雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
Osada N, Uno Y, Mineta K, Kameoka Y, Takahashi I, Terao K.	Ancient genome-wide admixture extends beyond the current hybrid zone between Macaca fascicularis and M. mulatta.	Mol Ecol.	19(14)	2884-95	2010 Jul Epub 2010 Jun 23.
Toyoda M, Yamazaki-Inoue M, Itakura Y, Kuno A, Ogawa T, Yamada M, Akutsu H, Takahashi Y, Kanzaki S, Narimatsu H, Hirabayashi J, Umezawa A.	Lectin microarray analysis of pluripotent and multipotent stem cells.	Genes Cells	16(1)	1-11	2011
Higuchi A, Ling QD, Ko YA, Chang Y, Umezawa A.	Biomaterials for the Feeder-Free Culture of Human Embryonic Stem Cells and Induced Pluripotent Stem Cells.	Chem Rev		dx.doi.org/10.1021/cr1003612	2011
Inamura M, Kawabata K, Takayama K, Tashiro K, Sakurai F, Katayama K, Toyoda M, Akutsu H, Miyagawa Y, Okita H, Kiyokawa N, Umezawa A, Hayakawa T, Furue MK, Mizuguchi H.	Efficient Generation of Hepatoblasts From Human ES Cells and iPS Cells by Transient Overexpression of Homeobox Gene HEX.	Mol Ther	19(2)	400-407	2011
Nishino K, Toyoda M, Yamazaki-Inoue M, Makino H, Fukawatase Y, Chikazawa E, Takahashi Y, Miyagawa Y, Okita H, Kiyokawa N, Akutsu H, Umezawa A.	Defining hypo-methylated regions of stem cell-specific promoters in human iPS cells derived from extra-embryonic amnions and lung fibroblasts.	PLoS One	5(9)	e13017	2010
Toyoda M, Hamatani T, Okada H, Matsumoto K, Saito H, Umezawa A.	Defining cell identity by comprehensive gene expression profiling.	Curr Med Chem	17(28)	3245-3252	2010
Hatta T, Murayama T, Narita K, Sumi E, Yokode M	Trend Analysis of Informed Consent Research in Clinical Trials: Comprehensive Retrieval via Electronic Databases.	Jpn J Clin Pharmacol Therapeutics.	42	21-25	2011
Shimada K, Mikami Y, Murayama T, Yokode M, Fujita M, Kita T, Kishimoto C.	Atherosclerotic plaques induced by marble-burying behavior are stabilized by exercise training in experimental atherosclerosis.	Int J Cardiol.			2011 in press

Shimada K, Murayama T, Yokode M, Kita T, Fujita M, Kishimoto C.	Olmesartan, a novel angiotensin II type 1 receptor antagonist, reduces severity of atherosclerosis in apolipoprotein E deficient mice associated with reducing superoxide production.	Nutr Metab Cardiovasc Dis. 2011 in press			2011 in press
Horie T, Ono K, Horiguchi M, Nishi H, Nakamura T, Nagao K, Kinoshita M, Kuwabara Y, Marusawa H, Iwanaga Y, Hasegawa K, Yokode M, Kimura T, Kita T.	MicroRNA-33 encoded by an intron of sterol regulatory element-binding protein 2 (Srebp2) regulates HDL in vivo.	Proc Natl Acad Sci U S A.	107	17321-17326	2010
村山敏典, 横出正之	【はじめての臨床応用研究】 知っておきたい臨床応用への制度　高度医療と先進医療	遺伝子医学MOOK別冊はじめての臨床応用研究		26-34	2010
Iwamori M, Iwamori Y, Adachi S, Nomura T.	Excretion into feces of asialo GM1 in the murine digestive tract and Lactobacillus johnsonii exhibiting binding ability toward asialo GM1. A possible role of epithelial glycolipids in the discharge of intestinal bacteria.	Glycoconj. J.			E-published Dec. 21, 2010
Taisei Nomura.	Biological Consequence and Health Concern from Low Dose and Low Dose Rate Radiations in Mice and Humans.	Health Physics.	100	266-268	2011 (in press)
野村大成、梁　治子、足立成基、時田偉子、堀家なな緒、畑中英子、菊谷理絵、中島裕夫、本行忠志、藤川和男、伊藤哲夫、落合俊昌、行徳淳一郎、若命浩二	宇宙環境の人体影響評価と防護に関する研究	Space Util. Res.	27		2011 (in press)
A. Yogo, T. Maeda, T. Hori, H. Sakaki, K. Ogura, M. Nishiuchi, A. Sagisaka,H. Kiriyama, H. Okada, S. Kanazawa, T. Shimomura, Y. Nakai, M. Tanoue F. Sasao, P. R. Bolton, M.Murakami, T. Nomura, S. Kawanishi, and K. Kondo	Measurement of relative biological effectiveness of protons in human cancer cells using a laser-driven quasimonoenergetic proton beamline	APPLIED PHYSICS LETTERS	98-5370 1		2011 (in press)

Dirks WG, MacLeod RA, Nakamura Y, Kohara A, Reid Y, Milch H, Drexler HG, Mizusawa H.	Cell line cross-contamination initiative: an interactive reference database of STR profiles covering common cancer cell lines.	Int J Cancer	126(1)	303-304	2010
Capes-Davis A, Theodosopoulos G, Atkin I, Drexler HG, Kohara A, Macleod RA, Masters JR, Nakamura Y, Reid YA, Reddel RR, Freshney RI	Check your cultures! A list of cross-contaminated or misidentified cell lines.	Int J Cancer	127(1)	1-8	2010
American Type Culture Collection Standards Development Organization Workgroup ASN-0002.	Cell line misidentification: the beginning of the end.	Nat Rev Cancer.	10(6)	441-448	2010
Barallon R, Bauer SR, Butler J, Capes-Davis A, Dirks WG, Elmore E, Furtado M, Kline MC, Kohara A, Los GV, Macleod RA, Masters JR, Nardone M, Nardone RM, Nims RW, Price PJ, Reid YA, Shewale J, Sykes G, Steuer AF, Stotts DR, Thomson J, Taraporewala Z, Alston-Roberts C, Kerrigan L.	Recommendation of short tandem repeat profiling for authenticating human cell lines, stem cells, and tissues.	In Vitro Cell Dev Biol Anim.	46(9)	727-732	2010
Mimura S, Kimura N, Hirata M, Tateyama D, Hayashida M, Umezawa A, Kohara A, Nikawa H, Okamoto T, Furue MK.	Growth factor-defined culture medium for human mesenchymal stem cells.	Int J Dev Biol.			in press

書籍

著者氏名	論文タイトル名	書籍全体の編集者名	書　籍　名	出版社名	出版地	出版年	ページ
Nomura T.	Transgenerational health concerns from radiation in mice and humans.	Bersimbay RI, Au W	Genome-environment interactions and genetic toxicology. 15th Alexander Hollaender Course	Eurasian National University Press	Astana	2010	19-23

Ⅶ 研究成果の刊行物 ・ 別冊

MOLECULAR ECOLOGY

Molecular Ecology (2010) 19, 2884–2895 doi: 10.1111/j.1365-294X.2010.04687.x

Ancient genome-wide admixture beyond the current hybrid zone between *Macaca fascicularis* and *M. mulatta*

NAOKI OSADA,*† YASUHIRO UNO,‡ KATSUHIKO MINETA,§ YOSUKE KAMEOKA,* ICHIRO TAKAHASHI* and KEIJI TERAO¶

**Department of Biomedical Resources, National Institute of Biomedical Innovation, 7-6-8 Saito-Asagi, Ibraki, Osaka 567-0085, Japan, †Department of Population Genetics, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 441-8540, Japan, ‡Pharmacokinetics and Bioanalysis Center, Shin Nippon Biomedical Laboratories, Ltd, 16-1 Minami-Akasaka, Kainan, Wakayama 642-0017, Japan, §Graduate School of Information Science and Technology, Hokkaido University, N14W9 Sapporo 060-0814, Japan, ¶Tsukuba Primate Research Center, National Institute of Biomedical Innovation, 1 Hachimandai, Tsukuba 305-0843, Japan*

Abstract

***Macaca fascicularis* and *Macaca mulatta* are two of the most commonly used laboratory macaques, yet their genetic differences at a genome-wide level remain unclear. We analysed the multilocus DNA sequence data of 54 autosomal loci obtained from *M. fascicularis* samples from three different geographic origins and *M. mulatta samples* of Burmese origin. *M. fascicularis* shows high nucleotide diversity, four to five times higher than humans, and a strong geographic population structure between Indonesian-Malaysian and Philippine macaques. The pattern of divergence and polymorphism between *M. fascicularis* and *M. mulatta* shows a footprint of genetic exchange not only within their current hybrid zone but also across a wider range for more than 1 million years. However, genetic admixture may not be a random event in the genome. Whereas randomly selected genic and intergenic regions have the same evolutionary dynamics between the species, some cytochrome oxidase P450 (CYP) genes (major chemical metabolizing genes and potential target genes for local adaptation) have a significantly larger species divergence than other genes. By surveying *CYP3A5* gene sequences of more than a hundred macaques, we identified three nonsynonymous single nucleotide polymorphisms that were highly differentiated between the macaques. The mosaic pattern of species divergence in the genomes may be a consequence of genetic differentiation under ecological adaptation and may be a salient feature in the genomes of nascent species under parapatry.**

Keywords: cytochrome oxidase P450, isolation with migration model, macaque, speciation

Received 5 January 2010; revision received 20 April 2010; accepted 27 April 2010

Introduction

Macaque monkeys are frequently used for a variety of biological studies, such as those on infectious diseases, pharmacology, and tissue engineering (Sibal & Samson 2001). Among the species of the genus *Macaca*, which includes as many as 19 species (Fooden 1976), *M. fascicularis* and *M. mulatta* are two of the most commonly used macaques in laboratories worldwide. They are classified as members of the *fascicularis* subgroup, principally based on the shape of the male genitalia (Fooden 1976). *M. fascicularis*, which is called the cynomolgus, long-tail, or crab-eating macaque, lives in Southeast Asia, including the Indonesian Islands, Philippine Islands, and Indochina, whereas *M. mulatta*, which is also known as the rhesus macaque, lives in more northern and western regions of the continent (Fig. S1, Supporting

Correspondence: Naoki Osada, Fax: +81 (55) 981-6793; E-mail: nosada@lab.nig.ac.jp

information). The current distribution of the two species is mostly allopatric, but partially overlapped around northern Indochina. Although *M. fascicularis* and *M. mulatta* show distinct morphological and behavioral differences, previous studies using morphological and genetic data have suggested a possible hybridization between the species in current populations (Fooden 1964; Tosi *et al.* 2002; Hamada *et al.* 2006; Kanthaswamy *et al.* 2008; Stevison & Kohn 2009).

Because of the prominent position of macaques in biomedical studies, the draft genome sequence of *M. mulatta* was published in 2007 as the third primate genome sequence (Gibbs *et al.* 2007). Although the draft genome sequence is derived from a single Indian *M. mulatta*, previous studies have shown that responses against toxins and pathogens vary considerably within and between macaque species. For example, Indian and Chinese *M. mulatta* have been shown to be differentially susceptible to Simian immunodeficiency virus (SIV) infection (Trichel *et al.* 2002). Similarly, *M. fascicularis* and *M. mulatta* are differentially susceptible to certain species of malaria (Wheatley 1980; Matsumoto *et al.* 2000). Therefore, evaluating and monitoring the intra- and inter-species diversity of macaques are important tasks for achieving a high reliability and reproducibility of laboratory experiments using macaques (Stevison & Kohn 2008). The amount of genetic diversity and geographic population structure within *M. mulatta* using genome-wide SNP data have been studied by many researchers (Magness *et al.* 2005; Ferguson *et al.* 2007; Hernandez *et al.* 2007; Malhi *et al.* 2007; Street *et al.* 2007). In contrast to *M. mulatta*, however, the level of polymorphism in *M. fascicularis* has been investigated mainly at sex-linked loci (Melnick *et al.* 1993; Hayasaka *et al.* 1996; Tosi *et al.* 2002; Smith *et al.* 2007; Blancher *et al.* 2008; Kanthaswamy *et al.* 2008; Bonhomme *et al.* 2009) or using microsatellite markers (Kanthaswamy *et al.* 2008; Bonhomme *et al.* 2009); only a few studies have characterized the nucleotide variation at the single nucleotide polymorphism (SNP) level (Street *et al.* 2007; Stevison & Kohn 2009).

The evolutionary history of bifurcated species may be too complex to be reconstructed, especially when the two species are closely related and the diversity within each species is large. Former studies that compared a large number of macaque transcripts have shown that their average genetic divergence is approximately 0.4–0.5% and that they diverged around 1 Ma (Osada *et al.* 2008). This estimate is based on the assumption that the two species have been under complete genetic isolation after their speciation. Recent advances in methods for analysing multilocus DNA sequence data have allowed us to infer the mode of speciation, that is whether two species diverged under complete geographic isolation (allopatric model) or gradually diverged with some extent of gene flow between nascent species (parapatric model) (Osada & Wu 2005; Hey & Nielsen 2007). The relatively older speciation event between humans and chimpanzees (≈6 **Mya**) may obscure this speciation pattern owing to intra-locus recombination and recurrent mutation (Osada & Wu 2005; Innan & Watanabe 2006; Patterson *et al.* 2006; Hobolth *et al.* 2007). Therefore, inferring the mode of speciation in macaques and its impact on their genome sequences will help us to understand a complex speciation pattern in primates. Here we investigated the amount of genetic diversity at 54 randomly chosen autosomal loci and inferred a geographic population structure for *M. fascicularis*. In addition, the genetic divergence between *M. fascicularis* and *M. mulatta* was quantified using these 54 loci.

We also investigated the amount of polymorphism at seven cytochrome P450 (CYP) loci. In a wild habitat, CYP genes may be responsible for detoxifying chemical substrates in leaves and fruits, and for the synthesis of internal hormones and steroids. They are also important genes for drug metabolism and are important to toxicology in pharmaceutical research. For example, a previous study on human *CYP3A5* polymorphisms has suggested that human polymorphisms in *CYP3A5* are associated with the salt response and that this is mediated through the deoxidation of cortisol (Thompson *et al.* 2004). Therefore, CYP genes represent candidate genes on which natural selection could affect pattern of genomic differentiation.

Materials and methods

Sample collection and DNA sequencing

Eight, seven, and nine unrelated female *M. fascicularis* individuals that have wild-caught parents from Indonesia, Peninsular Malaysia, and the Philippines, respectively, were used for the polymorphism studies. Five male *M. mulatta* of Burmese origin were also investigated. As two of the five *M. mulatta* shared either their parents or grandparents with other individuals, these samples were excluded from further analyses. The monkeys were cared for and handled according to the guidelines established by the Tsukuba Primate Research Center, Japan. Detailed sample information is given in Table S1 (Supporting information). The DNA from the macaques was extracted from blood samples. The DNA samples from *M. fascicularis* were amplified using REPLI-g (Qiagen, USA). In total, 61 loci were amplified using a high fidelity DNA polymerase (PrimeStar Taq: TakaraBio, Japan), and the PCR products were purified using Ampure (Agencourt Bioscience, USA). The PCR

primers were designed based on the rhesus macaque draft genome sequence (rheMac2). All of the primer sequences used in the study are presented in Table S2 (Supporting information). The DNA sequence of the PCR products was determined from both DNA strands using an ABI 3730 sequencer (Applied Biosystems). If the sample carried more than one heterozygous indel in the target region, the sequences were not used for further analyses. The DNA sequences have been deposited in a public database (DDBJ/Embl/Genbank accession nos. AB380123–AB381873). For the *CYP3A5* polymorphism study, blood samples from 38 *M. fascicularis* from Indochina, 40 *M. fascicularis* from Indonesia, and 34 *M. mulatta from China,* all imported via commercial animal breeders, were used (Uno *et al.* 2010).

Sequence analysis

All of the SNPs were screened using the ABI sequence analysis software and confirmed by a visual inspection of the chromatograms. The DNA sequences were aligned using the ClustalW and the MEGA4.0 program and corrected by a visual inspection (Thompson *et al.* 1994; Tamura *et al.* 2007). The summary statistics for each locus, such as Watterson's θ, the nucleotide diversity (π), Tajima's D, F_{ST}, and Nei's d_{xy}, were estimated using the DnaSP 4.2 software package (Rozas *et al.* 2003). The population structure was inferred using the STRUCTURE program with a burn-in length of 10 000 and 100 000 sampling steps (Pritchard *et al.* 2000). An admixture model was assumed, because each population is not supposed to be completely isolated. Although we repeated the analysis assuming a nonadmixture model, the results were essentially same (data not shown). The multidimensional scaling plot in Fig. 2 was drawn using Kimura's pairwise distances (Kimura 1980). The analysis of molecular variance (AMOVA) was performed using Arlequin 3.1 (Excoffier *et al.* 1992). In order to find outlier loci showing high or low genetic differentiation, the BayeScan program was used with a default parameter setting (Foll & Gaggiotti 2008).

Test of population demographic models

For the demographic study, we used only the control autosomal loci, and not the CYP loci. To estimate population parameters in the isolation with migration (IM) model, we inferred the haplotypes using the PHASE2.1 program (Stephens *et al.* 2001) and estimated the population parameters using the IMa software (Hey & Nielsen 2007). All of the sequence data were trimmed using the IMgc program so as not to contain possible recombined DNA fragments (Woerner *et al.* 2007). For each of the intra- and inter-species analysis, three independent runs of Markov chain Monte Carlo (MCMC) with 1 million sampling steps were carried out. Twenty MCMC chains with 1 500 000 steps of burn-in for the intra-species analysis, and ten MCMC chains with 1 million steps of burn-in for the inter-species analysis were run. Because the posterior distributions of all three runs were similar in each analysis, they were summed for parameter estimation. The infinite site (IS) nucleotide substitution model was assumed. For the test of Tajima's D statistics, we generated datasets under constant population size using a coalescent simulator (Hudson 2002), given the nucleotide diversity, sample size, number of loci, and fragment length of each locus. The P values for the test statistics are based on 1000 iterations of the coalescent simulation.

Prediction of three-dimensional protein structure

In order to assign the location of highly differentiated sites in CYP3A5 between the species, we predicted a CYP3A5 protein structure of *M. fascicularis*. A homology model for the macaque CYP3A5 protein was built based on the crystal structure of the human CYP3A4 [PDB ID: 1TQN (Yano *et al.* 2004)]. The sequence alignment, comparative protein modelling and energy minimization were performed using the SWISS-MODEL server (Arnold *et al.* 2006) with the default parameters. UCSF Chimera (Pettersen *et al.* 2004) was used for the three-dimensional protein visualization.

Results

The genetic diversity and population structure of M. fascicularis

We determined the nucleotide sequences of 54 autosomal loci from 24 *M. fascicularis* samples. The exact sampling location of these *M. fascicularis* was unknown, but the countries of origin (Indonesia, Peninsular Malaysia, or the Philippines) were identified. These 54 loci were selected to be distributed throughout the autosomes: 27 were located in coding sequence (CDS) regions that encompassed at least one exon, whereas 27 were located in intergenic sequence (IGS) regions that were located at least 100 kb away from any annotated genes. The functions of the genes in the CDS regions were randomly selected. The summary of the genetic diversity is shown in Table 1, and the summary statistics for each locus are presented in Table S2 (Supporting information). In total, we identified 703 SNPs from the 40.3 kb region (17.4 SNPs per kb) in the 24 *M. fascicularis* individuals.

On average, the nucleotide diversity (π) in the CDS and IGS regions was 0.26% (SE, 0.13%) and 0.35% (SE,

Table 1 Genetic diversity in *M. fascicularis* subpopulations and *M. mulatta* population across 54 control loci

Population group	N^a	θ^b	π^c	Tajima's D	$\pi_{CDS}{}^d$	$\pi_{IGS}{}^e$
Indonesian *M. fascicularis*	16	0.00345 (0.00187)	0.00306 (0.00166)	−0.316 ($P = 0.064$)	0.00258 (0.00127)	0.00355 (0.00189)
Malaysian M. *fascicularis*	14	0.00345 (0.00164)	0.00315 (0.00167)	−0.327 ($P = 0.046$)	0.00274 (0.00139)	0.00357 (0.00189)
Philippine *M. fascicularis*	18	0.00176 (0.00109)	0.00209 (0.00125)	0.626 ($P < 10^{-3}$)	0.00175 (0.00117)	0.00244 (0.00127)
All M. *fascicularis*	48	0.00387 (0.00175)	0.00306 (0.00157)	−0.636 ($P < 10^{-3}$)	0.00259 (0.00127)	0.00352 (0.00172)
Burmese *M. mulatta*	6	0.00254 (0.00282)	0.00245 (0.00264)	−0.380 ($P = 0.014$)	0.00223 (0.00117)	0.00277 (0.00197)

[a]Number of sampled chromosomes.
[b]Average number of segregating sites per site [Watterson's θ *(Watterson 1975)*], standard deviation in the parenthesis.
[c]Average nucleotide diversity per site, standard deviation in the parenthesis.
[d]Average nucleotide diversity per site in the coding regions.
[e]Average nucleotide diversity per site in the intergenic regions.

0.17%), respectively, and 0.31% in total. The nucleotide diversity in the IGS regions was four to five times higher than that of the current human population (Levy *et al.* 2007). The source of this high diversity is not a simple reflection of the geographic population structure, because the Indonesian and Malaysian samples, even though they were collected from a single country, have the same amount of genetic diversity compared with the whole *M. fascicularis* dataset (see Table 1).

In order to estimate the level of geographic differentiation among *M. fascicularis* populations, we first measured F_{ST} among the three populations. The populations from the Philippines showed a large genetic differentiation from the other two populations; the mean F_{ST} value was 0.158 between the Indonesian and Philippine macaques, and 0.182 between the Malaysian and Philippine macaques. However, the Indonesian and Malaysian populations showed little genetic differentiation between them (mean $F_{ST} = 0.039$). This result is not surprising because Indonesia and Malaysia had been periodically connected and formed Sundaland until the end of the last glacial age, but the Philippines were not connected (Heaney 1991). We also estimated the population structure using a Bayesian algorithm in the STRUCTURE program (Pritchard *et al.* 2000). Again, the Philippine population has a distinct genetic structure as shown in Fig. 1. In Fig. 1, we assigned the number of population ancestries (K) to be two. Increasing the value of K resulted in a much lower probability (Fig. S2, Supporting information), indicating we did not observe any population differentiation between the Indonesian and Malaysian macaques with a given dataset.

The divergence between M. fascicularis *and* M. mulatta

To quantify the genetic divergence between *M. fascicularis* and *M. mulatta*, five captive-born *M. mulatta* samples of Burmese origin were sequenced across the 54 loci. However, two individuals that shared some ancestry with the other individuals were excluded from the following analyses. The summary statistics are given in Table 1. Of 225 SNPs discovered in *M. mulatta* (5.59 SNPs per kb), 84 were shared between *M. fascicularis* and *M. mulatta*, and

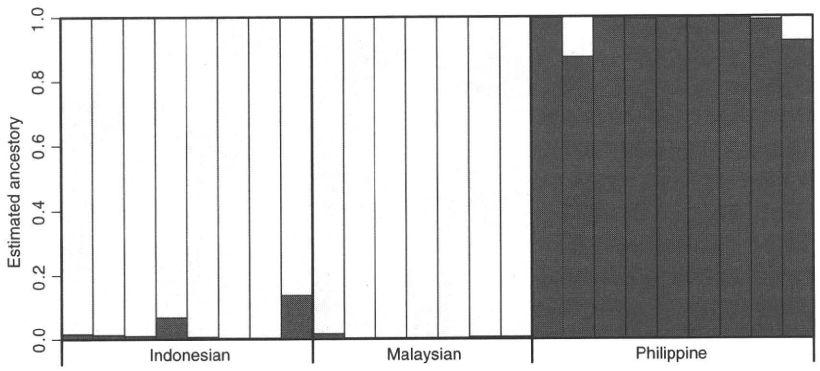


Fig. 1 Population structure of *M. fascicularis*. Population structure for 24 *M. fascicularis* samples was estimated using the STRUCTURE program (Pritchard *et al.* 2000). The vertical bars represent each *M. fascicularis* individual. The number of clusters is assumed to be two (K = 2). Samples were separated by the thick lines according to their geographic origins. Estimated ancestries are shown by white and grey colour.

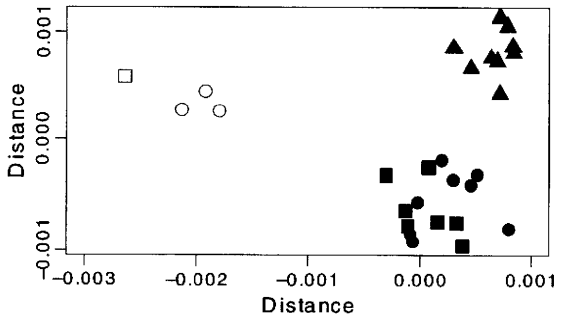


Fig. 2 Multidimensional scaling plot of the genetic distance among *M. fascicularis* and *M. mulatta* individuals. The closed circles (●), squares (■), and triangles (▲) represent each individual of Indonesian, Malaysian, and Philippine *M. fascicularis*, respectively. Indian (draft genome sequence as a representative) and Burmese *M. mulatta* individuals are marked as the opened square (□) and circles (○), respectively. Pairwise distances were estimated by Kimura's two parameters method (Kimura 1980).

141 were found only in rhesus macaques. The average nucleotide diversity of the Burmese *M. mulatta* in the IGS regions was 0.28%. Therefore, nucleotide diversity in Burmese *M. mulatta* is about 80% of that in *M. fascicularis*. Figure 2 shows a multidimensional scaling plot of the individuals from three *M. fascicularis* and *M. mulatta* populations, measured by the averaged genetic distance for all loci. The rhesus genome sequence derived from an Indian population was also added to the plot. The figure illustrates that *M. fascicularis* consists of highly variable individuals. All sampled macaques clustered into three distinct subpopulations: Indonesian-Malaysian *M. fascicularis*, Philippine *M. fascicularis*, and Indian-Burmese *M. mulatta*.

The average genetic divergence between *M. fascicularis* and *M. mulatta* (Nei's d_{xy} distance (Nei 1987)) in the IGS regions was 0.44%, and significantly higher than the nucleotide diversity within each species ($P = 0.033$, Wilcoxon test). We performed AMOVA, using the four-population data (Indonesian-Malaysian *M. fascicularis*, Philippine *M. fascicularis*, Burmese *M. mulatta*, and Indian *M. mulatta*). The total molecular variance was partitioned into the variance among individuals (70%), among populations (10%), and among species (20%), indicating that the variation among species exceeded the variation among populations. Nevertheless, throughout the 40.3 kb region in the survey, we found only one fixed nucleotide difference between the two macaque species. There are two explanations for the many shared SNPs and few fixed changes between species: (i) the incomplete lineage sorting of ancestral polymorphisms and (ii) the historical gene flow between the two species. Although ancestral polymorphisms may be largely responsible for the shared SNPs, we wished to test whether there was any contribution of inter-species gene flow to the genetic differentiation. In particular, by excluding the samples from current hybrid zone, we would be able to investigate the effect of historical gene flow between the species.

In order to quantify the relative contribution of the ancestral polymorphisms and gene flow to the observed amount of shared SNPs between the two macaques, we conducted MCMC simulations assuming the isolation with migration (IM) model (Hey & Nielsen 2007). To exclude candidate SNPs under natural selection, we applied the BayeScan program, which finds genetic variants showing significantly high or low F_{ST} (Foll & Gaggiotti 2008). Since there was no SNP having the Bayes factor higher than 10 (strong signature of directional selection) or less than 0.1 (strong signature of balancing selection) in the 54 loci, we included all SNPs to the population parameter estimation.

Because estimating population parameters among many subpopulations may be possible but is computationally too cumbersome, we separately estimated population parameters using data within species (between Indonesian-Malaysian and the Philippine *M. fascicularis*) and between species (between Indonesian-Malaysian *M. fascicularis* and Burmese *M. mulatta*). Since cryptic population structure within species may distort the parameter estimation between species, we excluded the samples of *M. fascicularis* in the Philippines for the between-species analysis. All estimated parameters and marginal distributions of posterior probability are presented in Table 2 and Fig. S2 (Supporting information), respectively. The population parameters were scaled by the mutation rate that was estimated using the human genome sequence as an outgroup, assuming a 25-million-year divergence between humans and macaques (Stewart & Disotell 1998) and a 6-year generation time in macaques. The results showed that the two macaque species started to split at 1.52 Ma (1.23–1.86 Ma, 95% CI) and have been exchanging their genes. The separation time between Indonesian-Malaysian and Philippine macaques was estimated to be 1.69 Ma (1.30–2.02 Ma, 95% CI). Population migration rate ($2N_em$) was estimated to be relatively small but significantly positive values: 0.44 from Indonesian-Malaysian *M. fascicularis* to *M. mulatta* and 0.35 from *M. mulatta* to Indonesian-Malaysian *M. fascicularis*. We rigorously tested the significance of the amount of inter-species gene flow using the nested model where the migration rate parameters toward either direction or both directions are fixed to zero. In all cases, null models of no migration were rejected ($P < 10^{-16}$, from *M. fascicularis* to *M. mulatta*; $P < 10^{-19}$, from *M. mulatta* to *M. fascicularis*; $P < 10^{-196}$, no migration for both directions). Between Indonesian-Malaysian and Philippine macaques, migration rate is

Table 2 Estimates of population parameter using MCMC method

	Estimated values[a]	95% CI
Between Indonesian-Malaysian (Sunda) and Philippine *M. fascicularis*		
N_1 (*Sunda*)	197 000	(169 000–229 000)
N_2 (*Philippine*)	23 000	(13 000–36 000)
N_a (ancestral)	9 000	(400–40 000)
$2N_1m_1$ (*Sunda* to *Philippine*)	2.23	(0.93–3.37)
$2N_2m_2$ (*Philippine* to *Sunda*)	0.56	(0.30–1.08)
Separation time (years)	1 690 000	(1 300 000–2 020 000)
Between Indonesian-Malaysian *M. fascicularis* and Burmese *M. mulatta*		
N_1 (*M. fascicularis*)	181 400	(155 900–210 000)
N_2 (*M. mulatta*)	110 000	(82 000–144 000)
N_a (ancestral)	33 000	(13 000–57 000)
$2N_1m_1$ (*M. fascicularis* to *M. mulatta*)	0.44	(0.21–0.75)
$2N_2m_2$ (*M. mulatta* to *M. fascicularis*)	0.35	(0.14–0.67)
Separation time (years)	1 520 000	(1 230 000–1 860 000)

[a]The values are the peak of marginal distributions. The parameters were scaled assuming mutation rate per year per locus is 6.26×10^{-7} and generation time is 6 years.

2.23 from Indonesian-Malaysian to Philippine *M. fascicularis* and 0.56 vice versa, which exceeded those between species. Population size of Indonesian-Malaysian *M. fascicularis*, Philippine *M. fascicularis*, Burmese *M. mulatta*, and common ancestors of them were estimated to be 181 000–197 000, 23 000, 110 000, and 9000–33 000, respectively.

Demography of macaque populations

Samples from the Philippines exhibited a different pattern for the SNP frequency distribution. We observed negative Tajima's *D* values in the Indonesian-Malaysian *M. fascicularis* and Burmese *M. mulatta*, which agrees with the population size expansion and weak gene flow between the species in the IM analysis. By contrast, the Tajima's *D* statistic in the Philippine population was highly skewed toward positive values compared with the other populations (Table 1). A positive Tajima's *D* value is due to an excess of intermediate frequency SNPs relative to rare SNPs, suggesting that the population size of Philippine macaques did not increase from a very small founder population (see 'Discussion'). We found that the distribution significantly deviated from the neutral expectation under a model of constant population size and no migration, using standard coalescent simulations ($P < 10^{-3}$).

Identifying highly differentiated loci between species

In the above analyses, we showed that the two macaque species have been exchanging their genes for a considerable period of time. The situation of these two macaques may thus match speciation under parapatry (Wu 2001). While genetic admixture tends to homogenize the genetic differentiation between species, local adaptation can act as a force that maintains the standing differentiation. When natural selection acts on genetic differentiation under parapatric speciation, we expect systematic differences in the species divergence between functional and nonfunctional regions, e.g. CDS and IGS (see Wu 2001; Osada & Wu 2005). We test the hypothesis using the method developed by Osada & Wu (2005). In our control loci, however, we found no significant difference in the genomic differentiation between the CDS and IGS regions, indicating that the randomly selected genic regions have similar evolutionary dynamics to the nongenic regions. Therefore, we investigated genomic regions that may be more relevant to local adaptation. Because *M. fascicularis* and *M. mulatta* are popular subjects for pharmacological research, we selected seven CYP genes to test for natural selection on their genetic differentiation. CYP genes are major metabolism genes for many internal and external substrates across a wide range of organisms.

We sequenced one randomly selected region for each CYP gene. Although many of the statistics within population, such as nucleotide diversity and Tajima's *D*, were not significantly different between CYP and control CDS regions, some of the patterns of divergence between species in the CYP regions were quite dissimilar to those in the control CDS regions (Table 3). The species divergence at the silent sites of the CYP genes was significantly greater than that of the other CDS regions ($P = 0.0034$, Wilcoxon test). In particular, *CYP2D6* shows the highest inter-species divergence among all the sequenced loci (see Table 3). The BayeScan analysis identified four SNPs that showed strong evidence (Bayes factor > 10) of natural selection, one in *CYP3A5* and three in *CYP2D6*. Interestingly, we found one fixed change between species in the first intron of *CYP3A5*.

Because we did not find any nonsynonymous substitutions in the partially sequenced region of *CYP3A5* (exon 1), we surveyed SNPs in the entire coding region of *CYP3A5* using an additional 74 Indochinese *M. fascicularis*, 80 Indonesian *M. fascicularis*, and 68 Chinese *M. mulatta*. In total, we found 38 SNPs within the three populations. For these SNPs, the ancestral alleles were

Table 3 Polymorphism and divergence in seven CYP loci

Gene	Region (bp)	π_f[a]	π_m[a]	d_{xy}[b]	F_{ST}[c]	A/S[d]	D_f[e]	Region
CYP1A2	696	0.00859	0.01222	0.01236	0.1236	18/10	−1.134	exon2, intron
CYP2A6	799	0.00508	0.00687	0.00881	0.2238	1/4	−0.353	exon5, intron
CYP2C9	809	0.00443	0.00597	0.00956	0.4072	1/5	−0.594	exon1, intron
CYP2C18	731	0.00516	0.00223	0.00569	0.2258	0/3	−0.404	exon2, intron
CYP2D6	861	0.00510	0.00523	0.01654	0.6603	4/4	−0.244	exon8, 9, intron
CYP3A5	715	0.00037	0.00430	0.00422	0.5263	0/0	−1.668	exon1, intron
CYP3A7	798	0.00499	0.00635	0.00653	0.1231	0/0	−0.220	3′ UTR, intergenic
other CDSs[f]	760.1	0.00388	0.00339	0.00496	0.2822	30/59	−0.676	

[a]Nucleotide diversity at silent (synonymous and noncoding) sites.
[b]Silent substitution rate per site averaged by all *fascicularis–mulatta* pairwise comparison.
[c]F_{ST} between *M. fascicularis* and *M. mulatta*.
[d]Number of nonsynonymous (A) and synonymous (S) SNPs in *M. fascicularis*.
[e]Tajima's *D* statistics for *M. fascicularis*.
[f]Average of 27 autosomal control CDS regions.

determined using a baboon cDNA sequence from a public database. Of these 38 SNPs, 9 are nonsynonymous, 17 are synonymous, 1 is nonsense, and 9 are noncoding (Uno *et al.* 2010). We found three nonsynonymous SNPs and one synonymous SNP that have a high derived allele frequency in either of the species. These high derived allele frequency sites showed an excess of nonsynonymous sites compared with polymorphisms within species ($P = 0.048$, Fisher's exact test). For two SNPs (D88E and T230I, see Fig. 3), Indonesian *M. fascicularis* and Burmese *M. mulatta* were completely segregated at the nonsynonymous sites, whereas a few Indochinese *M. fascicularis* carried *mulatta*-type alleles, which may be due to the ongoing introgression between these two populations. As shown in Fig. 3, the region spanning exon 3 to 6 showed a high level of genetic differentiation between species, but a moderate level of polymorphism within species.

We inferred a functional significance of the observed three nonsynonymous SNPs using the three-dimensional protein structure of CYP3A5. The protein structure of CYP3A5 was modelled using the crystal structure of human CYP3A4, which is highly similar to CYP3A5 (83% amino acid identity). All three amino

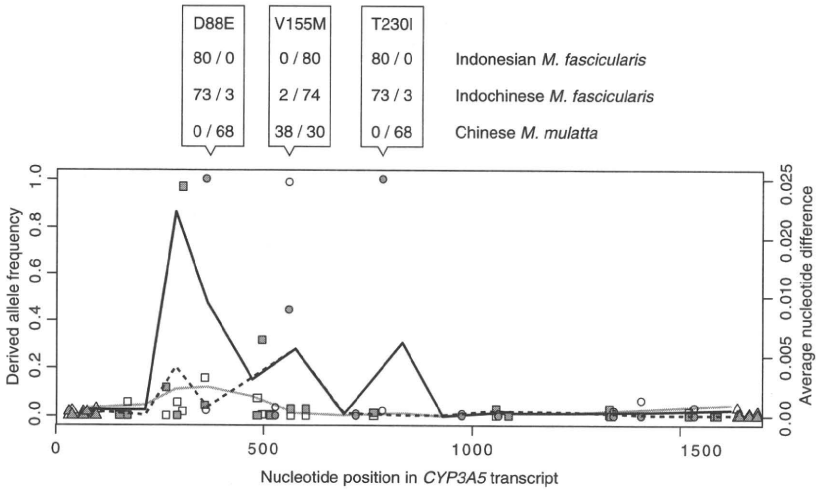


Fig. 3 Pattern of polymorphisms and divergence between *M. fascicularis* and *M. mulatta* across 13 *CYP3A5* exons. Derived allele frequency of noncoding, synonymous, and nonsynonymous SNPs are marked as triangles (▲), squares (■), and circles (●), respectively. Frequency in *M. fascicularis* is shown by opened symbols and frequency *M. mulatta* in by closed symbols. Boundary of exons is shown by the dashed line. Solid, dashed, and grey lines represent the average divergence between species (d_{xy}), nucleotide diversity in *M. fascicularis* (π_f), and nucleotide diversity in *M. mulatta* (π_m), in each exon, respectively. For three nonsynonymous SNPs that are highly differentiated between species, the numbers of ancestral (left) and derived (right) alleles in Indonesian *M. fascicularis*, Indochina *M. fascicularis*, and Burmese *M. mulatta* are written in the upper space of the figure.

acid variants are located in helical regions on the protein surface. Interestingly, the D88E and T230I polymorphisms are assigned to helix B and G′, respectively (see Fig. S4). Previous studies have shown that the B–C loop and helix G′ form a channel for substrates and are responsible for substrate specificity (Williams *et al.* 2004; Ekroos & Sjogren 2006).

Discussion

Genetic diversity of M. fascicularis

In this report, we surveyed the genetic diversity of *M. fascicularis* and inferred its genetic relationship to *M. mulatta*. As shown in Table 1, *M. fascicularis* has a considerable amount of genetic variation within and among its geographic populations. In particular, *M. fascicularis* in the Philippines exhibited a population structure distinct from the other populations of this species.

Gene flow between M. fascicularis *and* M. mulatta

The estimation of population parameters under the IM model suggests that there has been a significant amount of inter-species gene flow. Recent studies have also revealed nonzero rates of gene flow between *M. fascicularis* and *M. mulatta* using different samples (Kanthaswamy *et al.* 2008; Bonhomme *et al.* 2009; Stevison & Kohn 2009). Previous studies also showed strong unidirectional gene flow ($2N_em \approx 10$) from *M. mulatta* to *M. fascicularis*. Our estimation of the migration rate, however, was much smaller ($2N_em < 1$) than the previous estimates, and did not significantly differ in direction. One of the reasons is that we did not analyse Indochina *M. fascicularis* samples in the IM analysis, which may be under strong ongoing gene introgression. The significance of our study is that we show that lineage sharing between the two species beyond the hybrid zone, and this unlikely is due to ancestral polymorphism.

We should note that we did not sample all of the extant populations of the macaques. The IM model assumes that there are no hidden population structures or unsampled populations in the analysis. Unfortunately, this assumption is not realistic and hardly satisfied in the macaque species, because they are distributed over wide geographic ranges and their population sizes are relatively large. Nonetheless, a significant amount of gene flow from *M. mulatta* to *M. fascicularis* in the past is highly likely because the estimated migration rates are fairly high, and our results are consistent with the previous studies using different populations and different types of data.

This study agrees with the study of Stevison & Kohn (2009) that the start of the separation of these two species occurred about 1.23 Ma, which was slightly more recent than the previous estimates of around 2 Ma deduced from the mitochondrial locus (Hayasaka *et al.* 1996; Blancher *et al.* 2008), and much older than the estimates of 43 Ka using nuclear microsatellite markers (Bonhomme *et al.* 2009). If we assume that the human–macaque divergence occurred 35 Ma instead of 25 Ma, the speciation time estimate would become 2.13 Ma. In addition, the discordance partially relies on whether we consider ancestral polymorphisms. Previous studies using the mitochondrial locus did not consider an effect of ancestral polymorphisms on the species divergence time. Since the depth of a genealogy is a combination of the species divergence time after speciation and the coalescent time in the ancestral population, one tends to overestimate speciation time without considering ancestral polymorphisms. We also have to note that the IM model assumes that migration rates and effective population sizes for males and females are equal, which may be violated to some extent, as suggested by previous studies (Tosi *et al.* 2000, 2003).

Our analysis suggests a gene introgression between species over 1 million years. The period was mostly occupied by the middle to late Pleistocene. During that period, the Indonesian Islands were connected to mainland Asia and formed Sundaland at least twice (Heaney 1991). Interestingly, we did not observe any strong population structure between *M. fascicularis* from Indonesia and Peninsular Malaysia, and they harboured a very high genetic diversity within populations. We suggest that, during the glacial periods, the climatic change may have altered the natural habitats of *M. fascicularis*, and effectively mixed a gene pool of *M. fascicularis* in Sundaland by the dispersal from glacial refugia or migration of local populations. Through these migration events, the hybrid zone of *M. fascicularis* and *M. mulatta* may have shifted, expanded, or contracted, and genes migrated from *M. mulatta* could have spread within *M. fascicularis* in Sundaland and vice versa.

In this study, we used only three to eight individuals of macaques for each population to infer the population structure and demographic history. Although a small sample size might cause the biased estimation of parameters, recent studies surveying a large number of loci showed that a small number of samples does not seriously affect the estimation of population parameters (e.g., Shi *et al.* 2010). In an extreme case, we could infer the history of populations from the whole genome sequences of a few individuals (Huff *et al.* 2010). However, in future studies, sampling of macaques from a wider geographic range would be invoked to understand the detailed genetic structure of extant macaque populations.

Genetic differentiation of Philippine M. fascicularis

Because the Philippine Islands have been unconnected or only partially connected to the mainland during the glacial ages, previous studies have suggested that the low genetic diversity in Philippine macaques may be due to a founding event with a very small number of founders by rafting or human introduction (Smith *et al.* 2007). However, by analysing the multilocus DNA data, we found that the Tajima's *D* statistic in the Philippine population was significantly positive (Table 1). Under the population bottleneck model, whether Tajima's *D* statistic become positive or negative depends on the level of reduction, the time of population bottleneck, and the time after the bottleneck (Tajima 1989; Fay & Wu 1999). If the reduction of population size is severe, Tajima's *D* value drops quickly after the bottleneck and recovers to an equilibrium state after a long time. Therefore, we suggest that a single founding event with a small number of individuals to the Philippine Islands is unlikely.

We should note that the previous studies using mitochondrial markers have shown much smaller genetic diversity of Philippine macaques (Smith *et al.* 2007; Blancher *et al.* 2008). Our estimation of the divergence time between Indonesian-Malaysian and Philippine *M. fascicularis* was comparable to the divergence time between *M. fascicularis* and *M. mulatta*. Considering that there has been a high level of gene flow between the *M. fascicularis* populations, many loci would have been migrated between the populations while some loci still remain highly diverged.

The data might be explained by multiple founding events from Sundaland to the Philippines. We hypothesize that the initial isolation of the Philippine population started around 1–2 Ma. After the initial isolation, there might have been additional migration events from Sundaland to the Philippines. The short coalescence branches of the mitochondrial genomes in Philippine macaques may indicate that the recently migrated mitochondrial genomes spread into the population by natural selection or genetic drift. If population contraction or bottleneck had occurred in the Philippines, the chance of fixation of migrated mitochondrial alleles would increase. The previous studies suggested the male-biased migration pattern of macaques, which predicts a deep geographic structure in the mitochondrial genome than the rest of the genome (Tosi *et al.* 2000). However, we observed the opposite pattern, i.e., shallow genealogies in mitochondrial genome and deep genealogies in some other loci. The pattern indicates that the migration to the Philippines might not have been male biased. The analysis using multiple individuals from multiple islands may elucidate the complex population structure and demography of the Philippine macaques.

Inference of natural selection on CYP genes

In contrast to the control regions, CYP genes showed a significantly high genetic divergence between the species. We selected the *CYP3A5* gene for the deep sequencing analysis because, among the CYP loci we surveyed, *CYP3A5* was the only locus that contained fixed nucleotide changes between the macaques. Here we focused on the differentiation between *M. fascicularis* and *M. mulatta* in the *CYP3A5* gene. The detailed polymorphism analysis within *M. fascicularis*, including a functional validation, was presented in Uno *et al.* (2010). Some evidence shows that the unusual pattern of species differentiation in *CYP3A5* has been driven by adaptive evolution in each species. First, highly differentiated sites were significantly enriched with nonsynonymous SNPs, compared with the other sites segregating within the populations. In addition, two of the three highly differentiated nonsynonymous sites are likely to contribute to the substrate specificity of CYP3A5, considering the predicted conformational structure of the protein. The other site, V155M, may or may not be involved in substrate recognition.

The unusual pattern of SNPs in the CYP genes indicates that the beneficial alleles in one species have not been advantageous, or perhaps have been deleterious, in the other species, and the migration to the other species has been impeded by the species boundary. The pattern may well fit the model that the two species have evolved under parapatry, where nascent species have gradually diverged with a genetic connection (Wu 2001). Both our MCMC estimation of the gene flow rate and the generally high species divergence in the CYP genes corroborate this model.

Conclusion

In conclusion, the genetic structure of *M. fascicularis* and *M. mulatta*, although they have been widely used for a variety of biomedical research and deemed as nominal species, may be very complex because of the past genetic admixture and effect of natural selection on their genomes. For some loci, a *M. fascicularis* individual may be genetically closer to *M. mulatta* than the conspecific siblings in the next cage. Therefore, we have to be aware of their genetic heterogeneity and control their genetic background to attain highly reliable biomedical studies. Our samples showed that 70% of the total genetic variance in the macaques is attributed to the genetic variance within regional populations, indicating that the restriction of sampling populations for

biomedical studies does not significantly help. Conversely, the high genetic diversity of macaques and their heterogeneous pattern of genetic admixture will give us an opportunity to search for causative genetic differences for important adaptive traits within and between species. The CYP genes are good candidates for such studies.

Acknowledgements

We thank Dr Tetsuro Matano (University of Tokyo) for the blood samples of *M. mulatta*. We wish to acknowledge valuable comments from three anonymous reviewers. This study was supported by a Health Science Research Grant from the Ministry of Health, Labor and Welfare of Japan and a Grant-in-Aid for Young Scientists (B) from the Ministry of Education, Culture, Sports, Science and Technology, Japan (19770073).

Conflict of interests

The authors declare that there is no conflict of interests in the manuscript.

References

Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, **22**, 195–201.

Blancher A, Bonhomme M, Crouau-Roy B *et al.* (2008) Mitochondrial DNA sequence phylogeny of 4 populations of the widely distributed cynomolgus macaque (*Macaca fascicularis fascicularis*). *Journal of Heredity*, **99**, 254–264.

Bonhomme M, Cuartero S, Blancher A, Crouau-Roy B (2009) Assessing natural introgression in 2 biomedical model species, the rhesus macaque (*Macaca mulatta*) and the long-tailed macaque (*Macaca fascicularis*). *Journal of Heredity*, **100**, 158–169.

Ekroos M, Sjogren T (2006) Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proceedings of the National Academy of Sciences, USA*, **103**, 13682–13687.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Fay J, Wu C (1999) A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Molecular Biology and Evolution*, **16**, 1003–1005.

Ferguson B, Street SL, Wright H *et al.* (2007) Single nucleotide polymorphisms (SNPs) distinguish Indian-origin and Chinese-origin rhesus macaques (*Macaca mulatta*). *BMC Genomics*, **8**, 43.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Fooden J (1964) Rhesus and crab-eating macaques: intergradation in Thailand. *Science*, **143**, 363–364.

Fooden J (1976) Provisional classifications and key to living species of macaques (primates: Macaca). *Folia Primatol (Basel)*, **25**, 225–236.

Gibbs RA, Rogers J, Katze MG *et al.* (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222–234.

Hamada Y, Urasopon N, Hadi I, Malaivijitnond S (2006) Body size and proportions and pelage color of free-ranging *Macaca mulatta* from a zone of hybridization in northeastern Thailand. *International Journal of Primatology*, **27**, 497–513.

Hayasaka K, Fujii K, Horai S (1996) Molecular phylogeny of macaques: implications of nucleotide sequences from an 896-base pair region of mitochondrial DNA. *Molecular Biology and Evolution*, **13**, 1044–1053.

Heaney LR (1991) A synopsis of climatic and vegetational change in Southeast Asia. *Climatic Change*, **19**, 53–61.

Hernandez RD, Hubisz MJ, Wheeler DA *et al.* (2007) Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. *Science*, **316**, 240–243.

Hey J, Nielsen R (2007) Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of National Academy of Sciences, USA*, **104**, 2785–2790.

Hobolth A, Christensen OF, Mailund T, Schierup MH (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genetics*, **3**, e7.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Huff CD, Xing J, Rogers AR, Witherspoon D, Jorde LB (2010) Mobile elements reveal small population size in the ancient ancestors of *Homo sapiens*. *Proceedings of the National Academy of Sciences USA*, **107**, 2147–2152.

Innan H, Watanabe H (2006) The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. *Molecular Biology and Evolution*, **23**, 1040–1047.

Kanthaswamy S, Satkoski J, George D *et al.* (2008) Hybridization and stratification of nuclear genetic variation in *Macaca mulatta* and *M. fascicularis*. *International Journal of Primatology*, **29**, 1295–1311.

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

Levy S, Sutton G, Ng PC *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biology*, **5**, e254.

Magness CL, Fellin PC, Thomas MJ *et al.* (2005) Analysis of the *Macaca mulatta* transcriptome and the sequence divergence between *Macaca* and human. *Genome Biology*, **6**, R60.

Malhi RS, Sickler B, Lin D *et al.* (2007) MamuSNP: a resource for Rhesus Macaque (*Macaca mulatta*) genomics. *PLoS ONE*, **2**, e438.

Matsumoto J, Kawai S, Terao K *et al.* (2000) Malaria infection induces rapid elevation of the soluble Fas ligand level in serum and subsequent T lymphocytopenia: possible factors responsible for the differences in susceptibility of two species of *Macaca* monkeys to *Plasmodium coatneyi* infection. *Infection and Immunity*, **68**, 1183–1188.

Melnick DJ, Hoelzer GA, Absher R, Ashley MV (1993) mtDNA diversity in rhesus monkeys reveals overestimates of

divergence time and paraphyly with neighboring species. *Molecular Biology and Evolution*, **10**, 282–295.
Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.
Osada N, Wu CI (2005) Inferring the mode of speciation from genomic data: a study of the great apes. *Genetics*, **169**, 259–264.
Osada N, Hashimoto K, Kameoka Y *et al.* (2008) Large-scale analysis of *Macaca fascicularis* transcripts and inference of genetic divergence between *M. fascicularis* and *M. mulatta*. *BMC Genomics*, **9**, 90.
Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D (2006) Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, **441**, 1103–1108.
Pettersen EF, Goddard TD, Huang CC *et al.* (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, **25**, 1605–1612.
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
Rozas J, Sanchez-Delbarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.
Shi W, Ayub Q, Vermeulen M *et al.* (2010) A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Molecular Biology and Evolution*, **27**, 385–393.
Sibal LR, Samson KJ (2001) Nonhuman primates: a critical role in current disease research. *ILAR Journal*, **42**, 74–84.
Smith DG, Mcdonough JW, George DA (2007) Mitochondrial DNA variation within and among regional populations of longtail macaques (*Macaca fascicularis*) in relation to other species of the fascicularis group of macaques. *American Journal of Primatology*, **69**, 182–198.
Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.
Stevison LS, Kohn MH (2008) Determining genetic background in captive stocks of cynomolgus macaques (*Macaca fascicularis*). *Journal of Medical Primatology*, **37**, 311–317.
Stevison LS, Kohn MH (2009) Divergence population genetic analysis of hybridization between rhesus and cynomolgus macaques. *Molecular Ecology*, **18**, 2457–2475.
Stewart CB, Disotell TR (1998) Primate evolution: in and out of Africa. *Current Biology*, **8**, R582–588.
Street SL, Kyes RC, Grant R, Ferguson B (2007) Single nucleotide polymorphisms (SNPs) are highly conserved in rhesus (*Macaca mulatta*) and cynomolgus (*Macaca fascicularis*) macaques. *BMC Genomics*, **8**, 480.
Tajima F (1989) The effect of change in population size on DNA polymorphism. *Genetics*, **123**, 597–601.
Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.
Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.
Thompson EE, Kuttab-Boulos H, Witonsky D *et al.* (2004) CYP3A variation and the evolution of salt-sensitivity variants. *American Journal of Human Genetics*, **75**, 1059–1069.
Tosi AJ, Morales JC, Melnick DJ (2000) Comparison of Y chromosome and mtDNA phylogenies leads to unique inferences of macaque evolutionary history. *Molecular Phylogenetics and Evolution*, **17**, 133–144.
Tosi AJ, Morales JC, Melnick DJ (2002) Y-chromosome and mitochondrial markers in *Macaca fascicularis* indicate introgression with Indochinese *M. mulatta* and a biogeographic barrier in the isthmus of Kra. *International Journal of Primatology*, **23**, 161–178.
Tosi AJ, Morales JC, Melnick DJ (2003) Paternal, maternal, and biparental molecular markers provide unique windows onto the evolutionary history of macaque monkeys. *Evolution*, **57**, 1419–1435.
Trichel AM, Rajakumar PA, Murphey-Corb M (2002) Species-specific variation in SIV disease progression between Chinese and Indian subspecies of rhesus macaque. *Journal of Medical Primatology*, **31**, 171–178.
Uno Y, Matsushita A, Osada N *et al.* (2010) Genetic variants of *CYP3A4* and *CYP3A5* in cynomolgus and rhesus macaques. *Drug Metabolism and Disposition*, **38**, 209–214.
Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.
Wheatley BP (1980) Malaria as a possible selective factor in the speciation of macaques. *Journal of Mammology*, **61**, 307–311.
Williams PA, Cosme J, Vinkovic DM *et al.* (2004) Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science*, **305**, 683–686.
Woerner AE, Cox MP, Hammer MF (2007) Recombination-filtered genomic datasets by information maximization. *Bioinformatics*, **23**, 1851–1853.
Wu CI (2001) The genic view of the process of speciation. *Journal of Evolution Biology*, **14**, 851–865.
Yano JK, Wester MR, Schoch GA *et al.* (2004) The structure of human microsomal cytochrome P450 3A4 determined by X-ray crystallography to 2.05-A resolution. *Journal of Biological Chemistry*, **279**, 38091–38094.

Naoki Osada is an Assistant Professor in the Department of Population Genetics at the National Institute of Genetics and interested in the process of genome evolution and speciation of various organisms. Yasuhiro Uno is a Group Leader of Genome Research Group at Shin Nippon Biomedical Laboratories (SNBL) and studies the genetic basis of drug metabolism in laboratory macaques. Katsuhiko Mineta is an Associate Professor at Hokkaido University. He is interested in the evolution of a gene network. Yosuke Kameoka and Ichiro Takahashi are both Senior Researchers in the Department of Disease Bioresources Research at the National Institute of Biomedical Innovation. They study genome resources of *Macca fascicularis*. Keiji Terao was a Director of Tsukuba Primate Research Center at the National Institute of Biomedical Innovation and worked for establishing biomedical research resources of non-human primates.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Geographic distribution of fascicularis group macaques by Fooden (1976). Possible hybrid zone is shown in blue color.

Fig. S2 Population structure of *M. fascicularis*. Population structure for 24 *M. fascicularis* samples was estimated using STRUCTURE program. The vertical bars represent each *M. fascicularis* individual. The number of clusters is assumed to be three ($K = 3$).

Fig. S3 Marginal posterior density plots of population parameters. (A) Population size of Indonesian-Malaysian *M. fascicularis* (blue), Philippine *M. fascicularis* (red), and common ancestors (black). (B) Migration rate per mutation from Indonesian-Malaysian to Philippine *M. fascicularis* (blue) and from Philippine to Indonesian-Malaysian *M. fascicularis* (red). (C) Divergence time between Indonesian-Malaysian and Philippine *M. fascicularis*. D) Population size of Indonesian-Malaysian *M. fascicularis* (blue), Burmese *M. mulatta* (red), and common ancestors (black). (E) Migration rate per mutation from Indonesian-Malaysian *M. fascicularis* to *M. mulatta* (blue) and from *M. mulatta* to Indonesian-Malaysian *M. fascicularis* (red). (F) Divergence time between Indonesian-Malaysian *M. fascicularis* and *M. mulatta*.

Fig. S4 Predicted three-dimensional protein structure of macaque CYP3A5. The model was based on the crystal structure of human CYP3A4. The heme group is shown as ball and stick representation. Three amino acid sites highly differentiated between *M. fascicularis* and *M. mulatta* are shown in the figure. Two sites (D88E and T230I) may be involved in the substrate specificity of macaque CYP3A5.

Table S1 Statistics for all loci

Table S2 Sample information

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Genes to Cells

Lectin microarray analysis of pluripotent and multipotent stem cells

Masashi Toyoda[1], Mayu Yamazaki-Inoue[1], Yoko Itakura[2], Atsushi Kuno[2], Tomohisa Ogawa[3], Masao Yamada[3], Hidenori Akutsu[1], Yuji Takahashi[1], Seiichi Kanzaki[1], Hisashi Narimatsu[2], Jun Hirabayashi[2] and Akihiro Umezawa[1]*

[1]*Department of Reproductive Biology, National Institute for Child Health and Development, 2-10-1 Okura, Setagaya-ku, Tokyo 157-8535, Japan*

[2]*Research Center for Medical Glycoscience, National Institute of Advanced Industrial Science and Technology, AIST Tsukuba Central 2, Tsukuba, Ibaraki 305-8568, Japan*

[3]*GP BioSceinces Ltd, 1-3-3, Azamino-Minami, Aoba-ku, Yokohama, Kanagawa 225-0012, Japan*

Stem cells have a capability to self-renew and differentiate into multiple types of cells; specific markers are available to identify particular stem cells for developmental biology research. In this study, we aimed to define the status of somatic stem cells and the pluripotency of human embryonic stem (hES) and induced pluripotent stem (iPS) cells using a novel molecular methodology, lectin microarray analysis. Our lectin microarray analysis successfully categorized murine somatic stem cells into the appropriate groups of differentiation potency. We then classified hES and iPS cells by the same approach. Undifferentiated hES cells were clearly distinguished from differentiated hES cells after embryoid formation. The pair-wise comparison means based on 'false discovery rate' revealed that three lectins -*Euonymus europaeus* lectin (EEL), *Maackia amurensis* lectin (MAL) and *Phaseolus vulgaris* leucoagglutinin [PHA(L)]- generated maximal values to define undifferentiated and differentiated hES cells. Furthermore, to define a pluripotent stem cell state, we generated a discriminant for the undifferentiated state with pluripotency. The discriminant function based on lectin reactivities was highly accurate for judgment of stem cell pluripotency. These results suggest that glycomic analysis of stem cells leads to a novel comprehensive approach for quality control in cell-based therapy and regenerative medicine.

Introduction

Stem cells produce almost every tissue of the human body. In general, they have the ability to divide and self-renew and to differentiate into various cell types. Stem cells have varying degrees of differentiation potential: (i) totipotency (ability to form the embryo and the trophoblast of the placenta) like fertilized eggs (zygotes); (ii) pluripotency (ability to differentiate into almost all cells that arise from the three germ layers) like human embryonic stem (hES) cells and induced pluripotent stem (iPS) cells; (iii) multipotentiality (capability of producing a limited range of differentiated cell lineages upon their location) like most tissue-based stem cells; and (iv) unipotentiality (ability to generate one cell type) like cells such as the epidermal stem cells and the spermatogonial cells of the testis. That is, a hierarchy of stem cells exists. In addition, human ES cell lines show variation in differentiation propensity (Osafune *et al.* 2008). iPS cells, another type of pluripotent stem cell, have been generated from somatic cells of different origin by retroviral transduction of four transcription factors (Takahashi *et al.* 2007; Yu *et al.* 2007). The established iPS cells have a wider variety of differentiation ability and gene expression when compared to ES cells (Aoi *et al.* 2008; Lee *et al.* 2009; Kaichi *et al.* 2010). However, a small proportion of these stem cells sometimes show spontaneous differentiation during serial passage. Therefore, to realize the potential for iPS cells to be utilized for cell therapy and as a valuable tool for drug discovery, it is necessary to monitor the status of these stem cells and to define

Communicated by: Takashi Tada
**Correspondence*: umezawa@1985.jukuin.keio.ac.jp

DOI: 10.1111/j.1365-2443.2010.01459.x

their exact stage during processes of growth and/or differentiation.

Glycosylation is a critical post- or co-translational modification found in more than 50% of eukaryotic proteins (Budnik *et al.* 2006). Thus, the glycome, which represents the total set of glycans expressed in a cell, is believed to be information-rich, as it varies among cell types, stages of development and differentiation, and even in the malignant transformation processes (Varki 1993). Lectins have long been used as tools to characterize cell surface glycans, such as for blood-group typing, tissue staining, lectin-probed blotting and flow cytometry (Sharon & Lis 2004). The use of lectins in glycan profiling provides considerable advantages. A modern technology to discriminate glycan profiling is lectin microarray analysis, which is an emerging technology that enables ultrasensitive detection of multiplex lectin–glycan interactions (Angeloni *et al.* 2005; Kuno *et al.* 2005; Pilobello *et al.* 2005). The system developed by Kuno *et al.* (2005) is based on a unique principle, that is, the evanescent-field fluorescence-detection principle, which has been used extensively for biosensors to study real-time binding events on the glass slide surfaces. Thus, the evanescent-field methods have greater advantage to analyze relatively weak interactions between lectins and glycoproteins in a liquid phase at equilibrium. Furthermore, this method is applicable for the analysis of the physiological and pathological status of crude glycoproteins extracted from mammalian cells (Ebe *et al.* 2006; Kuno *et al.* 2008) and cell surfaces (Tateno *et al.* 2007). Although the number of probes in lectin microarray is much smaller than in mRNA expression arrays, lectin microarray analysis enables high-throughput and sensitive analysis of a large set of biological samples and provides a snapshot of cell profiling. In this study, we further developed lectin microarray technology to define the status of somatic and pluripotent stem cells. The glycan-based comprehensive approach promises to be of great value, complementing more established methods such as gene expression analysis and epigenetic analysis.

Results

Lectin microarray analysis of mouse mesenchymal cells

Mesenchymal stem cells are multipotent and therefore may be useful in cell-based therapy along with ES cells and iPS cells. Mesenchymal stem cell (MSC) lines [(9-15c), osteoblasts (KUSA-A1), chondroblasts (KUM5) and preadipocytes (H-1/A)] were established from mouse bone marrow and were shown to retain potency both *in vivo* and *in vitro* (Umezawa *et al.* 1991; Matsumoto *et al.* 2005; Sugiki *et al.* 2007). To investigate their carbohydrate structures, we carried out a lectin microarray analysis of the cell membrane proteins. We quantified lectin signal using 'Array-Pro Analyzer' software and calculated the average net intensities of three spots for each lectin on the chip (Fig. 1A). Experiments with each cell line were performed in triplicate or quadruplicate. Four mesenchymal cell lines with different potencies showed differential lectin reactivities. 9-15c MSCs showed strong reactivity to wheat germ agglutinin (WGA), *Lycopersicon esculentum* lectin (LEL), concanavalin A (ConA), *Sambucus nigra* agglutinin (SNA) and *Ricinus communis* agglutinin I (RCA120) (Fig. 1A and Fig. S1 in Supporting Information). These signal intensities by lectin microarray were consistent with mean fluorescent intensities by flow cytometric analysis (Fig. 1B). We then performed hierarchical clustering analysis and principal component analysis (PCA) on the signal values of each lectin (Fig. 1C, D). H-1/A preadipocytes can be distinguished by KUM5 chondroblasts by lectin reactivities of GSL1A4, GSL1B4, BPL, PWM and MPA (PC1 axis), and 9-15c MSCs can be distinguished by KUSA-A1 osteoblasts by SNA. These cell types were reproducibly categorized into independent distinct groups.

Lectin microarray analysis of human mesenchymal cells

Human MSCs harvested from a variety of tissues have the capability to differentiate into numerous tissue lineages despite the fact that they may have tissue-specific characteristics. To clarify relationship between the tissue-specific characters of mesenchymal cells and glycomics, we performed lectin microarray analysis (LecChip™: Fig. S1 in Supporting Information) of mesenchymal cells derived from various tissues (Fig. 2A). Signal intensities by lectin microarray were consistent with the mean fluorescent intensities analysis determined by flow cytometric analysis (Fig. 2B). Hierarchical clustering analysis showed that human embryonic carcinoma NCR-G3 cells were reproducibly categorized into an independent group (red color in Fig. 2C), which is distinct from a group of mesenchymal cells derived from a variety of tissues (green color in Fig. 2C). In mesenchymal cells, bone marrow-, placenta- and extra finger-derived mesenchymal cells were categorized into distinct groups labeled in yellow, orange and blue, respectively (Fig. 2C).

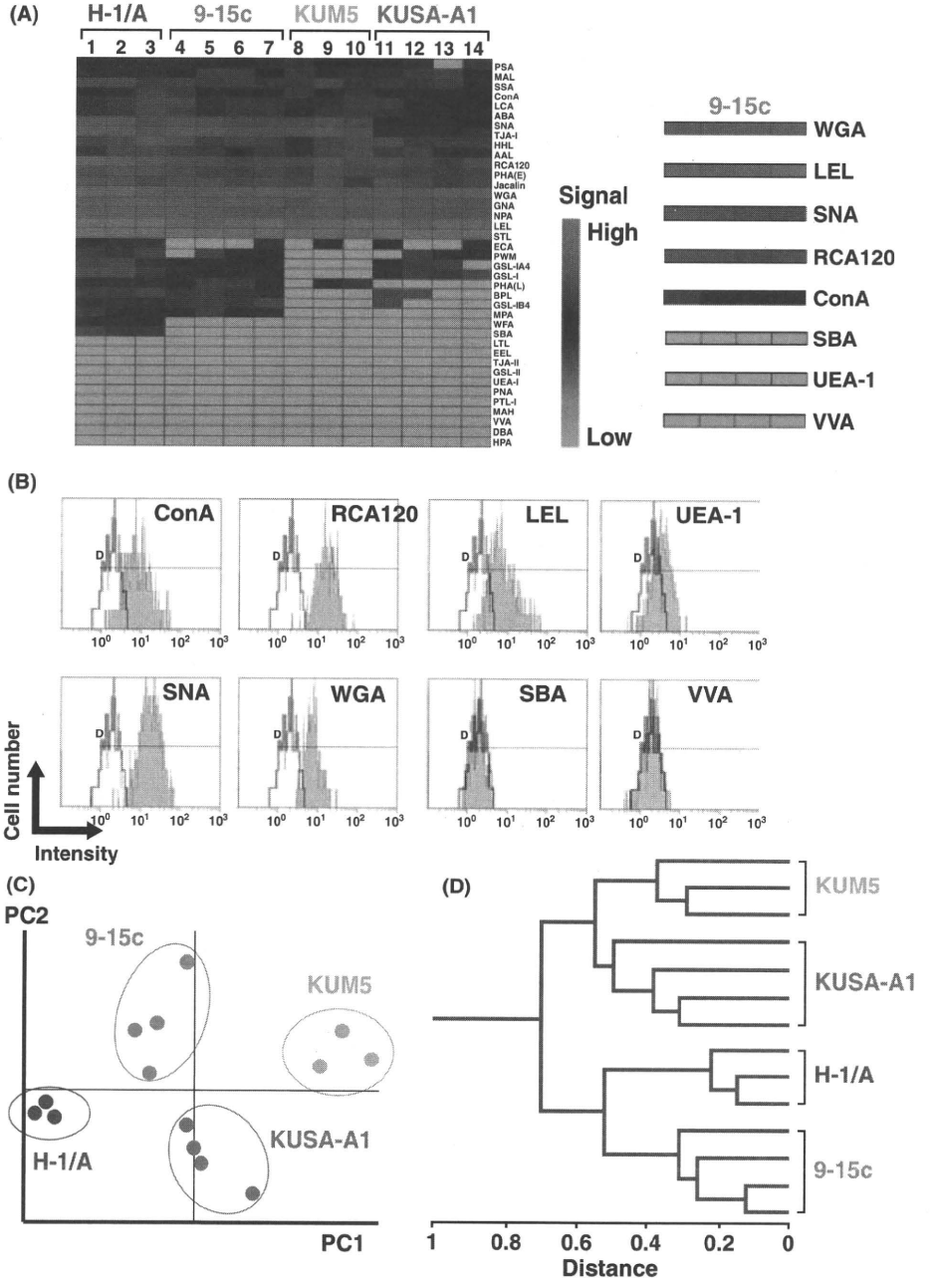
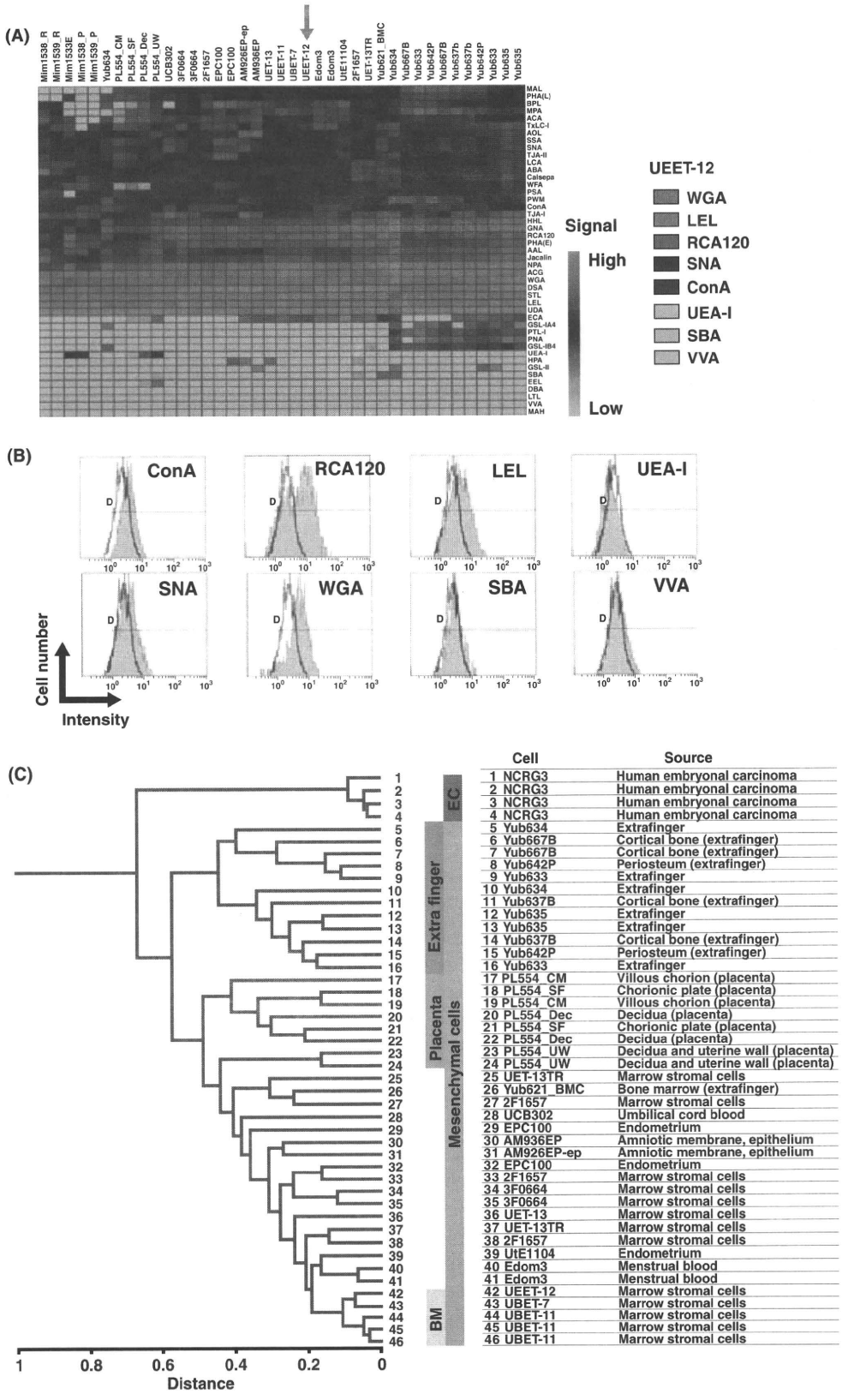


Figure 1 Lectin microarray analysis of mouse mesenchymal cells. (A) Heat map of 9-15c multipotent cells, KUSA-A1 osteoblasts, KUM5 chondroblasts and H-1/A preadipocytes. (B) Flow cytometric analysis of 9-15c multipotent cells using each lectin probe. Mean fluorescent intensities by flow cytometric analysis are consistent with signal intensities by lectin microarray. Nonshaded and shaded areas indicate reactivity of antibodies for isotype controls and that of antibodies for cell surface markers, respectively. (C) Principal component analysis of lectin microarray on mouse bone marrow–derived mesenchymal cells. Each cell is reproducibly subcategorized into groups of mesenchymal cell types. (D) Hierarchical clustering analysis of lectin microarray on mouse bone marrow–derived mesenchymal cells.

Human mesenchymal cells reacted to (i) *Pisum sativum* agglutinin (PSA), *Lens culinaris* agglutinin (LCA), *Aspergillus oryzae* lectin (AOL) and *Aleuria aurantia* lectin (AAL) that bind to Fucα1-6GlcNAc; (ii) SNA, *Sambucus sieboldiana* agglutinin (SSA) and *Trichosanthes japonica* agglutinin I (TJA-I) that bind to



Cell	Source
1 NCRG3	Human embryonal carcinoma
2 NCRG3	Human embryonal carcinoma
3 NCRG3	Human embryonal carcinoma
4 NCRG3	Human embryonal carcinoma
5 Yub634	Extrafinger
6 Yub667B	Cortical bone (extrafinger)
7 Yub667B	Cortical bone (extrafinger)
8 Yub642P	Periosteum (extrafinger)
9 Yub633	Extrafinger
10 Yub634	Extrafinger
11 Yub637B	Cortical bone (extrafinger)
12 Yub635	Extrafinger
13 Yub635	Extrafinger
14 Yub637B	Cortical bone (extrafinger)
15 Yub642P	Periosteum (extrafinger)
16 Yub633	Extrafinger
17 PL554_CM	Villous chorion (placenta)
18 PL554_SF	Chorionic plate (placenta)
19 PL554_CM	Villous chorion (placenta)
20 PL554_Dec	Decidua (placenta)
21 PL554_SF	Chorionic plate (placenta)
22 PL554_Dec	Decidua (placenta)
23 PL554_UW	Decidua and uterine wall (placenta)
24 PL554_UW	Decidua and uterine wall (placenta)
25 UET-13TR	Marrow stromal cells
26 Yub621_BMC	Bone marrow (extrafinger)
27 2F1657	Marrow stromal cells
28 UCB302	Umbilical cord blood
29 EPC100	Endometrium
30 AM936EP	Amniotic membrane, epithelium
31 AM926EP-ep	Amniotic membrane, epithelium
32 EPC100	Endometrium
33 2F1657	Marrow stromal cells
34 3F0664	Marrow stromal cells
35 3F0664	Marrow stromal cells
36 UET-13	Marrow stromal cells
37 UET-13TR	Marrow stromal cells
38 2F1657	Marrow stromal cells
39 UtE1104	Endometrium
40 Edom3	Menstrual blood
41 Edom3	Menstrual blood
42 UEET-12	Marrow stromal cells
43 UBET-7	Marrow stromal cells
44 UBET-11	Marrow stromal cells
45 UBET-11	Marrow stromal cells
46 UBET-11	Marrow stromal cells