IV. 研究成果の刊行物・別冊

Journal of Human Genetics (2011), 1–2

www.nature.com/jhg

COMMENTARY

A commentary on Implication of gene copy number variation in health and diseases

Tadashi Kaname

Journal of Human Genetics advance online publication, 8 December 2011; doi:10.1038/jhg.2011.140

Genetic and structural variations in the genome account for most of the differences between the genomes of individual humans.

Among these variations, single nucleotide polymorphisms (SNPs) are well characterized and widely used for genome-wide association studies (GWAS) of common human diseases and drug metabolism because of the availability of platforms for large-scale assay, such as microarray systems.

SNPs are divided into several types based on allele frequency in the general population, including rare SNPs (<0.5%) and common SNPs (>1–5%).[1] On the basis of the common disease-common variants hypothesis and because of the eraly types of microarrays covering common SNPs, GWAS using common SNPs have been undertaken to identify disease-related alleles.[1] However, discoveries from GWAS using common SNPs could explain only 2–15% of heritable variation in disease risk. It has been speculated that there are considerable missing heritability elements.[1]

Progress in microarray technology and the Human Genome Project resulted in the discovery of variations in gene copy number in the human genome; these variations were termed copy number variations (CNVs).[2] CNV is defined as the presence of multiple copies of genome segments >1 kb up to several Mb in size among individuals owing to deletions, insertions, inversions, duplications or complex recombinations.[2] More than 7000 CNVs were mapped in the human genome.[2] Similar to SNPs, human genome has common and rare CNVs.[3]

Dr T Kaname, Department of Medical Genetics, University of the Ryukyus Graduate School of Medicine, 207 Uehara, Nishihara, Okinawa 903-0215, Japan.
E-mail: tkaname@med.u-ryukyu.ac.jp

As mapped CNVs were associated with nearly 3000 genes[2] and diversity of CNVs was believed to affect human variation, CNVs were expected to account for the missing heritability. Many groups have addressed the presence of CNVs in the human genome and their association with common diseases including neuropsychiatric, infectious, autoimmune, cardiovascular and metabolic diseases.

Recent advances in defining relationships between CNVs and health, diseases and drug response, were summarized by Almal and Padh.[3] In their review, the authors clearly and concisely described the general aspects of CNVs and the relationships between human diseases, or drug metabolism and CNVs. To date, many studies have evidenced the involvement of CNVs in many human diseases.

Although mouse models of CNVs were not mentioned in the review, some reports described the establishment of transgenic models of CNVs to validate the effects in mice, which might be important. For example, mouse models of 16p11.2 CNVs exhibited dosage-dependent changes in gene expression, viability, brain architecture and behavior that mimicked autism in humans.[4] Therefore, for assessing the effects or associations of CNV, increase in direct experimentation on mice is anticipated.

Till date, a lot of population-based studies have reported that CNVs in some pharmacogenetic genes, such as CYP2D6 and CYP2C19, have a clear role in drug efficacy and toxicity.[5] In contrast, some recent GWAS of common CNVs reported rather pessimistic results in common diseases as follows.[6,7]

Common CNV-based GWAS in 16 000 cases of eight common diseases—bipolar disorder, breast cancer, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type-1 diabetes and type-2 diabetes—and 3000 controls revealed that common CNVs were unlikely to contribute to the genetic basis of these diseases.[7]

In addition, another study based on the assessment of the patterns of linkage disequilibrium between CNVs and SNPs suggested that for complex traits the heritability void left by GWAS could not be compensated by common CNVs.[6]

As both authors concluded, other platforms or resources of genetic variants could elucidate more disease susceptibilities. Maps of precise nucleotide CNVs and rare CNVs, as well as rare SNPs would be important resources for association studies of human diseases.[1]

However, the resolution of CNVs on chromosomes (mapping of CNVs) has not been sufficiently high, and has been biased because of limitations in microarray technology. More recently, CNV studies have become more precise to the level of nucleotide resolution and have allowed the discovery of rare CNVs by resequencing personal genomes using the next-generation sequencer.[8–10]

The Structural Variation Group of the 1000 Genomes Project constructed a map of CNVs on the basis of whole genome DNA sequencing data from 185 human genomes, which were resequenced using next-generation sequencer.[10] The map encompassed 22 035 deletions and 6000 additional structural variations, including insertions and tandem duplications. In addition, more than half of the structural variations were mapped to nucleotide resolution.

Accumulation of such precise and accurate data of CNVs, including personal or rare CNVs, will facilitate disease association studies, and enable the discovery of new disease susceptibilities and drug responses.

1 Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J. *et al.* Finding the missing heritability of complex diseases. *Nature.* **461,** 747–753 (2009).
2 Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T D. *et al.* Global variation in copy number in the human genome. *Nature.* **444,** 444–454 (2006).
3 Almal, S. H. & Padh, H. Pharmacogenetics: implication of gene copy number variation. *J. Hum. Genet.* (e-pub ahead of print 20 September 2011; doi:10.1038/jhg.2011.108).
4 Horev, G., Ellegood, J., Lerch, J. P., Son, Y- E. E., Muthuswamy, L., Vogel, H. *et al.* Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proc. Natl Acad. Sci. USA.* **108,** 17076–81 (2011).
5 Johansson, I. & Ingelman-Sundberg, M. CNVs of human genes and their implication in pharmacogenetics. *Cytogenet. Genome. Res.* **123,** 195–204 (2008).
6 Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature.* **464,** 704–712 (2010).
7 The Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16000 cases of eight common diseases and 3000 shared controls. *Nature.* **464,** 713–720 (2010).
8 Conrad, D. F., Bird, C., Blackburne, B., Lindsay, S., Mamanova, L., Lee, C. *et al.* Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nat. Genet.* **42,** 385–391 (2010).
9 Sudmant, P. H., Kitzman, J. O., Antoncacci, F., Alkan, C., Malig, M., Tsalenko, A. *et al.* Diversity of human copy number variations and multicopy genes. *Science.* **330,** 641–646 (2010).
10 Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C. *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature.* **470,** 59–65 (2011).

Journal of Human Genetics (2012) 57, 207–211

www.nature.com/jhg

ORIGINAL ARTICLE

Missense mutations in the DNA-binding/dimerization domain of *NFIX* cause Sotos-like features

Yuriko Yoneda[1], Hirotomo Saitsu[1], Mayumi Touyama[2], Yoshio Makita[3], Akie Miyamoto[4], Keisuke Hamada[5], Naohiro Kurotaki[6], Hiroaki Tomita[7], Kiyomi Nishiyama[1], Yoshinori Tsurusaki[1], Hiroshi Doi[1], Noriko Miyake[1], Kazuhiro Ogata[5], Kenji Naritomi[8] and Naomichi Matsumoto[1]

Sotos syndrome is characterized by prenatal and postnatal overgrowth, characteristic craniofacial features and mental retardation. Haploinsufficiency of *NSD1* causes Sotos syndrome. Recently, two microdeletions encompassing *Nuclear Factor I-X (NFIX)* and a nonsense mutation in *NFIX* have been found in three individuals with Sotos-like overgrowth features, suggesting possible involvements of *NFIX* abnormalities in Sotos-like features. Interestingly, seven frameshift and two splice site mutations in *NFIX* have also been found in nine individuals with Marshall–Smith syndrome. In this study, 48 individuals who were suspected as Sotos syndrome but showing no *NSD1* abnormalities were examined for *NFIX* mutations by high-resolution melt analysis. We identified two heterozygous missense mutations in the DNA-binding/dimerization domain of the NFIX protein. Both mutations occurred at evolutionally conserved amino acids. The c.179T>C (p.Leu60Pro) mutation occurred *de novo* and the c.362G>C (p.Arg121Pro) mutation was inherited from possibly affected mother. Both mutations were absent in 250 healthy Japanese controls. Our study revealed that missense mutations in *NFIX* were able to cause Sotos-like features. Mutations in DNA-binding/dimerization domain of NFIX protein also suggest that the transcriptional regulation is abnormally fluctuated because of *NFIX* abnormalities. In individuals with Sotos-like features unrelated to *NSD1* changes, genetic testing of *NFIX* should be considered.
Journal of Human Genetics (2012) **57,** 207–211; doi:10.1038/jhg.2012.7; published online 2 February 2012

Keywords: DNA-binding/dimerization domain; missense mutation; *NFIX*; Sotos syndrome

INTRODUCTION

Sotos syndrome (MIM #117550) is an overgrowth syndrome characterized by tall stature and/or macrocephaly, distinctive facial appearance and mental retardation.[1] A *de novo* t(5;8)(q35;q24.1) translocation in a patient with Sotos syndrome revealed disruption of *NSD1* at 5q35. Subsequent identification of nonsense, frameshift and submicroscopic deletion mutations of *NSD1* in patients with Sotos syndrome clearly showed that haploinsufficiency of *NSD1* causes Sotos syndrome.[2] *NSD1* encodes nuclear receptor-binding SET domain protein 1, which functions as a histone methyltransferase that activates and represses transcription through chromatin modification.[3] The diagnosis of Sotos syndrome is established by confirming *NSD1* abnormalities,[4] and abnormalities of *NSD1* causes up to 90% of Sotos syndrome cases. However, a part of patients with suspected Sotos syndrome are known to show no abnormalities in *NSD1*,[5] suggesting involvement of another gene.

Recently it was reported that two patients with Sotos-like overgrowth features possessed microdeletions encompassing *Nuclear Factor I-X (NFIX)* at 19p13.2. In addition, a nonsense mutation in *NFIX* was identified in one patient with Sotos-like features.[6] Interestingly, frameshift and donor-splice site mutations were also identified in Marshall–Smith syndrome (MIM 602535) that is osteochondysplasia syndrome characterized by accelerated skeletal maturation, relative failure to thrive, respiratory difficulties, mental retardation and unusual facial features.[7] Therefore, *NFIX* mutations could cause either Sotos-like features or Marshall–Smith syndrome. Whereas the transcripts possessing the nonsense mutation in a patient with Sotos-like features suffered from the nonsense-mediated mRNA decay, transcripts of mutated alleles (by a donor-spice site and two frameshift mutations) in patients with Marshall–Smith syndrome escaped from the nonsense-mediated mRNA decay surveillance and could be translated, suggesting that haploinsufficiency of *NFIX* leads to

[1]Department of Human Genetics, Yokohama City University Graduate School of Medicine, Yokohama, Japan; [2]Department of Pediatrics, Okinawa Child Development Center, Okinawa, Japan; [3]Education Center, Asahikawa Medical University, Asahikawa, Japan; [4]Department of Pediatrics, Hokkaido Asahikawa Habilitation Center for Disabled Children, Asahikawa, Japan; [5]Department of Biochemistry, Yokohama City University Graduate School of Medicine, Yokohama, Japan; [6]Department of Neuropsychiatry, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan; [7]Department of Biological Psychiatry, Tohoku University Graduate School of Medicine, Sendai, Japan and [8]Department of Medical Genetics, University of the Ryukyus Faculty of Medicine, Nishihara, Japan
Correspondence: Dr N Matsumoto, Department of Human Genetics, Yokohama City University Graduate School of Medicine, Fukuura 3-9, Kanazawa-ku, Yokohama 236-0004, Japan.
E-mail: naomat@yokohama-cu.ac.jp
Received 9 September 2011; revised 21 November 2011; accepted 5 January 2012; published online 2 February 2012

Sotos-like features and dominant-negative effects of the truncated NFIX proteins cause Marshall–Smith syndrome.[6]

In this study, we screened for *NFIX* mutations in 48 Japanese patients who were suspected as Sotos syndrome, but showed neither deletions nor mutations in *NSD1*. Detailed genetic and clinical data are presented.

MATERIALS AND METHODS

Subjects

A total of 48 patients suspected as Sotos syndrome were analyzed for *NFIX* mutations. *NSD1* investigation by sequencing and fluorescent *in situ* hybridization analysis was negative in these patients. In this study, the patients presenting with cardinal features of Sotos syndrome (specific craniofacial features, intellectual disability and overgrowth to same extent) but showing no *NSD1* abnormalities are referred as those with 'Sotos-like features'. Experimental protocols were approved by the Committee for Ethical issues at Yokohama City University School of Medicine. All individuals were investigated in agreement with the requirements of Japanese regulations.

Mutation analysis

Genomic DNA was isolated from peripheral blood leukocytes according to standard methods. DNA for mutation screening was amplified by illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Buckinghamshire, UK). Sequencing of exon 1 and high-resolution melting curve (HRM) analysis of exon 2–9 covering the *NFIX* coding region (GenBank accession number NM_002501.2) were performed. For exon 1, the 12 µl PCR mixture contained 30 ng DNA, 0.3 µM each primer, 0.4 mM each dNTP, 1× PCR buffer for KOD FX and 0.3 U KOD FX polymerase (Toyobo, Osaka, Japan). For exons 2–9, real-time PCR and HRM analysis were serially performed in 12 µl mixture on Rotor-Gene Q (QIAGEN, Hilden, Germany). For exon 7, the PCR mixture contained 30 ng DNA, 0.3 µM each primer, 0.4 mM each dNTP, 0.36 µl SYTO9 (Invitrogen, Carlsbad, CA, USA), 0.4 mM each dNTP, 1× PCR buffer for KOD FX and 0.3 U KOD FX polymerase (Toyobo). For the remaining exons, the PCR mixture contained 30 ng DNA, 0.25 µM each primer, 0.36 µl SYTO9 (Invitrogen), 0.2 mM each dNTP, 1× ExTaq buffer and 0.375 U ExTaq HS (Takara, Otsu, Japan). Primers and conditions of PCR are shown in Supplementary Table 1. The PCR products showing an aberrant melting curve were sequenced. All the novel mutations in DNA amplified by GenomiPhi were verified by sequencing of PCR products using genomic DNA as a template. Mutations were checked in 250 Japanese normal controls (500 alleles) by HRM analysis.

Parentage testing

For the family showing *de novo* mutations, parentage was confirmed by microsatellite analysis as previously described.[8] Biological parentage was judged if more than four informative markers were compatible and other uninformative markers showed no discrepancies.

Prediction of functional effect

The effect of the mutations for protein features was predicted by following web-based prediction tools: SIFT (http://sift.jcvi.org/), PolyPhen (http://genetics.bwh.harvard.edu/pph/), PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/), Mutation Taster (http://www.mutationtaster.org/) and Align GVGD (http://agvgd.iarc.fr/agvgd_input.php).

RESULTS

NFIX mutations

Two heterozygous missense mutations were identified. The c.179T>C (p.Leu60Pro) mutation in patient 1 were not found in her parents, indicating that the mutation occurred *de novo* (Figure 1a). Biological parentage was confirmed by several microsatellite markers (data not shown). The c.362G>C (p.Arg121Pro) mutation in patient 2 was found in his mother (Figure 1a). These two mutations occurred at evolutionary conserved amino acids (Figure 1b) and were absent in 250 Japanese normal controls. Interestingly, the missense changes were located in DNA-binding/dimerization domain of the NFIX protein (Figure 1c). Evaluation with web-based prediction tools strongly suggested that these substitutions are pathogenic (Supplementary Table 2).

Clinical information of the patients

Patient 1 is a product of unrelated healthy parents. The body weight at birth was 2816 g (−0.6 s.d.), height 48.8 cm (0 s.d.) and OFC 33.5 cm (+0.3 s.d.). Neonatal hypotonia was recognized. At 17 months of age, her weight was 9.24 kg (−0.5 s.d.), height 84.9 cm (+2 s.d.) and OFC 48 cm (+1.2 s.d.). The facial appearance showed long/narrow and triangular face, high forehead, midface hypoplasia, prominent ears, epicanthal folds, strabismus, down-slanting palpebral fissures, short nose with anteverted nares, prominent long philtrum, everted lower lip and narrow palate (Figure 1d). Large hands/feet, prominent fingertips, pectus excavatum were also noted. Her primary dentition started at 7 months of age and was completed by 17 months of age. Bone age was estimated as 3 years at 17 months of age and as 5 years at 3 years of age. Bullet-shaped phalanges, which are typical features of Marshall–Smith syndrome, were not observed. She was initially diagnosed as Sotos syndrome. She showed mental retardation and severe developmental delay with developmental quotients of 19. Scoliosis was noted at 18 months of age and surgically treated for several times. Complex partial seizures were noted at 4 years of age and were controlled with phenytoin and zonisamide. At present (17 years of age), prognathia was observed (Figure 1e). Her weight was 40 kg (−2 s.d.) and height 156.5 cm (−0.2 s.d.).

Patient 2 is a male at age of 20 years. The birth weight was 2938 g (−0.4 s.d.), height 51 cm (+0.8 s.d.) and OFC 35.5 cm (+1.4 s.d.). Respiratory insufficiency was noted, but no visceral malformations were pointed out. Bilateral tubing therapy was performed for recurrent bilateral exudative otitis media at 4 years of age. At 14 years of age, his weight was 58.1 kg (+0.6 s.d.) and height 185.7 cm (+3.5 s.d.). Mental retardation was evident as the IQ score (Tanaka–Binet intelligence test) was 59. Craniofacial features included high forehead, down-slanting palpebral fissures and prognathia. He was suspected as Sotos syndrome. His mother showed tall stature, suggesting that c.362G>C led to overgrowth in the mother. Unfortunately, further details of clinical features in the mother are unavailable. Clinical information of two patients is summarized in Table 1.

DISCUSSION

NFIX is a member of the nuclear factor I (NFI) family proteins, which are implicated as site-specific DNA-binding proteins known to function in viral DNA replication and gene expression regulation.[9] NFI proteins form homo- or heterodimers and bind to the palindromic DNA consensus sequence through its N-terminal DNA-binding/dimerization domain.[10] Point mutations in DNA-binding/dimerization domain of NFI protein have been shown to cause loss of dimerization, DNA-binding and replication activities,[11] highlighting the importance of structural integrity of DNA-binding/dimerization domain. It has been reported that the DNA binding domain of SMADs and NFI transcription factors shared considerable structural similarity, and the secondary structure of the DNA-binding domain of NFI was estimated based on that of SMADs.[12] In this study, we identified two heterozygous missense mutations, the c.179T>C (p.Leu60Pro) and the c.362G>C (p.Arg121Pro), in the DNA-binding/dimerization domain. Of note, two mutations are estimated to be localized within α-helical region of DNA-binding domain and at evolutionally conserved amino acids between SMADs and NFI.[12] In addition, two mutations cause substitutions to a proline residue,

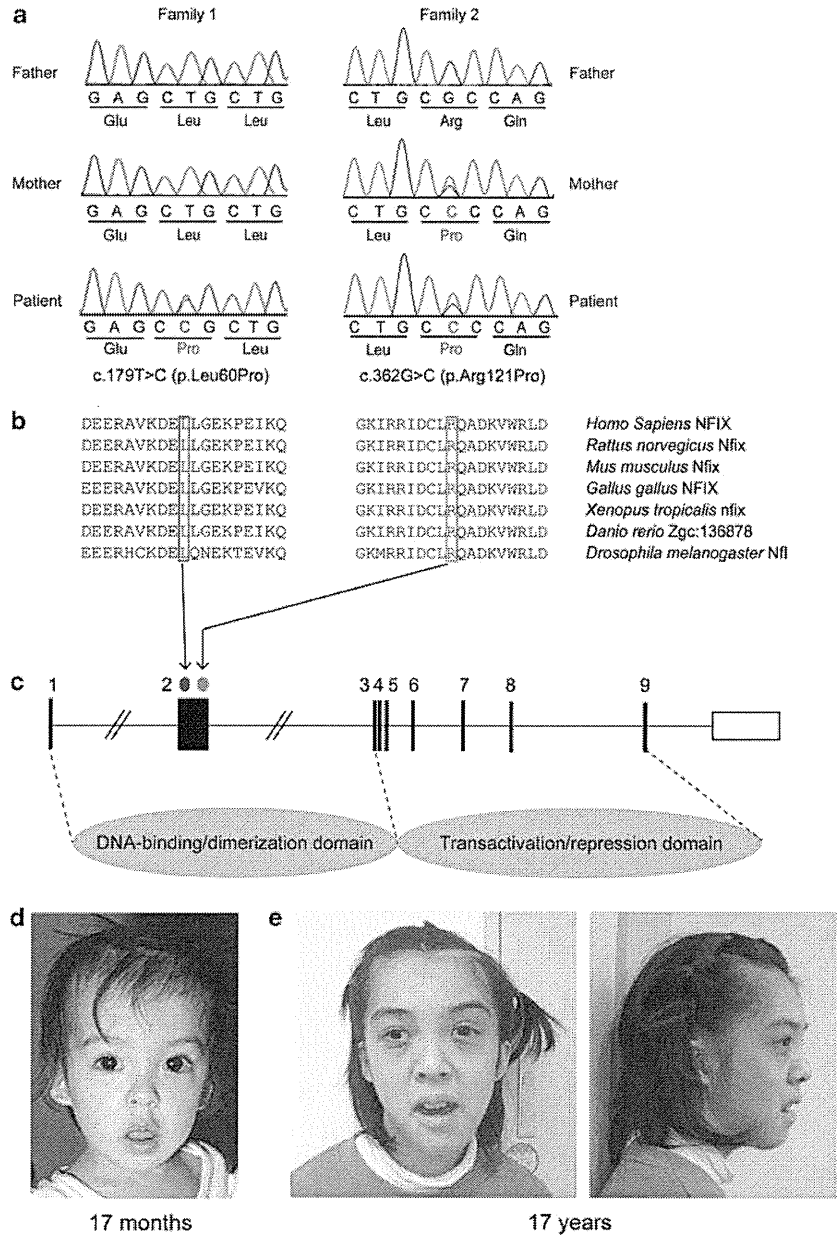

Figure 1 Missense mutations in *NFIX* in individuals with Sotos-like features. **(a)** Electropherogram of family 1 (left) and family 2 (right). The c.179T>C (p.Leu60Pro) mutation occured *de novo*. The c.362G>C (p.Arg121Pro) mutation was inherited from his mother. **(b)** An amino-acid sequence alignments of NFIX protein including amino-acid positions 60 and 121. Protein sequences were obtained through the NCBI protein database and multiple sequence alignment was performed by CLUSTALW web site (http://clustalw.ddbj.nig.ac.jp/). **(c)** Schematic representation of *NFIX* consisting of nine exons. UTR and coding exons are indicated by open and filled rectangles, respectively. The location of mutations is indicated by red (c.179T>C) and blue (c.362G>C) dots. At the bottom, C-terminal DNA-binding/dimerization domain and N-terminal transactivation/repression domain are depicted. Both the c.179T>C and c.362G>C mutations are located in exon 2 encoding a part of DNA-binding/dimerization domain. **(d)** Facial appearance of patient 1 at 17 months of age, showing long/narrow and triangular face, down slanting, short nose with anteverted nares and everted lower lip. **(e)** At 17 years of age, prognathia was noted in patient 1.

which is a unique amino acid characterized by imino radical. Proline has a pyrrolidine ring that restricts the available conformational space; therefore, it has effects on chain conformation and the process of protein folding.[13] Thus, it is very likely that two mutations could affect DNA-binding activity of NFIX protein through conformational changes of the DNA-binding domain.

Because *NFIX* mutations could cause both Marshall–Smith syndrome and Sotos-like features,[6] it is great concern to which of them two patients with missense mutations could be classified. Main clinical features of Sotos syndrome are childhood overgrowth including tall stature and/or macrocephaly, characteristic face and mental retardation. Other minor features are scoliosis, hypotonia in infancy, seizures,

Table 1 Clinical features of two patients with missense mutations in *NFIX*

				Reported by Malan et al.[6]		
		Patient 1	*Patient 2*	*Patient A*	*Patient B*	*Patient C*
Genetics	*NFIX deletion/mutation*	*c.179T>C*	*c.362G>C*	*del 19p13.3*	*del 19p13.3*	*c.568C>T*
Epidemiology	Age at last evaluation (years)	17	14	14	10	27
	Sex	F	M	M	M	F
	Mat/pat age	48/52	?/?	31/33	25/30	31/31
Prenatal growth	Birth weight (g)	2816 (−0.6 s.d.)	2938 (−0.4 s.d.)	4500 (>95)	3110 (10–50)	3600 (50–90)
	Birth height (cm)	48.8 (0 s.d.)	51 (+0.8 s.d.)	53 (95)	49 (50)	52 (95)
	OFC (cm)	33.5 (+0.3 s.d.)	35.5 (+1.4 s.d.)	38 (>95)	33.5 (10)	37.5 (>95)
Postnatal growth	Weight (kg)	9.24 (−0.5 s.d.)[a]	58.1 (+0.6 s.d.)[b]	>P98	>P98	>P98
	Height (cm)	84.9 (+2 s.d.)[a]	185.7 (+3.5 s.d.)[b]	>P98	>P98	>P98
Development						
SS	Autistic traits	−	−	+	+	+
	Behavioral anomalies	NA	−	+	+	+
	Motor retardation	+	+	+	−	−
	Hypotonia	+	+	+	+	−
Overlapped	Mental retardation	+	+	+	+	+
	Degree of delay	DQ19	IQ42	NA	NA	NA
	Speech delay	+	+	+	+	+
	First words (months)	24	18	NA	NA	NA
Craniofacial features						
SS	Long/narrow face	+	−	+	+	+
	Down-slanting palpebral fissures	+	+	+	−	+
	Small mouth	NA	−	+	−	+
	Prognathia	+	+	+	−	−
Overlapped	High forehead	+	+	+	+	+
MSS	Everted lower lip	+	−	+	−	+
	Underdeveloped midface	+	−	NA	NA	NA
	Proptosis	NA	−	NA	NA	NA
	Short nose	+	−	NA	NA	NA
	Prominent premaxilla	NA	−	NA	NA	NA
	Gum hypertrophy	+[c]	−	NA	NA	NA
	Retrognathia	−	−	NA	NA	NA
Eyes						
SS	Hypermotropia	−	−	+	+	−
	Strabismus	+	−	+	−	+
	Nystagmus	−	−	−	−	+
	Astigmatism	NA	NA	−	+	−
MSS	Myopia	NA	−	NA	NA	NA
	Blue sclerae	NA	−	NA	NA	NA
Musculo-skeletal abnormalities						
SS	Abdominal wall hypotonia	−	−	+	−	+
	Pectus excavatum	+	−	+	+	−
	Coxa valga	−	−	+	+	−
Overlapped	Scoliosis	+	−	+	−	+
	Advanced bone age	+	NA	+	+	+
MSS	Abnormal bone maturation	NA	NA	NA	NA	NA
	Bone fractures	−	−	NA	NA	NA
	Kyphosis	−	−	NA	NA	NA
	Umbilical hernia	−	−	NA	NA	NA

Abbreviations: F, female; M, male; Mat/pat, maternal/paternal; MSS, Marshall–Smith syndrome; NA, not ascertained; OFC, Occipitofrontal circumference; SS, Sotot's syndrome.
Growth of patients 1 and 2 is indicated with s.d. and that of patients in the report of Malan *et al.*[6] is indicated with percentile.
[a]At 17 months.
[b]At 14 years.
[c]Suggested the possibility of the adverse drug reaction.

cardiac defect and genitourinary anomalies.[5] On the other hand, main clinical features of Marshall–Smith syndrome are moderate to severe developmental delay with absent or limited speech, unusual behavior, disharmonic bone maturation, respiratory compromise secondary to upper airway obstruction, short stature and kyphoscoliosis.[14] One of remarkable differences between Sotos syndrome and Marshall–Smith syndrome is facial appearances. Although both syndromes has high forehead, Sotos syndrome has a long/narrow face, triangular shaped face with a prominent chin, down-slanting of the palpebral fissures,[1,4–5] whereas Marshall–Smith syndrome has proptosis, underdeveloped midface and prominent premaxilla.[7,14] In patient 1, although some characteristic features of Marshall–Smith syndrome such as everted lower lip, short nose and midface hypoplasia were observed, overall facial appearance, overgrowth features at 17 month of age, scoliosis, hypotonia and seizures were consistent with Sotos syndrome. Similarly, in patient 2, the facial appearance, tall stature and macrocephaly were consistent with Sotos syndrome. In both patients, their body weights were relatively low in comparison with their heights. This is consistent with the fact that, throughout childhood and early adolescence, the height was usually more significantly increased than weight in Sotos patients.[15] In addition, our patients did not show respiratory difficulties, one of specific features in Marshall–Smith syndrome, which cause early death in the neonatal period or early infancy.[7] Thus missense mutations in the DNA-binding/dimerization domain, which may lead to loss of transcriptional regulation by NFIX protein, could cause Sotos-like syndrome in two patients.

Many clinical features including tall statue, mental retardation, speech delay and high forehead are shared between our patients and three patients reported by Malan *et al.*[6] with *NFIX* abnormalities. The recognizable difference is autistic traits. Autistic traits are not observed in our patients but all of Malan *et al.*'s[6] patients. Thus there is a possibility that autistic traits are caused by haploinsufficiency of *NFIX* in Malan *et al.*'s[6] patients, but not by misssense mutations in the DNA-binding/dimerization domain. However, identification of a greater number of cases with *NFIX* mutations is required to confirm this hypothesis.

In conclusion, our report provides further evidences that *NFIX* is a causative gene for Sotos-like features. Abnormalities of *NSD1* are found in majority of Sotos syndrome cases and aberration of other genes including *NFIX* may be found in the minority of Sotos syndrome/Sotos-like features. Genetic testing of *NFIX* should be considered in such patients if no *NSD1* abnormalities were identified.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank the patients and their family members for their participation in this study. This work was supported by Research Grants from the Ministry of Health, Labour and Welfare (HS, N Miyake and N Matsumoto) and the Japan Science and Technology Agency (N Matsumoto), a Grant-in-Aid for Young Scientist from the Japan Society for the Promotion of Science (HS, HD and N Miyake) and a Grant-in-Aid for Scientific Research from Japan Society for the Promotion of Science (N Matsumoto).

1 Leventopoulos, G., Kitsiou-Tzeli, S., Kritikos, K., Psoni, S., Mavrou, A., Kanavakis, E. *et al.* A clinical study of Sotos syndrome patients with review of the literature. *Pediatr. Neurol.* **40,** 357–364 (2009).
2 Kurotaki, N., Imaizumi, K., Harada, N., Masuno, M., Kondoh, T., Nagai, T. *et al.* Haploinsufficiency of *NSD1* causes Sotos syndrome. *Nat. Genet.* **30,** 365–366 (2002).
3 Rayasam, G. V., Wendling, O., Angrand, P. O., Mark, M., Niederreither, K., Song, L. *et al. NSD1* is essential for early post-implantation development and has a catalytically active SET domain. *EMBO J.* **22,** 3153–3163 (2003).
4 Visser, R. & Matsumoto, N. in *Inborn Errors of Development* (eds Epstein, C. J., Erickson, R. P., Wynshaw-Boris, A.) 1032–1037 (Oxford University Press, New York, 2008).
5 Tatton-Brown, K. & Rahman, N. Sotos syndrome. *Eur. J. Hum. Genet.* **15,** 264–271 (2007).
6 Malan, V., Rajan, D., Thomas, S., Shaw, A. C., Louis Dit Picard, H., Layet, V. *et al.* Distinct effects of allelic *NFIX* mutations on nonsense-mediated mRNA decay engender either a Sotos-like or a Marshall-Smith syndrome. *Am. J. Hum. Genet.* **87,** 189–198 (2010).
7 Adam, M. P., Hennekam, R. C., Keppen, L. D., Bull, M. J., Clericuzio, C. L., Burke, L. W. *et al.* Marshall-Smith syndrome: natural history and evidence of an osteochondrodysplasia with connective tissue abnormalities. *Am. J. Med. Genet. A.* **137,** 117–124 (2005).
8 Saitsu, H., Kato, M., Mizuguchi, T., Hamada, K., Osaka, H., Tohyama, J. *et al. De novo* mutations in the gene encoding STXBP1 (MUNC18-1) cause early infantile epileptic encephalopathy. *Nat. Genet.* **40,** 782–788 (2008).
9 Gronostajski, R. M. Roles of the NFI/CTF gene family in transcription and development. *Gene* **249,** 31–45 (2000).
10 Kruse, U. & Sippel, A. E. Transcription factor nuclear factor I proteins form stable homo- and heterodimers. *FEBS Lett.* **348,** 46–50 (1994).
11 Armentero, M. T., Horwitz, M. & Mermod, N. Targeting of DNA polymerase to the adenovirus origin of DNA replication by interaction with nuclear factor I. *Proc. Natl. Acad. Sci. USA* **91,** 11537–11541 (1994).
12 Stefancsik, R. & Sarkar, S. Relationship between the DNA binding domains of SMAD and NFI/CTF transcription factors defines a new superfamily of genes. *DNA Seq.* **14,** 233–239 (2003).
13 MacArthur, M. W. & Thornton, J. M. Influence of proline residues on protein conformation. *J. Mol. Biol.* **218,** 397–412 (1991).
14 Shaw, A. C., van Balkom, I. D., Bauer, M., Cole, T. R., Delrue, M. A., Van Haeringen, A. *et al.* Phenotype and natural history in Marshall-Smith syndrome. *Am. J. Med. Genet. A* **152A,** 2714–2726 (2010).
15 Cole, T. R. & Hughes, H. E. Sotos syndrome: a study of the diagnostic criteria and natural history. *J. Med. Genet.* **31,** 20–32 (1994).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

RESEARCH ARTICLE

AMERICAN JOURNAL OF medical genetics PART A

Spectrum of *MLL2* (*ALR*) Mutations in 110 Cases of Kabuki Syndrome

Mark C. Hannibal,[1,2] Kati J. Buckingham,[1] Sarah B. Ng,[3] Jeffrey E. Ming,[4] Anita E. Beck,[1,2] Margaret J. McMillin,[2] Heidi I. Gildersleeve,[1] Abigail W. Bigham,[1] Holly K. Tabor,[1,2] Heather C. Mefford,[1,2] Joseph Cook,[1] Koh-ichiro Yoshiura,[5] Tadashi Matsumoto,[5] Naomichi Matsumoto,[6] Noriko Miyake,[6] Hidefumi Tonoki,[7] Kenji Naritomi,[8] Tadashi Kaname,[8] Toshiro Nagai,[9] Hirofumi Ohashi,[10] Kenji Kurosawa,[11] Jia-Woei Hou,[12] Tohru Ohta,[13] Deshung Liang,[14] Akira Sudo,[15] Colleen A. Morris,[16] Siddharth Banka,[17] Graeme C. Black,[17] Jill Clayton-Smith,[17] Deborah A. Nickerson,[3] Elaine H. Zackai,[4] Tamim H. Shaikh,[18] Dian Donnai,[17] Norio Niikawa,[13] Jay Shendure,[3] and Michael J. Bamshad[1,2,3]*

[1]Department of Pediatrics, University of Washington, Seattle, Washington

[2]Seattle Children's Hospital, Seattle, Washington

[3]Department of Genome Sciences, University of Washington, Seattle, Washington

[4]Department of Pediatrics, The Children's Hospital of Philadelphia, The University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania

[5]Department of Human Genetics, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan

[6]Department of Human Genetics, Yokohama City University Graduate School of Medicine, Yokohama, Japan

[7]Department of Pediatrics, Tenshi Hospital, Sapporo, Japan

[8]Department of Medical Genetics, University of the Ryukyus, Okinawa, Japan

[9]Department of Pediatrics, Dokkyo Medical University, Koshigaya Hospital, Saitama, Japan

[10]Division of Medical Genetics, Saitama Children's Medical Center, Saitama, Japan

[11]Division of Clinical Genetics, Kanagawa Children's Medical Center, Yokohama, Japan

[12]Department of Pediatrics, Chang Gung Children's Hospital, Taoyuan, Taiwan, Republic of China

[13]Research Institute of Personalized Health Sciences, Health Sciences University of Hokkaido, Hokkaido, Japan

[14]National Laboratory of Medical Genetics, Xiangya Hospital, Central South University, Republic of China

[15]Department of Pediatrics, Sapporo City General Hospital, Sapporo, Japan

[16]University of Nevada School of Medicine, Las Vegas, Nevada

[17]Department of Genetic Medicine, Manchester Academic Health Sciences Centre, University of Manchester, England

[18]Department of Pediatrics, University of Colorado, Denver, Colorado

Received 25 February 2011; Accepted 30 March 2011

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

Grant sponsor: National Institutes of Health/National Heart Lung and Blood Institute; Grant number: 5R01HL094976; Grant sponsor: National Institutes of Health/National Human Genome Research Institute; Grant numbers: 5R21HG004749, 1RC2HG005608, 5RO1HG004316, T32HG00035; Grant sponsor: National Institute of Heath/National Institute of Environmental Health Sciences; Grant number: HHSN273200800010C; Grant sponsor: National Institute of Neurological Disorders and Stroke; Grant number: RO1NS35102; Grant sponsor: NIHR Manchester Biomedical Research Centre; Grant sponsor: Ministry of Health, Labour and Welfare of Japan; Grant sponsor: Japan Science and Technology Agency; Grant sponsor: Society for the Promotion of Science; Grant sponsor: Life Sciences Discovery Fund; Grant numbers: 2065508, 0905001; Grant sponsor: Washington Research Foundation; Grant sponsor: National Institutes of Health/National Institute of Child Health and Human Development; Grant numbers: 1R01HD048895, 5K23HD057331.

Mark C. Hannibal, Kati J. Buckingham, and Sarah B. Ng contributed equally to this work.

*Correspondence to:
Michael J. Bamshad, M.D., Department of Pediatrics, University of Washington School of Medicine, Box 356320, 1959 NE Pacific Street, Seattle, WA 98195. E-mail: mbamshad@u.washington.edu

Published online 10 June 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI 10.1002/ajmg.a.34074

Kabuki syndrome is a rare, multiple malformation disorder characterized by a distinctive facial appearance, cardiac anomalies, skeletal abnormalities, and mild to moderate intellectual disability. Simplex cases make up the vast majority of the reported cases with Kabuki syndrome, but parent-to-child transmission in more than a half-dozen instances indicates that it is an autosomal dominant disorder. We recently reported that Kabuki syndrome is caused by mutations in *MLL2*, a gene that encodes a Trithorax-group histone methyltransferase, a protein important in the epigenetic control of active chromatin states. Here, we report on the screening of 110 families with Kabuki syndrome. *MLL2* mutations were found in 81/110 (74%) of families. In simplex cases for which DNA was available from both parents, 25 mutations were confirmed to be de novo, while a transmitted *MLL2* mutation was found in two of three familial cases. The majority of variants found to cause Kabuki syndrome were novel nonsense or frameshift mutations that are predicted to result in haploinsufficiency. The clinical characteristics of *MLL2* mutation-positive cases did not differ significantly from *MLL2* mutation-negative cases with the exception that renal anomalies were more common in *MLL2* mutation-positive cases. These results are important for understanding the phenotypic consequences of *MLL2* mutations for individuals and their families as well as for providing a basis for the identification of additional genes for Kabuki syndrome. © 2011 Wiley-Liss, Inc.

Key words: Kabuki syndrome; *MLL2*; *ALR*; Trithorax group histone methyltransferase

How to Cite this Article:
Hannibal MC, Buckingham KJ, Ng SB, Ming JE, Beck AE, McMillin MJ, Gildersleeve HI, Bigham AW, Tabor HK, Mefford HC, Cook J, Yoshiura K-i, Matsumoto T, Matsumoto N, Miyake N, Tonoki H, Naritomi K, Kaname T, Nagai T, Ohashi H, Kurosawa K, Hou J-W, Ohta T, Liang D, Sudo A, Morris CA, Banka S, Black GC, Clayton-Smith J, Nickerson DA, Zackai EH, Shaikh TH, Donnai D, Niikawa N, Shendure J, Bamshad MJ. 2011. Spectrum of *MLL2* (*ALR*) mutations in 110 cases of Kabuki syndrome.
Am J Med Genet Part A 155:1511–1516.

INTRODUCTION

Kabuki syndrome (OMIM#147920) is a rare, multiple malformation disorder characterized by a distinctive facial appearance, cardiac anomalies, skeletal abnormalities, and mild to moderate intellectual disability. It was originally described by Niikawa et al. [1981] and Kuroki et al. [1981] in 1981, and to date, about 400 cases have been reported worldwide [Niikawa et al., 1988; White et al., 2004; Adam and Hudgins, 2005]. The spectrum of abnormalities found in individuals with Kabuki syndrome is diverse, yet virtually all affected persons are reported to have similar facial features consisting of elongated palpebral fissures, eversion of the lateral third of the lower eyelids, and broad, arched eyebrows with lateral sparseness. Additionally, affected individuals commonly have severe feeding problems, failure to thrive in infancy, and height around or below the 3rd centile for age in about half of cases.

We recently reported that a majority of cases of Kabuki syndrome are caused by mutations in *mixed lineage leukemia 2* (*MLL2*; OMIM#602113), also known as either *MLL4* or *ALR* [Ng et al., 2010]. *MLL2* encodes a SET-domain-containing histone methyltransferase important in the epigenetic control of active chromatin states [FitzGerald and Diaz, 1999]. Exome sequencing revealed that 9 of 10 individuals had novel variants in *MLL2* that were predicted to be deleterious. A single individual had no mutation in the protein-coding exons of *MLL2*, though in retrospect, his phenotypic features are somewhat atypical of Kabuki syndrome. In a larger validation cohort screened by Sanger sequencing, we found *MLL2* mutations in approximately two-thirds of 43 Kabuki cases, suggesting that Kabuki syndrome is genetically heterogeneous.

Herein we report on the results of screening *MLL2* for mutations in 110 families with one or more individuals affected with Kabuki syndrome in order to: (1) characterize the spectrum of *MLL2* mutations that cause Kabuki syndrome; (2) determine whether *MLL2* genotype is predictive of phenotype; (3) assess whether the clinical characteristics of *MLL2* mutation-positive cases differ from *MLL2* mutation-negative cases; and (4) delineate the subset of Kabuki cases that are *MLL2* mutation-negative for further gene discovery studies.

MATERIALS AND METHODS

Subjects

Referral for inclusion into the study required a diagnosis of Kabuki syndrome made by a clinical geneticist. From these cases, phenotypic data were collected by review of medical records, phone interviews, and photographs. These data were collected from five different clinical genetics centers in three different countries and over a protracted period of time and forwarded for review to two of the authors (M.B. and M.H.). Data on certain phenotypic characteristics including stature, feeding difficulties, and failure to thrive was not uniformly collected or standardized. Therefore, we decided to be conservative in our analysis and use only phenotypic traits that could be represented by discrete variables (i.e., presence or absence) and for which data were available from at least 70% of cases. In addition, these clinical summaries were de-identified and therefore facial photographs were unavailable from most cases studied. Written consent was obtained for all participants who provided identifiable samples. The Institutional Review Boards of Seattle Children's Hospital and the University of Washington approved all studies. A summary of the clinical characteristics of 53 of these individuals diagnosed with Kabuki syndrome has been reported previously [Ng et al., 2010].

Mutation Analysis

Genomic DNA was extracted using standard protocols. Each of the 54 exons of *MLL2* was amplified using Taq DNA polymerase (Invitrogen, Carlsbad, CA) following manufacturer's recommendations and using primers previously reported [Ng et al., 2010]. PCR products were purified by treatment with exonuclease I (New England Biolabs, Inc., Beverly, MA) and shrimp alkaline phosphatase (USB Corp., Cleveland, OH), and products were sequenced using the dideoxy terminator method on an automated sequencer (ABI 3130xl). The electropherograms of both forward and reverse strands were manually reviewed using CodonCode Aligner (Dedham, MA). Primer sequences and conditions are listed in Supplementary Table I.

For *MLL2* mutation-negative samples, DNA was hybridized to commercially available whole-genome tiling arrays consisting of one million oligonucleotide probes with an average spacing of 2.6 kb throughout the genome (SurePrint G3 Human CGH Microarray 1 × 1 M, Agilent Technologies, Santa Clara, CA). Twenty-one probes on this array covered *MLL2* specifically. Data were analyzed using Genomics Workbench software according to manufacturer's instructions.

RESULTS

All 54 protein-coding exons and intron–exon boundaries of *MLL2* were screened by Sanger sequencing in a cohort of 110 kindreds with Kabuki syndrome. This cohort included 107 simplex cases (including a pair of monozygotic twins) and 3 familial (i.e., parent-offspring) cases putatively diagnosed with Kabuki syndrome. Seventy novel *MLL2* variants that were inferred to be disease-causing were identified in 81/110 (74%) kindreds (Fig. 1 and Supplementary Table II online). These 81 mutations included 37 nonsense mutations (32 different sites and five sites with recurrent mutations), 3 in-frame deletions or duplications (2 different sites and 1 site with a recurrent mutation), 22 frameshifts (22 different sites), 16 missense mutations (11 different sites and 4 sites with recurrent mutations), and 3 splice consensus site (or intron–exon boundary) mutations. None of these variants were found in dbSNP (build 132), the 1000 Genomes Project pilot data, or 190 chromosomes from individuals matched for geographical ancestry. In total, pathogenic variants were found at 70 sites. Additionally, there were 10 sites at which recurrent mutations were observed.

For 25 simplex cases in which we identified *MLL2* mutations, DNA was available from both unaffected parents, and in each case the mutation was confirmed to have arisen de novo (Supplementary Table II online). These included 14 nonsense, 5 frameshift, 3 missense, 2 splice site mutations, and 1 deletion. De novo events were confirmed at 6 of the 10 sites where recurrent mutations were noted. In addition to the 81 kindreds in which we identified causal *MLL2* mutations, we found two *MLL2* variants in each of three simplex cases. In each case, neither *MLL2* mutation could unambiguously

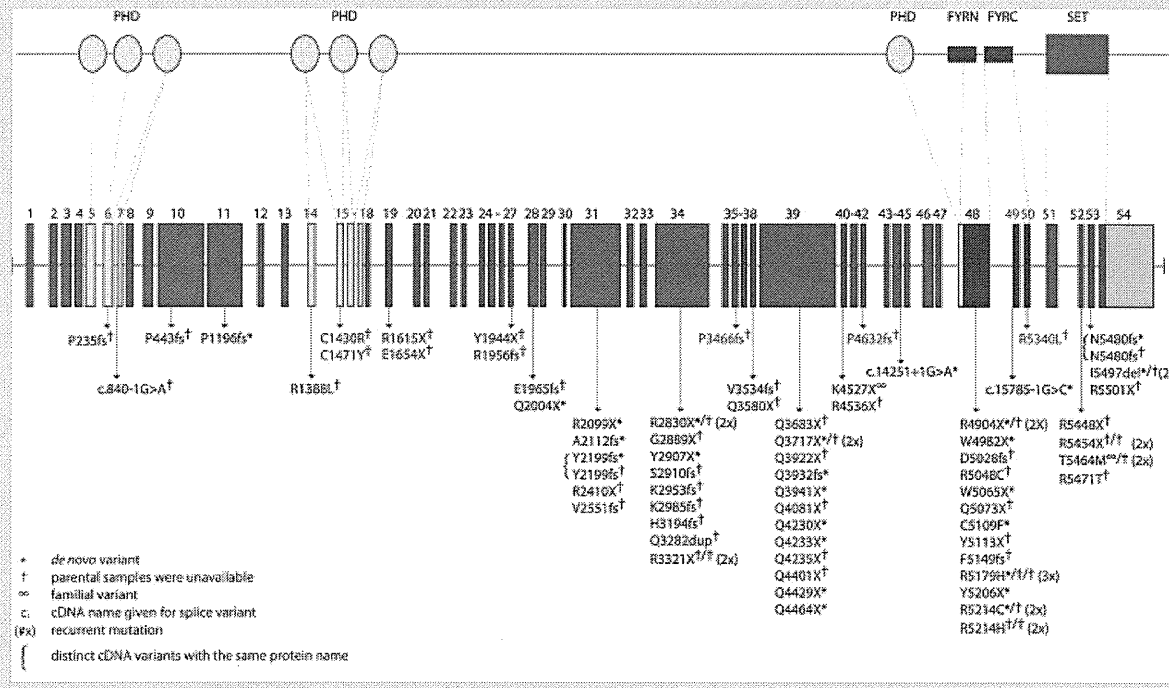

FIG. 1. Genomic structure and allelic spectrum of *MLL2* mutations that cause Kabuki syndrome. *MLL2* is composed of 54 exons that include untranslated regions (orange) and protein coding sequence (blue) including 7 PHD fingers (yellow), FYRN (green), FYRC (green), and a SET domain (red). Arrows indicate the locations of 81 mutations affecting 70 sites found in 110 families with Kabuki syndrome including: 37 nonsense, 22 frameshifts, 16 missense, 3 in-frame deletions/duplications, and 3 splice-site mutations. Asterisks indicate mutations that were confirmed to be de novo and crosses indicate cases for which parental DNA was unavailable. Figure adapted from Ng et al. [2010].

be defined as disease-causing (Supplementary Table II online). In one case, we found both a 21 bp in-frame insertion in exon 39 and a 1 bp insertion in exon 46 predicted to cause a frameshift. However, the unaffected mother also carried the 21 bp insertion suggesting that this is a rare polymorphism, and that the 1 bp deletion is the pathogenic mutation responsible for Kabuki syndrome.

Apparent disease-causing variants were discovered in nearly half (i.e., 22/54) of all protein-coding exons of *MLL2* and in virtually every region known to encode a functional domain (Fig. 1). However, the distribution of variants appeared non-random as 13 and 12 novel variants were identified in exons 48 and 39, respectively. These sites accounted for 25, or more than one-third, of all the novel *MLL2* variants and 31/81 mutations that cause Kabuki syndrome in our cohort. Eleven of the 12 pathogenic variants in exon 39 were nonsense mutations and occurred in regions that encode long polyglutamine tracts.

Four of the families studied herein had two individuals affected with Kabuki syndrome. A pair of monozygous twins with a c.15195G>A nonsense mutation were concordant for mild developmental delay, congenital heart disease, preauricular pits, and palatal abnormalities, but discordant for hearing loss, and a central nervous system malformation. Concordance for mild developmental delay between an affected parent and child was observed in two families with *MLL2* mutations, one with a nonsense mutation, c.13579A>T, p.K4527X, and the other with a missense mutation, c.16391C>T, p.T5464M that was also found in a simplex case. No *MLL2* mutation was found in the remaining affected parent and child pair (Fig. 2).

To examine the relationship between genotype and phenotype, we first compared the frequency of developmental delay, congenital heart disease, cleft lip and/or palate, and structural renal defects between *MLL2* mutation-positive versus *MLL2* mutation-negative cases. No significant difference was observed between groups for three of these four phenotypes (Table Ia). However, renal anomalies were observed in 47% (31/66 cases) of *MLL2* mutation-positive cases compared to 14% (2/14 cases) of *MLL2* mutation-negative cases and this difference was statistically significant ($\chi^2 = 5.1$, df = 1, $P = 0.024$). In 35 cases in two clinical cohorts for whom more complete phenotypic data were available, short stature was observed in 54% (14/26) of *MLL2* mutation-positive cases compared to 33% (3/19 cases) of *MLL2* mutation-negative cases. We also divided the *MLL2* mutation-positive cases into those with nonsense and frameshift mutations and those with missense mutations and compared the frequency of developmental delay, congenital heart disease, cleft lip and/or palate, and structural renal defects between groups. No significant differences were observed between groups (Table Ib).

In 26 independent cases of Kabuki syndrome, including one parent-offspring pair, no *MLL2* mutation was identified. Both persons in the mother–child pair had facial characteristics consistent with Kabuki syndrome (Fig. 2), mild developmental delay, and no major malformations. The mother is of Cambodian ancestry and her daughter is of Cambodian and European American ancestry. In general, most of the *MLL2* mutation-negative Kabuki cases had facial characteristics (Fig. 3) similar to those of the *MLL2* mutation-positive Kabuki cases, and a similar pattern of major malformations (Table I) with the exception of fewer renal abnormalities.

TABLE I. Phenotypic Traits Grouped by *MLL2* Mutation Status (a) and Type (b)

Trait	*MLL2* +	*MLL2* −
Intellectual disability	74/74 (100%)	19/20 (95%)
Mild	51/74 (69%)	10/20 (50%)
Moderate	18/74 (24%)	4/20 (20%)
Severe	4/74 (5%)	3/20 (15%)
Cleft palate, CL/CP	29/72 (40%)	8/18 (44%)
Congenital heart defect	36/71 (51%)	8/19 (42%)
Renal abnormality	31/66 (47%)	2/14 (14%)

Trait	Truncating (N = 59)	Missense (N = 16)
Intellectual disability	54/54 (100%)	15/15 (100%)
Mild	36/54 (67%)	11/15 (73%)
Moderate	13/54 (24%)	4/15 (27%)
Severe	5/54 (9%)	0/15
Cleft palate, CL/CP	23/54 (43%)	3/14 (21%)
Congenital heart defect	30/54 (55%)	4/13 (30%)
Renal anomaly	9/44 (20%)	2/12 (17%)

We screened the *MLL2* mutation-negative cases by aCGH for large deletions or duplications that encompassed *MLL2*. Abnormalities were found in four cases. In one case, a 1.87 kb deletion of chromosome 5 (hg18, chr5:175,493,803–177,361,744) that included *NSD1* and had breakpoints in flanking segmental duplications identical to the microdeletion commonly found in Sotos syndrome, was found. This suggests that this individual has Sotos syndrome, not Kabuki syndrome [Kurotaki et al., 2002]. A second case had a novel 977-kb deletion of chromosome 19q13 (hg18, chr19:61,365,420–62,342,064) encompassing 20 genes. The majority of genes within the deleted region are zinc finger genes, some of which are known to be imprinted in both human and mouse. A third case had a complex translocation t(8;18)(q22;q21). Finally, a fourth case was found to have extra material for the entire chromosome 12. Average log 2 ratio across chromosome 12 was 0.49, most likely representing mosaic aneuploidy of chromosome 12. No aCGH abnormalities were observed in 21 cases and aCGH failed for one case.

DISCUSSION

We have expanded the spectrum of mutations in *MLL2* that cause Kabuki syndrome and explored the relationship between *MLL2* genotype and some of the major, objective phenotypic characteristics of Kabuki syndrome. The majority of variants found to cause Kabuki syndrome are either novel nonsense or frameshift mutations, and appear to arise de novo. While mutations that cause Kabuki syndrome are found throughout the *MLL2* gene, there appear to be at least two exons (39 and 48) in which mutations are identified with a considerably higher frequency. Mutations in these two exons account for nearly half of all mutations found in *MLL2*, while the length of these exons represents ~24% of the *MLL2* open reading frame (ORF). Furthermore, exon 48, the exon in which mutations are most common, comprises only ~7% of the

FIG. 2. Facial photographs of mother and daughter with Kabuki syndrome in whom no causative mutation in *MLL2* was identified. Both have mild developmental delay and no known major malformations.

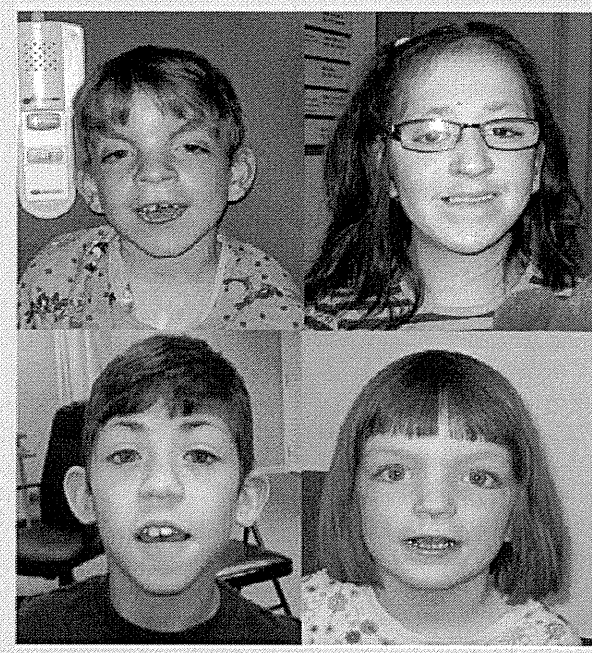

FIG. 3. Facial photographs of four children diagnosed with Kabuki syndrome in whom no causative mutation in *MLL2* was found. The photograph in the upper left was reprinted from Ng et al. [2010].

MLL2 ORF. Exon 39 contains several regions that encode long polyglutamine tracts suggesting the presence of a mutational hotspot, although no such explanation is obvious for exon 48. A stepwise approach in which these regions are the first screened might be a reasonable approach to diagnostic testing. However, capture of all introns, exons, and nearby *MLL2* regulatory regions followed by next-generation sequencing would be more comprehensive and likely to be less costly over the long term.

Comparison of four of the objective clinical characteristics of *MLL2* mutation-negative versus *MLL2* mutation-positive cases allowed us to explore both the relationship between *MLL2* genotype and Kabuki phenotype and the phenotype of *MLL2* mutation-negative cases. Overall, the clinical characteristics of *MLL2* mutation-positive cases did not differ significantly from *MLL2* mutation-negative cases with the exception that renal anomalies were more common in *MLL2* mutation-positive cases. Similarly, we observed no significant phenotypic—including the severity of developmental delay—differences between individuals grouped by mutation type. However, the phenotypic data available to us for analysis was limited and, for many cases, we lacked specific information about each malformation present. Furthermore, the most typical phenotypic characteristic, the distinctive facial appearance, was not compared in detail between cases although it would be of interest to study facial images "blinded" to mutation status to investigate its power to predict genotype. Analysis of genotype–phenotype relationships using both a larger set of Kabuki cases, and with access to more comprehensive phenotypic information would be valuable.

No *MLL2* mutation could be identified in 26 of the cases referred to us with a diagnosis of Kabuki syndrome. In three of these cases, aCGH identified structural variants that could be of clinical significance although additional investigation is required. A fourth case had the classical deletion observed in individuals with Sotos syndrome, and in retrospect it appears that this case was included in the cohort erroneously. The 22 remaining cases, including 1 parent-offspring pair, represent individuals with fairly classic phenotypic features of Kabuki syndrome without a *MLL2* mutation. This observation suggests that Kabuki syndrome is genetically heterogeneous. To this end, in these 22 cases, we sequenced the protein-coding exons of *UTX*, a gene that encodes a protein that directly interacts with *MLL2* but no pathogenic changes were found (data not shown). Exome sequencing of a subset of these *MLL2* mutation-negative cases to identify other candidate genes for Kabuki syndrome is underway.

Whether Kabuki syndrome is the most appropriate diagnosis for the *MLL2* mutation-negative cases is unclear. Some of the *MLL2* mutation-negative cases appear to have a facial phenotype that differs somewhat from that of the *MLL2* mutation-positive cases. Whether these *MLL2* mutation-negative cases diagnosed by expert clinicians should be considered Kabuki syndrome, a variant thereof, or a separate disorder remains to be determined. Our opinion is that

there is simply not yet enough information to make an informed decision about this issue.

Most of the mutations in *MLL2* are predicted to result in haploinsufficiency. However, it is unclear by what mechanism(s) haploinsufficiency of *MLL2* could cause Kabuki syndrome. *MLL2* encodes a histone 3 lysine 4 (H3K4) methyltransferase, one of at least 10 proteins (genes for which have not to our knowledge yet been screened in Kabuki cases in which *MLL2* mutations were not found) that have been identified to specifically modify the lysine residue at the fourth amino acid position of the histone H3 protein [Kouzarides, 2007]. MLL2 has a SET domain near its C-terminus that is shared by yeast Set1, *Drosophila* Trithorax (TRX) and human MLL1 [FitzGerald and Diaz, 1999]. MLL2 appears to regulate gene transcription and chromatin structure in early development [Prasad et al., 1997]. In mice, loss of *MLL2* results in embryonic lethality before E10.5, and while $Mll2^{+/-}$ mice are viable, they are smaller than wild-type [Ng et al., 2010].

Kabuki syndrome is the most common of a small, but growing group of multiple malformation syndromes accompanied by developmental delay that are caused by mutations in genes that encode proteins involved in histone methylation [De Sario, 2009]. The most notable of these is CHARGE syndrome, which is one of the syndromes often considered in the differential diagnosis of children ultimately diagnosed with Kabuki syndrome. CHARGE syndrome is caused by mutations in *CHD7*, which encodes a chromodomain protein that recognizes the trimethylated H3K4 side chain [Vissers et al., 2004]. Other disorders caused by defects of histone methylation status include several intellectual disability syndromes, some of which are also characterized by malformations (e.g., cleft lip/palate) that overlap with those found in individuals with Kabuki syndrome.

Kabuki syndrome is one of the most common causes of heritable developmental delay. Discovery that mutations in *MLL2* are the most common cause of Kabuki syndrome highlights the role that disrupted regulation of histone methylation plays as a cause of human birth defects. Characterizing the spectrum of mutations in *MLL2* is a small but important first step toward understanding the mechanism(s) that underlies Kabuki syndrome.

ACKNOWLEDGMENTS

We thank the families for their participation and the Kabuki Syndrome Network for their support. Our work was supported in part by grants from the National Institutes of Health/National Heart Lung and Blood Institute (5R01HL094976 to D.A.N. and J.S.), the National Institutes of Health/National Human Genome Research Institute (5R21HG004749 to J.S., 1RC2HG005608 to M.J.B., D.A.N., and J.S.; and 5RO1HG004316 to H.K.T.), National Institute of Heath/National Institute of Environmental Health Sciences (HHSN273200800010C to D.N.), National Institute of Neurological Disorders and Stroke (RO1NS35102 to C.A.M.), NIHR Manchester Biomedical Research Centre (D. D.), Ministry of Health, Labour and Welfare (K.Y., N.M., T.O., and N.N.), Ministry of Health, Labour and Welfare of Japan (N.M), Japan Science and Technology Agency (N.M.), Society for the Promotion of Science (N.M.), the Life Sciences Discovery Fund (2065508 and 0905001), the Washington Research Foundation, and the National Institutes of Health/National Institute of Child Health and Human Development (1R01HD048895 to M.J.B. and 5K23HD057331 to A.E.B.). S.B.N. is supported by the Agency for Science, Technology and Research, Singapore. A.W.B. is supported by a training fellowship from the National Institutes of Health/National Human Genome Research Institute (T32HG00035).

REFERENCES

Adam MP, Hudgins L. 2005. Kabuki syndrome: A review. Clin Genet 67: 209–219.

De Sario A. 2009. Clinical and molecular overview of inherited disorders resulting from epigenomic dysregulation. Eur J Med Genet 52:363–372.

FitzGerald KT, Diaz MO. 1999. ML L2: A new mammalian member of the trx/MLL family of genes. Genomics 59:187–192.

Kouzarides T. 2007. Chromatin modifications and their function. Cell 128:693–705.

Kuroki Y, Suzuki Y, Chyo H, Hata A, Matsui I. 1981. A new malformation syndrome of long palpebral fissures, large ears, depressed nasal tip, and skeletal anomalies associated with postnatal dwarfism and mental retardation. J Pediatr 99:570–573.

Kurotaki N, Imaizumi K, Harada N, Masuno M, Kondoh T, Nagai T, Ohashi H, Naritomi K, Tsukahara M, Makita Y, Sugimoto T, Sonoda T, Hasegawa T, Chinen Y, Tomita Ha, Kinoshita HA, Mizuguchi A, Yoshiura T, Ki K, Ohta T, Kishino T, Fukushima Y, Niikawa N, Matsumoto N. 2002. Haploinsufficiency of NSD1 causes Sotos syndrome. Nat Genet 30:365–366.

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J. 2010. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 42:790–793.

Niikawa N, Matsuura N, Fukushima Y, Ohsawa T, Kajii T. 1981. Kabuki make-up syndrome: A syndrome of mental retardation, unusual facies, large and protruding ears, and postnatal growth deficiency. J Pediatr 99:565–569.

Niikawa N, Kuroki Y, Kajii T, Matsuura N, Ishikiriyama S, Tonoki H, Ishikawa N, Yamada Y, Fujita M, Umemoto H, et al. 1988. Kabuki make-up (Niikawa-Kuroki) syndrome: A study of 62 patients. Am J Med Genet 31:565–589.

Prasad R, Zhadanov AB, Sedkov Y, Bullrich F, Druck T, Rallapalli R, Yano T, Alder H, Croce CM, Huebner K, Mazo A, Canaani E. 1997. Structure and expression pattern of human ALR, a novel gene with strong homology to ALL-1 involved in acute leukemia and to Drosophila trithorax. Oncogene 15:549–560.

Vissers LE, van Ravenswaaij CM, Admiraal R, Hurst JA, de Vries BB, Janssen IM, van der Vliet WA, Huys EH, de Jong PJ, Hamel BC, Schoenmakers EF, Brunner HG, Veltman JA, van Kessel AG. 2004. Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. Nat Genet 36:955–957.

White SM, Thompson EM, Kidd A, Savarirayan R, Turner A, Amor D, Delatycki MB, Fahey M, Baxendale A, White S, Haan E, Gibson K, Halliday JL, Bankier A. 2004. Growth, behavior, and clinical findings in 27 patients with Kabuki (Niikawa-Kuroki) syndrome. Am J Med Genet Part A 127A:118–127.

OPEN ACCESS Freely available online

PLoS one

Population Genetic Structure of Peninsular Malaysia Malay Sub-Ethnic Groups

Wan Isa Hatin[1], Ab Rajab Nur-Shafawati[1], Mohd-Khairi Zahri[1], Shuhua Xu[3], Li Jin[3], Soon-Guan Tan[5], Mohammed Rizman-Idid[4,6], Bin Alwi Zilfalil[1,2]*, The HUGO Pan-Asian SNP Consortium[¶]

1 Human Genome Centre, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia, **2** Department of Pediatrics, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia, **3** Chinese Academy of Sciences and Max Planck Society (CAS-MPG) Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, **4** Institute of Biological Sciences, Faculty of Science, Universiti Malaya, Kuala Lumpur, Malaysia, **5** Department of Cell and Molecular Biology, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, Selangor, Malaysia, **6** Centre of Research for Computational Sciences and Informatics in Biology, Bioindustry, Environment, Agriculture and Healthcare (CRYSTAL), Faculty of Science, Universiti Malaya, Kuala Lumpur, Malaysia

Abstract

Patterns of modern human population structure are helpful in understanding the history of human migration and admixture. We conducted a study on genetic structure of the Malay population in Malaysia, using 54,794 genome-wide single nucleotide polymorphism genotype data generated in four Malay sub-ethnic groups in peninsular Malaysia (*Melayu Kelantan, Melayu Minang, Melayu Jawa* and *Melayu Bugis*). To the best of our knowledge this is the first study conducted on these four Malay sub-ethnic groups and the analysis of genotype data of these four groups were compiled together with 11 other populations' genotype data from Indonesia, China, India, Africa and indigenous populations in Peninsular Malaysia obtained from the Pan-Asian SNP database. The phylogeny of populations showed that all of the four Malay sub-ethnic groups are separated into at least three different clusters. The *Melayu Jawa, Melayu Bugis* and *Melayu Minang* have a very close genetic relationship with Indonesian populations indicating a common ancestral history, while the *Melayu Kelantan* formed a distinct group on the tree indicating that they are genetically different from the other Malay sub-ethnic groups. We have detected genetic structuring among the Malay populations and this could possibly be accounted for by their different historical origins. Our results provide information of the genetic differentiation between these populations and a valuable insight into the origins of the Malay sub-ethnic groups in Peninsular Malaysia.

Citation: Hatin WI, Nur-Shafawati AR, Zahri M-K, Xu S, Jin L, et al. (2011) Population Genetic Structure of Peninsular Malaysia Malay Sub-Ethnic Groups. PLoS ONE 6(4): e18312. doi:10.1371/journal.pone.0018312

Editor: Henry Harpending, University of Utah, United States of America

Received October 7, 2010; **Accepted** March 3, 2011; **Published** April 5, 2011

Funding: This research was supported by Fundamental Research Grants Scheme (FRGS) project (Grant Number: 203/PPSP/6170025) and Universiti Sains Malaysia (USM) Fellowship Scheme. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: zilfalil@kb.usm.my

¶ The HUGO Pan-Asian SNP Consortium authors with their affiliations are listed in Text S1.

Introduction

Malays (*Melayu*) are an ethnic group who speak Malayo-Polynesian language which is a member of the Austronesian family [1,2]. They predominantly inhabit the Malay Peninsula, the east coast of Sumatra and the coast of Borneo [1]. In Peninsular Malaysia, the Malays consist of various sub-ethnic groups which are believed to have different ancestral origins based on their migrations centuries ago [3]. The Malay Peninsula was once a very strategic port and trading centre, connecting Indochina and the Indonesian archipelago [4]. However, migrating populations from surrounding areas has further confounded the investigation of the origin of Malays.

This study aims to investigate whether the different Malay sub-ethnic groups originate from a single population or several populations by exploring the possibility of genetic structuring. The Malay populations in the western (*Melayu Minang*) and southern parts (*Melayu Jawa* and *Melayu Bugis*) of the Peninsular Malaysia were believed to have had more historical and cultural links with the populations from the Indonesian archipelago compared to the Malay populations in north-eastern regions (*Melayu Kelantan*). The existence of Chinese and Indian in the Malay Peninsula with different timelines throughout the centuries brought varying degrees of cultural influences and genetics admixtures to the Malay populations. Substantial influx of Chinese and Indians were started only during the British colonial era to work as laborers in the tin mines and the plantation industry that were mainly concentrated on the west coast of peninsula [5]. Prior to British colonization, Chinese and Indian traders had established strong trading links with the Malay Peninsula. These early contacts did not cause large scale migration but intermarriage and integration between them and the Malays were common [5]. Moreover, the Indians had been conspicuous in the region very much earlier, since the period of the ancient Hindu Malay kingdoms which arose in the 2nd century such as *Chi Tu, Gangga Negara, Kadaram* and *Langkasuka* that controlled much of the northern Malay Peninsula [6]. These early Malay states were heavily influenced by concepts of religion, government and arts that were brought by the Indians and traces of this influence can still be found in Malays culture despite the later influence of Islam [6,7].

In addition, the existence of indigenous *Orang Asli* (aboriginal peoples) populations in the peninsula such as the Negritos (*Jahai* and *Kensui*) and Proto-Malays (*Temuan*) have also raised questions as to whether they are associated with the first wave of human migration from Africa, or belong to the more recent events of Asian human evolution [8,9]. The Negritos, who speak the Aslian languages which are part of the Austro-Asiatic language family, are of Australo-Malesian affinity and share some common physical features with African pygmy populations, including short stature, woolly hair and dark skin [8,10]. These nomadic hunter-gatherers are believed to be the earliest settlers and original coastal inhabitants of the Malay Peninsula but the arrival of newcomers forced this group further inland, resulting in them being isolated in forested hilly regions, mainly in northern part of Peninsular Malaysia [10,11]. Meanwhile, the Proto-Malays who arrived later than Negritos in 2000 BC were seafaring people and settled mostly in the central and southern regions of Peninsular Malaysia [1,11]. They are Austronesian speakers apart from one tribe, (the *Semelai*) who speak Aslian [9] and embrace people who are similar in appearance to the Malays but of diverse origins, some probably having entered the region by sea in recent centuries whilst others may have been living in the peninsula for thousands of years [9,10]. In contrast, the present-day Malays of the Malay Peninsula are described as Deutero-Malays, the descendants of the Proto-Malays who had admixed with Siamese, Javanese, Sumatran, Indian, Thai, Arab and Chinese traders [12]. However, according to Fix [10], the original Deutero-Malays migrated from southern China (after the migration of the Proto-Malays) over 1500 years ago and their intermarriages with the Proto-Malays and traders of the ancient trade routes resulted in the diverse recent Deutero-Malay populations that became known currently as the Malays. Hence, Malay population structure analysis would not just provide the information on the genetic differentiation between the populations but would also provide insight into the relationship with the indigenous populations in Peninsular Malaysia.

In this study, we used single nucleotide polymorphisms (SNPs) due to their amenability to high throughput genotyping [13]. SNPs are valuable genetic markers for revealing the evolutionary history of populations [14,15,16]. In this analysis more than 54,000 SNPs loci that were shared by 434 individuals were screened to investigate the distribution of genetic variation and population genetic structure of the Malay populations. This number is sufficient to estimate population genetic parameters with statistical confidence [14,17].

The distance-based approaches that are used in this analysis can detect fine-scale population structure of our studied populations and are not computationally demanding compared to model-based approaches [18]. Inference based on this method does not depend on the modeling assumptions and also requires no special marker selection criteria. In addition, the SNPs analyses with distance-based method are very fast, efficient, robust and able to handle relatively small sample sizes, especially when investigating isolated populations that comprised of few individuals [18,19,20]. We implemented these methods to investigate the population genetic structure of four Malay sub-ethnic groups; *Melayu Kelantan*, *Melayu Minang*, *Melayu Jawa*, and *Melayu Bugis* in Peninsular Malaysia. We included the indigenous Proto-Malay and Negrito populations to determine the degree of their genetic relatedness to the Malays.

Materials and Methods

The population sampling of Peninsular Malaysia Malays were done by following the inclusion and exclusion criteria (Table 1). The SNPs genotype data of 71 unrelated individual of four Malay sub-ethnic groups namely Melayu Kelantan, Melayu Minang, Melayu Jawa and Melayu Bugis were generated by Affymetrix GeneChip Mapping Xba 50 K Array, a microarray chip that enabled researchers to screen over 50,000 SNPs loci in each individual. A total of 58,960 SNPs that have been genotyped for all the sampled individuals were screened under the strict criteria of data quality control. Samples with a call rate below than 90% were excluded from further analysis and after the assessment, 4,166 SNPs (7%) were filtered out (Unmapped to Affymetrix annotation file, chromosome X SNPs and intersection SNPs with downloaded Pan-Asian SNP genotypes), leaving a total of 54,794 autosomal SNPs as the final genotype data for each individual to be used in further analyses.

The additional 11 populations (Table 1) comprised of Proto-Malays (*Temuan*), Negritos (*Jahai* and *Kensui*), Indonesian Malays (*Melayu*, *Jawa* and *Toraja*), Yunnan Chinese (*Jinuo* and *Wa*), South-West Indians (who speak in *Marathi* and *Telegu* language) and Africans (*Yoruba*) were obtained from the Pan-Asian SNP database [21] (http://www4a.biotec.or.th/PASNP). All of these genotype data were generated from DNA samples that were collected with informed and written consent and approved by local ethics committees (Research and Ethics (Human) Committee, School of Medical Sciences, Universiti Sains Malaysia (USM)) and institutional review board (IRB) of the respective countries.

Table 1. List of all studied populations with the location and sample ID.

Population (No. of samples)	Sample ID	Location
Malaysian Melayu[a]:		Peninsular Malaysia:
Kelantan (18)	MY-KN	Kelantan
Minang (20)	MY-MN	Negeri Sembilan
Jawa (19)	MY-JV	Johor
Bugis (14)	MY-BG	Johor
Proto-Malay[b]:		
Temuan (49)	MY-TM	Negeri Sembilan
Negrito[b]:		
Jahai (50)	MY-JH	Perak
Kensui (30)	MY-KS	Kedah
Indonesian[b]:		Indonesia:
Melayu (12)	ID-ML	Sumatra
Jawa (19)	ID-JV	Java Island
Toraja (20)	ID-TR	Sulawesi
Chinese[b]:		China:
Jinuo (29)	CN-JN	Yunnan
Wa (56)	CN-WA	Yunnan
Indian[b]:		India:
Marathi (14)	IN-WL	Maharashtra
Telugu (24)	IN-DR	Andhra Pradesh
African[b]:		Africa:
Yoruba (60)	YRI	Nigeria

[a]The inclusion criteria are; the sampled individual of a population must be at least three generations of the same population, no parental admixture and communicate daily in the local dialect. The exclusion criteria are those that contradict the inclusion criteria.

[b]The genotype data that were obtained from the Pan-Asian SNP Consortium database.

doi:10.1371/journal.pone.0018312.t001

Allele frequency and genetic distance based on Fixation Index Statistic (Fst) [22] were calculated by PEAS v1.0 [23]. MEGA 4 software [24] and two programs, Neighbor and Consense from PHYLIP 3.67 [25], were implemented to construct the Neighbor Joining tree [26] using all 54,794 autosomal SNPs, shared by 434 individuals from 15 populations. The tree was rooted using *Yoruba* (YRI) as outgroup. Bootstrapping test was performed 1000 times, whereby branches with less than 80% bootstrap values have been dissolved. Multi-dimensional scale analysis was done by SPSS 13 and represented in Euclidean distance three dimension (3D) model.

Results and Discussion

Fst [22] is a method to show population genetic structure by partitioning genetic variance within populations relative to between populations. The Neighbor Joining tree (Figure 1) based on the genetic distance measure of Fst for the 15 studied populations showed strongly supported nodes (>95% of bootstrap values) and was rooted using *Yoruba* (YRI) as an outgroup. *Yoruba* is an ethnic group from Nigeria and serve as an outgroup to the non-African populations in our study. As the sampling procedure stringently followed the inclusion and exclusion criteria that emphasized the three generations without any different ethnic admixture rule for an individual to be considered as a valid subject for this study, we assumed that there was no recent admixture or gene flow among all the studied populations.

In the Fst tree (Figure 1), each of the Malay sub-ethnic groups in Peninsular Malaysia; *Melayu Kelantan* (MY-KN), *Melayu Minang* (MY-MN), *Melayu Jawa* (MY-JV) and *Melayu Bugis* (MY-BG) is monophyletic, thus establishing there is substructure among Malays. However, the tree does not support the designation of Malays as a monophyletic group since the MY-KN were on a separated clade from other Malay populations, and the MY-JV were more closely related to Proto-Malays and Chinese than other Malays. Generally, the populations were assigned into three different clusters (Cluster I, II, and III) instead of forming a single Malays cluster. Different ancestry across the Malay groups is likely, as they are a paraphyletic class.

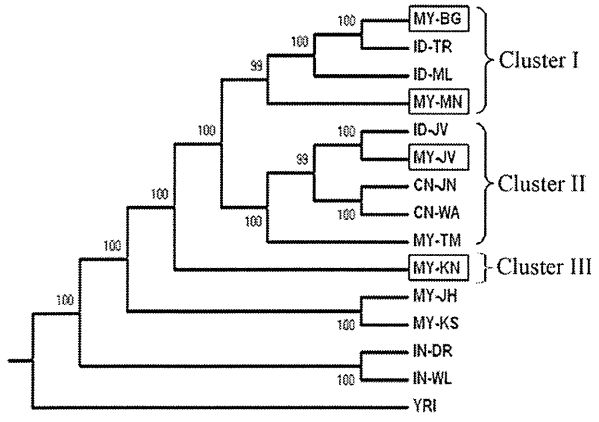

Figure 1. Neighbor-Joining tree of 15 populations based on Fst measurement. In the square boxes are the four studied Malay sub-ethnic groups. Numbers at each branch are represent the percentage value of a thousand bootstrap replications and branches with bootstrap values less than 80% were condensed. The tree suggests a diverse origin of the Malay sub-ethnic groups that forms Cluster I, II and III. Notably, the position of MY-KN in Clade III is the most basal among other studied Malays supported by 100% of bootstrap replicates. There is a distinct genetic difference between the indigenous *Orang Asli* populations; the Negritos is oldest among the peopling groups in Malay Peninsula, whereas the Proto-Malays shared a common ancestry or have had some mixing with the Chinese and Javanese populations.
doi:10.1371/journal.pone.0018312.g001

In Cluster I, the MY-MN was grouped with Indonesian *Melayu* (ID-ML), whereas MY-BG was grouped with Indonesian *Toraja* (ID-TR). The topology may reflect the migrations of MY-MN and MY-BG to Malay Peninsula from Sumatra and Sulawesi, which are also the geographic origins of ID-ML and ID-TR, respectively. Between these populations, MY-MN appeared as the more basal group than MY-BG, which may suggest populations in Sumatra may have separated earlier than those in Sulawesi.

Cluster II grouped the *Melayu Jawa* (MY-JV) together with Indonesian *Jawa* (ID-JV), which suggest past migration between these populations, or common ancestry. Both populations cluster with the Chinese groups (CN-JN and CN-WA). The Chinese may have had more widespread admixture with the *Jawa* people rather than other Malays, Malaysian and Indonesian in this study. As both Malaysian and Indonesian *Jawa* have very close genetic relationship with the Chinese, it could be postulated that the mixture happened before the migration event of the *Jawa* people to Malay Peninsula around 15th century [27].

In Cluster III, the *Melayu Kelantan* (MY-KN) were basal compared to other Malays on the tree. Interestingly it formed an independent clade and placed outside, rather than within the two mentioned clades. The topology might suggest that MY-KN may have had an ancestry that is more divergent than those of the other Malay populations. This could also be attributed to their geographical location at the northern part of Peninsular Malaysia, which would account for their limited links with populations from the Indonesian archipelago. In contrast, MY-MN, MY-BG and MY-JV, have settled on the western and southern regions of the peninsula in proximity to the Indonesian archipelago.

The other explanation for the paraphyletic nature of the Malay class could be admixture of MY-KN with Indian populations (represented by IN-DR and IN-WL). The influence of Hinduism from India was historically very great and the Malays were largely 'Indianized' before they were converted to Islam [7]. Although Hinduism also existed in some of the Indonesian islands (eg. Java Island), it was more predominant among the cultures of populations in mainland Southeast Asia such as Thailand, Cambodia, Myanmar which had more direct contact with the Indian populations [28]. And, the northern part of Peninsular Malaysia had more historical connections with these civilizations [7,29] since centralization of the ancient Indianized kingdoms had occurred in that region for centuries in the early millennium.

Possible admixture between Malays and Indians was first shown genetically using biochemical markers [30]. Even though the admixture could have occurred during the British colonial period from the 19th to the middle of the 20th century when massive migration of Indian laborers to the west coast of Peninsular Malaysia to work on the railroad and in the rubber and oil palm plantation industries took place [5], we believed that the admixture between MY-KN and Indians was very ancient and had happened during the early existence of the Malays. According to the 2010 Malaysian population census, Malaysia's population is about 28.9 million and the Indian community is the smallest of the three main ethnic groups, comprising 6.8% of the population, with most of them residing in the western and north-western regions of Peninsular Malaysia which are the location of the big cities and large urban areas in the country. In Kelantan state which is the origin of MY-KN, the total population is about 1.67 million and the percentage of the Indian community is only 0.2% of the population. The Indians are not a large component of the

Kelantan population either during or after the British colonial era, as it is an agrarian state with lush paddy fields and rustic fishing villages without any plantation industry to attract the Indian immigrants to this north-eastern part of the peninsula [5].

Regarding the phylogenetic affinities of aboriginal peoples in Malaysia, it was revealed that the Proto-Malay, *Temuan* (MY-TM) population was more related to the Chinese and Malays, especially with *Jawa* populations than the Negritos, represented by the *Kensui* (MY-KS) and *Jahai* (MY-JH) populations. This topology is consistent with the fact that the tribal Proto-Malays are believed to have migrated from Yunnan, China about 4,000–6,000 years ago [1]. They were once probably people of coastal Borneo who expanded into Sumatra and the Malay Peninsula as a result of their seafaring way of life [2,31]. Thus, our results may provide a genetic evidence of the pre-historic migration of Proto-Malays from Yunnan, China.

On the other hand, the Negritos are regarded as the earliest inhabitants of the Malay Peninsula and are probably descended from the Hoabinhians, as their mtDNA variation shows strong evidence for indigenous origins within Malay Peninsula, with time depth of ~60,000 years ago [9,32]. However, their origin and the route of their migration to Asia is still a matter of great speculation [33,34]. Nevertheless, the suggestion of their origin via a southern route of migration from South India is also plausible, as the recent mtDNA studies on 'relict' populations of Southeast Asia and the Andaman and Nicobar Islands also point to the human dispersals through the southern exit route [32,35,36]. The phylogenetic tree showed a concordance with the facts mentioned above, where the position of the Negritos on the tree were placed at the most basal position among the Malay, Indonesian and Chinese populations. The genetic relationship of the Malays and aboriginal peoples has not only provided some additional insight to the initial peopling in Malay Peninsula, but also may allow one to gauge admixture due to more recent migration.

The genetic structure of Malays in the Fst tree is recapitulated by multi dimensional scale (MDS) analysis in three dimensions (3D) model as shown in Figure 2. Notably, all four Malay sub-ethnic groups are well separated into three different sub-clusters, although they still remained in the same dimensional platform (dimension 3) indicating an existence of substructure within the Malay population. Malay sub-ethnic groups clustered together with Chinese populations (CN-JN and CN-WA) on the middle area of the dimension 3 platform and far separated from three other group populations which are *Yoruba*, Indians (IN-DR and IN-WL) and Negritos (MY-JH and MY-KS) which are far more diversified than the modern Malays.

Conclusion

One of the goal of population genetics to understand the nature and extent of human population structure [37]. We have utilized the distance-based clustering method to show the population genetic structure of Malays and the existence of differences among them. The detected substructure of Malays of Peninsular Malaysia indicates the existence of genetic heterogeneity in the population that might relate to the diverse origins and histories. This study has performed investigation of more comprehensive Malay populations that were not included in the report by Pan-Asian SNP research (PASNPI) [21]. The inclusion of the indigenous populations in this study have shown genetic affinities which have not been revealed in previous studies.

Our results illustrate the potential to investigate further the peopling of Peninsular Malaysia by including more ethnic groups not covered in this study. For a culturally mixed country such as

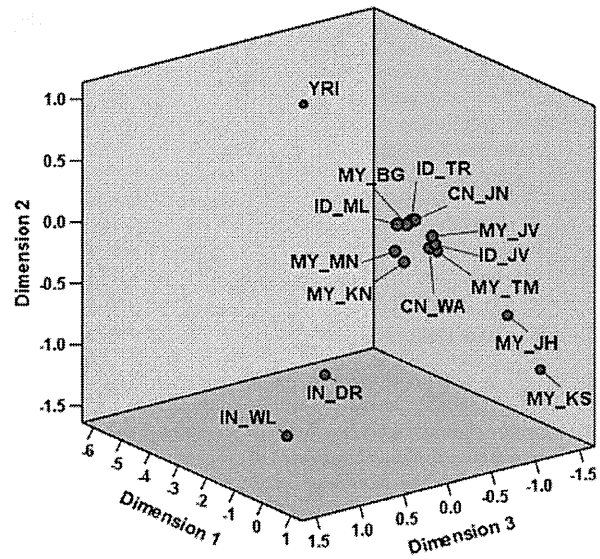

Figure 2. MDS analysis in three dimension model recapitulated the pattern of Fst tree. The 3D MDS showed that all four Malay populations are well separated into three different sub-clusters, although still remained in the same cluster and dimensional platform. They are far separated from three other group populations which are Yoruba, Indian (IN-DR and IN-WL) and Negrito (MY-JH and MY-KS) which are far more diversified than the modern Malays.
doi:10.1371/journal.pone.0018312.g002

Malaysia, where people of various ethnicity practice different lifestyles under many different environments, the knowledge of population genetic substructure is important for proper design of case control association studies and for identifying disease predisposing alleles that may differ across ethnic groups [38]. Only by characterizing genetic variation among individuals and populations, can we gain a better understanding of differential susceptibility to disease, differential response to pharmacological agents and complex interaction of genetic and environmental factors in producing phenotypes [38].

Supporting Information

Text S1 The participants of the HUGO Pan-Asian SNP Consortium are arranged alphabetically by surname.
(DOC)

Acknowledgments

We acknowledged the contributions by the other members of this study group from the School of Health Sciences and the School of Dental Sciences, Universiti Sains Malaysia. We thank all the subjects who have participated in this research and those who have helped us in the data collection. Special thanks to the UKM Medical Molecular Biology Institute (UMBI), Universiti Kebangsaan Malaysia and Matrix Analytical Sdn. Bhd., Malaysia for allowing us to use their laboratory facilities.

Author Contributions

Conceived and designed the experiments: NSAR ZMK PASNPI. Performed the experiments: NSAR ZMK PASNPI. Analyzed the data: HWI. Contributed reagents/materials/analysis tools: SX LJ PASNPI. Wrote the paper: HWI. Population sampling: NSAR HWI ZBA PASNPI. Interpretation of results: HWI SX RIM. Critical review of manuscript: TSG ZBA RIM SX LJ HWI NSAR.

References

1. Bellwood PS (1997) Prehistory of the Indo-Malaysian Archipelago. Honolulu, Hawaii: University of Hawai'i Press. x, 384 p., [340] p. of plates p.
2. Omar AH (2004) Languages and Literature. The Encyclopedia of Malaysia.
3. Paul W (1961) The Golden Khersonese: Studies in the Historical Geography of the Malay Peninsula before AD 1500. Kuala Lumpur: University of Malaya Press.
4. Jacq-Hergoualc'h M, Hobson V The Malay Peninsula: Crossroads of the Maritime Silk Road (100 BC-1300 AD): BRILL. 607 p.
5. (2008) Malaysia, Singapore, Brunei, and the Philippines. New York: Marshall Cavendish Reference. 1584 p.
6. Arasaratnam S Indians in Malaysia and Singapore: Published for the Institute of Race Relations, London, by Oxford University Press (Bombay). 214 p.
7. I Syukri (2002) Sejarah Kerajaan Melayu Patani/Ibrahim Syukri. Bangi: Universiti Kebangsaan Malaysia. pp 131, 129.
8. Allen FA (1879) The Original Range of The Papuan and Negrito Races. The Journal of the Anthropological Institute of Great Britain and Ireland 8: 38–50.
9. Hill C, Soares P, Mormina M, Macaulay V, Meehan W, et al. (2006) Phylogeography and ethnogenesis of aboriginal Southeast Asians. Mol Biol Evol 23: 2480–2491.
10. Fix AG (1995) Malayan paleosociology: implications for patterns of genetic variation amongst the Orang Asli. American Anthropology 97: 313–323.
11. Carey I (1976) Orang Asli: the aboriginal tribes of peninsular Malaysia. Kuala Lumpur; New York: Oxford University Press.
12. Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bosch E, et al. (1998) Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. Am J Hum Genet 63: 1824–1838.
13. Brookes AJ (1999) The essence of SNPs. Gene 234: 177–186.
14. Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. TRENDS in Ecology and Evolution 18: 249–256.
15. Kuhner MK, Beerli P, Yamato J, Felsenstein J (2000) Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics 156: 439–447.
16. Petkovski E, Keyser C, Ludes B, Hienne R (2003) Validation of SNPs as markers for individual identification. International Congress Series 1239: 33–36.
17. Nielsen R (2000) Estimation of population parameters and recombination rates using single nucleotide polymorphisms. Genetics 154: 931–942.
18. Gao X, Starmer J (2007) Human population structure detection via multilocus genotype clustering. BMC Genet 8: 34.
19. Mihaescu R, Levy D, Pachter L (2009) Why Neighbor-Joining Works. Algorithmica 54: 1–24.
20. Crandall K, Lagergren J, Simonsen M, Mailund T, Pedersen C (2008) Rapid Neighbour-Joining. Algorithms in Bioinformatics: Springer Berlin/Heidelberg. pp 113–122.
21. The-HUGO-Pan-Asian-SNP-Consortium (2009) Mapping Human Genetic Diversity in Asia. Science 326: 1541–1545.
22. Weir BS, Hill WG (2002) Estimating F-statistics. Annu Rev Genet 36: 721–750.
23. Xu SH, Gupta S, Jin L (2010) PEAS V1.0: a package for elementary analysis of SNP data. Molecular Ecology Resources 10: 1085–1088.
24. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
25. Felsenstein J (2007) PHYLIP: Phylogeny Inference Package. University of Washington.
26. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.
27. Taylor JG, ebrary Inc (2003) Indonesia peoples and histories. New Haven: Yale University Press. p xxi, 420.
28. Wolters OW, Reynolds CJ, Cornell University. Southeast Asia P (2008) Early Southeast Asia: selected essays. Ithaca, New York: Southeast Asia Program, Cornell University.
29. Allen SJ (1998) History, Archaeology, and the Question of Foreign Control in Early Historic-Period Peninsular Malaysia. International Journal of Historical Archaeology 2: 261–289.
30. Teng YS, Tan SG (1979) Genetic evidence of gene flow from Indians to Malays. Journal of Human Genetics 24: 1–8.
31. Rahman NHSNA (1998) Early History. Encyclopedia of Malaysia. 46 p.
32. Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, et al. (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308: 1034–1036.
33. Consortium THP-AS (2009) Mapping human genetic diversity in Asia. Science 326: 1541–1545.
34. Kashyap V, Sitalaximi T, Sarkar B, Trivedi R (2003) Molecular relatedness of the aboriginal groups of Andaman and Nicobar Islands with similar ethnic populations. The International Journal of Human Genetics 3: 5–11.
35. Majumder PP (2008) Genomic inferences on peopling of south Asia. Curr Opin Genet Dev 18: 280–284.
36. Thangaraj K, Chaubey G, Kivisild T, Reddy AG, Singh VK, et al. (2005) Reconstructing the origin of Andaman Islanders. Science 308: 996.
37. Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, et al. (2003) Human population genetic structure and inference of group membership. Am J Hum Genet 72: 578–589.
38. Tishkoff SA, Kidd KK (2004) Implications of biogeography of human populations for 'race' and medicine. Nature Genetics 36.

Please cite this article in press as: Okada et al., *SMOC1* Is Essential for Ocular and Limb Development in Humans and Mice, The American Journal of Human Genetics (2011), doi:10.1016/j.ajhg.2010.11.012

ARTICLE

SMOC1 Is Essential for Ocular and Limb Development in Humans and Mice

Ippei Okada,[1,14] Haruka Hamanoue,[1,2,14] Koji Terada,[3] Takaya Tohma,[4] Andre Megarbane,[5] Eliane Chouery,[5] Joelle Abou-Ghoch,[5] Nadine Jalkh,[5] Ozgur Cogulu,[6] Ferda Ozkinay,[6] Kyoji Horie,[7] Junji Takeda,[7,8] Tatsuya Furuichi,[9,10] Shiro Ikegawa,[9] Kiyomi Nishiyama,[1] Satoko Miyatake,[1] Akira Nishimura,[1] Takeshi Mizuguchi,[1,15] Norio Niikawa,[11,12] Fumiki Hirahara,[2] Tadashi Kaname,[13] Koh-ichiro Yoshiura,[12] Yoshinori Tsurusaki,[1] Hiroshi Doi,[1] Noriko Miyake,[1] Takahisa Furukawa,[3] Naomichi Matsumoto,[1,*] and Hirotomo Saitsu[1,*]

Microphthalmia with limb anomalies (MLA) is a rare autosomal-recessive disorder, presenting with anophthalmia or microphthalmia and hand and/or foot malformation. We mapped the MLA locus to 14q24 and successfully identified three homozygous (one nonsense and two splice site) mutations in the SPARC (secreted protein acidic and rich in cysteine)-related modular calcium binding 1 (*SMOC1*) in three families. *Smoc1* is expressed in the developing optic stalk, ventral optic cup, and limbs of mouse embryos. *Smoc1* null mice recapitulated MLA phenotypes, including aplasia or hypoplasia of optic nerves, hypoplastic fibula and bowed tibia, and syndactyly in limbs. A thinned and irregular ganglion cell layer and atrophy of the anteroventral part of the retina were also observed. Soft tissue syndactyly, resulting from inhibited apoptosis, was related to disturbed expression of genes involved in BMP signaling in the interdigital mesenchyme. Our findings indicate that SMOC1/Smoc1 is essential for ocular and limb development in both humans and mice.

Introduction

Microphthalmia with limb anomalies (MLA [MIM 206920]), also known as Waardenburg anophthalmia syndrome or ophthalmoacromelic syndrome, is a rare autosomal-recessive disorder first described by Waardenburg.[1] It is characterized by ocular anomalies ranging from mild microphthalmia to true anophthalmia and by limb anomalies such as oligodactyly, syndactyly, and synostosis of the 4th and 5th metacarpals.[2–4] The genetic cause for MLA has remained unknown.

It is widely known that secreted signaling molecules such as Sonic hedgehog (Shh), wingless-type MMTV integration site family (Wnt), transforming growth factor β (Tgf-β), bone morphogenetic proteins (Bmps), and fibroblast growth factor (Fgf) are involved in the development of many organs and tissues, including the eyes and limbs.[5,6] In particular, mutations in *BMP4* (MIM 112262) have resulted in anophthalmia with systemic manifestations, including polydactyly and/or syndactyly (also known as micropthalmia, syndromic 6, MCOPS6 [MIM 607932]),[7] highlighting importance of BMP signaling in both the developing eye and limb.

SMOC1 (MIM 608488), which encodes SPARC (secreted protein acidic and rich in cysteine)-related modular calcium binding 1, is a member of the SPARC (also known as BM-40) matricellular protein family that modulates cell-matrix interaction by binding to many cell-surface receptors, the extracellular matrix, growth factors, and cytokines.[8,9] SMOCs are extracellular glycoproteins with five domains: an N-terminal follistatin-like (FS) domain, two thyroglobulin-like (TY) domains, a domain unique to SMOC, and an extracellular calcium-binding (EC) domain.[9] SMOC1 is widely expressed in various tissues with localization to basement membranes.[9,10] Although the biological function of SMOC1 remains largely unknown, it has been recently reported that *Xenopus* smoc protein, the ortholog of human SMOC1, acts as a BMP antagonist,[11] suggesting that human SMOC1 can also modulate BMP signaling.

Here, we demonstrate that *SMOC1* mutations cause MLA. We also show that *Smoc1* null mice recapitulated

[1]Department of Human Genetics, Yokohama City University Graduate School of Medicine, 3-9 Fukuura, Kanazawa-ku, Yokohama 236-0004, Japan; [2]Department of Obstetrics and Gynecology, Yokohama City University Graduate School of Medicine, 3-9 Fukuura, Kanazawa-ku, Yokohama 236-0004, Japan; [3]Department of Developmental Biology, Osaka Bioscience Institute, 6-2-4 Furuedai, Suita, Osaka 565-0874, Japan; [4]Division of Pediatrics, Okinawa Prefectural Nanbu Medical Center & Children's Medical Center, 118-1 Ikyoku, Arakawa, Haebaru, Okinawa 901-1193, Japan; [5]Medical Genetics Unit, St. Joseph University, Beirut 1104-2020, Lebanon; [6]Department of Pediatrics, Ege University Faculty of Medicine, 35100 Bornova-Izmir, Turkey; [7]Department of Social and Environmental Medicine, Graduate School of Medicine, Osaka University, 2-2 Yamadaoka, Suita, Osaka 565-0871, Japan; [8]Center for Advanced Science and Innovation, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan; [9]Laboratory for Bone and Joint Disease, Center for Genomic Medicine, RIKEN, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan; [10]Laboratory Animal Facility, Research Center for Medical Sciences, Jikei University School of Medicine, 3-25-8, Nishi-Shimbashi, Minato-ku, Tokyo 105-8461, Japan; [11]Research Institute of Personalized Health Sciences, Health Sciences University of Hokkaido, Ishikari-Tobetsu, Hokkaido 061-0293, Japan; [12]Department of Human Genetics, Nagasaki University Graduate School of Biomedical Sciences, Sakamoto 1-12-4, Nagasaki 852-8523, Japan; [13]Department of Medical Genetics, University of the Ryukyus Faculty of Medicine, 207 Uehara, Nishihara, Okinawa 903-0215, Japan
[14]These authors contributed equally to this work
[15]Current address: Laboratory of Biochemistry and Molecular Biology, National Cancer Institute, National Institutes of Health, Building 37, Room 6050, Bethesda, MD 20892, USA
*Correspondence: naomat@yokohama-cu.ac.jp (N.M.), hsaitsu@yokohama-cu.ac.jp (H.S.)
DOI 10.1016/j.ajhg.2010.11.012.