

The population genomic landscape of human genetic structure, admixture history and local adaptation in Peninsular Malaysia

Lian Deng · Boon Peng Hoh · Dongsheng Lu · Ruiqing Fu · Maude E. Phipps · Shilin Li · Ab Rajab Nur-Shafawati · Wan Isa Hatin · Endom Ismail · Siti Shuhada Mokhtar · Li Jin · Bin Alwi Zilfalil · Christian R. Marshall · Stephen W. Scherer · Fahd Al-Mulla · Shuhua Xu

Received: 7 March 2014 / Accepted: 29 May 2014 / Published online: 11 June 2014
© Springer-Verlag Berlin Heidelberg 2014

Abstract Peninsular Malaysia is a strategic region which might have played an important role in the initial peopling and subsequent human migrations in Asia. However, the genetic diversity and history of human populations—especially indigenous populations—inhabiting this area remain poorly understood. Here, we conducted a genome-wide study using over 900,000 single nucleotide polymorphisms (SNPs) in four major Malaysian ethnic groups (MEGs; Malay, Proto-Malay, Senoi and Negrito), and

made comparisons of 17 world-wide populations. Our data revealed that Peninsular Malaysia has greater genetic diversity corresponding to its role as a contact zone of both early and recent human migrations in Asia. However, each single Orang Asli (indigenous) group was less diverse with a smaller effective population size (N_e) than a European or an East Asian population, indicating a substantial isolation of some duration for these groups. All four MEGs were genetically more similar to Asian populations than to other continental groups, and the divergence time between MEGs and East Asian populations (12,000—6,000 years ago) was also much shorter than that between East Asians and Europeans.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-014-1459-8) contains supplementary material, which is available to authorized users.

L. Deng · D. Lu · R. Fu · S. Xu (✉)
Max Planck Independent Research Group on Population Genomics, Chinese Academy of Sciences and Max Planck Society (CAS-MPG) Partner Institute for Computational Biology (PICB), Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
e-mail: xushua@picb.ac.cn

B. P. Hoh · S. S. Mokhtar
Faculty of Medicine, Institute of Medical Molecular Biotechnology, Jalan Hospital, Universiti Teknologi MARA, Sungai Buloh Campus, 47000 Sungai Buloh, Selangor, Malaysia

M. E. Phipps
Jeffrey Cheah School of Medicine and Health Sciences, Monash University Malaysia, Sunway city, 46150 Selangor, Malaysia

S. Li · L. Jin
Ministry of Education (MOE) Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China

A. R. Nur-Shafawati · W. I. Hatin
Human Genome Center, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kelantan, Malaysia

E. Ismail
Faculty of Science and Technology, School of Biosciences and Biotechnology, Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia

B. A. Zilfalil
Department of Pediatrics, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kelantan, Malaysia

C. R. Marshall · S. W. Scherer
The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada

S. W. Scherer
McLaughlin Centre and Department of Molecular Genetics, University of Toronto, Toronto, Canada

F. Al-Mulla
Department of Pathology, Faculty of Medicine, Kuwait University, 13110 Safat, Kuwait

Thus, Malaysian Orang Asli groups, despite their significantly different features, may share a common origin with the other Asian groups. Nevertheless, we identified traces of recent gene flow from non-Asians to MEGs. Finally, natural selection signatures were detected in a batch of genes associated with immune response, human height, skin pigmentation, hair and facial morphology and blood pressure in MEGs. Notable examples include *SYN3* which is associated with human height in all Orang Asli groups, a height-related gene (*PNPT1*) and two blood pressure-related genes (*CDH13* and *PAX5*) in Negritos. We conclude that a long isolation period, subsequent gene flow and local adaptations have jointly shaped the genetic architectures of MEGs, and this study provides insight into the peopling and human migration history in Southeast Asia.

Introduction

Recent advances in genotyping and sequencing technologies have facilitated genome-wide investigations of human genetic variation and provided new insights into population history and genotype–phenotype relationships. Large-scale international collaborative efforts such as the International HapMap Project (The International HapMap Consortium 2005, 2007) (<http://www.hapmap.org>) and the 1000 Genomes Project (The 1000 Genomes Project Consortium 2011) (<http://www.1000genomes.org>) have produced the most detailed catalogues to date of human DNA variation that may be useful for future evolutionary and medical studies. Although these projects covered world-wide populations, such as those located in or with ancestry from Europe, East and South Asia, West Africa, and the Americas, few Southeast Asian populations were included. Thus, to understand regional human demographic history and local adaptation, more Southeast Asia data are needed (Lu and Xu 2013).

Our earlier studies and those of other groups suggest that Peninsular Malaysia was critical in the peopling of East and Southeast Asia (The HUGO Pan-Asian SNP Consortium 2009). Thus, it is conceivable that differences in diet, climate, and exposure to pathogens among ethnically and geographically diverse Malaysian ethnic groups may have produced distinct selection pressures. Therefore, population genomic studies of Malaysian populations may advance our understanding of human population history and the genetic basis of adaptation to diverse environments.

Proto-Malay, Senoi and Negrito (Semang) are the three main Orang Asli (indigenous) groups consisting of 18 ethnic subgroups in Peninsular Malaysia (6 subgroups for each main group) (Ang et al. 2012). Among these three groups, Senoi is the largest, and Negrito is the smallest (Lim et al. 2010; Ang et al. 2011). Traits that best distinguish among groups are skin color, hair, and height. Traditionally,

Negritos live as hunter-gatherers; Proto-Malays are predominantly farmers, fisherfolk, factory workers and small traders; whereas Senoi communities are involved in plantation and factory work, trading, and to a small extent, hunting and gathering (Fix 1995; Ang et al. 2012).

Previous investigations based on mitochondrial DNA (mtDNA) and Y chromosome indicated that Negritos are genetically closer to Southeast Asian populations than to hunter-gatherers from other continents, e.g., African Pygmies (Thangaraj et al. 2003), although Negritos are very similar to Pygmies with respect to appearance and living environment. Studies of African Pygmies suggested that their phenotypically smaller body size arose from adaptation to tropical rain forests, including food limitation, high temperature and other challenges, and that a convergent adaptation occurred to Negritos (Perry and Dominy 2009; Jarvis et al. 2012; Mendizabal et al. 2012; Lachance et al. 2012); however, the complexity of this prevented us from excluding the possibility that the two dark-skin-color groups shared a common ancestor. Indeed, previous studies indicated that Negritos might be the earliest settlers in Southeast Asia (Endicott et al. 2003; Thangaraj et al. 2005; Reich et al. 2011).

To extend our knowledge and provide the first comprehensive landscape of population history and local adaptations in ethnically diverse Malaysians, we conducted a genome-wide study of the four Malaysian ethnic groups using the Affymetrix Genome-Wide Human SNP Array 6.0, which interrogates more than 900 K single nucleotide polymorphisms (SNPs) encompassing the entire genome. We first measured population genetic diversity among Malaysian ethnic groups, on both the haplotype (haplotype diversity) and SNP (heterozygosity) levels. We also investigated population demography by examining the effective population size (N_e) against time. We next analyzed the population structure of the four ethnic groups together with the other world-wide population samples, to elucidate their genetic relationships and time of divergence with surrounding ethnic groups living in East Asia, Europe and Africa. We then examined the implications of genetic structure and admixture of Malaysian ethnic groups to understand the initial peopling of East Asia and the subsequent human migrations in Asia. Finally, we performed a genome-wide scan for the local adaptation signatures and identified genes related to immune response, human height, skin pigmentation, hair morphology and blood pressure in each ethnic group.

Materials and methods

Populations and samples

After obtaining approvals from all local university research ethics committees, the Ministry of Health and Department

of Indigenous Development (Jabatan Kemajuan Orang Asli, JAKOA), and acquiring informed consent from each participant, blood samples of individuals from four Malaysian ethnic groups, containing 17 Malays (MLY, from Melayu Kelantan), 4 Proto-Malays (PML, from Temuan), 17 Senois (SNI, from Temiar), and 46 Negritos (NGO, from 6 subgroups: Lanoh, Kintak, Jehai, Kensiu, Mendriq and Bateq), were collected from different regions of Malaysia. When possible, as determined by national identity cards and interviews with participants and their chieftains, individuals selected were offspring of non-consanguineous marriages between members of the same ethnicity within three generations. These procedures were also in accordance with the ethical standards of the Helsinki Declaration of 1975 (revised in 2000). For the purpose of making comparisons, we examined recent published genome-wide data from HapMap 3 (Olivier 2003) including 116 YRI (Yoruba in Ibadan, Nigeria), 112 CEU (Utah residents with Northern and Western European ancestry from the CEPH collection), 137 CHB (Han Chinese in Beijing, China), 113 JPT (Japanese in Tokyo, Japan), and 88 GIH (Gujarati Indians in Houston, Texas); and unrelated samples from Human Genome Diversity Project (HGDP) (Li et al. 2008) (<http://www.hagsc.org/hgdp/>): 5 San from Namibia and 17 Papuan from New Guinea. In addition, we integrated samples from Southern Chinese ethnic groups, including 9 Dong, 8 Li, 8 Yao, and 6 Zhuang, of which the genotyping data were previously reported by Lou et al. (2011) and from two Kuwait populations including 42 individuals with Arabian ancestry (ARB) and 22 individuals with Persian ancestry (IRN).

Retrieval of genome-wide association studies data

A list of genes associated with complex traits was compiled from published genome-wide association studies (GWASs) via the National Human Genome Research Institute (NHGRI) GWAS catalog available online (<http://www.genome.gov/26525384>), including 619 genes in total, of which 480 genes are associated with 9 traits of particular interest: facial morphology, hair color, hair morphology, height, skin pigmentation, skin sensitivity to the sun, melanoma, immune response, and malaria. These complex traits are the most distinctive traits in Malaysian populations, especially in indigenous populations. In field work, we observed substantial differences in the incidence of hypertension between NGO and other groups, so we included 139 genes associated with blood pressure and cardiovascular disease in this compiled gene list, with the intent of exploring the possible mechanism of high incidence of high blood pressure in NGO.

Genotyping and data quality control

Genotyping of 84 Malaysian samples with the Affymetrix Genome-Wide Human SNP Array 6.0 was performed

according to the “48 Sample Protocol” (Affymetrix, *Genome-Wide Human SNP Nsp/Sty 6.0 User Guide, Rev. 3, 2008, P/N 702504*). *CEL files containing raw intensity data were analyzed with Birdsuite version 1.5.3 (Korn et al. 2008). There are 2,973 duplicate SNPs among the 893,634 SNPs on the array according to RS numbers. One of each pair with the fewest missing genotypes was kept for further data filtration. Finally, data of 827,204 autosomal SNPs common to all Malaysian ethnic populations were used for downstream analyses.

To minimize bias from the limited sample size, we combined all 6 NGO subgroups. Then two data filtration strategies were performed separately for each population. First, we excluded nearly 100,000 SNPs with missing rates >0.1, and 1 NGO individual with missing rates >0.1, yielding a first dataset of 721,082 SNPs in 83 individuals (17 MLY, 4 PML, 17 SNI, and 45 NGO). We generated the second dataset by another more rigid filtration, involving 353,440 SNPs with missing rates <0.05 and minor allele frequency >0.02, and 60 individuals (17 MLY, 4 PML, 17 SNI, and 22 NGO) with missing rates <0.05. Both datasets comprised autosomal SNPs meeting the criteria in all the 4 populations.

A similar procedure was applied to the 31 Southern Chinese samples and 64 Middle East samples genotyped by Affymetrix Genome-Wide Human SNP Array 6.0 following the standard protocols. For each population, samples with missing rates >0.1 and SNPs with missing rates >0.1 were removed. Finally, 688,865 autosomal SNPs shared by 9 Dong, 8 Yao, 8 Li, and 6 Zhuang samples, and 725,896 autosomal SNPs shared by 41 ARB samples (one sample was removed due to a calling rate less than 85 %) and 22 IRN samples were kept for subsequent analyses.

Besides, samples from HGDP and HapMap 3 database were also filtered separately. Finally, 1,451,219 SNPs (missing rates <0.05) shared by YRI, CEU, CHB, JPT and GIH, and 638,931 SNPs (missing rates <0.1) shared by Papuan and San were included. No individual was excluded due to the missing rates >0.1.

For different purposes of analysis, several combined sub-datasets were generated by employing different data filtration criteria (Table S1).

Haplotype diversity and population heterozygosity

Haplotypes for the 22 autosomes were inferred for each individual from its genotypes with BEAGLE (Browning and Browning 2009) version 3.3.2. We used 651,695 SNPs shared by YRI, CEU, CHB, JPT and 4 Malaysian ethnic groups to estimate the haplotype frequency with a sliding window ranging from 10 to 250 kb along the genome, and measured haplotype diversity based on these estimated haplotype frequencies following Nei (1987). This measure

of haplotype diversity is analogous to the heterozygosity at a single SNP. To provide an overall measurement for each population, we averaged SNP-specific heterozygosity as the population heterozygosity.

Effective population size and divergence time

Totally there were 214,852 SNPs shared by YRI, CEU, CHB, JPT and samples from Malaysian ethnic groups. In the Malaysian samples, one of the 22 Jehai (JEH) individuals was included in the 22 NGO. We phased the genotypes of the 214,852 SNPs that were all biallelic to estimate the effective population size (N_e) and divergence time (Mcevoy et al. 2011). To estimate N_e , r^2 was calculated using PLINK (Purcell et al. 2007) version 1.07 to measure linkage disequilibrium (LD) between SNPs. In a finite population, genetic drift and recombination may be the two main contributors to the LD pattern across the genome, with the former increasing the LD and the latter leading to LD decay. Therefore, recent N_e can be reflected on the LD between distant SNPs considering only the effect of recombination, whereas ancient N_e might chiefly affect the LD of adjacent SNPs. Knowing N_e , we could estimate the divergence time between any two populations, because they confound genetic drift under the neutral evolution theory, which determines the population genetic differentiation as measured by F_{ST} .

Analysis of population structure

The 289,998 autosomal SNPs shared by YRI, CEU, CHB, JPT and Malaysian ethnic groups were used for calculating the global F_{ST} (Weir and Hill 2002) to measure population differentiation, with standard deviations over loci calculated by bootstrapping with 1,000 re-sampling replications, and for constructing the phylogenetic tree with PHYLIP version 3.69 (<http://www.phylip.com/>) and MEGA (Tamura et al. 2011) version 5.05. Then, 25,347 independent SNPs randomly sampled from the overall 289,998 SNPs were used for principal component analysis (PCA) at the individual level with EIGENSTRAT (Price et al. 2006) version 2.0, and also for genetic structure analysis which was performed using a model-based clustering method with STRUCTURE (Evanno et al. 2005) version 2.3. For population structure analysis, we randomly selected 80 samples from YRI, CEU, CHB and JPT (20 for each), to create a comparable population size to Malaysian ethnic groups.

Population admixture analysis

To detect population admixture, 145,852 SNPs shared by all Malaysian samples, and those from HGDP and the HapMap 3 database, were integrated for the *4 Population Test*

(Reich et al. 2009) which was conducted with TreeMix (Pickrell and Pritchard 2012) version 1.1. Then, to estimate population admixture time, we also incorporated samples from Southern Chinese ethnic groups and Kuwaitis but excluded samples from HGDP, and applied an LD-based method [ALDER (Loh et al. 2013)] to the shared 463,912 SNPs.

F_{ST} , locus-specific branch length, integrated haplotype score calculation, and gene functional enrichment analysis

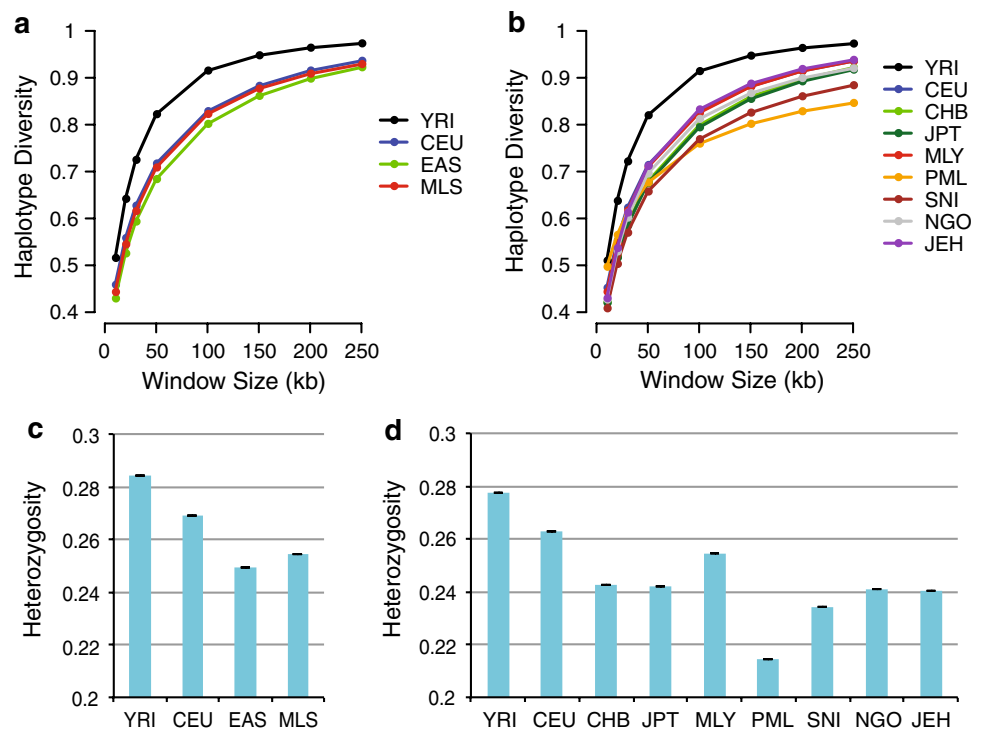
We took 651,695 SNPs shared by YRI, CEU, CHB, JPT and Malaysian ethnic groups for the unbiased estimate of SNP-specific pairwise F_{ST} according to Weir and Hill (2002), and for computing locus-specific branch length (LSBL) (Shriver et al. 2004). We used phased genotypes of 700,359 SNPs with a known ancestral allele to calculate integrated haplotype scores (iHS) for each Malaysian ethnic group. Standardized iHS statistics (Sabeti et al. 2007) were calculated for each Malaysian group separately using the R package *rehh* (Gautier and Vitalis 2012). First, we identified a list of SNPs significantly different between each Malaysian ethnicity and other groups (by F_{ST} and LSBL analyses) or within each Malaysian ethnic group (by iHS analysis). Then we focused on the top signatures, i.e. the top 0.1 % SNPs showing the most extreme differentiation in each comparing pair on F_{ST} and LSBL statistics, and the top 1 % genomic regions in each single group on iHS. Correspondingly, genes covering those significant SNPs or falling within those genomic regions were regarded as candidate genes. Next, functional enrichment analysis was performed to identify putative signatures under local adaptation using DAVID Bioinformatics Resources 6.7 (<http://david.abcc.ncifcrf.gov/>). It was conducted independently on different lists of candidate genes identified in different pairs of populations, using 5 categories: functional category (SP_PIR_KEYWORDS), gene ontology (GOTERM_BP_FAT), pathways (KEGG_PATHWAY), protein domains (INTERPRO) and disease (OMIM_DISEASE).

Results

Genetic diversity of Malaysian ethnic groups

Population genetic diversity differs greatly across continents. As expected, YRI harbored the most diverse haplotypes and the largest heterozygosity among all continental groups (Fig. 1; Table S2). Collectively, Malaysian ethnicities were genetically more diverse than East Asians with an identical sample size, reflecting the complex Malaysian population history. However, individual population patterns were different across Malaysia. Except MLY, the 3

Fig. 1 Comparison of genetic diversity of populations based on haplotype diversity and average SNP heterozygosity. Haplotype diversity of **a** groups of populations and **b** single populations calculated through 10–250 kb sliding windows across the genome. Heterozygosity of **c** groups of populations and **d** single populations. MLS contains MLY, PML, SNI and NGO; EAS contains CHB and JPT. All groups are of equal sample size in **a**, **c**, and 17 samples for each population in **a**, **c**, and 17 samples for each population, except PML (four samples) in **b**, **d**. (95 % confidence intervals *error bars*)



Orang Asli groups (PML, SNI and NGO) were not obviously more diverse than East Asians, and especially, SNI and PML were less diverse than any world-wide populations which indicated longer isolation and inbreeding. To examine whether the potential ascertainment bias could affect our analysis of genetic diversity, we randomly sampled 142,766 SNPs based on the minor allele frequency (MAF) spectrum of the full sequence data obtained from the 1000 Genomes Project (Fig. S1), and used the 142,711 SNPs with MAF larger than 0.05 in all the populations to re-calculate haplotype diversity and heterozygosity. We observed slight difference when we compared results based on re-sampling data and those obtained from the original data. Therefore, ascertainment bias could exist. However, estimation of haplotype diversity was not significantly affected, since our comparisons of relative diversity showed the similar trend as in the original data (Fig. S2). To correct possible overestimations of genetic diversity caused by population substructures in NGO, we did independent estimation for JEH, a sub-tribe of NGO, which had the largest sample size among all NGO subgroups in this study. The genetic diversity observed in JEH and that in the entire NGO group were very similar.

This genetic diversity pattern can be observed with estimations of N_e based on the LD pattern (Fig. S3; Table S3). Great spatial and temporal variance of N_e is depicted in Fig. 2a. Figure 2b shows the harmonic mean of the estimated N_e , representing the long-term N_e , for each population from 250 to 5 KYA. These ranged from 13,900 in YRI

to 1,200 in PML, with Orang Asli groups apparently having a much smaller N_e than all the other populations. Overall, our results supported the “Out of Africa” hypothesis since all the non-African populations showed a significant decline of N_e (Fig. 2b) suggesting they have experienced a shared bottleneck event, although some population specific bottlenecks might also occur subsequently. As recent as 20,000 years ago, a rapid expansion of African (YRI), European (CEU), East Asian (CHB and JPT) and Malay (MLY) occurred and continued until present. However, a slight population decrease was observed in the Malaysian Orang Asli groups who numbered less than 2,500 nearly 10 to 5 KYA (Fig. 2a). Malay group may have expanded due to gene flow from Orang Asli tribes or from other European or East Asian populations. Considering the harmonic mean of N_e estimate from 100 to 30 KYA, we observed an excess of N_e in Negritos (NGO and JEH) compared to East Asians (CHB and JPT) (Fig. 2c). These results supported conclusions from archeological and anthropological studies that Negritos could be descendants of more ancient migrants to Asia among modern Asians. A long isolation and a reliance on hunting and gathering may have kept the Negrito populations size relatively small.

Genetic relationships and population structure

To understand causes of unique population genetic diversity patterns in Malaysia, we investigated genetic relationships and population structures of Malaysian populations.

Fig. 2 Estimation of effective population size (N_e). **a** N_e over time. N_e of Malaysian Orang Asli populations decreased in the recent thousands of years. The harmonic mean of N_e , **b** from 4–200 KYA and **c** from 25–100 KYA. For each population, N_e is the average over recombination distance classes **b** from 0.005–0.25 cM and **c** from 0.015–0.035 cM. (Standard deviations *error bars*)

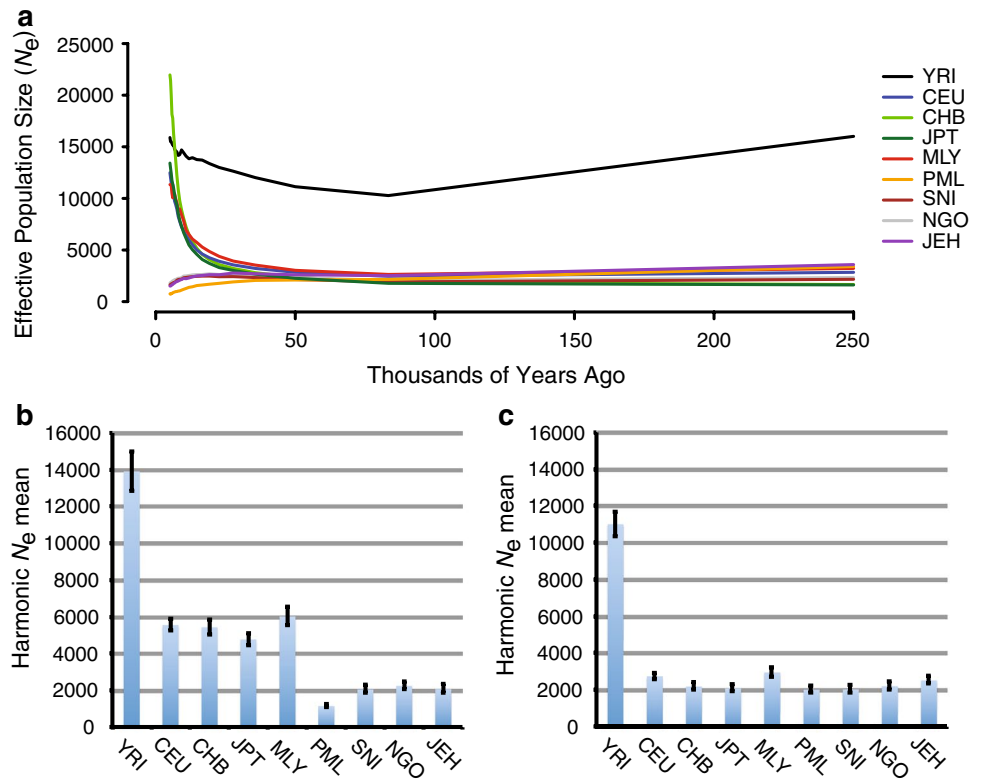
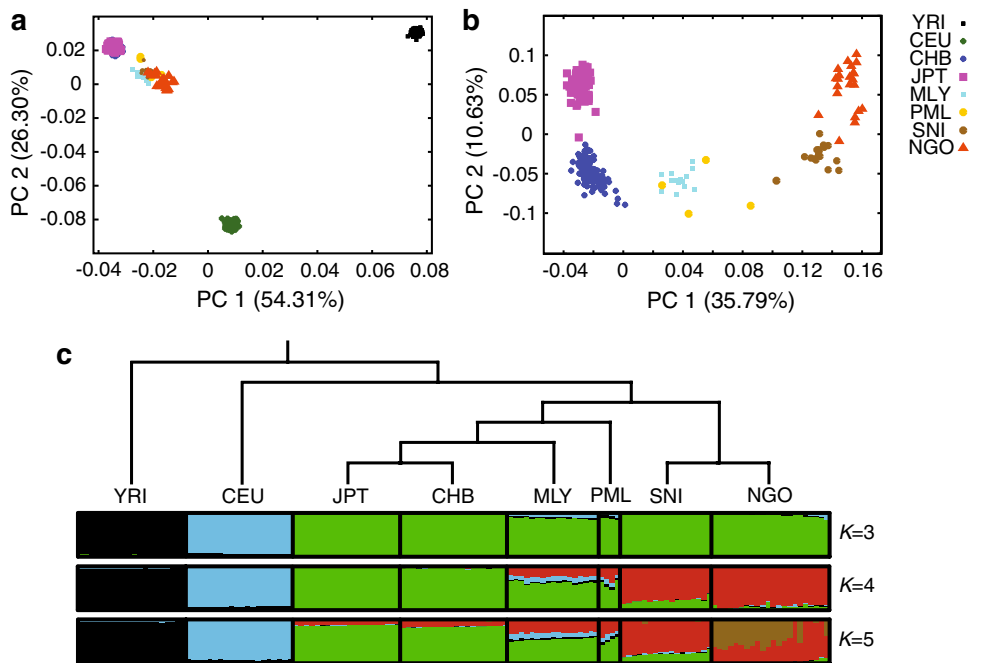


Fig. 3 Population structure and genetic relatedness of eight populations. **a** PCA of 538 individuals representing eight world-wide populations showing that Malaysian populations are genetically closer to East Asians. **b** PCA of 310 individuals representing 6 Asian populations showing a separation among East Asian, Malay&Proto-Malay, and Senoi&Negrito along PC1. **c** Population structure inferred from STRUCTURE shown in the plot adjacent to the corresponding branch in the tree. The number of clusters is indicated by the value of K



Initial PCA of YRI, CEU, CHB, JPT and Malaysian ethnic groups revealed continental genetic variations (Fig. 3a). The Asian cluster was divided into three main clusters (East Asian, Malays&Proto-Malays, and Senois&Negritos) on PC1, with MLY&PML lying between East Asians

(CHB and JPT) and NGO&SNI, suggesting that Malays and Proto-Malays may have experienced admixture from East Asians (Fig. 3b). NGO was widely dispersed along PC2, suggesting possible substructure within it. Meanwhile, SNI was distinguished from NGO to some extent in

Table 1 Divergence time of each pair of populations

<i>T</i> (KYA)	YRI	CEU	CHB	JPT	MLY	PML	SNI	NGO
YRI	–	[58.69, 61.67]	[62.96, 66.84]	[62.05, 65.92]	[66.91, 70.50]	[54.94, 58.78]	[58.98, 63.23]	[58.10, 61.86]
CEU	59.64	–	[23.86, 25.16]	[23.01, 24.15]	[22.43, 23.61]	[16.72, 17.36]	[19.50, 20.44]	[19.35, 20.29]
CHB	64.29	24.34	–	[1.97, 2.10]	[5.84, 6.22]	[7.76, 8.24]	[10.00, 10.66]	[11.30, 12.05]
JPT	63.35	23.40	2.02	–	[7.16, 7.60]	[7.89, 8.33]	[7.27, 8.05]	[11.11, 11.78]
MLY	67.72	22.66	5.93	7.25	–	[6.88, 7.28]	[8.17, 8.70]	[9.49, 10.07]
PML	56.34	16.96	7.95	8.05	6.94	–	[4.90, 5.23]	[5.76, 6.05]
SNI	60.41	19.77	10.15	10.20	8.24	5.01	–	[3.64, 3.91]
NGO	59.19	19.55	11.50	11.25	9.52	5.79	3.67	–

Lower matrix data represent divergence time calculated across genetic distance from 0.005 to 0.25 cM. Upper matrix data in the bracket represent 95 % confidence intervals

YRI, Yoruba in Ibadan, Nigeria; CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; EAS, East Asian including CHB and JPT; MLY, Malaysian population including MLY, PML, SNI and NGO; CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo, Japan; MLY, Malay in Malaysia; PML, Proto-Malay in Malaysia; SNI, Senoi in Malaysia; NGO, Negrito in Malaysia

Fig. 3b, but remained close to NGO even at higher level PCs excluding East Asians (Fig. S4), which clearly distinguished SNI&NGO and MLY&PML. Even though NGO had substructures that were perhaps caused by long subgroup isolation, on each PC level, they were clustered as a whole group with little overlap with other populations, suggesting that the difference between NGO and other populations was larger than that within NGO. Thus, combining all NGO subgroups together to compensate for the limited sample size in each subgroup was acceptable.

A neighbor-joining tree was constructed based on the global F_{ST} of each pairwise population (Fig. 3c; Table S4). Similar genetic relationships among these populations were revealed as that depicted in PCA, and the geographic relationships were fully reflected in the phylogenetic relationships. According to the tree topology, MLY is closer to CHB than to PML, but the global F_{ST} between MLY and CHB ($F_{ST} = 0.015$) and that between MLY and PML ($F_{ST} = 0.016$) were not significantly different (Students' t test, $T = 0.95$, $p > 0.1$). The distance between CHB and PML ($F_{ST} = 0.027$) was nearly the sum of those from MLY to CHB and to PML ($F_{ST} = 0.031$), suggesting the admixture into the modern MLY from East Asians and PML.

Next, we investigated the genetic makeup of Malaysian ethnic groups using hierarchical STRUCTURE analysis. At $K = 3$, all individuals were segregated into three continental groups with all Malaysian populations united with East Asians by a dominant component. At $K = 4$, a new component, accounting for a major proportion in NGO, appeared in all Malaysian ethnic groups, decreasing from SNI&NGO to MLY&PML, then to East Asians. From this gradient, we can infer a decrease of East Asian genomic contribution from MLY&PML to SNI&NGO in Malaysia, which was confirmed in the population structure analysis for only Asian populations (Fig. S5a), and only Malaysian

ethnic groups (Fig. S5b). With increasing K (from $K = 4$ to $K = 8$, data not shown), additional components emerged in NGO. These observations suggest that Malaysian population shares a common origin with East Asians.

Population divergence time of Malaysian ethnic groups

To understand how a general genomic landscape of relationships among all sample groups was established, we estimated the divergence time of each pair of groups with an N_e - and F_{ST} -based method (see “Materials and methods”). Divergence time (T) estimation was shown in Table 1. After the “Out of Africa” event, non-Africans diverged into Europeans and Asians, with the divergence time of 60 KYA between Africans and Europeans, and 64 KYA between Africans and East Asians. Then various populations derived from the Asian group: the divergence time between Malaysian populations with Africans, Europeans and Asians declined ($T_{YRI-Malaysian} \approx 60$ KYA, $T_{CEU-Malaysian} \approx 20$ KYA, $T_{CHB-Malaysian} \approx 10$ KYA). In these four Malaysian ethnic groups, the shortest divergence time was that between SNI and NGO (4 KYA). And SNI seemed to be closer to MLY&PML than NGO did, with the T estimated as 5–8 KYA, which was smaller than the divergence time between NGO and MLY&PML around 6–10 KYA. Collectively, compared to the time of the splits of different continental groups brought about by “Out of Africa”, the divergence within Malaysian population occurred more recently, and Malaysian ethnic groups shared an evolutionary branch with East Asians.

Population admixture

Southeast Asia is a region of great historical and genetic diversity and it has undergone extensive population

Table 2 Measurement of gene flow from East and South Asians, Europeans and Africans to Malaysian populations, and those among Malaysian populations

	MLY		SNI		NGO	
	Date (generations ago)	Proportion (%)	Date (generations ago)	Proportion (%)	Date (generations ago)	Proportion (%)
East Asia–Malaysia						
CHB	8.3 ± 3.6	11.9 ± 2.1	26.1 ± 7.3	11.6 ± 2.0	32.8 ± 3.3	21.3 ± 1.5
JPT	NA	NA	34.4 ± 8.2	11.7 ± 1.5	36.8 ± 2.8	19.3 ± 1.1
Dong	17.5 ± 5.2	16.2 ± 2.6	28.8 ± 9.7	13.8 ± 2.2	33.9 ± 2.9	22.6 ± 1.3
Yao	13.9 ± 3.6	29.2 ± 5.1	34.5 ± 8.3	17.0 ± 1.9	35.1 ± 3.1	26.3 ± 1.4
Li	17.8 ± 8.6	20.4 ± 5.7	23.1 ± 6.6	13.5 ± 2.1	33.5 ± 3.2	24.6 ± 1.4
Zhuang	NA	NA	40.0 ± 8.3	15.9 ± 2.4	34.6 ± 2.7	25.5 ± 1.2
South Asia–Malaysia						
GIH	28.1 ± 5.9	14.9 ± 1.6	36.7 ± 17.9	6.0 ± 1.5	45.5 ± 4.3	9.2 ± 0.5
Middle East–Malaysia						
ARB	35.3 ± 5.1	8.1 ± 0.8	42.9 ± 13.4	3.5 ± 0.6	46.3 ± 4.6	5.0 ± 0.4
IRN	33.1 ± 5.2	10.2 ± 1.0	40.3 ± 13.9	4.2 ± 0.7	44.6 ± 4.5	5.7 ± 0.5
Europe–Malaysia						
CEU	36.3 ± 5.6	7.9 ± 0.8	47.9 ± 18.2	3.2 ± 0.8	46.4 ± 5.4	4.5 ± 0.4
Africa–Malaysia						
YRI	40.4 ± 6.3	1.6 ± 0.2	27.8 ± 11.5	1.1 ± 0.3	45.8 ± 5.1	1.6 ± 0.1
Within Malaysia						
MLY	–	–	21.0 ± 6.9	18.8 ± 3.9	26.7 ± 2.6	31.1 ± 1.6
PML	NA	NA	31.2 ± 7.2	13.9 ± 2.1	28.3 ± 3.0	19.8 ± 1.5

Populations in the first column are gene flow donors, and those in the first row are gene flow receptors

GIH, Gujarati Indians in Houston, Texas; ARB, Arab descendants from Kuwait; IRN, Iranian descendants from Kuwait; NA: data are not available (due to small sample size). Other population IDs appear in Table 1

admixture events (The HUGO Pan-Asian SNP Consortium 2009). Consistently, with population structure analysis (Fig. 3c), admixture patterns were observed in Malaysian populations, even though Malaysians shared overall similarity with East Asians. To illustrate, at every level of K , Malaysian ethnicities, especially MLY and PML, harbored a small proportion of European and African components. To confirm the admixture in Malaysian groups, we applied the 4 Population Test based on the phylogenetic tree (Fig. 3c). In these analyses, to minimize the inter-population effect between our target populations and others, we selected San and Papuan that are indigenous populations in South Africa and New Guinea, respectively, as reference populations because they are both distant from other world-wide populations (Magalhães et al. 2012).

We performed the 4 Population Test on San, YRI/CEU, Malaysian ethnic groups and Papuan, which was denoted as f_4 (San, X; M, Papuan) with M representing one of the four Malaysian ethnic groups, and X as YRI or CEU. Inconsistent with the tree, the values were all negative for MLY, PML, SNI and NGO ($p < 0.01$) whether considering YRI or CEU (Table S5), implying possible gene flow from both Africans and Europeans to the four ethnic groups. So the phylogenetic tree reconstructed as such directly from

modern genomic data did not provide an excellent fit to the evolutionary relationships of these populations. Some admixture events did happen between Malaysian populations and other continental groups as indicated by STRUCTURE analysis.

To estimate the admixture proportion and date, we examined the LD decay in the four Malaysian admixed groups. Based on the fact that LD decays at a constant rate as recombination breaks down the contiguous chromosomal blocks inherited from the ancestral mixing populations, we could date the admixture events by admixture LD between SNPs (Chakraborty 1988) (See “Materials and methods”). East Asians contributed most to the Malaysian populations compared to Africans and Europeans, reaching 12–29 % in MLY, 12–17 % in SNI and 19–26 % in NGO. Southern Chinese ethnic groups, e.g., Dong, Li, Yao and Zhuang, had more interactions with Malaysian populations due to proximity to Malaysia (Table 2). Genetic admixture between East Asians and Malay occurred ~8–18 generations (200–450 years) ago, which was later than the admixture between Malay and Orang Asli groups which was estimated to be around 21–31 generations (525–775 years) ago, and also much later than that between Orang Asli groups and people with East Asian ancestry which happened ~23–40

generations (575–1,000 years) ago, assuming a generation time of 25 years which was applied throughout this study. These results indicated that Malaysian Orang Asli groups had an even earlier interaction with East Asian ancestry than ancestors of Malay people, which was not expected considering the fact that geographical location of Malay is much closer to where the Orang Asli groups residing. However, it was unlikely that the East Asian ancestry in Orang Asli groups was via gene flow from Malay, since the estimation of gene flow from East Asians to Orang Asli groups (about 20 %, Table 2) was much higher than that from East Asian to Malay (about 10 %, Table 2).

Taking YRI as the ancestral population, we observed some gene flow with a proportion around 1–2 % in MLY, SNI and NGO. The dates are 40, 28 and 46 generations (1,000, 700 and 1,150 years) ago, for MLY, SNI and NGO, respectively (Table 2). The European mixture proportions are larger than those of Africans, ranging from 3 % in SNI to 8 % in MLY. Our estimates for the dates are 36 generations (900 years) ago for MLY, 48 generations (1,200 years) ago for SNI, and 46 generations (1,150 years) ago for NGO. The European contribution to Malaysians could have mainly resulted from more recent colonization events, starting from 1600s beginning with the Portuguese, followed by the Dutch and eventually the British. Because ancient literature suggested interactions of Middle East and India with the Malay land (Andaya 2002), we replaced CEU with Arabs and Iranians from Kuwait and Indians, and obtained a relatively more obvious admixture estimate (Table 2). However, we failed to measure the gene flow to PML, which could not be large according to the STRUCTURE analysis. Overall, although Malaysian populations are closer to other Asian populations genetically, they harbored a fraction of African or European genomes, resulting presumably from recent admixture with Africans and Europeans.

Although we could not detect genetic interactions within Malaysia because of the lack of proper reference populations, we documented gene flow among Malaysian populations. On one hand, they are geographically close enough to create frequent admixture, and on the other hand, some admixture patterns were easily found in the PCA and population structure analyses (Fig. 3). Because ALDER is unsuitable for estimating the mixture between populations that are genetically too close, we simply measured the gene flow from MLY and PML to SNI and NGO. We found that MLY and PML have the largest mixture proportion (14–31 %) with SNI and NGO among all the analyzed populations, and that they mixed most recently, around 21–31 generations (525–775 years) ago.

Our estimation of admixture time indicated the genetic contribution of western Eurasian ancestry (from European) to Malaysian people could have started more than

1,000 years ago, which was much earlier than historical records (Andaya 2002). However, it might be not that beyond expectation since the genetic admixture often occurred before it was seen and recorded. To this end, our analysis identified recent gene flow to Peninsular Malaysia from western Eurasian and African populations, and these results suggested a great impact of population admixture on the genetic makeup of human groups in Peninsular Malaysia and surrounding regions.

Signatures of local adaptation in Malaysian ethnic groups

Besides the unique genetic structure and human migration, local adaptation contributes to the genetic diversity of Malaysians, because it allows allele accumulation and thus provides competitive advantages in specific environments of different geographic regions. The four Malaysian ethnic groups in this study differ significantly in appearance (Ang et al. 2012), and the most distinctive phenotypes are height, skin color, and hair. More interestingly, the average blood pressure was higher in NGO compared to PML and SNI ($p < 0.0001$), with the highest average blood pressure reaching 140/91 mmHg in Bateq, an NGO subgroup. The average blood pressure of NGO was also higher than the world-wide normal blood pressure according to WHO guidelines (120/80 mmHg is normal, and $\geq 140/90$ mmHg is high blood pressure; from <http://www.euro.who.int/en/home>).

To identify signatures of local adaptations in Malaysians, we applied both allele frequency-based (F_{ST} and LSBL) and haplotype-based (iHS) methods to search for footprints of natural selection in genomic data (see “Materials and methods”). We took F_{ST} as the primary statistic in our study to utilize the most intuitive and direct information in the data, since the accuracy of iHS may be affected by fluctuations in sample size and SNP density. The significant terms in which the candidate genes identified by F_{ST} -based method enriched (Benjamini FDR corrected $p < 0.05$) are summarized in Table S6. We found that candidate genes in all the comparisons were more significantly related to nervous system development, which mainly involved neuron differentiation and projection, and synapse organization and transmission, reflecting the neural adaptation caused by various stimuli in different environment. Also, many genes were enriched in the annotation cluster for ion transport. For instance, all Malaysian ethnicities differ substantially from East Asians and Europeans with respect to the calcium homeostasis, which could affect many aspects of human development and metabolism (formation of bones/teeth, muscle fiber contraction, and hormone secretion). Particularly, upon comparing NGO with other populations (PML, CEU, CHB and JPT), we found signatures playing functions on sodium transport. Sodium

transport, specifically the $\text{Na}^+ - \text{K}^+$ pump, helps to maintain blood pressure, and the increase of sodium in the extracellular space may lead to hypertension (Llaurado 1983; The DASH-Sodium Collaborative Research Group 2001). This may explain the high blood pressure observed in NGO samples. Interestingly, signals of some immunoglobulin-related genes showed in Malaysian ethnicities, probably due to exposure to a distinct microbial environment. In addition, genes associated with cell adhesion and epidermal growth factor (EGF) were also enriched in the Malaysian ethnic groups.

Nearly 70 % of the candidate genes identified in the F_{ST} -based analysis were confirmed by the LSBL-based method, and this proportion was much higher than that of candidate genes identified in both F_{ST} -based and iHS-based analyses, i.e., only 30 % of the candidate genes identified by F_{ST} -based method were replicated in iHS-based analyses (Fig. S6). However, we did obtain a list of genes underlying possible natural selection which have been identified by all the three methods. Our functional annotation showed that these genes played various roles in different biological processes. Some candidate genes can support functional adaptations of indigenous populations to their life in the rain forests. First, due to the specific diets of Orang Asli groups, especially hunter-gatherers, several genes contributing to nutrient uptake and even those related to type 2 diabetes and obesity had significant natural selection signatures, such as *LRP2* in SNI and NGO, *DLC1* in NGO, *ZMAT4* in SNI, and *TCF7L2* in PML. Second, tropical forest environments that have distinct temperatures and ultraviolet radiation (which can produce unique external injuries) contribute to differences in indigenous populations, especially with regard to genetic variants associated with UV and temperature stimuli responses, wound healing, sensory and visual perception, and behavioral response. For example, in candidate genes of SNI, *PTPRK* and *XYLT1* are related to abiotic stimulus responses, and *NRG1* is related to wound healing. In PML, *KCNMA1* and *TTC8* function in sensory organ development. We also identified genes on some hunter-gatherer-specific phenotypes (height and facial morphology). Notable examples are *PNPT1* and *ETV6* in NGO, and *MAML2* in SNI, all of which are related to adult human height, and *PLEKHAI* coding for facial morphology identified in SNI. Moreover, some genes we identified were candidate genes in previous genome-association studies, mainly for neurological control and cancer. For instance, *RGL1* and *A2BP1* were thought to be related to attention deficit hyperactivity disorder (ADHD) (Anney et al. 2008; Sonuga-barke et al. 2008), and they were identified in MLY and NGO here, respectively. Other significant genes repeatedly identified in F_{ST} -based and LSBL-based analyses, and in F_{ST} -based and iHS-based analyses are summarized in Table S7 and Table S8, respectively.

Candidate genes for complex traits

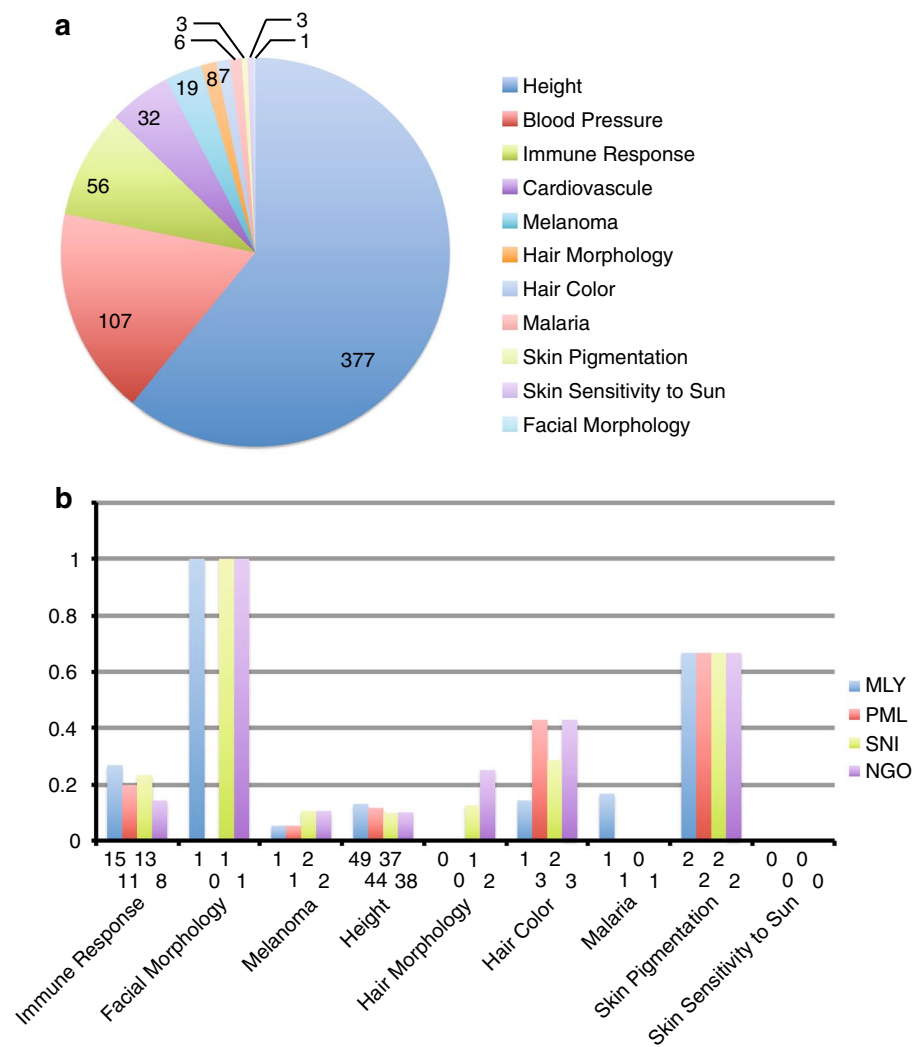
Although we identified genes with significant natural selection signatures related to many important biological processes, background noise prevented identification of genes contributing to the phenotypic diversity we observed in Malaysian ethnic groups from de novo screening. These promising signatures could have been masked by: (1) the long history of the Orang Asli groups; and (2) the complex demographic events, especially population admixture, in Peninsular Malaysia. Therefore, we applied a GWAS gene scan in each population to determine if we could identify previously reported genes related to the distinct traits in Malaysian populations, including skin pigmentation, hair and facial morphology, height, immune response, and blood pressure. In the compiled GWAS gene list (see “Materials and methods”), 377 of the 619 genes were related to human height, comprising the majority of this gene list, followed by 107 genes related to blood pressure (Fig. 4a). Mapping of GWAS genes included: (1) mapping GWAS genes related to seven traits in appearance, malaria, and immune response in all the four Malaysian populations; and (2) mapping blood pressure-associated genes and cardiovascular genes in NGO.

First, a small proportion of the total candidate genes in each population were related to those complex traits, with the largest proportion (3.7 %) in MLY and the least (3.3 %) in SNI. Meanwhile, GWAS genes identified as significant in our study could comprise only a small proportion of the GWAS gene list, most of which were height-related genes, followed by immune response-related genes, and a few genes related to hair, skin, and facial morphology in different populations (Fig. 4b). The top ten most significant GWAS genes identified in each population are summarized in Table 3. Despite their statistical significance, they did not rank high when compared with other candidate genes.

Of the 56 candidate GWAS genes in NGO, 58 in SNI, and 61 in PML, a height-related gene, *SYN3*, was remarkable in both F_{ST} and LSBL analyses (Tables 3, S9). In addition, *SYN3* fell in the region with ~30 % SNPs showing $\text{liHSI} > 2$ in NGO and SNI. In PML, the region encompassing it harbored 93.3 % significant SNPs. *PNPT1*, which is also associated with height, had an even stronger signature than *SYN3* in NGO according to F_{ST} , LSBL and iHS analyses. For MLY, a notable example showing signatures of potential natural selection is *RUNX1*, which is responsible for the immune response. With the above observations, we suggested that genetic adaptations of appearance such as height might have occurred in Malaysian indigenous groups.

Next, identification of genes linked to high blood pressure and related cardiovascular disorders in NGO individuals was performed. Screening the compiled gene list

Fig. 4 GWAS genes identified as signatures of local adaptation. **a** GWAS genes are from published GWASs via the National Human Genome Research Institute (NHGRI) GWAS catalog available online, and related to 11 complex traits with the number of genes in each trait showed in the pie. **b** Proportion of GWAS genes that are located in the regions with evidence of local adaptation. The number of genes identified is shown below each bar



revealed that several blood pressure-related genes were significant on F_{ST} in NGO (Table 4), such as *CDH13* and *PAX5*. *CDH13* encodes a member of the cadherin superfamily that protects vascular endothelial cells from apoptosis due to oxidative stress and is associated with resistance to atherosclerosis (Ivanov et al. 2004). *PAX5* encodes a member of the paired box (PAX) family of transcription factors and *PAX5* proteins are important regulators in B cell differentiation, neural development and spermatogenesis (Adams et al. 1992; Emelyanov et al. 2002).

Discussion

Southeast Asia covers an area of 1.6 million square miles and is home to 600 million people. A wide variety of ethnic groups exists, accounting for its diverse range of languages and religions. Peninsular Malaysia is located in the heart of Southeast Asia, and is very diverse in both geography and culture. The entire Malaysian population comprises

numerous ethnic groups, including Malays, Chinese, Indians and different indigenous groups, and as such is a complex multilingual and multicultural society. The wealth of human diversity in Malaysia makes it a representative region in Southeast Asia. Taking advantage of the high-density genotyping data in four major and representative Malaysian ethnic groups (MLY, PML, SNI and NGO), we conducted the first comprehensive genome-wide study and provided a genomic landscape of populations in Peninsular Malaysia with respect to their evolutionary history and local adaptation.

Compared with European and East Asian populations, the overall genetic diversity in Peninsular Malaysia as represented by the four major groups is much higher, but lower in each single indigenous group (Fig. 1; Table S2). This can be seen in the patterns of N_e . For example, the N_e of MLY reaches about 6,000, but it is only less than 3,000 in Malaysian indigenous groups, which is much lower than any other world-wide populations (Fig. 2), possibly due to: (1) the inbreeding and long isolation of indigenous groups;

Table 3 Top 10 GWAS genes identified in 4 Malaysian ethnic groups by F_{ST} -based analysis

Group	Gene	Chr	F_{ST}			LSBL	iHS	Trait	
			Comparing group	No. SNPs	Rank				Top F_{ST}
NGO	<i>C6orf173</i>	6	CHB/JPT/SNI	8/8/8	4/4/106	0.75/0.76/0.39	+	H	
	<i>PNPT1</i>	2	CEU/YRI	2/2/1	8/47/10	0.84/0.92	+	+	H
	<i>SYN3</i>	22	YRI/PML	1/4	10/11	0.97/0.86	+	+	H
	<i>CDH13</i>	16	MLY/CHB/JPT	6/9/7	11/25/27	0.58/0.67/0.65	+	+	H
	<i>ETV6</i>	12	JPT/MLY/CHB/ PML/SNI/CEU	3/3/2/2/1	15/57/87/96/172/268	0.67/0.49/0.58/ 0.79/0.36/0.67	+	+	H
	<i>ANTXR1</i>	2	CEU	2	23	0.81	+		H
	<i>PAX3</i>	2	SNI/JPT/MLY	1/1/1	27/142/178	0.46/0.56/0.43	+		FM
	<i>BNC2</i>	9	CEU/MLY	8/1	29/319	0.79/0.40	+	+	H
	<i>UQCC</i>	20	JPT/CHB/CEU	7/5/1	33/44/110	0.64/0.62/0.73	+		H
	<i>GDF5</i>	20	JPT/CHB/CEU	9/6/2	35/46/112	0.64/0.62/0.73	+		H
SNI	<i>SYN3</i>	22	YRI/PML/CHB/JPT	2/2/2/2	3/47/97/197	0.99/0.80/0.65/0.61	+	+	H
	<i>RUNX1</i>	21	PML/JPT/CHB	2/1/1	6/201/314	0.88/0.61/0.57	+		IR
	<i>SLC24A5</i>	15	CEU	1	8	0.90			SP
	<i>SPSB1</i>	1	MLY/JPT/CHB	1/2/2	10/16/23	0.56/0.76/0.74	+	+	IR
	<i>CDH13</i>	16	PML/CHB/JPT/YRI	7/8/6/1	14/187/211/281	0.86/0.61/0.61/0.89	+		H
	<i>UNC13C</i>	15	CEU/PML/NGO	2/5/2	15/156/247	0.89/0.71/0.34	+		IR
	<i>PLEKHA1</i>	10	CHB/JPT/MLY	2/2/2	18/26/299	0.76/0.73/0.39	+	+	H
	<i>SLIT3</i>	5	CHB/JPT/MLY	3/3/2	25/106/270	0.74/0.66/0.40	+		H
	<i>DNM3</i>	1	MLY	8	25	0.52	+		H
	<i>PAX3</i>	2	NGO	1	27	0.46			FM
PML	<i>SYN3</i>	22	CHB/JPT/NGO/YRI/SNI	2/2/4/2 2	1/5/11/34/47	0.95/0.94/0.86/0.99/0.80	+	+	H
	<i>FREM1</i>	9	MLY/CHB/CEU	5/3/3	6/209/299	0.77/0.83/0.93	+		H
	<i>RUNX1</i>	21	SNI/JPT/CHB/MLY/YRI	2/3/4/2/2	6/34/54/57/125	0.88/0.90/0.87/0.71/0.98	+		IR
	<i>CDH13</i>	16	SNI/CHB/CEU/JPT/YRI	7/3/4/3/1	14/139/263/295/400	0.86/0.85/0.94/0.83/0.96	+	+	H
	<i>SLC24A4</i>	14	SNI/NGO	2/1	33/272	0.81/0.77	+		HC
	<i>SLC24A5</i>	15	CEU	1	37	0.98	+		SP
	<i>OCA2</i>	15	JPT/CHB	4/5	40/56	0.90/0.87	+	+	HC
	<i>FOLH1</i>	11	MLY	1	41	0.71	+	+	H
	<i>GPR158</i>	10	CEU	2	50	0.96			IR
	<i>SUPT3H</i>	6	NGO	1	73	0.81	+		H
MLY	<i>SLC45A2</i>	5	CEU	1	2	0.87			M, HC, SP
	<i>RUNX1</i>	21	CHB/JPT/PML/YRI	1/2/2/3	4/23/57/228	0.58/0.54/0.71/0.88	+	+	IR
	<i>FREM1</i>	9	PML	5	6	0.77	+		H
	<i>SYN3</i>	22	YRI/CHB	1	6	0.98	+		H
	<i>SPSB1</i>	1	SNI	1	10	0.56	+		IR
	<i>SLC24A5</i>	15	CEU	1	11	0.84			SP
	<i>CDH13</i>	16	NGO/YRI/CHB/CEU/JPT	6/2/2/1/1	11/76/141/223/293	0.58/0.91/0.40/0.71/0.39	+		H
	<i>PDE11A</i>	2	YRI	6	20	0.94		+	H
	<i>DNM3</i>	1	SNI	8	25	0.52	+		H
	<i>ZBTB38</i>	3	YRI	2	33	0.93		+	H

The data separated by a slash in each column are corresponding to the results obtained in populations showing in the second column. For example, in the first gene in NGO, *C6orf173*, the No. SNPs (8/8/8) denotes 8 SNPs were identified in MLY, CHB and JPT, respectively; the top ranks of SNPs in the three populations are 4, 4 and 106, respectively; the F_{ST} values of top SNPs are 0.746, 0.760 and 0.389, respectively

Population IDs appear in Table 1. + represents that the gene in the second column is also significant in LSBL-based or iHS-based analysis

Chr chromosome, No. SNPs number of SNPs falling in the gene region, H height, HC hair color, FM facial morphology, M Melanoma, SP skin pigmentation, IR immune response

Table 4 Blood pressure-related genes detected by F_{ST} analysis in Negrito

Gene	Comparing group	Chr	No. SNPs	Rank	Top F_{ST}	Top SNP
<i>CDH13</i>	MLY/CHB/JPT	16	6/9/7	11/25/27	0.58/0.67/0.65	rs4783307/rs12933997/rs12933997
<i>PAX5</i>	PML	9	3	24	0.86	rs7466692
<i>STK39</i>	CEU/MLY/SNI	2	1/1/1	55/269/246	0.76/0.41/0.34	rs11897461/rs6740826/rs6740826
<i>EBF1</i>	PML	5	1	68	0.82	rs6895454
<i>MYO16</i>	SNI	13	2	85	0.40	rs484455
<i>CNNM2</i>	PML	10	2	98	0.79	rs17115327,rs3740388
<i>AS3MT</i>	MLY	10	4	137	0.44	rs10748835,rs11191438
<i>MTHFR</i>	MLY	1	1	166	0.43	rs1801131
<i>PHACTR1</i>	CEU	6	3	169	0.70	rs10484269
<i>SLC4A7</i>	CHB	3	3	186	0.53	rs13078798
<i>TBX5</i>	SNI	12	1	224	0.35	rs11067076
<i>CACNB2</i>	SNI	10	1	271	0.33	rs4237348
<i>LMAN1L</i>	CEU	15	2	298	0.66	rs3743486
<i>ZNF259</i>	MLY	11	1	303	0.40	rs6589566
<i>CYP17A1</i>	MLY	10	1	311	0.40	rs10786712
<i>ARID3B</i>	CEU	15	2	349	0.65	rs10459601,rs7497036

Population IDs appear in Table 1. Other abbreviations, including data separated by a slash appear in Table 3

and (2) underestimations resulting from limited samples of Malaysian ethnic groups.

The complex patterns of genetic diversity in modern populations are results of demographic history and selective effects that have acted to adapt different populations to their environments (Balaresque et al. 2007). We therefore attempted to elucidate the unique genetic diversity patterns in Malaysian using two investigative approaches. First, we investigated the population structure and history, and then we focused on the identification of signatures of local adaptation for the different Malaysian ethnicities in this study. Generally, all Malaysian ethnic groups are more closely related to East Asians than to Europeans and Africans, among which Malays and Proto-Malays harbored more Asian ancestries than Senois and Negritos (Fig. 3). We inferred the divergence time of Africans and non-Africans as ~60 KYA (Table 1). Compared to the time of “Out of Africa”, the divergence of Malaysian populations from East Asians (~10 KYA) occurred more recently than that between Europeans or Africans and East Asians. In particular, our estimation provided genetic evidence for a split between ancestors of NGO and SNI occurring nearly 4,000 years ago, suggested by language analyses in ancient literature (Benjamin 1976). Although Malaysian population shared more recent common ancestry with East Asians, gene flows from Africans and Europeans were detected, implicating some admixture events. A small proportion of gene flows occurred from Africans and Europeans to Malaysians, and correspondingly, we observed those shared genetic components between Malaysians, especially Malays and Proto-Malays, with Europeans and

Africans in the STRUCTURE analysis (Fig. 3c). In addition, we identified genes with signatures of natural selection on distinct appearance traits of different Malaysian ethnic groups (Fix 1995; Ang et al. 2012), especially height and skin pigmentation, and the high blood pressure in NGO. Notable examples, such as *SYN3* and *PNPT1*, which are both associated with human height, and *CDH13* and *PAX5* related to blood pressure, were found to play some roles in local adaptations of Malaysian indigenous populations. All these observations allow us to conclude that a long period of isolation, subsequent gene flow, and local adaptations jointly shaped the genetic architectures of the present-day communities living in Peninsular Malaysia.

This study provides a comprehensive picture of Malaysian ethnic groups and explains the reasons accounting for the genetic diversity in the population. In a previous study, we examined a data set with much lower density SNPs in Malaysian Malay group and we explored the population history of Malay (Hatin et al. 2011). With respect to genetic architecture and population history of Malay, most of the results and conclusions are supported by this study based on higher density data. For instance, previously we already identified gene flow to the Malay from their surrounding populations including Chinese and Indians. Here, by the admixture analysis in the current study, we confirmed the admixture between Malaysian populations and the surrounding neighborhoods, and furthermore, we could quantitatively estimate the gene flow from East Asians (10–30 %) and Indians (14.9 %) to the Malay, respectively. In addition, we also confirmed that, compared with Negritos,

Proto-Malays were more genetically related to Malays and Chinese (Fig. 3c).

Although the HUGO Pan-Asia Initiative revealed the great genetic diversity of Southeast Asian and its important role in human evolutionary history (The HUGO Pan-Asian SNP Consortium 2009; Ngamphiw et al. 2011), the population histories accounting for this variance remained poorly understood. Our findings on population admixture are consistent with the previous studies providing molecular evidence for the admixture of the three Malaysian indigenous groups with European (Ang et al. 2012). On one hand, the European gene flows mentioned in those studies are fully confirmed. But on the other, our results might have been biased by insufficient sample size and the limited number of SNPs integrated for this part of analysis. Geographical location, colonization and trade are main reasons for the population admixture in Peninsular Malaysia. Chinese traders first visited Peninsular Malaysia in the mid-fifth century, followed by Arab and Persian (known as Iranian today) merchants in the early eighth century (Andaya 2002). In the fourteenth century, the Malacca Sultanate developed international trade with the Yuan Dynasty, attracting an increasing number of traders from other countries, such as Chinese, Persians, Arabs and Gujarats. Since 1600s, Malaysia had been colonized by many European countries, such as Portugal, the Netherlands and Britain, resulting in the gene flow from Europeans to Malaysian populations. Since much higher gene flow was detected in Malaysian ethnic groups especially Malay (Table 2) from Middle East and Indian populations than that from Europeans, it is likely that the European genetic component in Malaysians was brought into Malay via people from Middle East and India. Indeed, the average genetic difference between Malaysian populations and Middle East population ($F_{ST} = 0.09$) is slightly smaller than that between Malaysian populations and Europeans ($F_{ST} = 0.10$). According to historical records, the first millennium and a half AD was the period of Indianization of Peninsular Malaysia (Andaya 2002). Then Indians began migrating to Malaysia as laborers in the early nineteenth century forced by the British during colonial time, which had a huge impact on the demographics, culture and society of Malaysia, and the Indian community has become one of the main ethnic groups in Malaysia nowadays (Hirschman 1975; Saw 1988). Taken together, we suggest that the Malaysian population is of great historical and genetic complexity, which might have resulted from extensive admixture events. Such complexity might cloud the estimation of divergence time, which was estimated under the assumption of no population admixture, and present difficulties with the identification of local adaptation signatures.

Local adaptations in different populations are largely driven by environmental pressures and may confer fitness

advantages. For example, a strong capacity of nutrient absorption, wound healing, and sensory and visual perception may help adapt to the hunter-gathering life. Some genes related to these phenotypes have already been identified in our previous study when we compared Orang Asli populations or Southeast Asian populations with others (Qian et al. 2013), and some of them were also confirmed in this study, such as *DLC*, *CDH13*, *RGL1* and *A2BP1*. The short stature of Negritos might be more suitable for hot tropical environments, offering better thermoregulation, reduced caloric requirements, and greater mobility (Migliano et al. 2007; Perry and Dominy 2009). Previous studies revealed the GH1-IGF1 pathway as a candidate for the short stature in hunter-gatherers from both Africa and Southeast Asia (Clavano-Harding et al. 1999; Dávila et al. 2002). Here *IGF1* and *GHR* participating in the GH1-IGF1 pathway were identified in the analyses of F_{ST} or LSBL. More often, height-related genes were previously identified in African pygmies only (Jarvis et al. 2012; Mendizabal et al. 2012; Lachance et al. 2012), some of which were also significant in this study of Malaysian Orang Asli groups (Table S10). Besides, previous association studies suggested that some skin pigmentation-related genes may affect the skin color of different Orang Asli groups, e.g., *SLC45A2* and *SLC24A5* (Ang et al. 2012). These genes were as significant as we expected in our F_{ST} analysis. Both fell in the top 0.1 % signatures in four Malaysian populations compared with Europeans (Table S10). However, these signatures were not statistically significant in iHS analysis, suggesting that *SLC45A2* and *SLC24A5* could be actually resulted from selection in Europeans rather than in Malay populations. Other reported genes, such as *MC1R* and *ASIP*, did not have observable signatures likely because they were identified in association studies of populations with substantially different genetic background from those are studied here. Specifically, *TYR* was identified to be related to skin variations between European and South Asians (Sulem et al. 2007).

Two challenges were apparent with this study. First, our estimations are based on small samples and this may produce bias for some analyses, such as allele frequency calculation and haplotype investigation. As such, N_e and divergence time of these ethnic groups might be underestimated. We randomly sampled 20 individuals from YRI, CEU, JPT and CHB, respectively, to make them comparable to Malaysian populations, and re-estimated N_e and divergence time for them. Consequently, N_e for the recent thousands of years is much smaller than those of non-sampled populations, and there is also an underestimation in divergence time (Fig. S7; Table S11). With small samples, we could not accurately calculate allelic frequency and gene flow. For example, with four Proto-Malay samples, we could hardly observe an exponential decay of the weighted LD

curve (Loh et al. 2013) while estimating population admixture irrespective of the reference ancestral populations we used. Thus, we could not measure gene flow from other populations to Proto-Malay. However, PCA and STRUCTURE analyses were accurate; they are not so sensitive to sample size fluctuations. To minimize this kind of bias, we standardized the F_{ST} value in the comparison of each pairwise population, accounting for their distinct sample size. A more comprehensive collection of larger numbers of samples across a wider geographic range will be necessary, especially to understand genomics of Malaysian Orang Asli groups, but it must be appreciated that some of these groups are small and on the verge of extinction. Therefore, in reality, large samples per ethnic group are unavailable which probably prevent improving statistical power by increasing sample size.

The second challenge was the lack of suitable methods for the various ethnic groups with complex historical backgrounds. Our estimation of N_e was based on LD pattern of a population. However, the bottleneck effect on the non-African populations might lead to the overestimation of LD before the bottleneck event(s), thus resulting in the underestimation of N_e and divergence time. For example, both archeological and genetic evidence suggested the “out of Africa” occurred around 100,000 years ago (Armitage et al. 2011; Gronau et al. 2011), which is earlier than our estimation (~60 KYA), although the worldwide pattern of population genetic diversity and relationships would not be much affected. Despite the absolute divergence time could be underestimated, our analysis supported that the divergence between Malaysian ethnic groups and East Asians was much more recent. Existing methods for admixture analyses lack sufficient power to provide convincing results in the investigation of those complex groups. In the admixture analysis, on one hand, we did not detect the gene exchange within Malaysia by the *4 Population Test* because these four ethnic groups are so close genetically and we could not find proper reference populations that were totally independent—but not too distant to them—to form a certain tree topology. This challenge may lead to underestimation of the admixture time. Because our estimation is based on the hybrid isolation (HI) model (Jin et al. 2012), which might not be representative of the complex real cases, and the method we used did not take more information, such as admixture proportion, into account, it is likely that real admixture events may have started at an earlier time than that we estimated. As for studying the local adaptations, the complex admixture events happened in the Malaysian ethnic groups may mask the real promising signatures of local adaptation, which could have confounded our identification for those signals. More sophisticated methods of constructing

population histories and deciphering natural selection signatures will assist researchers and improve the understanding of human population diversity in the future.

With respect to genomic data, next generation sequencing (NGS) has allowed genetic studies to address biological questions at a genome-wide scale (Koboldt et al. 2013). Whole-genome sequencing data can be obtained for future investigations into the basis of genetic structure and variance in Asia. Current Malay sequence data in public repositories is from Singapore Malays (Wong et al. 2013) and the original locations for the samples are not available. Thus, it does not make much sense to integrate these data into our current analysis. Our microarray data provided information for us to understand population genomic landscapes of Malaysian ethnic groups, while NGS data can improve resolution in some particular respects in the future.

In summary, these preliminary data provide a comprehensive picture of the population genomics landscape in Malaysia, a strategic area with an important role in human migration history. Natural selection signatures identified here may offer a good candidate list for further replication and verification. Our findings advance the understanding of genetic diversity in Peninsular Malaysia and offer insight into the peopling and human migration history in Southeast Asia. These data may pave the way for future evolutionary and scientific studies.

Acknowledgments The authors would like to thank many of the group members for their helpful discussions and LetPub for its linguistic assistance during the preparation of this manuscript. We thank M. Seielstad for assistance in reviewing the manuscript and his comments. These studies were supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040100), by the National Science Foundation of China (NSFC) grants (91331204 and 31171218). S.X. is Max-Planck Independent Research Group Leader and member of CAS Youth Innovation Promotion Association. S.X. also gratefully acknowledges the support of the National Program for Top-notch Young Innovative Talents of The “Ten-Thousand-Talents” Project and the support of K.C.Wong Education Foundation, Hong Kong. B.P.H., M.E.P. and the Malaysian investigators were supported by research grant Ministry of Science, Technology and Innovation (MOSTI) grant erBiotek Grant # 100-RM/BIOTEK 16/6/2 B (1/2011) and [100-RMI/GOV 16/6/2 (19/2011)]. B.A.Z also received funding from the following research grants: APEX Delivering Excellence 2012 (DE 2012) grant: (1002/PPSP/910343), USM short term grant: (304/PPSP/61311034) and MOSTI (ER-BIOTEK) grant: (304/PPSP/6150113/K105). S.W.S. was supported by the University of Toronto McLaughlin Centre and The Hospital for Sick Children Foundation. S.W.S. holds the GlaxoSmithKline-CIHR Chair in Genome Sciences at the University of Toronto and The Hospital for Sick Children. GSK has no influence on how research funding is allocated and spent. None of the funders had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest The authors have declared that no competing interests exist.

References

- Adams B, Dörfler P, Aguzzi A et al (1992) Pax-5 encodes the transcription factor BSAP and is expressed in B lymphocytes, the developing CNS, and adult testis. *Genes Dev* 6:1589–1607
- Andaya LY (2002) Orang Asli and the Melayu in the history of the Malay Peninsula. *J Malayan Branch R Asiat Soc* 75:23–48
- Ang KC, Leow JWH, Yeap WK et al (2011) Phylogenetic relationships of the Orang Asli and Iban of Malaysia based on maternal markers. *Genet Mol Res* 10:640–649
- Ang KC, Ngu MS, Reid KP et al (2012) Skin color variation in Orang Asli tribes of Peninsular Malaysia. *PLoS One* 7:e42752
- Anney RJL, Lasky-su J, Colm Ó et al (2008) Conduct disorder and ADHD: evaluation of conduct problems as a categorical and quantitative trait in the international multicentre ADHD genetics study. *Am J Med Genet B Neuropsychiatr Genet* 147B:1369–1378
- Armitage SJ, Jasim SA, Marks AE et al (2011) The southern route “out of Africa”: evidence for an early expansion of modern humans into Arabia. *Science* 331:453–456
- Balaresque PL, Ballereau SJ, Jobling MA (2007) Challenges in human genetic diversity: demographic history and adaptation. *Hum Mol Genet* 16:R134–R139
- Benjamin G (1976) Austroasiatic subgroupings and prehistory in the Malay Peninsula. In: Jenner P, Thompson LC, Starosta S (eds) *Austroasiatic Studies*, University of Hawaii, Honolulu, pp 37–128
- Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 84:210–223
- Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci USA* 85:9119–9123
- Clavano-Harding AB, Ambler GR, Cowell CT et al (1999) Initial characterization of the GH-IGF axis and nutritional status of the Ati Negritos of the Philippines. *Clin Endocrinol (Oxf)* 51:741–747
- Dávila N, Shea BT, Omoto K et al (2002) Growth hormone binding protein, insulin-like growth factor-I and short stature in two pygmy populations from the Philippines. *J Pediatr Endocrinol Metab* 15:269–276
- Emelyanov AV, Kovac CR, Sepulveda MA, Birshstein BK (2002) The interaction of Pax5 (BSAP) with Daxx can result in transcriptional activation in B cells. *J Biol Chem* 277:11156–11164
- Endicott P, Gilbert MTP, Stringer C et al (2003) The genetic origins of the Andaman islanders. *Am J Hum Genet* 72:178–184
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14:2611–2620
- Fix AG (1995) Malayan paleosociology: implications for patterns of genetic variation among the Orang Asli. *Am Anthropol* 97:313–323
- Gautier M, Vitalis R (2012) rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* 28:1176–1177
- Gronau I, Hubisz MJ, Gulko B et al (2011) Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet* 43:1031–1034
- Hatin WI, Nur-shafawati AR, Zahri M et al (2011) Population genetic structure of Peninsular Malaysia Malay sub-ethnic groups. *PLoS One* 6:e18312
- Hirschman C (1975) *Ethnic and social stratification in Peninsular Malaysia*. American Sociological Association, Washington, DC
- Ivanov D, Philippova M, Tkachuk V et al (2004) Cell adhesion molecule T-cadherin regulates vascular cell adhesion, phenotype and motility. *Exp Cell Res* 293:207–218
- Jarvis JP, Scheinfeldt LB, Soi S et al (2012) Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genet* 8:e1002641
- Jin W, Wang S, Wang H et al (2012) Exploring population admixture dynamics via empirical and simulated genome-wide distribution of ancestral chromosomal segments. *Am J Hum Genet* 91:1–14
- Koboldt DC, Steinberg KM, Larson DE et al (2013) The next-generation sequencing revolution and its impact on genomics. *Cell* 155:27–38
- Korn JM, Kuruvilla FG, McCarroll SA et al (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* 40:1253–1260
- Lachance J, Vernot B, Elbers CC et al (2012) Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* 150:457–469
- Li JZ, Absher DM, Tang H et al (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104
- Lim LS, Ang KC, Mahani MC et al (2010) Mitochondrial DNA polymorphism and phylogenetic relationships of Proto Malays in Peninsular Malaysia. *J Biol Sci* 10:71–83
- Llaurado JG (1983) Cellular sodium transport and hypertension: a new hypothesis. *West J Med* 139:715–716
- Loh P-R, Lipson M, Patterson N et al (2013) Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193:1233–1254
- Lou H, Li S, Yang Y et al (2011) A map of copy number variations in Chinese populations. *PLoS One* 6:e27341
- Lu D, Xu S (2013) Principal component analysis reveals the 1000 Genomes Project does not sufficiently cover the human genetic diversity in Asia. *Front Genet* 4:1–9
- Magalhães TR, Casey JP, Conroy J et al (2012) HGDP and HapMap analysis by Ancestry Mapper reveals local and global population relationships. *PLoS One* 7:e49438
- Mcevoy BP, Powell JE, Goddard ME, Visscher PM (2011) Human population dispersal “Out of Africa” estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome Res* 21:821–829
- Mendizabal I, Marigorta UM, Lao O, Comas D (2012) Adaptive evolution of loci covarying with the human African Pygmy phenotype. *Hum Genet* 131:1305–1317
- Migliano AB, Vinicius L, Mirazo M (2007) Life history trade-offs explain the evolution of human pygmies. *Proc Natl Acad Sci USA* 104:20216–20219
- Nei M (1987) *Molecular evolutionary genetics*. Columbia University, New York
- Ngamphiw C, Assawamakin A, Xu S et al (2011) PanSNPdb: the Pan-Asian SNP genotyping database. *PLoS One* 6:e21451
- Olivier M (2003) A haplotype map of the human genome. *Physiol Genomics* 13:3–9
- Perry GH, Dominy NJ (2009) Evolution of the human pygmy phenotype. *Trends Ecol Evol* 24:218–225
- Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8:e1002967
- Price AL, Patterson NJ, Plenge RM et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909
- Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
- Qian W, Deng L, Lu D, Xu S (2013) Genome-wide landscapes of human local adaptation in Asia. *PLoS One* 8:e54224
- Reich D, Thangaraj K, Patterson N et al (2009) Reconstructing Indian population history. *Nature* 461:489–494

- Reich D, Patterson N, Kircher M et al (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* 89:516–528
- Sabeti PC, Varilly P, Fry B et al (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913–918
- Saw SH (1988) *The population of Peninsular Malaysia*. Singapore University, Singapore
- Shriver MD, Kennedy GC, Parra EJ et al (2004) The genomic distribution of population substructure in four populations using 8, 525 autosomal SNPs. *Hum Genomics* 1:274–286
- Sonuga-barke EJS, Lasky-su J, Neale BM, et al (2008) Does parental expressed emotion moderate genetic effects in ADHD? An exploration using a genome wide association scan. *Am J Med Genet B Neuropsychiatr Genet* 147B:1359–1368
- Sulem P, Gudbjartsson DF, Stacey SN et al (2007) Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* 39:1443–1452
- Tamura K, Peterson D, Peterson N et al (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28:2731–2739
- Thangaraj K, Singh L, Reddy AG et al (2003) Genetic affinities of the Andaman islanders, a vanishing human population. *Curr Biol* 13:86–93
- Thangaraj K, Chaubey G, Kivisild T et al (2005) Reconstructing the origin of Andaman Islanders. *Science* 308:996
- The 1000 Genomes Project Consortium (2011) A map of human genome variation from population scale sequencing. *Nature* 467:1061–1073
- The DASH-Sodium Collaborative Research Group (2001) Effects on blood pressure of reduced dietary sodium and the dietary approaches to stop hypertension (DASH) diet. *N Engl J Med* 344:3–10
- The HUGO Pan-Asian SNP Consortium (2009) Mapping human genetic diversity in Asia. *Science* 326:1541–1545
- The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437:1299–1320
- The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861
- Weir BS, Hill WG (2002) Estimating F-statistics. *Annu Rev Genet* 36:721–750
- Wong L-P, Ong RT-H, Poh W-T et al (2013) Deep whole-genome sequencing of 100 southeast Asian Malays. *Am J Hum Genet* 92:52–66