The HUGO Journal
a SpringerOpen Journal

ORIGINAL RESEARCH Open Access

A genome wide pattern of population structure and admixture in peninsular Malaysia Malays

Wan Isa Hatin[1], Ab Rajab Nur-Shafawati[1], Ali Etemad[2], Wenfei Jin[3], Pengfei Qin[3], Shuhua Xu[3], Li Jin[3], Soon-Guan Tan[4], Pornprot Limprasert[5], Merican Amir Feisal[6,7], Mohammed Rizman-Idid[6], Bin Alwi Zilfalil[1,2*] and The HUGO Pan-Asian SNP Consortium

Abstract

Background: The Malays consist of various sub-ethnic groups which are believed to have different ancestral origins based on their migrations centuries ago. The sub-ethnic groups can be divided based on the region they inhabit; the northern (*Melayu Kedah* and *Melayu Kelantan*), western (*Melayu Minang*) and southern parts (*Melayu Bugis* and *Melayu Jawa*) of Peninsular Malaysia. We analyzed 54,794 autosomal single nucleotide polymorphisms (SNPs) which were shared by 472 unrelated individuals from 17 populations to determine the genetic structure and distributions of the ancestral genetic components in five Malay sub-ethnic groups namely *Melayu Bugis*, *Melayu Jawa*, *Melayu Minang*, *Melayu Kedah*, and *Melayu Kelantan*. We also have included in the analysis 12 other study populations from Thailand, Indonesia, China, India, Africa and *Orang Asli* sub-groups in Malay Peninsula, obtained from the Pan Asian SNP Initiative (PASNPI) Consortium and International HapMap project database.

Results: We found evidence of genetic influx from Indians to Malays, more in *Melayu Kedah* and *Melayu Kelantan* which are genetically different from the other Malay sub-ethnic groups, but similar to Thai *Pattani*. More than 98% of these northern Malays haplotypes could be found in either Indians or Chinese populations, indicating a highly admixture pattern among populations. Nevertheless, the ancestry lines of Malays, Indonesians and Thais were traced back to have shared a common ancestor with the Proto-Malays and Chinese.

Conclusions: These results support genetic admixtures in the Peninsular Malaysia Malay populations and provided valuable information on the enigmatic demographical history as well as shed some insights into the origins of the Malays in the Malay Peninsula.

Keywords: Malays; Single nucleotide polymorphisms; Genetic structure; Admixture; Haplotypes

Background

The knowledge of population genetic structure and genetic ancestry hold great potential towards better understanding of the differential susceptibility to disease, response to drugs and complex interaction of genetic and environment factors (Collin et al. 2003; Campbell and Tishkoff 2008). Recent studies have highlighted the importance of characterizing the genetic make-up of admixed populations (Sankararaman et al. 2008; Bryca et al. 2010 and Patterson et al. 2010. The admixtures within individual may affect the interactions between complex genes with other genes and environmental factors, and at the same time, also affect the susceptibility of individual to particular diseases (Collin et al. 2003; Tang et al. 2005; Lao et al. 2006 and Hunley et al. 2009). In addition, the analysis of genetic variations also provides detail knowledge for understanding the ancient human demographic history. The enigmatic history of Malays as well as their morphological features that exhibit fusion from various ethnicities and cultural background (Rahman 1998; Andaya 2001; Reid 2001; Hussein et al. 2007 and Omar 2004) have made them a uniquely complex population and intriguing subject to be studied.

In Peninsular Malaysia, the Malays form the majority of the population (63.1%) followed by Chinese (24.6%)

* Correspondence: zilfalil@gmail.com
[1]Human Genome Centre, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kelantan, Malaysia
[2]Department of Pediatrics, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kelantan, Malaysia
Full list of author information is available at the end of the article

and Indians (7.3%) (Jabatan Perangkaan Malaysia 2010). The intermarriage and integration among them for centuries had given complex admixtures in genome of Malays. Moreover, the Malays also known to have various sub-ethnic groups due to different ancestral origins based on their migrations centuries ago (Paul 1961). Thus, it is important to understand the definition of Malays either sociologically or anthropologically, in order to select the sampling populations which were relevant to the aim of this study. Sociologically, Malays are Malaysian citizen born to a Malaysian citizen who professes the religion of Islam, habitually speaks the Malay language, conforms to Malay custom and is domiciled in Malaysia (Constitution of Malaysia). Anthropologically, the Malays are described as an ethnic group of Austronesian people who speak Malayo-Polynesian language that belong to the Southern Mongoloid group of races and predominantly inhabit the Malay Peninsula (comprises of southern Thailand, Peninsular Malaysia, and the island of Singapore), south coast of Myanmar, eastern Sumatra, the coast of Borneo and the smaller islands between these locations - collectively known as the *Alam Melayu*. These locations today are part of the modern nations of Malaysia (Peninsular and Eastern Malaysia), Indonesia, Singapore, Brunei, Southern Myanmar and Southern Thailand (Omar 2004; Bellwood 1997). The existences of indigenous *Orang Asli* (aboriginal peoples) populations in the Peninsular Malaysia such as the *Semang* and Proto-Malays have also raised questions as to what extent they have contributed to the uniquely admixed gene pool of Malays (Bellwood 1993). The relationship between the Malays and the *Orang Asli*, especially with the *Semang* who are believed to be the earliest settlers and original coastal inhabitants of the Malay Peninsula (Allen 1879; Carey 1976 and Fix 1995) were important to be studied in order to identified the origin of Malays as well as the occupancy of prehistoric human populations in this region.

The previous study has shown that there is genetic substructure among Malays (Hatin et al. 2011). The *Melayu Kelantan* in north-eastern regions was genetically different from other Malay populations in the western (*Melayu Minang*) and southern parts (*Melayu Jawa* and *Melayu Bugis*) of the Peninsular Malaysia (Hatin et al. 2011). Beside, close genetic relationship of the *Melayu Kelantan* with the Indians and the *Orang Asli Semang* (*Jahai* and *Kensiu*) was also established (Hatin et al. 2011). Against these backgrounds, we conducted this study to investigate the extent of admixture in Malays, especially in northern Malays of Peninsular Malaysia using a model-based clustering method. The model-based methods attempt to more directly reconstruct historical events. This method is computationally intensive but it is explicit where the assumptions are stated, not hidden. In addition, we performed haplotype sharing analysis to consider the question of whether any outlier is migrants, experienced admixture or ancient population. Therewith, we included two more populations from northern part of peninsula, which are *Melayu Kedah* and Thai *Pattani* to verify the divergence pattern of the northern Malays.

Results

Pattern of genetic variations among populations

The genetic variations within and between five Peninsular Malaysia Malay sub-ethnic groups and other studied populations were characterized by the pair-wise Fst between populations, followed by the non-parametric Multi-Dimensional Scale (MDS) analysis. The table of pair-wise Fst value that has been multiplied with 1000 is shown in Table 1. All of the genetic distance values that showed closer relationship between populations were shaded in gray color. The genetic divergence between five of the Peninsular Malaysia Malay sub-ethnic groups shows significant difference of the *Melayu Bugis* from the other Malays which is substantially closer to Indonesian *Toraja* (Fst = 0.019).

Melayu Kelantan and *Melayu Kedah* were genetically close to each other (Fst = 0.015). Meanwhile, the genetic divergence between *Melayu Jawa* and *Melayu Minang* (Fst = 0.021) showed that they were closer to *Melayu Kedah* and *Melayu Kelantan* than to each other. Interestingly, these four Malay sub-ethnic groups were significantly closer to Proto-Malays *Temuan*, Indonesian *Jawa* and Chinese *Wa* from Yunnan, China. The genetic distances between the Proto-Malays *Temuan*, Indonesian *Jawa*, and Chinese *Wa* to each other were also substantially lower than to any other populations, even between population within their groups. Although the Thai *Pattani* samples were also close to these group, but they were much closer to *Melayu Kedah* relative to the other Peninsular Malaysia Malays (Fst = 0.018).

In relation with the *Semang* group, both the *Melayu Kedah* and *Melayu Kelantan* samples were relatively closer to the *Jahai* and *Kensiu* than any other Malays. Similarly, both the *Semang* samples were closer to the samples of Indonesian *Jawa* and Chinese *Wa* than any other populations. It is also noted that the genetic distance between *Melayu Kedah* and *Melayu Kelantan* with the Indians, especially with *Telugu* were considerable smaller than to any other populations.

The MDS analysis for 17 populations was performed in two dimensions (2D) and three dimensions (3D) based on Fst genetic distance method as shown in Figure 1. The genetic variation of Malays showed by the pair-wise Fst was recaptured by the MDS scatter plot. The MDS scatter plot in 2D platform (Figure 1A) exhibited that all the Peninsular Malaysia Malay, Indonesian, Thai, Proto-Malay and Chinese populations were scattered closely at the below-right corner of the plot, which are near to the

Table 1 Pair-wise Fst (x 1000) between the Malay sub-ethnic groups and other populations in this study

	MY-BG	MY-JV	MY-MN	MY-KN	MY-KD	TH-PT	MY-TM	MY-JH	MY-KS	ID-JV	ID-ML	ID-TR	CN-JN	CN-WA	IN-WL	IN-DR	YRI
MY-BG																	
MY-JV	24																
MY-MN	24	21															
MY-KN	23	**19**	**18**														
MY-KD	22	**18**	**18**	**15**													
TH-PT	26	21	23	21	**18**												
MY-TM	26	**19**	**21**	**18**	**18**	**23**											
MY-JH	42	34	35	32	31	37	30										
MY-KS	53	47	46	42	41	48	41	23									
ID-JV	22	**16**	**19**	**17**	**17**	**21**	**17**	**32**	44								
ID-ML	26	25	23	23	23	28	25	41	53	23							
ID-TR	**19**	22	21	21	21	25	23	40	52	20	22						
CN-JN	31	25	27	24	23	28	25	40	52	23	31	28					
CN-WA	24	**18**	**20**	**17**	**17**	**21**	**17**	**32**	44	**15**	24	22	17				
IN-WL	57	54	46	**42**	**39**	48	50	57	61	52	57	56	57	51			
IN-DR	51	48	40	**36**	**33**	42	44	50	55	46	51	50	51	45	17		
YRI	112	109	102	99	97	106	104	111	116	107	112	110	112	106	88	84	

*The bold numbers indicate close genetic relationship between the populations.

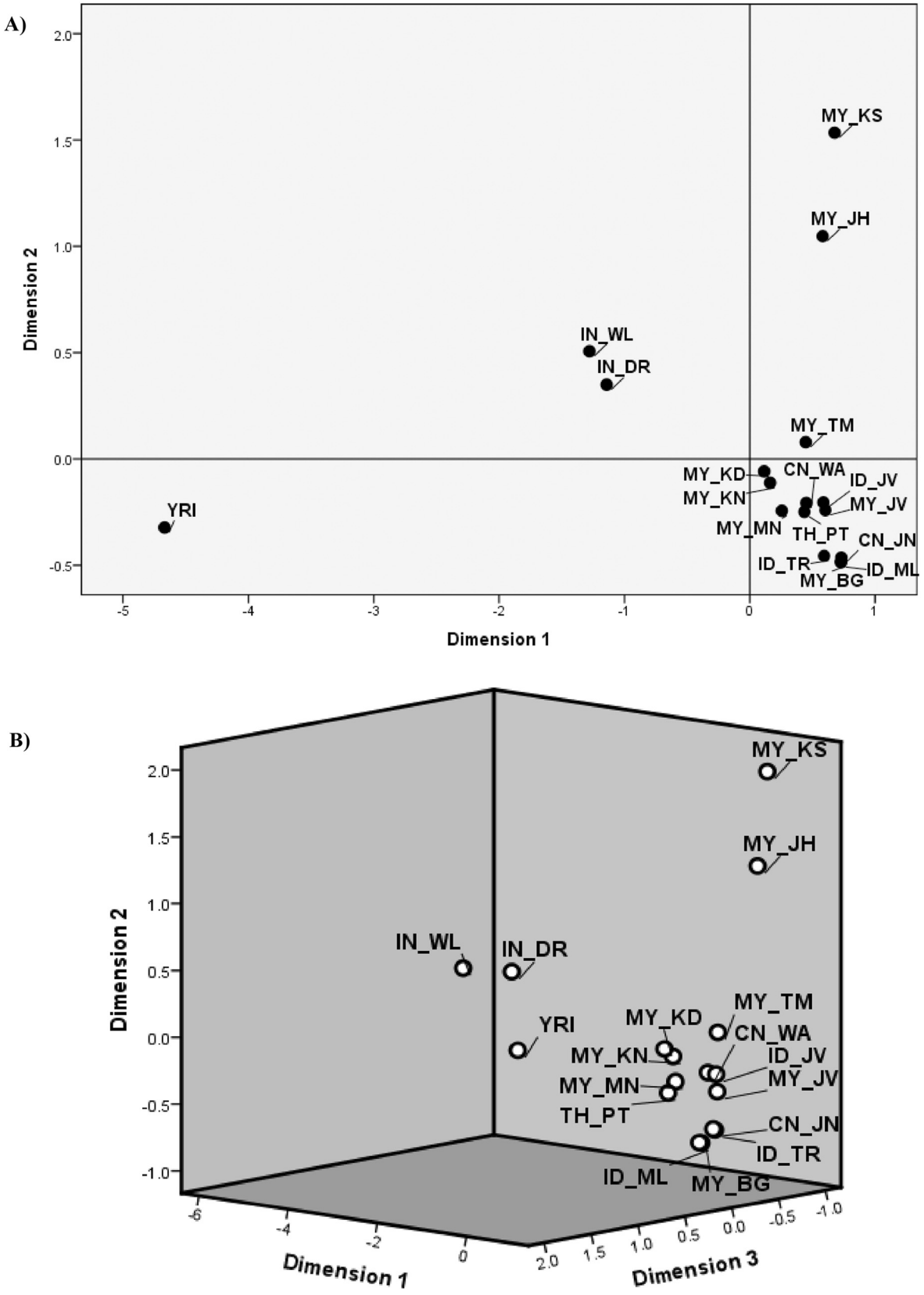

Figure 1 MDS analysis for 17 populations based on Fst. A) two dimensions (2D) and **B)** three dimensions (3D).

intersection of axes. They were far separated from three other group populations which are *Yoruba*, Indians and *Semang* that are far more diversified than the modern Malays. The same pattern also can be seen in the 3D MDS plot (Figure 1B) where all five Peninsular Malaysia Malay sub-ethnic groups were 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 Peninsular Malaysia Malays. The MDS analysis showed that there were possible admixtures among *Melayu Kedah*, *Melayu Kelantan*, *Melayu Minang* and Thai *Pattani*. In the case of *Orang Asli* group, the genetic structure clearly appeared in the Proto-Malay *Temuan* with possible admixture to *Jawa* populations and Chinese *Wa*.

Population genetic structure and ancestry

The assignment of each of the 472 individuals sampled from 17 pre-defined populations into genetically inferred clusters of $K = 2$ to $K = 10$ is shown in Figure 2. Each individual was represented by a thin vertical line, which was partitioned into K color segment that represent the individual's estimated Q fractions in K clusters. Each population was labeled below the figure and separated by the solid line. The results showed that individuals from the same pre-defined population shared almost similar Q values. Any pre-defined populations which shared similar distinctive Q values were merged into inferred cluster, as shown in Table 2.

The most probable number of ancestral clusters was determined by the maximum value of the *Ln(Pr)* of K and by careful observation and comparison of each of the Qs of Ks from multiple runs between the same and different sampling datasets using the SSC. The value of *Ln(Pr)* was observed to increase until $K = 6$ and started to become relatively inconsistent at $K = 7$ onwards. The SSC scores were greater than 0.95 in all cases of $K < 6$ while for larger Ks ($K > 5$) the SSC were slightly lower with a minimal value of 0.90. In analysis of $K > 6$, the splitting orders of clusters varied across different runs and different datasets. However, for the same cluster mode, the SSC of membership coefficient estimates were still high (>0.90). Therefore, the $K = 6$ was considered as the most statistically supported by the data in all of the sampling datasets, depicted by six different colors in such a way as that given by the Q value (Table 2), corresponds to the fraction of genome inferred to have ancestry in the cluster.

At $K = 2$, all samples were separated into two distinct cluster of African (*Yoruba*) in black color, and non-African populations that were grouped into a yellow colored cluster. At $K = 3$, a newly cluster in green color represent *Semang* samples from Peninsular Malaysia (*Jahai* and *Kensiu*). The green component of the *Semang* also could be seen slightly in the Proto-Malay *Temuan*, Thai *Pattani* and two northern Malay sub-ethnic groups; *Melayu Kedah* and *Melayu Kelantan*. The Indian populations remained in the yellow cluster, although parts of their genome were also partitioned into the green component. At $K = 4$, all the Indian samples were assigned into a red colored cluster, which is exclusively separated from the yellow colored cluster of the non-African populations. It should be noted here that Indian proportions in the red colored fractions also appeared in three Peninsular Malaysia Malay sub-ethnic groups; *Melayu Kedah*, *Melayu Kelantan* and *Melayu Minang*, as well as in Thai *Pattani*. At $K = 5$, another cluster in pink color was apparent. This cluster mainly existed in the Proto-Malay *Temuan* and small proportions could be seen in both of the Chinese samples (*Jinuo* and *Wa*).

The structure of the Malays became apparent at $K = 6$, where a new cluster, denoted by the light blue color was confined mainly in the Peninsular Malaysia Malays, Thais and Indonesians samples. The light blue fractions are prominent in the Proto-Malay *Temuan* but rather slightly in the Chinese samples. Interestingly, the Indian components could be seen clearly in the samples of *Melayu Kedah*, *Melayu Kelantan*, *Melayu Minang*, and Thai *Pattani*. The yellow colored cluster mainly belonged to the Chinese samples although the proportions were also associated with the Malays, Proto-Malays, Thais and Indonesians indicating a common ancestor origin for these populations. The proportions of Q for each population in each of the six inferred clusters are shown in Table 2.

Higher K values revealed other clusters as shown in the Q plot at Figure 2. These clusters were generally confined to single populations; orange proportions in *Semang Jahai* for $K = 7$, while purple proportions mainly in Chinese *Jinuo* at $K = 8$. The splitting order of clusters varied greatly across different runs and different data sets. The newly derived clusters of $K = 9$ and $K = 10$ started to lose biological meanings as the real clusters and produced relatively lower Q proportions with unstable patterns in the graph of *Ln(Pr)*.

Nevertheless, the higher the number of *Ks*, the more it resembles or represent the modern genome of studied populations. As shown at $K = 10$ (Figure 2), the genetic structure in the genome of Malays appeared; 1) the *Melayu Bugis* were more delineate to Indonesian *Melayu* and *Toraja*, 2) the *Melayu Jawa* were similar to the Indonesian *Jawa*, with significant components of Chinese in their genomes, 3) the *Melayu Kelantan* and *Melayu Kedah* were more resemble to Thai *Pattani*, with significant admixture from Indian components.

HS pattern in Malays SNPs genotype data

Partitions of the gene pool of Malays by HSAs were conducted in the framework of a three-population comparison

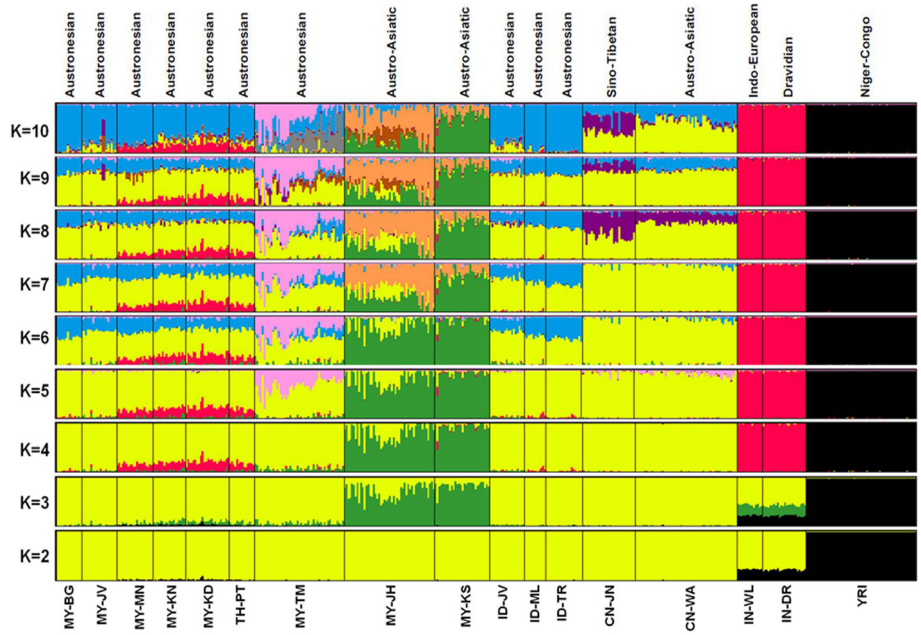

Figure 2 The estimated population structure and ancestral membership coefficients of each of the 472 individuals for K = 2 to K = 10 from dataset S2. The linguistic family of populations were showed at the top of the figure while the name of populations were showed below the figure. Each population was separated by the solid line and each individual was represented by a thin vertical line, which was partitioned into *K* color segment that represent the individual's estimated *Q* fractions in *K* clusters.

as shown in Figure 3. The haplotypes of Malays (MY) were identified as four categories compared with Chinese (CN) and Indians (IN), that is, private in MY (MY), shared with CN only (MY-CN), shared with IN only (MY-IN), and shared with all three populations (MY–CN–IN). Generally, more than 98% of MY haplotypes could be found in either CN or IN with more contributions from IN populations. In the bins 50 kb to 200 kb, both MY (Figure 3A) and CN (Figure 3B) has less than 2% of private haplotypes, whereas IN have more than 3% of private haplotypes (Figure 3C). The same pattern also observed in population samples of northern Malays (NMY) (Figure 3D) that consist of *Melayu Kedah*, *Melayu Kelantan* and Thai *Pattani*. We further confirmed the HS pattern of peninsula Malays (PMY) data which comprised of all five Peninsular Malaysia Malay sub-ethnic groups including Thai *Pattani* compared with the *Orang Asli* Proto-Malays (PM) and *Semang* (NG). The PMY (Figure 3G) has slightly higher percentage of private haplotypes compared to both of the PM (Figure 3H) and NG (Figure 3I). All of these HS percentages were calculated without taking into account the frequencies of distinct haplotypes.

STRUCTURE and HSAs phylogeny

The phylogenetic trees based on Cavalli DC and Nei's DA genetic distances that were reconstructed using allele frequencies in each ancestral component inferred by Bayesian algorithm from the STRUCTURE analyses as well as the HSAs phylogeny are shown in Figure 4. The trees (Figure 4A and 4B) which reflected the identified ancestral clusters ($K = 6$) from STRUCTURE analyses consistently showed that the last split of branches was the clade of Malays, Indonesians and Thais (MIT), clustered together with PM. However, the trees revealed two slightly different topologies. The Cavalli DC showed simultaneously evolutionary divergence between the group NG and CN populations and the group of MIT and PM. Whilst, the topology of Nei's DA support the subsequent divergence of evolutionary processes of the group of populations, which is more similar to the pattern of splitting orders in STRUCTURE analysis. In the phylogeny analysis of HSAs, Figure 4C showed that PM is a bigger haplotypes donor to PMY's gene pool and much more related to PMY compared to NG. The HSAs phylogeny of the NMY (Figure 4D) has confirmed the closer genetic relationship with IN compared to CN, perhaps due to long-term admixture between both populations. These HSAs phylogeny patterns were concordance with the admixture analyses using the STRUCTURE program as described above.

Discussion

Quantifying genetic distance is the main aspect in population genetic study, especially to characterize population structure and identify substructures (Lao et al. 2006; Rosenberg et al. 2002; Weir et al. 2005 and Tishkoff et al.

Table 2 Proportion of membership coefficient (*Q*) for each of population in each of the six inferred clusters (*K* = 6)

Country & Ethnicity (Sample size)	Location/State	Population ID	Clusters (*K* = 6)					
			Malays	Proto-Malays	*Semang*	Chinese	Indians	African
Malaysia:								
Malays*:								
Jawa (19)	Johor	MY-JV	0.419	0.004	0.010	0.665	0.004	0.001
Bugis (14)	Johor	MY-BG	0.255	0.027	0.044	0.561	0.008	0.001
Minang (20)	Negeri Sembilan	MY-MN	0.318	0.011	0.019	0.525	0.125	0.002
Kelantan (18)	Kelantan	MY-KN	0.222	0.018	0.054	0.542	0.162	0.002
Kedah (24)	Kedah	MY-KD	0.206	0.028	0.031	0.527	0.208	0.001
Proto-Malay[a]:								
Temuan (49)	Negeri Sembilan	MY-TM	0.101	0.361	0.053	0.478	0.006	0.001
Negritos[a]:								
Jahai (50)	Perak	MY-JH	0.012	0.010	0.808	0.168	0.002	0.000
Kensui (30)	Kedah	MY-KS	0.018	0.006	0.926	0.035	0.015	0.001
Thailand*:								
Pattani (14)	Pattani	TH-PT	0.237	0.013	0.039	0.570	0.140	0.001
Indonesia[a]:								
Jawa (19)	Java	ID-JV	0.251	0.029	0.048	0.663	0.008	0.001
Melayu (12)	Sumatera	ID-ML	0.378	0.005	0.018	0.586	0.012	0.001
Toraja (20)	Sulawesi	ID-TR	0.444	0.004	0.009	0.540	0.003	0.001
China[a]:								
Jinuo (29)	Yunnan	CN-JN	0.016	0.005	0.015	0.959	0.005	0.001
Wa (56)	Yunnan	CN-WA	0.032	0.005	0.021	0.936	0.005	0.001
India[a]:								
Marathi (14)	Maharashtra	IN-WL	0.002	0.003	0.006	0.006	0.979	0.004
Telugu (24)	Andra Pradesh	IN-DR	0.003	0.003	0.011	0.006	0.975	0.002
Africa[b]:								
Yoruba (60)	Nigeria	YRI	0.001	0.001	0.003	0.001	0.002	0.992

[a]The genotype data obtained from the database of PASNPI Consortium (http://www4a.biotec.or.th/PASNP/).
[b]The genotype data obtained from International HapMap Consortium (http://hapmap.ncbi.nlm.nih.gov/).
*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.

2009). Previous studies have also shown the importance of genetic distance assessments, such as in inference of migration patterns (Li et al. 2008; Ramachandran et al. 2005; Deshpande et al. 2009 and Laval et al. 2010) as well as in the need to adjust for population stratification in association studies Bryca et al. 2010; Patterson et al. 2010; Miclaus et al. 2009 and Li et al. 2010). This study was conducted to infer the population structure of Malays that may or may not have shared ancestry from other study populations. Therefore, the genetic distance based on Fst was chosen to measure the genetic variation distributions of studied populations. Fst is a powerful method to show population genetic structure by partitioning genetic variance within populations relative to between populations (Weir and Hill 2002; Weir and Cockerham 1984).

Many studies have shown that genetic distance of global populations correlates with geographic distance between populations, which refers to a situation called Isolation-by-Distance (IBD) (Li et al. 2008; Tishkoff et al. 2009; Ramachandran et al. 2005; Prugnolle et al. 2005 and Gonder et al. 2007). Nevertheless, based on the genetic distance analysis of this study, the IBD model can be applied to a particular group of populations, but not to the Malays group. The close genetic relationship between *Melayu Kedah* and *Melayu Kelantan* were mostly reflected to their geographic origin at the northern part of the peninsula, likewise Thai *Pattani* have smaller value of genetic distance to both of the northern Malays (Table 1). However, in the case of *Melayu Bugis* and *Melayu Jawa*, in which both population samples have

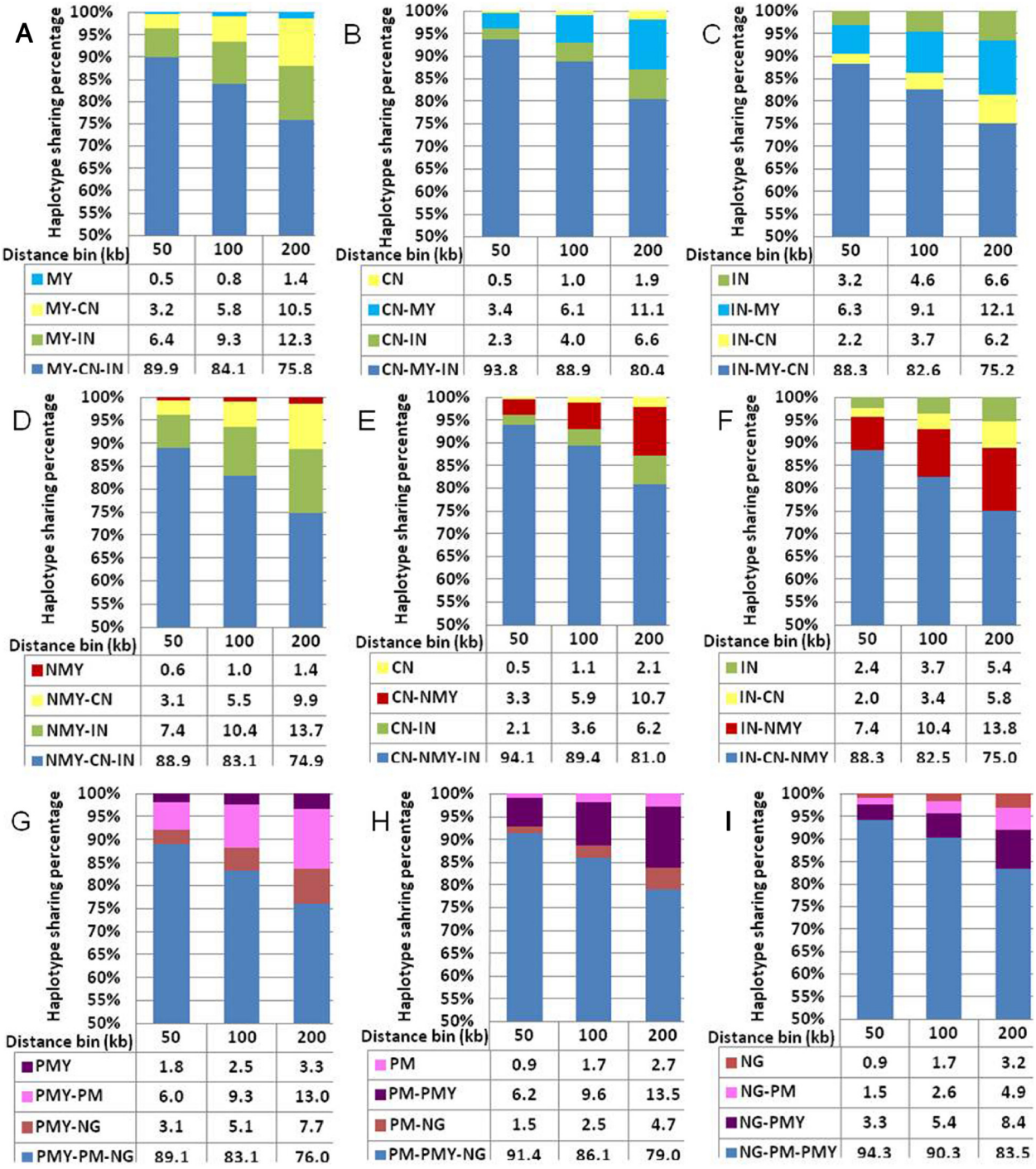

Figure 3 Haplotype Sharing (HS) percentage of the Malays (MY), northern Malays (NMY), peninsula Malays (PMY), Chinese (CN), Indians (IN), Proto-Malays (PM) and *Semang* (NG). Haplotypes in population A were identified by HSA as four classes: 1) private in population A; 2) shared with population B only; 3) shared with population C only; and 4) shared with all the three populations. **A)** MY; **B)** CN; **C)** IN; **D)** NMY; **E)** CN; **F)** IN; **G)** PMY; **H)** PM and **I)** NG. HS proportions were obtained by sampling 76 chromosomes 100 times without replacement and calculated without considering the frequencies of distinct haplotypes.

been collected in the same state at the southern part of the peninsula, the result was in contrast with IBD. As shown in Table 1, the *Melayu Bugis* were genetically distant from other Malays but closely related to Indonesian *Toraja* from South Sulawesi, while the *Melayu Jawa* were significantly closer to Indonesian *Jawa* from Central Java. In this

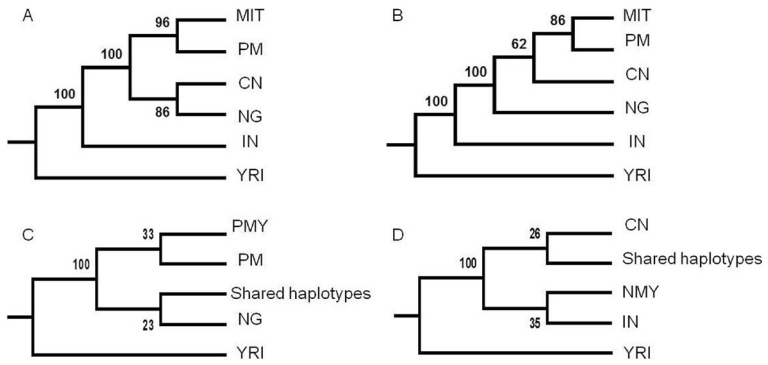

Figure 4 Phylogenetic trees of STRUCTURE analyses and Haplotype Sharing Analyses (HSAs). The phylogenetic trees re-constructed based on two types of genetic distance methods, which are **A)** Cavalli DC and **B)** Nei's DA. The clade of MIT consists of Malays, Indonesians and Thais, PM is Proto-Malays, CN is Chinese, NG is *Semang*, IN is Indians and YRI is *Yoruba*. The phylogenetic trees based on haplotype sharing distance from 100 kb bins of HSAs were showed by; **C)** PMY is private haplotypes found only in Malays samples of peninsula; PM is private haplotypes found only in Proto-Malays samples; NG is private haplotypes found only in *Semang* samples; Shared haplotypes is found in all PMY, PM and NG samples. **D)** NMY is private haplotypes found only in northern Malay samples; CN is private haplotypes found only in Chinese samples; IN is private haplotypes found only in Indians samples; Shared haplotypes is found in all NMY, CN and IN samples; YRI is the African haplotypes that were used as outgroup.

situation, the genetic distance corresponds largely to shared ancestry between populations, as been discussed in many previous studies (Hatin et al. 2011; Rosenberg et al. 2002; Jorde and Wooding 2004 and Consortium THP-AS 2009).

According to the history of Malay Peninsula and Indonesia, the migration of Indonesian people into mainland peninsula were very common whether due to trading activities or seeking asylum from civil war and colonial invasion (Taylor 2003; Sainuddin 2003 and Munoz 2006). For instance, the modern *Melayu Jawa* in peninsula are descendants of *Jawa* people that originated from Java Island. They migrated to the state of Johor and Selangor around the 15th century to avoid conflicts due to civil war. The large scale migration of *Jawa* into peninsula was during British colonial era in between of 1880 to 1930 for seeking a better life from harsh invasion of the Dutch. The first influx of *Melayu Bugis* into Peninsular Malaysia from their origin in South Sulawesi was in the 17th century. At that time, the Dutch were expanding their trade and political control on Sulawesi. Arising conflicts between the colonial and colonized people were inevitable and most of the *Bugis* people migrated to Johor state at the southern part of the peninsula. Later on, they also settled in Selangor state for over a few centuries. Since then, the *Melayu Bugis* and Indonesian *Toraja* which originated from the same geographic origin were geographically separated for hundreds of years.

It is also known from the history and anthropological evidence, the *Melayu Minang* were originally *Minang kabau* people from West Sumatra. They migrated to Malay Peninsula in the 14th century, long before the arrival of *Bugis* and *Jawa* people. They started the colonization of Negeri Sembilan, a state in the middle of the western part of Peninsular Malaysia and at present day are known as *Melayu Minang*. However, in relation to the other Malays, the *Melayu Minang* were genetically closer to both of *Melayu Kedah* and *Melayu Kelantan*, than to *Melayu Jawa* (Table 1). Here, the genetic distance between those Malay sub-ethnic groups may have been affected by the presence of genetic admixture in their genetic data from the same mixing populations. In quantifying genetic distance among populations that experienced admixture from the same mixing populations, interpreting the relationships can be tricky as the admixture reduces the average of genetic distance between them (Handley et al. 2007; Halder et al. 2008; Auton et al. 2009 and Moorjani et al. 2011).

It is interesting to see the close genetic relatedness among Malays (except for *Melayu Bugis*) with the Proto-Malay *Temuan*, Indonesian *Jawa* and Chinese *Wa* from Yunnan, China. This may imply a common origin for all of those populations regardless of their historical migration patterns as been reported (Bellwood 1997 and Omar 2004). But it is more intriguing to see how genetically close the *Melayu Kedah* and *Melayu Kelantan* are to the Indian *Telugu* and *Marathi* compared to any other populations in this study. This result has conformed to the historical contact between those populations with the Indians (Arasaratnam 1970).

Apart from being close to Proto-Malay *Temuan*, the *Melayu Kedah* and *Melayu Kelantan* also showed relatedness to the earliest aboriginal people of peninsula in view of having the smallest value of genetic distance to both of the *Semang* sub-groups, *Jahai* and *Kensiu* compared with any other Malays. The Indonesian *Jawa* and Chinese *Wa* also shared the same lower value to both of the *Semang*, compared to other group of studied populations. This may imply two probable situations that could explained the genetic relatedness of those populations.

Firstly, a common origin for all of those populations with subsequent evolutionary divergence due to pre-historical migration and become closer with subsequent recent migration as been implied by the HUGO-PASNP Consortium (Consortium THP-AS 2009). Secondly, they might have diverse origin from simultaneously pre-historical evolutionary divergence (Cavalli-Sforza et al. 1994) but admixed due to recent migration patterns as been explained by (Andaya 2001). To be certain about which are the most favorable demographic histories among them, further analysis was conducted in model-based approach using ancestral components that will be discussed subsequently.

Meanwhile, the MDS has been widely used for the analysis of proximity data among objects to reveal the hidden structure underlying the data (Steyvers 2002; Borg and Groenen 2005) especially in DNA microarray data (Tzeng et al. 2008). In this study, the genetic structure of Malays showed by Fst was successfully recapitulated in two and three dimensions (2D and 3D) models as shown in Figure 2. The Malay populations are shown explicitly as three sub-clusters on both of the 2D and 3D platforms, signifying an existence of substructure within the Malays. This could be achieved by finding the disposition of studied populations that are compatible with the given genetic distances among them on a map. The Euclidean distance is used to represent the transformed data in such a way that the MDS clustering matches the original data as much as possible even in a smaller number of dimensions (Borg and Groenen 2005). However, the distance-based method could not provide the ancestral membership coefficients of the admixture among the populations. Hence, we implemented further analysis to determine the genetic admixture and ancestry of the Malays.

Previous studies have shown the robustness of the STRUCTURE software in inferring the population structure and ancestry in variety of population data (Xu et al. 2010; Consortium THP-AS 2009; Bamshad et al. 2003; Rosenberg et al. 2005 and Witherspoon et al. 2007). In this study, the assignment of individuals into inferred clusters was in accordance to historical and demographical background of studied populations. At $K = 2$, all individuals of the 17 pre-defined populations was predominantly separated into African and non-African ancestries, indicating that the modern human dispersal originated from Africa (Cavalli-Sforza et al. 1994; Bowcock et al. 1994). The emergence of cluster that is specific to *Orang Asli Semang* of Malay Peninsula as early as at $K = 3$, supports the arrival of the *Semang* into SEA region via the first wave from Africa as postulated by many researchers (Allen 1879; Carey 1976; Fix 1995; Consortium THP-AS 2009; Kashyap et al. 2003 and Hill et al. 2006). The fact that their cluster occurred before the Indian cluster at $K = 4$, may be caused by the great effect of genetic drifts in their genetic data due to population bottleneck event and later exhibit the founder effects. As they are extremely geographically isolated and conserved from the outside world, they have preserved their ancestral allele state since the divergence. This is unlike the Indians, who have widespread admixtures with Europeans (Brahmachari et al. 2008). Histories of language shifting are also common in some of the Indian populations, as shown by the Indian *Marathi* in this analysis.

It is also noted that the ancestral component of *Semang* could be seen in Indians, Proto-Malays, Thai *Pattani* and both *Melayu Kedah* and *Melayu Kelantan*. However, based on the Q proportions of $K = 6$ as the ancestral cluster, the component was not significantly high with merely 0.05 in *Melayu Kelantan* and Proto-Malays, whilst much lower in the other Malays. Interestingly, the admixture coefficients of the Indian ancestral component exist in both of the *Melayu Kedah* and *Melayu Kelantan*, as well as in *Melayu Minang* and Thai *Pattani*. The highest coefficients were in *Melayu Kedah* with 0.21, followed by *Melayu Kelantan* 0.16, Thai *Pattani* 0.14 and *Melayu Minang* 0.12.

In population HSAs, the Indians have higher percentage of population private haplotypes than the Chinese and the Malays. This pattern is more compatible with the scenario of population admixture in Indians. However, high level of haplotype diversity is not just expected in an admixture population with divergent ancestries, but also in an ancient population as it has long time to accumulate much more private haplotypes. The HS pattern of Indians with the Chinese and Malays was relatively lower than the HS of those two groups with each other. The reason that could cause to these HS pattern perhaps due to the main separation between Indians and Asian populations dates to about 60,000 years ago (Cavalli-Sforza and Feldman 2003) while the populations in Southeast Asia and East Asia (China) have very close connections in the more recent past, either due to Neolithic expansions from China into mainland Southeast Asia and Island Southeast Asia or somewhat earlier migrations in the late Pleistocene or Early Holocene due to climate change and sea-level changes (Ricaut et al. 2006; Bellwood and Dizon 2008 and Hung 2008).

The STRUCTURE results exhibited very close estimates with the HSAs results, suggesting major contribution of Indian haplotypes in the northern Malays. The centralization of the ancient Indianized kingdoms had occurred in mainland Southeast Asia such as Thailand, Cambodia, and Myanmar for centuries in the early millennium (Tarling 1999; Stark 2006). Although Hinduism also existed in some of the Indonesian islands (eg. Sumatra and Java Island), it was more predominant among the populations in mainland region and the northern part of Malay Peninsula (Shuhaimi

1984; Allen 1997 and Syukri 2002). The existence of Indian ancestral component within these northern Malay populations is relevant to their early historical contacts with the Indians. The long-term historical contacts between Malays and Indians, may explain the higher admixture coefficients in both *Melayu Kedah* and *Melayu Kelantan*.

Furthermore, the existence of the Indians haplotypes in the gene pool of the northern Malays may signify that they are the oldest Malay populations in Malay Peninsula as the Indians had been conspicuous in the region very much earlier, since the proto-historic times. The ancient Hindu Malay kingdoms which arose approximately in 100 before common era (BCE) to 7th century CE such as *Chi Tu, Langkasuka* and *Kadaram* have controlled much of the northern Malay Peninsula (Arasaratnam 1970). The Indian influxes continued to expand during the subsequent empires of *Srivijaya* and *Majapahit* (Paul 1961). These early Malay states were heavily influenced by concepts of religion, government and arts that were brought by the Indians. The proto-historic of Malay Peninsula ended in the beginning of 15th century CE with the emergence of Malacca Sultanate. Malacca that encompassed most of modern day Peninsular Malaysia, Singapore and a great portion of eastern Sumatra thrived into the most important entrepôts in Southeast Asia and a hub of Islamic studies, spreading Islam to Malay Archipelago in 16th century CE. Still, traces of the Indian influence can be found in Malay culture until today (Arasaratnam 1970; Shuhaimi 1984 and Syukri 2002).

Possible admixture between Malays and Indians could also have occurred during the British colonial period from the 19th to the middle of the 20th century. However, the Indians are not a large component of the Kedah and Kelantan population either during or after the British colonial era as most of them reside 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 the Kelantan state which is the origin of the *Melayu Kelantan*, the total population is about 1.67 million and the percentage of the Indian community is only 0.2% of the population. In the Kedah state which is the origin of the *Melayu Kedah*, the total population is about 2.04 million and the percentage of the Indian community is also slightly higher than in Kelantan with 6.6% of the population (Hunley et al. 2009). Moreover, 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. Hence, we believed that the admixture in both *Melayu Kedah* and *Melayu Kelantan* with Indians was ancient and has occurred during the early existence of the Malays.

The ancestral component of Proto-Malay *Temuan* appeared at $K = 5$, while the ancestral component of Malays emerged at $K = 6$ and both of the components also existed in the Chinese individuals, especially in Chinese *Wa* at respective clusters. In those inferred clusters, the ancestral component of Chinese (yellow component) was predominant in all the Malays, Thais, and Indonesians as well as in Proto-Malays itself. This is in accordance to the historical and anthropological evidences of the migration pattern of Proto-Malays from Yunnan, southern mainland China (Bellwood 1997; Carey 1976). The yellow component also might be related with a large Neolithic input from China into mainland and island Southeast Asia due to the expansion of agriculture and animal domestication (Bellwood and Oxenham 2008; Bellwood 2011). Although the Austronesian dispersal did not originated from the early farming dispersal, but it was a peripheral result of the demographic impetus and technological advancement by the developments of food production in mainland East Asia (Bellwood 2011). Furthermore, the peopling of pacific region started from earlier migration of modern human expansion from Africa throughout much of Southeast Asia during a period of relatively stable climate and sea-level from 45,000 year before present (YBP) to 40,000 YBP. The extreme climate and rapidly changes of sea-level during the Last Glacial Maximum lead to decrease the expansion of human populations from 33,000 YBP – 16,000 YBP (Bird et al. 2004; Forster 2004). Later, the post-glacial expansion in coastal settlement arose concurrently with the development of coastal ecosystems and environments due to the slow rise of the sea-level. However, the sea-level fluctuations inhibited the coastal settlement and the drop in sea level in the mid-Holocene may have caused widespread human expansion throughout Oceania (Bird et al. 2004; Pope and Terrell 2008). The yellow component suggests that despite having great admixture or genetic differences due to genetic drift, all of these populations have a common ancestor, which is referred to as Southern Mongoloid group of races.

In relation to the modern Malays, it is known that Malays have been previously referred to as admixed Deutero-Malays, which are the descendants of the Proto-Malays who had admixture with other populations, such as Arab, Sumatran and Siamese (Sainuddin 2003). Other sources have postulated that the Deutero-Malays originally migrated through the southern part of China, and reached the Malay Peninsula about 1500 to 2000 years ago, after the arrival of the Proto-Malays (Fix 1995). According to Kasimin (Kasimin 1991), compared with the arrival of Indians and Chinese to Malay Peninsula, the Deutero-Malays were the earliest to be settled. Then, the vast and subsequent influxes of other populations to peninsula, mainly due to trading activities had integrated the Deutero-Malays into admixtures. These Deutero-Malays are known as the present day Malays. Given the vague historical facts on the origin of Malays with a fine line to differentiate

between Proto-Malays and Deutero-Malays, it is still evident that these populations have profoundly close genetic relationship and shared a common ancestor with the Chinese.

The ancestral clusters or the most biologically sensible number of clusters which captured the structure of the given genotyped data was identified as six clusters (Table 2). In this ancestral clusters (K = 6), any pre-defined populations which shared similar distinctive *Q* values were merged into the inferred cluster. These clusters associate mainly to self-identified geographic origin or geographic proximity, linguistics family and their ancestry. The first cluster is referred as the cluster of Austronesian speakers and includes all the Malays, Thais and Indonesians. The second cluster was exclusive to Proto-Malays ancestry, which is *Temuan* in this study. The fact that these Proto-Malays are different from the other Austronesian speakers although belonging to the same linguistic family, perhaps due to evolution forces such as genetic drift or selection pressures that have reshuffled their genetic components. Yet, both these clusters still have conserved great proportions of Chinese ancestral component, testifying their Southern Mongoloid morphological features.

The third cluster belongs to the Austro-Asiatic speakers, which are the *Semang Jahai* and *Kensiu*. The fourth cluster associates with Chinese ancestry and belongs to the Chinese *Jinuo* and *Wa* from Yunnan, China. However, these Chinese populations speak different languages; *Wa* speak Austro-Asiatic whilst *Jinuo* speak Sino-Tibetan language. These two populations are the indigenous populations in Yunnan and as shown by the HUGO-PASNP Consortium (Consortium THP-AS 2009), both populations were clustered among indigenous Thais populations who also speak Austro-Asiatic language. This indicates a history of language shifting in the Chinese *Jinuo* (Dutton and Tryon 1994). Although sharing the same linguistic family, a great difference between Chinese *Wa* and the *Semang* group in Malay Peninsula is likely due to their historical divergence with wide range of demographic histories (Consortium THP-AS 2009). The fifth cluster belongs to Indian ancestry, consist of *Marathi* and *Telugu*. Again, despite having different linguistic family, both the Indians were clustered in the same ancestral cluster. Lastly,the sixth cluster consisted of only *Yoruba* with African ancestry who speak Niger-Congo language.

The population genetic structure of Malays and other studied populations became more apparent at higher number of clusters, but it also might not representative of anything and just residuals of the methodology. Thus, the interpretation must be made carefully as well as must be supported by other evidence, either historical or anthropological. In this analysis, it is clear that whole-genome data of *Melayu Bugis* were more delineated to Indonesian *Melayu* and *Toraja* from Indonesia with very minimal admixture. The *Melayu Jawa* were similar to the Indonesian *Jawa*, with significant components of Chinese in their genomes. The *Melayu Kelantan* and *Melayu Kedah* resemble more to Thai *Pattani*, with significant admixture from Indian components. Although the *Melayu Minang* also shared the Indian ancestral component, they were slightly different at this higher level of clusters as they exhibit almost none of the Chinese ancestral component. This may be due to different demographic histories compared to other Malays as has been reflected by their unique maternally cultural and traditional rule, called "*Adat Pepatih*" (Reid 2001; Ricklefs 2001). This unique sociological cultural may have had greater effect of pressure selections in the *Melayu Minang* population. The result of admixture analysis was concordance with the previous study of human leukocyte antigen (HLA) polymorphism and population structure analysis of Malays (Hatin et al. 2011; Edinur et al. 2009).

The resulting *Q* plot from STRUCTURE and HSAs partition plots did not revealed the relationship among the components in term of evolutionary history. Thus, phylogenetic trees were reconstructed to further refine the analysis. The topology of tree produced by Cavalli DC contradicted most known patterns of historical migrations. Based on historical facts, it is unlikely that Malays and Proto-Malays have a simultaneously historical divergence with the *Semang* group. Many evidences have shown that the original inhabitants of the Malay Peninsula are the *Semang*. For instance, the oldest Paleolithic human skeleton estimated about 11,000 years old, was reported to have genetic similarities with the *Semang* (Majid 2005). The other topology, which was produced by Nei's DA is more favored to reflect the evolution histories among the study populations. Previous genetic studies have postulated the northwards migration of SEA people to central and eastern Asia (Consortium THP-AS 2009; Su et al. 1999) before gradually migrated back southwards (Rahman et al. 1998; Andaya 2001; Omar 2004; Bellwood 1997 and Fix 1995). The phylogeny analyses of HSAs indicated greater similarity where the Malays are more related to Proto-Malays than to *Semang*. Although the ancestry line of Malays were traced back to the Proto-Malays and the Chinese, the Indians have contributed more haplotypes to the northern Malays that may resulted to the *Melayu Kedah* and *Melayu Kelantan* to be genetically different from the other Malays.

Conclusions

The genetic clustering by model-based approach has successfully showed the admixture and ancestral coefficients within the studied populations that is in line to the historical backgrounds, which cannot be achieved by the distance-based method. The need to characterize the genetic make-up of this admixture proportions, especially in

genetic-based medical studies is very important as it clearly affect the gene pool of a particular population. Thus, this study suggests that a larger scale research of targeted admixed populations on other ethnic groups in Malaysia should be conducted in the near future.

Methods

Population samples and genotype data

All genotype data were generated from DNA samples that were collected with informed and written consent and approved by local ethics committees in Malaysia (Research and Ethics (Human) Committee, School of Medical Sciences, Universiti Sains Malaysia (USM) and Medical Ethics Committee, Pusat Perubatan Universiti Malaya (PPUM)), China (Ethical Committee, Chinese National Human Genome Centre (CNHGC) at Shanghai, PR China), Indonesia (Research Ethics Commission, Eijkman Institute for Molecular Biology, Indonesia), India (Human Ethics Committee, Institute of Genomics and Integrative Biology (IGIB)) and Thailand (Ethics Committee Faculty of Medicine, Prince of Songhkla University, Thailand.

In this study, samples were carefully selected by the inclusion and exclusion criteria that emphasized the three generations without any different ethnic admixture. All datasets used in the study were derived from 17 populations representing six linguistic families that consisted of 472 unrelated individuals from five Malay sub-ethnic groups (*Melayu Bugis*, *Melayu Jawa*, *Melayu Minang*, *Melayu Kedah*, and *Melayu Kelantan*); three *Orang Asli* sub-groups (*Jahai*, *Kensui* and *Temuan*); one Thai population (*Pattani*), three Indonesian populations (*Melayu*, *Jawa* and *Toraja*); two from Yunnan, China (*Jinuo* and *Wa*); two from India that were assigned based on their language (*Telugu* and *Marathi*); and one from Nigeria, Africa (*Yoruba*). The map of the Asian continent depicting geographic locations of the sampling populations is shown in Figure 5. All samples were assigned anonymously and code identified at analysis and data point as shown in Table 2.

The Affymetrix GeneChip Mapping Xba 50 K Arrays were used to genotype single nucleotide polymorphisms (SNPs) on a genome-wide scale for 109 unrelated individuals of the five Peninsular Malaysia Malay sub-ethnic groups and one Thai *Pattani*. Meanwhile, the additional genotype data of 363 unrelated individuals from other 11 populations were obtained from the database of the Pan Asian SNP Initiative (PASNPI) Consortium, except for *Yoruba* that were obtained from the International HapMap Consortium.

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. A total of 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.

Genetic differentiation between populations

The genetic divergence between studied populations were determined by Fixation Index Statistic (Fst) as described by Weir and Hill (Weir and Hill 2002). Package for Elementary Analysis of SNP data v1.0 (PEAS) (Xu

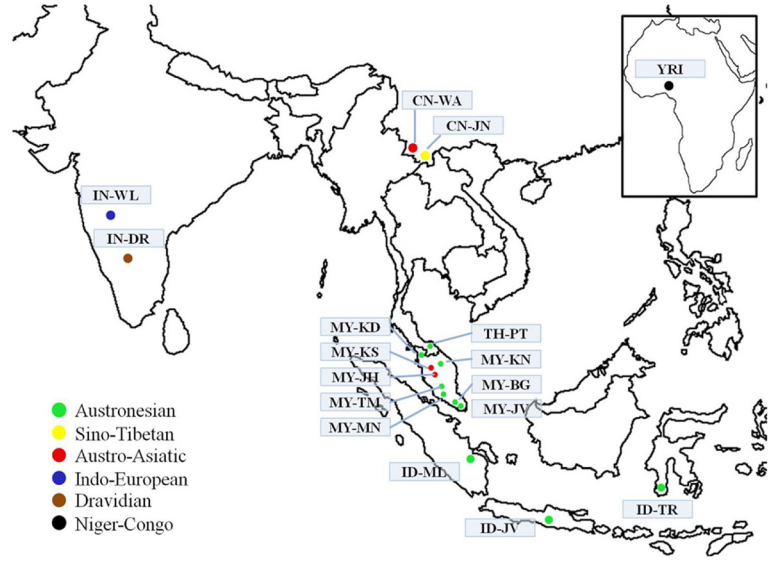

Figure 5 Map of the Asian continent depicting geographic locations of the sampled populations in six countries. The small box on the upper right corner of the figure shows African continent. The colors denoted the linguistic family of the populations.

et al. 2010) was used to calculate allele frequency and genetic distance between populations. The genetic structures of population were assessed by a multivariate statistical technique such as MDS analysis using SPSS 18. The input data was the genetic distance matrix of pair-wise Fst. In this analysis, the convergence of the S-Stress value was set to 0.001 and the iterations were set to a maximum of 30. The number of dimensions employed was two dimensions and then increased to three dimensions.

Admixture analyses by STRUCTURE

We used STRUCTURE v2.3.3 (Pritchard et al. 2000), a model-based clustering software which implements the Markov Chain Monte Carlo (MCMC) algorithm within a Bayesian framework to estimate the genetic structure and distribution of ancestral component for each individual of studied populations. The admixture model and correlated allele frequencies between populations in the parameter setting of STRUCTURE analysis are powerful to detect subtle population genetic structures, such as in highly admixed populations (Falush et al. 2003). It assigns individuals into pre-specified clusters (*K*) with estimated membership coefficient (*Q)* for each cluster which were fitted with posterior probabilities of *Pr (X|K)* solely based on the given genotyped data *(X)* without incorporating any other population information (Pritchard et al. 2000; Falush et al. 2003). The value of *K* that maximized the value of *Pr (X|K)* which showed by the graph of *Ln(Pr)* over the run of analysis is the most probable number of ancestral clusters (Pritchard et al. 2003).

We created sub-datasets from the full SNPs dataset using between marker distance (BMD) as the threshold value in the re-sampling procedures and a total of five sampling datasets (S1-S5) have been produced. The average of the BMD is 550 kb and each dataset contain approximately 3700 number of SNPs that were evenly distributed across 22 autosomal chromosomes. The sub-datasets were created because of STRUCTURE's limitation which does not deal with strong background linkage disequilibrium in the data (Falush et al. 2003). The used of sub-datasets also worth to cut the analysis time due to the computational intensity of the STRUCTURE analysis that is time consuming (Falush et al. 2003).

We ran a series of analysis in STRUCTURE from $K = 2$ to $K = 10$ and number of iterations were set to 10 times for each *K*s and each datasets in order to verify the consistency of the results. Hence, we have submitted a total of 450 running analyses (5 sub-datasets × 9 inferred cluster x 10 iterations = 450) with 30,000 burn-in length and 20,000 MCMC iterations in each analysis to STRUCTURE software. The distribution of the alpha parameter showed a relatively constant distribution indicating convergence after 20,000 iterations. The estimated *Q* matrices from the STRUCTURE outputs were carefully observed and compared using the symmetric similarity coefficients (SSC) and there were no big differences in the estimation of *Q*s for all runs. The SSC was computed via permutation analyses of *Q* matrices for any number of clusters from multiple runs or multiple datasets generated by STRUCTURE software. The analyses were implemented by Cluster Matching and Permutation Program (CLUMPP) (Jakobsson and Rosenberg 2007). A program called *distruct* (Rosenberg 2004) was used to provide much finer control of the graphic plot of *Q*. It displays each individual as a line segment that partitioned into *K* colored components, which represent the individual's *Q* in the *K* clusters.

Haplotype-sharing analyses (HSAs)

The fast PHASE v1.2 (Scheet and Stephens 2006) was used to estimate haplotypes for each individual from 54,794 SNPs data. The number of random starts of the EM algorithm (−T) was set to 20, and the number of iterations of EM algorithm (−C) was set to 50. The software provides an estimation of the true underlying patterns of haplotype structure and to enhance the accuracy of the analysis, population labels were applied during the model fitting procedure (Scheet and Stephens 2006). The percentages of haplotype sharing (HS) among populations were determined based on (Xu et al. 2009) by HaploSharing program of the PEAS v1.0. The analysis binned the inferred haplotypes within particular size of windows and let a window slide by half of the other window size each time, considering the substantial variation of recombination across human genome (The International HapMap Consortium 2007; Li et al. 2008).

In this study, we adopted three sizes of sliding window (50 kb, 100 kb and 200 kb) to estimate the HS in three-population framework. According to (Xu et al. 2009), if there were three populations, A, B, and C, the haplotypes of one population can be identified as four categories when compared with those of the other two populations, regardless of the haplotype frequency. For instance, the haplotypes of population A are classified into four haplotype categories: 1) haplotypes are private in population A (denoted by H_{AP}), 2) haplotypes are common in populations A and B but not in population C (H_{AB}), 3) haplotypes are in common in populations A and C but not in population B (H_{AC}), and 4) haplotypes are common in all populations (H_{ABC}). The haplotypes for populations B and C can be similarly defined.

In the HSAs of this study, the particular populations were merged into a group pursuant to the result of STRUCTURE analysis. For the first HSA, all Peninsular Malaysia Malays, Thai *Pattani* and Indonesians were merged into a group named Malays (MY), the Chinese (CN) group consisted of *Jinuo* and *Wa*, while the group of Indians (IN) comprised of *Marathi* and *Telugu*. This

analysis was done to indicate the proportion of HS among the three groups and to identify which group was a bigger genetic contributor to Malays group. To represent the divergence pattern of the northern Peninsular Malaysia Malays, we combined the *Melayu Kedah*, *Melayu Kelantan* and Thai *Pattani* into a group named northern Malays (NMY) and implemented the second HSA with the Chinese and Indians groups. The third HSA was done to examine the relationship of the peninsula Malays (PMY) group consisted of Peninsular Malaysia Malays and Thai *Pattani* with the indigenous populations, which are the *Orang Asli Semang* (*Jahai* and *Kensiu*) grouped as Negrito (NG) and the Proto-Malays *Temuan* labeled as Proto-Malays (PM).

However, the varying sample size among populations could affect the HS results. Taking this point into consideration, we performed a procedure called Non-Replace Sampling (NRS) in the HaploSharing program. The number of haplotypes in each genomic window on 76 chromosomes (38 individuals is the minimal sample size in this study) were counted for each population. The bootstrap replicate for the sampling procedure was 100 times and the results were averaged for each window. Any haplotype that observed less than twice in this analysis was excluded.

Phylogeny analyses

The phylogenetic trees of HSAs were based on four-population framework of HS estimation as African *Yoruba* (YRI) samples were added to serve as an outgroup for the rooted trees. The haplotype sharing distances among the group of populations from 100 kb bins were calculated based on Fst (Weir and Hill 2002). The distance based population trees were reconstructed using the Neighbor Joining algorithm (Saitou and Nei 1987) by Molecular Evolutionary Genetics Analysis 4 (MEGA) (Tamura et al. 2007) and two programs from Phylogenetic Inference Package 3.67 (Phylip) (Felsenstein 2007) which are Neighbor and Consense. The phylogenetic trees of STRUCTURE analyses based on Bayesian algorithm were reconstructed using the estimated allele frequencies of the inferred clusters. The PEAS program called *ClusterDis* was used to calculate two types of genetic distances among the inferred clusters, named Cavalli-Sforza Chord Distance (Cavalli DC) (Cavalli-Sforza and Edwards 1967) and the Nei's Matrix Distance (Nei's DA) (Nei et al. 1983). Bootstrapping test was performed for 1000 times and the phylogenetic trees were rooted using YRI, assuming that the exit point of human diaspora in Africa was correct.

Abbreviations

MDS: Multi-dimensional scale; PASNPI: Pan Asian SNP Initiative; MIT: Malays, Indonesians and Thais; IBD: Isolation-by-distance; BCE: Before common Era; YBP: Year before present; HLA: Human Leukocyte Antigen; USM: Universiti Sains Malaysia; PPUM: Pusat Perubatan Universiti Malaya; CNHGC: Chinese National Human Genome Centre; SNPs: Single nucleotide polymorphisms; CLUMPP: Cluster matching and permutation program; SSC: Symmetric similarity coefficients; BMD: Between marker distance; MCMC: Markov Chain Monte Carlo; UKM: Universiti Kebangsaan Malaysia; UMBI: Medical Molecular Biology Institute; 2D: Two dimensions; 3D: Three dimensions; HS: Haplotype sharing; MY: Malays; NMY: Northern Malays; PMY: Peninsula Malays; CN: Chinese; IN: Indians; PM: Proto-Malays; YRI: African yoruba; NG: Semang; HSAs: Haplotype sharing analyses; NRS: Non-Replace Sampling; NG: Negrito.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HWI contributed in conception of study design, collection of data, drafted the manuscript and analysis and interpretation of data. NSAR contributed in conception of study design, participated in the collection of data, carried out the lab works, drafted the manuscript and analysis of data. AE drafted the manuscript. JWF contributed in analysis and interpretation of data. QPF contributed in analysis and interpretation of data. XS contributed in analysis and interpretation of data and drafting the manuscript. JL drafted the manuscript. TSG drafted the manuscript. LP drafted the manuscript. FMA contributed in conception of study design, drafting manuscript, interpretation of data and approval of manuscript. IMR contributed conception of study design, drafting manuscript and interpretation of data. ZBA contributed in conception of study design, drafting manuscript, analysis and interpretation of data, collection of data and approval of manuscript. All authors read and approved the final manuscript.

Authors' information

The participants of the HUGO Pan-Asian SNP Consortium are arranged alphabetically by surname: Mahmood Ameen Abdulla,[1] Ikhlak Ahmed,[2] Anunchai Assawamakin,[3,4] Jong Bhak,[5] Samir K. Brahmachari,[2] Gayvelline C. Calacal,[6] Amit Chaurasia,[2] Chien-Hsiun Chen,[7] Jieming Chen,[8] Yuan-Tsong Chen,[7] Jiayou Chu,[9] Eva Maria C. Cutiongco-de la Paz,[10] Maria Corazon A. De Ungria,[6] Frederick C. Delfin,[6] Juli Edo,[1] Suthat Fuchareon,[3] Ho Ghang,[5] Takashi Gojobori,[11,12] Junsong Han,[13] Sheng-Feng Ho,[7] Boon Peng Hoh,[14] Wei Huang,[15] Hidetoshi Inoko,[16] Pankaj Jha,[2] Timothy A. Jinam,[1] Li Jin,[17,38] Jongsun Jung,[18] Daoroong Kangwanpong,[19] Jatupol Kampuansai,[19] Giulia C. Kennedy,[20,21] Preeti Khurana,[22] Hyung-Lae Kim,[18] Kwangjoong Kim,[18] Sangsoo Kim,[23] Woo-Yeon Kim,[5] Kuchan Kimm,[24] Ryosuke Kimura,[25] Tomohiro Koike,[11] Supasak Kulawonganunchai,[4] Vikrant Kumar,[8] Poh San Lai,[26,27] Jong-Young Lee,[18] Sunghoon Lee,[5] Edison T. Liu,[8] Partha P. Majumder,[28] Kiran Kumar Mandapati,[22] Sangkot Marzuki,[29] Wayne Mitchell,[30,31] Mitali Mukerji,[2] Kenji Naritomi,[32] Chumpol Ngamphiw,[4] Norio Niikawa,[40] Nao Nishida,[25] Bermseok Oh,[18] Sangho Oh,[5] Jun Ohashi,[25] Akira Oka,[16] Rick Ong,[8] Carmencita D. Padilla,[10] Prasit Palittapongarnpim,[33] Henry B. Perdigon,[6] Maude Elvira Phipps,[1,34] Eileen Png,[8] Yoshiyuki Sakaki,[35] Jazelyn M. Salvador,[6] Yuliana Sandraling,[29] Vinod Scaria,[2] Mark Seielstad,[8] Mohd Ros Sidek,[14] Amit Sinha,[2] Metawee Srikummool,[19] Herawati Sudoyo,[29] Sumio Sugano,[37] Helena Suryadi,[29] Yoshiyuki Suzuki,[11] Kristina A. Tabbada,[6] Adrian Tan,[8] Katsushi Tokunaga,[25] Sissades Tongsima,[4] Lilian P. Villamor,[6] Eric Wang,[20,21] Ying Wang,[15] Haifeng Wang,[15] Jer-Yuarn Wu,[7] Huasheng Xiao,[13] Shuhua Xu,[38] Jin Ok Yang,[5] Yin Yao Shugart,[39] Hyang-Sook Yoo,[5] Wentao Yuan,[15] Guoping Zhao,[15] Bin Alwi Zilfalil,[14] Indian Genome Variation Consortium[21]Department of Molecular Medicine, Faculty of Medicine, and the Department of Anthropology, Faculty of Arts and Social Sciences, University of Malaya, Kuala Lumpur, 50603, Malaysia. [2]Institute of Genomics and Integrative Biology, Council for Scientific and Industrial Research, Mall Road, Delhi 110007, India. [3]Mahidol University, Salaya Campus, 25/25 M. 3, Puttamonthon 4 Road, Puttamonthon, Nakornpathom 73170, Thailand. [4]Biostatistics and Informatics Laboratory, Genome Institute, National Center for Genetic Engineering and Biotechnology, Thailand Science Park, Pathumtani 12120, Thailand. [5]Korean BioInformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology (KRIBB), 111 Gwahangno, Yuseong-gu, Deajeon 305–806, Korea. [6]DNA Analysis Laboratory, Natural Sciences Research Institute, University of the Philippines, Diliman, Quezon City 1101, Philippines. [7]Institute of Biomedical Sciences, Academia Sinica, 128 Sec 2 Academia Road Nangang, Taipei City 115, Taiwan. [8]Genome Institute of Singapore, 60 Biopolis Street 02–01, 138672, Singapore. [9]Institute of Medical Biology, Chinese Academy of Medical Science, Kunming, China. [10]Institute of Human Genetics, National Institutes of Health, University of the Philippines Manila, 625 Pedro Gil Street, Ermita Manila 1000, Philippines. [11]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka 411–8540, Japan. [12]Biomedicinal Information

Research Center, National Institute of Advanced Industrial Science and Technology, 2–42 Aomi, Koto-ku, Tokyo 135–0064, Japan. [13]National Engineering Center for Biochip at Shanghai, 151 Li Bing Road, Shanghai 201203, China. [14]Human Genome Center, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia. [15]MOST-Shanghai Laboratory of Disease and Health Genomics, Chinese National Human Genome Center Shanghai, 250 Bi Bo Road, Shanghai 201203, China. [16]Department of Molecular Life Science Division of Molecular Medical Science and Molecular Medicine, Tokai University School of Medicine, 143 Shimokasuya, Isehara-A Kanagawa-Pref A259-1193, Japan. [17]State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, 220 Handan Road, Shanghai 200433, China. [18]Korea National Institute of Health, 194, Tongil-Lo, Eunpyung-Gu, Seoul, 122–701, Korea. [19]Department of Biology, Faculty of Science, Chiang Mai University, 239 Huay Kaew Road, Chiang Mai 50202, Thailand. [20]Genomics Collaborations, Affymetrix, 3420 Central Expressway, Santa Clara, CA 95051, USA. [21]Veracyte, 7000 Shoreline Court, Suite 250, South San Francisco, CA 94080, USA. [22]The Centre for Genomic Applications (an IGIB-IMM Collaboration), 254 Ground Floor, Phase III Okhla Industrial Estate, New Delhi 110020, India. [23]Soongsil University, Sangdo-5-dong 1–1, Dongjak-gu, Seoul 156–743, Korea. [24]Eulji University College of Medicine, 143–5 Yong-du-dong Jung-gu, Dae-jeon City 301–832, Korea. [25]Department of Human Genetics, Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113–0033, Japan. [26]Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, National University Hospital, 5 Lower Kent Ridge Road, 119074, Singapore. [27]Population Genetics Lab, Defence Medical and Environmental Research Institute, DSO National Laboratories, 27 Medical Drive, 117510, Singapore. [28]Indian Statistical Institute (Kolkata) 203 Barrackpore Trunk Road, Kolkata 700108, India. [29]Eijkman Institute for Molecular Biology, Jl. Diponegoro 69, Jakarta 10430, Indonesia. [30]Informatics Experimental Therapeutic Centre, 31 Biopolis Way, 03–01 Nanos, 138669, Singapore. [31]Division of Information Sciences, School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, 639798, Singapore. [32]Department of Medical Genetics, University of the Ryukyus Faculty of Medicine, Nishihara, 207 Uehara, Okinawa 903–0215, Japan. [33]National Science and Technology Development Agency, 111 Thailand Science Park, Pathumtani 12120, Thailand. [34]Monash University (Sunway Campus), Jalan Lagoon Selatan, 46150 Bandar Sunway, Selangor, Malaysia. [35]RIKEN Genomic Sciences Center, W502, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230–0045, Japan. [36]Department of Biochemistry, University of Hong Kong, 3/F Laboratory Block, Faculty of Medicine Building, 21 Sasson Road, Pokfulam, Hong Kong. [37]Laboratory of Functional Genomics, Department of Medical Genome Sciences Graduate School of Frontier Sciences, University of Tokyo (Shirokanedai Laboratory), 4-6-1 Shirokanedai, Minato-ku, Tokyo 108–8639, Japan. [38]Chinese Academy of Sciences-Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Rd., Shanghai 200031, China. [39]Genomic Research Branch, National Institute of Mental Health, National Institutes of Health, 6001 Executive Boulevard, Bethesda, MD 20892 USA. [40]Research Institute of Personalized Health Sciences, Health Sciences University of Hokkaido, Tobetsu 061–0293, Japan

Acknowledgements

We would like to acknowledge the contributions made by other members of this study group from the School of Health Sciences and the School of Dental Sciences, Universiti Sains Malaysia. We appreciate all the subjects who have participated in this research and those who have helped us in the data collection. Special thanks to the Universiti Kebangsaan Malaysia (UKM), Medical Molecular Biology Institute (UMBI), and Matrix Analytical Sdn. Bhd Malaysia, for allowing us to use their laboratory facilities. Also, CAS-MPG Partner Institute Computational Biology, Fudan University, Shanghai, China for helping us with the data analysis; and PASNPI Consortium for allowing us to use their genotype data. This study was funded by the following research grants: **APEX Delivering Excellence 2012 (DE 2012) grant:** (1002/PPSP/910343) entitled "USM As Anchor for The Malaysian Node of Human Variome Project" and the **USM short term grant:** (304/PPSP/61311034) entitled "DNA Profiling of the Kelantan, Kedah and Pattani Malays Using SNPs Microarray" and **MOSTI (ER-BIOTEK) grant:** (304/PPSP/6150113/k105) entitled "Genome Variations and their Importance in Understanding Evolution, Migration and Health in Multi-Ethnic Population of Malaysia".

Author details

[1]Human Genome Centre, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kelantan, Malaysia. [2]Department of Pediatrics, School of Medical Sciences, Universiti Sains Malaysia, 16150 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, 200031 Shanghai, China. [4]Department of Cell and Molecular Biology, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400 Selangor, Malaysia. [5]Human Genetics Unit, Department of Pathology, Faculty of Medicine, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand. [6]Institute of Biological Sciences, 50603 Kuala Lumpur, Malaysia. [7]Centre of Research for Computational Sciences and Informatics in Biology, Bioindustry, Environment, Agriculture and Healthcare (CRYSTAL),Faculty of Science, Universiti Malaya, 50603, Kuala Lumpur, Malaysia.

Received: 4 May 2014 Accepted: 19 May 2014
Published online: 30 October 2014

References

Allen FA (1879) The original range of the Papuan and negrito races. J Anthropol Institute of Great Britain and Ireland 8:38–50

Allen J (1997) Inland Angkor, coastal Kedah: landscapes, subsistence systems, and state development in early Southeast Asia. Bull Indo-Pacific Prehist Assoc 16:79–87

Andaya LY (2001) The search for the origins of Melayu. J Southeast Asian Stud 32:315–330

Arasaratnam S (1970) Indians in Malaysia and Singapore. Published for the Institute of Race Relations, London, by Oxford University Press (Bombay, p 214, http://hdl.handle.net/10108/5999

Auton A, Bryca K, Boyko AR, Lohmueller KE, Novembre J, Reynolds A, Indap A, Wright MH, Degenhardt JD, Gutenkunst RN, King KS, Nelson MR, Bustamante CD (2009) Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res 19:795–803

Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB (2003) Human population genetic structure and inference of group membership. Am J Hum Genet 72:578–589

Bellwood P (1993) Cultural and biological differentiation in Peninsular Malaysia: the last 10,000 years. Asian Perspect 32:37–60

Bellwood PS (1997) Prehistory of the Indo-Malaysian Archipelago. University of Hawai'i Press, Honolulu, Hawaii, http://press.anu.edu.au/titles/prehistory-of-the-indo-malaysian-archipelago/

Bellwood P (2011) Holocene population history in the Pacific region as a model for worldwide food producer dispersals. Current Anthropology 52(4):363–S378

Bellwood P, Dizon E (2008) Austronesian cultural origins: out of Taiwan, via the Batanes Islands, and onwards to western Polynesia. In: Sanchez-Mazas A, Blench R, Ross MD, Peiros I, Lin M (eds) Past human migrations in East Asia: matching archaeology, linguistics and genetics. Routledge, London, pp 23–39

Bellwood P, Oxenham M (2008) The Expansions of Farming Societies and the Role of the Neolithic Demographic Transition. In: The Neolithic Demographic Transition and its Consequences. Springer, Dordrecht, pp 13–34

Bird MI, Hope G, Taylor D (2004) Populating PEP II: the dispersal of humans and agriculture through Austral-Asia and Oceania. Quat Int 118:145–163

Borg I, Groenen PJF (2005) Modern Multidimensional Scaling: Theory and Applications. Springer, New York

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

Brahmachari SK, Majumder P, Mukerji M, Habib S, Dash D, Ray K, Indian Genome Variation Consortium (2008) Genetic landscape of the people of India: a canvas for disease gene exploration. J Genet 87(1):3–20

Bryca K, Autona A, Nelsonb MR, Oksenbergc JR, Hauserc SL, Williams S, Froment A, Bodo JM, Wambebe C, Tishkoff SA, Bustamante CD (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci 107:786–791

Campbell MC, Tishkoff SA (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9:403–433

Carey I (1976) Orang Asli : The Aboriginal Tribes of Peninsular Malaysia. New York: Oxford University Press, Kuala Lumpur, http://trove.nla.gov.au/version/12940665

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Am J Hum Genet 19:233–257
Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. Nat Genet 33:266–275
Cavalli-Sforza LL, Piazza A, Menozzi P (1994) History and Geography of Human Genes. Princeton University Press, Princeton
Collin FS, Green ED, Guttmacher AE, Guyer MS (2003) A vision for the future of genomic research. Nature 422:835–847
Consortium TIHM (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861
Consortium THP-AS (2009) Mapping human genetic diversity in Asia. Science 326:1541–1545
Deshpande O, Batzoglou S, Feldman MW, Cavalli-Sforza LL (2009) A serial founder effect model for human settlement out of Africa. Proc R Soc B 276:291–300
Dutton T, Tryon DT (1994) Language contact and change in the Austronesian world. Walter de Gruyter 77
Edinur HA, Zafarina Z, Spínola H, Nurhaslindawaty AR, Panneerchelvam S, Norazmi MN (2009) HLA polymorphism in six Malay subethnic groups in Malaysia. Hum Immunol 70:518–526
Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587
Felsenstein J (2007) PHYLIP: Phylogeny Inference Package. University of Washington, Washinton, http://evolution.genetics.washington.edu/phylip/doc/main.html Accessed April 2013
Fix AG (1995) Malayan paleosociology: implications for patterns of genetic variation amongst the Orang Asli. Am Anthropol 97:313–323
Forster P (2004) Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. Philosophical Transactions of the Royal Society B: Biological Sciences 359:255–264
Gonder MK, Mortensen HM, Reed FA, Sousa AD, Tishkoff SA (2007) Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol 24:757–768
Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T (2008) A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat 29:648–658
Handley LJL, Manica A, Goudet J, Balloux F (2007) Going the distance: human population genetics in a clinal world. Trends Genet 23:432–439
Hatin WI, Nur-Shafawati AR, Zahri MK, Xu S, Jin L, Tan SG, Rizman-Idid M, Zilfalil BA, Pan-Asian HUGO, Consortium SNP (2011) Population genetic structure of Peninsular Malaysia Malay sub-ethnic groups. PLoS One 6(4):e18312
Hill C, Soares P, Mormina M, Macaulay V, Meehan W, Blackburn J, Clarke D, Raja JM, Ismail P, Bulbeck D, Oppenheimer S, Richards M (2006) Phylogeography and ethnogenesis of aboriginal Southeast Asians. Mol Biol Evol 23:2480–2491
Hung HC (2008) Migration and Cultural Interaction in Southern Coastal China, Taiwan and the Northern Philippines, 3000 BC to AD 1. Dissertation, Canberra National University
Hunley KL, Healy ME, Long JC (2009) The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: Implications for biological race. Am J Phys Anthropol 139:35–46
Hussein T, Ibrahim S, Alias R (2007) Malaysia Negara Kita, 2nd edn. MDC Co, Kuala Lumpur
Jabatan Perangkaan Malaysia (2010) Penduduk Mengikut Jantina, Kumpulan Etnik dan Umur. Percetakan dokumen kerajaan Malaysia, Malaysia. Putrajaya, www.statistics.gov.my/portal/download_Buku_Tahunan/files/BKKP/Buku_Tahunan_Perangkaan_Malaysia_2010.pdf. June 2011
Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806
Jorde LB, Wooding SP (2004) Genetic variation, classification and race. Nat Genet 36:S28–S33
Kashyap V, Sitalaximi T, Sarkar B, Trivedi R (2003) Molecular relatedness of the aboriginal groups of Andaman and Nicobar Islands with similar ethnic populations. Int J Hum Genet 3:5–11
Kasimin A (1991) Religion and Social Change Among the Indigenous People of the Malay Peninsular. Dewan Bahasa dan Pustaka, Kuala Lumpur, p 326
Lao O, Duijn KV, Kersbergen P, Knijff PD, Kayser M (2006) Proportioning whole-genome single-nucleotide–polymorphism diversity for the identification of geographic population structure and genetic ancestry. Am J Hum Genet 78:680–689
Laval G, Patin E, Barreiro LB, Quintana-Murci L (2010) Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. PLoS One 5(4):e10284
Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104
Li M, Reilly MP, Rader DJ, Wang LS (2010) Correcting population stratification in genetic association studies using a phylogenetic approach. Bioinformatics 26:798–806
Majid Z (2005) The Excavation and Analyses of the Perak Man Buried in Gua Gunung Runtuh, Lenggong, Perak. The Perak Man and Other Prehistoric Skeletons of Malaysia. Penerbit Universiti Sains Malaysia, Penang, pp 1–32
Miclaus K, Wolfinger R, Czika W (2009) SNP selection and multidimensional scaling to quantify population structure genetic. Epidemiology 33:488–496
Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, Burns E, Ostrer H, Price AL, Reich D (2011) The history of African gene flow into southern Europeans, Levantines, and Jews. PLoS Genet 7(4):e1001373
Munoz PM (2006) Early kingdoms of the Indonesian archipelago and the Malay Peninsula. Editions Didier Millet, Kuala Lumpur, p 392
Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J Mol Evol 19:153–170
Omar AH (2004) Languages and Literature. The Encyclopedia of Malaysia. http://www.edmbooks.com/Book/6993/The-Encyclopedia-of-Malaysia-Languages-and-Literature.html
Patterson N, Petersen DC, Van der Ross RE, Sudoyo H, Glashoff RH, Marzuki S, Reich D, Hayes VM (2010) Genetic structure of a unique admixed population: implications for medical research. Hum Mol Genet 19:411–419
Paul W (1961) The Golden Khersonese: Studies in the Historical Geography of the Malay Peninsula before AD 1500. University of Malaya Press, Kuala Lumpur, http://search.library.wisc.edu/catalog/ocm01101422
Pope KO, Terrell JE (2008) Environmental setting of human migrations in the circum-Pacific region. J Biogeogr 35:1–21
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
Pritchard JK, Wen W, Falush D (2003) Documentation for STRUCTURE software: Version 2. Available from http://pritch.bsd.uchicago.edu/structure.html
Prugnolle F, Manica A, Balloux F (2005) Geography predicts neutral genetic diversity of human populations. Curr Biol 15(5):R159–R160
Rahman NHSNA, Miksic JN, Hasan K, Bellwood P, Datan I, Andaya BW (1998) The Encyclopedia of Malaysia: Early History. Archipelago Press, Singapore, p 144
Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A 102:15942–15947
Reid A (2001) Understanding Melayu (Malay) as a source of diverse modern identities. J Southeast Asian Stud 32:295–313
Ricaut F, Bellatti M, Lahr MM (2006) Ancient Mitochondrial DNA From Malaysian Hair Samples:Some Indications of Southeast Asian Population Movements. Am J Hum Biol 18:654–667
Ricklefs MC (2001) A History of Modern Indonesia since c. 1200. Stanford University Press, Stanford
Rosenberg NA (2004) Distruct: a program for the graphical display of population structure. Mol Ecol Notes 4:137–138
Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385
Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW (2005) Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet 1(6):e70
Sainuddin S (2003) Titas tamadun Melayu. Quantum Books, Perak
Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425
Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in admixed populations. Am J Hum Genet 82:290–303
Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629–644
Shuhaimi NH (1984) Art, Archaeology and the Early Kingdoms in the Malay Peninsula and Sumatra: c 400–1400 A.D. Dissertation. University of London, London, p 651

Stark MT (2006) Early mainland Southeast Asian landscapes in the first millennium A.D. Annu Rev Anthropol 35:407–432
Steyvers M (2002) Multidimensional Scaling. In: Encyclopedia of cognitive science. Stanford University. Macmillan Reference Ltd, Stanford
Su B, Xiao J, Underhill P, Deka R, Zhang W, Akey J, Huang W, Shen D, Lu D, Luo J, Chu J, Tan J, Shen P, Davis R, Cavalli-Sforza LL, Chakraborty R, Xiong M, Du R, Oefner P, Chen Z, Jin L (1999) Y-Chromosome evidence for a northward migration of modern humans into eastern Asia during the last ice age. Am J Hum Genet 65:1718–1724
Syukri I (2002) Sejarah Kerajaan Melayu Patani. Universiti Kebangsaan Malaysia, Bangi, p 131
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
Tang H, Quertermous T, Rodriguez B, Kardia SLR, Zhu X, Brown A, Pankow JS, Province MA, Hunt SC, Boerwinkle E, Schork NJ, Risch NJ (2005) Genetic structure, self-identified race/ethnicity, and confounding in case–control association studies. Am J Hum Genet 76:268–275
Tarling N (1999) The Cambridge history of Southeast Asia, vol 1, part 1. Cambridge University Press, Cambridge, p 368
Taylor JG (2003) Indonesia: Peoples and Histories. Yale University Press, New Haven and London
Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo JM, Doumbo O, Ibrahim M, Juma AT, Kotze MJ, Lema G, Moore JH, Mortensen H, Nyambo TB, Omar SA, Powell K, Pretorius GS, Smith MW, Thera MA, Wambebe C, Weber JL, Williams SM (2009) The genetic structure and history of Africans and African Americans. Science 324:1035–1044
Tzeng J, Lu HHS, Li WH (2008) Multidimensional scaling for large genomic data sets. BMC Bioinformatics 9:179
Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370
Weir BS, Hill WG (2002) Estimating F-statistics. Annu Rev Genet 36:721–750
Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. Genome Res 15:1468–1476
Witherspoon DJ, Wooding S, Rogers AR, Marchani EE, Watkins WS, Batzer MA, Jorde LB (2007) Genetic similarities within and between human populations. Genetics 176:351–359
Xu S, Jin W, Li J (2009) Haplotype-sharing analysis showing uyghurs Are unlikely genetic donors. Mol Biol Evol 26:2197–2206
Xu SH, Gupta S, Jin L (2010) PEAS V1.0: a package for elementary analysis of SNP data. Mol Ecol Resour 10:1085–1088

doi:10.1186/s11568-014-0005-z
Cite this article as: Hatin *et al.*: A genome wide pattern of population structure and admixture in peninsular Malaysia Malays. *The HUGO Journal* 2014 **8**:5.