Genomic Insights of Bruneian Malays

Mirza Azmi[1], Lie Chen[1], Adi Idris[2] and Zen H. Lu[1*]

[1] PAPRSB Institute of Health Sciences, Universiti Brunei Darussalam, Jalan Tungku Link, BE1410 Brunei Darussalam

[2]Griffith Centre for Cell and Gene Therapy, Griffith University, Gold Coast Campus, QLD 4222, Australia

*Corresponding author:

Zen H. Lu

Email: zenhuat.lu@ubd.edu.bn

Conflicts of Interest:

All authors hereby declare that they have no conflicts of interest with the works to be published in this manuscript.

Abstract

The Malays and their many sub-ethnic groups collectively make up one of the largest population groups in Southeast Asia. However, their genomes, especially those from Brunei, remain very much underrepresented and understudied. We analysed here the publicly available WGS and genotyping data of two and 39 Bruneian Malay individuals, respectively. NGS reads from the two individuals were first mapped against the GRCh38 human reference genome and their variants called. Of the total ~5.28 million short nucleotide variants and indels identified, ~217K of them were found to be novel; with some predicted to be deleterious and may be associated with risk factors of common non-communicable diseases in Brunei. Unmapped reads were next mapped against the recently reported novel Chinese and Japanese genomic contigs and *de novo* assembled. ~227 Kbp genomic sequences missing in GRCh38 and a partial open reading frame encoding a potential novel small zinc finger protein were successfully discovered. Interestingly, although the Malays in Brunei and Singapore share as many as ~4.38 million common variants, principal component and admixture analysis comparing the genetic structure of the local Malays against other Asian population groups suggested that the Malays in Brunei are genetically closer to some Filipino ethnic groups than the Malays in Malaysia and Singapore. Taken together, our works have provided a first comprehensive insight into the genomes of the Bruneian Malay population.

Introduction

The Malay as an Austronesian ethnic group is made up of various sub-ethnic populations residing mainly in the Southeast Asian countries of Brunei Darussalam, Malaysia, Singapore and Indonesia. Despite them being one of the largest ethnic groups in the region, no Malay genome was included in either the earlier HapMap or the later 1000 Genomes Project [1, 2]. In fact, to date, only a handful of genomic studies have been conducted in Singapore, Malaysia and Indonesia [3-10]. Although valuable information in areas such as population genetic structure, rare or novel genetic variations, pharmacogenomics and genetic disease risk factors have been gained, much remain to be uncovered. This is especially true for the different Malay sub-ethnic groups in Brunei where their genomics landscape remains virtually unexplored till now.

With a population of about 430 thousand, Brunei Darussalam, located on the northern coastline of the Borneo Island, is the smallest country (in term of population size) in Southeast Asia. Approximately 65% of the population in Brunei are Malays, which politically, historically and culturally are made up of seven indigenous groups namely Melayu Brunei, Melayu Tutong, Melayu Belait, Kedayan, Murut, Dusun, and Bisaya. The last three are also found throughout the island of Borneo, including the neighbouring countries of Malaysia and Indonesia. Studies have suggested that they are likely to be genetically related to the Amis in Taiwan who might have migrated to Borneo through the Philippines [6, 10, 11].

Furthermore, a few recent studies have reported genomic sequences which are unique to their specific East Asian populations but missing in the human reference genome [12-16]. Not only have these sequences provided better insights into the population genetic

structure between the Southern and Northern Han Chinese, for example, but they also allowed identification of several novel spliced variant transcripts of genes. Therefore, the existence of novel sequences in the Malay genome is not totally unlikely and together with the genetic variants identified from the various Malay sub-ethnic groups, they should provide a more accurate genetic stratification of the Malay populations.

In this study, the recently published whole genome sequencing (WGS) and genotyping data of several Bruneian Malays were reanalysed. Short variants as well as copy number variations (CNVs) were called and compared with those reported in other Malay groups in the region. Population genetic structure analysis was also conducted by comparing a subset of the single nucleotide variants (SNVs) against several Asian population groups to study the genetic architecture of the Bruneian Malays. Finally, attempts were made to search for novel genomic sequences within the Bruneian Malay genomes. Findings from the present study have for the first time shown the existence of genetic differences between the Bruneian Malays and the other Austronesian Malay groups in Southeast Asia.

Materials and Methods

Subjects

Genomic data of 41 Bruneian Malays used in this study were obtained from two published studies; namely WGS data of two Dusun female individuals generated as part of the Simons Genome Diversity Project, and 730K genotyping data of 20 Dusun and 17 Murut individuals, respectively, from a Southeast Asian population structure study [6, 10] (Table

S1). In addition, the 700K AncestryDNA®'s genotyping data of a local mixed-race Malay-European mother-son duo (private contribution) was also included in the population structure analysis.

Variant calling and annotation

A two-pronged approach was adopted to analyse the genomic data (Figure S1). The quality-assessed WGS reads were first mapped against the GRCh38 human reference genome with decoy sequences [17] using the maximal exact matches (MEM) algorithm of BWA (version 0.7.17) [18]. In addition to the default mapping options, the insert size was set to 400 bp and shorter split and unpaired paired-end reads were both tagged as secondary mapping.

Three different variant callers; namely GATK (version 4.1.4.1) [19], BCFtools (version 1.9) [20], and FreeBayes (version 1.0.2) [21] were then employed to call both SNVs and insertions/deletions (INDELs) from the mapped genomes. The respective GATK's "best practices" workflows for identifying somatic and mitochondrial variants were adopted here. In brief, the mapped reads were name- or position-sorted, mates fixed, duplicates marked, and base calls recalibrated against known variant datasets obtained from 1000 Genomes Project [2], dbSNP (build 154), Singapore Genome Variation Project (SGVP) [3] and Singapore Sequencing Malay Project (SSMP) [4]. The Singaporean variants were converted from the GRCh37- to GRCh38-based coordinates using GATK's "LiftoverVcf" command. Autosomal variants from each of the two recalibrated BAM files were next called using the GATK's "HaplotypeCaller" and jointly genotyped using "GenotypeGVCFs". Quality scores of these variants were subjected to further

recalibration against the same known variant datasets and those obtained from GRCh38-lifted-over HapMap. Variants which did not fit the quality statistics were discarded.

When calling short variants using BCFtools, the “mpileup” command was first used to calculate the likelihood of genotypes at specific positions with the minimum mapping quality and number of gapped reads supporting INDEL set to 30 and 5, respectively. Raw variants were next generated using the “call” command with the multiallelic calling algorithm switched on and the genotype quality (GQ) and posterior probabilities (GP) calculated before filtering away the low-quality ones (QUAL<Q30, GQ<Q30, and DP<10×).

Short variants on reads with mapping quality ≥Q30 were also called from the two BAM files using another haplotype-based variant caller, FreeBayes, running on default parameters. BCFtools was then used to filter the raw variants as previously described.

In order to minimise the limitation of each of the three variant callers [22, 23], a consensus set of variants was generated by intersecting the three filtered VCF files using the “isec” command of the BCFtools. In addition, INDELs longer than 40 bp were also removed to further improve the confidence of the consensus variants.

The genotyping data of the 37 Bruneian Malays were extracted from the larger PLINK-formatted dataset and transformed into a VCF-formatted file using the program PLINK (version 1.9) [24] while a custom script was used to transform the variant tables from AncestryDNA® to their respective VCF-formatted files. GRCh37-based genomic coordinates of all variants were finally lifted as described previously.

All variants called from the 41 Malays' WGS and genotyping datasets were next merged using BCFtools before being annotated against genomic features in various databases using the "annotate" command of BCFtools and AnnoVar [25] (Table S2). Features added included associated dbSNP's accession numbers; ENSEMBL's gene transcripts; GnomAD's allelic frequency of variants; GWAS's and ClinVar's clinical association, e.g. genetic diseases or pharmacogenomics; and potential functional impacts of non-synonymous variants from SIFT and PolyPhen-2. Biological and molecular functions of genes with potential deleterious variants, which were defined as non-synonymous coding variants annotated as 'deleterious' by both SIFT and PolyPhen-2 or classified as 'high impact' by ENSEMBL, were investigated by submitting them to the Gene Ontology online web-server (http://geneontology.org/). On the other hand, mitochondrial non-synonymous variants were uploaded to the MITOMAP web-server (https://www.mitomap.org/) [26] and their functional impacts predicted using APOGEE [27].

In addition, basic variant statistics such as number of SNPs, INDELs, allele frequency, depth distribution, Het/Homo and Ts/Tv ratios were calculated using the statistical function of either BCFtools or RTG tool [28]. Variant density distributed across each chromosome was calculated based on a 1 Mbp sliding window using the SNPdensity command of VCFtool [29] and plotted using the CMplot R package [30].

CNV Calling

cn.MOPs was employed to identify CNVs in the two WGS datasets [31]. Read depths within a 1-Kbp sliding window were first calculated using the "getReadCountsfromBAM" function across each of the 22 autosomes before they were normalised and plotted with the "segplot" function. CNVs meeting the following four criteria, namely (1) read depth

≥10×; (2) spanning across 3 or more segments/windows; (3) log2 value for copy number gains ≥0.8; and (4) log2 value for copy number ≤-2.8; were then extracted and annotated using the AnnotSV (https://www.lbgi.fr/AnnotSV/) web-server [32].

Comparative Genomic Analysis

The Malay variant data files of 89 and 25 genotyped individuals from SGVP and Morseburg's study, respectively, and 96 whole genome sequenced individuals from SSMP were transformed, when necessary, and merged as previously described into two separate VCF files according to the country of origin. These were then compared against variants of the 41 Bruneian Malays using the "isec" command of BCFtools to identify shared and unique variants among the Malays in the three countries. Potential functional impacts of those that were unique to the Bruneian Malays were further investigated as previously described.

To gain more insights into the population genetic structure of the Malays in Brunei, a principal component (PCA) and an admixture analysis were performed on a subset of SNP from 1,499 individuals, including the 41 Bruneians, belonging to 22 different population groups (Table S3) in South, East, and Southeast Asia. As a control, the AncestryDNA®'s genotyping variants of a European relative of the two mixed-race Bruneian Malays were included in both analyses. Again, whenever necessary, SNP files were transformed into GRCh38-based VCF files as previously described.

The PCA was conducted according to a method adopted from a previous study investigating the European population genetic structure [33]. Prior to calculating the principal components of the Asian SNP data, a filtering step using PLINK was performed

to increase the examining resolution of the SNPs while avoiding the removal of excessive number of true positives, *i.e*., SNPs with low linkage disequilibrium (LD). The filtering criteria included (1) missingness rate of ≥10%, *i.e*., below 90% genotyping rate to ensure all the SNPs were comparable across all the different population groups; and (2) genotypic r^2, which is a common measurement for LD, ≥80% within a 50-SNP sliding window increasing at 5 SNP per step to ensure the selected SNPs were all independent from one another as well as non-arbitrary. The first ten principal components of each individual with the filtered SNPs were then measured using the PLINK's command "pca" and those which could explain over 80% of the variance were selected and plotted against each other using the R package, ggplot2 [34].

An "unsupervised" admixture analysis on the same filtered Asian SNP datasets was next conducted using the default settings of an admixture analysis tool, ADMIXTURE [35]. The analysis assumes no prior knowledge on ancestral origin and historical context of the observed genotypic profiles. The number of major ancestral origins (K) in each of the Asian population groups were first estimated and cross-validated before the hypothetical ancestral proportions (Q) of each of the 1,499 individuals were calculated at K=4,5 and 6. A bar plot representing their hypothetical ancestral proportions was subsequently plotted using ggplot2.

Uncovering Novel Sequences in the Bruneian Malay Genome

Reads from the WGS datasets of the two Bruneian Malays which failed to be mapped to the human reference genome were subjected to further investigation in an attempt to search for novel sequences that may be unique to the Bruneian Malay genome. Since the sequenced genomic DNAs were extracted from saliva, some of these unmapped

reads were likely to be of microbial origins and they were removed through two rounds of microbial mapping (Figure S2A). They were first BWA-mapped as previously described against a collection of microbial genomes downloaded from the Human Oral Microbiome Database (HOMD) (http://www.homd.org/ftp/all_oral_genomes/current/). Unmapped reads were then extracted and converted to fastq files using GATK's SamToFastq before been fed into the metagenomics analysis package Kraken2 [36] to further fish out as many of the microbial reads as possible. Kraken2 uses a k-mer-based approach to compare reads against the MiniKraken2 (ftp://ftp.ccb.jhu.edu/pub/data/kraken2dbs/old/minikraken2_v1_8GB_201904.tgz) microbial database to identify those of microbial origins.

The presence of genomic sequences potentially novel to the Bruneian Malay was finally investigated by subjecting the final set of unmapped reads to a workflow involving mapping against East Asian sequences and *de-novo* assembly (Figure S2B). A novel sequence is defined here as one that is missing from the human reference genome.

The recently reported novel ~12 Mbp HX1 Chinese [12] and ~6 Mbp JRG Japanese [15] sequences were chosen as the "reference genomes" for the mapping as comparative genomic analysis here showed that they are genetically the closest population groups, among those publicly available, to the Bruneian Malay. Again, a similar BWA mapping strategy used earlier was applied here. However, the output BAM files were more stringently filtered (>Q30 mapping quality, >30×coverage depth, and mapped region spanning over the length of a read (*i.e.* >100 bp)) to increase the confidence that the sequences were indeed similar to that of the Chinese/Japanese. Variant calling and filtering was next carried out as before using BCFtools' "mpileup"/"call" and "filter" (QUAL

and GQ ≥30) commands. Consensus sequences representative of the Bruneian Malay were then generated using GATK's "FastaAlternateReferenceMaker" which replaced the reference alleles with the filtered variants.

Next, *de novo* assembly of the unmapped reads was performed using a K-mer-based assembler, MEGAHIT [37]. k-mer values ranging between 17 and 31, as recommended by a best K-mer predictor, KmerGenie [38], with an increment of 2 (parameter: --k-min 17 --k-max 31 --k-step 2) were attempted. MEGAHIT by default removes redundant contigs by merging those sharing ≥95% similarity and trims edges with low-coverage (<4×). The newly assembled contigs were then BLASTed with default parameters against the GRCh38 human reference genome to ensure they were of human origin. Sequences were considered human when they shared ≥90% identity with the matched human genome. Those with ≤90% sequence identity were BLASTed once more but against the NCBI's non-redundant (NR) nucleotide database. This time, only contigs with ≥95% identity to human fosmids- and BAC-cloned human sequences were considered likely novel Bruneian Malay sequences.

Contigs derived from the mapping and *de novo* assembly were next BLASTed against each other to remove redundant sequences. The final set of human contigs were further investigated for their gene coding potential by BLASTing their 6-frame translations against the online NCBI's non-redundant protein database.

Results

Genetic variants of the Bruneian Malays

The high-quality WGS paired-end reads from each the two Bruneians could be mapped successfully across >94% the human reference genome at an average depth of 37x and 47x, respectively (Table S4). The remaining ~6% falling mainly within known gaps or challenging chromosomal regions, such as centromere and telomeres (Figure S3).

A combined total of ~5.07, million consensus variants were called by the three variants callers from the two Bruneian Malays (Table 1). Of these, ~217K, consisting mainly of INDELs, were novel and have not been reported in dbSNP. The fact that both Het/Homo and Ts/Tv ratios were within the ranges reported elsewhere indicates that reliable variant callings have been performed [4, 5, 9]. However, the slightly lower Het/Homo ratio in comparison to other population groups seems to suggest lower heterozygosity in the genetics of the Dusun Malays than other Malay groups in the region. Higher homozygosity has in fact been observed in smaller population and less admixed groups [39], which is likely true for the Bruneian Dusun Malays.

The SNPs and INDELs were found to be unevenly distributed throughout the whole genome with an overall mean SNP and INDEL density of 1,498 SNPs/Mbp and 145 INDELs/Mbp, respectively, across the two individuals (Figure 1, Table S5 & S6). A total of 22 and 20 SNP- and INDELS dense regions, *i.e.* ≥2x mean density, respectively, could be clearly seen in the plots. Not unexpected, the human leucocyte antigen (*HLA*) locus (chr6:29-33Mbp) was found to be highly diverse. Interestingly, in addition to the *HLA* locus, there were only two other SNP-dense regions (chr6:5-9Mb & chr16:77-79Mb), which have previously been reported elsewhere [4, 40].

A total of 1,192 CNVs consisting of 392 copy number gains and 859 copy number losses were also detected in the 22 autosomes of the two Bruneian Malays. Sizes of these CNVs

ranged from the minimum 3 Kbp set by the caller to the longer 159 Kbp. 62 of them were found to cover 72 different protein-coding genes with some of them spanning over more than one gene and a number of these genes are known to be potentially associated with diseases such as cancer and cardiovascular diseases (Table S7).

Variants called from the two WGS datasets were next analysed together with that obtained from 39 genotyped individuals. A total of 5,276,758 short variants consisting of 4,776,365 SNPs and 500,393 INDELs were yielded (Table 2). 35,783 of these were protein-coding variants and they could be split almost equally into synonymous and non-synonymous variants. Among the latter variants, 2,094 were predicted to be potentially deleterious impacting 1,718 genes. Gene Ontology analysis revealed that these genes are involved in cellular processes such immune defence, protein modification, transcriptional regulation, cell signalling and molecular transport (Figure 2). It is, however, important to note that deleterious variants with minor allele frequencies <0.05 are found in only small number among the 41 Malay individuals (Table 3). Such variants are predicted to have a higher likelihood of being the risk alleles of diseases [41]. Indeed, such alleles with potential risk association with cancer, heart disease and pharmacogenomics could be identified among the Bruneian Malays.

Distinct admixed genetic structure of the Bruneian Malays.

To gain a better understanding on the population genetics of Bruneian Malays in the context of the wider Asian population groups, *the first ten* principal components (PC) of 583,453 SNPs from each of the 1,499 East, South, and Southeast Asian individuals and a control individual with European ancestry were calculated. A two-dimension PCA plot of the first two *PC*s*, which together could* explain over 80% of the variance*,* shows that

PC1 separate*d* the South Asians from the other population groups while PC2 separate*d* the overlapping East and Southeast Asia (Figure *3A*). The extreme outlier on PC1 *was* the control European individual. The genetically less varied and, hence, the packed South Asian cluster included *various* population groups *residing in* the Indian Subcontinent and other countries. On the other hand, the North-South overlapping between the tightly clustered East Asian groups and the more spread-out cluster of the Southeast Asian*s* indicate*d* a probable flow of genetic materials among the different groups in the region.

Although culturally and linguistically the most similar among all the Malay groups in Southeast Asia, subtle genetic differences among the Malays in Brunei, Singapore, and Malaysia could be inferred from their PCA clustering. Unsurprisingly, the two Bruneian Malay sub-ethnic groups, the Murut and the Dusun, are genetically highly similar to each other as they compactly crowded together within the Southeast Asian cluster. On the other hand, the Malays from Malaysia and Singapore, which themselves formed another compact subcluster, are slightly further apart from the Bruneian Malays. Regardless, the two Malays are not that distinct from each other. In addition, the Malays from Brunei, Singapore, and Malaysia are also genetically closely related to the two Filipino ethnic groups, namely Luzon and Vizaya. Many of them actually overlapped with one another in the PCA plot. Interestingly, as one of the oldest indigenous groups in the Philippines, the Igorots themselves formed another compact subcluster which lied furthest away from the other Southeast Asians.

An “unsupervised” admixture analysis was then conducted under the assumption that the 22 Asian populations are made up of four to six major ancestral groups, *i.e*., K values of 4, 5 and 6. Of the three K values, K=5 exhibited the lowest cross-validation error, though

the difference between each K values was mostly negligible. Therefore, the hypothetical ancestral proportions (Q) of each of the 1,499 individuals, including the control, were first calculated with all three K values. At K=4, differences in the ancestral components between majority of the South Asians and the European control were almost indistinguishable and separation among the different population groups became clearer only at K=5 and above (data not shown). While K=5 has the lowest cross-validation error, K=6 appears to be in best agreement with the findings from PC analysis (Figure 3B).

While K2 constituted the major ancestral components among all the Malay population groups in Brunei, Singapore, and Malaysia, the subtle genetic differences observed in the PCA segregation became clearer in the admixture analysis. The genetic ancestry of the Bruneian Malays is made up largely of two components of K2 and K6, amounting to 91% and 95% in the Dusun and Murut, respectively, and small proportions of K1 and K3. In contrast, majority of the Singaporean and Malaysian Malays share a highly similar genetic admixture pattern containing all six ancestral components, including the European K4 and South Asian K5 ancestries. Unlike their Bruneian cousins, the K2 component of the Singaporean and Malaysian Malays was found to be present in a higher (>57%) proportion while K6 contributed only 26% of the admixed make-up. In fact, the result here suggested that Bruneian Malays actually share a closer genetic ancestry background with the Vizaya and Luzon from Philippines. It is, therefore, clear that although the three Malay population groups in Southeast Asia may share an almost identical culture and language, differences do exist in their genetic structure with the Bruneian Malays having a distinct admixture pattern.

Novel Sequences in the Bruneian Malay Genome

A total of ~93 million unmapped or ~3% of the raw reads from the two WGS datasets were obtained after initial mapping against the human reference and microbial genomes. Of these, ~16 million were mapped to ~146 Kbp sequences making up of ~9.2 and ~137 Kbp novel Chinese and Japanese contigs, respectively. 170 short variants, including both SNPs and INDELs, were called from the mapping; suggesting the potential commonality of these sequences among the three population groups.

Furthermore, although more than 19K contigs with a cumulative length of ~53 Mbp could be assembled from the "leftover" reads, only 58 of them totalling ~82 Kbp were found to share >95% BLAST search sequence identity with either some regions of the GRCh38 human genome or fosmids- and BAC-cloned human sequences, *i.e.,* they are of human origin. The "discarded" contigs were likely to belong to some other yet-to-be verified human, unculturable microbial or misassembled sequences.

The mapping and *de-novo* assembly approaches had, therefore, yielded 227,763 bp novel Bruneian sequences which are missing in the GRCh38 human reference genome. When the gene-coding potential of these sequences were further investigated, an open reading frame out of a total 473 was found to share 92% sequence identity across a 106-residue region with a macaque's hypothetical small zinc finger protein which is predicted to be involved in the homeostasis of zinc ions (Figure S4).

Discussions

Genomic Landscape of Bruneian Malays

The identification of more than 5.2 million variants from 41 Bruneian Malay individuals has enabled, for the first time, their genomic landscape to be studied in detailed and compared against other Malay groups in the region. More than 216K of these variants have never been reported in dbSNP. Crucially, many of them were found to have minor allele frequency <0.001. This subset of rare variants is likely to be unique to the Bruneian Dusun Malay. Similar pattern has also been seen in studies assessing the allele frequency variants in different population groups [42]. Although most of these rare variants are expected to have benign functional impacts, a few may in fact be the population-specific disease risk alleles. In fact, ~200 novel variants were found to be located on exonic regions of each of the Bruneian genome and majority of them are of non-synonymous types. Genes harbouring these non-synonymous variants include those that are known to be associated with genetic diseases, such as cancer, cardiovascular diseases, and diabetes mellitus, and pharmacogenetic markers. However, more genotypic and phenotypic data from larger cohorts will be needed to provide the necessary resolution power for studies such as GWAS to establish the association between these rare variants and diseases.

In addition, a number of variant-dense loci which have not been reported in other population groups have also been identified in the two Bruneian Malay individuals. For instance, chr10 and 13 were found to harbour the highest variant density. This finding is different from the analysis done using the 1kGP data in which chr16 and 1 were found to have the highest and lowest SNP density, respectively [40]. It is important, however, to realise the potential biases introduced into the present study with only two samples while the majority samples in the 1kGP were European.

Subtle Genetic Differences between the Malays from Borneo and Malay Peninsula

The Malays are one of the largest Austronesian population groups spreading over Island Southeast Asia and as far as South Africa. Although the different sub-ethnic groups share considerable cultural and linguistic ties, they are not genetically homogenous [6, 11, 43]. Adding to this genetic diversity, population genetic structure analysis conducted here has, for the first time, unveiled the differences in genetic ancestries among the Murut and Dusun Malays in Brunei and the presumably most closely related Malay groups in Singapore and Malaysia. Both PCA and admixture analysis revealed that the local Malays share a closer genetic ancestry background with the Filipino Vizaya and Luzon than the Singaporean and Malaysian Malays. The handful Malaysians who were found to share near identical genetic ancestries with this Bruneian-Filipino group most likely belong to either the Malay or one of the indigenous groups in the Malaysian Bornean state of Sabah. In fact, Yew *et al*. (2018) [11] reported that Malays from East Malaysia, specifically those residing in the state of Sabah, share a common ancestry with the Filipinos. Given the close geographical proximity, genetic admixture among these people is not unexpected. Interestingly, a recent study on population genetic structure of the various Indonesian ethnic groups reported that the Sulawesians there are more closely related to the Bruneian Dusuns and Muruts than any other groups on the Indonesian Archipelago [43]. While it is likely that the Bruneian Dusun and Murut Malays may also share highly similar genetic ancestries with other yet to be studied sub-ethnics groups on the various islands of Borneo, the Philippines and Indonesia, evidence to date seems to suggest that this

group of “Malay” was likely to have spread from the Philippines west- and south-ward to as far as the coastal regions of Borneo and Sulawesi only.

The presence of population-specific novel sequences in the Bruneian Malay genome

Although these sequences have not been validated using Sanger sequencing or PCR, their high mapping quality to partial human reference genome and similarity to both the novel East Asian and fosmid- or BAC-cloned human sequences indicate that they have to be novel genomic sequences harboured by the Bruneians.

The fact that ~146 Kbp of the novel ~227 Kbp sequences found in the Bruneian Malay genome are highly similar to that of the Chinese and Japanese corroborates well with findings in our admixture analysis. Indeed, some genomic sequences which are absent in the human reference genome have now be shown to be Asian-specific and they are shared among different Asian population groups. Shi *et al*. (2016) [12] found that only a quarter of the Chinese HX1 novel sequences are absent in previously reported Asian genomes, suggesting that majority of them are likely to be present in other Asian populations. Similarly, more than half of the Japanese JRG novel sequences reported by Nagasaki *et al*. (2019) [15] were also found in the *de-novo* assembled Korean genome by Seo *et al*. (2016) [16]. On the other hand, since both Chinese and Japanese studies have reported novel sequences in the Mbp range, it is probable that more Malay-specific genomic sequences have yet to be uncovered. In addition, the presence of an ORF encoding a novel human homologue of a primate’s small zinc finger protein have added more weight to the importance of finding such population-specific novel sequences.

Conclusion

This is to our best knowledge the first and most comprehensive genetics and genomics analysis of the Malays in Brunei. In addition to adding ~5.2 million variants to the local Malay population and the discovery of ~227 Kbp novel genomic sequences, our studies have also shown the existence of subtle differences in the population genetic structure among the different Malay groups in Southeast Asia. Hence, a more refined stratification of these groups using variants from larger cohorts will be necessary should the benefits of medical genetics and genomics are to be fully realised in the region.

Data availability

The consensus genetic variants for the 41 Bruneian Malays are available from: https://genome-asia.ucsc.edu/s/Mirza%20Azmi/BNMalayGRCh38.

Acknowledgements

We are grateful to the local individual who willingly share with us his family's genotyping data from AncestryDNA®.

Author Contributions

MA designed the study, performed the bioinformatics analysis, and drafted the manuscript. AI and LC provided valuable insights and discussion on human and medical genetics. ZHL conceived and designed the study, supervised the works, and drafted the manuscript. All authors reviewed the manuscripts.

Funding

This work was funded by a Universiti Brunei Darussalam's Competitive Research Grant (No: UBD/OAVCR/CRGWG(017)/171001).

References

1. International HapMap C. The International HapMap Project. Nature. 2003;426:789-96.
2. 1000 Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68-74.
3. Teo YY, Sim X, Ong RT, Tan AK, Chen J, Tantoso E, et al. Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res. 2009;19:2154-62.
4. Wong LP, Ong RT, Poh WT, Liu X, Chen P, Li R, et al. Deep whole-genome sequencing of 100 southeast Asian Malays [Research Support, Non-U.S. Gov't]. Am J Hum Genet. 2013;92:52-66.
5. Wu D, Dou J, Chai X, Bellis C, Wilm A, Shih CC, et al. Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. Cell. 2019;179:736-49 e15.
6. Morseburg A, Pagani L, Ricaut FX, Yngvadottir B, Harney E, Castillo C, et al. Multi-layered population structure in Island Southeast Asians. European journal of human genetics : EJHG. 2016;24:1605-11.
7. Consortium HP-AS, Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, et al. Mapping human genetic diversity in Asia. Science. 2009;326:1541-5.
8. Fu R, Mokhtar SS, Phipps ME, Hoh BP, Xu S. A genome-wide characterization of copy number variations in native populations of Peninsular Malaysia. European journal of human genetics : EJHG. 2018;26:886-97.
9. Salleh MZ, Teh LK, Lee LS, Ismet RI, Patowary A, Joshi K, et al. Systematic pharmacogenomics analysis of a Malay whole genome: proof of concept for personalized medicine. PloS one. 2013;8:e71554.
10. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations [Article]. Nature. 2016;538:201-06.
11. Yew CW, Hoque MZ, Pugh-Kitingan J, Minsong A, Voo CLY, Ransangan J, et al. Genetic relatedness of indigenous ethnic groups in northern Borneo to neighboring populations from Southeast Asia, as inferred from genome-wide SNP data. Annals of human genetics. 2018;82:216-26.
12. Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, et al. Long-read sequencing and *de novo* assembly of a Chinese genome. Nat Commun. 2016;7:12065.
13. Du Z, Ma L, Qu H, Chen W, Zhang B, Lu X, et al. Whole Genome Analyses of Chinese Population and *de novo* Assembly of A Northern Han Genome. Genomics, proteomics & bioinformatics. 2019;17:229-47.
14. Cao H, Wu H, Luo R, Huang S, Sun Y, Tong X, et al. *De novo* assembly of a haplotype-resolved human genome. Nat Biotechnol. 2015;33:617-22.
15. Nagasaki M, Kuroki Y, Shibata TF, Katsuoka F, Mimori T, Kawai Y, et al. Construction of JRG (Japanese reference genome) with single-molecule real-time sequencing. Hum Genome Var. 2019;6:27.

16. Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, et al. De novo assembly and phasing of a Korean human genome [Letter]. Nature. 2016;538:243-47.
17. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30:2843-51.
18. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv e-prints. 2013;1303:3997.
19. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data [Research Support, N.I.H., Extramural]. Genome Res. 2010;20:1297-303.
20. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987-93.
21. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv. 2012;*1207*:3907.
22. Schilbert HM, Rempel A, Pucker B. Comparison of Read Mapping and Variant Calling Tools for the Analysis of Plant NGS Data. Plants (Basel). 2020;9:1-14.
23. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020;12:91.
24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559-75.
25. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
26. Lott MT, Leipzig JN, Derbeneva O, Xie HM, Chalkia D, Sarmady M, et al. mtDNA Variation and Analysis Using Mitomap and Mitomaster. Current protocols in bioinformatics. 2013;44:1 23 1-26.
27. Castellana S, Fusilli C, Mazzoccoli G, Biagini T, Capocefalo D, Carella M, et al. High-confidence assessment of functional impact of human mitochondrial non-synonymous genome variations by APOGEE. PLoS Comput Biol. 2017;13:e1005628.
28. Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, et al. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. bioRxiv. 2015:023754.
29. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156-8.
30. Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, et al. rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool for Genome-wide Association Study. Genomics, proteomics & bioinformatics. 2021;19:619-28.
31. Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, et al. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. 2012;40:e69.

32. Geoffroy V, Herenger Y, Kress A, Stoetzel C, Piton A, Dollfus H, et al. AnnotSV: an integrated tool for structural variations annotation. Bioinformatics. 2018;34:3572-74.
33. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456:98-101.
34. Wickham H. ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). Springer: New York, 2016.
35. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655-64.
36. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257.
37. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674-6.
38. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014;30:31-7.
39. Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. Runs of homozygosity: windows into population history and trait architecture. Nat Rev Genet. 2018;19:220-34.
40. Telenti A, Pierce LC, Biggs WH, di Iulio J, Wong EH, Fabani MM, et al. Deep sequencing of 10,000 human genomes. Proceedings of the National Academy of Sciences of the United States of America. 2016;113:11901-06.
41. Kido T, Sikora-Wohlfeld W, Kawashima M, Kikuchi S, Kamatani N, Patwardhan A, et al. Are minor alleles more likely to be risk alleles? BMC Med Genomics. 2018;11:3.
42. Moore CB, Wallace JR, Wolfe DJ, Frase AT, Pendergrass SA, Weiss KM, et al. Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. PLoS Genet. 2013;9:e1003959.
43. Hudjashov G, Karafet TM, Lawson DJ, Downey S, Savina O, Sudoyo H, et al. Complex Patterns of Admixture across the Indonesian Archipelago. Mol Biol Evol. 2017;34:2439-52.
44. Shen GQ, Li L, Girelli D, Seidelmann SB, Rao S, Fan C, et al. An LRP8 variant is associated with familial and premature coronary artery disease and myocardial infarction. Am J Hum Genet. 2007;81:780-91.
45. Bogari NM, Aljohani A, Amin AA, Al-Allaf FA, Dannoun A, Taher MM, et al. A genetic variant c.553G > T (rs2075291) in the apolipoprotein A5 gene is associated with altered triglycerides levels in coronary artery disease (CAD) patients with lipid lowering drug. BMC Cardiovasc Disord. 2019;19:2.
46. Lee JY, Kim J, Kim SW, Park SK, Ahn SH, Lee MH, et al. BRCA1/2-negative, high-risk breast cancers (BRCAX) for Asian women: genetic susceptibility loci and their potential impacts. Scientific reports. 2018;8:15263.
47. Henriquez-Hernandez LA, Murias-Rosales A, Gonzalez-Hernandez A, de Leon AC, Diaz-Chico N, Fernandez-Perez L. Distribution of TYMS, MTHFR, p53 and MDR1 gene polymorphisms in patients with breast cancer treated with neoadjuvant chemotherapy. Cancer Epidemiol. 2010;34:634-8.

48. Marano M, Bottaro G, Goffredo B, Stoppa F, Pisani M, Marinaro AM, et al. Deferasirox-induced serious adverse reaction in a pediatric patient: pharmacokinetic and pharmacogenetic analysis. Eur J Clin Pharmacol. 2016;72:247-8.

Figure legends

Figure 1A. Variant density plot of the two Malay genomes. The SNPs (A) and INDELs (B) were distributed across the entire genome except in the unmapped regions and known gaps of the human reference genome. Regions with high variant densities are coloured red while those in green depict regions with low variant densities.

Figure 2. Gene Ontology analysis on genes harbouring deleterious variants. The Sankey map shows the association of the genes with their protein classes on the middle and biological functions on the right. The genes encode for a variety of protein classes involving in different biological processes.

Figure 3. Population genetic structure of 22 Asian ethnic groups. (A) Principal component analysis and (B) admixture analysis of ~583K variants among the 1499 individuals. BM, Burma/Myanmar; BN, Brunei; EAS, East Asians; CDX, Chinese Dai; CHB, Han Chinese in Beijing; CHS, Southern Han Chinese; JPT, Japanese; KHV, Kinh Vietnamese, EU, Europe; MY, Malaysia; PH, Philippines; SAS, South Asians; BEB, Bengali; GIH, Gujarati; ITU, Telugu; PJL, Punjabi; STU, Sri Lankan Tamil; SG, Singapore; VN, Vietnam.

Table 1. Variant statistics called from the WGS data of the two Bruneian Malays

Variants	**Combined**	**Individual 1**	**Individual 2**
	Nuclear Chromosomes		
Total**:**	5,069,437	3,935,796	3,935,869
*Variants in the primary assembly of GRCh38 only	[illegible]4%)	3,551,224 (90.0%)	3,544,852 (89.9%)
INDELs (% total)	[illegible])	384,572 (10.0%)	391,017 (10.1%)
Het/Homo ratio	1.23	1.25	1.22
SNP Ts/Tv ratio	2.09	2.09	2.09
Novel variants (% total):	216,837 (4.3%)	161,427 (4.1%)	167,745 (4.3%)
SNPs (% novel)	7,565 (3.5%)	4,468 (2.8%)	5,497 (3.8%)
INDELs (% novel)	209,272 (96.5)	156,959 (97.2%)	162,248 (96.2%)
MAF <0.001 (% total)	148,481 (2.9%)	77,660 (2.0%)	80,143 (2.0%)
RNA genes* (% total)	614,368 (12.1%)	477,073 (12.1%)	477,184 (12.1%)
Protein-coding genes* (% total):	1,910,236 (37.7%)	1,478,330 (37.6)	1,478,164 (37.6)
Intron (% coding genes)	1,743,454 (34.4%)	1,349,176 (34.3%)	1,348,714 (34.3%)
Exon (% coding genes)	29,591 (0.6%)	22,865 (0.6%)	22,770 (0.6%)
5' UTR (% coding genes)	9,700 (0.2%)	7,526 (0.2%)	7,624 (0.2%)
3' UTR (% coding genes)	42,748 (0.8%)	33,332 (0.8%)	33,095 (0.8%)
Splicing (% coding genes)	190 (<0.01%)	143 (<0.01%)	147 (<0.01%)

Table 2. Summary of the protein-coding variants found in the 41 Bruneian Malays.

Functional annotation	**Combined**	**WGS**	**Genotyping**
Short variants:	5,276,758	5,069,437	697,223
SNP	4,776,365	4,569,044	697,223
INDELs	500,393	500,393	0
Protein-coding variants:	35,783	29,591	17,952
Synonymous (%)	17,900 (50.0%)	14,811 (50.1%)	9,242 (51.5%)
Non-synonymous (%):	17,753 (49.6%)	14,652 (49.5%)	8,690 (48.4%)
Missense (%)	17,111 (96.4%)	14,039 (95.8%)	8,611 (99.1%)
Stop-gain SNPs (%)	172 (1.0%)	146 (1.0%)	66 (0.8%)
Stop-loss SNPs (%)	35 (0.2%)	32 (0.2%)	13 (0.1%)
Frameshift INDELs (%)	154 (0.8%)	154 (1.1%)	0 (<0.1%)
Non-frameshift INDELs (%)	281 (1.6%)	281 (1.9%)	0 (<0.1%)
Unclassified (%)	130 (0.4%)	128 (0.4%)	20 (0.1%)

Table 3. List of deleterious short variants in the local Malays associated with disease risk factors.

dbSNP ID	Gene	ENSEMBL Transcript	Mutation	Global AF□	EAS AF□	BN AF□ (# of individual*)	Associated risk factors	Reference
rs5174	*LRP8*	ENST00000306052	c.G2855A: p.R952Q	0.28	0.03	0.027 (37)	Myocardial infarction	[44]
rs2075291	*APOA5*	ENST00000542499	c.G553T: p.G185C	0.0027	0.0713	0.013 (37)	Hypertriglyceridemia	[45]
rs766173	*BRCA2*	ENST00000380152	c.A865C: p.N289H	0.0385	0.0932	0.041 (37)	Breast cancer	[46]
rs1801133	*MTHFR*	ENST00000376585	c.C788T: p.A263V	0.2719	0.2873	0.041 (39)	Increased drug toxicity of cyclophosphamide and fluorouracil (cancer)	[47]
rs4148323	*UGT1A1*	ENST00000305208	c.G211A: p.G71R	0.0092	0.1419	0.03 (37)	Increased drug toxicity to deferasirox (iron overload)	[48]

* The number of Bruneian Malays with genotype information of a given variant. Not all the Malays have complete genotype information of all the variants, especially the genotyping data. Therefore, this number was included alongside the Bruneian allele frequency (BN AF) to show the depth of the calculation.

□ AFs were based on GnomAD

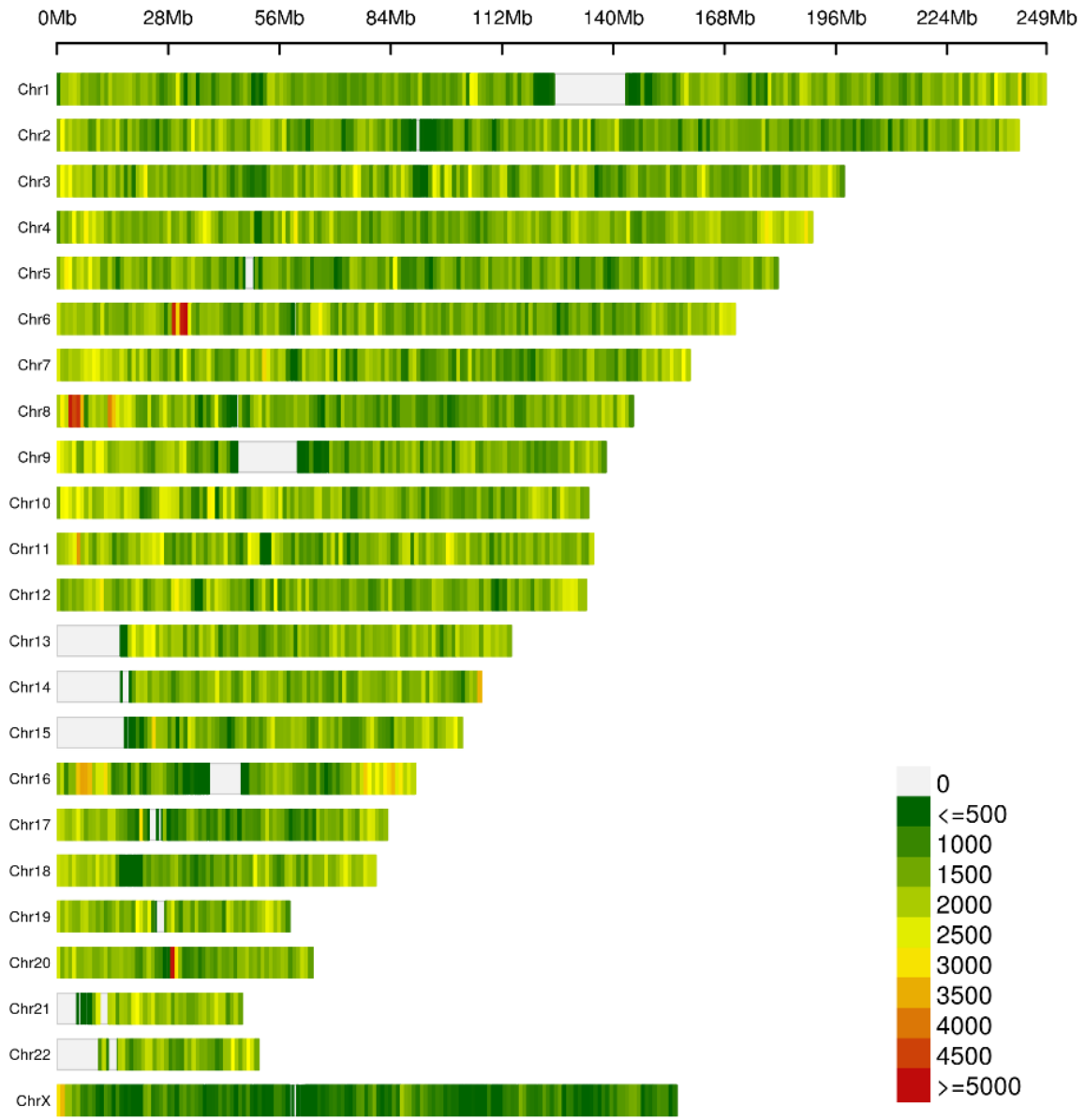
0Mb
28Mb
56Mb
84Mb
112Mb
140Mb
168Mb
196Mb
224Mb
249Mb
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
0
<=500
1000
1500
2000
2500
3000
3500
4000
4500
>=5000

Figure 1A

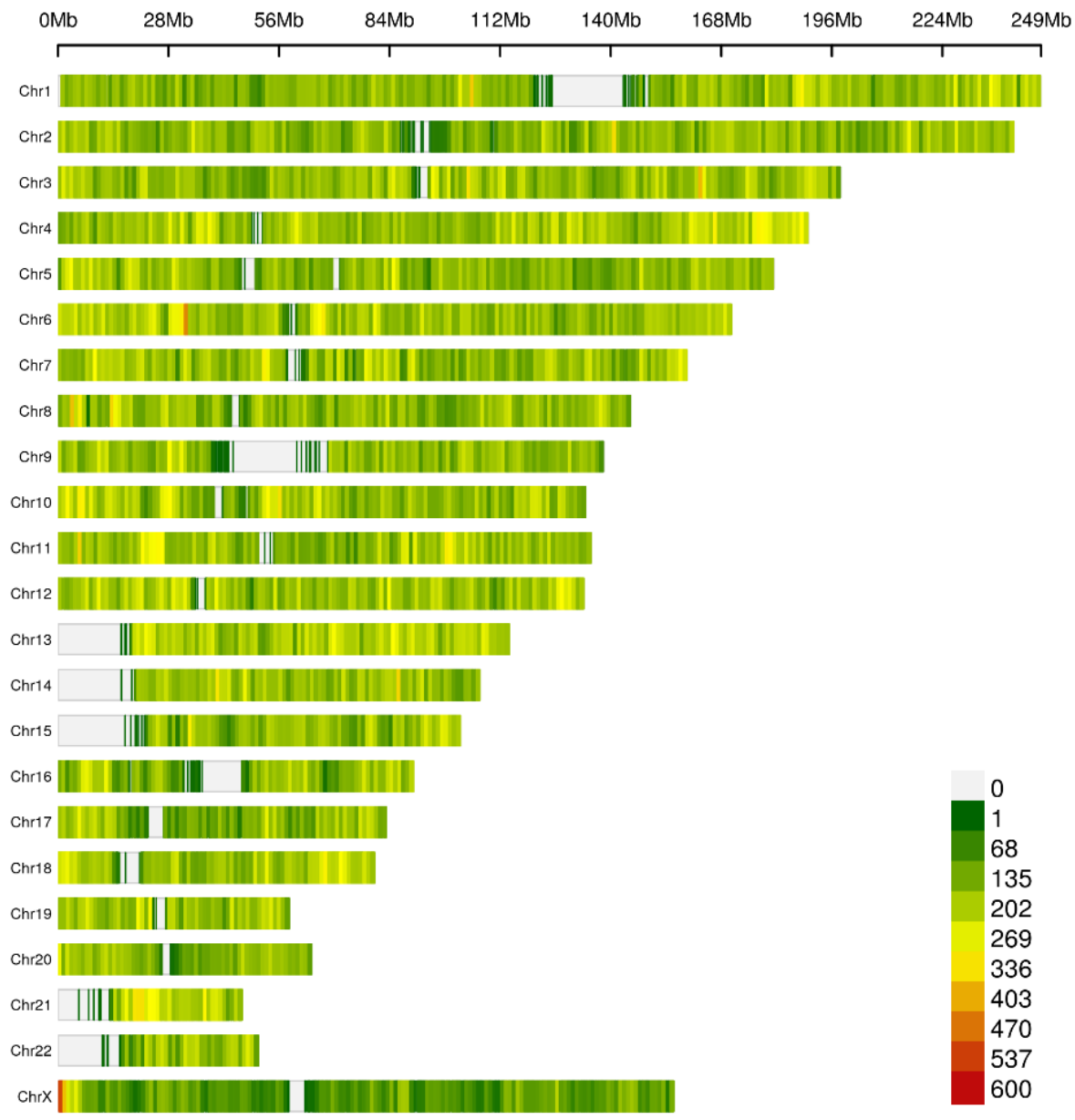
0Mb
28Mb
56Mb
84Mb
112Mb
140Mb
168Mb
196Mb
224Mb
249Mb
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
0
1
68
135
202
269
336
403
470
537
600

Figure 1B

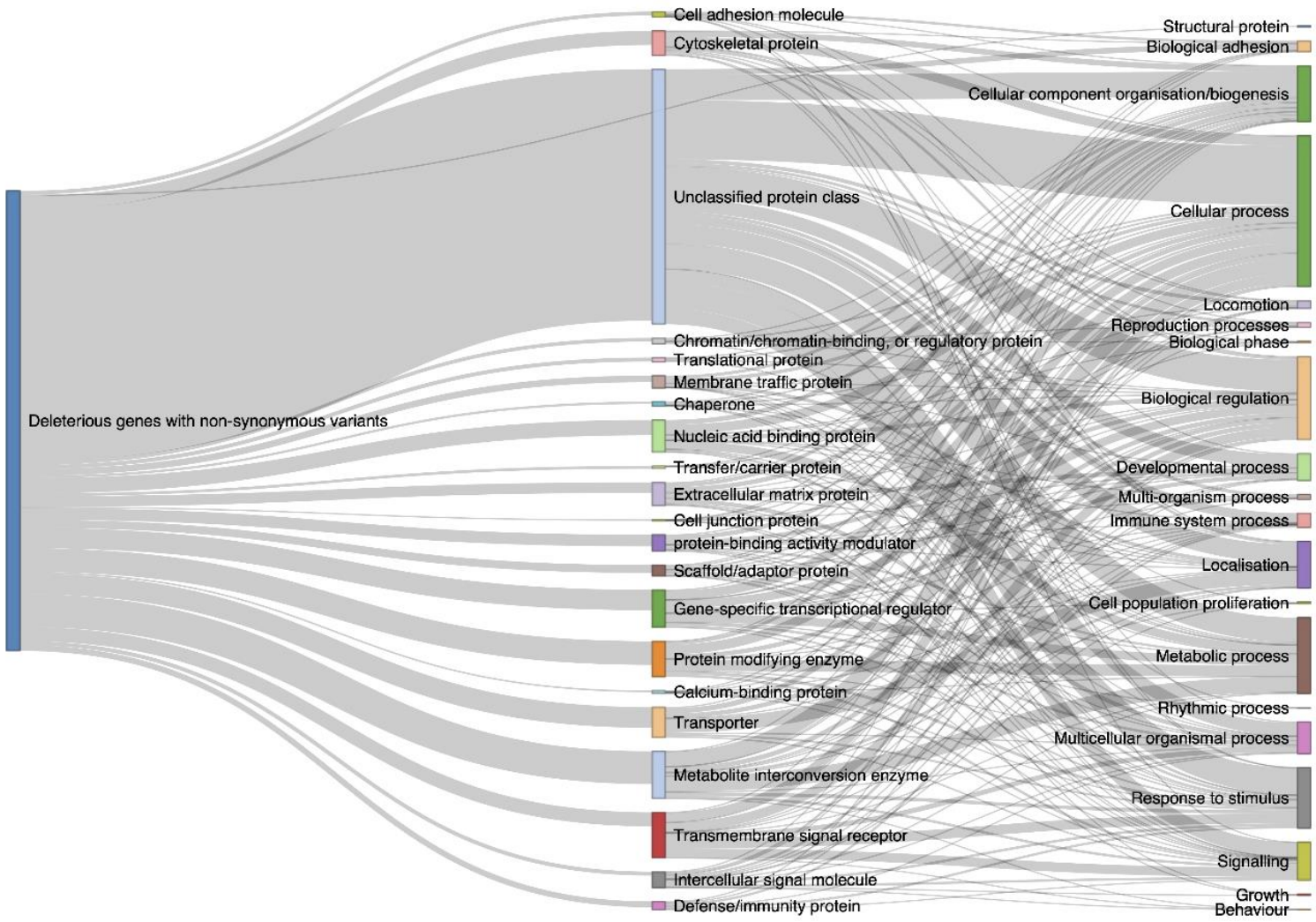
Deleterious genes with non-synonymous variants
Cell adhesion molecule
Cytoskeletal protein
Unclassified protein class
Chromatin/chromatin-binding, or regulatory protein
Translational protein
Membrane traffic protein
Chaperone
Nucleic acid binding protein
Transfer/carrier protein
Extracellular matrix protein
Cell junction protein
protein-binding activity modulator
Scaffold/adaptor protein
Gene-specific transcriptional regulator
Protein modifying enzyme
Calcium-binding protein
Transporter
Metabolite interconversion enzyme
Transmembrane signal receptor
Intercellular signal molecule
Defense/immunity protein
Structural protein
Biological adhesion
Cellular component organisation/biogenesis
Cellular process
Locomotion
Reproduction processes
Biological phase
Biological regulation
Developmental process
Multi-organism process
Immune system process
Localisation
Cell population proliferation
Metabolic process
Rhythmic process
Multicellular organismal process
Response to stimulus
Signalling
Growth
Behaviour

Figure 2

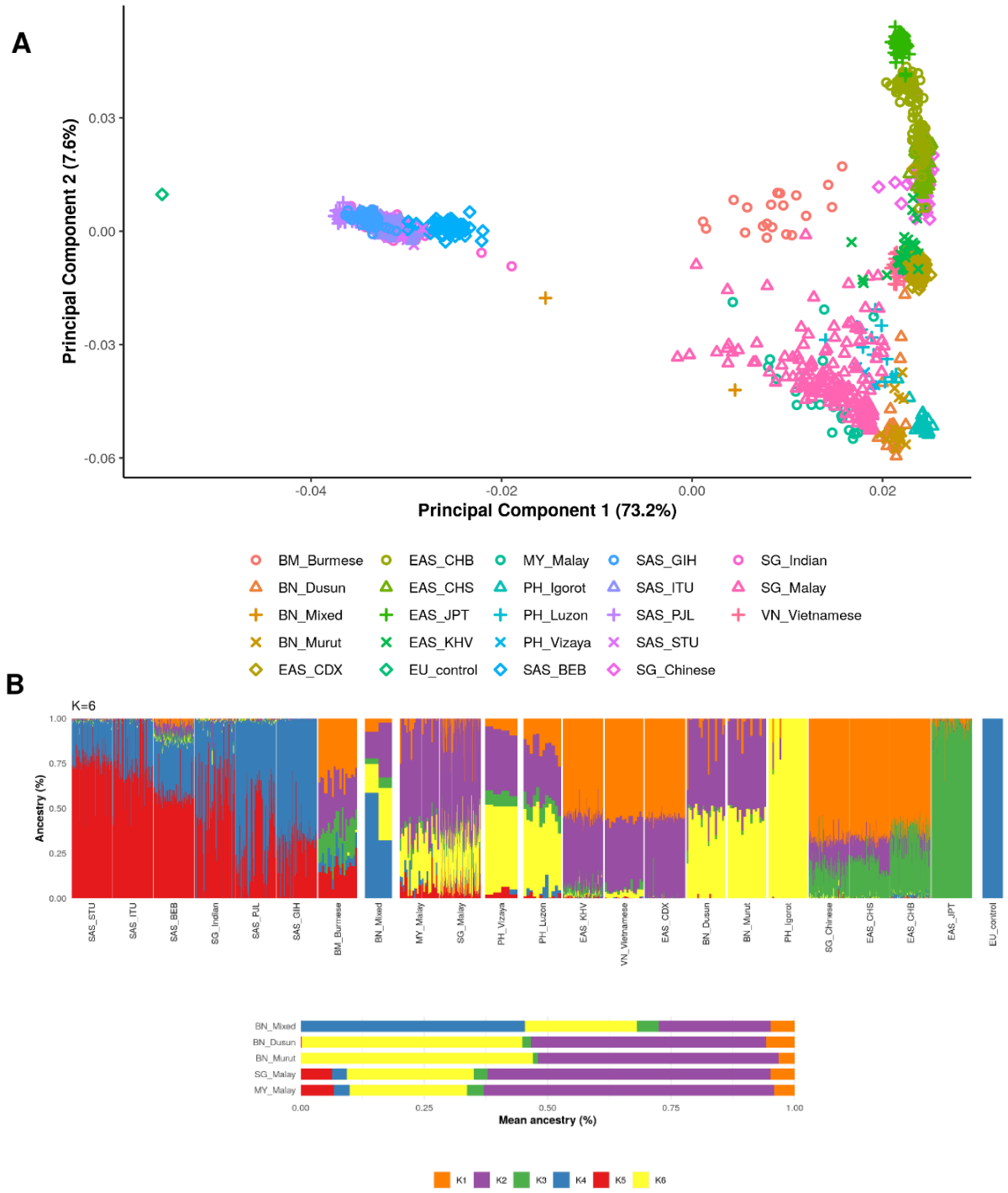
A
Principal Component 2 (7.6%)
0.03
0.00
-0.03
-0.06
-0.04
-0.02
0.00
0.02
Principal Component 1 (73.2%)
BM_Burmese
BN_Dusun
BN_Mixed
BN_Murut
EAS_CDX
EAS_CHB
EAS_CHS
EAS_JPT
EAS_KHV
EU_control
MY_Malay
PH_Igorot
PH_Luzon
PH_Vizaya
SAS_BEB
SAS_GIH
SAS_ITU
SAS_PJL
SAS_STU
SG_Chinese
SG_Indian
SG_Malay
VN_Vietnamese
B
K=6
Ancestry (%)
1.00
0.75
0.50
0.25
0.00
SAS_STU
SAS_ITU
SAS_BEB
SG_Indian
SAS_PJL
SAS_GIH
BM_Burmese
BN_Mixed
MY_Malay
SG_Malay
PH_Vizaya
PH_Luzon
EAS_KHV
VN_Vietnamese
EAS_CDX
BN_Dusun
BN_Murut
PH_Igorot
SG_Chinese
EAS_CHS
EAS_CHB
EAS_JPT
EU_control
BN_Mixed
BN_Dusun
BN_Murut
SG_Malay
MY_Malay
0.00
0.25
0.50
0.75
1.00
Mean ancestry (%)
K1
K2
K3
K4
K5
K6

Figure 3