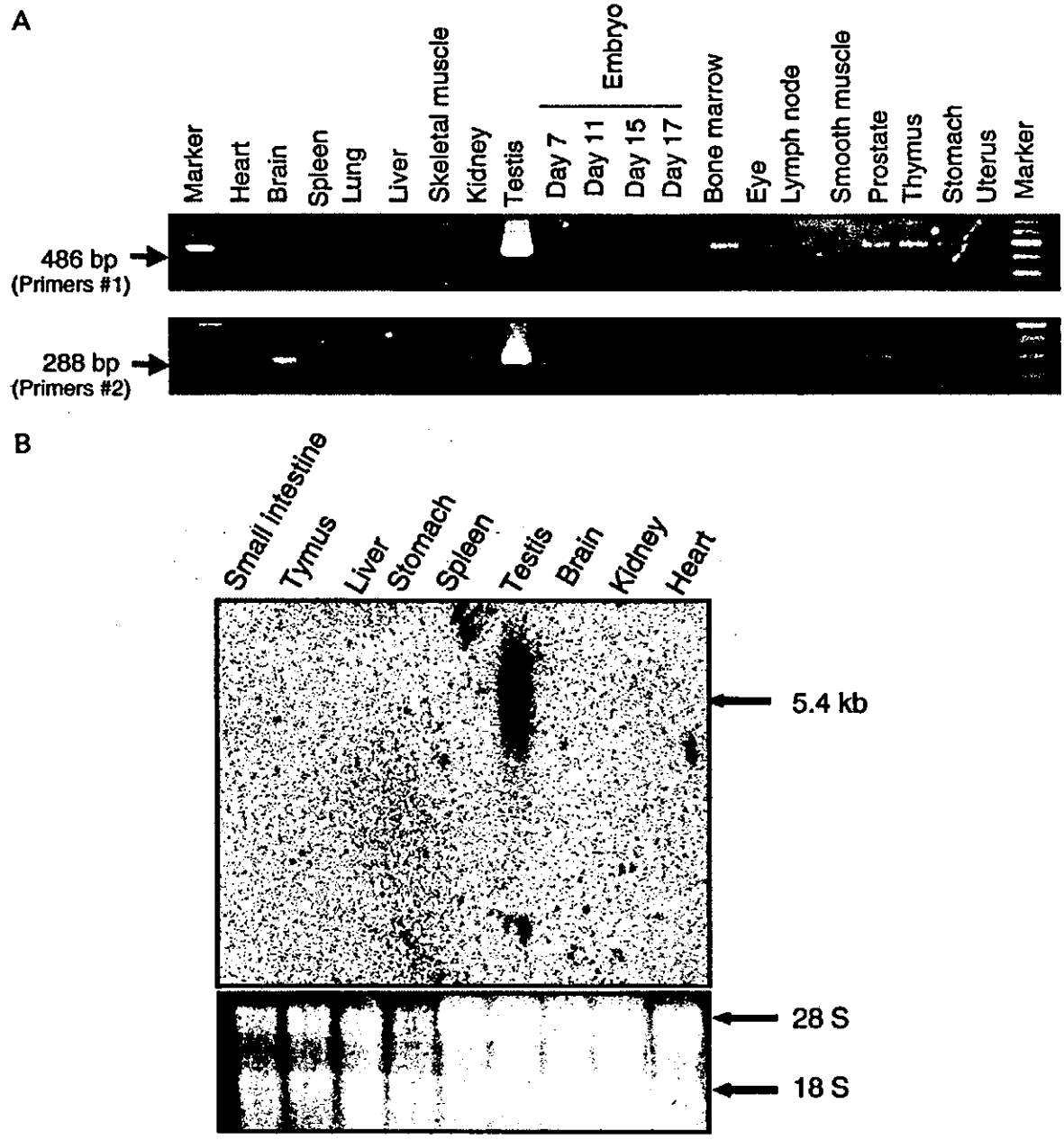


Fig. 5. Expression of the mouse *Abcc12* gene in different tissues. The Abcc12 transcript was detected by PCR (A) and by Northern hybridization (B) as described in Section 2. For the PCR detection (A), two sets of primers (#1 and #2; Fig. 1) were used. The resulting PCR products were 486 and 288 bp, as indicated by arrows. For the Northern hybridization (B), RNA (15 μg/lane) prepared from mouse tissues was fractionated by electrophoresis in 1.0% (w/v) agarose gels and visualized by ethidium bromide (bottom). 18S and 28S rRNAs are indicated by arrows. Northern hybridization (top) with a ^{32}P-labeled probe was carried out as described in Section 2. The detected Abcc12 mRNA (5.4 kb) is indicated by an arrow.

Abcc12 was observed in the testis. Relatively lower expression was detected in the brain, bone marrow, eye, lymph node, prostate, thymus, stomach, and uterus in the adult mouse. Northern blot hybridization (Fig. 5B) clearly demonstrates the predominant expression of mouse Abcc12 in the testis, being consistent with the results of RT–PCR (Fig. 5A). The transcript size of mouse Abcc 12 was about 5.4 kb.

3.5. Localization of mouse Abcc 12 in the testis

To elucidate the expression site of Abcc12 in the mouse testis, we have carried out laser-captured microdissection and RT–PCR. The seminiferous tubules and the interstitium were dissected, and RNA was extracted to prepare cDNA (see Section 2). PCR was carried out with the same primer

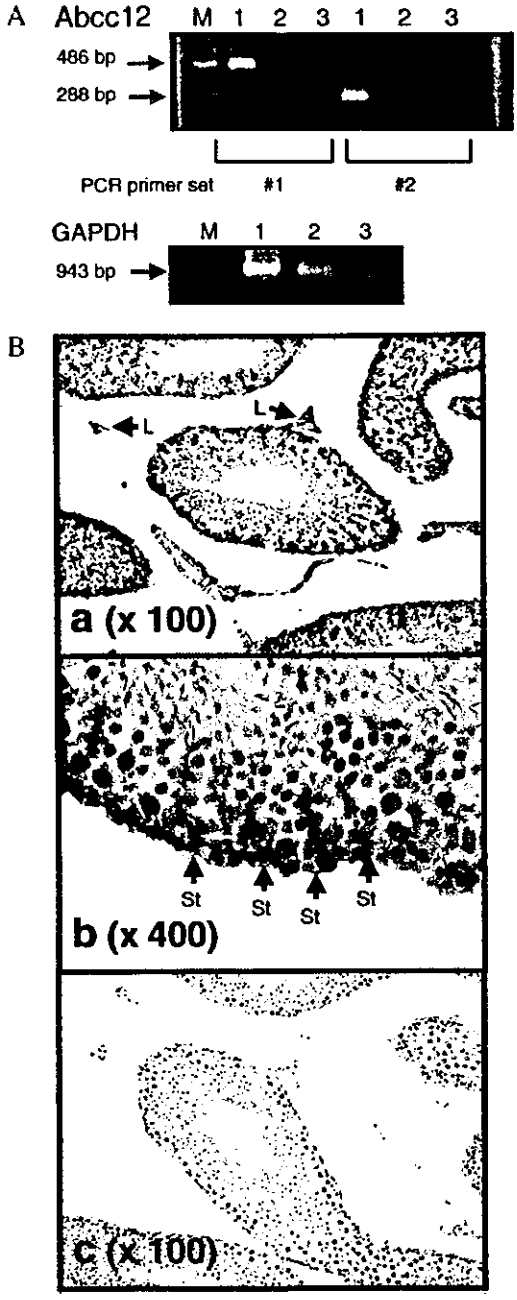


Fig. 6. Detection of the Abcc12 transcript in the mouse testis by means of laser-captured microdissection and RT–PCR (A) as well as by in situ hybridization (B). (A) Abcc12 and GAPDH transcripts were detected by PCR with RT reaction products prepared from micro-dissected samples. For the PCR detection, two sets of primers (#1 and #2; Fig. 1) were used, and the resulting PCR products were 486 and 288 bp, as indicated by arrows. The 943 bp product of GAPDH is the positive control for the PCR reaction. Lane M, DNA size markers; lane 1, seminiferous tubules; lane 2, stroma cells; lane 3, without RT reaction products. (B) The Abcc12 transcript in the mouse testis was detected by in situ hybridization as described in Section 2. Panels a and b show the results of hybridization with the anti-sense probe, whereas panel c shows the negative control, i.e. hybridization with the sense probe. Magnifications are indicated in parentheses. Arrows indicate Leydig (L) and Sertoli (St) cells.

Fig. 7. Schematic illustration for alternative splicing of the *Abcc12* gene. Exons are numbered according to the sequence of the cloned Abcc12 cDNA. Variant A cDNA has one extra exon (133 bp) with a stop codon (TGA) between exons 16 and 17. Variant B cDNA has one extra exon (99 bp) with a stop codon (TAG) between exons 13 and 14. The exon 15 of the FANTOM 2 cDNA (ID = 4932443H13) has a 727 bp extension, as compared with the exon 15 of Abcc12 cDNA.

sets #1 and #2 as described above. As shown in Fig. 6A, Abcc12 expression was exclusively high in the seminiferous tubules, whereas little expression was detected in the interstitium. To gain further insight into cell type-specific expression of the *Abcc12* gene, we carried out in situ hybridization. Fig. 6B depicts the results of the in situ hybridization, demonstrating that the expression of Abcc12 was high in Sertoli cells of the seminiferous tubules (Fig. 6B, panel b). In addition, expression of Abcc12 was also detected in Leydig cells of the interstitium (Fig. 6B, panel a) under our hybridization conditions. No hybridization signal was observed with the sense probe, as the negative control (Fig. 6B, panel c).

3.6. Splicing variants of Abcc12

During the cloning of Abcc12 cDNA in the present study, we have discovered two variant forms of Abcc12 (GenBank accession numbers: AF514414 and AF514415 for variants A and B, respectively). Fig. 7 summarizes the configurations of those variants of the Abcc12 transcript together with the partial cDNA (ID = 4932443H13) reported in the FANTOM 2 database. The cDNAs of both variants A and B consist of 30 exons. As shown in Fig. 7, the variant A cDNA has an extra exon (133 bp) located between exons 16 and 17. Although the variant A cDNA consists of a total of 30 exons, the variant A is considered to encode a short peptide (775 amino acid residues), because the extra exon has a translation stop codon, TGA (Fig. 7). Likewise, the variant B cDNA has one extra exon (99 bp) with a stop codon (TAG) between exons 13 and 14 (Fig. 7), and, therefore, it also encodes a short peptide (687 amino acid residues). On the other hand, the FANTOM 2 cDNA (ID = 4932443H13) cloned by the 5′-oligo-cap method (Carninci et al., 1996) has an extension (121 bp) at the

5′-end of the cDNA, as compared with the cloned Abcc12 cDNA (data not shown). It is noteworthy that the exon 15 of the FANTOM 2 cDNA is different from that of the Abcc12 cDNA cloned in this study, although the other exons 2–14 are identical. Indeed, the exon 15 in the FANTOM 2 cDNA is 727 bp larger than the exon 15 of Abcc12 cDNA, but it encodes a translation stop codon (TGA) in the extended sequence (Fig. 7).

4. Discussion

4.1. Molecular characteristics of mouse Abcc12 cDNA

In the present study, we have cloned and characterized the cDNA of a new mouse ABC transporter, named Abcc12. The cloned cDNA was 4511 bp long and comprised a 4101 bp open reading frame. The deduced peptide consists of 1367 amino acid moieties, carrying two sets of Walker A, Walker B (Walker et al., 1982), and signature C (Higgins, 1992) motifs within the peptide (Fig. 2A). Based on the ATP binding cassettes and the putative trans-membrane spanning domains (Fig. 2B), Abcc12 is regarded as a 'full' ABC protein. The amino acid sequence of the Abcc12 protein deduced from the cloned cDNA exhibits the highest identity (84.5%) to human ABCC12 among all of the members of the ABCC subfamily hitherto identified in the human and the mouse (Table 1 and Fig. 3). Indeed, the hydropathy profile of mouse Abcc12 is virtually the same as that of human ABCC12 (Fig. 2B). From these results, it could be concluded that mouse Abcc12 is the orthologue of human ABCC12. In addition, our data suggest that the cDNA sequence of human ABCC12 (MRP9) recently reported by Bera et al. (2002) may be a splicing variant form, since exons 5 and 16 are missing in their sequence.

Based on the phylogenic relationship deduced from the amino acid sequence identities, the ABCC subfamily could be clustered into four classes (Fig. 3). For example, class A involves human ABCC1, ABCC2, ABCC3, and ABCC6, as well as mouse Abcc1, Abcc2, and Abcc6. These ABC transporters appear to function as conjugate transporters, e.g., GS-X pumps and/or multi-specific organic anion transporters (cMOAT) (Ishikawa, 1992; Borst and Oude Elferink, 2002). On the other hand, class B includes human ABCC8 (SUR1), ABCC9 (SUR2), and mouse Abcc9, which are sulfonylurea receptors coupled with potassium channels, i.e., Kir 6.1 or Kir 6.2. Human CFTR (ABCC7), ABCC10, mouse Abcc7 (mouse CFTR), and Abcc10 are involved in class C. Mutations in the *CFTR* gene are known to be the cause of cystic fibrosis, an autosomal recessive genetic disorder affecting a number of organs, including the lungs, airways, pancreas, and sweat glands (http://www.genet.sickkid.on.ca/cftr/). The physiological function of ABCC10 in this class is not known at the present time.

According to this clustering, the mouse Abcc12 belongs to class D, which involves human ABCC4, ABCC5, ABCC11, and ABCC12, as well as mouse Abcc4 and Abcc5 (Fig. 3). Recent studies demonstrated that human both ABCC4 and ABCC5 transport nucleotide analogues (Schuetz et al., 1999; Wijnholds et al., 2000; Jedlitschky et al., 2000; Chen et al., 2001). ABCC5 reportedly does not confer multidrug resistance when over-expressed in human embryonic kidney 293 cells (McAleer et al., 1999). Because of the similarity of the amino acid sequences, it is assumed that human ABCC11 and ABCC12 are functionally related to ABCC4 or ABCC5.

4.2. Mouse Abcc12 gene: an orthologue of human ABCC12 gene

Our conclusion that mouse Abcc12 is the orthologue of human ABCC12 is supported by similarities in the location and organization of those genes, as well. The present study provides evidence that the open reading frame in the mouse Abcc12 cDNA consists of 29 exons, as does the human ABCC12 cDNA (Yabuuchi et al., 2001; Tammur et al., 2001). In addition, the mouse *Abcc12* and human *ABCC12* genes (29 exons and introns) span 62 and 63 kb, respectively. The mouse *Abcc12* gene is located between two microsatellite markers, D8Mit347 and D8Mit348, on the chromosome 8D3 locus. This locus reportedly contains many conserved linkage homologies with human chromosome 16q12.1 (Serikawa et al., 1998), where the human *ABCC12* gene has recently been discovered (Yabuuchi et al., 2001; Tammur et al., 2001). Being consistent with this idea, the chromosomal location of the mouse *Siah 1* gene and its distance (167 kb) from the Abcc12 gene is conserved in the human chromosome 16q12.1 where both *SIAH 1* and *ABCC12* genes are located. The human *SIAH 1* gene (Hu et al., 1997) encodes a 282-amino-acid protein with 76% amino acid identity to the *Drosophila* SINA protein which is involved in the *ras* signaling pathway to mediate the R7 photoreceptor formation in the *Drosophila* eye (Carthew and Rubin, 1990). *Siah 1a* is one of the mouse orthologue genes and is mapped on the chromosome 8D3 locus (Holloway et al., 1997). Taken together, it is strongly suggested that the mouse *Abcc12* gene is closely related to the human *ABCC12* gene in terms of both the protein structure and the organization of the gene.

It is of importance, however, to note that in spite of the tandem location of both *ABCC11* and *ABCC12* genes on human chromosome 16q12.1, there was no mouse orthologue gene corresponding to the human *ABCC11* at that mouse chromosomal locus. In addition, there was no putative *Abcc11* gene detected even by an extensive search throughout the currently available mouse genome data. Thus, it appears that the *Abcc11* gene is absent from the mouse genome.

4.3. Tissue-specific expression of the mouse Abcc12 gene

We have previously reported that the expression of the

human *ABCC12* gene was widely distributed in various tissues, including testis, brain, liver, lung, kidney, thymus, prostate, ovary, colon, and leukocytes as well as in several fetal tissues (Yabuuchi et al., 2001). In contrast, the present study demonstrates that the mouse *Abcc12* gene is expressed at high levels exclusively in the testis (Fig. 5). The reason for such differences in organ-specific expression profiles between mouse Abcc12 and human ABCC12 is not known, but may be eventually explained by analysis of the promoter regions of those genes.

In the present study, by means of laser-captured microdissection combined with RT–PCR as well as in situ hybridization, the *Abcc12* transcript was detected in Sertoli cells of the seminiferous tubules in the mouse testis (Fig. 6A,B). In addition, in situ hybridization further revealed the expression of the *Abcc12* in Leydig cells, as well (Fig. 6B). Accumulating evidence suggests that the blood-testis barrier plays an important role in protecting the germ cells from harmful influences. To date it has been reported that ABCB1 (P-glycoprotein or MDR1) is expressed in luminal capillary endothelium and on the myoid-cell layer around the seminiferous tubule (Bart et al., 2002), whereas ABCC1 (MRP) is located basolaterally on both Sertoli and Leydig cells (Wijnholds et al., 1998). These ABC transporters are regarded as the first line players in the body's detoxification system. In this context, Abcc12 is also considered to play a role as a member of such a detoxification system, or it may be involved in the transport of endogenous substances in the testis. The physiological function and substrate specificity of Abcc12 remains to be elucidated.

4.4. Concluding remarks

Northern blot analysis revealed that mRNA with a size of 5.4 kb is the major transcript of mouse Abcc12 in the testis (Fig. 5B). In the present study, however, we have detected the existence of at least two splice variants for mouse Abcc12 (Fig. 7). In addition, the results of the FANTOM 2 project (The FANTOM Consortium, 2002) demonstrate that there is another splicing variant form that encodes a shorter peptide of Abcc12 (Fig. 7). These data suggests that mouse Abcc12 is transcribed into multiple forms by means of alternative splicing.

In the previous paper, we demonstrated that the human ABCC12 gene is transcribed into several splice variants (Yabuuchi et al., 2001). Recently, Bera et al. (2002) reported that the human ABCC12 (MRP9) is expressed as two major transcripts of 4.5 and 1.3, and that the 4.5 kb transcript is highly expressed in the epithelial cells of breast cancer. Transcript of the ABCC12 gene were detected in cell lines of carcinoma and adenocarcinoma originating from breast, lung, colon pancreas and prostate, as well (Yabuuchi et al., 2001), suggesting that expression of the ABCC12 gene may be up-regulated during carcinogenesis. Therefore, it is of great interest to study how alternative splicing is regulated in the expression of the human ABCC12 and mouse Abcc12 genes.

Acknowledgements

The authors thank Ms. Yukiko Saito (University of Tokyo Medical School) for her technical assistance in the preparation of tissue samples. This study was supported by research grants entitled 'Studies on the genetic polymorphism and function of pharmacokinetics-related proteins in Japanese population' (H12-Genome-026) and 'Toxicoproteomics: expression of ABC transporter genes and drug–drug interactions' (H14-Toxico-002) from the Japanese Ministry of Health and Welfare as well as by a Grant-in-Aid for Creative Scientific Research (No. 13NP0401) and a research grant (No. 14370754) from the Japan Society for the Promotion of Science. In addition, this study was supported, in part, by the institutional core grant of the 21st Century COE Program from the Ministry of Education, Culture, Sports, Science and Technology.

References

Bart, J., Groen, H.J.M., van der Graaf, W.T.A., Hollema, H., Hendrikse, N.H., Vaalburg, W., Sleijfer, D.T., de Vries, E.G.E., 2002. An oncological view on the blood-testis barrier. Lancet Oncol. 3, 357–363.

Bera, T.K., Lee, S., Salvatore, G., Lee, B., Pastan, I., 2001. MRP8, a new member of ABC transporter superfamily, identified by EST database mining and gene prediction program, is highly expressed in breast cancer. Mol. Med. 7, 509–516.

Bera, T.K., Iavarone, C., Kumar, V., Lee, S., Lee, B., Pastan, I., 2002. MRP9, an unusual truncated member of the ABC transporter superfamily, is highly expressed in breast cancer. Proc. Natl. Acad. Sci. USA 99, 6997–7002.

Borst, P., Oude Elferink, R., 2002. Mammalian ABC transporters in health and disease. Annu. Rev. Biochem. 71, 537–592.

Carninci, P., Kvam, C., Kitamura, A., Okazaki, Y., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., Muramatsu, M., Hayashizaki, Y., Schneider, C., 1996. High-efficiency full-length cDNA by biotinylated CAP trapper. Genomics 37, 327–336.

Carthew, R.W., Rubin, G.M., 1990. Seven in absentia, a gene required for specification of R7 cell fate in the Drosophila eye. Cell 63, 561–577.

Chen, Z.S., Lee, K., Kruh, G.D., 2001. Transport of cyclic nucleotides and estradiol 17-beta-D-glucuronide by multidrug resistance protein 4. Resistance to 6-mercaptopurine and 6-thioguanine. J. Biol. Chem. 276, 33747–33754.

Cole, S.P., Bhardwaj, G., Gerlach, J.H., Mackie, J.E., Grant, C.E., Almquist, K.C., Stewart, A.J., Kurz, E.U., Duncan, A.M., Deeley, R.G., 1992. Overexpression of a transporter gene in a multidrug-resistant human lung cancer cell line. Science 258, 1650–1654.

Dean, M., Rzhetsky, A., Allikmets, R., 2001. The human ATP-binding cassette (ABC) transporter superfamily. Genome Res. 11, 1156–1166.

Higgins, C.F., 1992. ABC transporters: from microorganisms to man. Annu. Rev. Cell Biol. 8, 67–113.

Holloway, A.J., Della, N.G., Fletcher, C.F., Largespada, D.A., Copeland, N.G., Jenkins, N.A., Bowtell, D.D., 1997. Chromosomal mapping of five conserved murine homologs of the Drosophila RING finger gene Seven-in-absentia. Genomics 41, 160–168.

Hu, G., Chung, Y.-L., Glover, T., Valentine, V., Look, A.T., Featon, E.R.,

1997. Characterization of human homologs of the Drosophila seven in absentia (sina) gene. Genomics 46, 103–111.

Ishikawa, T., 1989. ATP/Mg^{2+}-dependent cardiac transport system for glutathione S-conjugates: A study using rat heart sarcolemma vesicles. J. Biol. Chem. 264, 17343–17348.

Ishikawa, T., 1992. The ATP-dependent glutathione S-conjugate export pump. Trends Biochem. Sci. 17, 463–468.

Ishikawa, T., 2003. Multidrug resistance: genomics of ABC transporters. Encyclopedia of Human Genome, Nature Publishing Group, in press.

Jedlitschky, G., Burchell, B., Keppler, D., 2000. The multidrug resistance protein 5 functions as an ATP-dependent export pump for cyclic nucleotides. J. Biol. Chem. 275, 30069–30074.

Klein, I., Sarkadi, B., Varadi, A., 1999. An inventory of the human ABC proteins. Biochim. Biophys. Acta 1461, 237–262.

Kozak, M., 1991. An analysis of vertebrate mRNA sequences: intimations of translational control. J. Cell Biol. 115, 887–903.

Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.

Lee, W.-L., Tay, A., Ong, H.-T., Goh, L.-M., Monaco, A.P., Szepetowski, P., 1998. Association of infantile convulsions with paroxysmal dyskinesias (ICCA syndrome): confirmation of linkage to human chromosome 16p12–q12 in a Chinese family. Hum. Genet. 103, 608–612.

Leier, I., Jedlitschky, G., Buchholtz, U., Cole, S.p.C., Deeley, R.G., Keppler, D., 1994. The MRP encodes an ATP-dependent export pump of leukotriene C4 and structurally related conjugates. J. Biol. Chem. 269, 27807–27810.

McAleer, M.A., Breen, M.A., White, N.L., Matthews, N., 1999. pABC11 (also known as MOAT-C and MRP5), a member of the ABC family of proteins, has anion transporter activity but does not confer multidrug resistance when overexpressed in human embryonic kidney 293 cells. J. Biol. Chem. 274, 23541–23548.

Mouse Genome Sequencing Consortium, 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.

Müller, M., Meijer, C., Zaman, G.J., Borst, P., Scheper, R.J., Mulder, N.H., de Vries, E.G.E., Jansen, P.L.M., 1994. Overexpression of the gene encoding the multidrug resistance-associated protein results in increased ATP-dependent glutathione S-conjugate transport. Proc. Natl. Acad. Sci. USA 91, 13033–13037.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Schuetz, J., Connelly, M.C., Sun, D., Paibir, S., Flynn, P.M., Srinivas, R.V., Kumar, A., Fridland, A., 1999. MRP4: A previously unidentified factor in resistance to nucleoside-based antiviral drugs. Nature Med. 5, 1048–1051.

Serikawa, T., Cui, Z., Yokoi, N., Kuramoto, T., Kondo, Y., Kitada, K., Guenet, J.-L., 1998. A comparative genetic map of rat, mouse and human genomes. Exp. Anim. 47, 1–9.

Tammur, J., Prades, C., Arnould, I., Rzhetsky, A., Hutchinson, A., Adachi, M., Schuetz, J.D., Swoboda, K.J., Ptacek, L.J., Rosier, M., Dean, M., Allikmets, R., 2001. Two new genes from the human ATP-binding cassette transporter superfamily, ABCC11 and ABCC12, tandemly duplicated on chromosome 16q12. Gene 273, 89–96.

The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team, 2002. Analysis of the mouse transcriptom based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563–573.

Tomita, H., Nagamitsu, S., Wakui, K., Fukushima, Y., Yamada, K., et al., 1999. Paroxysmal kinesigenic choreoathetosis locus maps to chromosome 16p11.2–p12.1. Am. J. Hum. Genet. 65, 1688–1697.

Walker, J.E., Saraste, M., Runswick, M.J., Gay, N.J., 1982. Distantly related sequences in the α and β subunits of ATP synthetase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J. 1, 945–951.

Wijnholds, J., Scheffer, G.L., van der Valk, M., van der Valk, P., Beijnen, J.H., Scheper, R.J., Borst, P., 1998. Multidrug resistance protein 1 protects the oropharyngeal mucosal layer and the testicular tubules against drug-induced damage. J. Exp Med. 188, 797–808.

Wijnholds, J., Mol, C.A., van Deemter, L., de Haas, M., Scheffer, G.L., Baas, F., Beijnen, J.H., Scheper, R.J., Hatse, S., De Clercq, E., Balzarini, J., Borst, P., 2000. Multidrug-resistance protein 5 is a multispecific organic anion transporter able to transport nucleotide analogs. Proc. Natl. Acad. Sci. USA 97, 7476–7481.

Yabuuchi, H., Shimizu, H., Takayanagi, S., Ishikawa, T., 2001. Multiple splicing variants of two new human ATP-binding cassette transporters, ABCC11 and ABCC12. Biochem. Biophys. Res. Commun. 288, 933–939.

Yabuuchi, H., Takayanagi, S., Yoshinaga, K., Taniguchi, N., Aburatani, H., Ishikawa, T., 2002. ABCC13, an unusual truncated ABC transporter, is highly expressed in fetal human liver. Biochem. Biophys. Res. Commun. 299, 410–417.

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

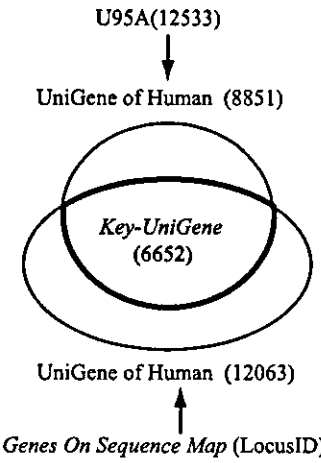


Fig. 1. Correspondence between probeIDs and LocusIDs. To associate gene locus information with gene expression profiles, probeIDs on the Affymetrix U95A oligonucleotide arrays and the LocusIDs on Genes On Sequence Map (*Homo sapiens* build 27) of NCBI were translated into UniGenes. We utilized the 12,063 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*.

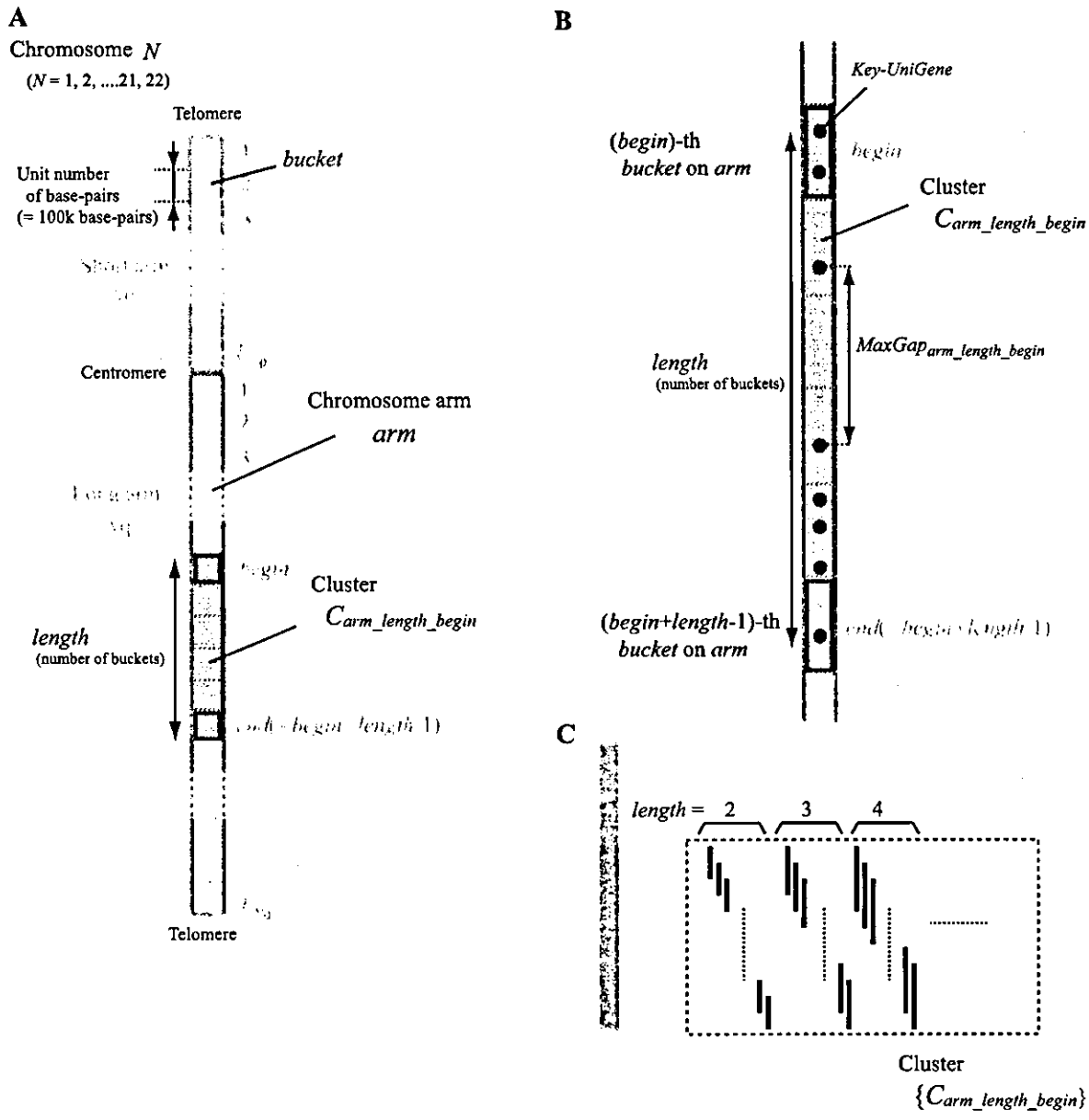


Fig. 2. Formation of clusters of genes with chromosomal proximity. *A*: for easier handling of the gene locus information, each chromosome arm region was quantized by unit region called "bucket" whose length was 100 kbp, and the Key-UniGenes were assigned the corresponding buckets according to their 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.

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

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

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.

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 $n_{arm\_length\_begin}$.

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.

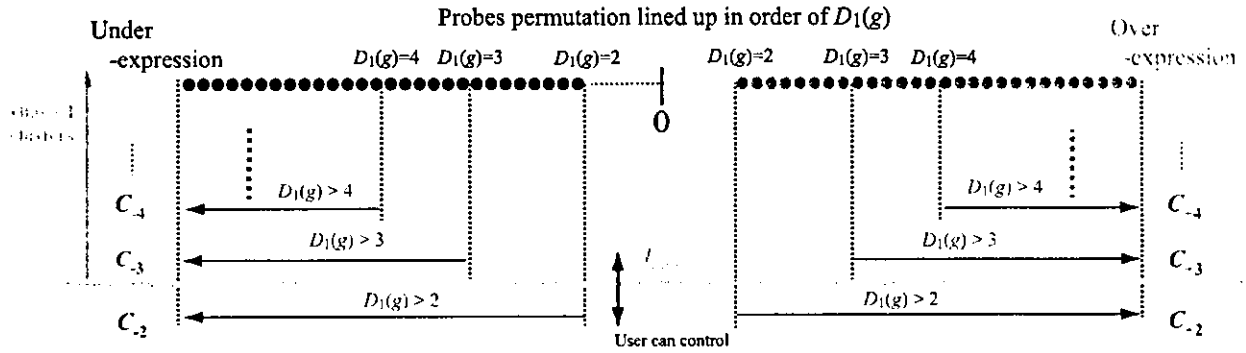


Fig. 3. Probe permutation arranged in order of the difference in gene expression level between squamous cell lung carcinomas (SQs) and normal lungs (NLs). Probes on the U95A arrays are lined up in order of the $D_1(g)$ level, which represents the difference in the gene expression level between SQs and 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.

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

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

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

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

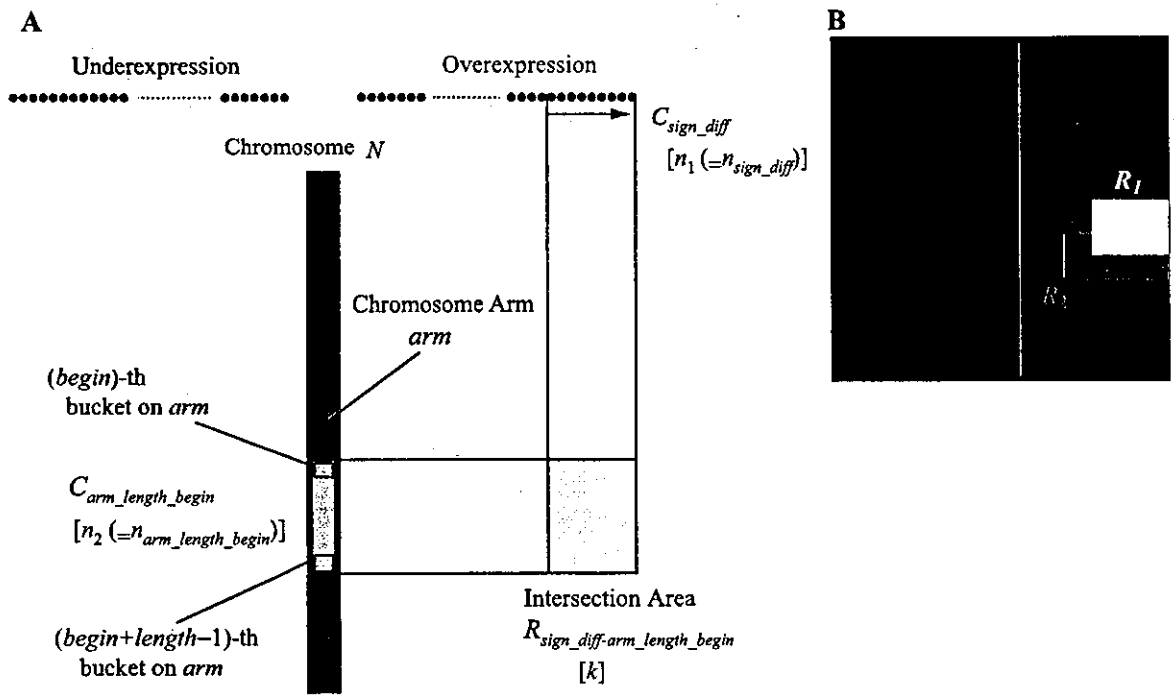


Fig. 4. Clusters of genes specific to the group of SQs vs. clusters of genes with proximity on chromosomes. *A*: to detect expression imbalance regions, it is necessary to search for genes with both cancer specificity and chromosomal proximity. The fundamental algorithm of the EIM is to evaluate statistically 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.

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

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,

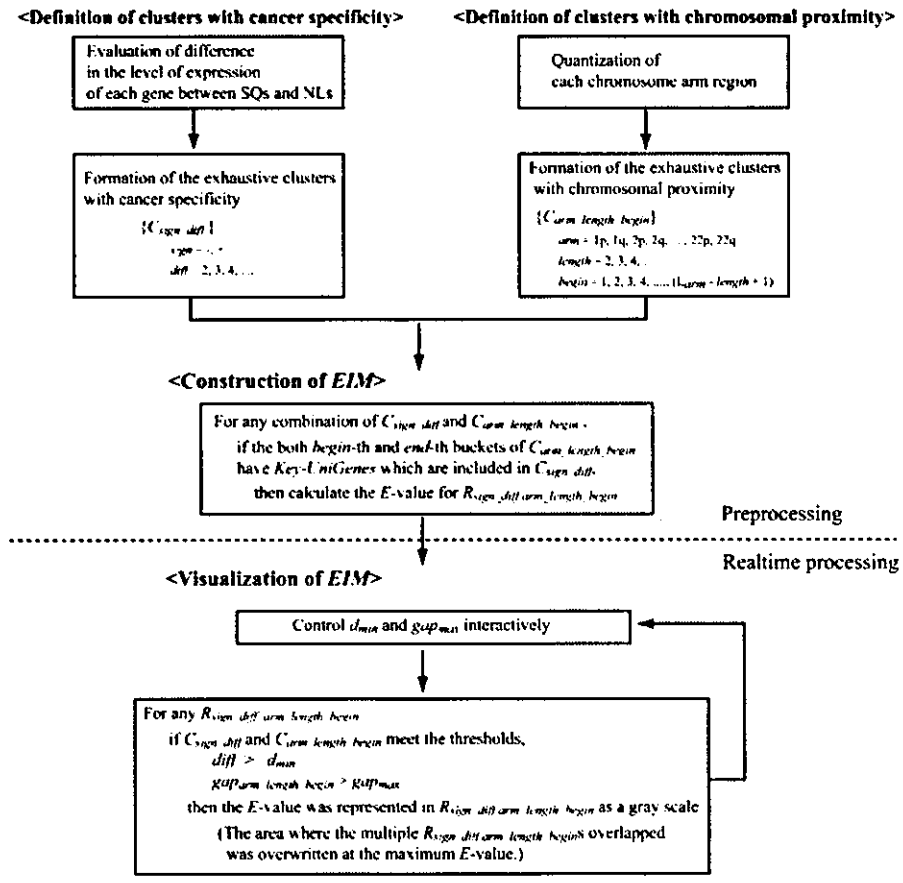


Fig. 5. Flowchart for construction of the EIM for detecting expression imbalance regions specific to SQs. This flowchart provides details of the steps of the EIM for detecting expression imbalance regions specific to SQs. For the steps of "Definition of clusters with cancer specificity," please refer 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}.

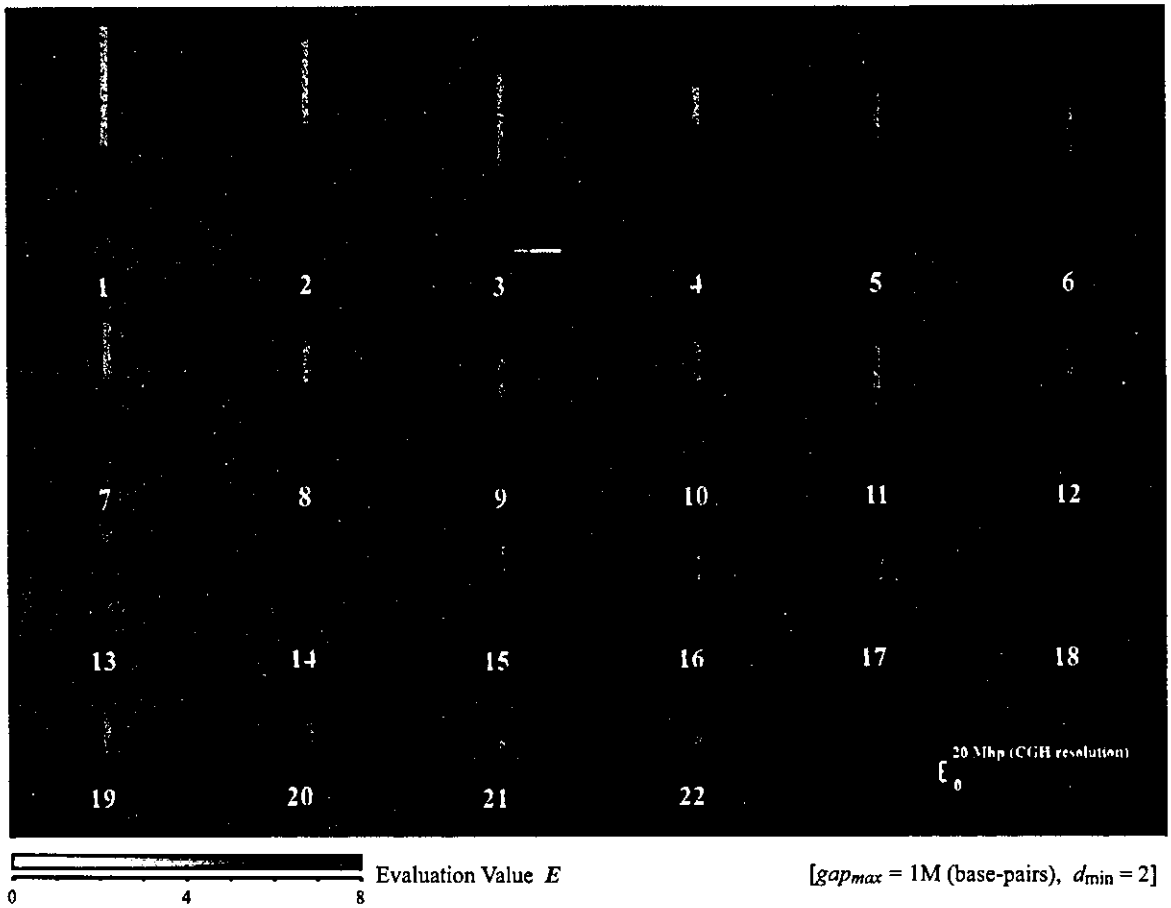


Fig. 6. The EIM applied for detecting expression imbalance regions specific to SQs. The regions of under- and overexpression in SQs were visualized on the *left* and *right* side, respectively, as gray regional signals. All statistical evaluation values of any combinations between the exhaustive uncertainty 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 ($gap_{3q\_1694\_5}$) 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).

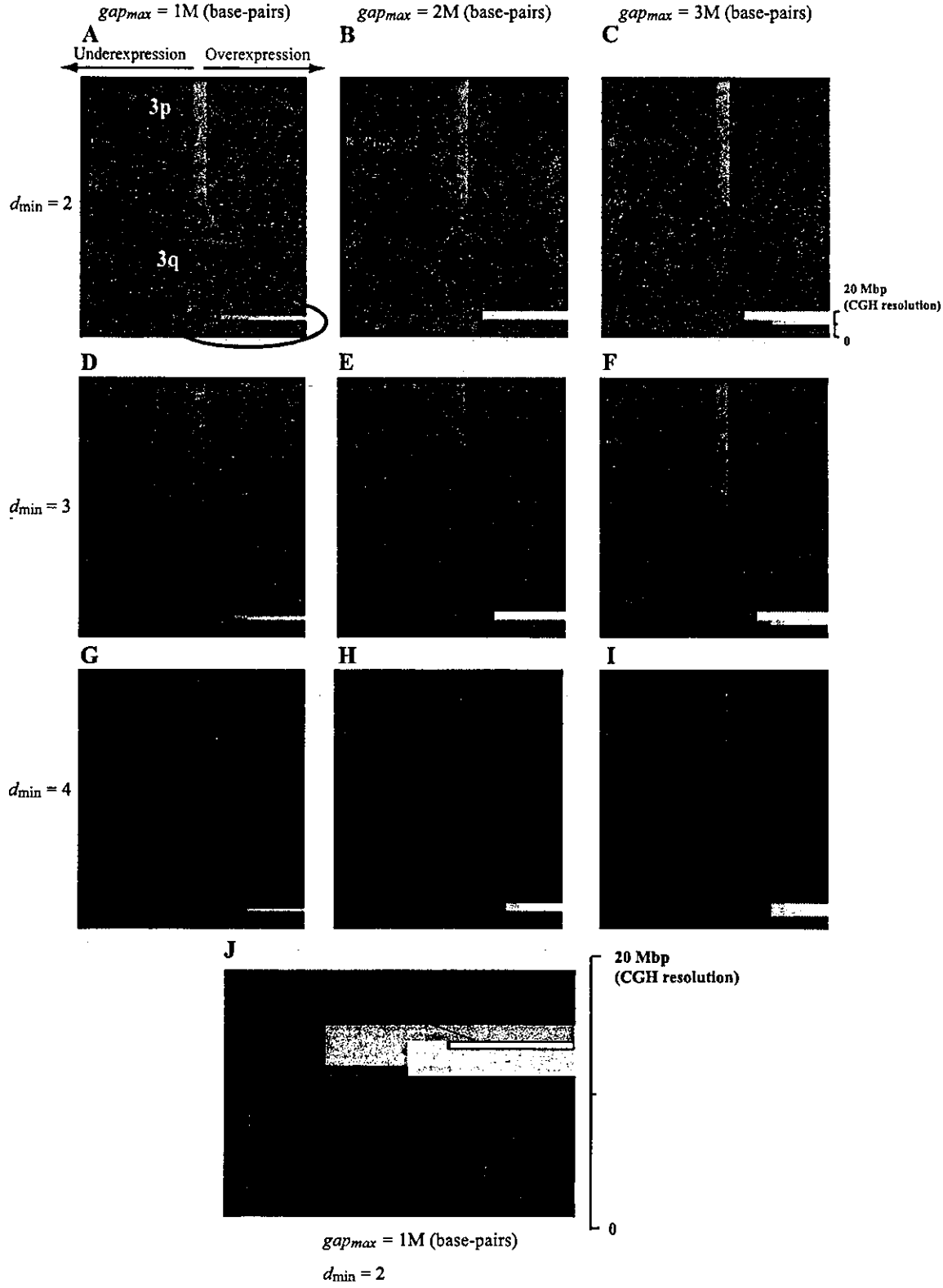


Table 3. *Clusters of probes with under- or overexpression profiles in each squamous cell lung carcinoma*

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.

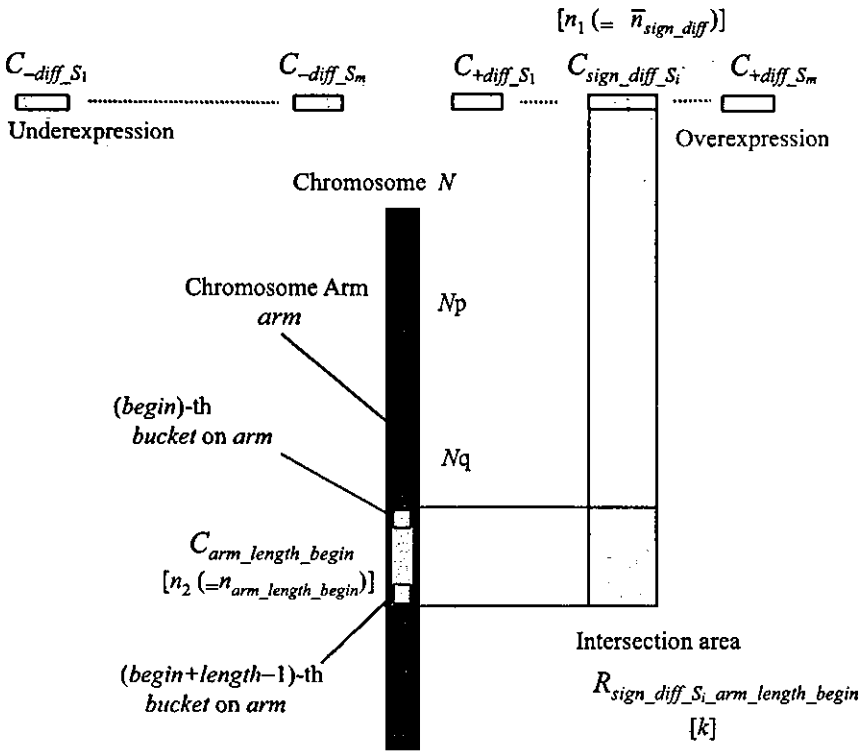


Fig. 8. Individual-specimen clusters vs. locus clusters. In a manner similar to the EIM for detecting expression imbalance of SQ specimen group, that for detecting individual differences in expression imbalance among SQ specimens was also constructed. In a SQ specimen S_i (i = 1, 2,. . ., 21), probes 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

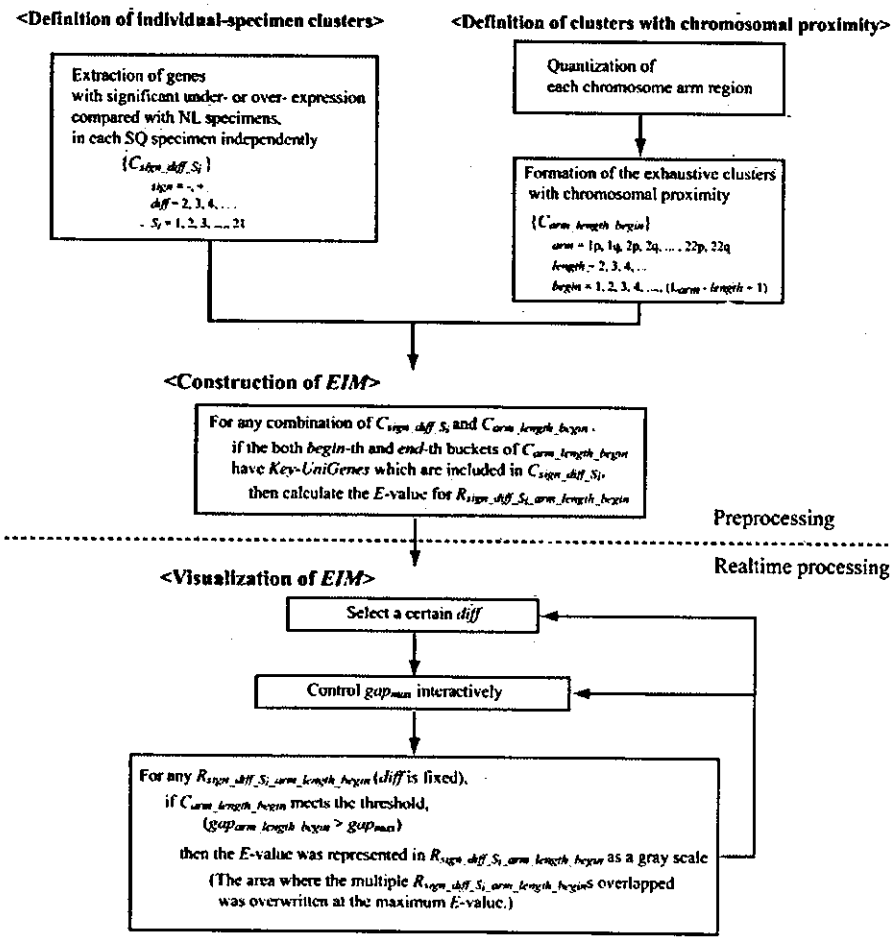


Fig. 9. Flowchart for construction of the EIM for detecting individual differences in expression imbalance among SQs. This flowchart provides details of the steps of the EIM for detecting individual differences in expression imbalance among SQs. For the step of "Definition of clusters with chromosomal 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.

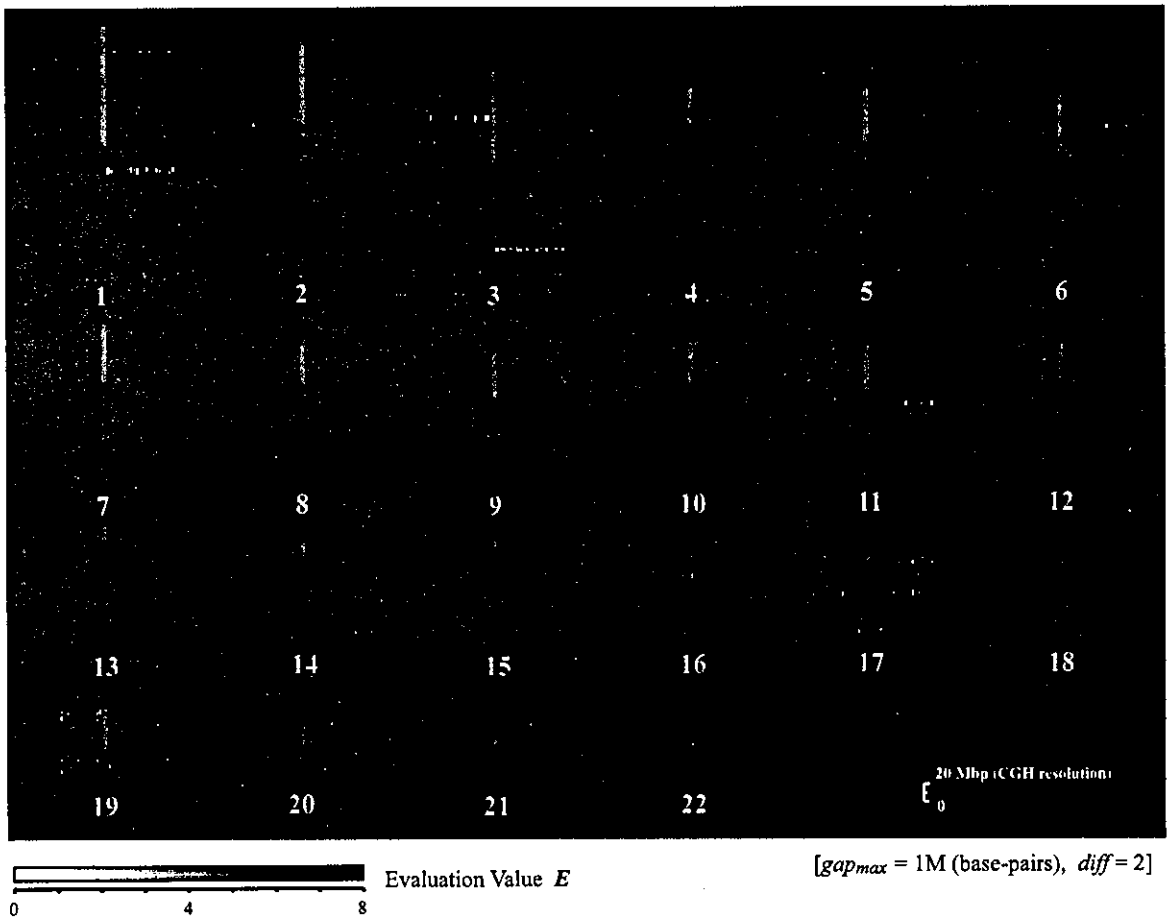
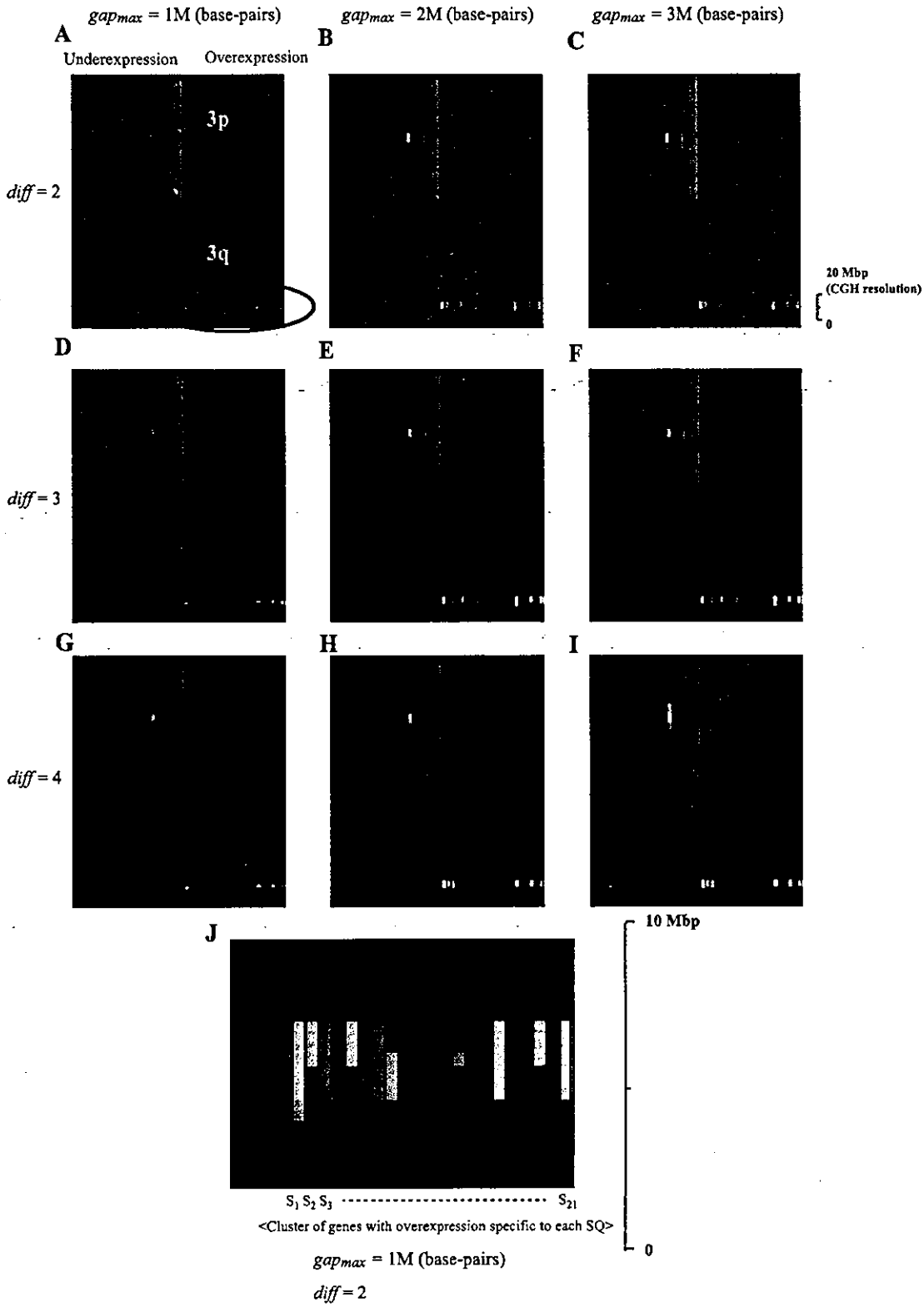


Fig. 10. The EIM for detecting individual difference of expression imbalance among SQs. The EIM was applied for detecting individual differences of expression imbalance among the SQs. Regions of underexpression and overexpression were visualized on the *left* and *right* side, respectively, as gray regional 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).



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).

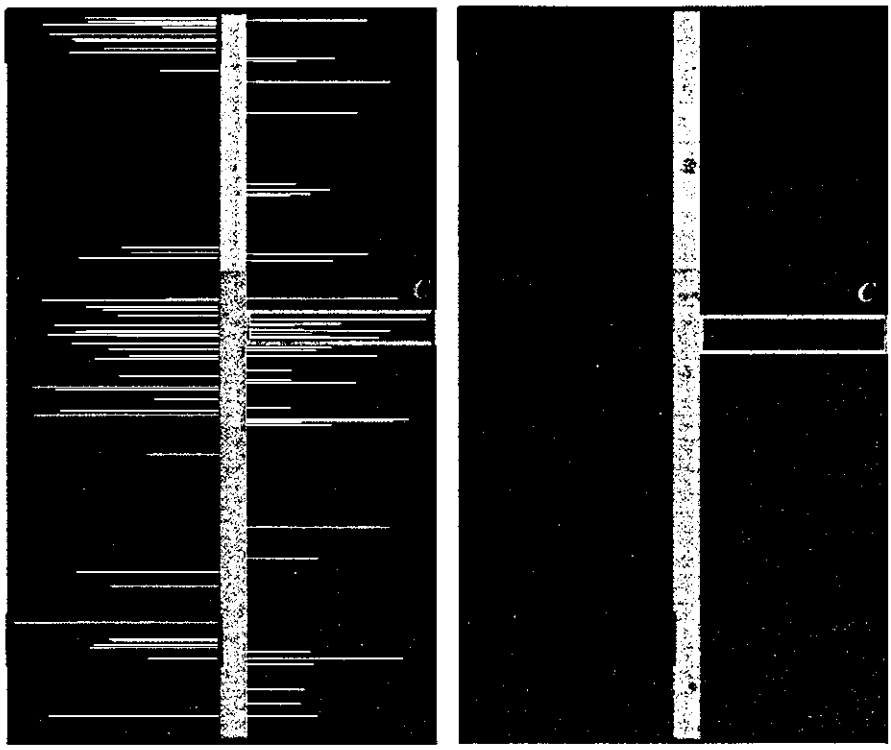
The high-resolution spatial map of expression profiles described in this report, i.e., the EIM, has several significant advantages. Its validity is clearly shown by the fact that many known loci with high frequent genomic losses or gains were detected by regional signals obtained with high resolution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