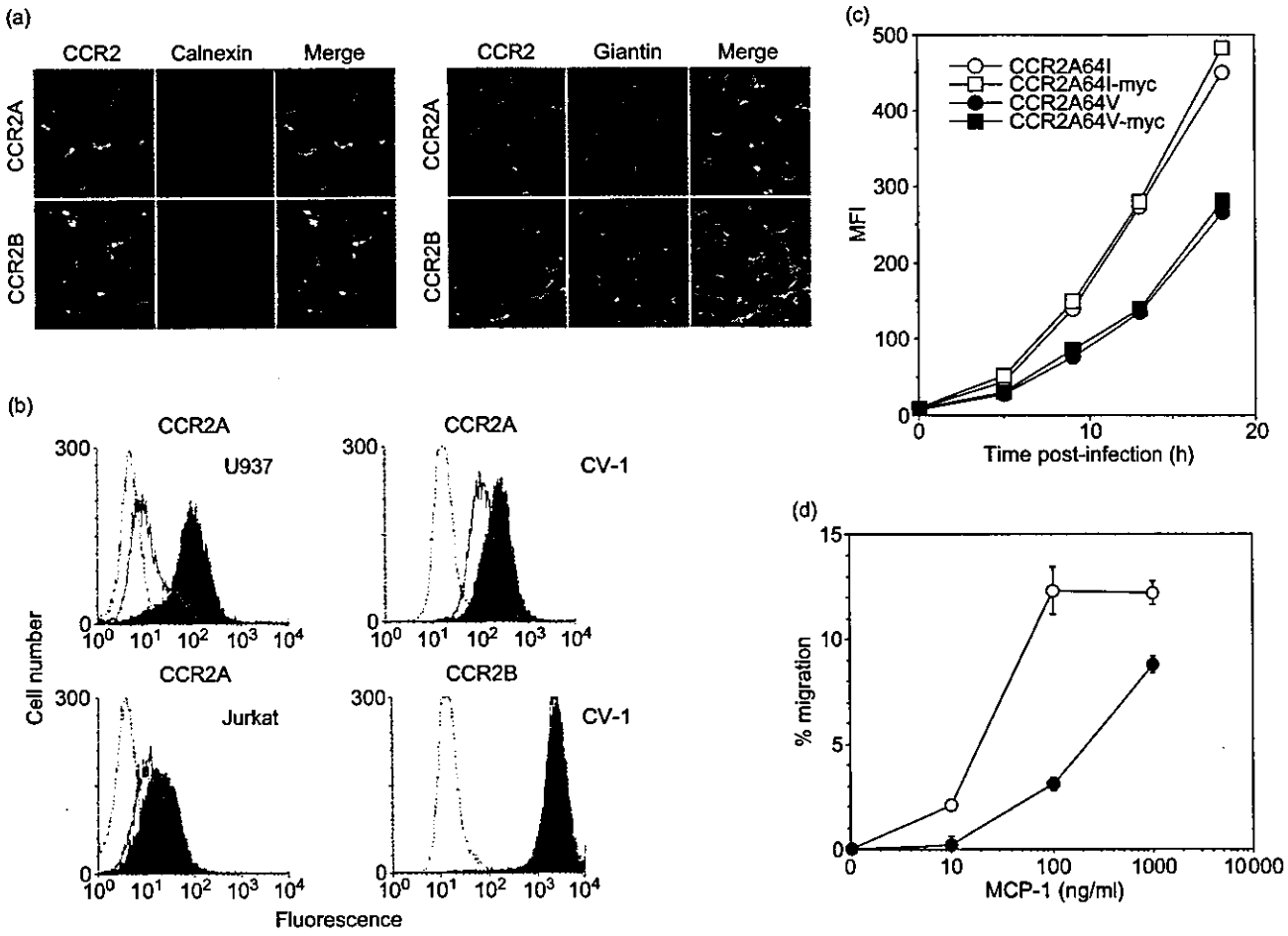


Fig. 2. (a) Subcellular distribution of CCR2A-64V and CCR2B-64V in CV1 cells. SeV vector (SeV) was used to express the CCR2A-64V and CCR2B-64V molecules. Cells were fixed and permeabilized before staining with MAB150 anti-CCR2 mouse MAb followed by FITC-labelled anti-mouse IgG. Cells were then re-stained with anti-calnexin or anti-giantin rabbit polyclonal antibody followed by Cy5-labelled anti-rabbit IgG, and analysed by confocal laser microscopy. (b) Surface expression of CCR2A-64V (green) and CCR2A-64I (red) in U937, CV1 or Jurkat cells. Cells infected with the parental Z strain served as a negative control (black). In lower right panel, green and red indicates CCR2B-64V and CCR2B-64I, respectively. (c) The cell surface expression of CCR2A-64I (open circles), CCR2A-64I-myc (open squares), CCR2A-64V (filled circles), and CCR2A-64V-myc (filled squares) at 5, 9, 12 and 18 h after infection by SeV. MFI indicates mean fluorescence intensity of each sample. (d) Chemokine receptor activity of recombinant CCR2A-64V and CCR2A-64I. Jurkat cells infected with SeV expressing CCR2A-64V (closed circles) or CCR2A-64I (open circles) migrated in response to increasing concentration of MCP-1. Data points are means of triplicate determination with standard deviations.

more efficiently than those expressing CCR2A-64V. These results are in good agreement with the observation that expression of CCR2A-64I is higher than that of CCR2A-64V.

CCR2A-64I is more stable than CCR2A-64V

Differential levels of expression between CCR2A-64V and CCR2A-64I prompted us to compare the rate of degradation of those proteins in pulse–chase experiments. For this purpose, we used recombinant SeV expressing CCR2A-64V-myc or CCR2A-64I-myc. Comparison of immunoprecipitated materials from ^{35}S-labelled CV1 cells expressing CCR2A-64V-myc and CCR2A-64I-myc showed that almost identical levels of CCR2A-64V-myc and CCR2A-64I-myc proteins were synthesized during the 30-min labelling period (t = 0) (Fig. 3a). However, CCR2A-64V-myc proteins appeared to degrade more rapidly than CCR2A-64I-myc proteins. The half-life of CCR2A-64I-myc was approximately 90 min, whereas that of CCR2A-64V-myc was approximately 50 min in CV1 cells (Fig. 3b). More prominent results were obtained when we used U937 cells, as the half-life of CCR2A-64I-myc was approximately 60 min, whereas that of CCR2A-64V-myc was approximately 18 min in U937 cells. This finding is in a good agreement with the observation

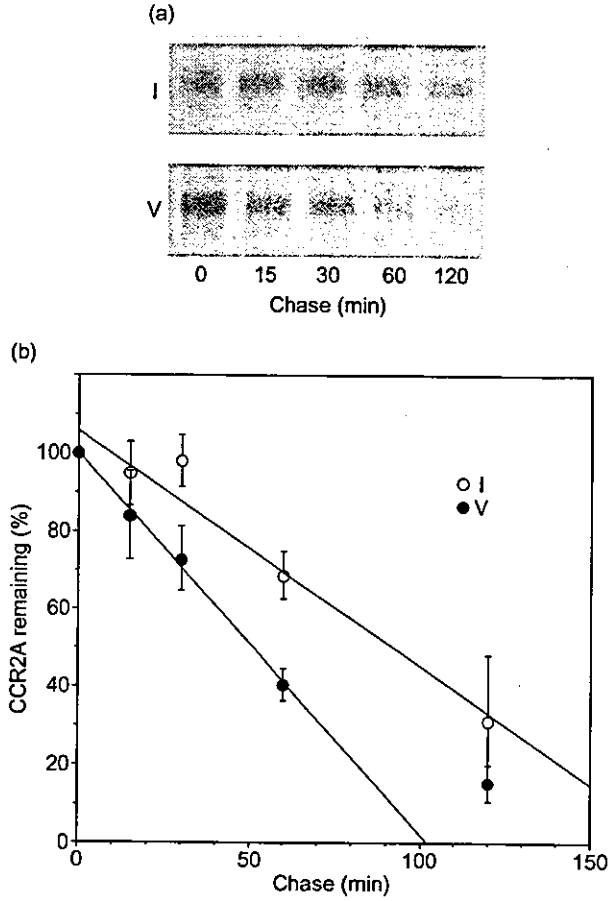


Fig. 3. CCR2A-64I is more stable than CCR2A-64V. CV1 cells were infected with SeV expressing CCR2A-64V-myc and CCR2A-64I-myc for 9 h. Cells were labelled for 30 min and then harvested following the chase time indicated. (a) Representative gels of pulse–chase analysis. (b) PhosphorImager analysis of the gels shown in (a). Open and closed circles denote cells infected with SeV expressing CCR2A-64V-myc and CCR2A-64I-myc, respectively. Data points are means of four independent experiments with standard deviations.

that the difference in cell surface expression levels between CCR2A-64V and CCR2A-64I was greater in U937 cells than in CV-1 cells (Fig. 2b). These results indicate that higher cell surface expression of CCR2A-64I was due to increased stability of CCR2A-64I. On the other hand, we failed to detect any significant difference in the half-life between CCR2B-64V and CCR2B-64I (data not shown).

CCR5 but not CXCR4 expression was more severely blocked by co-expression of CCR2A-64I than by co-expression of CCR2A-64V

To determine whether or not CCR2A has a dominant-negative effect on the expression of major HIV-1 receptor molecules, we first inoculated SeV expressing CCR2A-64V or CCR2A-64I in CV1 cells and incubated the cells for 9 h at 37°C. The cells were then superinfected with recombinant Vac expressing CCR5, CXCR4, or CD4. Five hours after Vac infection, surface expression of CCR5, CXCR4, or CD4 were examined by flow cytometry. As shown in Fig. 4a, the CCR5 MFI of cells co-infected with parental Z strain of SeV was 391, while that of the cells co-infected with SeV expressing CCR2A-64V was 297, indicating that co-expression of CCR2A-64V significantly reduced levels of CCR5 expression on the cell surface. This dominant-negative effect on CCR5 expression was more prominent when SeV expressing CCR2A-64I were used (MFI, 145) than SeV expressing CCR2A-64V were used. The same results were obtained when we used recombinant SeV expressing CCR2A-64V-myc and CCR2A-64I-myc (MFI, 300 and 179, respectively). Similar results were obtained when CV1 cells were inoculated with Vac expressing CCR5 5 h after infection by SeV expressing CCR2A, as the CCR5 MFI on cells co-infected with Z, SeV expressing CCR2A-64V, and Sev expressing CCR2A-64I, was 299, 205, and 160, respectively. Furthermore, the dominant-negative effect of CCR2A on CCR5 expression was also observed when T cell line H9 was used. The CCR5 MFI on H9 cells co-infected with Z, SeV expressing CCR2A-64V, and SeV expressing CCR2A-64I was 263, 230 and 195, respectively. In contrast, the cell surface expression of CXCR4, another major co-receptor, as well as that of CD4, the main receptor of HIV-1, were not affected by CCR2A-64V or CCR2A-64I (Fig. 4a). In contrast with CCR2A, neither CCR2B-64V nor CCR2B-64I affected the surface expression of CCR5 (Fig. 4b).

HIV-1 coreceptor activity of CCR5 was more dramatically reduced by co-expression of CCR2A-64I than by co-expression of CCR2A-64V

To assess the effect of CCR2A-64I on HIV-1 infection, we examined the ability of cells expressing both CCR2A and CCR5 molecules to support CD4-dependent cell fusion mediated by an HIV-1 envelope protein of the R5 strain SF162. For this purpose, we prepared CV1 cells expressing both CCR5 and CCR2A as described in Fig. 4a, and mixed those cells with mouse L cells expressing HIV-1 envelope protein. As shown in Fig. 5a, the envelope-mediated cell fusion activity of CCR5 was more dramatically reduced by co-expression of CCR2A-64I than by that of CCR2A-64V.

We also inoculated a live SF162 strain of HIV-1 into CD4 positive MT4 cells expressing both CCR5 and CCR2A. As shown in Fig. 5b, MT4 cells expressing CCR5 and CCR2A-64V supported SF162 replication better than those expressing CCR5 and CCR2A-64I.

Co-immunoprecipitation of CCR2A and CCR5

Many seven-transmembrane receptors, including che-

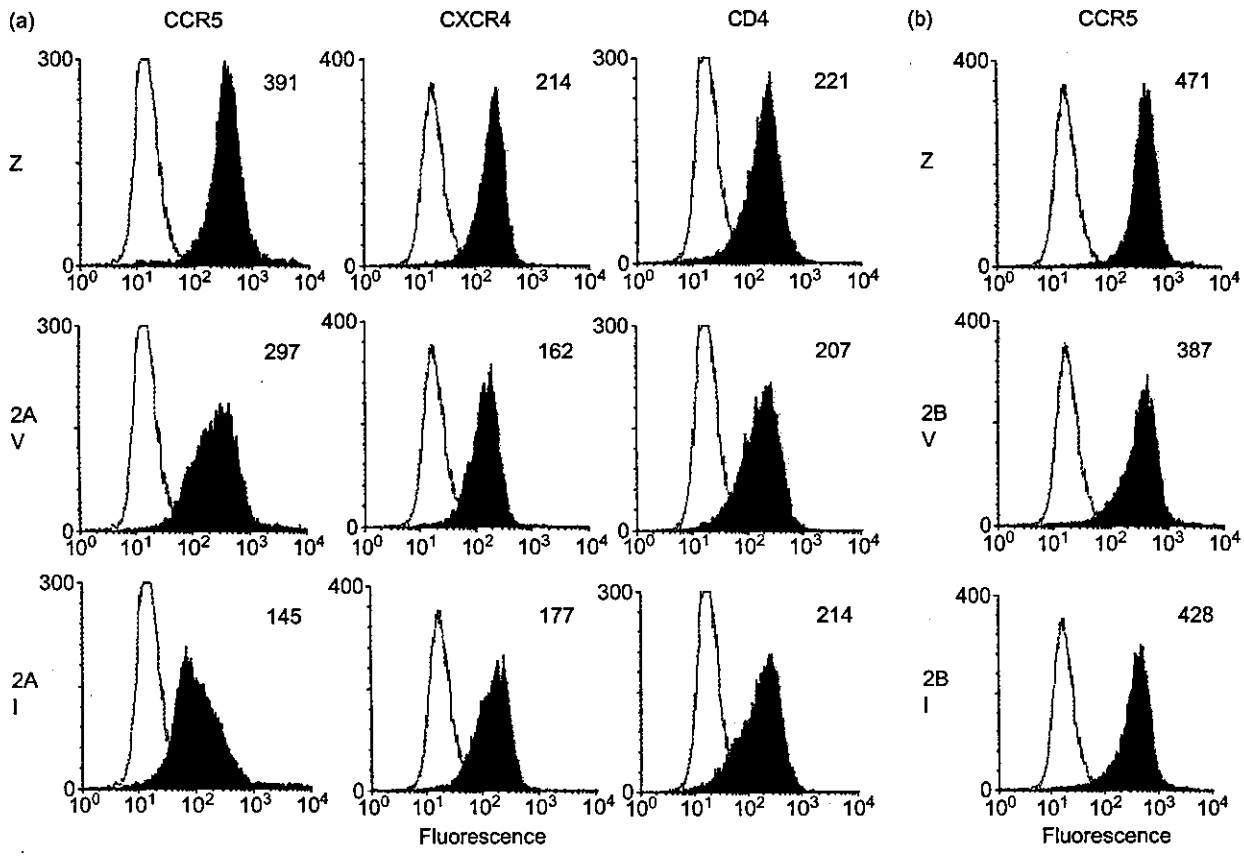


Fig. 4. (a) Effect of CCR2A-64V and CCR2A-64I on HIV-1 corecepter expression. Vac vectors were used to express CCR5, CXCR4 and CD4 in the CV1 cells inoculated with SeV expressing CCR2A-64V or CCR2A-64I. Z denotes the wild-type SeV. Five hours after infection, cells were stained with MAb against CCR5, CXCR4, or CD4. Flow cytometry was used to determine surface expression levels. The number in each panel indicates mean fluorescence intensity. (b) Effect of CCR2B-64V and CCR2B-64I on CCR5 expression.

mokine receptors, have been reported to form homo-oligomers. CCR2A is highly homologous to CCR5 (68% at the amino acid level), and formation of heterodimers between CCR2B and CCR5 was reported previously [20]. The dominant-negative effect of CCR2A on CCR5 expression shown in Figs 4a, 5a and 5b raised the possibility of heterodimer formation between CCR2A and CCR5. To test this hypothesis, we used SeV expressing CCR2A-64V-myc or CCR2A-64I-myc, and Vac expressing HA-tagged version of CCR5 (CCR5-HA). Anti-myc and anti-HA immunoprecipitates from cell lysates were developed in Western blots by using anti-HA or anti-myc antibodies. As expected, CCR5-HA was detected by anti-HA antibody in anti-myc-derived immunoprecipitates from CCR5-HA and CCR2A-64V-myc co-expressed cell lysates as well as from CCR5-HA and CCR2A-64I-myc co-expressed cell lysates. At the same time, CCR2A-64V-myc and CCR2A-64I-myc were detected by anti-myc antibody in anti-HA-derived immunoprecipitates of CCR5-HA and CCR2A-64V-myc co-expressed cell lysates and in that of CCR5-HA and CCR2A-64I-myc co-expressed cell lysates (Fig. 5c). These results clearly indicate that CCR2A formed heterodimers with CCR5.

In CCR5-HA expressing cells, we consistently observed two types of CCR5-HA molecules with different electrophoretic mobility. When we used anti-HA antibody to precipitate CCR5-HA directly, most of the CCR5-HA molecules migrated at approximately 38 kDa. In contrast, most of the CCR5-HA molecules that co-precipitated with CCR2A-64V-myc or CCR2A-64I-myc migrated at 37 kDa. We speculated that the CCR5-HA of 38 kDa represented authentic CCR5 molecules and that of 37 kDa represented immature forms of CCR5. To verify the maturation process of CCR5, we labelled the cells infected with Vac expressing CCR5-HA by [^{35}S]-methionine for 30 min and harvested those cells following chase periods ranging from 15 to 60 min. As shown in Fig. 5d, the 37-kDa CCR5-HA could be detected only after the labelling period (0 min). This result suggests that CCR2A binds to premature forms of CCR5 and

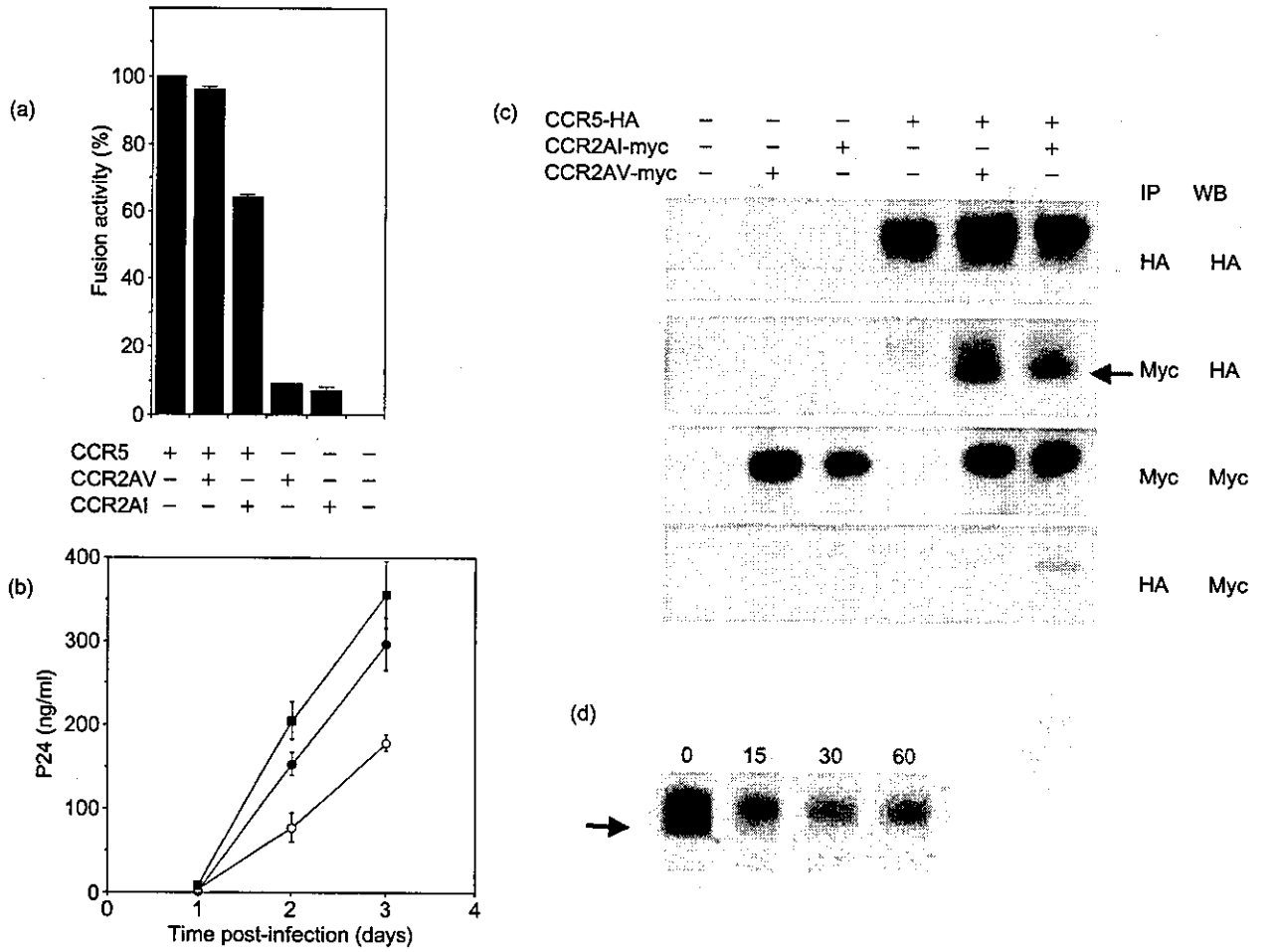


Fig. 5. (a) Coreceptor activity of CCR5 in CCR2A-64V or CCR2A-64I co-expressed cells. SeV vector was used to express CCR2A-64V or CCR2A-64I, and Vac vector was used to express CCR5 as described in Fig. 4. HIV-1 coreceptor activity of each sample was measured using the method described in Materials and methods. The wild-type Vac WR strain was used as a CCR5-negative control, and the wild-type SeV Z strain was used as the CCR2A-negative control. (b) MT4 cells were co-infected with SeV expressing CCR5 and SeV expressing CCR2A-64V (filled circles), CCR2A-64I (open circles), or parental Z strain (filled squares). Five hours after infection, cells were inoculated with an HIV-1 strain SF162. (c) Co-immunopresipitation of CCR2A and CCR5. Recombinant Vac expressing CCR5-HA or parental WR strain (−) was superinfected in CV1 cells infected with SeVs expressing CCR2A-64V-myc, CCR2A-64I-myc, or the parental Z strain (−). Immunoprecipitation and Western blot analysis were performed by using anti-HA or anti-myc antibody. An arrow indicates 37-kDa CCR5-HA molecules. (d) Pulse–chase analysis of CCR5 molecules. A recombinant Vac expressing CCR5-HA was inoculated into CV1 cells. An arrow indicates 37-kDa CCR5-HA molecules.

interferes with the maturation process of CCR5 molecules in cytoplasm.

Discussion

Many independent cohort studies have affirmed the AIDS-delaying effects of the *CCR2-64I* allele [4–8], but the molecular mechanism of this protective effect had not yet been elucidated. In the present study, we demonstrated that a valine to isoleucine substitution at position 64 increased stability of CCR2A but not of CCR2B molecules in cells. When co-expressed with the major HIV-1 co-receptor CCR5, CCR2A-64I more severely interfered with cell surface expression as well as HIV-1 co-receptor activity of CCR5 than CCR2A-64V. Furthermore, CCR2A was shown to co-precipitate with immature form of CCR5. These results suggest that CCR2A binds to CCR5 in the cytoplasm and dominantly interferes with CCR5 maturation and surface expression. On the other hand, the 64I substitution did not affect the level of CCR2B expression, being consistent with results published previously [9,10]. We speculate that increased ability of CCR2A-64I to down modulate CCR5 expression

might be a possible cause of delay in HIV-1 disease progression in patients with this allele. Alternatively, it is also possible that immune cell trafficking and/or signalling might be affected by CCR2A stabilization, leading to a delay in HIV-1 diseases.

Previously, Mellado *et al.* reported that CXCR4 could dimerize with CCR2B-64I variants but not with wild-type CCR2B-64V upon stimulation with SDF-1 and MCP-1. Based on this finding, they proposed that this ability of CCR2B-64I to heterodimerize with CXCR4 may cause a delay in AIDS progression [20]. However, several independent cohort studies have shown that the effects of the *CCR2-64I* allele were more pronounced in earlier stages of disease than in latter stages [5,8,21]. In a Dutch cohort, delay in HIV-1 disease progression was more pronounced before the emergence of X4 variants and was not observed after the emergence of X4 variants in individuals with the CCR2-64I allele [6]. Therefore, it is unlikely that CCR2B-64I/CXCR4 heterodimerization is the main cause of delay in AIDS progression in individuals with CCR2-64I.

Previous studies exploring the oligomerization of chemokine receptors also yielded controversial results. Rodrigues-Frade *et al.* reported that CCR2B forms homodimers upon stimulation by MCP-1 [22]. Other studies, however, have shown that CCR5 [23,24] and CXCR4 [25] can form homodimers without any stimulation by their ligands. Although we did not test whether or not stimulation with MCP-1 and/or RANTES increases hetero-oligomer formation between CCR2A and CCR5, our present results support the latter model that chemokine receptors may form oligomers without stimulation by their ligands.

In addition to AIDS pathogenesis, the CCR2-64I allele was reported to be associated with lower risks of coronary artery calcification [26] and acute rejection in renal transplantation [27]. Our present results shed light onto possible mechanisms of the association of this allele with such diverse human phenotypes. It is now widely accepted that monocyte attachment to cardiovascular wall is the first event implicated in atherogenesis of coronary arteries [28,29]. Since monocytes are known to express both CCR2A and CCR2B [13], an increased stability of CCR2A resulting from the 64I substitution may interfere with the function of CCR2B in monocytes, leading to decreased monocyte invasion to cardiovascular walls. With respect to acute rejection in renal transplantation, CCR5 is known to play an important role in both rejection of renal transplantation [30] and experimental graft-versus-host disease models [31]. Therefore, it is possible that an increased ability of CCR2A-64I to interfere with CCR5 expression can cause a decreased frequency of acute rejection after renal transplantation in recipients with this allele.

Previous studies have failed to show a statistically significant difference in levels of CCR5 expression on stimulated or non-stimulated peripheral blood mononuclear cells between CCR2-64I homozygotes and CCR2-64V homozygotes [9,10,32], although a slight reduction was noted in CCR2-64I homozygotes. In fact, we also failed to observe a statistically significant reduction of CCR5 levels on peripheral CD4 cells of homozygotes of CCR2-64I (data not shown). CCR2 is reported to be expressed on monocytes/macrophages [33], basophils [34,35], B cells [36], NK cells [37], dendritic cells [38,39], and a limited population of T cells [40]. Although we observed very few CCR2 cells in peripheral blood mononuclear cells, Bartoli *et al.* reported that numerous mononuclear cells in tonsil expressed CCR2A [41]. It may be possible that specific cell types expressing both CCR2A and CCR5 in tonsil or lymph nodes play an important role in AIDS pathogenesis and are responsible for the delay in HIV-1 diseases observed in patients with CCR2-64I.

Acknowledgements

pGIT7 beta-gal was kindly supplied by E. Berger. We thank D. Chao for critical discussion and S. Bando for technical assistance.

Sponsorship: Supported by grants from the Human Science Foundation, the Ministry of Education, Culture, Sports, Science, and Technology, and the Ministry of Health, Labour and Welfare, Japan.

References

1. Doranz BJ, Rucker J, Yi Y, Smyth RJ, Samson M, Peiper SC, *et al.* **A dual-tropic primary HIV-1 isolate that uses fusin and the beta-chemokine receptors CKR-5, CKR-3, and CKR-2b as fusion cofactors.** *Cell* 1996, **85**:1149–1158.
2. Rucker J, Edinger AL, Sharron M, Samson M, Lee B, Berson JF, *et al.* **Utilization of chemokine receptors, orphan receptors, and herpesvirus- encoded receptors by diverse human and simian immunodeficiency viruses.** *J Virol* 1997, **71**:8999–9007.
3. Penton-Rol G, Cota M, Polentarutti N, Luini W, Bernasconi S, Borsatti A, *et al.* **Up-regulation of CCR2 chemokine receptor expression and increased susceptibility to the multitropic HIV strain 89.6 in monocytes exposed to glucocorticoid hormones.** *J Immunol* 1999, **163**:3524–3529.
4. Smith MW, Dean M, Carrington M, Winkler C, Huttley GA, Lomb DA, *et al.* **Contrasting genetic influence of CCR2 and CCR5 variants on HIV-1 infection and disease progression. Hemophilia Growth and Development Study (HGDS), Multicenter AIDS Cohort Study (MACS), Multicenter Hemophilia Cohort Study (MHCS), San Francisco City Cohort (SFCC), ALIVE Study.** *Science* 1997, **277**:959–965.
5. Kostrikis LG, Huang Y, Moore JP, Wolinsky SM, Zhang L, Guo Y, *et al.* **A chemokine receptor CCR2 allele delays HIV-1 disease progression and is associated with a CCR5 promoter mutation.** *Nat Med* 1998, **4**:350–353.
6. van Rij RP, de Roda Husman AM, Brouwer M, Goudsmit J, Coutinho RA, Schuitemaker H. **Role of CCR2 genotype in the clinical course of syncytium-inducing (SI) or non-SI human immunodeficiency virus type 1 infection and in the time to**

conversion to SI virus variants. *J Infect Dis* 1998, **178**: 1806–1811.

7. Ioannidis JP, Rosenberg PS, Goedert JJ, Ashton LJ, Benfield TL, Buchbinder SP, *et al.* **Effects of CCR5-Delta32, CCR2-64I, and SDF-1 3'A alleles on HIV-1 disease progression: An international meta-analysis of individual- patient data.** *Ann Intern Med* 2001, **135**:782–795.
8. Mulherin SA, O'Brien TR, Ioannidis JP, Goedert JJ, Buchbinder SP, Coutinho RA, *et al.* **Effects of CCR5-Delta32 and CCR2-64I alleles on HIV-1 disease progression: the protection varies with duration of infection.** *AIDS* 2003, **17**:377–387.
9. Lee B, Doranz BJ, Rana S, Yi Y, Mellado M, Frade JM, *et al.* **Influence of the CCR2-V64I polymorphism on human immunodeficiency virus type 1 coreceptor activity and on chemokine receptor function of CCR2b, CCR3, CCR5, and CXCR4.** *J Virol* 1998, **72**:7450–7458.
10. Mariani R, Wong S, Mulder LC, Wilkinson DA, Reinhart AL, LaRosa G, *et al.* **CCR2-64I polymorphism is not associated with altered CCR5 expression or coreceptor function.** *J Virol* 1999, **73**:2450–2459.
11. Mummidi S, Ahuja SS, Gonzalez E, Anderson SA, Santiago EN, Stephan KT, *et al.* **Genealogy of the CCR5 locus and chemokine system gene variants associated with altered rates of HIV-1 disease progression.** *Nat Med* 1998, **4**: 786–793.
12. Charo IF, Myers SJ, Herman A, Franci C, Connolly AJ, Coughlin SR. **Molecular cloning and functional expression of two monocyte chemoattractant protein 1 receptors reveals alternative splicing of the carboxyl-terminal tails.** *Proc Natl Acad Sci USA* 1994, **91**:2752–2756.
13. Wong LM, Myers SJ, Tsou CL, Gosling J, Arai H, Charo IF. **Organization and differential expression of the human monocyte chemoattractant protein 1 receptor gene. Evidence for the role of the carboxyl-terminal tail in receptor trafficking.** *J Biol Chem* 1997, **272**:1038–1045.
14. Sanders SK, Crean SM, Boxer PA, Kellner D, LaRosa GJ, Hunt SW, 3rd. **Functional differences between monocyte chemotactic protein-1 receptor A and monocyte chemotactic protein-1 receptor B expressed in a Jurkat T cell.** *J Immunol* 2000, **165**:4877–4883.
15. Louisirirotchanakul S, Liu H, Roongpisuthipong A, Nakayama EE, Takebe Y, Shioda T, *et al.* **Genetic analysis of HIV-1 discordant couples in Thailand: association of CCR2 64I homozygosity with HIV-1-negative status.** *J Acquir Immune Defic Syndr* 2002, **29**:314–315.
16. Kato A, Sakai Y, Shioda T, Kondo T, Nakanishi M, Nagai Y. **Initiation of Sendai virus multiplication from transfected cDNA or RNA with negative or positive sense.** *Genes Cells* 1996, **1**:569–579.
17. Shioda T, Nakayama EE, Tanaka Y, Xin X, Liu H, Kawana-Tachikawa A, *et al.* **Naturally occurring deletional mutation in the C-terminal cytoplasmic tail of CCR5 affects surface trafficking of CCR5.** *J Virol* 2001, **75**:3462–3468.
18. Shioda T, Kato H, Ohnishi Y, Tashiro K, Ikegawa M, Nakayama EE, *et al.* **Anti-HIV-1 and chemotactic activities of human stromal cell-derived factor 1alpha (SDF-1alpha) and SDF-1beta are abolished by CD26/dipeptidyl peptidase IV-mediated cleavage.** *Proc Natl Acad Sci USA* 1998, **95**:6331–6336.
19. Nakayama EE, Shioda T, Tatsumi M, Xin X, Yu D, Ohgimoto S, *et al.* **Importance of the N-glycan in the V3 loop of HIV-1 envelope protein for CXCR-4- but not CCR-5-dependent fusion.** *FEBS Lett* 1998, **426**:367–372.
20. Mellado M, Rodriguez-Frade JM, Vila-Coro AJ, de Ana AM, Martinez AC. **Chemokine control of HIV-1 infection.** *Nature* 1999, **400**:723–724.
21. Michael NL, Louie LG, Rohrbaugh AL, Schultz KA, Dayhoff DE, Wang CE, *et al.* **The role of CCR5 and CCR2 polymorphisms in HIV-1 transmission and disease progression.** *Nat Med* 1997, **3**:1160–1162.
22. Rodriguez-Frade JM, Vila-Coro AJ, de Ana AM, Albar JP, Martinez AC, Mellado M. **The chemokine monocyte chemoattractant protein-1 induces functional responses through dimerization of its receptor CCR2.** *Proc Natl Acad Sci USA* 1999, **96**: 3628–3633.
23. Benkirane M, Jin DY, Chun RF, Koup RA, Jeang KT. **Mechanism of transdominant inhibition of CCR5-mediated HIV-1 infection by ccr5delta32.** *J Biol Chem* 1997, **272**:30603–30606.
24. Issafras H, Angers S, Bulenger S, Blanpain C, Parmentier M, Labbe-Jullie C, *et al.* **Constitutive agonist-independent CCR5 oligomerization and antibody- mediated clustering occurring at physiological levels of receptors.** *J Biol Chem* 2002, **277**: 34666–34673.
25. Babcock GJ, Farzan M, Sodroski J. **Ligand-independent dimerization of CXCR4, a principal HIV-1 coreceptor.** *J Biol Chem* 2003, **278**:3378–3385.
26. Valdes AM, Wolfe ML, O'Brien EJ, Spurr NK, Gefter W, Rut. A, *et al.* **Val64Ile polymorphism in the C-C chemokine receptor 2 is associated with reduced coronary artery calcification.** *Arterioscler Thromb Vasc Biol* 2002, **22**:1924–1928.
27. Abdi R, Tran TB, Sahagun-Ruiz A, Murphy PM, Brenner BM, Milford EL, *et al.* **Chemokine receptor polymorphism and risk of acute rejection in human renal transplantation.** *J Am Soc Nephrol* 2002, **13**:754–758.
28. Ross R. **The pathogenesis of atherosclerosis: a perspective for the 1990s.** *Nature* 1993, **362**:801–809.
29. Hanke H, Lenz C, Finking G. **The discovery of the pathophysiological aspects of atherosclerosis–a review.** *Acta Chir Belg* 2001, **101**:162–169.
30. Segerer S, Cui Y, Eitner F, Goodpaster T, Hudkins KL, Mack M, *et al.* **Expression of chemokines and chemokine receptors during human renal transplant rejection.** *Am J Kidney Dis* 2001, **37**:518–531.
31. Murai M, Yoneyama H, Harada A, Yi Z, Vestergaard C, Guo B, *et al.* **Active participation of CCR5(+)CD8(+) T lymphocytes in the pathogenesis of liver injury in graft-versus-host disease.** *J Clin Invest* 1999,104:49-57.
32. Shieh B, Liau YE, Hsieh PS, Yan YP, Wang ST, Li C. **Influence of nucleotide polymorphisms in the CCR2 gene and the CCR5 promoter on the expression of cell surface CCR5 and CXCR4.** *Int Immunol* 2000, **12**:1311–1318.
33. Fantuzzi L, Borghi P, Ciolli V, Pavlakis G, Belardelli F, Gessani S. **Loss of CCR2 expression and functional response to monocyte chemotactic protein (MCP-1) during the differentiation of human monocytes: role of secreted MCP-1 in the regulation of the chemotactic response.** *Blood* 1999, **94**:875–883.
34. Ochensberger B, Tassera L, Bifrare D, Rihs S, Dahinden CA. **Regulation of cytokine expression and leukotriene formation in human basophils by growth factors, chemokines and chemotactic agonists.** Eur *J Immunol* 1999, **29**:11–22.
35. Iikura M, Miyamasu M, Yamaguchi M, Kawasaki H, Matsushima K, Kitaura M, *et al.* **Chemokine receptors in human basophils: inducible expression of functional CXCR4.** *J Leukoc Biol* 2001, **70**:113–120.
36. Frade JM, Mellado M, del Real G, Gutierrez-Ramos JC, Lind P, Martinez AC. **Characterization of the CCR2 chemokine receptor: functional CCR2 receptor expression in B cells.** *J Immunol* 1997, **159**:5576–5584.
37. Polentarutti N, Allavena P, Bianchi G, Giardina G, Basile A, Sozzani S, *et al.* **IL-2-regulated expression of the monocyte chemotactic protein-1 receptor (CCR2) in human NK cells: characterization of a predominant 3.4-kilobase transcript containing CCR2B and CCR2A sequences.** *J Immunol* 1997, **158**:2689–2694.
38. Sallusto F, Schaerli P, Loetscher P, Schaniel C, Lenig D, Mackay CR, *et al.* **Rapid and coordinated switch in chemokine receptor expression during dendritic cell maturation.** Eur *J Immunol* 1998, **28**:2760–2769.
39. Vanbervliet B, Homey B, Durand I, Massacrier C, Ait-Yahia S, de Bouteiller O, *et al.* **Sequential involvement of CCR2 and CCR6 ligands for immature dendritic cell recruitment: possible role at inflamed epithelial surfaces.** Eur *J Immunol* 2002, **32**:231–242.
40. Rabin RL, Park MK, Liao F, Swofford R, Stephany D, Farber JM. **Chemokine receptor responses on T cells are achieved through regulation of both receptor expression and signaling.** *J Immunol* 1999, **162**:3840–3850.
41. Bartoli C, Civatte M, Pellissier JF, Figarella-Branger D. **CCR2A and CCR2B, the two isoforms of the monocyte chemoattractant protein-1 receptor are up-regulated and expressed by different cell subsets in idiopathic inflammatory myopathies.** *Acta Neuropathol (Berl)* 2001, **102**:385–392.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Genomics 84 (2004) 185–192
GENOMICS
www.elsevier.com/locate/ygeno

Dynamic regulation of gene expression by the Flt-1 kinase and Matrigel in endothelial tubulogenesis

Satsuki Kobayashi,[a,b] Emi Ito,[c,d] Reiko Honma,[c,d] Yoshihisa Nojima,[b] Masabumi Shibuya,[a] Shinya Watanabe,[c] and Yoshiro Maru[a,e,*]

[a] *Division of Genetics, The Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo 108-8639, Japan*
[b] *Department of Medicine and Clinical Science, School of Medicine, University of Gunma, Maebashi, Gunma 371-8511, Japan*
[c] *Department of Clinical Informatics, Graduate School of Medicine and Dentistry, Tokyo Medical and Dental University, Bunkyo-ku, Tokyo 113-8519, Japan*
[d] *Japan Biological Informatics Consortium, Chuo-ku, Tokyo 104-0032, Japan*
[e] *Department of Pharmacology, Tokyo Women's Medical University, Shinjuku-ku, Tokyo 162-8666, Japan*

Received 4 November 2003; accepted 12 February 2004

Available online 21 March 2004

Abstract

A nontubulogenic endothelial cell line, NP31, can be transformed by the active form of the Flt-1 kinase (BCR-FLTm1) into Tb3 cells, which show a tubulogenic property only when cultured in Matrigel. By utilizing this strict dependence of NP31 on BCR-FLTm1 and Matrigel for experimental angiogenesis, we performed microarray analyses under several conditions and found 97 genes whose dynamically regulated profiles of gene expression are divided into nine groups, in two major clusters. In one major cluster, gene expression is interdependently regulated by BCR-FLTm1 or Matrigel. The second major cluster contains genes whose expression patterns under BCR-FLTm1 influence are reversed by Matrigel. Based on these gene expression patterns in NP31 driven by BCR-FLTm1 and/or Matrigel, we propose a model in which sequential and alternate stimulation by BCR-FLTm1 and Matrigel induces cooperative regulation of subsets of genes. Microarray analyses of Tb3 under 11 different conditions revealed 5 candidate genes whose gene expression regulation is most closely associated with tubulogenesis.

Keywords: VEGF-R-1; Flt-1; VEGF; Endothelium; Vascular; Capillaries; Angiogenesis; Neovascularization; Microarray analysis of gene expression; Matrigel

Angiogenesis is a complicated process in which new vessels sprout out of the preexisting ones. Among numerous angiogenic molecules documented so far, one of the most essential factors is the vascular endothelial growth factor (VEGF) [1]. VEGF is currently known to bind two distinct tyrosine kinase receptors (VEGF-Rs), KDR and Flt-1. Flt-1 binds not only VEGF but also placenta growth factor (PlGF), which heterodimerizes with VEGF under some conditions. Ligand-activated Flt-1 is capable of transphosphorylating KDR and vice versa. Thus, cross talk between those two receptors could be possible at both ligand and receptor levels [2]. The embryonic lethality of knockout mice of either of those VEGF-Rs has left a fundamental question of what their functions are in adults. In addition to the cross talk mentioned above, there is some uniqueness to Flt-1. While the affinity to VEGF is approximately 10 times higher in Flt-1 than in KDR, VEGF-induced autophosphorylation activity is much weaker in Flt-1 than in KDR [3]. Given the naturally occurring soluble form of Flt-1, which retains only the ligand-binding ability, an inhibitory function of Flt-1 on KDR has been proposed. However, we have also shown, by utilizing the molecularly engineered mice in which the tyrosine kinase domain of Flt-1 was specifically destroyed, that Flt-1 is also involved in pathological angiogenesis such as tumor progression [4]. A cDNA microarray analysis, which tried to profile the gene expression in network formation by human umbilical vein endothelial cells (HUVEC) in Matrigel, showed up-regulation of both PlGF and Flt-1 [5,6]. To simplify the multi-ligand/receptor system, we have previously established a constitutively activated ligand-independent form of the Flt-1 kinase (BCR-FLTm1) [7]. BCR-FLTm1 has a tubulogenic potential not only in endothelial cells but also in fibroblastic cells [7,8].

* Corresponding author. Fax: +81-35269-7417.
E-mail address: ymaru@research.twmu.ac.jp (Y. Maru).

0888-7543/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2004.02.009

The application of microarray analysis to endothelial cell biology has been reported [2,5,6,9]. In most of the cases HUVEC were stimulated by VEGF or by collagen in combination with VEGF. One of the difficulties with HUVEC is that their sensitivity to VEGF or collagen depends on cell conditions. In addition, HUVEC show capillary morphogenesis even in the absence of VEGF or other angiogenic growth factors. Furthermore, the multi-ligand/receptor system complicates the interpretation of the results unless gene knockout cells are utilized [2].

Here we have utilized the endothelial cell line NP31, established in our laboratory, and examined gene expression profiles when NP31 cells were stimulated by BCR-FLTm1 and/or Matrigel.

Results

Microarray analyses of endothelial NP31 cells

BCR-FLTm1 promoted nontubulogenic endothelial NP31 cells to differentiate into tubulogenic cells (Tb3) (Fig. 1). Morphological differences were observed between NP31 and Tb3 cells in normal cultures on type I collagen (Figs. 1a and 1d) and were most prominent when the cells were cultured in Matrigel (Figs. 1b and 1e). Even in the absence of any other growth factors (Fig. 1f), Tb3 cells were capable of forming endothelial tubules in Matrigel. However, Tb3 cells failed to show tubulogenesis in three-dimensional culture in type I collagen (data not shown). Therefore, we initially examined gene expression profiles that were specific to Tb3 cells cultured in Matrigel versus type I collagen. We determined that the expression levels of 97 of 3824 genes were significantly changed as follows (Fig. 2): To narrow down the group of genes whose expression levels changed specifically in response to the expression of BCR-FLTm1, NP31 and Tb3 cells were stimulated by Matrigel. We found that expression levels of 281 genes were altered by a factor of 1 ($\log_2$ ratio) in Tb3 cells in Matrigel over type I collagen (Fig. 2, *a*). Then we subtracted genes whose expressions were altered in NP31 cells by Matrigel stimulation (Fig. 2, *b*). To subtract further BCR-FLTm1-mediated effects on gene expression in the absence of Matrigel, we also performed a virtual microarray analysis between Tb3 and NP31 cells on type I collagen (Fig. 2, *c*). Eventually, we determined 97 of 3824 genes whose expression was significantly changed in a BCR-FLTm1- and/or Matrigel-specific manner (Fig. 2).

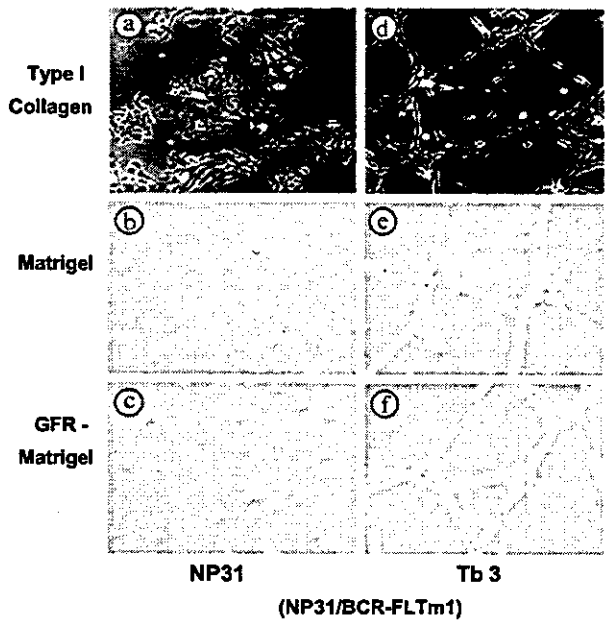


Fig. 1. Tubulogenesis by Tb3 cells. NP31 cells transformed by BCR-FLTm1 (Tb3) show morphologically distinguishable cell shapes on type I collagen plates from NP31 cells (compare a and d). (b and e) When plated onto Matrigel, Tb3 cells formed capillary-like networks (e), while the original NP31 cells remained aggregated (b). (c and f) Growth factor-reduced (GFR) Matrigel basically gave the same results as Matrigel.

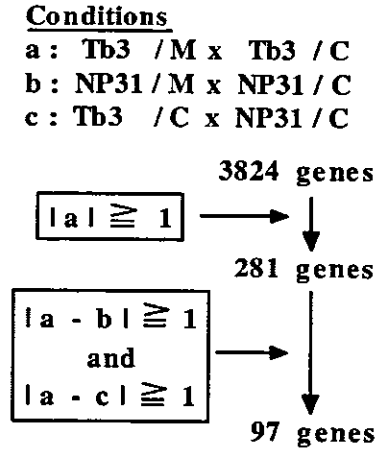


Fig. 2. Selection of genes whose expression levels were altered in a BCR-FLTm1- and/or Matrigel-specific manner. Three sets of $\log_2$ values (log ratio) are shown as (*a*) Tb3 on Matrigel over Tb3 on type I collagen, (*b*) NP31 on Matrigel over NP31 on type I collagen, and (*c*) Tb3 on type I collagen over NP31 on type I collagen. We extracted genes whose absolute value of log ratio was 1 and greater than 1 from the data set (*a*). Then we selected genes that satisfied the following conditions: (i) the absolute value of difference between data sets (*a*) and (*b*) for each gene was 1 and greater than 1 and (ii) the absolute value of difference between data sets (*a*) and (*c*) for each gene was 1 and greater than 1. Eventually, these operations gave 97 genes whose expression levels were altered in a BCR-FLTm1- and/or Matrigel-specific manner. The eventually selected data for 97 genes were subjected to hierarchical clustering analysis.

Forty-six genes, in red, with expression levels higher than 1 ($\log_2$ ratio) were determined to be up-regulated, and 51 genes, in blue, with expression levels lower than −1 ($\log_2$ ratio) were down-regulated and are shown in Fig. 3. Up-regulated genes are documented in detail in the literature and we found that 10 of the 46 up-regulated genes were already known to be involved in angiogenesis, which include MIP2 [10], Csf3, GM-CSF [11], Dpp4 [12], Egr1

[2,5,9,13,14], pJunB [15–17], ATF3 [18–20], ceruloplasmin [21,22], metallothionein [23,24], and NOS2 [25,26].

The reliability of the pattern of gene expression in the microarray was supported by quantitative reverse transcriptase- polymerase chain reaction (RT-PCR) analyses, and results of representative genes are shown in Fig. 4.

On the basis of hierarchical clustering analyses of microarray data, we grouped those 97 genes whose expression levels were altered in a BCR-FLTm1- and/or Matrigel-specific manner into nine categories (A–I in Fig. 3). Two major clusters (A–D and E–I), with an exception of R-97, were recognized as shown in the dendrogram in Fig. 3. The large clusters comprising the categories A–D and E–I included genes that were up- or down-regulated, respectively, when the extracellular matrix for Tb3 cells was changed from type I collagen to Matrigel to induce tubule formation (column (1)). In clusters A and E, neither BCR-FLTm1 nor Matrigel alone could induce alterations in gene expression. Therefore, those subsets of genes were interdependently regulated by BCR-FLTm1 and Matrigel. In clusters B, D, G, and I, the direction of regulation by either BCR-FLTm1 or Matrigel alone was the same as what we observed in column (1) and therefore the regulation of gene expression was cooperative but without strict interdependence. Interestingly, in clusters C, F, and H, the regulatory direction by at least one of two stimulations, BCR-FLTm1 or Matrigel, was reversed in column (1). For example, in cluster C (R-43), BCR-FLTm1 down-regulated its expression on type I collagen, which was reversed or up-regulated on Matrigel.

Trials to define components in Matrigel that are essential in tubulogenesis

Matrigel consists of growth factors and matrix proteins. To define the critical components that induce up- and down-regulation of gene expression observed in tubulogenesis by Tb3 cells, we stimulated Tb3 cells with individual factors, including laminin, fibronectin, type I collagen, epidermal growth factor (EGF), basic fibroblast growth factor (bFGF), nerve growth factor (NGF), platelet-derived growth factor (PDGF), insulin-like growth factor-1 (IGF-1), and transforming growth factor-β (TGF-β). Comparison of gene expression before and after stimulation of Tb3 cells by any single component, or by any single matrix protein in combination with a mixture of all of the growth factors, did not result in the gene expression pattern observed under Matrigel stimulation (data not shown, Fig. 5, (a), (b), and (c)). In addition, Tb3 cells displayed tubulogenesis in growth factor-reduced Matrigel, which gave a pattern of gene expression similar to that in the complete Matrigel, suggesting that growth factors described above are not critical and the BCR-FLTm1-derived signal alone may be sufficient in this biological system.

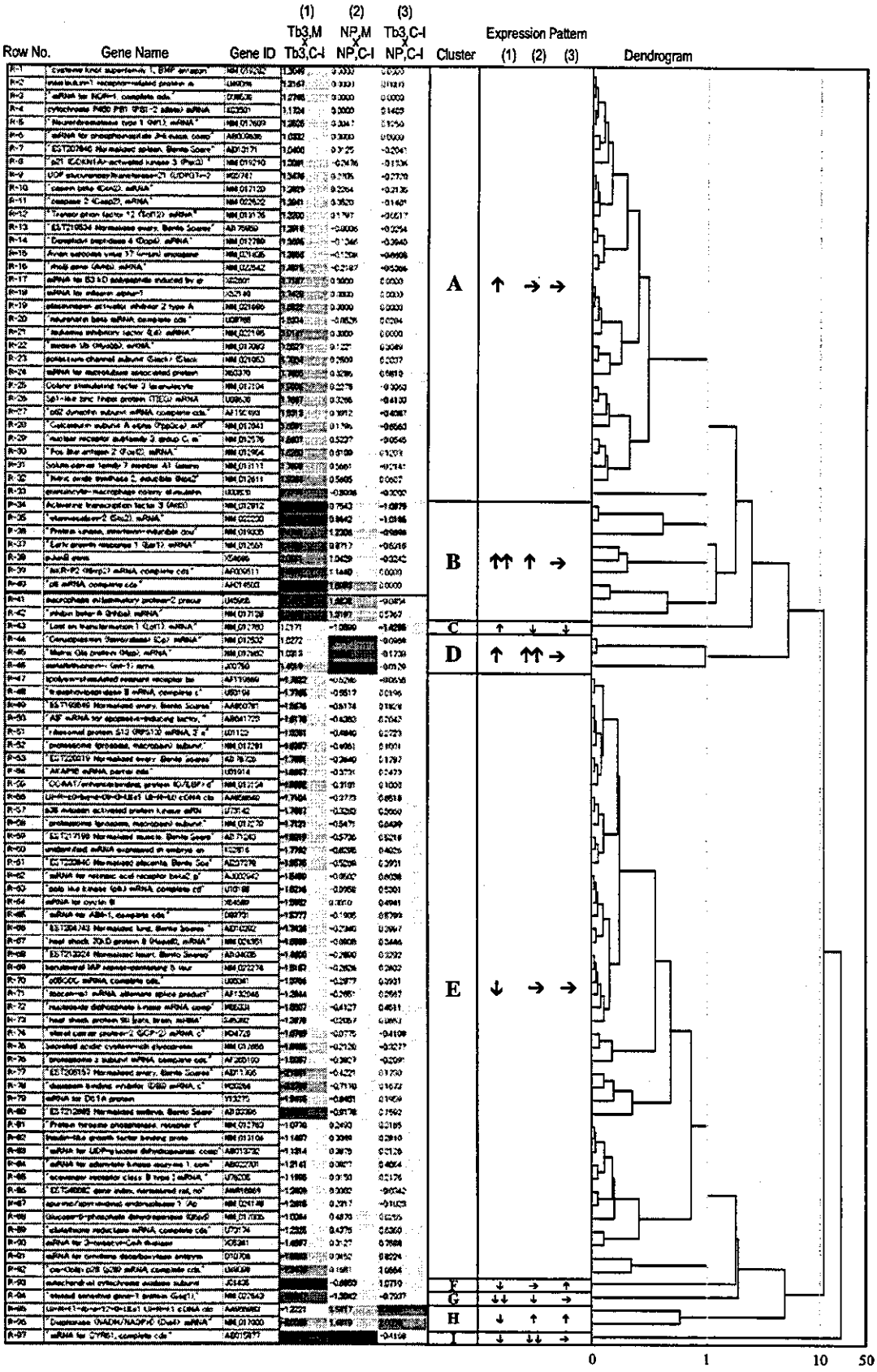
The remaining possibilities that promote the dynamic regulation of gene expression observed in Matrigel-treated Tb3 cells include unknown factors contained in Matrigel or certain secretory molecules produced in Tb3 cells after Matrigel stimulation. Therefore, we treated Tb3 cells cultured on type I collagen with supernatants collected from Matrigel culture with (Fig. 5 (k)) or without (Fig. 5 (e)) Tb3 cells. As shown in the dendrogram in Fig. 5, the gene expression pattern in growth factor-reduced Matrigel (f) was the closest to that in Matrigel (g). The second closest was the washed-out Matrigel (d) or type I collagen with supernatants from Matrigel culture in the absence of Tb3 cells. Considering that tubulogenesis was observed only under conditions (f) and (g), those data indicate that growth factors were not essential for tubulogenesis and that certain soluble factors from Matrigel, as well as what was left in Matrigel after wash, were indispensable to obtain the gene expression patterns in the complete Matrigel.

Selection of genes whose expression was altered only under tubulogenic conditions

Although we have failed to specify the critical component(s) in Matrigel that induces invading tubulogenesis, 11 conditions (Fig. 5, columns (a)–(k)) that were applied to Tb3 cultures to profile gene expression gave us a chance to narrow down genes whose expression profile was found only under tubulogenic conditions, columns (f) and (g). We display the 97 genes that were selected in Fig. 3 in column g′ with red and blue arrows indicating up- and down-regulated genes, respectively (Fig. 5 (g′)). Tb3 cells could not form tubules under the rest of the conditions (data not shown). Among those 97 genes, we found the following genes whose expression levels were significantly altered only under the tubulogenic conditions ((f) and (g)) but not under others, in the order of registered name and number: matrix Gla protein (Mgp) (NM_012862), colony-stimulating factor 3 (Csf3) (NM_017104), A-kinase anchor protein 95 (U01914), cysteine-rich 61 (CYR61) (AB015877), and metallothionein-1 (mt-1) (J00750). Whatever their functions are in tubulogenesis, those 5 genes belong to clusters A, D, and E (with the exception of CYR61 in cluster I), in which gene expression was interdependently regulated by both BCR-FLTm1 and Matrigel.

Discussion

The application of microarray analyses showed that one of the molecular mechanisms underlying our endothelial tubulogenesis model is gene expression interdependently regulated by BCR-FLTm1 and Matrigel as observed in the typical clusters of A and E. There are also clusters of genes in which Matrigel reverses the effects of BCR-FLTm1 on gene expression. This could be performed by a certain transcriptional repressor(s). Among the up-regulated genes in those groups in Fig. 3 (1), ATF3 is the only transcription factor with a repressor function. The JunB gene was also up-regulated by Matrigel and belongs to cluster B under our


(1) Tb3,M / Tb3,C-I
(2) NP,M / NP,C-I
(3) Tb3,C-I / NP,C-I
Row No.
Gene Name
Gene ID
Cluster
Expression Pattern
(1) (2) (3)
Dendrogram
A
B
C
D
E
F
G
H
I
0
1
10
50

criteria. In stress responses such as ionizing radiation, p38 and JNK activate promoters of ATF3 and c-Jun through Jun/ATF sites to simultaneously up-regulate their transcription, allowing ATF3 to modulate growth arrest [27]. It may be an ideal experiment to test if ATF3 expression in an inducible expression system would reverse the BCR-FLTm1-induced gene expression in cluster F (R-93). Another candidate may be p8, a stress-associated protein with a DNA-binding ability. The property of p8 to regulate the cell cycle negatively may be related to the behavior of quiescent endothelial cells that stop proliferation during tubulogenesis. Increased activities of cyclin-dependent kinase (CDK) 2 and CDK4 in P8-deficient cells are associated with down-regulation of a CDK inhibitor, p27, which has been shown to be important in growth arrest due to cell-to-cell contact [28]. Tb3 cells are weakly transformed, as they give small colonies in soft agar [8]. However, this growth-promoting activity is suppressed in Matrigel once the tubulogenesis is accomplished [8]. This could be reflected in the oppositely regulated gene expression profiles in the clustered genes in C, F, and H.

We suppose that Tb3 cells have three essential biological properties necessary for angiogenesis, growth, invasion, and differentiation with cell cycle arrest, and therefore represent the leading edge of the angiogenic tubules where endothelial cells are growing by differentiating. Based on the results of microarray analyses described above, we assume that sequentially occurring alternate stimuli of the active VEGF-R kinase (BCR-FLTm1) and Matrigel work in concert to up- and down-regulate subsets of genes in this simplified biological system (Fig. 6). The BCR-FLTm1-activated cells acquire a growth advantage and concomitant production and activation of matrix metalloprotease 2 (MMP2) as previously described [8], which supposedly enables cells immediately before cell division to escape from Matrigel contact and then, after cell division, to regain contact with Matrigel. We propose the idea of an angiogenic unit that is a sequence of events in the order of (1) BCR-FLTm1 (VEGF-R activation)–(2) MMP2 (matrix degradation)–(3) cell division (generation of new cells that participate in tubule formation)–(4) Matrigel. One round of the angiogenic unit consists of two phases. Phase 1 could be represented by the expression profile shown in Fig. 3 (3) and phase 2 by Fig. 3 (1) (Fig. 6). In phase 2 both up- and down-regulated genes were found. In either case cooperation between BCR-FLTm1 and Matrigel was observed. Interestingly, BCR-FLTm1-regulated genes in phase 1 (clusters C, F, H, and I) are oppositely regulated by Matrigel in phase 2. For one round to be sequentially followed by the next round of angiogenic unit, genes both up- and down-regulated by Matrigel need to be returned back to the baseline level of gene expression and to acquire sensitivity again to both stimuli. Molecular events that underlie this process still remain uncovered.

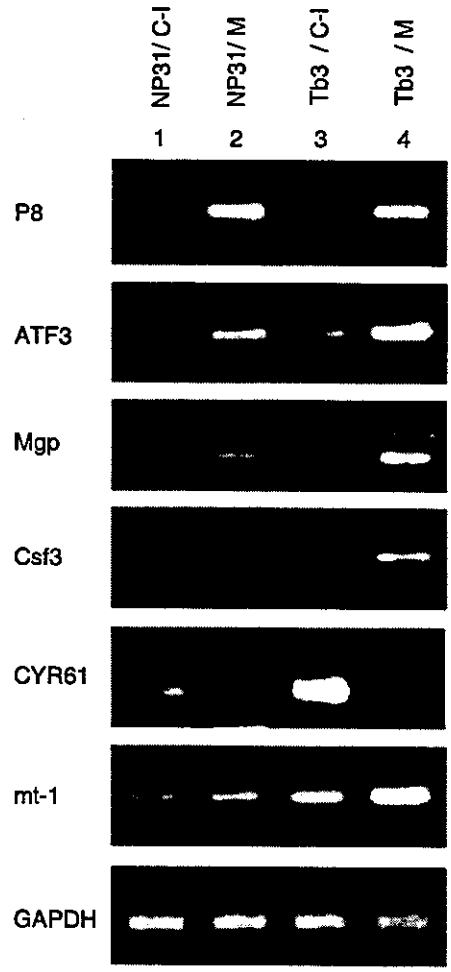


Fig. 4. RT-PCR analysis of representative genes in microarray. Total RNAs from NP31 cells (lanes 1 and 2) or Tb3 cells (lanes 3 and 4) on type I collagen (C–I) (lanes 1 and 3) or on Matrigel (M) (lanes 2 and 4) were subjected to RT-PCR analyses for P8 (R-40, group B in Fig. 3), ATF3 (R-34, group B), matrix Gla protein (Mgp) (R-45, group D), colony-stimulating factor 3 (Csf3) (R-25, group A), cysteine-rich 61 (CYR61) (R-97, group I), metallothionein-1 (mt-1) (R-46, group D), and control GAPDH.

We initially thought that extracellular matrix proteins such as collagen or laminin might be responsible for this restoration. However, stimulation by matrix proteins gave transcriptional patterns that are totally different from that by Matrigel. High-affinity integrins are reported to be recruited to the leading edge of the angiogenic tubule. However, once cell-to-matrix contacts are achieved (+/+ status shown in

Fig. 3. Dendrogram and grouping of 97 selected genes based on clusters. Hierarchical clustering analysis of 97 selected genes described in Fig. 2 is shown. Nine clusters (A–I) in which expression patterns differ in columns (1)–(3) are shown. For example, in cluster A, neither BCR-FLTm1 (3) nor Matrigel (2) altered the expression significantly, but the combined effect up-regulated expression of genes R-1 to R-33.

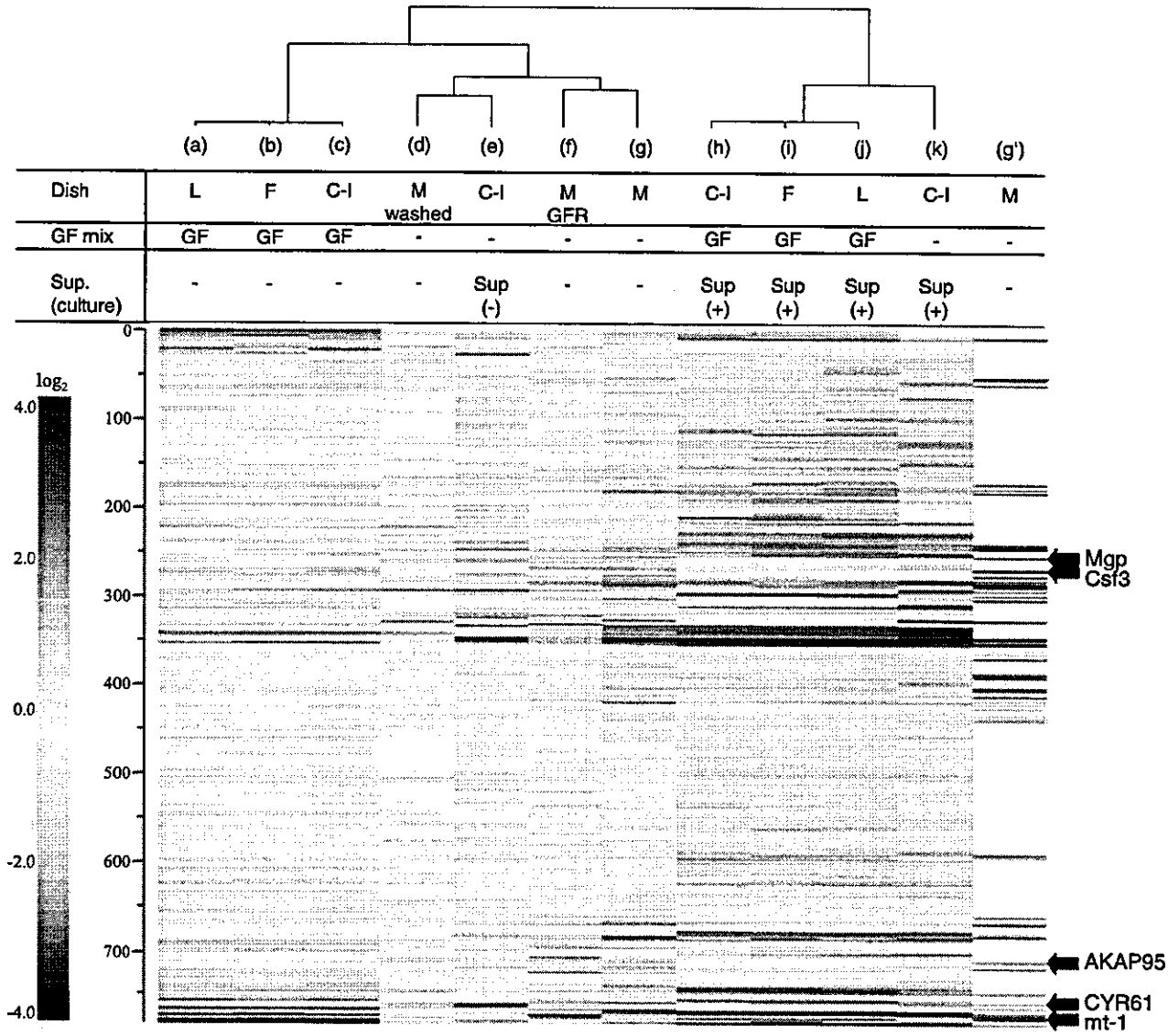


Fig. 5. Microarray analyses of Tb3 cells cultured under 11 different conditions. Tb3 cells were stimulated by a variety of conditions. Matrix proteins include laminin (L), fibronectin (F), type I collagen (C–I), Matrigel (M), and growth factor-reduced Matrigel (GFR). A mixture of growth factors (GF) or supernatants (Sup) collected from Matrigel cultures with (+) or without (−) Tb3 cells was added. Poly(A)$^+$ RNA extracted from Tb3 cells cultured under these 11 conditions and a common reference RNA extracted from Tb3 cells cultured on type I collagen were labeled with cyanine 5 and cyanine 3, respectively. The samples labeled with the two colors were mixed together and hybridized to microarrays (see also Materials and methods). A dendrogram representing the result of clustering is presented only for the sample conditions. Column g′ shows genes depicted in Fig. 3 as up- or down-regulated in comparison of Tb3 on Matrigel against Tb3 on type I collagen in purple and sky blue strips, respectively. Red and blue arrows indicate genes up- and down-regulated, respectively, only under conditions in columns f and g.

Fig. 6), there should be a counteracting system that inhibits integrin signaling, in other words, the existence of a factor(s) that can switch +/+ status to −/− in which integrin is inhibited and cells are susceptible to restimulation by angiogenic factors. Recently a dynamic control of integrin activation by class 3 semaphorins (SEMA3) has been shown. Production of SEMA3 in angiogenic factor-stimulated endothelial cells antagonizes integrin activation [29].

Although our dissection of Matrigel components by microarray analyses failed to specify any single known factor that makes an essential contribution to tubulogenesis, the genes specifically altered in common to columns (f) and (g) (associated with tubule formation) include Mgp, GM-CSF, CYR61, and mt-1. All of them have been reported to be not only produced in endothelial cells under certain conditions but also actively involved in angiogenesis [11,23,24,30–32].

It is reported that while VEGF stimulated tubule formation in human cerebral microvascular endothelial cells, PlGF alone was ineffective but augmented VEGF-driven tubulo-

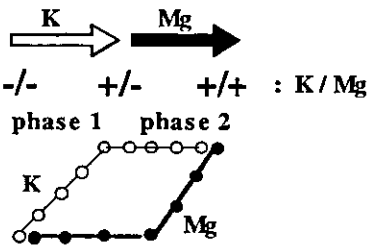


Fig. 6. Hypothetical angiogenic unit in tubule formation. A proposed model of dynamic regulation in gene expression in tubulogenesis is shown. BCR-FLTm1 (K for kinase) stimulates NP31 cells in phase 1, while there is no stimulation by Matrigel (Mg) yet (or Matrigel-activated signals are supposedly returned back to their baseline level) (−/− to +/−, an open arrow and open circles indicate an increase in K signals, while Mg signals, shown by a black arrow and closed circles, remain inactivated). In phase 2, Matrigel stimulates NP31 cells that are already stimulated by BCR-FLTm1 (+/− to +/+, a black arrow and closed circles indicate an increase in Matrigel signals, while BCR-FLTm1 signals, shown by open circles, remain activated). Along with those two stimulations, nine clusters of genes are coordinately and differentially up- and down-regulated.

genesis [2]. Studies with PlGF-knockout mice have revealed that PlGF and VEGF induce transcriptional profiles distinct from each other. The narrowed-down genes included Egr1, which has been reported to be a PlGF-regulated gene. Therefore, in addition to Flt-1-specific signaling, BCR-FLTm1 may substitute for KDR-mediated signaling in Tb3 cells.

Materials and methods

Cell cultures

NP31 and Tb3 cells were cultured as described before [8]. NP31 cells were established from sinusoidal endothelial cells of rat liver in primary culture by introduction of SV40 large T antigen. Although expression of endothelial markers such as VEGF receptors is retained in NP31 cells, they lost the property to form tubules in Matrigel for unknown reasons. Culture plates coated with type I collagen (Sumitomo, Akita, Japan), fibronectin, laminin, or Matrigel (Becton–Dickinson Bioscience) were purchased. Conditions to stimulate Tb3 cells were the following: PDGF 10 ng/ml (Genzyme/Techne), EGF 50 ng/ml (R&D), TGF-β 20 ng/ml (R&D), IGF 50 ng/ml (R&D), bFGF 10 ng/ml (Pepro Tech EC Ltd.), NGF 100 ng/ml (Takara, Shiga, Japan), or a mixture of those growth factors for 24 h. Culture supernatants were collected from Matrigel cultures in the presence or absence (Sup(−)) of Tb3 cells for 24 h. Matrigel was used as washed Matrigel after collection of Sup(−).

Microarray analysis and RT-PCR

A set of synthetic polynucleotides (80-mers) representing 3824 rat genes (MicroDiagnostic, Tokyo, Japan) was arrayed on a glass slide (S9115; Matsunami, Kishiwada, Japan) with a custom-made arrayer. Poly(A)$^+$ RNA was prepared from cells with Trizol reagent (Invitrogen, CA, USA) and a Poly(A) Purist Kit (Ambion, TX, USA), according to the manufacturer's instructions. Two micrograms of poly(A)$^+$ RNA was subjected to labeling with cyanine 5–dUTP or cyanine 3–dUTP (Perkin–Elmer, MA, USA). Labeling, hybridization, and subsequent washes of microarrays were performed with a Labeling & Hybridization Kit (MicroDiagnostic, Tokyo, Japan), according to the manufacturer's instructions. Hybridization signals were measured with a GenePix 400A scanner (Axon Instruments, CA, USA) and then processed into primary expression ratios (ratios of cyanine 5-labeled to cyanine 3-labeled samples) by GenePix Pro software (Axon Instruments). The primary expression ratios were converted into $\log_2$ values as secondary expression ratios ($\log_2$ ratio).

To determine genes whose expression is regulated in a BCR-FLTm1- and/or Matrigel-specific manner, microarray data were processed as follows: Poly(A)$^+$ RNA extracted from Tb3 cells cultured on Matrigel, NP31 cells cultured on Matrigel, and NP31 cells cultured on type I collagen was labeled with cyanine 5 (red) during first-strand cDNA synthesis. As a common reference sample, poly(A)$^+$ RNA extracted from Tb3 cells cultured on type I collagen was labeled with cyanine 3 (green). The cyanine 5-labeled sample was equally mixed with the cyanine 3-labeled common reference sample (Tb3 on type I collagen) and hybridized to microarrays containing 3824 rat genes. For each gene, an expression ratio against the common reference sample was calculated as a $\log_2$ value (log ratio). Next, to obtain virtual expression ratios of NP31 cells cultured on Matrigel against NP31 cells cultured on type I collagen, we subtracted log ratios derived from NP31 on type I collagen/Tb3 on type I collagen from log ratios derived from NP31 on Matrigel/Tb3 on type I collagen for all genes. Moreover, to obtain virtual expression ratios of Tb3 cells on type I collagen against NP31 cells on type I collagen, we reciprocally converted log ratios derived from NP31 on type I collagen/Tb3 on type I collagen for all genes. To eliminate influence of nonspecific hybridization signals from all data sets, we deleted genes that satisfied the following condition: the absolute value of the difference between individual log ratio and mean average of log ratios among three data sets for each gene was 0.5 or smaller.

Hierarchical clustering analysis of $\log_2$ ratios was performed with an MDI gene expression analysis software package (MicroDiagnostic, Tokyo, Japan).

Microarray analysis of Tb3 cells under 11 conditions was performed as follows: For each gene, an expression ratio against the common reference was calculated and converted into a $\log_2$ value. All 11 data sets were assembled together and subjected to the subsequent filtering operations. First, we selected genes whose absolute value of log ratio was 1 or greater in at least 1 of 11 data sets. Second, from the selected genes, we deleted genes that satisfied the following condition: the absolute value of the difference between individual

log ratio and mean average of log ratios among 11 data sets for each gene was 1 or less. After the completion of these operations, selected data were subjected to two-dimensional clustering analysis for sample conditions and genes.

Primers for RT-PCR analysis were, in the order of forward and reverse primers, AACAGGCAAGACTTTG-GAG and GTTGTACAGTTTATTGTTACTG for p8, CGAGCGAAGACTGGAGCAAAATGATG and GC-GGCCGCATTCAGTAAGGACTCCCCAATTG for ATF3, ACACCCGAGACCATGAAGAG and CTGCCTGAAG-TAGCGGTTGT for Mgp, CCTAGCAGGCATTTCCTCTG and GCCTTCTCTCTCTGCTCCAA for Csf3, CAAGAA-ATGCAGCAAGACCA and CCGGGCTCCAGTACTAT-GAA for CYR61, and CTGCCTTCTTGTCGCTTACA and GGAGGTGTACGGCAAGACTC for mt-1.

References

[1] N. Ferrara, H.P. Gerber, J. LeCouter, Nat. Med. 9 (2003) 669–676.
[2] M. Autiero, et al., Nat. Med. 9 (2003) 936–943.
[3] M. Shibuya, Cell Struct. Funct. 26 (2001) 25–35.
[4] S. Hiratsuka, Y. Maru, A. Okada, M. Seiki, T. Noda, M. Shibuya, Cancer Res. 61 (2001) 1207–1213.
[5] S.E. Bell, A. Mavila, R. Salazar, K.J. Bayless, S. Kanagala, S.A. Maxwell, G.E. Davis, J. Cell Sci. 114 (2001) 2755–2773.
[6] J. Kahn, et al., Am. J. Pathol. 156 (2000) 1887–1900.
[7] Y. Maru, S. Yamaguchi, M. Shibuya, Oncogene 16 (1998) 2585–2595.
[8] Y. Maru, H. Hirosawa, M. Shibuya, Eur. J. Cell Biol. 79 (2000) 130–143.
[9] M. Abe, Y. Sato, Angiogenesis 4 (2001) 289–298.
[10] J.A. Belperio, M.P. Keane, D.A. Arenberg, C.L. Addison, J.E. Ehlert, M.D. Burdick, R.M. Strieter, J. Leukocyte Biol. 68 (2000) 1–8.
[11] I.R. Buschmann, H.J. Busch, G. Mies, K.A. Hossmann, Circulation 30 (2003) 30.
[12] Z. Zukowska-Grojec, et al., Circ. Res. 83 (1998) 187–195.
[13] I.L. Szabo, R. Pai, B. Soreghan, M.K. Jones, D. Baatar, H. Kawanaka, A.S. Tarnawski, J. Physiol. Paris 95 (2001) 379–383.
[14] L.M. Khachigian, V. Lindner, A.J. Williams, T. Collins, Science 271 (1996) 1427–1431.
[15] L.E. Dike, D.E. Ingber, J. Cell Sci. 109 (Pt 12) (1996) 2855–2863.
[16] H. Shirotani-Ikejima, K. Kokame, T. Hamuro, G. Bu, H. Kato, T. Miyata, Biochem. Biophys. Res. Commun. 299 (2002) 847–852.
[17] T. Shono, M. Ono, H. Izumi, S.I. Jimi, K. Matsushima, T. Okamoto, K. Kohno, M. Kuwano, Mol. Cell. Biol. 16 (1996) 4231–4239.
[18] Y. Cai, C. Zhang, T. Nawa, T. Aso, M. Tanaka, S. Oshiro, H. Ichijo, S. Kitajima, Blood 96 (2000) 2140–2148.
[19] J. Kawauchi, C. Zhang, K. Nobori, Y. Hashimoto, M.T. Adachi, A. Noda, M. Sunamori, S. Kitajima, J. Biol. Chem. 277 (2002) 39025–39034.
[20] T. Yin, G. Sandhu, C.D. Wolfgang, A. Burrier, R.L. Webb, D.F. Rigel, T. Hai, J. Whelan, J. Biol. Chem. 272 (1997) 19943–19950.
[21] A. Bianchini, G. Musci, L. Calabrese, J. Biol. Chem. 274 (1999) 20265–20270.
[22] C.K. Mukhopadhyay, E. Ehrenwald, P.L. Fox, J. Biol. Chem. 271 (1996) 14773–14778.
[23] L.L. Pearce, K. Wasserloos, C.M. St Croix, R. Gandley, E.S. Levitan, B.R. Pitt, J. Nutr. 130 (2000) 1467S–1470S.
[24] M. Penkowa, J. Carrasco, M. Giralt, A. Molinero, J. Hernandez, I.L. Campbell, J. Hidalgo, J. Cereb. Blood Flow Metab. 20 (2000) 1174–1189.
[25] P. Rafice, et al., J. Biol. Chem. 277 (2002) 35605–35615.
[26] G. Garcia-Cardena, J. Folkman, J. Natl. Cancer Inst. 90 (1998) 560–561.
[27] J. Kool, M. Hamdi, P. Cornelissen-Steijger, A.J. Eb van der, C. Terleth, H. van Dam, Oncogene 22 (2003) 4235–4242.
[28] S. Vasseur, A. Hoffmeister, A. Garcia-Montero, G.V. Mallo, R. Feil, S. Kuhbandner, J.C. Dagorn, J.L. Iovanna, Oncogene 21 (2002) 1685–1694.
[29] G. Serini, et al., Nature 424 (2003) 391–397.
[30] S.J. Leu, Y. Liu, N. Chen, C.C. Chen, S.C. Lam, L.F. Lau, J. Biol. Chem. (2003).
[31] K.I. Bostrom, Z. Kardiol. 89 (Suppl 2) (2000) 69–74.
[32] S.J. Leu, S.C. Lam, L.F. Lau, J. Biol. Chem. 277 (2002) 46248–46255.

Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones

Tadashi Imanishi[1], Takeshi Itoh[1,2], Yutaka Suzuki[3,68], Claire O'Donovan[4], Satoshi Fukuchi[5], Kanako O. Koyanagi[6], Roberto A. Barrero[5], Takuro Tamura[7,8], Yumi Yamaguchi-Kabata[1], Motohiko Tanino[1,7], Kei Yura[9], Satoru Miyazaki[5], Kazuho Ikeo[5], Keiichi Homma[5], Arek Kasprzyk[4], Tetsuo Nishikawa[10,11], Mika Hirakawa[12], Jean Thierry-Mieg[13,14], Danielle Thierry-Mieg[13,14], Jennifer Ashurst[15], Libin Jia[16], Mitsuteru Nakao[3], Michael A. Thomas[17], Nicola Mulder[4], Youla Karavidopoulou[4], Lihua Jin[5], Sangsoo Kim[18], Tomohiro Yasuda[11], Boris Lenhard[19], Eric Eveno[20,21], Yoshiyuki Suzuki[5], Chisato Yamasaki[1], Jun-ichi Takeda[1], Craig Gough[1,7], Phillip Hilton[1,7], Yasuyuki Fujii[1,7], Hiroaki Sakai[1,7,22], Susumu Tanaka[1,7], Clara Amid[23], Matthew Bellgard[24], Maria de Fatima Bonaldo[25], Hidemasa Bono[26], Susan K. Bromberg[27], Anthony J. Brookes[19], Elspeth Bruford[28], Piero Carninci[29], Claude Chelala[20], Christine Couillault[20,21], Sandro J. de Souza[30], Marie-Anne Debily[20], Marie-Dominique Devignes[31], Inna Dubchak[32], Toshinori Endo[33], Anne Estreicher[34], Eduardo Eyras[15], Kaoru Fukami-Kobayashi[35], Gopal R. Gopinath[36], Esther Graudens[20,21], Yoonsoo Hahn[18], Michael Han[23], Ze-Guang Han[21,37], Kousuke Hanada[5], Hideki Hanaoka[1], Erimi Harada[1,7], Katsuyuki Hashimoto[38], Ursula Hinz[34], Momoki Hirai[39], Teruyoshi Hishiki[40], Ian Hopkinson[41,42], Sandrine Imbeaud[20,21], Hidetoshi Inoko[1,7,43], Alexander Kanapin[4], Yayoi Kaneko[1,7], Takeya Kasukawa[26], Janet Kelso[44], Paul Kersey[4], Reiko Kikuno[45], Kouichi Kimura[11], Bernhard Korn[46], Vladimir Kuryshev[47], Izabela Makalowska[48], Takashi Makino[5], Shuhei Mano[43], Regine Mariage-Samson[20], Jun Mashima[5], Hideo Matsuda[49], Hans-Werner Mewes[23], Shinsei Minoshima[50,52], Keiichi Nagai[11], Hideki Nagasaki[51], Naoki Nagata[1], Rajni Nigam[27], Osamu Ogasawara[3], Osamu Ohara[45], Masafumi Ohtsubo[52], Norihiro Okada[53], Toshihisa Okido[5], Satoshi Oota[35], Motonori Ota[54], Toshio Ota[22], Tetsuji Otsuki[55], Dominique Piatier-Tonneau[20], Annemarie Poustka[47], Shuang-Xi Ren[21,37], Naruya Saitou[56], Katsunaga Sakai[5], Shigetaka Sakamoto[5], Ryuichi Sakate[39], Ingo Schupp[47], Florence Servant[4], Stephen Sherry[13], Rie Shiba[1,7], Nobuyoshi Shimizu[52], Mary Shimoyama[27], Andrew J. Simpson[30], Bento Soares[25], Charles Steward[15], Makiko Suwa[51], Mami Suzuki[5], Aiko Takahashi[1,7], Gen Tamiya[1,7,43], Hiroshi Tanaka[33], Todd Taylor[57], Joseph D. Terwilliger[58], Per Unneberg[59], Vamsi Veeramachaneni[48], Shinya Watanabe[3], Laurens Wilming[15], Norikazu Yasuda[1,7], Hyang-Sook Yoo[18], Marvin Stodolsky[60], Wojciech Makalowski[48], Mitiko Go[61], Kenta Nakai[3], Toshihisa Takagi[3], Minoru Kanehisa[12], Yoshiyuki Sakaki[3,57], John Quackenbush[62], Yasushi Okazaki[26], Yoshihide Hayashizaki[26], Winston Hide[44], Ranajit Chakraborty[63], Ken Nishikawa[5], Hideaki Sugawara[5], Yoshio Tateno[5], Zhu Chen[21,37,64], Michio Oishi[45], Peter Tonellato[65], Rolf Apweiler[4], Kousaku Okubo[5,40], Lukas Wagner[13], Stefan Wiemann[47], Robert L. Strausberg[16], Takao Isogai[10,66], Charles Auffray[20,21], Nobuo Nomura[40], Takashi Gojobori[1,5,67*], Sumio Sugano[3,40,68]

1 Integrated Database Group, Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan, 2 Bioinformatics Laboratory, Genome Research Department, National Institute of Agrobiological Sciences, Ibaraki, Japan, 3 Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan, 4 EMBL Outstation—European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom, 5 Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Shizuoka, Japan, 6 Nara Institute of Science and Technology, Nara, Japan, 7 Integrated Database Group, Japan Biological Information Research Center, Japan Biological Informatics Consortium, Tokyo, Japan, 8 BITS Company, Shizuoka, Japan, 9 Quantum Bioinformatics Group, Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute, Kyoto, Japan, 10 Reverse Proteomics Research Institute, Chiba, Japan, 11 Central Research Laboratory, Hitachi, Tokyo, Japan, 12 Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto, Japan, 13 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, 14 Centre National de la Recherche Scientifique (CNRS), Laboratoire de Physique Mathematique, Montpellier, France, 15 The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom, 16 National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America, 17 Department of Biological Sciences, Idaho State University, Pocatello, Idaho, United States of America, 18 Korea Research Institute of Bioscience and Biotechnology, Taejeon, Korea, 19 Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden, 20 Genexpress—CNRS—Functional Genomics and Systemic Biology for Health, Villejuif Cedex, France, 21 Sino-French Laboratory in Life Sciences and Genomics, Shanghai, China, 22 Tokyo Research Laboratories, Kyowa Hakko Kogyo Company, Tokyo, Japan, 23 MIPS—Institute for Bioinformatics, GSF—National Research Center for Environment and Health, Neuherberg, Germany, 24 Centre for Bioinformatics and Biological Computing, School of Information Technology, Murdoch University, Murdoch, Western Australia, Australia, 25 Medical Education and Biomedical Research Facility, University of Iowa, Iowa City, Iowa, United States of America, 26 Genome Exploration Research Group, RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, Kanagawa, Japan, 27 Medical College of Wisconsin, Milwaukee, Wisconsin, United States of America, 28 HUGO Gene Nomenclature Committee, University College London, London, United Kingdom, 29 Genome Science Laboratory, RIKEN, Saitama, Japan, 30 Ludwig Institute of Cancer Research, Sao Paulo, Brazil, 31 CNRS, Vandoeuvre les Nancy, France, 32 Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, 33 Department of Bioinformatics, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan, 34 Swiss Institute of Bioinformatics, Geneva, Switzerland, 35 Bioresource Information Division, RIKEN BioResource Center, RIKEN Tsukuba Institute, Ibaraki, Japan, 36 Genome Knowledgebase, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, 37 Chinese National Human Genome Center at Shanghai, Shanghai, China, 38 Division of Genetic Resources, National Institute of Infectious Diseases, Tokyo, Japan, 39 Graduate School of Frontier Sciences, Department of Integrated Biosciences, University of Tokyo, Chiba, Japan, 40 Functional Genomics Group, Biological Information Research Center, National Institute

of Advanced Industrial Science and Technology, Tokyo, Japan, **41** Department of Primary Care and Population Sciences, Royal Free University College Medical School, University College London, London, United Kingdom, **42** Clinical and Molecular Genetics Unit, The Institute of Child Health, London, United Kingdom, **43** Department of Genetic Information, Division of Molecular Life Science, School of Medicine, Tokai University, Kanagawa, Japan, **44** South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa, **45** Kazusa DNA Research Institute, Chiba, Japan, **46** RZPD Resource Center for Genome Research, Heidelberg, Germany, **47** Molecular Genome Analysis, German Cancer Research Center-DKFZ, Heidelberg, Germany, **48** Pennsylvania State University, University Park, Pennsylvania, United States of America, **49** Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Osaka, Japan, **50** Medical Photobiology Department, Photon Medical Research Center, Hamamatsu University School of Medicine, Shizuoka, Japan, **51** Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan, **52** Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan, **53** Department of Biological Sciences, Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology, Kanagawa, Japan, **54** Global Scientific Information and Computing Center, Tokyo Institute of Technology, Tokyo, Japan, **55** Molecular Biology Laboratory, Medicinal Research Laboratories, Taisho Pharmaceutical Company, Saitama, Japan, **56** Department of Population Genetics, National Institute of Genetics, Shizuoka, Japan, **57** Human Genome Research Group, Genomic Sciences Center, RIKEN Yokohama Institute, Kanagawa, Japan, **58** Columbia University and Columbia Genome Center, New York, New York, United States of America, **59** Department of Biotechnology, Royal Institute of Technology, Stockholm, Sweden, **60** Biology Division and Genome Task Group, Office of Biological and Environmental Research, United States Department of Energy, Washington, D.C., United States of America, **61** Faculty of Bio-Science, Nagahama Institute of Bio-Science and Technology, Shiga, Japan, **62** Institute for Genomic Research, Rockville, Maryland, United States of America, **63** Center for Genome Information, Department of Environmental Health, University of Cincinnati, Cincinnati, Ohio, United States of America, **64** State Key Laboratory of Medical Genomics, Shanghai Institute of Hematology, Rui-Jin Hospital, Shanghai Second Medical University, Shanghai, China, **65** PointOne Systems, Wauwatosa, Wisconsin, United States of America, **66** Graduate School of Life and Environmental Sciences, University of Tsukuba, Ibaraki, Japan, **67** Department of Genetics, Graduate University for Advanced Studies, Shizuoka, Japan, **68** Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, Tokyo, Japan

The human genome sequence defines our inherent biological potential; the realization of the biology encoded therein requires knowledge of the function of each gene. Currently, our knowledge in this area is still limited. Several lines of investigation have been used to elucidate the structure and function of the genes in the human genome. Even so, gene prediction remains a difficult task, as the varieties of transcripts of a gene may vary to a great extent. We thus performed an exhaustive integrative characterization of 41,118 full-length cDNAs that capture the gene transcripts as complete functional cassettes, providing an unequivocal report of structural and functional diversity at the gene level. Our international collaboration has validated 21,037 human gene candidates by analysis of high-quality full-length cDNA clones through curation using unified criteria. This led to the identification of 5,155 new gene candidates. It also manifested the most reliable way to control the quality of the cDNA clones. We have developed a human gene database, called the H-Invitational Database (H-InvDB; http://www.h-invitational.jp/). It provides the following: integrative annotation of human genes, description of gene structures, details of novel alternative splicing isoforms, non-protein-coding RNAs, functional domains, subcellular localizations, metabolic pathways, predictions of protein three-dimensional structure, mapping of known single nucleotide polymorphisms (SNPs), identification of polymorphic microsatellite repeats within human genes, and comparative results with mouse full-length cDNAs. The H-InvDB analysis has shown that up to 4% of the human genome sequence (National Center for Biotechnology Information build 34 assembly) may contain misassembled or missing regions. We found that 6.5% of the human gene candidates (1,377 loci) did not have a good protein-coding open reading frame, of which 296 loci are strong candidates for non-protein-coding RNA genes. In addition, among 72,027 uniquely mapped SNPs and insertions/deletions localized within human genes, 13,215 nonsynonymous SNPs, 315 nonsense SNPs, and 452 indels occurred in coding regions. Together with 25 polymorphic microsatellite repeats present in coding regions, they may alter protein structure, causing phenotypic effects or resulting in disease. The H-InvDB platform represents a substantial contribution to resources needed for the exploration of human biology and pathology.

Introduction

The draft sequences of the human, mouse, and rat genomes are already available (Lander et al. 2001; Marshall 2001; Venter et al. 2001; Waterston et al. 2002). The next challenge comes in the understanding of basic human molecular biology through interpretation of the human genome. To display biological data optimally we must first characterize the genome in terms of not only its structure but also function and diversity. It is of immediate interest to identify factors involved in the developmental process of organisms, non-protein-coding functional RNAs, the regulatory network of gene expression within tissues and its governance over states of health, and protein–gene and protein–protein interactions. In doing so, we must integrate this information in an easily accessible and intuitive format. The human genome may encode only 30,000 to 40,000 genes (Lander et al. 2001; Venter et al. 2001), suggesting that complex interde-

Received December 19, 2003; Accepted April 1, 2004; Published April 20, 2004
DOI: 10.1371/journal.pbio.0020162

Abbreviations: 3D, three-dimensional; AS, alternative splicing; CAI, codon adaptation index; dbSNP, Single Nucleotide Polymorphism Database; DDBJ, DNA Data Bank of Japan; EC, Enzyme Commission; EMBL, European Molecular Biology Laboratories; EST, expressed sequence tag; FANTOM, Functional Annotation of Mouse; FLcDNA, full-length cDNA; FLJ, Full-Length Long Japan; FTHFD, formyltetrahydrofolate dehydrogenase; GO, Gene Ontology; GTOP, Genomes TO Protein structures and functions database; H-Angel, Human Anatomic Gene Expression Library; H-Inv or H-Invitational, Human Full-Length cDNA Annotation Invitational; H-InvDB, H-Invitational Database; iAFLP, introduced amplified fragment length polymorphism; NCBI, National Center for Biotechnology Information; ncRNAs, non-protein-coding RNAs; OMIM, Online Mendelian Inheritance in Man; ORF, open reading frame; PDB, Protein Data Bank; RefSeq, Reference Sequence Collection; SMO, Similarity, Motif, and ORF; SNP, single nucleotide polymorphism

Academic Editor: Richard Roberts, New England Biolabs

*To whom correspondence should be addressed. E-mail: tgojobor@genes.nig.ac.jp

pendent gene regulation mechanisms exist to account for the complex gene networks that differentiate humans from lower-order organisms. In organisms with small genomes, it is relatively straightforward to use direct computational prediction based upon genomic sequence to identify most genes by their long open reading frames (ORFs). However, computational gene prediction from the genomic sequence of organisms with short exons and long introns can be somewhat error-prone (Ashburner 2000; Reese et al. 2000; Lander et al. 2001).

Previous efforts to catalogue the human transcriptome were based on expressed sequence tags (ESTs) used for the identification of new genes (Adams et al. 1991; Auffray et al. 1995; Houlgatte et al. 1995), chromosomal assignment of genes (Gieser and Swaroop 1992; Khan et al. 1992; Camargo et al. 2001), prediction of genes (Nomura et al. 1994), and assessment of gene expression (Okubo et al. 1992). Recently, Camargo et al. (2001) generated a large collection of ORF ESTs, and Saha et al. (2002) conducted a large-scale serial analysis of gene expression patterns to identify novel human genes. The availability of human full-length transcripts from many large-scale sequencing projects (Nomura et al. 1994; Nagase et al. 2001; Wiemann et al. 2001; Yudate 2001; Kikuno et al. 2002; Strausberg et al. 2002) has provided a unique opportunity for the comprehensive evaluation of the human transcriptome through the annotation of a variety of RNA transcripts. Protein-coding and non-protein-coding sequences, alternative splicing (AS) variants, and sense–antisense RNA pairs could all be functionally identified. We thus designed an international collaborative project to establish an integrative annotation database of 41,118 human full-length cDNAs (FLcDNAs). These cDNAs were collected from six high-throughput sequencing projects and evaluated at the first international jamboree, entitled the Human Full-length cDNA Annotation Invitational (H-Invitational or H-Inv) (Cyranoski 2002). This event was held in Tokyo, Japan, and took place from August 25 to September 3, 2002.

Efforts which have been made in the same area as the H-Inv annotation work include the Functional Annotation of Mouse (FANTOM) project (Kawai et al. 2001; Bono et al. 2002; Okazaki et al. 2002), Flybase (GOC 2001), and the RIKEN *Arabidopsis* full-length cDNA project (Seki et al. 2002). In our own project, great effort has been taken at all levels, not only in the annotation of the cDNAs but also in the way the data can be viewed and queried. These aspects, along with the applications of our research to disease research, distinguish our project from other similar projects.

This manuscript provides the first report by the H-Inv consortium, showing some of the discoveries made so far and introducing our new database of the human transcriptome. It is hoped that this will be the first in a long line of publications announcing discoveries made by the H-Inv consortium. Here we describe results from our integrative annotation in four major areas: mapping the transcriptome onto the human genome, functional annotation, polymorphism in the transcriptome, and evolution of the human transcriptome. We then introduce our new database of the human transcriptome, the H-Invitational Database (H-InvDB; http://www.h-invitational.jp), which stores all annotation results by the consortium. Free and unrestricted access to the H-Inv annotation work is available through the database. Finally, we summarize our most important findings thus far in the H-Inv project in Concluding Remarks.

Results/Discussion

Mapping the Transcriptome onto the Human Genome

Construction of the nonredundant human FLcDNA database. We present the first experimentally validated nonredundant transcriptome of human FLcDNAs produced by six high-throughput cDNA sequencing projects (Ota et al. 1997, 2004; Strausberg et al. 1999; Hu et al. 2000; Wiemann et al. 2001; Yudate 2001; Kikuno et al. 2002) as of July 15, 2002. The dataset consists of 41,118 cDNAs (H-Inv cDNAs) that were derived from 184 diverse cell types and tissues (see Dataset S1). The number of clones, the number of libraries, major tissue origins, methods, and URLs of cDNA clones for each cDNA project are summarized in Table 1. H-Inv cDNAs include 8,324 cDNAs recently identified by the Full-Length Long Japan (FLJ) project. The FLJ clones represent about half of the H-Inv cDNAs (Table 1). The policies for library selection and the results of initial analysis of the constituent projects were reported by the participants themselves: the Chinese National Human Genome Center (CHGC) (Hu et al. 2000), the Deutsches Krebsforschungszentrum (DKFZ/MIPS) (Wiemann et al. 2001), the Institute of Medical Science at the University of Tokyo (IMSUT) (Suzuki et al. 1997; Ota et al. 2004), the Kazusa cDNA sequence project of the Kazusa DNA Research Institute (KDRI) (Hirosawa et al. 1999; Nagase et al. 1999; Suyama et al. 1999; Kikuno et al. 2002), the Helix Research Institute (HRI) (Yudate et al. 2001), and the Mammalian Gene Collection (MGC) (Strausberg et al. 1999; Moonen et al. 2002), as well as FLJ mentioned earlier (Ota et al. 2004). The variation in tissue origins for library construction among these six groups resulted in rare occurrences of sequence redundancy among the collections. In a recent study, the FLJ project has described the complete sequencing and characterization of 21,243 human cDNAs (Ota et al. 2004). On the other hand, the H-Inv project characterized cDNAs from this project and six high-throughput cDNA producers by using a different suite of computational analysis techniques and an alternative system of functional annotation.

The 41,118 H-Inv cDNAs were mapped on to the human genome, and 40,140 were considered successfully aligned. The alignment criterion was that a cDNA was only aligned if it had both 95% identity and 90% length coverage against the genome (Figure 1). The mean identity of all the alignments between 40,140 mapped cDNAs and genomic sequences was 99.6 %, and the mean coverage against the genomic sequence was 99.6%. In some cases, terminal exons were aligned with low identity or low coverage. For example, 89% of internal exons have identity of 99.8% or higher, while only 78% and 50% of the first and last exons do, respectively. These alignments with low identity or low coverage seemed to be caused by the unsuccessful alignments of the repetitive sequences found in UTR regions and the misalignments of 3′ terminal poly-A sequences. Although better alignments could be obtained for these sequences by improving the mapping procedure, we concluded that the quality of the FLcDNAs was high overall.

Due to redundancy and AS within the human transcriptome, these 40,140 cDNAs were clustered to 20,190 loci

Table 1. Summary of cDNA Resources

cDNA Sequence Provider*	Number of cDNAs (Without Redundancy)	Number of Library Origins	Major Tissue Library Origins	Method	URL	Reference
CHGC	758 (754)	30	Adrenal gland, hypothalamus, CD34+ stem cell	Selecting FLcDNA clones from EST libraries	http://www.chgc.sh.cn/	Hu et al. 2000
DKFZ/MIPS	5,555 (5,521)	14	Testis, brain, lymph node	Selecting FLcDNA clones from 5′- and 3′- EST libraries	http://mips.gsf.de/projects/cdna	Wiemann et al. 2001
FLJ/HRI	8,066 (8,057)	46	Teratocarcinoma, placenta, whole embryo	Oligo-capping method and selection by one-pass sequences	http://www.hri.co.jp/HUNT/	Ota et al. 1997, 2004; Yudate et al. 2001
FLJ/IMSUT	12,585 (12,560)	81	Brain, testis, bone marrow	Oligo-capping method and selection by one-pass sequences	http://cdna.ims.u-tokyo.ac.jp/	Suzuki et al. 1997; Ota et al. 2004
FLJ/KDRI	348(342)	1	Spleen	Selection by one-pass sequences	http://www.kazusa.or.jp/NEDO/	Ota et al. 2004
KDRI	2,000 (2,000)	9	Brain	In vitro protein synthesis and selection by one-pass sequences	http://www.kazusa.or.jp/huge/	Hirosawa et al. 1997; Nagase et al. 1999; Suyama et al. 1999; Kikuno et al. 2002
MGC/NIH	11,806(11,414)	69	Placenta, lung, skin	Selecting gene candidates from 5′-EST libraries	http://mgc.nci.nih.gov/	Strausberg et al. 1999

*FLcDNA data were provided for H-Inv project by the FLJ project of NEDO (URL: http://www.nedo.go.jp/bio-e/) and six high-throughput cDNA clone producers Chinese National Human Genome Center (CHGC), the Deutsches Krebsforschungszentrum (DKFZ/MIPS), Helix Research Institute (HRI), the Institute of Medical Science in the University of Tokyo (IMSUT), the Kazusa DNA Research Institute (KDRI), and the Mammalian Gene Collection (MGC/NIH).

DOI: 10.1371/journal.pbio.0020162.t001

(H-Inv loci). For the remaining 978 unmapped cDNAs, we conducted cDNA-based clustering, which yielded 847 clusters. The clusters created had an average of 2.0 cDNAs per locus (Table 2). The average was only 1.2 for unmapped clusters, probably because many of these genes are encoded by heterochromatic regions of the human genome and show limited levels of gene expression. The gene density for each chromosome varied from 0.6 to 19.0 genes/Mb, with an average of 6.5 genes/Mb. This distribution of genes over the genome is far from random. This biased gene localization concurs with the gene density on chromosomes found in similar previous reports (Lander et al. 2001; Venter et al. 2001). This indicates that the sampled cDNAs are unbiased with respect to chromosomal location. Most cDNAs were mapped only at a single position on the human genome. However, 1,682 cDNAs could be mapped at multiple positions (with mean values of 98.2% identity and 98.1% coverage). The multiple matching may be caused by either recent gene duplication events or artificial duplication of the human genome caused by misassembled contigs. In our study we have selected only the "best" loci for the cDNAs (see Materials and Methods for details).

In total, 21,037 clusters (20,190 mapped and 847 unmapped) were identified and entered into the H-InvDB. We assigned H-Inv cluster IDs (e.g., HIX0000001) to the clusters and H-Inv cDNA IDs (e.g., HIT000000001) to all curated cDNAs. A representative sequence was selected from each cluster and used for further analyses and annotation.

Comparison of the mapped H-Inv cDNAs with other annotated datasets. In order to evaluate the H-Inv dataset, we compared all of the mapped H-Inv cDNAs with the Reference Sequence Collection (RefSeq) mRNA database (Pruitt and Maglott 2001) (Figure 2). The RefSeq mRNA database consists of two types of datasets. These are the curated mRNAs (accession prefix NM and NR) and the model mRNAs that are provided through automated processing of the genome annotation (accession prefix XM and XR).

From the comparison, we found that 5,155 (26%) of the H-Inv loci had no counterparts and were unique to the H-Inv. All of these 5,155 loci are candidates for new human genes, although non-protein-coding RNAs (ncRNAs) (25%), hypothetical proteins with ORFs less than 150 amino acids (55%), and singletons (91%) were enriched in this category. In fact, 1,340 of these H-Inv-unique loci were questionable and require validation by further experiments because they consist of only single exons, and the 3′ termini of these loci align with genomic poly-A sequences. This feature suggests internal poly-A priming although some occurrences might be bona fide genes. The most reliable set of newly identified human genes in our dataset is composed of 1,054 protein-

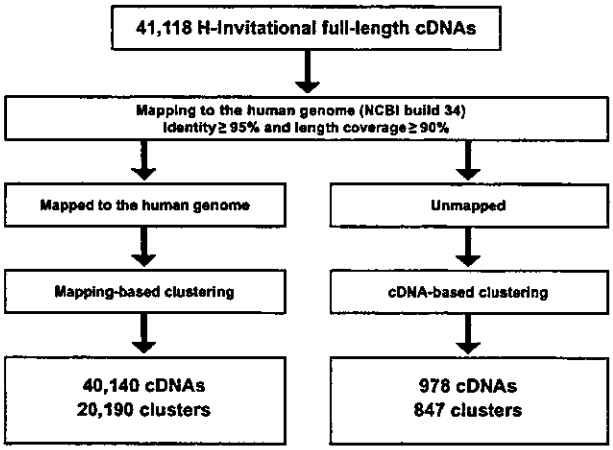


Figure 1. Procedure for Mapping and Clustering the H-Inv cDNAs

The cDNAs were mapped to the genome and clustered into loci. The remaining unmapped cDNAs were clustered based upon the grouping of significantly similar cDNAs.

DOI: 10.1371/journal.pbio.0020162.g001

coding and 179 non-protein-coding genes that have multiple exons. Therefore, at least 6.1% (1,233/20,190) of the H-Inv loci could be used to newly validate loci that the RefSeq datasets do not presently cover. These genes are possibly less expressed since the proportion of singletons (H-Inv loci consisting of a single H-Inv cDNA) was high (84%).

On the other hand, 78% (11,974/15,439) of the curated RefSeq mRNAs were covered by the H-Inv cDNAs. These figures suggest that further extensive sequencing of FLcDNA clones will be required in order to cover the entire human gene set. Nonetheless, this effort provides a systematic approach using the H-Inv cDNAs, even though a portion of the cDNAs have already been utilized in the RefSeq datasets.

It is noteworthy that H-Inv cDNAs overlapped 3,061 (17%) of RefSeq model mRNAs, supporting this proportion of the hypothetical RefSeq sequences. These newly confirmed 3,061 loci have a mean number of exons greater than RefSeq model mRNAs that were not confirmed, but smaller than RefSeq curated mRNAs. The overlap between H-Inv cDNAs and RefSeq model mRNAs was smaller than that between H-Inv cDNAs and RefSeq curated mRNAs. This suggests that the genes predicted from genome annotation may tend to be less expressed than RefSeq curated genes, or that some may be artifacts. All these results highlight the great importance of comprehensive collections of analyzed FLcDNAs for validat-

Table 2. The Clustering Results of Human FLcDNAs onto the Human Genome

Chromosome	Number of Loci	Number of cDNAs	Number of cDNAs/Locus	Number of Loci/Mb
1	1,998	4,057	2.0	8.1
2	1,408	2,791	2.0	5.8
3	1,224	2,455	2.0	6.1
4	809	1,527	1.9	4.2
5	920	1,851	2.0	5.1
6	1,027	1,912	1.9	6.0
7	1,008	1,994	2.0	6.4
8	761	1,448	1.9	5.2
9	817	1,630	2.0	6.0
10	863	1,705	2.0	6.4
11	1,116	2,245	2.0	8.3
12	1,014	2,071	2.0	7.7
13	394	743	1.9	3.5
14	626	1,363	2.2	5.9
15	693	1,415	2.0	6.9
16	865	1,851	2.1	9.6
17	1,110	2,245	2.0	13.6
18	334	593	1.8	4.4
19	1,210	2,378	2.0	19.0
20	536	1,124	2.1	8.4
21	197	379	1.9	4.2
22	480	985	2.1	9.7
X	646	1,173	1.8	4.2
Y	29	32	1.1	0.6
UN[a]	105	173	1.6	–
Unmapped	847	978	1.2	–
Total	21,037	41,118	2.0	–

[a]UN represents contigs that were not mapped onto any chromosome.

DOI: 10.1371/journal.pbio.0020162.t002

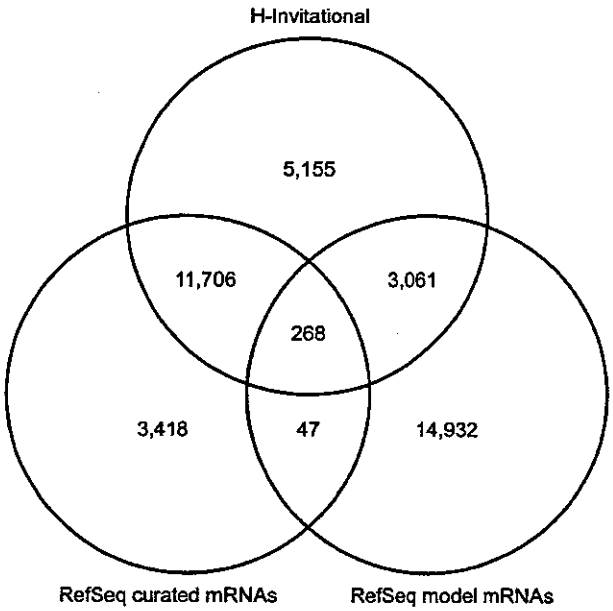


Figure 2. A Comparison of the Mapped H-Inv FLcDNAs and the RefSeq mRNAs

The mapped H-Inv cDNAs, the RefSeq curated mRNAs (accession prefixes NM and NR), and the RefSeq model mRNAs (accession prefixes XM and XR) provided by the genome annotation process were clustered based on the genome position. The numbers of loci that were identified by clustering are shown.

DOI: 10.1371/journal.pbio.0020162.g002

ing gene prediction from genome sequences. This may be especially true for higher organisms such as humans.

Incomplete parts of the human genome sequences. The existence of 978 unmapped cDNAs (847 clusters) suggests that the human genome sequence (National Center for Biotechnolgy Information [NCBI] build 34 assembly) is not yet complete. The evidence supporting this statement is twofold. First, most of those unmapped cDNAs could be partially mapped to the human genome. Using BLAST, 906 of the unmapped cDNAs (corresponding to 786 clusters) showed at least one sequence match to the human genome with a bit score higher than 100. Second, most of the cDNAs could be mapped unambiguously to the mouse genome sequences. A total of 907 unmapped cDNAs (779 clusters; 92%) could be mapped to the mouse genome with coverage of 90% or higher. If we adopted less stringent requirements, more cDNAs could be mapped to the mouse genome. The rest might be less conserved genes, genes in unfinished sections of the mouse genome, or genes that were lost in the mouse genome. Based on these observations, we conclude that the human genome sequence is not yet complete, leaving some portions to be sequenced or reassembled.

The proportion of the genome that is incomplete is estimated to be 3.7%–4.0%. The figure of 4.0% is based upon the proportion of H-Inv cDNA clusters that could not be mapped to the genome (847/21,037), while the 3.7% estimate is based on both H-Inv cDNAs and RefSeq sequences (only NMs). This statistic indicates that a minimum of one out of every 25–27 clusters appears to be unrepresented in the current human genome dataset, in its full form. Possible reasons for this include unsequenced regions on the human genome and regions where an error may have occurred during sequence assembly. If this is the case, this lends support to the use of cDNA mapping to facilitate the completion of whole genome sequences (Kent and Haussler 2001). For example, we can predict the arrangement of contigs based on the order of mapped exons. In addition we can use the sequences of unmapped exons to search for those clones that contain unsequenced parts of the genome. The mapping results of partially mapped cDNAs are thus quite useful.

Primary structure of genes on the human genome. Using the H-Inv cDNAs, the precise structures of many human genes could be identified based on the results of our cDNA mapping (Table S1). The median length of last exons (786 bp) was found to be longer than that of other exons, and the median length of first introns (3,152 bp) longer than that of other introns. These observed characteristics of human gene structures concur with the previous work using much smaller datasets (Hawkins 1988; Maroni 1996; Kriventseva and Gelfand 1999).

In the human genome, 50% of the sequence is occupied by repetitive elements (Lander et al. 2001). Repetitive elements were previously regarded by many as simply "junk" DNA. However, the contribution of these repetitive stretches to genome evolution has been suggested in recent works (Makalowski 2000; Deininger and Batzer 2002; Sorek et al. 2002; Lorenc and Makalowski 2003). The 21,037 loci of representative cDNAs were searched for repetitive elements using the RepeatMasker program. RepeatMasker indicated that 9,818 (47%) of the H-Inv cDNAs, including 5,442 coding hypothetical proteins, contained repetitive sequences. The existence of *Alu* repeats in 5% of human cDNAs was reported previously (Yulug et al. 1995). Our results revealed a significant number of repetitive sequences including *Alu* in the human transcriptome. Among them, 1,866 cDNAs overlapped repetitive sequences in their ORFs. Moreover, 554 of 1,866 cDNAs had repetitive sequences contained completely within their ORFs, including 81 cDNAs that were identical or similar to known proteins. This may indicate the involvement of repetitive elements in human transcriptome evolution, as suggested by the presence of Alu repeats in AS exons (Sorek et al. 2002) and the contribution to protein variability by repetitive elements in protein-coding regions (Makalowski 2000). We detected 2,254 and 5,427 cDNAs containing repetitive sequences in their 5′ UTR and 3′ UTR, respectively. The positioning of the repetitive elements suggests they play a regulatory role in the control of gene expression (Deininger and Batzer 2002) (see Table S1 or the H-InvDB for details).

AS transcripts. We wished to investigate the extent to which the functional diversity of the human proteome is affected by AS. In order to do this, we searched for potential AS isoforms in 7,874 loci that were supported by at least two H-Inv cDNAs. We examined whether or not these cDNAs represented mutually exclusive AS isoforms, using a combination of computational methods and human curation (see Materials and Methods). All AS isoforms that were supported independently by both methods were defined as the H-Inv AS dataset. Our analysis showed that 3,181 loci (40 % of the 7,874 loci) encoded 8,553 AS isoforms expressing a total of 18,612 AS exons. On average, 2.7 AS isoforms per locus were identified in these AS-containing loci. This figure represents