European Journal of Human Genetics (2018) 26:886–897
https://doi.org/10.1038/s41431-018-0120-8

ARTICLE

A genome-wide characterization of copy number variations in native populations of Peninsular Malaysia

Ruiqing Fu[1,2] · Siti Shuhada Mokhtar[3] · Maude Elvira Phipps[4] · Boon-Peng Hoh[1,5] · Shuhua Xu[1,2,6,7]

Received: 26 July 2017 / Revised: 20 November 2017 / Accepted: 1 February 2018 / Published online: 23 February 2018

Abstract

Copy number variations (CNVs) are genomic structural variations that result from the deletion or duplication of large genomic segments. The characterization of CNVs is largely underrepresented, particularly those of indigenous populations, such as the Orang Asli in Peninsular Malaysia. In the present study, we first characterized the genome-wide CNVs of four major native populations from Peninsular Malaysia, including the Malays and three Orang Asli populations; namely, Proto-Malay, Senoi, and Negrito (collectively called PM). We subsequently assessed the distribution of CNVs across the four populations. The resulting global CNV map revealed 3102 CNVs, with an average of more than 100 CNVs per individual. We identified genes harboring CNVs that are highly differentiated between PM and global populations, indicating that these genes are predominantly enriched in immune responses and defense functions, including *APOBEC3A_B*, beta-defensin genes, and *CCL3L1*, followed by other biological functions, such as drug and toxin metabolism and responses to radiation, suggesting some attributions between CNV variations and adaptations of the PM groups to the local environmental conditions of tropical rainforests.

These authors contributed equally: Ruiqing Fu, Boon-Peng Hoh.

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41431-018-0120-8) contains supplementary material, which is available to authorized users.

✉ Shuhua Xu
xushua@picb.ac.cn

1 Chinese Academy of Sciences (CAS), Key Laboratory of Computational Biology, Max Planck Independent Research Group on Population Genomics, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Shanghai 200031, China

2 University of Chinese Academy of Sciences, Beijing 100049, China

3 Institute of Medical Molecular Biotechnology, Faculty of Medicine, Universiti Teknologi MARA, Sungai Buloh Campus, Selangor, Malaysia

4 School of Medicine, Monash University Sunway Campus, Petaling Jaya, Malaysia

5 Faculty of Medicine and Health Sciences, UCSI University, Jalan Menara Gading, Taman Connaught, Cheras, Kuala Lumpur, Malaysia

6 School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China

7 Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

Introduction

Copy number variations (CNVs) are major genomic structural variations (SVs) spanning large regions of chromosomes, ranging from several hundred base pairs to several megabases [1–3]. CNVs may exert profound effects on phenotype, primarily via gene dosage effects [4–7], create novel fusion genes, alter the distance of a gene from a regulatory element, or modify the number of protein-coding exons within a gene [8]. These genetic variations have been associated with a number of systemic autoimmune diseases [9–11], neuropsychiatric disorders [12–14], as well as other complex disorders [15–17]. Owing to its recombination mechanisms [18–20], CNVs are generally observed as loss or gain of copies of a certain DNA segment compared to the reference genome. This alteration in copy number often harbors genes that are responsive towards environmental factors, such as the Amylase gene (*AMY*) [6] and beta-defensin (*DEFB*) [21]. Therefore, CNVs are likely to be subjected to natural selection [6, 22, 23].

Malaysia is a multi-ethnic, linguistic, and cultural country. The Malays (MLY) comprise the major community, accounting for 2/3 of the population, whereas the indigenous populations (locally known as "Orang Asli") comprise only ~0.6% of the total population in Malaysia

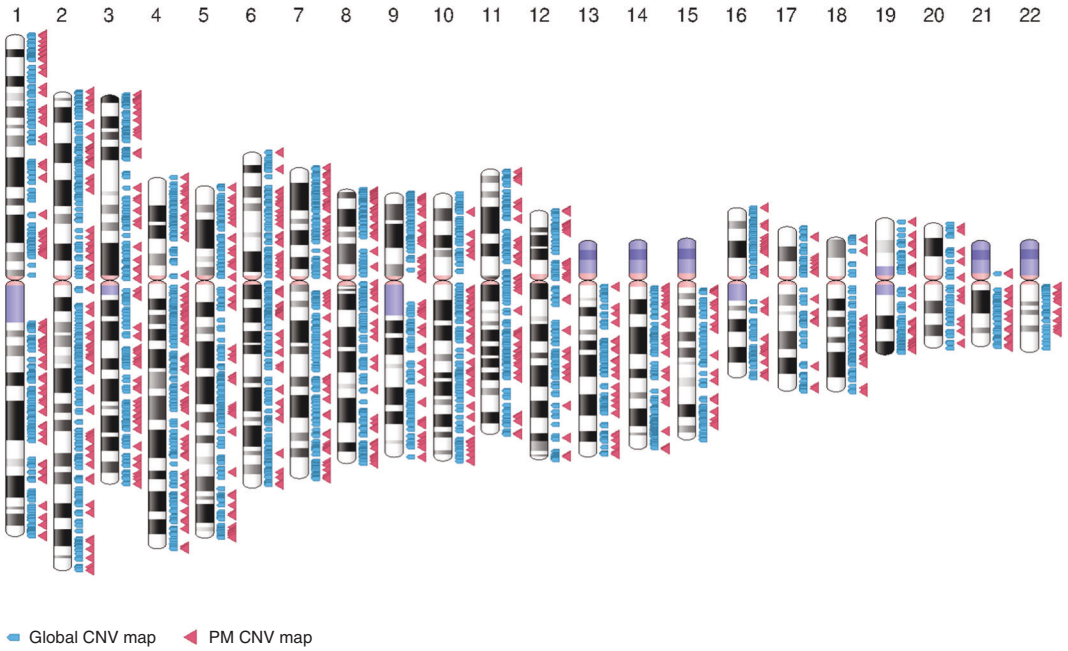

Fig. 1 CNV map of the native populations of Peninsular Malaysia (PM). The map was generated based on (i) the data set from PM populations only; and (ii) the four PM indigenous populations and along with the publicly available global HapMap populations

[24]. The Orang Asli populations are generally categorized into three main tribes; namely, Negrito (NGO), Senoi (SNI), and Proto-Malay (PML). The NGOs are phenotypically similar to, but genetically distinct from, the African Pygmies. Most NGOs still live as "semi-" hunter-gatherers in remote areas. The SNIs are traditionally believed to be a slash-and-burn farming community, living both plantation and hunter-gathering lifestyles, whereas the PMLs are mainly farmers and workers. Both SNIs and PMLs have recently undergone different levels of urbanization. Numerous genome-wide SNP genotyping studies have indicated that these indigenous communities may have experienced long periods of isolation [24–27]; however, the impact of such a process on the CNV architecture of these populations has been lacking.

While the majority of the populations in most parts of the world have been sampled, those in Southeast Asia (SEA) have rarely been included in such population studies [23]. To date, efforts in mapping CNVs of Southeast Asian populations is limited [28, 29], despite their contribution to genetic variation in the region, particularly the natives from Peninsular Malaysia (PM). Therefore, the present study aimed to construct a comprehensive CNV map for the four native populations of PM: MLY, NGO, SNI, and PML. Notably, we have included all six NGO sub-tribes from PM in this analysis; namely, Beteq, Mendriq, Kensiu, Jehai, Kintak, and Lanoh. The population structures of the native populations and the impact of local adaptation in shaping their genomes were assessed using the mapped CNVs.

Results

Characterization of CNVs

In this analysis, all calls for copy number polymorphisms (CNPs) and rare CNVs were integrated. Two sets of CNV maps were constructed, namely, (i) the four PM indigenous populations and along with the publicly available global HapMap populations (denoted herein as "global data set"); and (ii) the Peninsular Malaysia populations only (denoted herein as "PM data set").

The CNV calls were annotated to the human genome reference hg18. A total of 3102 copy number variable regions (CNVRs) were obtained from the global data set (denoted herein as global CNV map; Table S1), and 929 from the PM data set (denoted herein as PM CNV map; Table S2), respectively. On average, >100 CNVR events per individual were detected (Figure S1 and Table S3). The African population harbored most of the CNVRs (156 CNVR per individual), in agreement with the recently published CNV map [23], whereas the populations from PM harbored the lowest number of CNVRs (117–137 CNVRs per individual). This discrepancy may be attributed to lower CNV diversity in the PM populations, although we do not rule out the possibility of ascertainment bias. Within the four PM populations, the highest and lowest CNV loadings were observed in PML and NGO, respectively (141 CNVRs vs. 120 CNVRs per individual from the PM map). The vast majority (~80%) of the CNV events were

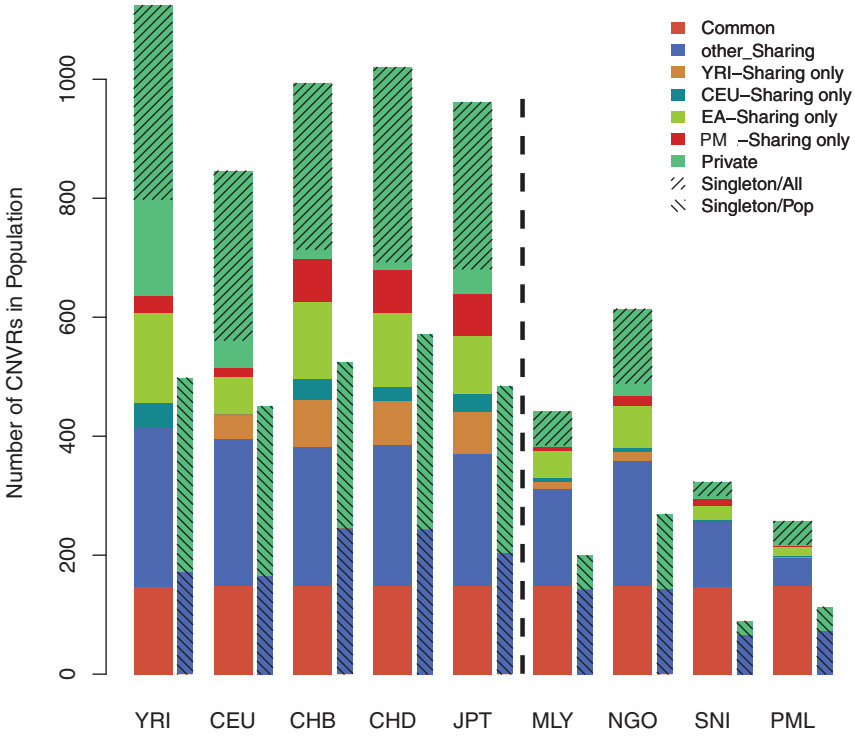

Fig. 2 CNVR sharing among all populations. CNVRs in each population are divided into different parts: common in every population (red, bottom), shared only with YRI (dark gold), CEU (forest green), East Asian populations (grass green), and Peninsular Malaysia populations (dark red), population-private (light green in the top), and others (blue). The thin bar represents the singletons (Pop) within each population (shadowed by the 135° sloping lines) and the singletons (All) counted by across all the samples are shadowed by the 45° sloping lines

deletions. The CNVs were mapped to autosomes (Fig. 1). The global CNV map covers 223.8 Mb (~7.8%) of the autosomal genome, with an average size of 72 kb (median: 18 kb), whereas the PM CNV map encompasses 59.8 Mb (~2.08%) of the autosomal genome, with a mean size of 64 kb (median: 16 kb). In contrast, on average, less than 1% of each individual genome was covered by CNVs (ranging from 0.48 to 1.5% for the global map). Compared to the Database for Genomic Variants (DGV) (dated July 23, 2015), the reproducibility of our CNVRs against the DGV records gradually decreased as the stringency of the overlapping criteria increased [30] (Figure S2).

The majority (>75% and >60% of the two CNV maps, respectively) of the CNVRs had low frequency (<1% for the global map and 5% for the PM map, considering that the total sample size is <100 for the PM populations) (Figure S3). We thus exercised caution as the number of available samples per population has restricted us to define "rare" CNVs (i.e., frequency <0.1%) in our CNV map (Figure S3), particularly for the PM CNV map. Furthermore, considering the small sample size of the PML tribe, we did not remove singletons from the CNV map.

Shared CNVs between populations

To elucidate the relationships among populations, CNVs that were shared between populations were assessed (Fig. 2). A total of 111 CNPs and 1740 CNVRs that were found in the global populations data set occurred only once in the PM populations, whereas 157 CNPs and 148 CNVRs were observed in all populations studied (herein denoted as common CNVs). Approximately 47% of the CNVs and 57% of the CNPs called were shared by at least two populations. On a separate note, the "population-private" CNVs, defined as CNVs that were observed only in one population (top green part of Fig. 2 and S4A), accounted for a substantial proportion of the CNVs in each population because: (i) the rare CNVs identified in this study were more likely to be detected only within a particular population; and (ii) the population-private CNVRs were mainly singletons (shadowed with 45° sloping lines in Fig. 2). Africans were found to harbor the largest number of such private CNVs, which is in line with the observations of Sudmant [23]. We also observed that the singletons (i.e., CNVs that occurred in only one sample in a particular population) within each population (the thin bar with 135° sloping lines in Fig. 2 and S4A) accounted for ~46% and 25% of the total number of CNVRs and CNPs for each population, respectively, many of which occurred once in a population but were common in other populations (thin blue bar in Fig. 2 and S4A). The East Asian (EA) populations (i.e., CHB, CHD, and JPT) shared relatively more reciprocal CNVs that were restricted to the PM populations and vice versa, suggesting that these populations are genetically closer than other populations. Within the four PM populations, the highest pairwise sharing of CNVs was detected between NGO and MLY (Figures S4B-S4D), possibly

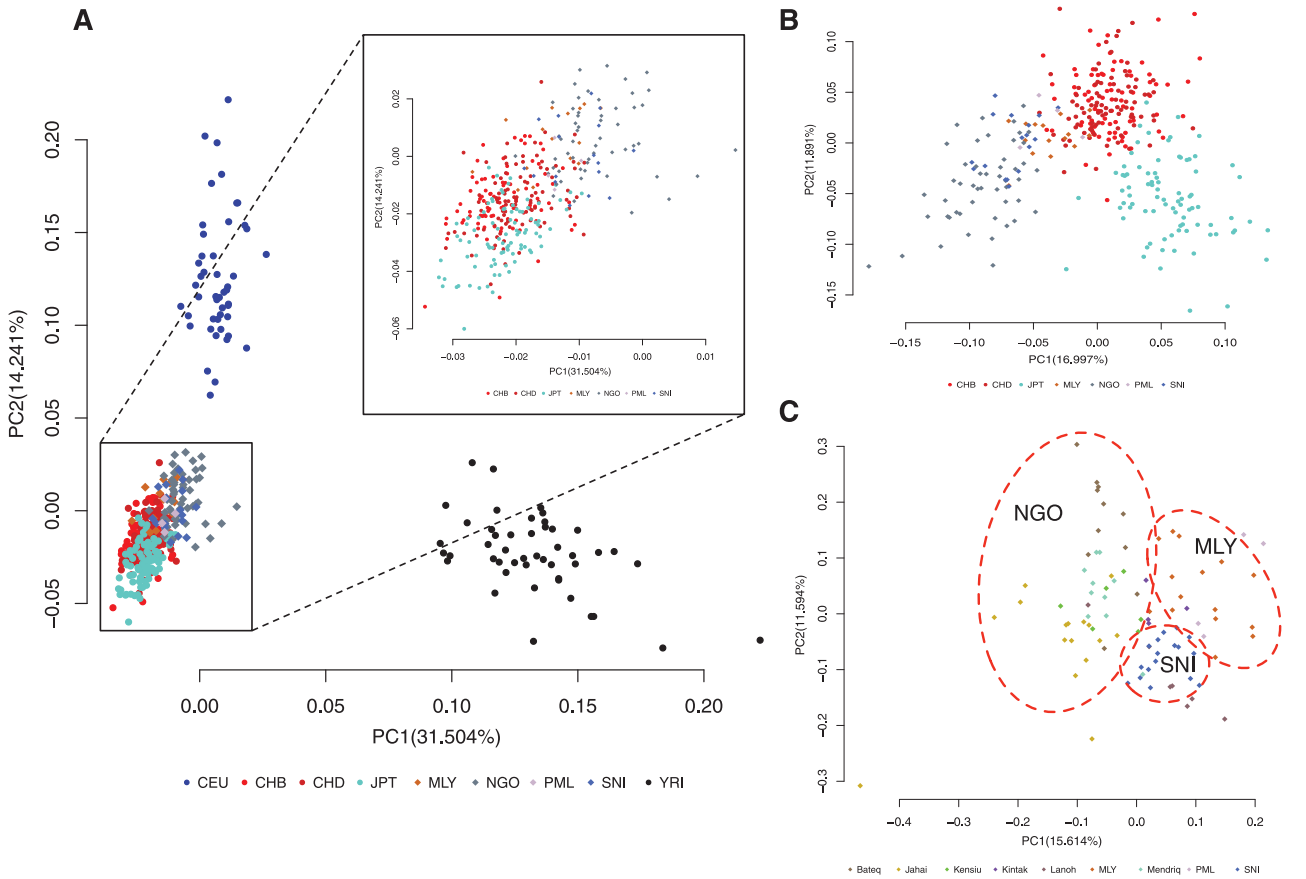

Fig. 3 Principal Component Analysis (PCA) using bi-allelic CNV (BiCNVs). Three independent PCAs were performed with different population data sets **a** global populations. The blow-up section shows the relationships between the East Asian populations; **b** Asian populations; **c** native populations of Peninsular Malaysia

because of the higher number of CNVs detected in these two populations (Fig. 2 and S4A).

Population structure and genetic relationships

The fine-scale population genetic structure and corresponding relationships were then assessed based of the generated CNV map. We conducted the principal component analysis (PCA) [31] using the shared biallelic CNVs (biCNVs) ($N = 878$) among the populations. As expected, the populations from three main continents were well separated along PC1 (Africans and non-Africans) and PC2 (Europeans and Asians). The PM populations were clustered close to the EA populations (Fig. 3a). We then looked at EA and PM populations and found three distinct clusters, namely, the Japanese Tokyo (JPT), HAN Chinese (CHB +CHD), and PM populations. MLY was clustered between the EA populations and the PM Austro-Asiatic populations, i.e., the NGO and the SNI (Fig. 3b), suggesting that these urban Malays may be genetically admixed between the EA and their neighboring natives. We then further zoomed into the genetic relationships of the four PM populations. With the 878 biCNVs identified, the Malays represented a distinct population separate from the NGOs and the SNIs, with a few outliers (Fig. 3c). The PCA also exhibited sub-structures within the six sub-tribes of Negritos (Figure S5). This is likely due to the reason that these indigenous hunter-gatherers may have been isolated from other populations for a long period of time, thereby resulting in a distinct genetic background, and that sub-structures might have occurred within the NGOs.

By including the commonly shared biCNVs among the studied populations, we performed STRUCTURE analysis [32–35] with K = 2–10 (with five replicates). Distinct clusters appeared as K increased (Fig. 4 and S6). When K = 3, the PM populations appeared to carry the same ancestral component as the EA populations, with some small amount of admixture of the African and European ancestral components (Figure S6). At K = 4, the distinguishable Austro-Asiatic component (here referring to NGO and SNI) appeared, and when K = 5, JPT was separated from the HAN Chinese (CHB and CHD). Collectively, this implies that the NGO and SNI are genetically unique,

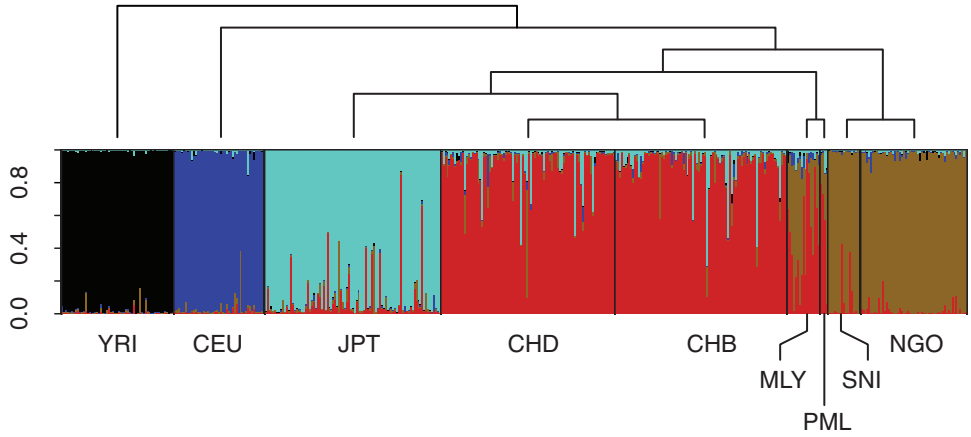

Fig. 4 Neighbor-joining tree and STRUCTURE analyses. The NJ-tree based on F_{ST} (upper panel) values is in agreement with the results of STRUCTURE analysis (K = 5; bottom panel)

whereas the MLY and PML exhibited admixed patterns, in agreement with the PCA results (Table S4).

Subsequently, population differentiation was measured using the unbiased F_{ST} statistic [36]. The pairwise population F_{ST} values were computed after CNV allele frequencies were inferred using an EM algorithm [37]. A neighbor-joining (NJ) tree was constructed with the F_{ST}-based distance [38, 39] (Fig. 4). The pattern of the clades coincided with the results of STRUCTURE analysis, and an earlier study using SNP genotyping [26]. The NJ tree revealed three major branches, with YRI as outlier. The EA nodes revealed that MLY and PML were more closely related to the EA populations, thereby suggesting some degree of gene flow from the EA populations, whereas NGO and SNI formed a unique clade distinct from the MLY and EA populations.

Linkage disequilibrium between CNVs and SNPs

A previous study has shown a decrease in linkage disequilibrium (LD) between CNPs and the flanking SNPs in Chinese populations compared to the European population [37]. However, whether the PM populations exhibit the same trend remains unclear. Thus, we investigated the 'taggability' (i.e., measuring the CNVs that are in high LD with flanking SNPs) of SNPs to CNPs in the PM populations. We exercised caution in our analysis as the small sample size of each PM population may result in large variations in LD. Therefore, we gathered the SNI, PML, and MLY as a group (denoted as SEA in the figures), and NGO as another. CN deletions that showed strong correlation ($r^2 \geq 0.8$) with the flanking SNPs (in either 3 Mb or 5 Mb searching windows) exhibited generally similar frequencies across populations, ranging from 58.36% (CEU), 56.86% (NGO), and CHB (54.87%) to SEA (47.76%), which was in agreement with the findings of a previous study [37] (Table S5A). Stronger taggibility of CN deletions has also been earlier reported [23].

In contrast to deletions, CEU had the lowest number of CN duplications that exhibited strong correlations with the flanking SNPs, possibly due to the reason that CEU has fewer CN "bi-duplications" (about 50% lower than those in CHB) (Table S5A). When calculating for CNVRs, CEU again showed the highest taggability among the investigated populations, even with a relatively smaller number of CNVRs (Table S5B).

Population-specific CNVs

Population isolation and forces such as natural selection and genetic drift may result in population-specific genomic variants. Therefore, we identified population-specific CNVs that occurred exclusively in the four PM populations or those that showed significantly higher frequencies than the global populations studied.

The number of CN deletions that were highly differentiated between PM (except PML) and the HapMap populations decreased when the geographical distance of the populations decreased (Fig. 5 and S8). Two CN duplications had significantly higher frequencies in SNI (CNP11172; chr7:g.138438_163742; 68.75% in SNI) and NGO (CNVR2705; chr16:g.(68534936_68545887)_(68818240_68824276); 45.45% in NGO) (Table S6). When the three EA populations were pooled together, more PM population-specific CNVs were revealed (Table S6). We acknowledge that the PM private CNVs were also included in the analysis, although they were primarily singletons (Table S7).

The "population-specific" CNVs identified showed high concordance with the records in DGV, implying the reproducibility of our CNV calls. We suspect that genetic drift or environmental pressure could have in part played a role in resulting the emergence of highly differentiated

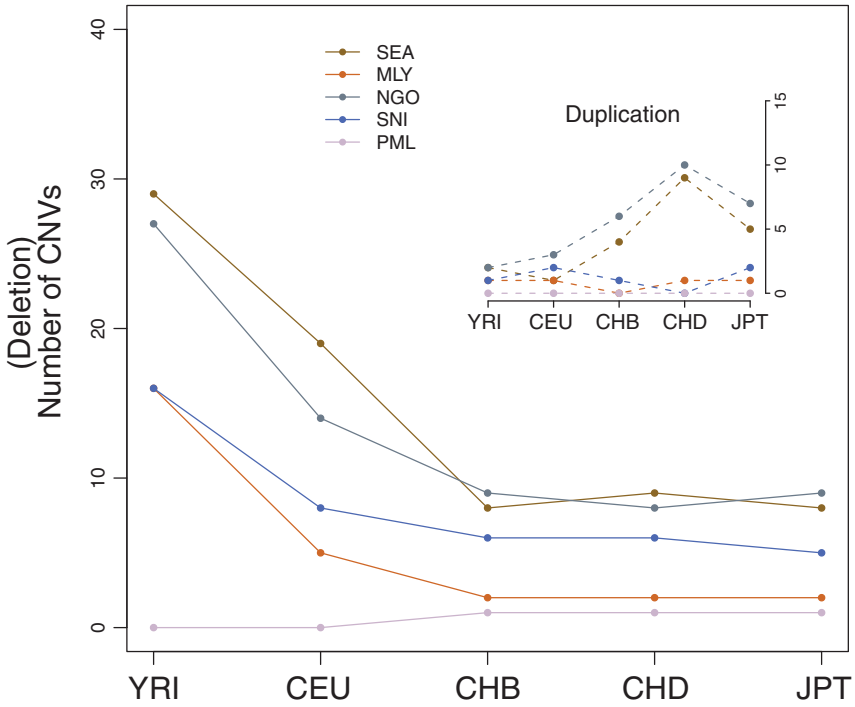

Fig. 5 Number of population-specific CNVs across all populations. The number of population-specific deletion CNPs between PM populations and HapMap global populations decreases as the population geographic distance decreases

CNVs in the PM populations compared to that in other global populations. In contrast, less than half of the "population-private" CNVRs could be found in DGV when a 50% reciprocal overlap criterion was applied, suggesting that there may be novel alleles that have yet to be detected, and thus complementing the enrichment of global CNV surveys.

Signatures of local adaptation

A recent study suggested that population-specific CNVs may be attributed, at least in part, to recent local adaptation of the hunter-gatherer Negritos [28]. We therefore explored further supporting evidence of signatures of local adaptation on the CNVs in our samples. The unbiased F_{ST} [36] was used to scan the genome-wide CNVs between pairs of populations from PM and global populations in this study (as reference), respectively. Then, the top 1% loci of the F_{ST} values of each pair were listed as candidates regions of local adaptation of the PM populations. Candidate genes underlying the putative regions of local adaptations were then annotated with RefSeq (date 08/2013) (Table S8). Annotations and the enrichment analyses were performed on the candidate genes using DAVID (version 6.7) (https://david.ncifcrf.gov/).

Because the number of genes overlapping CNPs was limited (Table S8A), we only assessed the enrichments in NGO, as well as the pooled PM populations (SEA). The enrichment scores were 6.85 and 5.38, respectively. The top enriched items referred to "defensin" (p values after Benjamini correction were 4.7e-12 and 7.2e-10, respectively; and the same for the following), "antibiotic" ($p = 6.6e-11$ and 1.0e-8), "antimicrobial" ($p = 5.8e-11$ and 8.7e-9), and "defense response to bacterium" (6.0e-9 and 1.6e-6). This is in agreement with the fact that the hot and humid climate of the tropical rainforest is preferred by various parasites and pathogens. Therefore, it is likely that the Orang Asli, particularly the hunter-gatherer Negritos, were exposed to various stresses against infections. As for CNVRs, which included both common CNPs and rare/de novo CNVs (Table S8B), the gene cluster involving "defensin" or response to bacterium remained at the top of the list (with enrichment scores>3 and p values<0.001 after Benjamini correction) in all PM populations. In addition, in SNI, genes related to sensory perception were identified, e.g., olfactory receptor (OR), *EYS*, and the family of the taste receptor protein (*TAS2Rs*).

The underlying genes in the population-specific CNVs were found to be enriched in the category "response to antimicrobiome" for both the NGOs and the SNIs (Table S6 and S7). Several candidate genes had drawn our attention (Table 1). Notably, *APOBEC3A_B*, a fusion of the *APOBEC3A* and *APOBEC3B* genes due to a deletion of the genomic region linking them, was reported to contribute to immunity by restricting the transmission of foreign DNA. A recent study suggested that the deletion of *APOBEC3B* is strongly associated with susceptibility to malaria [40]. Furthermore, the beta-defensin gene family, including

Table 1 Candidate genes harbouring the CNVRs with signals of local adaptation.*, Candidate genes derived from CNPs

Genes	F_{st}				Population specific CNV				Gene annotation
	MLY	NGO	SNI	PML	MLY	NGO	SNI	PML	
	YRI/CEU/CHB/CHD/JPT	YRI/CEU/CHB/CHD/JPT	YRI/CEU/CHB/CHD/JPT	YRI/CEU/CHB/CHD/JPT	YRI/CEU/CHB/CHD/JPT	YRI/CEU/CHB/CHD/JPT	YRI/CEU/CHB/CHD/JPT	YRI/ CEU/CHB/CHD/ JPT	
EYS	–/0.2596/0.2729/0.3853/–	—	0.5866/0.434/–/–/–	0.5692/0.6559/–/0.7874/–	—	1/1/0/0/0	1/1/0/0/0	–	sensory perception, cognition
APOBEC3B, APOBEC3A_B	—	—	0.4134/ 0.2962/–/–/–	0.6758/0.5211/–/–/–	—	1/1/0/0/0	1/1/0/0/0	–	immune response, associated with susceptibility to malaria
KCNIP4	—	—	–/0.6125/ 0.3825/0.386/0.3433	–/0.4251/–/–/–	—	0/1/0/0/0	0/1/1/1/1	–	ion transport, K-Ca channel
GSTM1, GSTM2	–/–/0.9217/0.9223/–	–/–/ 0.5697/0.5715/–	–/–/0.8943/0.8952/–	–/–/0.9496/0.9501/–	0/0/1/1/0	0/0/1/1/0	0/0/1/1/0	0/ 0/ 1/ 1/ 0	drug metabolism
CES1[a]	–/–/–/0.4707/0.4626	—	0.6878/–/–/0.7054/6986	—	1/0/0/1/1	1/0/0/0/0	1/0/0/1/1	–	response to toxin; drug metabolism
DPP6	0.4032/–/–/–/–	0.5136/–/–/–/–	0.3687/–/–/–/–	—	1/ 0/ 0/ 0/ 0	1/0/0/0/0	1/0/0/0/0	–	associated with amyotrophic lateral sclerosis (ALS)
TRIM16	—	—	0.3525/0.434/0.2993/0.3641/–	—	—	—	1/1/1/1/0	–	response to nutrient and toxin
PRB1	—	–/–/–/–/ 0.1974	—	—	—	0/0/1/0/1	—	–	related to human salivary glycoproteins
PRODH	—	–/–/0.2022/0.2042/0.2157	—	—	—	0/0/1/1/1	—	–	associated with hyperprolinemia type 1 and susceptibility to schizophrenia 4 (SCZD4)
DGCR6,DGCR9,DGCR5	—	–/–/0.2022/0.2042/0.2157	—	—	—	0/0/1/1/1	—	–	DiGeorge syndrome region
NOMO1[a]	–/–/0.429/0.5096/0.4703	–/–/0.451/0.5236/0.4879	–/–/ 0.2793/0.3671/0.3244	—	—	0/0/1/1/1	—	–	participating in the Nodal signaling pathway during vertebrate development
DEFB4A, DEFB103A, DEFB103B, DEFB104A, DEFB104B, DEFB105A, DEFB105B, DEFB106A, DEFB106B, DEFB107A, DEFB107B	–/–/0.2076/0.8232/0.667	–/–/0.3761/0.8097/0.7207	–/–/0.2725/0.8575/0.7169	–/–/–/0.9444/0.7855	—	0/0/0/1/1	0/0/0/0/1	–	Beta defensin, antibiotic, antimicrobial
OR4M2, OR4N3P, OR4N4	—	–/–/–/–/0.5107	–/–/–/–/0.3302	—	—	0/0/0/0/1	—	–	olfactory receptor activity
SPAG11A, SPAG11B	–/–/0.2076/0.8232/0.667	–/–/0.3761/0.8097/0.7207	–/–/0.2793/0.3671/0.3244	—	—	0/0/0/1/1	0/0/0/0/1	–	spermatogenesis, defense response to bacterium
GNLY	—	–/–/–/–/0.1918	—	—	—	0/0/0/1/1	—	–	

Table 1 (continued)

Genes	F_{st}				Population specific CNV				Gene annotation
	MLY YRI/CEU/CHB/ CHD/JPT	NGO YRI/CEU/CHB/ CHD/JPT	SNI YRI/CEU/CHB/ CHD/JPT	PML YRI/CEU/CHB/ CHD/JPT	MLY YRI/ CEU/ CHB/ CHD/JPT	NGO YRI/ CEU/ CHB/ CHD/JPT	SNI YRI/ CEU/ CHB/ CHD/JPT	PML YRI/ CEU/ CHB/ CHD/ JPT	
									defense response to bacterium and fungus
CD8A, CD8B	—	–/–/–/0.1918	—	—	—	0/0/0/1/1	—	–	immune response
CCL3L1, CCL3L3, CCL4L1, CCL4L2	—	–/–/0.2241/ 0.3923/–	—	—	—	0/0/1/1/0	—	–	defense response, inflammatory response, response to wounding and virus
MAT2A	—	–/–/–/0.1918	—	—	—	0/0/0/1/1	—	–	response to radiation
GGCX	—	–/–/–/0.1918	—	—	—	0/0/0/1/1	—	–	blood coagulation, wound healing
GLYAT	–/0.3813/ –/0.1948/–	—	—	—	0/1/0/0/0	—	—	–	response to toxin

[a]Candidate genes derived from CNPs

DEFB4B, *DEFB103A*, and *DEFB104B*, encode a group of defensins that respond to bacterial invasion. In addition, the chemokine gene *CCL3L1*, which is associated with susceptibility to HIV-1 infection and autoimmune diseases, is utilized as a signal for local adaptation. CNV duplications harboring the *CCL3L1* gene have occurred across all the studied populations, with the following frequencies in specific populations: CEU (97.87%), YRI (91.38%), NGO (78.18%), JPT (42.86%), CHB (38.20%), SNI (29.41%), PML (25.00%), and MLY (17.65%), thereby suggesting that the Negritos may have exerted a different pressure on this gene, compared to the other Peninsular Malaysians. CNVs that overlapped with the *CES1* gene were significantly differentiated in MLY and SNI, and this gene is responsible for the hydrolysis or transesterification of various xenobiotics and drug clearance. Besides, a gene related to responses to radiation, *MAT2A*, was also observed, and may be attributed to the exposure to sunlight (hence UV exposure) in the tropical region, which is an important environmental stress in tropical rainforests.

Discussion

The complex histories and multiple inflows of populations in PM and the environmental pressures due to the distinct climate of the tropical rainforest have essentially shaped the unique genetic diversity of the human populations in this region. Although the most comprehensive global CNV map has been reported, numerous regions remain unexplored, e.g., natives of PM [23]. This study has narrowed the gap of the current CNV map. Approximately 72% of the CNVs found in the PM map are novel.

One advantage of this study is that both populations, as well as samples are identical to our earlier study [26], which revealed the population genetic structure of the PM populations using SNP data. Therefore, the findings of this study can be thoroughly assessed and verified. Although different scales of genomic variants would capture different genomic information, the results of population genetic analyses using SNPs and CNV show high concordance. The STRUCTURE analysis demonstrates that a small proportion of the European component of the Malay population ($K = 4$) is in agreement with the findings of an earlier study [26], thereby serving as supporting evidence for this observed concordance. Collectively, our analysis supports the idea that the Malay population is likely an admixture population comprising two major ancestral components; namely, East Asian populations and Southeast Asian indigenous populations. In addition, the genetically closer affinity observed between SNI and NGO, as well as between the MLY and EA populations are in agreement with the findings of earlier reports [26, 41].

The taggability of the SNPs and CNVs was relatively higher in CEU than our populations (Table S5), which is in agreement with the results of earlier studies [37]. We agree with the postulation of Lou et al. [37] that higher taggability may be attributable to ascertainment bias in both SNP discovery and array design between Asian and European populations.

We acknowledge that some of the population-specific CNVs identified in this study did not match those identified in a previous investigation [28] (Table S10), which may be due to the following reasons: (i) Mokhtar et al. [28] used a very stringent criterion to construct the CNV set, which included overlapped calls of the 2 out of the 3 algorithms; therefore, some of the true positives might have gone undetected; (ii) the definition of population-specific CNVs differed between the two studies. They defined population-specific CNVs as those present in one but not in other populations, whereas we extended this to CNVs that are of significantly higher occurrence in one population compared to another.

Although selection signals for SNP in various populations have been extensively explored, such analyses for CNVs are relatively scarce. The candidate genes underlying the signals of local adaptation were significantly enriched with genes related to defense against microbes and immune responses and pathways involved in taste transduction, particularly those in NGO and SNI. Essentially, these findings coincide with the physical and environmental attribution of the nomadic lifestyle of the NGO and SNI during prehistoric days, which exposed them to various forms of transmissible diseases and undernourishment; hence, selective pressure may have responded against these challenges. The results of our enrichment analysis are in agreement with the results of Mokhtar et al. [28], in which enriched genes related to immune system processes were also identified. It is interesting to note that although different candidate genes were identified in these two studies, the same biological pathway was identified in both studies, implying that a complex host immune system is essential in tropical rainforests and that positive natural selection would have acted on these PM populations. We investigated whether these putative signals of local adaptation are correlated with those identified by SNPs [26]; however, no correlation was observed ($r^2 \geq 0.5$ and 3 Mb window size). Rather, some of these were mapped to putative CNV regions of local adaptation (Table S11). For instance, a CNV located at chr4:g.(68943659_68964800)_(69234339_69378740) harbored six candidate SNPs that were fixed (i.e., with 100% frequency) in MLY and SNI and that also showed high frequencies in NGO. While this observation further confirms the respective local adaptation signals, we acknowledge that the false-positive signals may be introduced due to the distribution of the probe design; hence, caution should be exercised when interpreting the findings. Considering the limitations of the current CNV data set, further confirmatory analyses are warranted.

Intriguingly only a limited number of candidate genes underlying the signatures of local adaptation were in agreement with the findings of Deng et al. [26] (Table S12). Further investigation revealed that all the CNV signals, except for one, were singletons. However, considering our limited sample size, we do not rule out the possibility that these genes may have been experienced multiple mutation mechanisms (i.e., SNPs and CNVs) in the evolutionary process.

In summary, we constructed the first CNV map of native populations from PM, thereby providing further insights to the CNV landscape of the human genome. We have also reported the first assessment of LD between CNVs and SNPs in the native populations of PM. Several putative candidate genes underlying the CNVs were proposed to have undergone local adaptation. We acknowledge that constraints of sample size and potential ascertainment bias could have confounded our results; however, the results on the CNV characteristics and population genetic structures are in agreement with the findings of earlier reports. Essentially, forces such as natural selection and genetic drift have shaped the unique genomic structure of the PM populations, thereby contributing to the emergence of distinct genomic variants that are involved in specific biological processes. Further investigations on other indigenous populations from Southeast Asia may facilitate in generating a denser CNV map for a better understanding of the evolution of the human genome and its related medical implications.

Materials and methods

Populations and samples

Peripheral blood samples of 100 unrelated individuals comprising 17 MLYs (from Kelantan), 62 NGOs (from Bateq, Mendriq, Kensiu, Jehai, Kintak, and Lanoh), 17 SNIs (from Temiar), and 4 PMLs (from Temuan) were collected from distinct regions of PM. Informed consents were obtained from the participants. All procedures were in accordance with ethical standards of the Research and Ethics committee as approved by the local Ethical Committees and Department of Orange Asli Development (Jabatan Kemajuan Orange Asli, JAKOA) and the Helsinki Declaration of 1975, as revised in 2000.

Genotyping, CNV detection, and quality control

Samples were genotyped with an Affymetrix Genome-wide SNP array 6.0 genotyping platform according to the manufacturer's instruction. CNVs were called using Birdsuite (1.5.5). The CNVs called comprised two major components: (i) copy number of the 1316 pre-defined CNPs (i.e., copy number polymorphisms); and (ii) rare/de novo CNV segments of each individual. Birdsuite also provides results combining the two parts [42]. The HapMap samples, including 58 YRIs (Yoruba in Ibadan, Nigeria), 47 CEUs (Utah residents with Northern and Western European ancestry from the CEPH collection), 89 CHBs (HAN Chinese in Beijing, China), 90 CHDs (Chinese in Metropolitan Denver, Colorado), and 91 JPTs (Japanese in Tokyo, Japan), were also included in this study and independently called using the same tool and same procedure. To minimize the bias introduced by sample size, we conditionally grouped the sub-tribes of NGO together in most of our analyses.

The threshold for QC call-rate was set at 0.80. Seven NGO samples that failed the QC were removed from subsequent analysis. The final data set consisted of 17 MLYs, 55 NGO; 11 Bateqs, 10 Mendriqs, 6 Kensius, 16 Jehais, 5 Kintaks, 7 Lanohs), 17 SNIs, 4 PMLs, 58 YRIs, 47 CEUs, 89 CHBs, 90 CHDs, and 91 JPTs (Table S13). Due to the uncertainty of the performance of Birdsuite on the sex chromosomes, only autosomal CNVs were called. All probe coordinates were mapped to the human genome assembly build36 (hg18).

Generation of the CNV map

One pre-defined CNP, CNP11594 (chr9: g.38906782_65412427), which spans an unusually large region of more than 26.5 Mb across the centromere of chromosome 9, was identified. We believe it is an artifact and thus was excluded from subsequent analyses. A total of 1290 autosomal CNPs were included in the library (Table S14). Then the cut-off value of 0.1 was set as confidence score to decide the true CNP callings, as suggested by Birdsuite.

For integrated CNPs and rare/de novo CNVs, segments with lengths <1 kb, confidence scores <5, or covered by <2 probes were first removed. Then, CNVRs were called by merging these individual CNV segments which consisted two probes that overlapped across individuals [3]. Consequently, a CNV map consisting of 3102 CNVRs was constructed for all the 468 individuals included in this study, which we refer to as the global CNV map, and a CNV map consisting of 929 CNVRs from PM populations, which we refer to as the PM CNV map, respectively.

The CNVs called have been submitted to dbVar with the accession number nstd156.

Population genetic analysis

Analysis of population structures and relatedness was conducted using PCA (EIGENSOFT version 5.0.2) [31] and STRUCTURE (STRUCTURE version 2.3.1) [32–35] of biallelic CNVs. The Expectation-Maximization (EM) algorithm was applied to calculate the allele frequency of the CNVs on the basis of Hardy-Weinberg equilibrium (HWE), followed by unbiased F_{ST} [36] for each pair of populations. The NJ tree was generated using PHYLIP [38], MEGA [39], and Dendroscope [43], respectively, based on the pairwise population F_{ST} values of the CNVs. Considering the limited sample size of PML, we bootstrapped the other 8 populations 100 times to construct a consensus tree using the same procedure (Figure S7A). A simple pairwise distance (i.e., the genotype differences between pairs of individuals, normalized by the total CNVs) between any pair of individuals to build another individual NJ tree was measured (Figure S7B). Missing data were excluded from the analysis.

Analysis of LD between CNVs and SNPs

We assessed for LD between CNVs and flanking SNPs within windows of 3–5 Mb. In this analysis, a biallelic model was applied. The CNVs were categorized into deletions and duplications. Analysis was performed using PLINK [44] (http://pngu.mgh.harvard.edu/purcell/plink/, version 1.07). When searching for the tagging SNPs of CNVs, the r^2 cut-offs were set to 0.2, 0.5, and 0.8 respectively. We were aware that the limited sample size of our study may introduce bias to the results; therefore, during analysis, the samples from PM were classified into (i) NGO and (ii) Peninsular Malaysian (denoted as SEA), which comprised MLY, SNI, and PML.

Population-specific CNVs

Two strategies were applied to define population-specific CNVs. First, deletions or duplications with significantly high frequencies in the test population were identified (Fisher's exact test, p value with Bonferroni correction; 3.8e-5 for CNPs and 1.6e-5 for CNVRs). Second, as a complement to the power deficiency of Fisher's test on the population-private CNVs, the population-private CNVs were also included in the analysis.

Signals of local adaptation and gene enrichment analysis

For all studied populations, the EM algorithm was used to infer CNV allele frequencies, and subsequently calculated the pairwise F_{ST} values for each CNV by comparing the populations from PM with the HapMap populations. Then, the CNVs were ranked according to their F_{ST} values for each pair, and identified the CNVs in the PM populations among the top 1% (Table S8). The population-specific CNVs were also taken into consideration.

Enrichment and functional analyses of the candidate genes harboring the CNVs that were identified as signals of local adaptation were performed using DAVID Bioinformatics Resources 6.7 (http://david.abcc.ncifcrf.gov/).

Acknowledgements We thank the Department of Orang Asli Development (JAKOA) and especially all subjects who voluntarily participated in this study. SX acknowledges financial support from the Strategic Priority Research Program (XDB13040100) and Key Research Program of Frontier Sciences (QYZDJ-SSW-SYS009) of the Chinese Academy of Sciences (CAS), the National Natural Science Foundation of China (NSFC) grant (91331204, 91731303, 31771388, and 31711530221), the National Science Fund for Distinguished Young Scholars (31525014), the National Key Research and Development Program (2016YFC0906403), and the Program of Shanghai Academic Research Leader (16XD1404700). B-PH acknowledges the Chinese Academy of Sciences President's International Fellowship Initiatives (2017VBA0008) awarded to him. This study is also supported by 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)] awarded to B-PH and MEP. SX is Max-Planck Independent Research Group Leader and member of CAS Youth Innovation Promotion Association. SX also gratefully acknowledges the support of the National Program for Top-notch Young Innovative Talents of The "Wanren Jihua" Project. We thank LetPub (www.letpub.com) for providing linguistic assistance during the preparation of this manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004;36:949–51.
2. Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. Science. 2004;305:525–8.
3. Redon R, Ishikawa S, Fitch KR, Feuk L, Redon R, Ishikawa S, et al. Global variation in copy number in the human genome. Nature. 2006;444:444–54.
4. Lupski JR, Stankiewicz P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet. 2005;1:0627–33.
5. Wong KK, deLeeuw RJ, Dosanjh NS, et al. A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet. 2007;80:91–104.
6. Perry GH, Dominy NJ, Claw KG, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 2007;39:1256–60.
7. Lupski JR, Wise CA, Kuwano A, et al. Gene dosage is a mechanism for Charcot-Marie-Tooth disease type 1A. Nat Genet. 1992;1:29–33.
8. Hollox EJ, Hoh B-P. Human gene copy number variation and infectious disease. Hum Genet. 2014;133:1217–33.
9. Fanciulli M, Norsworthy PJ, Petretto E, et al. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet. 2007;39:721–3.
10. Mamtani M, Anaya J-M, He W, Ahuja SK. Association of copy number variation in the FCGR3B gene with risk of autoimmune diseases. Genes Immun. 2010;11:155–60.
11. Molokhia M, Fanciulli M, Petretto E, et al. FCGR3B copy number variation is associated with systemic lupus erythematosus risk in Afro-Caribbeans. Rheumatology. 2011;50:1206–10.
12. Sebat J, Lakshmi B, Malhotra D, et al. Strong association of de novo copy number mutations with autism. Science. 2007;316:445–9.
13. Stefansson H, Rujescu D, Cichon S, et al. Large recurrent microdeletions associated with schizophrenia. Nature. 2008;455:232–6.
14. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M. Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet. 2008;40:880–5.
15. Pollex RL, Hegele RA. Copy number variation in the human genome and its implications for cardiovascular disease. Circulation. 2007;115:3130–8.
16. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437–55.
17. Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. Annu Rev Genet. 2011;45:203–26.
18. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. Pathogenetics. 2008;1:4.
19. Lam HYK, Mu XJ, Stütz AM, et al. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol. 2010;28:47–55.
20. Mills RE, Walter K, Stewart C, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011;470:59–65.
21. Gene D, Asia E, Hardwick RJ, et al. A worldwide analysis of beta-defensin copy number variation suggests recent selection of a high-expressing. Hum Mutat. 2011;67948. https://doi.org/10.1002/humu.21491.
22. Song H, Hu H, Seok I, Chung Y. Identifying copy number variants under selection in geographically structured populations based on F-statistics. Genom Inform. 2012;10:81–7.
23. Sudmant PH, Mallick S, Nelson BJ, et al. Global diversity, population stratification, and selection of human copy-number variation. Science 2015;349:aab3761.
24. Aghakhanian F, Yunus Y, Naidu R, et al. Unravelling the genetic history of negritos and indigenous populations of Southeast Asia. Genome Biol Evol. 2015;7:1206–15.
25. Liu X, Yunus Y, Lu D, et al. Differential positive selection of malaria resistance genes in three indigenous populations of Peninsular Malaysia. Hum Genet. 2015;134:375–92.
26. Deng L, Hoh BP, Lu D, et al. The population genomic landscape of human genetic structure, admixture history and local adaptation in Peninsular Malaysia. Hum Genet. 2014;133:1169–85.

27. Jinam Ta, Phipps ME, Saitou N. Admixture patterns and genetic differentiation in negrito groups from West Malaysia estimated from genome-wide SNP data. Hum Biol. 2013;85:173–88.
28. Mokhtar SS, Marshall CR, Phipps ME, et al. Novel population specific autosomal copy number variation and its functional analysis amongst Negritos from Peninsular Malaysia. PLoS ONE. 2014;9:e100371 https://doi.org/10.1371/journal.pone.0100371
29. Ku C-S, Pawitan Y, Sim X, et al. Genomic copy number variations in three Southeast Asian populations. Hum Mutat. 2010;31:851–7.
30. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. 2014;42:986–92.
31. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.
32. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.
33. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics. 2003;164:1567–87.
34. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. Mol Ecol Notes. 2007;7:574–8.
35. Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour. 2009;9:1322–32.
36. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution. 1984;38:1358.
37. Lou H, Li S, Yang Y, et al. A map of copy number variations in chinese populations. PLoS ONE. 2011;6:e27341 https://doi.org/10.1371/journal.pone.0027341
38. Felsenstein J. PHYLIP (Phylogeny Inference Package)Version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle. Cladistics. 2004;5:164–6.
39. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28:2731–9.
40. Jha P, Sinha S, Kanchan K, et al. Deletion of the APOBEC3B gene strongly impacts susceptibility to falciparum malaria. Infect Genet Evol. 2012;12:142–8.
41. Hatin WI, Nur-Shafawati AR, Zahri MK, et al. Population genetic structure of peninsular Malaysia Malay sub-ethnic groups. PLoS ONE. 2011;6:2–6.
42. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 2008;40:1253–60.
43. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012;61:1061–7.
44. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.