, the gene expression of enzymes that catabolize FA (*29–32*), namely carnitine: palmitoyl-CoA acyltransferase 1 (CPT1), acyl-CoA oxidase

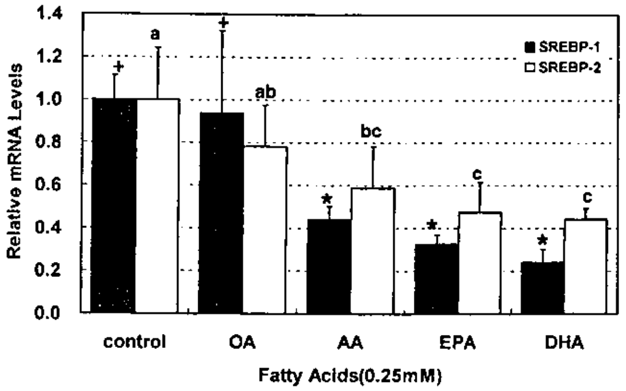


Fig. 1. Effect of PUFA on sterol regulatory element-binding protein (SREBP) mRNA expression in HepG2 cells. HepG2 cells were incubated with PUFA (0.25 mM) for 24 h. Total RNA was extracted, then mRNA expression levels of SREBPs were measured using real time RT-PCR as described in Materials and Methods. Relative mRNA levels were normalized to those of GAPDH. Values are means±SD ($n=3$). Mean values with different superscript letters in SREBP-1 expressions are significantly different ($p<0.05$). Different symbols ($^+$ and *) show significant differences in the SREBP-2 expressions ($p<0.05$).

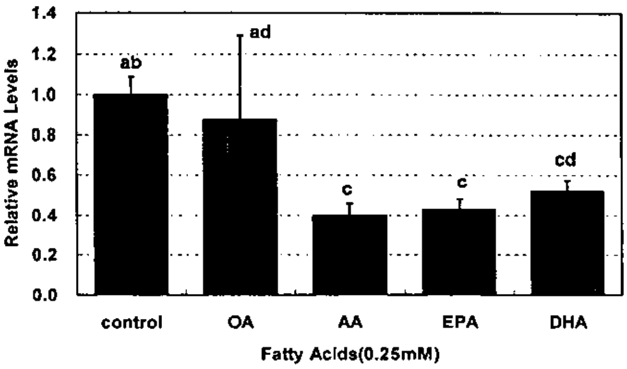


Fig. 2. Effect of PUFA on mevalonate pyrophosphate decarboxylase (MPD) expression in HepG2 cells. HepG2 cells were incubated with FA (0.25 mM) for 24 h. Total RNA was extracted, then mRNA expression of MPD were measured using real time RT-PCR as described in Materials and Methods. Relative mRNA levels were normalized to those of GAPDH. Values are means±SD ($n=3$). Mean values with different letters show significant differences in PUFA treatments ($p<0.05$).

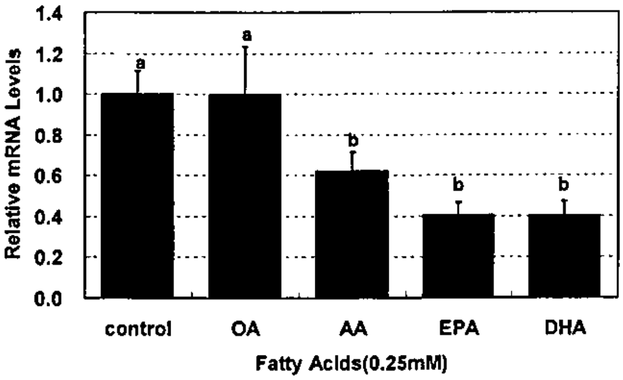


Fig. 3. Effect of PUFA on prostasin expression in HepG2 cells. HepG2 cells were incubated with FA (0.25 mM) for 24 h. Total RNA was extracted, then mRNA expression of prostasin were measured using real time RT-PCR as described in Materials and Methods. Relative mRNA levels were normalized to those of GAPDH. Values are means±SD ($n=3$). Mean values with different letters show significant differences in PUFA treatments ($p<0.05$).

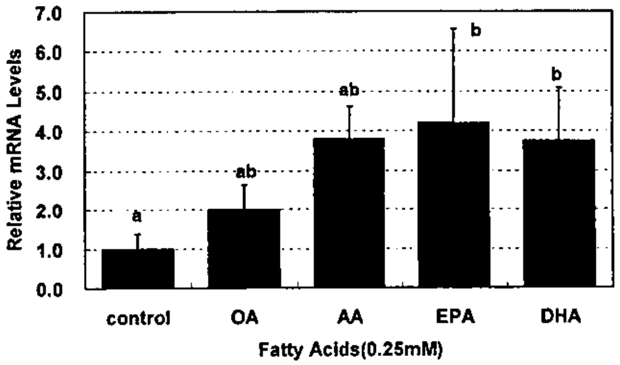


Fig. 4. Effect of PUFA on HTGL expression in HepG2 cells. HepG2 cells were incubated with FA (0.25 mM) for 24 h. Total RNA was extracted, then mRNA expression of HTGL were measured using real time RT-PCR as described in Materials and Methods. Relative mRNA levels were normalized to those of GAPDH. Values are means±SD ($n=3$ for control, OA, EPA and DHA, $n=2$ for AA). Mean values with different letters show significant differences in PUFA treatments ($p<0.05$).

(AOX) and acyl-CoA synthetase (ACS), which are induced by PUFA, did not change in our study. These enzymes are related to fatty acid oxidation and are generally believed to be regulated by PPAR (*5*, *33*, *34*). PPAR-α and -γ are located in the liver and adipocytes, respectively (*5*). The expression level of PPAR-α in human liver (*35*, *36*) is much lower than that in mouse liver, and over-expression of PPAR-α in HepG2 cells shows the induction of mitochondrial HMG-CoA synthase, CPT, and ACS mRNA (*37*). The present study detected only weak expression of PPAR-α and PPAR-γ in HepG2 cells (Table 2). Therefore, the effects might be obvious without regulation mediated by PPAR-α. We also showed that enzymes involved in the TCA cycle were up-regulated. Therefore, PUFA suppressed the synthesis of cholesterol and lipogenesis, but induced ATP generation by activation of the TCA cycle in HepG2 cells.

Takahashi et al. (*38*) have recently examined the effect of dietary fish oil on the gene expression profile in mouse liver using high-density oligonucleotide arrays. Although our findings were similar to theirs, they showed that immune reaction-related genes, antioxidant genes (several glutathione transferase, uncoupling protein 2 and Mn-superoxide dismutase) and genes involved in lipid catabolism were significantly up-regulated, indicating that dietary fish oil down-regulated the endogenous PPAR-α-activation system and increased the antioxidant gene expression that protects against excess ROS. Our data also suggested that PUFA induce antioxidant genes, such as metallothionein-IG and extracellular-superoxide dismutase (SOD3). However, the overall response to oxidation was much less and the expression of microsomal glutathione *S*-transferase and manganese superoxide dismutase (SOD2) were not significantly changed (Table 2). We believe that little oxidative stress was induced by adding PUFA to HepG2 cells even though the PUFA were extremely pure (99%). The induction of immunological and antioxidant genes in their study might have been caused by adaptation to excess ROS production, since they fed the diet containing a very high concentration of fish oil (60% of total energy intake) for 6 mo.

Prostasin is a new serine protease that was purified from seminal fluid, and its cDNA has been sequenced (*39*). Prostasin is expressed in the human prostate, kidney, and lung, as well as in body fluids, including seminal fluid and urine (*40*). The relationship between prostasin and prostate cancer has been investigated (*41–44*). Prostasin might act as an extracellular regulator of epithelial sodium channels (*44*). However its physiological role in humans is not known. Prostasin was significantly suppressed by PUFA in this study and an SRE was located in its upstream region of the gene (*45*), suggesting that prostasin plays an important role in processing some proteins in response to cellular cholesterol concentrations.

PUFA also affected the genes involved in cell proliferation and differentiation. Further analysis using the data obtained by this study is needed in order to clarify the mechanism. PUFA are thought to control gene transcription through several steps. Together with their metabolites, PUFA play important roles in signal transduction cascades, and as ligands for transcription factors. Gene chip analysis might provide useful clues to investigate not only continuous regulation, but also the interaction between many transcription factors, such as SREBP and LXR.

Acknowledgments

This work was supported in part by grants from the Ministry of Health, Labor and Welfare for Aging and Health, the Japan Health Sciences Foundation for Health Sciences focusing on Drug Innovation, the Danone Institute Japan and Grants-in Aid for Scientific Research (No. 14580138) from the Ministry of Education, Science, Culture, Sports and Technology of Japan.

REFERENCES

1) Spector AA, York MA. 1985. Membrane lipid composition and cellular function. *J Lipid Res* **26**: 1015–1035.

2) Nestel PJ. 1990. Effect of *n*-3 fatty acids on lipid metabolism. *Annu Rev Nutr* **10**: 149–167.

3) Brown MS, Goldstein JL. 1997. The SREBP pathway: Regulation of cholesterol metabolism by proteolysis of a membrane-bound transcription factor. *Cell* **89**: 331–340.

4) Worgall TS, Sturley SL, Seo T, Osborne TF, Deckelbaum RJ. 1998. Polyunsaturated fatty acids decrease expression of promoters with sterol regulatory elements by decreasing levels of mature sterol regulatory element-binding protein. *J Biol Chem* **273**: 25537–25540.

5) Desvegne B, Wahli W. 1999. Peroxisome proliferator-activated receptors: Nuclear control of metabolism. *Endocr Rev* **20**: 649–688.

6) Tobin KA, Steinger HH, Alberti S, Spydevold O, Auwerx J, Gustafsson JA, Nebb HI. 2000. Cross-talk between fatty acid and cholesterol metabolism mediated by liver X receptor-alpha. *Mol Endocrinol* **14**: 741–752.

7) Hertz R, Magenheim J, Berman I, Bar-Tana J. 1998. Fatty acyl-CoA thioesters are ligands of hepatic nuclear factor-4 alpha. *Nature* **392**: 512–516.

8) Roche E, Buteau J, Anieto I, Reig JA, Soria B, Prentki M. 1999. Palmitate and oleate induce the immediate-early response genes *c-fos* and *nur-77* in the pancreatic *β*-cell line INS-1. *Diabetes* **47**: 2007–2014.

9) Duplus E, Glorian M, Forest C. 2000. Fatty acid regulation of gene transcription. *J Biol Chem* **275**: 30749–30752.

10) Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 14: 1675–1680.

11) Lee CK, Klopp RG, Weindruch R, Prolla TA. 1999. Gene expression profile of aging and its retardation by caloric restriction. *Science* **285**: 1390–1393.

12) Horton JD, Shimano H, Hamilton RL, Brown MS, Goldstein JL. 1999. Disruption of LDL-receptor gene in transgenic SREBP-1a mice unmasks hyperlipidemia resulting from production of lipid-rich VLDL. *J Clin Invst* **103**: 1067–1076.

13) Shimomura I, Bashmakov Y, Horton JD. 1999. Increased levels of nuclear SREBP-1c associated with fatty livers in two mouse models of diabetes mellitus. *J Biol Chem* **274**: 30028–30032.

14) Horton JD, Shimomura I, Brown MS, Hammer RE, Goldstein JL, Shimano H. 1998. Activation of cholesterol synthesis in preference to fatty acid synthesis in liver and adipose tissue of transgenic mice overproduction sterol regulatory element-binding protein-2. *J Clin Invest* **101**: 2331–2339.

15) Seo T, Oelkers P, Giattina MR, Worgall TS, Sturley SL, Deckelbaum RJ. 2001. Differential modulation of ACAT1 and ACAT2 transcription and activity by long chain free fatty acids in cultured cell. *Biochemistry* **40**: 4756–4762.

16) Hirano R, Igarashi O, Kondo K, Itakura H, Matsumoto A. 2001. Regulation by long-chain fatty acids of the expression of cholesteryl ester transfer protein in HepG2 cells. *Lipids* **36**: 401–406.

17) Takabe W, Mataki C, Wada Y, Ishii M, Izumi A, Aburatani H, Kamakubo T, Niki E, Kodama T, Noguchi N. 2000. Gene expression induced by BO-654, probucol and BHQ in human endothelial cells. *J Atheroscler Thromb* **7**: 223–230.

18) Akiyoshi S, Ishii M, Nemoto N, Kawabata M, Aburatani H, Miyazono K. 2001. Targets of transcriptional regulation by transforming growth factor-beta: expression profile analysis using oligonucleotide arrays. *Jpn J Cancer Res* **92**: 257–268.

19) Xu J, Nakamura MT, Cho HP, Clarks SD. 1999. Sterol regulatory element binding protein-1 expression is suppressed by dietary polyunsaturated fatty acids. *J Biol Chem* **274**: 23577–23583.

20) Kim HJ, Takahashi M, Ezaki O. 1999. Fish oil feeding

decreases mature sterol regulatory element-binding protein-1 (SREBP-1) by down-regulation of SREBP-1c mRNA in mouse liver. *J Biol Chem* **274**: 25892–25898.

21) Yahagi H, Shimano H, Hatsu A, Amemiya-Kudo M, Okazaki H, Tamura Y, Izuka Y, Shinoiri F, Ohashi K, Osuga J, Harada K, Gotoda T, Nagai R, Ishibashi S, Yamada N. 1999. A crucial role of sterol regulatory element-binding protein-1 in the regulation of lipogenic gene expression by polyunsaturated fatty acids. *J Biol Chem* **274**: 35840–35844.

22) Xu J, Teran-Garcia M, Park JH, Nakamura T, Clarke SD. 2001. Polyunsaturated fatty acids suppress hepatic sterol regulatory element-binding protein-1 expression by accelerating transcript decay. *J Biol Chem* **276**: 9800–9807.

23) Sakakura Y, Shimano H, Sone H, Takahashi A, Inoue N, Toyoshima H, Suzuki S, Yamada N, Inoue K. 2001. Sterol regulatory element-binding proteins induce an entire pathway of cholesterol synthesis. *Biochem Biophys Res Commun* **286**: 176–183.

24) Yoshikawa T, Shimano H, Yahagi N, Ide T, Amemiya-Kudo M, Matsuzaka T, Nakakuki M, Tomita S, Okazaki H, Tamura Y, Iizuka Y, Ohashi K, Takahashi A, Sone H, Osuga J, Gotoda T, Ishibashi S, Yamada N. 2002. Polyunsaturated fatty acids suppress sterol regulatory element binding protein 1c promoter activity by inhibition of liver X receptor (LXR) binding to LXR response elements. *J Biol Chem* **277**: 1705–1711.

25) Zhang L, Ge L, Tran T, Stenn K, Prouty SM. 2001. Isolation and characterization of the human stearoyl-CoA desaturase gene promoter: requirement of a conserved CCAAT *cis*-element. *Biochem J* **357**: 183–193.

26) Kim H-J, Miyazaki M, Ntambi JM. 2002. Dietary cholesterol opposes PUFA-mediated repression of the stearoyl-CoA desaturase-1 gene by SREBP-1 independent mechanism. *J Lipid Res* **43**:1750–1757.

27) Matsuzaka T, Shimano H, Yahagi N, Amemiya-Kudo M, Yoshikawa T, Hasty AH, Tamura Y, Osuga J, Okazaki H, Iizuka Y, Takahashi A, Sone H, Gotoda T, Ishibashi S, Yamada N. 2002. Dual regulation of mouse $\Delta 5$- and $\Delta 6$ desaturase gene expression by SREBP-1 and PPARα. *J Lipid Res* **43**: 107–114.

28) Gauthier B, Robb M, Gaudet F, Ginsburg GS, MacPherson R. 1999. Characterization of a cholesterol response element (CRE) in the promoter of the cholesteryl ester transfer protein gene: functional role of the transcription factors SREBP-1a, and-2, and YY1. *J Lipid Res* **40**: 1284–1293.

29) Chatelain F, Kohl C, Esser V, McGarry JD, Girard J, Pegorier J-P. 1996. Cyclic AMP and fatty acids increase carnitine palmitoyltransferase I gene transcription in cultured fetal rat hepatocytes. *Eur J Biochem* **235**: 789–798.

30) Berthou L, Saladin R, Yaqoob P, Branellec D, Calder P, Frunchart J-C, Denefle P, Auwerx J, Staels B. 1995. Regulation of rat liver apolipoprotein A-I, apolipoprotein A-II and acyl-coenzyme A oxidase gene expression by fibrates and dietary fatty acids. *Eur J Biochem* **232**: 179–187.

31) Flatmark T, Niilsson A, Krannes J, Eikhom TS, Fukami NH. 1998. On the mechanism of Induction of the enzyme systems for peroxisomal B-oxidation of fatty acids in rat liver by diets rich in partially hydrogenated fish oil. *Biochim Biophys Acta* **962**: 122–130.

32) Ide T. 2001. Effect of dietary alpha-linolenic acid on the activity and gene expression of hepatic fatty acid oxidation enzymes. *Biofactors* **13**: 9–14.

33) Clarke SD, Jump DB. 1996. Polyunsaturated fatty acid regulation of hepatic gene transcription. *Lipids* **31**: S-7–S-11.

34) Price PT, Nelson CM, Clarke SD. 2000. Omega-3 polyunsaturated fatty acid regulation of gene expression. *Curr Opin Lipidol* **11**: 3–7.

35) Palmer CN, Hsu MH, Griffin KJ, Raucy JL, Johnson EF. 1998. Peroxisome proliferator activated receptor-alpha expression in human liver. *Mol Pharmacol* **53**: 14–22.

36) Gevois P, Torra IP, Chinetti G, Grotzinger T, Dubois G, Fruchart JC, Fruchart-Najib J, Leitersdorf E, Staels B. 1999. A truncated human peroxisome proliferator-activated receptor alpha splice variant with dominant negative activity. *Mol Endocrinol* **13**: 1535–1549.

37) Hsu MH, Savas U, Griffin KJ, Johnson EF. 2001. Identification of peroxisome proliferators-responsive human genes by elevated expression of the peroxisome proliferators-activated receptor α in HepG2 cells. *J Biol Chem* **276**: 27950–27958.

38) Takahashi M, Tsuboyama-Kasaoka N, Nakatani T, Ishii M, Tsutsumi S, Aburatani H, Ezaki O. 2002. Fish oil feeding alters liver gene expressions to defend against PPARα and ROS production. *Am J Physiol Gastrointest Liver Physiol* **282**: G338–G348.

39) Yu JX, Chao L, Chao J. 1995. Molecular cloning, tissue-specific expression, and cellular localization of human prostasin mRNA. *J Biol Chem* **270**: 13483–13489.

40) Laribi A, Berteau P, Gala J, Eschwege P, Benoit G, Tombal B, Schmitt F, Loric S. 2001. Blood-borne RT-PCR assay for prostasin-specific transcripts to identify circulating prostate cells in cancer patients. *Eur Urol* **39**: 65–71.

41) Chen LM, Hodge GB, Guarda LA, Welch JL, Greenberg NM, Chai KX. 2001. Down-regulation of prostasin serine protease: a potential invasion suppressor in prostate cancer. *Prostate* **48**: 93–103.

42) Mok SC, Chao J, Skates S, Wong K, Yiu GK, Muto MG, Berkowitz RS, Cramer DW. 2001. Prostasin, a potential serum marker for ovarian cancer: identification through microarray technology. *J Natl Cancer Inst* **93**: 1458–1464.

43) Donaldson SH, Hirsh A, Li DC, Holloway G, Chao J, Boucher RC, Gabriel SE. 2002. Regulation of the epithelial sodium channel by serine proteases in human airways. *J Biol Chem* **277**: 8338–8345.

44) Narikiyo T, Kitamura K, Adachi M, Miyoshi T, Iwashita K, Shiraishi N, Nonoguchi H, Chen LM, Chai KX, Chao J, Tomita K. 2002. Regulation of prostasin by aldosterone in the kidney. *J Clin Invest* **109**: 401–408.

45) Yu JX, Chao L, Ward DC, Chao J. 1996. Structure and chromosomal localization of the human prostasin (PRSS8) gene. *Genomics* **32**: 334–340.

Physiol Genomics 13: 31–46, 2003.
First published January 7, 2003; 10.1152/physiolgenomics.00116.2002.

Expression imbalance map: a new visualization method for detection of mRNA expression imbalance regions

Makoto Kano,[1] Kunihiro Nishimura,[2] Shumpei Ishikawa,[3] Shuichi Tsutsumi,[3] Koichi Hirota,[4] Michitaka Hirose,[4] and Hiroyuki Aburatani[3]

[1]*School of Engineering and* [2]*School of Information Science and Technology, University of Tokyo, Tokyo 113-8655; and* [3]*Genome Science Division, and* [4]*Intelligent Cooperative System, Department of Information Systems, Research Center for Advanced Science and Technology, University of Tokyo, 153-8904, Japan*

Submitted 4 September 2002; accepted in final form 20 December 2002

Kano, Makoto, Kunihiro Nishimura, Shumpei Ishikawa, Shuichi Tsutsumi, Koichi Hirota, Michitaka Hirose, and Hiroyuki Aburatani. Expression imbalance map: a new visualization method for detection of mRNA expression imbalance regions. *Physiol Genomics* 13: 31–46, 2003. First published January 7, 2003; 10.1152/physiolgenomics.00116.2002.—We describe the development of a new visualization method, called the expression imbalance map (EIM), for detecting mRNA expression imbalance regions, reflecting genomic losses and gains at a much higher resolution than conventional technologies such as comparative genomic hybridization (CGH). Simple spatial mapping of the microarray expression profiles on chromosomal location provides little information about genomic structure, because mRNA expression levels do not completely reflect genomic copy number and some microarray probes would be of low quality. The EIM, which does not employ arbitrary selection of thresholds in conjunction with hypergeometric distribution-based algorithm, has a high tolerance of these complex factors. The EIM could detect regionally underexpressed or overexpressed genes (called, here, an expression imbalance region) in lung cancer specimens from their gene expression data of oligonucleotide microarray. Many known as well as potential loci with frequent genomic losses or gains were detected as expression imbalance regions by the EIM. Therefore, the EIM should provide the user with further insight into genomic structure through mRNA expression.

gene expression profiling; allelic imbalance; chromosome mapping; hypergeometric distribution; computing methodologies

THE RECENT DEVELOPMENT of microarray technology has enabled simultaneous measurement of genome-wide expression profiles. Many research studies have revealed strong correlations between the expression profiles and cancer classifications. The next era of gene expression analysis would involve systematic integration of expression profiles and other types of gene information, such as locus, gene function, and sequence information. In particular, integration between expression profiles and locus information should be effective in detecting gene structural abnormalities such as genomic gains and losses.

In general, cancer progression is not a single but a multistep process and includes many genomic structural abnormalities. Among them, genomic gains and losses, particularly deletion of tumor suppressor genes and amplification of oncogenes, are associated with cancer progression and its malignant phenotype, although the affected lesion varies among different types of cancers. Comparative genomic hybridization (CGH) for detecting genome-wide abnormalities such as copy number changes, has been applied to various types of cancers (5), but its low resolution (~20 Mb, corresponding to about 200 genes) makes it difficult to identify the causal genes, the structural alternation of which is critical for cancer biological behavior.

Integration of gene expression profiles and gene locus information might allow detection of copy number changes at a much higher resolution. Several studies using oligonucleotide probe arrays suggested a strong relationship between genomic structural abnormalities and expression imbalances (underexpression or overexpression). Mukasa et al. (7) reported that the expression levels of a significant number of genes in the 1p region were reduced to about 50%, in oligodendrogliomas with 1pLOH. Furthermore, Virtaneva et al. (12) reported that acute myeloid leukemia with trisomy 8 was associated with overexpression of genes on chromosome 8. Recently, a genome-wide transcriptome map of non-small cell lung carcinomas based on gene expression profiles generated by serial analysis of gene expression (SAGE) was conducted (3). However, the simple spatial mapping of the expression profiles on chromosomal location sometimes hardly provides information about genomic structure for the following reasons: *1*) since some microarray probes are of low quality, the microarray signal intensities do not always reflect their target mRNA expression levels; and *2*) mRNA expression level does not completely reflect genomic copy number. The aim of the present study was to develop a new method with high tolerance of such complex factors, designed to detect regionally underexpressed or overexpressed genes in cancer specimens compared with the corresponding normal tissues. The expression imbalance region, constituted by

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: M. Kano, Tokyo Research Laboratory, IBM Japan, 1623-14 Shimotsuruma, Yamato-shi, Kanawaga 242-8502, Japan (E-mail: mkano@jp.ibm.com).

these genes, likely reflects genomic structural changes such as chromosomal gain and loss.

When developing the methodology that integrates the expression profiles and locus information, two significant problems have to be dealt with. First, a definition of what constitutes an expression imbalance region is not yet clarified. How many base pairs on chromosome should be considered as a genomic region (referred to below as chromosomal proximity)? To consider that a certain gene is differentially expressed in cancer and normal tissue, how much difference in the gene expression level is needed between the two (referred to below as cancer specificity)? It is generally very difficult to determine adequate thresholds for chromosomal proximity and cancer specificity. Arbitrary selection of thresholds would involve a risk of overlooking significant genes (that is, "threshold problem"). In addition, to detect expression imbalance regions, it is necessary to search for genes with both cancer specificity and chromosomal proximity. Because determining these two thresholds synergistically increases the risk of overlooking significant genes, the "threshold problem" is more critical in this case.

When selecting thresholds, several statistical theories such as hypothesis testing are helpful. However, commonly used statistical criteria are also arbitrarily determined. If thresholds are automatically determined based on statistical theory, the user cannot search more genes with potential significance, because the information of genes overlooked is almost unknown. Therefore, to detect as many significant genes as possible, a comprehensive presentation of the distribution of the "false balance" (that is, the balance of false negative and false positive) is quite significant rather than an attempt to seek potentially optimal statistical criterion.

Second, there are many candidate expression imbalance regions. Some of them may be a family of genes that are tandemly repeated and are under similar transcriptional regulations. To confirm that a candidate locus is biologically significant, human curation is necessary, using a variety of biological information. Therefore, it is important to present large genome-wide data in a comprehensive manner, indicating which genes are to be further examined. That is, a broadband interface between humans and computers is essential.

We focused on visualization technology as the key technology to solve these two problems. Visualization is effective in providing, genome-wide, the false-balance distribution and indication of the genes that are worth examining. The visualization used in our report would make it possible to present the images of all genes that have both cancer specificity and chromosomal proximity.

In this study, we developed a novel visualization method for detecting expression imbalance regions at much higher resolution than conventional technologies such as CGH, called the expression imbalance map (EIM). The EIM was applied to gene expression data of lung squamous cell carcinoma measured by oligonucleotide microarray and detected many known as well as potential loci with frequent genomic losses or gains as regional signal images on chromosomes (expression imbalance regions). In addition, the EIM could detect not only the expression imbalance common to all cancer specimens, but also individual differences among cancer specimens.

MATERIAL AND METHODS

Data Sets

In this article, the EIM is illustrated using the gene expression data of lung cancer from the study of Bhattacharjee et al. (1). In this experiment, total mRNA was extracted from histologically defined specimens of squamous cell lung carcinomas (abbreviated here as "SQ"; $n = 21$) and normal lung tissues (abbreviated here as "NL"; $n = 17$). The expression profiles were obtained using human U95A oligonucleotide probe arrays (GeneChip; Affymetrix, Santa Clara, CA). The SQ-NL gene expression data set (SQ, $n = 21$; NL, $n = 17$) was then analyzed using the EIM.

Feature Selection and Logarithmic Transformation

To compensate for distortion in the expression level, changes in the expression level were limited from 1 to 8,000. In addition, 4,083 probes with a mean expression above 50 and CV (CV = mean/standard deviation) above 0.2 were selected to eliminate potential low-quality probes. The common logarithm of the gene expression data was used for the following analysis.

Translation from Probe to UniGene

To associate gene locus information with gene expression profiles, each "probeID" on the U95A array was translated to UniGene, using information on the UniGene web site of the National Center for Biotechnology Information (NCBI), by referring to the corresponding original GenBank accession number of each probe set. Then, 11,334 of 12,533 probes on the U95A array were translated into 8,851 UniGenes.

Gene Locus Information

Gene locus information was obtained from the web sites for Genes On Sequence Map (*Homo sapiens* build 27) of NCBI and is defined as "LocusID." Among the LocusIDs on chromosome 1 to 22 of Genes On Sequence Map, the 12,063 LocusIDs, which had the corresponding UniGenes, were utilized to identify the chromosome locations of genes. Since the gene expression data utilized in this study were obtained from both sexes, the X and Y chromosomes were excluded. However, by using the data obtained from only males or females, the EIM can be applied to the analysis of chromosome X and Y. Since the 12,063 LocusIDs had one-to-one correspondence with UniGenes, they were translated into 12,063 UniGenes. However, only 6,652 of the 12,063 UniGenes were in common with the 8,851 UniGenes translated from the probes on the U95A array (Fig. 1). In this article, these 6,652 UniGenes are called "Key-UniGenes." The distributions of the UniGenes and Key-UniGenes on each arm of the chromosome are shown in Table 1. The number of total Key-UniGenes was defined as U (=6,652).

Quantization of Each Chromosome Arm Region

For easier handling of the gene locus information, each chromosome arm region was quantized by unit region called

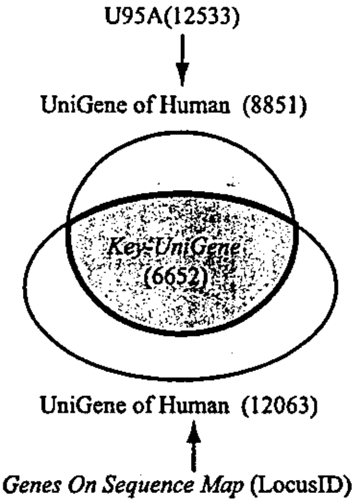


Fig. 1. Correspondence between probeIDs and LocusIDs. To associate gene locus information with gene expression profiles, probeIDs on the Affymetrix U95A oligonucleotide arrays and the LocusIDs on Genes On Sequence Map (*Homo sapiens* build 27) of NCBI were translated into UniGenes. We utilized the 12,063 LocusIDs, which had the corresponding UniGenes, on chromosome 1 to 22 of Genes On Sequence Map. The X and Y chromosomes were excluded, because the gene expression data utilized in this study were obtained from both sexes. Since these 12,063 LocusIDs had one-to-one correspondence with UniGenes, these were translated into 12,063 UniGenes. Out of 12,533 probes on the U95A array, 11,334 were translated into unduplicated 8,851 UniGenes, by referring to the corresponding original GenBank accession number of each probe set. Although the 12,063 UniGenes were obtained from Genes On Sequence Map, only 6,652 of the 12,063 UniGenes were in common with the 8,851 UniGenes translated from the probes on the U95A array. In this article, these 6,652 UniGenes are called "Key-UniGenes."

"bucket" whose length was 100,000 base pairs (100 kbp), and the Key-UniGenes were assigned the corresponding buckets according to their reading position (Fig. 2, *A* and *B*). A reading position indicates the start position for gene transcription and was obtained from Genes On Sequence Map. The number of buckets on chromosome arm *arm* was defined as L_{arm}.

Formation of Locus Cluster

To evaluate the proximity of genes on chromosome arm *arm*, the Key-UniGenes on the *length* neighbor buckets from (*begin*)-th were defined as a cluster $C_{arm\_length\_begin}$ (Fig. 2*A*). Repeating the sufficiently minute changes of *length* and *begin* formed the exhaustive uncertainty cluster sets of Key-UniGenes with chromosomal proximity (Fig. 2*C*). The EIM allows even clusters that overlap each other or include others. Therefore, all neighbor buckets in any area of each chromosome arm were defined as clusters. The number of Key-UniGenes in the cluster $C_{arm\_length\_begin}$ was defined as $n_{arm\_length\_begin}$. $C_{arm\_length\_begin}$ was defined for all

$$arm = 1p, 1q, 2p, 2q, \ldots, 22p, 22q$$

$$length = 2, 3, 4, \ldots \ [buckets]$$

$$begin = 1, 2, \ldots, (L_{arm} - length + 1)$$

In addition, to avoid considering a region that contains large gaps between genes as "one region," the gaps between the Key-UniGenes that lie next to each other in $C_{arm\_length\_begin}$ were calculated and the maximal gap was defined as $gap_{arm\_length\_begin}$ (Fig. 2*B*). The EIM allows the user to filter out the cluster(s) whose $gap_{arm\_length\_begin}$ is more than gap_{max}, which can be changed interactively. In other words, the user can exclude regions containing large gaps by controlling gap_{max}. When gap_{max} values were 500 kbp, 1 Mbp, 2 Mbp, and 3 Mbp, the percentages of the gaps that were less than gap_{max} were 77.6, 89.4, 96.0, and 98.2%, among all gaps between the Key-UniGenes that lie next to each other.

EIM for Detection of Expression Imbalance Specific To Squamous Cell Carcinomas

Clusters consisting of genes with expression profiles specific to SQs. Probes with expression profiles specific to SQs were extracted as a cluster from 4,083 probes of SQ-NL data sets. Although the EIM does not depend on the type of statistical method used for evaluating the difference between two groups, nonparametric tests such as the Mann-Whitney test have the advantage that no assumption is needed about the distribution of data, compared with parametric tests such as the *t*-test. Thus we explain the case of the Mann-Whitney test as an example.

More specifically, the difference in the level of expression of each gene between two groups (SQs and NLs) was defined using the statistical probability, *P*, of rank sum. Assume that there are two groups (G_a, $n = N_a$; G_b, $n = N_b$) and the rank sums in G_a and G_b are Sum_a and Sum_b, respectively, when all elements ($N_a + N_b$) are sorted in order. For simplicity, assume that Sum_a/N_a is greater than or equal to Sum_b/N_b. *P* is the probability of observing the rank sum of the N_a elements, which are randomly selected from all elements, to be more than Sum_a.

Table 1. *Number of the UniGenes and Key-UniGenes on Genes On Sequence Map*

Chr. Arm	UniGene Number	Key-UniGene Number (L_{arm})	Chr. Arm	UniGene Number	Key-UniGene Number (L_{arm})
1p	715	394	12p	211	107
1q	614	361	12q	488	289
2p	313	179	13p	0	0
2q	485	274	13q	218	127
3p	315	191	14p	0	0
3q	335	171	14q	411	228
4p	111	60	15p	0	0
4q	356	201	15q	379	197
5p	116	61	16p	254	130
5q	472	248	16q	244	123
6p	434	251	17p	218	130
6q	291	158	17q	513	290
7p	180	105	18p	52	34
7q	373	205	18q	135	76
8p	157	95	19p	391	199
8q	262	138	19q	481	249
9p	146	85	20p	122	53
9q	353	193	20q	245	124
10p	104	53	21p	0	0
10q	362	205	21q	137	83
11p	234	129	22p	0	0
11q	502	280	22q	334	176

Distributions of the UniGenes, which were obtained from Genes On Sequence Map (*Homo sapiens* build 27) of NCBI, and Key-UniGenes on each arm of the chromosome. Since the gene expression data utilized in this study were obtained from both sexes, the X and Y chromosomes were excluded. Key-UniGenes are the UniGenes that can be translated into from both the probes on the U95A oligonucleotide arrays and the LocusIDs on chromosome 1 to 22 of the Genes On Sequence Map. The total numbers of the UniGenes and Key-UniGenes are 12,063 and 6,652, respectively. Chr., chromosome; L_{arm}, number of "buckets" on chromosome arm *arm*.

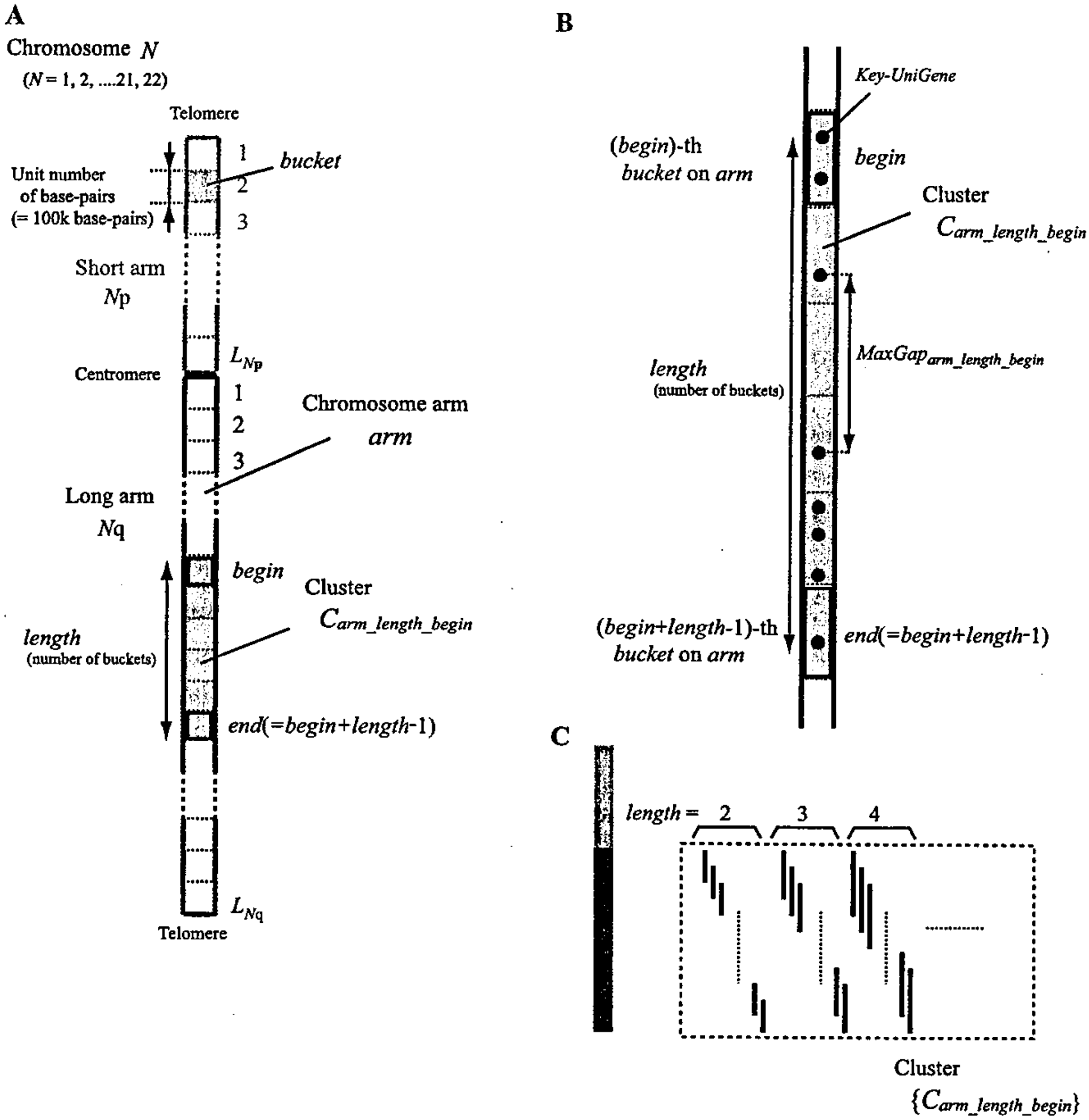


Fig. 2. Formation of clusters of genes with chromosomal proximity. *A*: for easier handling of the gene locus information, each chromosome arm region was quantized by unit region called "bucket" whose length was 100 kbp, and the Key-UniGenes were assigned the corresponding buckets according to their reading positions, which were obtained from Genes On Sequence Map (*Homo sapiens* build 27) of NCBI. The number of buckets on chromosome arm *arm* was defined as L_{arm}. To evaluate the proximity of genes on chromosome arm *arm*, the Key-UniGenes on the *length* neighbor buckets from (*begin*)-th were defined as a cluster $C_{arm\_length\_begin}$. *B*: to avoid considering a region containing large gaps between genes as "one region," the gaps between Key-UniGenes which lie next to each other in $C_{arm\_length\_begin}$ were calculated and the maximal gap was defined as $gap_{arm\_length\_begin}$. The expression imbalance map (EIM) allows the user to filter out the clusters whose $gap_{arm\_length\_begin}$ is more than gap_{max}, which can be changed interactively. In other words, the user can exclude regions containing large gaps by controlling gap_{max}. *C*: repeating the sufficiently minute changes of *length* and *begin* formed the exhaustive uncertainty cluster set of locus information. The EIM allows even the clusters that overlap each other or include others. Therefore, all neighbor buckets in any area of each chromosome arm were defined as clusters.

$$H(U,n_1,n_2,k) = 1 - \sum_{i=0}^{k-1} \frac{\binom{n_2}{i} \cdot \binom{U - n_2}{n_1 - i}}{\binom{U}{n_1}} \qquad (2)$$

When the H value is small, the overlap between C_{sign_diff} and $C_{arm\_length\_begin}$ is considered statistically significant. That is, if the H value is small, then the overlap did not occur accidentally. Thus the evaluation value, E, is defined as follows

$$E(U,n_1,n_2,k) = -\log_{10}H(U,n_1,n_2,k) \qquad (3)$$

For any combination of C_{sign_diff} and $C_{arm\_length\_begin}$, if both (*begin*)-th and (*begin* + *length* − 1)-th buckets of $C_{arm\_length\_begin}$ have the Key-UniGenes that are included in C_{sign_diff}, then their E values were calculated. This calculation was preprocessing for the EIM. Then, in real-time processing, if both C_{sign_diff} and $C_{arm\_length\_begin}$ met $d_{\min}$ and $gap_{\max}$, respectively, then the E value was represented in the intersection area $R_{sign\_diff\_arm\_length\_begin}$ as a gray scale. The user can control $d_{\min}$ and $gap_{\max}$ interactively. The area where the multiple $R_{sign\_diff\_arm\_length\_begin}$ values overlapped is overwritten at the maximum E value (Fig. 4*B*). A flowchart that details these steps is shown in Fig. 5. The EIM for detecting expression imbalance specific to SQs is shown in Fig. 6. In addition, Fig. 7 shows chromosome 3 of the EIM and the influence of $gap_{\max}$ and $d_{\min}$ on the detection of the expression imbalance regions specific to SQs.

EIM for Detection of Individual Differences in Expression Imbalance Among SQs

It is effective to extract probes with expression profiles specific to the group of cancers using statistical analyses, such as the Mann-Whitney analysis. However, because this type of analysis treats all specimens with the same pathological diagnosis as one group, the variation in a group is unobservable. This is sometimes a significant problem because cancer specimens generally have a great number of variations. Thus we also developed the EIM for detecting individual differences in expression imbalance among SQs.

Clusters of probes with expression imbalance in each SQ. The first step in the development of the EIM for detecting individual differences in expression imbalance among SQ specimens was to extract probes with under- or overexpression compared with NL specimens, in each SQ specimen independently. Assuming that the expression levels of a certain probe, g, in NL specimens have a lognormal distribution, if the expression level of a SQ specimen, S_i, is included in 100p% of sections on both sides of NL's distributions, its differential level D_2 was defined as follows

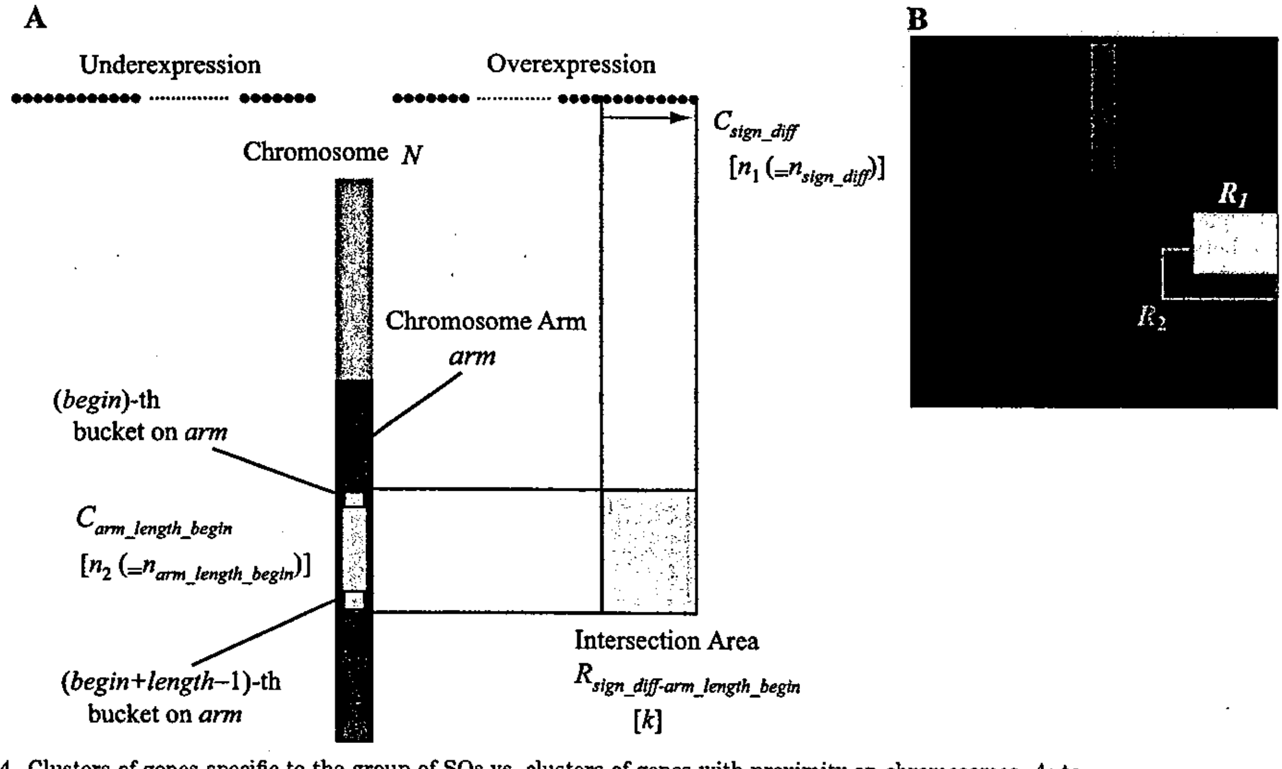


Fig. 4. Clusters of genes specific to the group of SQs vs. clusters of genes with proximity on chromosomes. *A*: to detect expression imbalance regions, it is necessary to search for genes with both cancer specificity and chromosomal proximity. The fundamental algorithm of the EIM is to evaluate statistically the overlaps between clusters of genes with cancer specificity and clusters of genes with chromosomal proximity. The clusters of probes with expression specific to the group of SQ, C_{sign_diff}, are arranged on the abscissa, and those of Key-UniGenes with proximity on chromosomes, $C_{arm\_length\_begin}$, on the ordinate. Among C_{sign_diff} values, the clusters of probes with underexpression and overexpression in SQs are arranged on the *left* and *right* side, respectively. The n_{sign_diff} and $n_{arm\_length\_begin}$ are the numbers of Key-UniGenes in C_{sign_diff} and $C_{arm\_length\_begin}$, respectively; k is the number of common Key-UniGenes both in C_{sign_diff} and $C_{arm\_length\_begin}$. The statistical significance of the overlap between C_{sign_diff} and $C_{arm\_length\_begin}$ was visualized in the intersection area $R_{sign\_diff\_arm\_length\_begin}$ as a gray scale. *B*: the area where the multiple $R_{sign\_diff\_arm\_length\_begin}$ overlapped was overwritten at the maximum E value. Therefore, when the E value of R_1 is higher than that of R_2, the area where R_1 and R_2 overlapped is overwritten at that of R_1.

Based on this P value, the differential level $D_1(g)$ in which g is the probe name was defined as follows

$$D_1(g) = -\log_{10}P \qquad (1)$$

Probes whose differential level D_1 was equal to or more than *diff* were defined as a cluster of probes with expression profiles specific to SQs, C_{sign_diff} (Fig. 3). The suffix *sign* indicates a differential direction (+, overexpression; −, underexpression in SQs). Repeating the sufficiently minute changes of *diff* formed the exhaustive uncertainty set of the clusters specific to SQs. C_{sign_diff} was defined for all

$$sign = -, +$$

$$diff = 2, 3, 4, \ldots$$

For example, C_{+3} was a cluster of probes whose differential level $D_1(g)$ of overexpression was 3 or more. The EIM was constructed by all the clusters C_{sign_diff} with *diff* greater than or equal to the minimum acceptable differential level $d_{\min}$ (Fig. 3). Since the default value of $d_{\min}$ is 2, all the clusters, C_{sign_diff}, would be utilized. The EIM allows the user to control $d_{\min}$ interactively for narrowing down the probes, if needed.

The numbers of probes, UniGenes, and Key-UniGenes of each cluster are shown in Table 2; n_{sign_diff} is the number of Key-UniGenes translated from probes of C_{sign_diff}. When multiple probes in a cluster could be mapped to a single UniGene, only the probe with the highest D_1 value was adopted. In addition, Fig. 3 shows probe permutations whose differential levels are 2 or more, arranged in the order of the differential level. Probes with under- and overexpression are arranged on the left and the right of Fig. 3, respectively.

Table 2. *Clusters of probes with expression profiles specific to the group of squamous cell lung carcinomas*

Differential Direction	Cluster Name (C_{sign_diff})	Probe Number	Key-UniGene Number (n_{sign_diff})
Underexpression (SQ < NL)			
	C_{-2}	1,007	668
	C_{-3}	844	567
	C_{-4}	642	429
	C_{-5}	448	301
	C_{-6}	283	188
	C_{-7}	83	61
Overexpression (SQ > NL)			
	C_{+2}	958	613
	C_{+3}	759	480
	C_{+4}	543	329
	C_{+5}	334	205
	C_{+6}	143	95
	C_{+7}	13	8

The probes (on the Affymetrix U95A arrays) whose expression profiles show significant difference between squamous cell lung carcinomas (SQs) and normal lung (NLs) were extracted as clusters, C_{sign_diff}. The suffix *sign* indicates the differential direction ("+" = overexpression; "−" = underexpression in SQs), and *diff* indicates a differential level D_1 in gene expression profiles between SQs and NLs. For example, C_{+3} is a cluster of probes whose differential level of overexpression is 3 or more. Repeating the sufficiently minute changes of *diff* formed the exhaustive set of the clusters consisting of genes with expression profiles specific to SQs. The numbers of probes and Key-UniGenes for each cluster are shown.

Construction of the EIM. To detect the expression imbalance regions, it is necessary to search for genes with both cancer specificity and chromosomal proximity. The fundamental algorithm of the EIM is to statistically evaluate the overlaps between clusters of genes with cancer specificity and clusters of genes with chromosomal proximity. The clusters specific to the group of SQs, C_{sign_diff}, are arranged on the abscissa, and the locus clusters, $C_{arm\_length\_begin}$, are on the ordinate, as shown in Fig. 4. The variable k is the number of common Key-UniGenes between C_{sign_diff} and $C_{arm\_length\_begin}$.

The variable k could be evaluated using the hypergeometric probability, H, for observing at least k common elements between randomly selected n_1 and n_2 elements among all U elements as follows, where n_1 is n_{sign_diff} and n_2 is $n_{arm\_length\_begin}$.

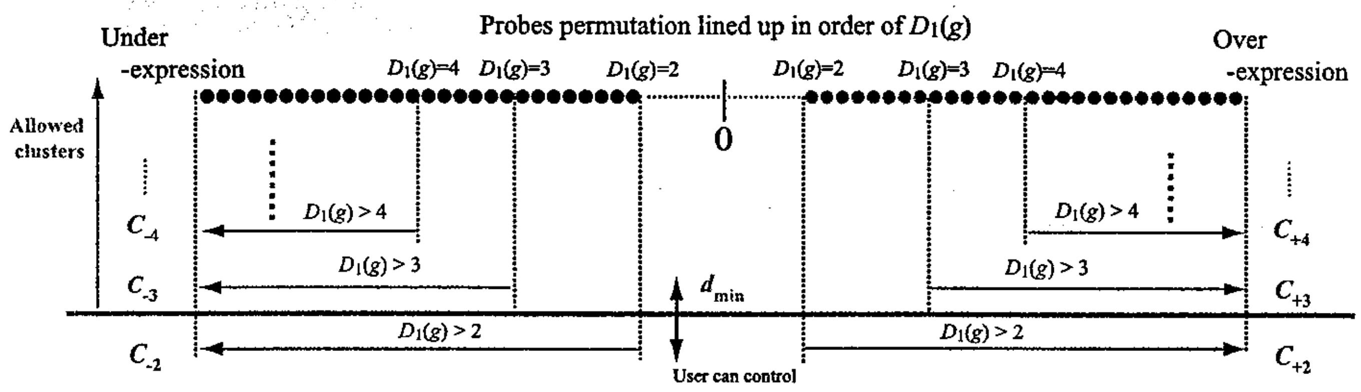


Fig. 3. Probe permutation arranged in order of the difference in gene expression level between squamous cell lung carcinomas (SQs) and normal lungs (NLs). Probes on the U95A arrays are lined up in order of the $D_1(g)$ level, which represents the difference in the gene expression level between SQs and NLs. Only probes with differential levels of 2 or more were arranged. Probes with underexpression and overexpression in SQs are arranged on the *left* and *right* side, respectively. Probes whose differential level $D_1(g)$ is equal to or more than *diff*, are defined as a cluster of probes with expression profiles specific to SQs, C_{sign_diff}. The suffix *sign* indicates the differential direction (+, overexpression; −, underexpression in SQs). Repeating the sufficiently minute changes of *diff* formed the exhaustive uncertainty set of the clusters specific to SQs. The EIM was constructed by all clusters C_{sign_diff} with *diff* that were greater than or equal to the minimum acceptable differential level $d_{\min}$. Since the default value of $d_{\min}$ is 2, all the clusters, C_{sign_diff}, would be utilized. The EIM allows the user to control $d_{\min}$ interactively for narrowing down the probes, if needed.

$$D_2(g,S_i) = -\log_{10}p \qquad (4)$$

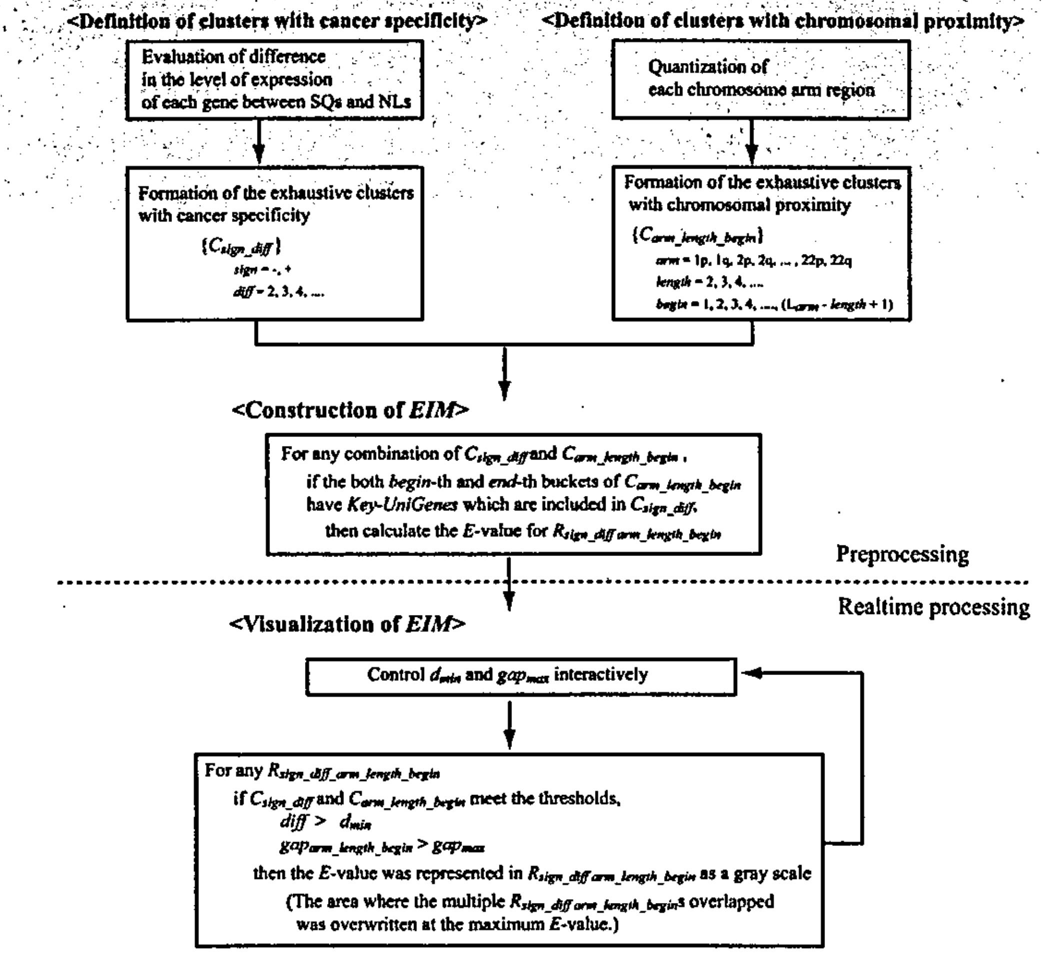
Regarding each SQ specimen S_i (i = 1, 2,. . ., 21), the probes whose differential levels $D_2(g,S_i)$ were equal to or more than *diff* were defined as the individual-specimen cluster, $C_{sign\_diff\_Si}$, where *sign* is the differential direction (+, overexpression; −, underexpression in each SQ specimen). $C_{sign\_diff\_Si}$ was defined for all

$$sign = -, +$$
$$diff = 2, 3, 4, \ldots$$
$$S_i = 1, 2, \ldots, 21$$

For example, C_{+2_Si} and C_{-2_Si} were clusters of probes whose expression of S_i were included in 1% of sections on both sides of NL's distributions. More specifically, C_{+2_Si} was a cluster of probes whose expression levels were equal to or higher than (ave_{NL} + 2.58 $stddev_{NL}$) in a specimen S_i, where ave_{NL} is the mean and $stddev_{NL}$ is the standard deviation of expression level in NL specimens. In the same manner, C_{-2_Si} was a cluster of probes whose expression levels were equal to or less than (ave_{NL} − 2.58 $stddev_{NL}$); $n_{sign\_diff\_Si}$ is the number of Key-UniGenes in $C_{sign\_diff\_Si}$. If multiple probes in a cluster could be mapped to single UniGene, then only the probe with the highest D_2 value was adopted. The average numbers, $\bar{n}_{sign\_diff}$, of $\{n_{sign\_diff\_Si}\}$(i = 1, 2,. . ., 21) are shown in Table 3.

Construction of the EIM. In a manner similar to the EIM for detecting expression imbalance of SQ group, that for detecting individual differences in expression imbalance among SQs was also constructed. The individual-specimen clusters, $C_{sign\_diff\_Si}$, were arranged on the abscissa with respect to each S_i, and the locus clusters on the ordinate (Fig. 8). Underexpression clusters were arranged on the left side and overexpression clusters on the right. Since the abscissa represented an array of S_i, it was impossible to represent *diff* on the abscissa like Fig. 4. Therefore, the EIM for individual specimen was visualized by $C_{sign\_diff\_Si}$ with a defined *diff*, and allowed the user to change *diff* interactively.

The number of common Key-UniGenes between $C_{sign\_diff\_Si}$ and $C_{arm\_length\_begin}$, k, could also be evaluated using $E(U, n_1, n_2, k)$ (*Eq. 3*), where n_1 was $\bar{n}_{sign\_diff}$ and n_2 was $n_{arm\_length\_begin}$. If the different specimens have the same number of genes with under- or overexpression on the same local region, then it is necessary to evaluate them as similar. Therefore, $\bar{n}_{sign\_diff}$ instead of $n_{sign\_diff\_Si}$ was used for the evaluation of the overlap between $C_{sign\_diff\_Si}$ and $C_{arm\_length\_begin}$. The E value for any combination of $C_{sign\_diff\_Si}$ and $C_{arm\_length\_begin}$ was calculated,



<Definition of clusters with cancer specificity>
Evaluation of difference in the level of expression of each gene between SQs and NLs
Formation of the exhaustive clusters with cancer specificity
$\{C_{sign\_diff}\}$
$sign$ = -, +
$diff$ = 2, 3, 4, ...
<Definition of clusters with chromosomal proximity>
Quantization of each chromosome arm region
Formation of the exhaustive clusters with chromosomal proximity
$\{C_{arm\_length\_begin}\}$
arm = 1p, 1q, 2p, 2q, ... , 22p, 22q
$length$ = 2, 3, 4, ...
$begin$ = 1, 2, 3, 4, ..., (L_{arm} - $length$ + 1)
<Construction of EIM>
For any combination of C_{sign_diff} and $C_{arm\_length\_begin}$,
if the both begin-th and end-th buckets of $C_{arm\_length\_begin}$
have Key-UniGenes which are included in C_{sign_diff},
then calculate the E-value for $R_{sign\_diff\,arm\_length\_begin}$
Preprocessing
Realtime processing
<Visualization of EIM>
Control d_{min} and gap_{max} interactively
For any $R_{sign\_diff\_arm\_length\_begin}$
if C_{sign_diff} and $C_{arm\_length\_begin}$ meet the thresholds,
$diff > d_{min}$
$gap_{arm\_length\_begin} > gap_{max}$
then the E-value was represented in $R_{sign\_diff\,arm\_length\_begin}$ as a gray scale
(The area where the multiple $R_{sign_diff\,arm_length_begin}$s overlapped
was overwritten at the maximum E-value.)

Fig. 5. Flowchart for construction of the EIM for detecting expression imbalance regions specific to SQs. This flowchart provides details of the steps of the EIM for detecting expression imbalance regions specific to SQs. For the steps of "Definition of clusters with cancer specificity," please refer to Fig. 3. For the steps of "Definition of clusters with chromosomal proximity," please refer to Fig. 2. For the steps of "Construction of the EIM" and "Visualization of EIM," please refer to Fig. 4. The user can interactively control the steps in real-time processing by changing gap_{max} and d_{min}.

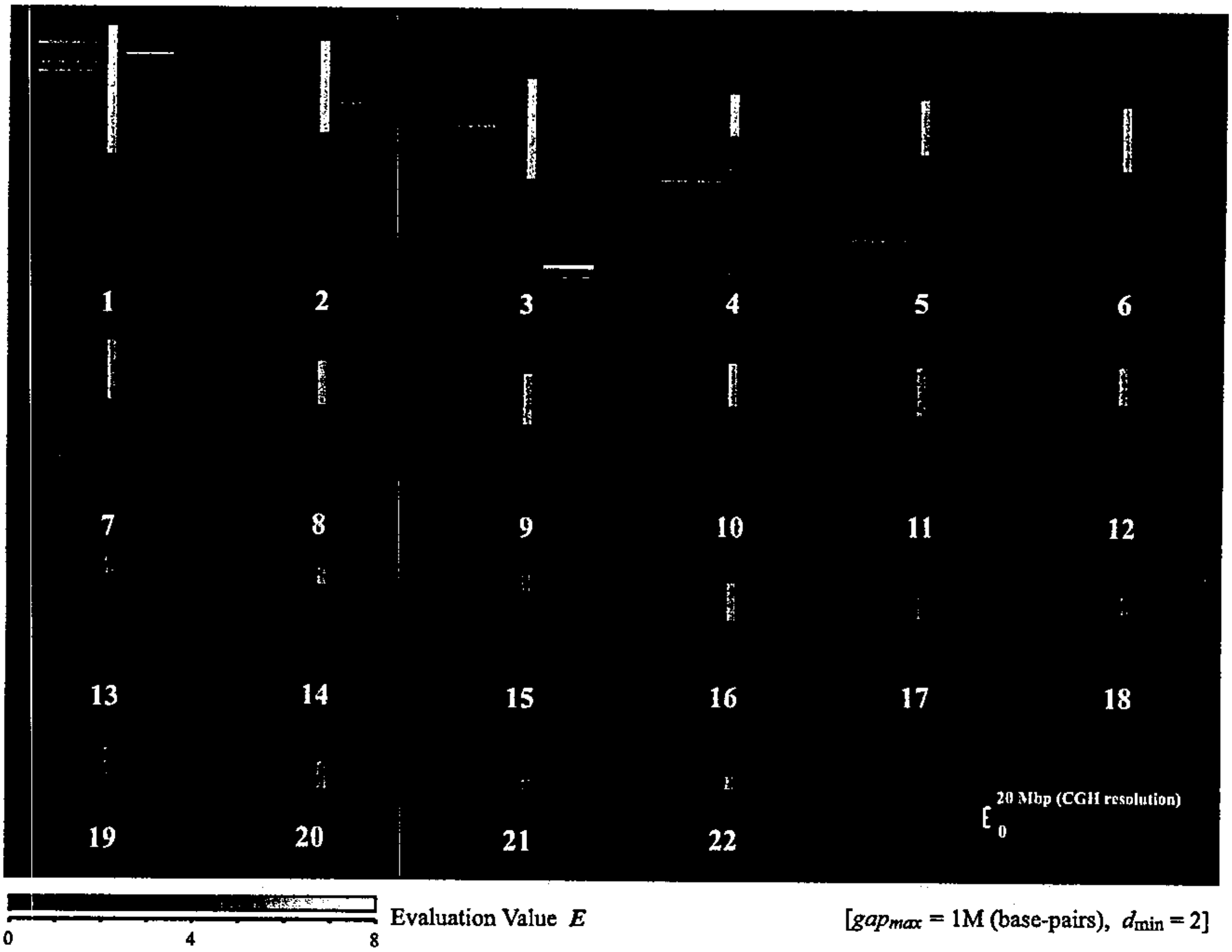


Fig. 6. The EIM applied for detecting expression imbalance regions specific to SQs. The regions of under- and overexpression in SQs were visualized on the *left* and *right* side, respectively, as gray regional signals. All statistical evaluation values of any combinations between the exhaustive uncertainty cluster sets of cancer specificity and chromosomal proximity are visualized on the EIM as the gradation of gray scale simultaneously. Each exhaustive uncertainty cluster set was formed by repetition of the sufficiently minute changes of the threshold of cancer specificity or chromosomal proximity. While the area with high luminance corresponds to the more probable expression imbalance region, the EIM enables the user to search as many genes as possible by referring to more expanded area with lower luminance. The EIM presented the most significant overexpression regions on 3q (the evaluation value E = 7.2), which is a well-known locus with frequent genomic gains, as detected by comparative genomic hybridization (CGH) (6, 8, 9). Note the high resolution of the EIM compared with CGH resolution (~20 Mbp).

Fig. 7. Expression imbalance regions specific to SQs on chromosome 3. *A–I*: chromosome 3 of the EIM and the influence of gap_{max} and d_{min} on the detection of the expression imbalance regions specific to SQs. The EIM represents the E values whose C_{sign_diff} and $C_{arm\_length\_begin}$ meet d_{min} and gap_{max}, respectively. The EIM allows the user to control gap_{max} and d_{min} interactively. The user can narrow down the possible expression imbalance regions by changing gap_{max} and d_{min}. Especially, as is shown in *A–I*, changing gap_{max}, which allows exclusion of regions containing large gaps between genes, markedly affected the detection of expression imbalance regions. *J*: the macrograph of the encircled *region A* from panel *A*. Intersection area $R_{+5\_3q\_1894\_5}$ shows the most significant overexpression region, which is a well-known locus with frequent genomic gains as previously detected by CGH (6, 8, 9). That is, the overlap (k = 6) between C_{+5} and $C_{3q\_1894\_5}$ was statistically the most significant (E = 7.2). C_{+5} was the cluster of probes with overexpression whose differential level $D_1(g)$ was more than 5 and its number of Key-UniGenes, n_{+5}, was 205. $C_{3q\_1894\_5}$ was the region from 189,400 to 189,900 kbp on chromosome 3 and contained 9 Key-UniGenes ($n_{3q\_1894\_5}$ = 9). The maximum gap ($gap_{3q\_1894\_5}$) between Key-UniGenes in $C_{3q\_1894\_5}$ was 146 kbp. In addition, all evaluation values of any combinations between the exhaustive uncertainty cluster sets of cancer specificity and chromosomal proximity are visualized simultaneously on the EIM as gradation of the gray scale. This gradation pattern could convey the distribution of the false balance to the user through visual perception and enabled the detection of as many significant genes as possible. In addition, note the high resolution of EIM compared with CGH resolution (~20 Mbp).

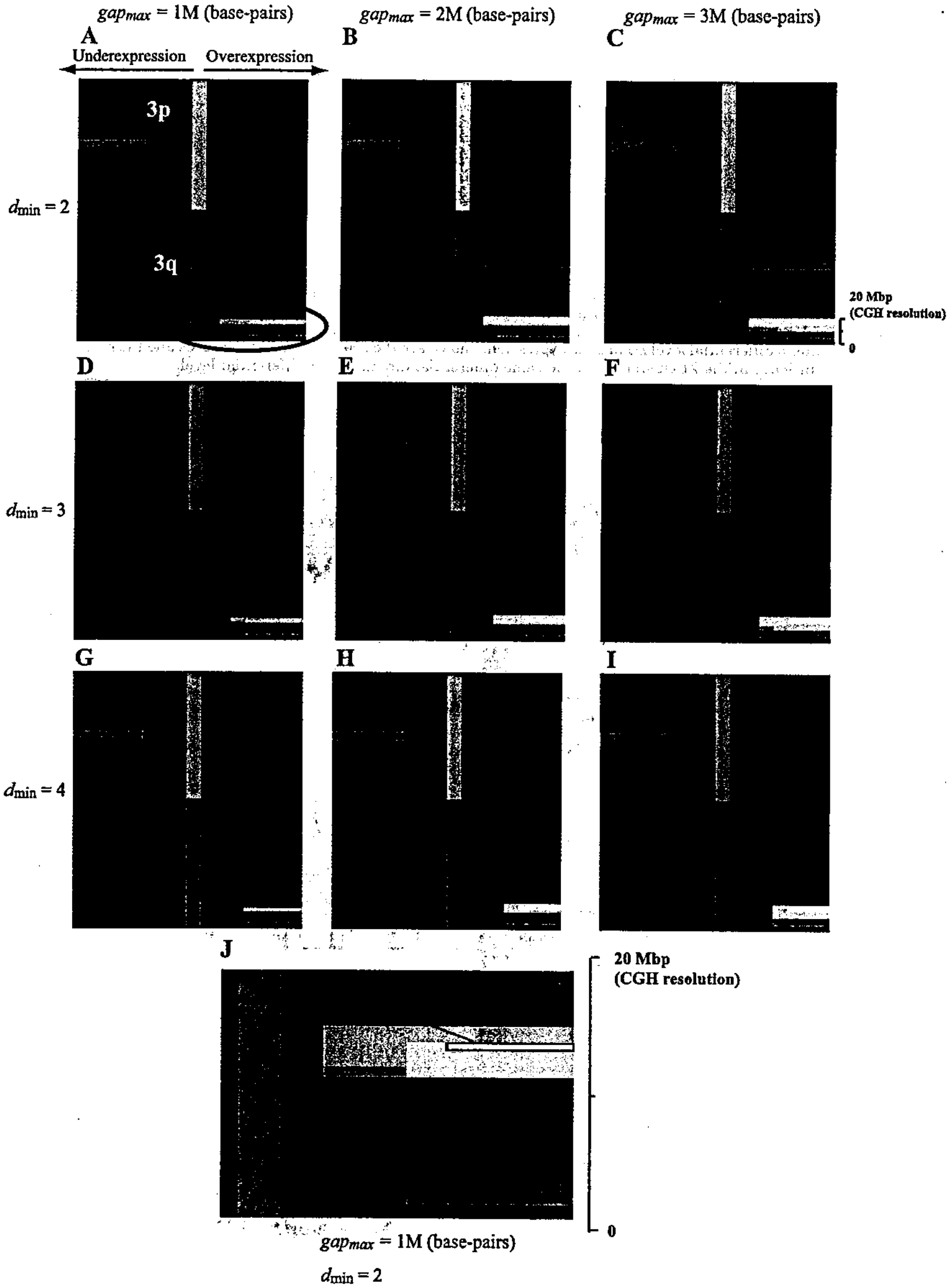

gap_{max} = 1M (base-pairs)
gap_{max} = 2M (base-pairs)
gap_{max} = 3M (base-pairs)
A
B
C
Underexpression
Overexpression
3p
3q
$d_{\min}$ = 2
20 Mbp
(CGH resolution)
0
D
E
F
$d_{\min}$ = 3
G
H
I
$d_{\min}$ = 4
J
20 Mbp
(CGH resolution)
0
gap_{max} = 1M (base-pairs)
$d_{\min}$ = 2

Table 3. *Clusters of probes with under- or overexpression profiles in each squamous cell lung carcinoma*

Differential Direction	Cluster Name ($C_{sign\_diff\_Si}$)	Avg. of Probe Number	Avg. of Key-UniGene Number ($\bar{n}_{sign\_diff}$)	SD of Key-UniGene Number
NL(17) > each SQ	C_{-2_Si}	669	447	103
	C_{-3_Si}	497	331	91
	C_{-4_Si}	387	259	82
	C_{-5_Si}	317	211	76
	C_{-6_Si}	268	181	70
NL(17) < each SQ	C_{+2_Si}	321	208	67
	C_{+3_Si}	188	120	48
	C_{+4_Si}	120	77	35
	C_{+5_Si}	81	50	25
	C_{+6_Si}	58	36	19

To detect individual differences in expression imbalance among 21 SQs, probes (on the U95A array) with under- or overexpression profiles in a SQ specimen, S_i (i = 1,2,. . .,21), compared with NLs were extracted as clusters, $C_{sign\_diff\_Si}$. This extraction was independently performed, regarding each SQ specimen. The suffix *sign* indicates the differential direction (+, overexpression; −, underexpression in each SQ specimen), *diff* indicates a differential level D_2 in gene expression. Shown are the average number of probes and the average and standard deviation (SD) of Key-UniGenes in the 21 clusters with the same differential direction and differential level.

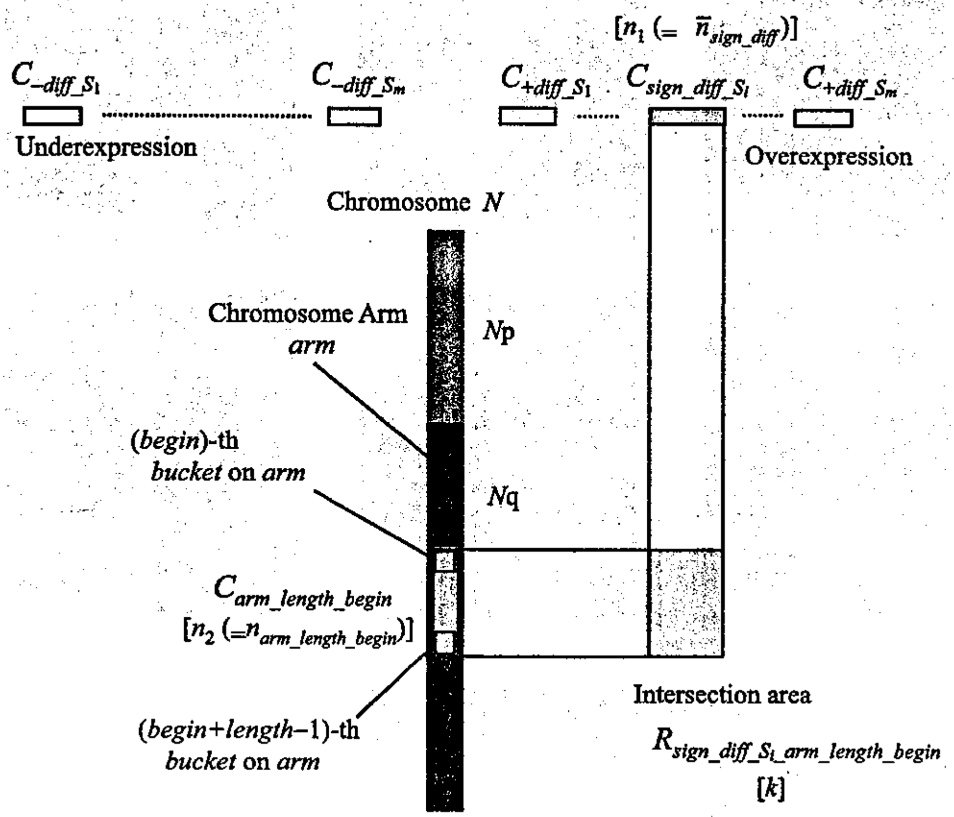


Fig. 8. Individual-specimen clusters vs. locus clusters. In a manner similar to the EIM for detecting expression imbalance of SQ specimen group, that for detecting individual differences in expression imbalance among SQ specimens was also constructed. In a SQ specimen S_i (i = 1, 2,. . ., 21), probes with expression whose differential level $D_2(g,S_i)$ was equal to or higher than *diff* compared with NL specimens were extracted as an individual-specimen cluster, $C_{sign\_diff\_Si}$. This extraction was independently performed with respect to each SQ specimen. The individual-specimen clusters, $C_{sign\_diff\_Si}$ values, were arranged on the abscissa with respect to each S_i, and the locus clusters, $C_{arm\_length\_begin}$ values, on the ordinate. Among $C_{sign\_diff\_Si}$ values, the clusters of under- and overexpression were arranged on the *left* and *right* side, respectively. Since the abscissa represented an array of S_i, it was impossible to represent *diff* on the abscissa like Fig. 4. Therefore, the EIM for individual specimen was visualized by $C_{sign\_diff\_Si}$ with a defined *diff*, and allowed the user to change *diff* interactively; $\bar{n}_{sign\_diff}$ is the average number of Key-UniGenes in $\{C_{sign\_diff\_Si}\}$(i = 1, 2,. . ., 21); $n_{arm\_length\_begin}$ is the number of Key-UniGenes in $C_{arm\_length\_begin}$; k is the number of common Key-UniGenes between $C_{sign\_diff\_Si}$ and $C_{arm\_length\_begin}$. The significance of overlap between $C_{sign\_diff\_Si}$ and $C_{arm\_length\_begin}$ was visualized in the intersection area $R_{sign\_diff\_Si\_arm\_length\_begin}$ as a gray scale.

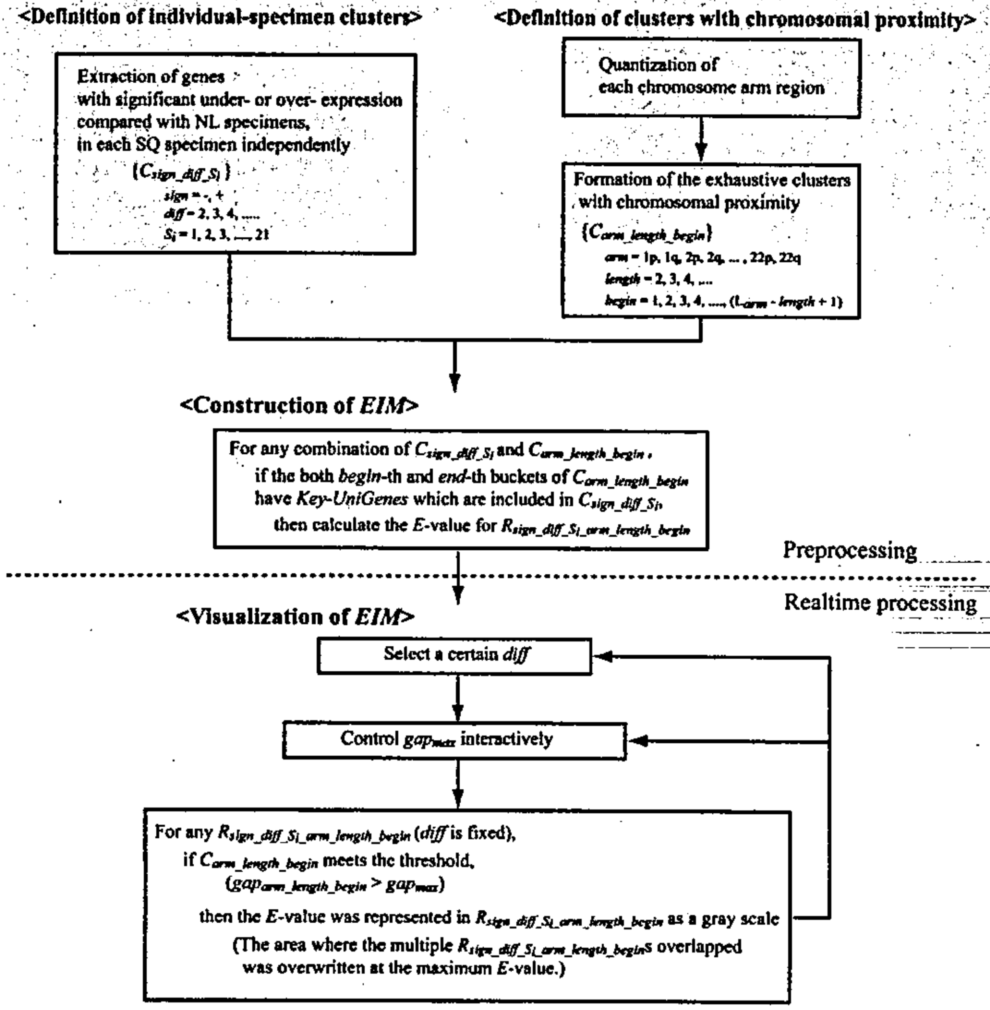
when both (*begin*)-th and (*begin* + *length* − 1)-th buckets of $C_{arm\_length\_begin}$ have the Key-UniGenes that are included in $C_{sign\_diff\_Si}$. This calculation was preprocessing for the EIM. Then, in real-time processing, after a certain *diff* was selected, each *E* value was represented in the intersection area, $R_{sign\_diff\_Si\_arm\_length\_begin}$, as a gray scale, if $C_{arm\_length\_begin}$ met $gap_{\max}$. The user can control *diff* and $gap_{\max}$ interactively.

A flowchart that details these steps is shown in Fig. 9. The EIM for detecting individual difference of expression imbalance among SQ specimens is shown in Fig. 10. Figure 11 shows chromosome 3 of the EIM and the influence of $gap_{\max}$ and *diff* on the detection of the individual differences in expression imbalance among SQs.

RESULTS AND DISCUSSION

Detection of Expression Imbalance Specific to SQs

The EIM showed the distribution of expression imbalance specific to SQs (Fig. 6). It is highly comparable to previous CGH data of lung cancer reported by other investigators (6, 8, 9). There are significant differences among these CGH data because of method variation and sample preparation (especially tumor fraction of clinical samples). So it may be of little importance to compare details with individual CGH experiments. However, the most frequent abnormal loci reported in most of these studies were also detected by the EIM as regional signal images on chromosomes (expression imbalance regions), such as loss of 3p, 4q, 5q, and 8p, and gain of 1q, 3q, and 12p (6, 8, 9). The major difference from the CGH image is that signals are detected in a more confined area, which reflects the high resolution of EIM. Figures 6, 7, 10, and 11 clearly show the high resolution of EIM compared with CGH image. Especially, the intersection area $R_{+5\_3q\_1894\_5}$ showed the most significant overexpression region on 3q (Fig. 7), which is reported to be the most frequent aberration



<Definition of individual-specimen clusters>
Extraction of genes with significant under- or over- expression compared with NL specimens, in each SQ specimen independently
{$C_{sign\_diff\_S_i}$}
sign = −, +
diff = 2, 3, 4, ...
S_i = 1, 2, 3, ..., 21
<Definition of clusters with chromosomal proximity>
Quantization of each chromosome arm region
Formation of the exhaustive clusters with chromosomal proximity
{$C_{arm\_length\_begin}$}
arm = 1p, 1q, 2p, 2q, ... , 22p, 22q
length = 2, 3, 4, ...
begin = 1, 2, 3, 4, ..., (L_{arm} - length + 1)
<Construction of EIM>
For any combination of $C_{sign\_diff\_S_i}$ and $C_{arm\_length\_begin}$,
if the both begin-th and end-th buckets of $C_{arm\_length\_begin}$
have Key-UniGenes which are included in $C_{sign\_diff\_S_i}$,
then calculate the E-value for $R_{sign\_diff\_S_i\_arm\_length\_begin}$
Preprocessing
Realtime processing
<Visualization of EIM>
Select a certain diff
Control gap_{max} interactively
For any $R_{sign\_diff\_S_i\_arm\_length\_begin}$ (diff is fixed),
if $C_{arm\_length\_begin}$ meets the threshold,
($gap_{arm\_length\_begin} > gap_{max}$)
then the E-value was represented in $R_{sign\_diff\_S_i\_arm\_length\_begin}$ as a gray scale
(The area where the multiple $R_{sign_diff_S_i_arm_length_begin}$s overlapped
was overwritten at the maximum E-value.)

Fig. 9. Flowchart for construction of the EIM for detecting individual differences in expression imbalance among SQs. This flowchart provides details of the steps of the EIM for detecting individual differences in expression imbalance among SQs. For the step of "Definition of clusters with chromosomal proximity," please refer to Fig. 2. For the step of "Construction of the EIM" and "Visualization of EIM," please refer to Fig. 8. In this type of EIM, since the abscissa represented an array of S_i, it was impossible to represent *diff* on the abscissa like Fig. 4. Therefore, the EIM for individual specimen was visualized by $C_{sign\_diff\_Si}$ with a defined *diff*, and allowed the user to change *diff* interactively. In addition, it is possible to exclude regions containing large gaps between genes by changing $gap_{\max}$ interactively.

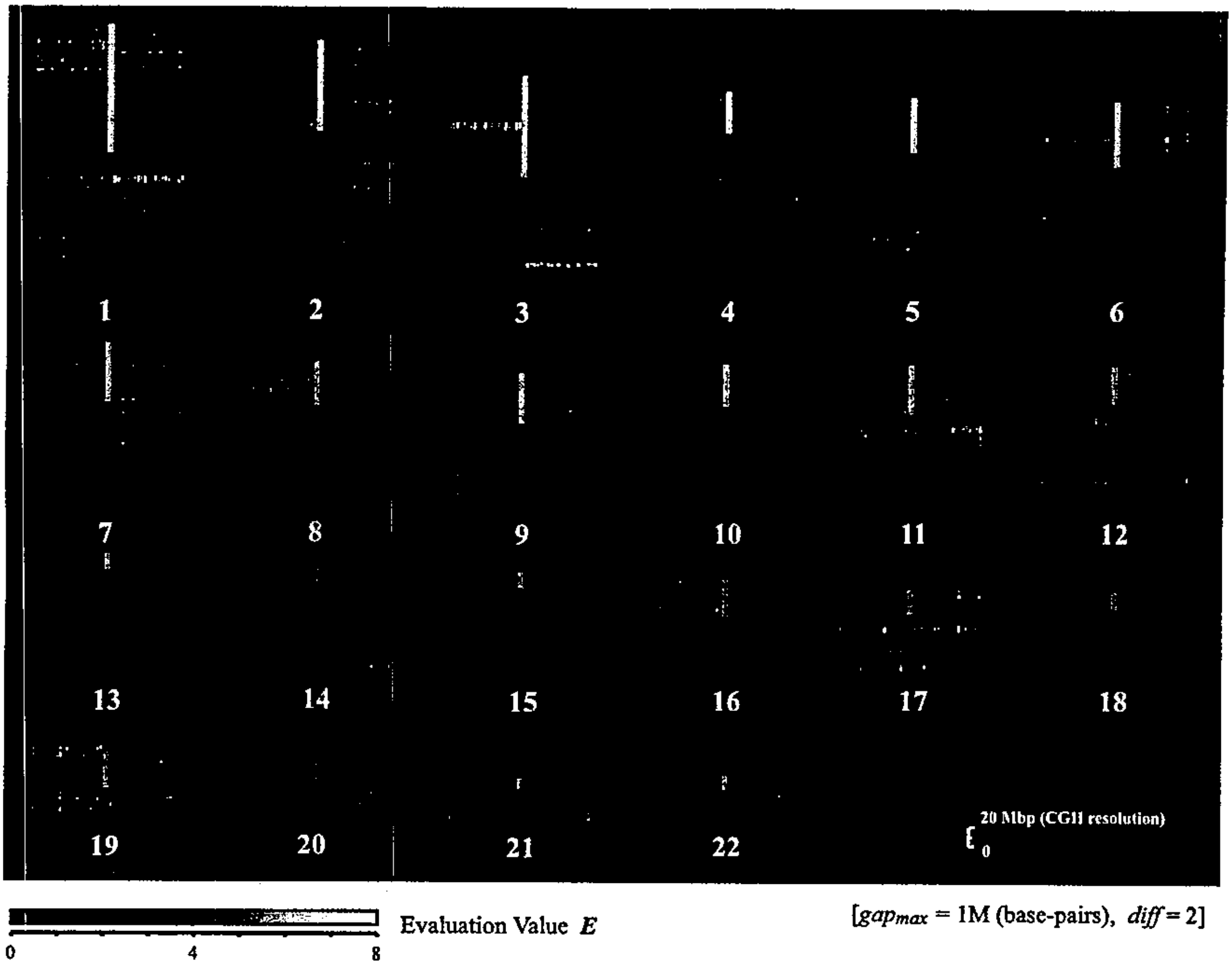
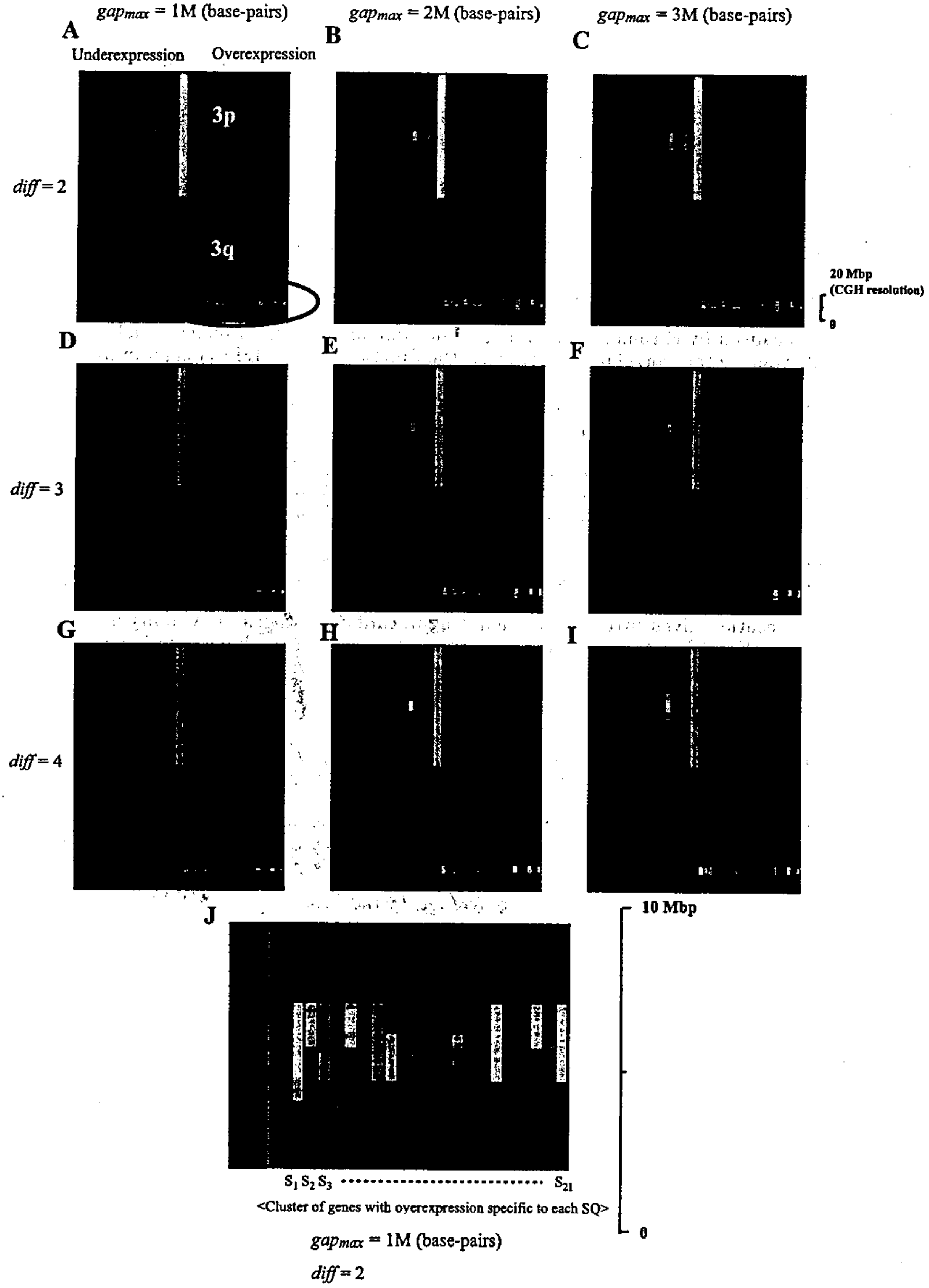


Fig. 10. The EIM for detecting individual difference of expression imbalance among SQs. The EIM was applied for detecting individual differences of expression imbalance among the SQs. Regions of underexpression and overexpression were visualized on the *left* and *right* side, respectively, as gray regional signals. The expression imbalance regions in each SQ were evaluated independently. Note the high resolution of EIM compared with CGH resolution (~20 Mbp).

in SQs by CGH (6, 8, 9). That is, the overlap (k = 6) between C_{+5} (the cluster of probes with overexpression whose differential level $D_1(g)$ is more than 5: n_{+5} = 205) and $C_{3q\_1894\_5}$ (the region from 189,400 to 189,900 kbp on chromosome 3: $n_{3q\_1894\_5} = 9, gap_{3q\_1894\_5} = 146$ kbp) was statistically the most significant (E = 7.2). Therefore, the overlap was evaluated using the hypergeometric probability for observing at least 6 (=k) common elements between randomly selected 205 (=n_{+5}) and 9 (=$n_{3q\_1894\_5}$) elements among 6,652 (=U) elements. The user can narrow down the possible expression imbalance regions by changing $gap_{\max}$ and $d_{\min}$ interactively. Especially, as is shown in Fig. 7, *A–I*, changing $gap_{\max}$, which allows exclusion of the regions containing large gaps between genes, markedly influenced the detection of expression imbalance regions. In addition, all evaluation values of any combinations between the exhaustive uncertainty cluster sets of cancer specificity and chromosomal proximity are visualized simultaneously on the EIM as gradation

Fig. 11. Individual difference of expression imbalance on chromosome 3. *A–I*: chromosome 3 of the EIM and the influence of $gap_{\max}$ and *diff* on the detection of individual differences in expression imbalance among SQs. With regard to each SQ specimen, the under- and overexpression regions were visualized on the *left* and *right* side, respectively. Since the expression imbalance regions in each SQ were evaluated independently, this type of EIM clarified the individual difference of the overexpression region on 3q, which was detected as the most significant region in the group of SQs by another type of EIM. The user can narrow down the possible expression imbalance regions by changing $gap_{\max}$ and *diff*. *J*: macrograph of the encircled *region A* from panel *A*. When $gap_{\max}$ was 1 Mbp and *diff* was 2, the EIM showed that 17 of 21 SQs had overexpression regions on 3q, which is comparable to other data sets by CGH (6, 8, 9). In addition, note the high resolution of the EIM compared with CGH resolution (~20 Mbp).


A
gapmax = 1M (base-pairs)
Underexpression
Overexpression
3p
3q
B
gapmax = 2M (base-pairs)
C
gapmax = 3M (base-pairs)
diff = 2
20 Mbp
(CGH resolution)
0
D
E
F
diff = 3
G
H
I
diff = 4
J
10 Mbp
S1 S2 S3 .... S21
<Cluster of genes with overexpression specific to each SQ>
0
gapmax = 1M (base-pairs)
diff = 2

of gray scale, which is clearly shown in Fig. 7*J*. This gradation pattern could convey the distribution of the false balance to the user through visual perception and enabled the detection of as many significant genes as possible.

Table 4 shows the gene list of $C_{3q\_1894\_5}$. Although this overexpression region strongly reflected the known genomic gain detected by CGH, several probes without overexpression were also detected on this region. There may be several reasons for this. First, since several probes with low quality were possibly included in this region, signal intensity does not always reflect their target mRNA expression levels. Improvement of the quality of probes would make it possible to detect the overexpression region more clearly. Second, mRNA expression levels would not completely reflect genomic copy number changes caused by chromosomal gain or loss, although there was strong correlation between them, because they are under various transcriptional control including feedback pathway of lost or gained genes themselves. Mukasa et al. (7) also reported that several genes without reduction of expression were detected in 1pLOH region of oligodendrogliomas. In addition, it should be stated that cancer tissues used here contained significant number of noncancerous stromal or inflammatory cells, which add noisy expression to cancer profiling.

Because of the complex factors discussed above, simple spatial mapping of the microarray expression profiles on chromosomal location gives little information about genomic structure (Fig. 12, *left*). In addition, it is very difficult to define adequate thresholds for cancer specificity and chromosomal proximity, because the distribution of "false balance" is unclear and the risk of overlooking significant genes by arbitrary selection of thresholds is high (i.e., the "threshold problem"). However, the EIM, using a new methodology without arbitrary selection of thresholds in conjunction with hypergeometric distribution-based algorithm, has a high tolerance of these complex factors and controls the risk of overlooking the expression imbalance regions. This advantage of the EIM over the simple spatial mapping is clearly shown in Fig. 12. The EIM detected the underexpression regions, *A* and *B*, and overexpression region, *C*, on chromosome 11, which are known loci with frequent genomic gain or genomic loss (6, 8, 9), although it was difficult to detect it from the simple spatial mapping of D_1 value.

Detection of Individual Difference in Expression Imbalance Among SQ Specimens

The analysis for extraction of probes with expression profiles specific to the group of cancer is very effective and popular. However, this type of analysis sometimes raises a critical problem because the individual difference among a group is unobservable. In this context, the function of the EIM to detect individual difference of expression imbalance in a group is very significant. Figure 11, *A–I*, shows that the user can narrow down the possible expression imbalance regions on chromosome 3 by changing $gap_{\max}$ and *diff* interactively. Furthermore, Fig. 11*J* shows the individual difference in the most significant overexpression regions on 3q ($gap_{\max}$ = 1 Mbp, *diff* = 2), where 17 of 21 SQs had overexpression regions, a finding comparable with other data sets analyzed by CGH (6, 8, 9).

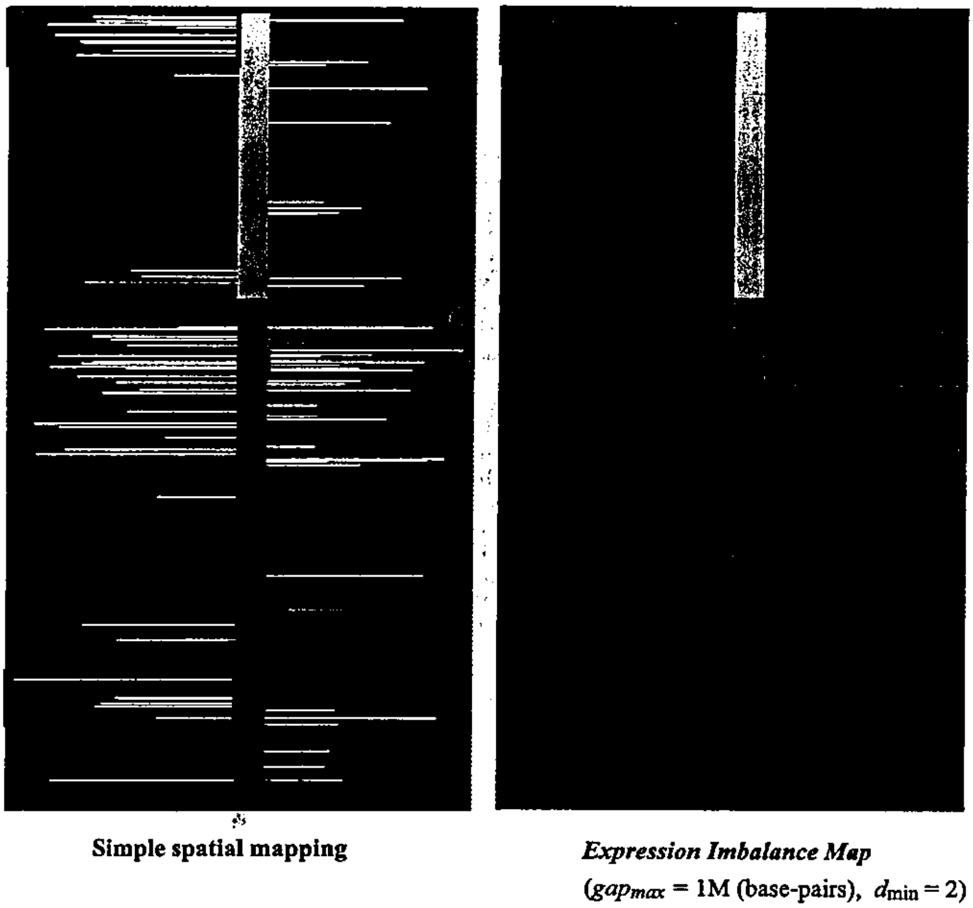
The high-resolution spatial map of expression profiles described in this report, i.e., the EIM, has several significant advantages. Its validity is clearly shown by the fact that many known loci with high frequent genomic losses or gains were detected by regional signals obtained with high resolution by this method.

Recently, several studies have been reported on microarray-based CGH for detecting genome-wide copy number changes (10). However, to our knowledge, no spatial mapping data obtained with such validity and genome-wide coverage have ever been reported previously from this array-CGH method. Experimental difficulty of genome hybridization and limited number of

Table 4. *Gene list of the overexpression region on 3q detected by the EIM*

Cancer Specificity	UniGene	Location, base pairs	Description
*	Hs.108660	189457995	ATP-binding cassette, subfamily C (CFTR/MRP), member_5
?	Hs.343882	189554055	CaM-KII inhibitory protein
x	Hs.129801	189604044	KIAA0604 gene product
x	Hs.1166	189609401	thrombopoietin (myeloproliferative leukemia virus oncogene ligand, megakaryocyte growth and development factor)
*	Hs.74619	189621219	proteasome (prosome, macropain) 26S subunit, non-ATPase, 2
x	Hs.141660	189658124	chloride channel 2 ·
*	Hs.211568	189734699	eukaryotic translation initiation factor 4 gamma, 1
?	Hs.146161	189735389	hypothetical protein MGC2408
*	Hs.153591	189832147	Not56 (*D. melanogaster*)-like protein
*	Hs.174044	189851048	dishevelled 3 (homologous to *Drosophila* dsh)
*	Hs.152936	189862279	adaptor-related protein complex 2, mu 1 subunit

The expression imbalance map (EIM) detected the most significant overexpression regions, $R_{+5\_3q\_1894\_5}$, on 3q in the SQs. This region is a known locus with frequent genomic gains (6, 8, 9). This table shows the gene list of intersection area $R_{+5\_3q\_1894\_5}$. $R_{+5\_3q\_1894\_5}$ evaluated the overlap between C_{+5} (the cluster of probes on the U95A oligonucleotide arrays with overexpression whose differential level are more than 5) and $C_{3q\_1894\_5}$ (the region from 189,400 to 189,900 kbp on chromosome 3: $gap_{3q\_1894\_5}$ = 146 kbp). Differential levels of the genes marked with an asterisk (*) were more than 5, and those of the genes with "x" were less than 5. The genes with "?" were not the Key-UniGenes but the UniGenes that were contained in Genes On Sequence Map.



Simple spatial mapping

Expression Imbalance Map

(gap_{max} = 1M (base-pairs), $d_{\min} = 2$)

Fig. 12. Advantages of the EIM over the simple spatial mapping of expression profiles. *Left*: a simple spatial mapping of D_1 value, which was calculated from the expression profiles of SQs, on chromosome 11. *Right*: the EIM of the same region. The EIM allowed detection of the underexpression regions, *A* and *B*, and overexpression region, *C*, on chromosome 11, which are known loci with genomic gain or genomic loss (6, 8, 9), although it is difficult to detect it by simple spatial mapping.

probes on CGH array eould be major problems for it. There may be several reasons for the successful result of our alternative approach, calculation of genomic structure from expression profile. The first reason is the use of the Affymetrix-type GeneChip. The large number of probes (12,533) available enables detection of a relatively short abnormal region (chromosomal loss can frequently affect areas as short as a few hundred kbp), although this method can be easily applied to other types of microarrays. The second reason, which is most important, is that the EIM is a visualization method using a new methodology without arbitrary selection of thresholds in conjunction with hypergeometric distribution-based algorithm. By processing the complex factors and the threshold problems which hinder user's visual perception of essential information, the EIM presents to the user a comprehensive visual image of whole genome-wide information, clearly indicating where expression imbalance regions are and which genes are to be examined. It has an obvious advantage over simple spatial mapping of the expression profiles. For further curation by the user, simple clicking of a selected expression imbalance region on the EIM image leads to a direct link to a file that contains the actual gene names of the region, their expression scores, and other biological information. In addition, if the user input the UniGene number of genes of interest, the EIM indicates its position on the chromosome. Therefore, the EIM can be a broadband interface that enables user's visual perception of complex data and further curation.

Using the EIM, we might be able to detect regional under- or overexpressions independent of copy number changes, such as gene methylation silencing and/or imprinting abnormality (11). In addition, by using the Kruskal-Wallis test (4), which is a rank sum test to deal with three or more data groups instead of Mann-Whitney test, the EIM can easily extend to multiple phenotypes.

In conjunction with the microdissection technique, which can isolate only tumor-cell-specific RNA (2), our EIM can more precisely detect potential genomic structural changes, which offer more diagnostic and therapeutic impact.

Conclusion

In this report, we describe the development of the expression imbalance map, or EIM, a visualization method without arbitrary selection of thresholds, in conjunction with hypergeometric distribution-based algorithm, for detecting expression imbalance regions. By using this method, many known as well as potential loci with high frequent genomic losses or gains were detected as regional signals with much higher resolution than conventional methods, such as CGH. The EIM can be a broadband interface which enables user's visual perception of complex data and further curation,