

Characterising private and shared signatures of positive selection in 37 Asian populations

Xuanyao Liu^{1,2}, Dongsheng Lu³, Woei-Yuh Saw^{2,4}, Philip J. Shaw⁵, Pongsakorn Wangkumhang⁵, Chumpol Ngamphiw⁵, Suthat Fucharoen⁶, Worachart Lert-itthiporn^{7,8}, Kwanrutai Chin-inmanu⁸, Tran Nguyen Bich Chau⁹, Katie Anders^{9,10}, Anuradhani Kasturiratne¹¹, H. Janaka de Silva¹², Tomohiro Katsuya¹³, Ryosuke Kimura¹⁴, Toru Nabika¹⁵, Takayoshi Ohkubo¹⁶, Yasuharu Tabara¹⁷, Fumihiko Takeuchi¹⁸, Ken Yamamoto¹⁹, Mitsuhiro Yokota²⁰, Dolikun Mamatyusupu²¹, Wenjun Yang²², Yeun-Jun Chung²³, Li Jin²⁴, Boon-Peng Hoh²⁵, Ananda R. Wickremasinghe¹¹, Rick Twee-Hee Ong², Chiea-Chuen Khor²⁶, Sarah J. Dunstan^{9,10,27}, Cameron Simmons^{9,10,28}, Sissades Tongsimas⁵, Prapat Suriyaphol⁸, Norihiro Kato¹⁸, Shuhua Xu^{3,29,30}, Yik-Ying Teo^{1,2,4,18,26,31,*}

¹ NUS Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore 117456

² Saw Swee Hock School of Public Health, National University of Singapore, Singapore 117549

³ Max Planck Independent Research Group on Population Genomics, Chinese Academy of Sciences and Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences Shanghai, China 200031

⁴ Life Sciences Institute, National University of Singapore, Singapore 117456

⁵ National Center for Genetic Engineering and Biotechnology, National Science and Technology Development Agency, Pathum Thani, Thailand 12120

⁶ Institute of Molecular Biosciences, Mahidol University, Nakhon Pathom, Thailand 73170

⁷ Molecular Medicine Graduate Program, Faculty of Science, Mahidol University, Bangkok, Thailand 10700

⁸ Division of Bioinformatics and Data Management for Research, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, Thailand 10700

⁹ Oxford University Clinical Research Unit, Hospital for Tropical Diseases, Ho Chi Minh City, Vietnam

¹⁰ Centre for Tropical Medicine, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK OX3 7BN

¹¹ Department of Public Health, Faculty of Medicine, University of Kelaniya, Ragama, Sri Lanka

¹² Department of Medicine, Faculty of Medicine, University of Kelaniya, Ragama, Sri Lanka

¹³ Department of Clinical Gene Therapy, Osaka University Graduate School of Medicine, Suita, Japan 565-0871

¹⁴ Department of Human Biology and Anatomy, Graduate School of Medicine, University of the Ryukyus, Nishihara-cho, Japan 903-0215

¹⁵ Department of Functional Pathology, Shimane University School of Medicine, Izumo, Japan 693-8501

¹⁶ Department of Hygiene and Public Health, Teikyo University School of Medicine, Tokyo, Japan 162-8655

¹⁷ Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan 606-8501

¹⁸ Department of Gene Diagnostics and Therapeutics, National Center for Global Health and Medicine, Tokyo, Japan 162-8655

¹⁹ Department of Medical Chemistry, Kurume University School of Medicine, Kurume, Japan 830-0011

²⁰ Department of Genome Science, School of Dentistry, Aichi Gakuin University, Nagoya, Japan 464-8651

²¹ College of the Life Sciences and Technology, Xinjiang University, Urumqi, China 830046

²² Key Laboratory of Reproduction and Heredity of Ningxia Region, Ningxia Medical University, Yinchuan, Ningxia, China 750004

²³ Integrated Research Center for Genome Polymorphism, Department of Microbiology, The Catholic University Medical College, Socho-gu, Seoul, Korea 137-701

- ²⁴ State Key Laboratory of Genetic Engineering and Ministry of Education (MOE) Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, China 200433
- ²⁵ Institute of Medical Molecular Biotechnology, Faculty of Medicine, Universiti Teknologi MARA, Sungai Buloh Campus, Selangor, Malaysia
- ²⁶ Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672
- ²⁷ The Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Melbourne, Australia
- ²⁸ Department of Microbiology and Immunology, The University of Melbourne, Melbourne, Australia
- ²⁹ School of Life Sciences and Technology, ShanghaiTech University, Shanghai, China
- ³⁰ Collaborative Innovation Center of Genetics and Development, Shanghai, China
- ³¹ Department of Statistics and Applied Probability, National University of Singapore, Singapore

Running title: Positive selection in Asia

Conflict of interest statement: The authors declare there is no conflict of interest that results from the submission of this manuscript.

Keywords: population structure, positive selection, Asian populations

* Corresponding author: Saw Swee Hock School of Public Health, MD1, 12 Science Drive 2, National University of Singapore, Singapore 117549. Email: statyy@nus.edu.sg. Tel: +65-6516 2760. Fax: +65 6872 3919.

ABSTRACT

The Asian Diversity Project (ADP) assembled 37 cosmopolitan and ethnic minority populations in Asia that have been densely genotyped across over half a million markers to study patterns of genetic diversity and positive natural selection. We performed population structure analyses of the ADP populations and divided these populations into four major groups based on their genographic information. By applying a highly sensitive algorithm haploPS to locate genomic signatures of positive selection, 140 distinct genomic regions exhibiting evidence of positive selection in at least one population were identified. We examined the extent of signal sharing for regions that were selected in multiple populations and observed that populations clustered in a similar fashion to that of how the ancestry clades were phylogenetically defined. In particular, populations predominantly located in South Asia underwent considerably different adaptation as compared to populations from the other geographical regions. Signatures of positive selection present in multiple geographical regions were predicted to be older and have emerged prior to the separation of the populations in the different regions. In contrast, selection signals present in a single population group tended to be of lower frequencies and thus can be attributed to recent evolutionary events.

INTRODUCTION

The HUGO Pan-Asian SNP Consortium (PASNP) was a successful model of a genomics research network in Asia which surveyed the genetic make-up of 1,928 participants from 73 populations and permitted an unprecedented insight into possible human migration across the Asian continent ¹. However, owing to the small number of participants in each population and the use of a low-density genotyping microarray, analyses of the PASNP data were mostly limited to broad analyses of population structure between the populations, and less is known about the cross-population micro-structures such as admixture, linkage disequilibrium (LD) variation, as well as within-population adaptive forces such as natural selection and population bottlenecks. In fact, the sparse SNP density confounded several attempts to investigate the role of natural selection in driving genetic diversity between these populations, regardless of whether these relied on comparing allele frequencies ² or locating long haplotypes ³.

As the largest continent on Earth spanning over 93.5° in latitude and 196.3° in longitude, Asia possesses tremendous diversity in climate and altitude. This geographical variation, coupled with the myriad cultures and languages in Asia, have exerted profound influences on the genetics of Asian populations, allowing these populations to adapt differentially to high altitude ⁴⁻⁸, diet⁹ and infectious diseases ^{3,10,11}. Given the vast expanse of Asia, populations that are geographically closer are expected to have adapted in a similar fashion, given the likelihood of more similar micro-climates and dietary practices, whereas the selection on populations further apart may substantially differ. However, to date no collection of genetic data spanning multiple Asian populations exists at sufficient genomic coverage to allow a comprehensive and systematic evaluation of local adaptation in Asia ^{2,12-14}.

The Asian Diversity Project (ADP) overcomes this paucity in data by enlisting scientists across Asia to cooperate and share resources for a deeper survey of population genetics in Asia. To date, ADP has

assembled genetic data for 37 cosmopolitan and ethnic minority populations from public databases ¹⁵⁻¹⁹ and private repositories (**Figure 1A**), where there are at least half a million SNP markers genotyped in each population and 33 of the 37 populations surveyed at least 10 individuals (**Table S1**).

Here, we aim to characterise the extent and sharing of positive selection in the 37 ADP populations. To do this, we first investigated the degree of genetic similarity between these populations with a comprehensive suite of population structure analyses, which supported the clustering of the Asian populations into four major population groups that corresponded broadly with the geographical distributions of the populations and present-day definitions of Asian populations. The genotype data for each population was subsequently phased with BEAGLE ²⁰, and the haplotype-based methods haploPS ¹⁴ was employed on the phased data to locate genomic signatures of positive selection. The regions identified to be under selection were summarised according to the degree of sharing across populations in the four groups. We believe this is the first survey of positive selection in Asia that possesses sufficient statistical power to allow a systematic evaluation of the extent of sharing signatures of local adaptation. The genotype data has been deposited at the European Genome-phenome Archive (EGA, <https://ega-archive.org>) under the accession number EGAS000001002100.

METHODS

Sample collection and genotyping in ADP

The ADP has integrated 37 Asian populations from publicly available databases, such as (i) the four populations with Asian ancestries in Phase 3 of the International HapMap Project ¹⁹ (HapMap: CHB, CHD, GIH, JPT); (ii) three populations from the Singapore Genome Variation Project ¹⁶ (SGVP: CHS, INS, MAS); and four North Eurasian populations (Kurd, Buryat, Kyrgystani, Mongolian) from the Jorde Lab ^{17,18} (http://jorde-lab.genetics.utah.edu/?page_id=23). DNA samples of Tajik, Uyghur, Kazakh, and

Kirgiz were collected from the Xinjiang Uyghur Autonomous Region in China, whereas DNA samples of Hui (30) were collected from the Ningxia Hui Autonomous Region in China. All procedures were followed in accordance with the ethical standards of the ethics committee of Fudan University and the Helsinki Declaration of 1975, as revised in 2000. Data and information of the other ethnic populations from China (Dong, Han, Li, Tibetan, Yao, Zhuang) were described elsewhere as indicated ^{21,22}. The samples from eight Japanese prefectures (Amagasaki, Ehime, Fukuoka, Kita-nagoya, Ohasama, Okinawa²³, Shimane, Tokyo) were coordinated through the National Center for Global Health and Medicine (NCGM) in Tokyo, on behalf of the Japanese SNP Consortium. The NCGM was also responsible for coordinating the Sri Lankan datasets which included Sri Lanka Tamils, Indian Tamils, and Sinhalese. A genetic dataset of 200 Kinh Vietnamese was made available from the Oxford University Clinical Research Unit in Ho Chi Minh City, Vietnam, while two independent Thai datasets were contributed respectively by the Siriraj Hospital, Mahidol University, and Genome Technology Research Unit, the National Center for Genetic Engineering and Biotechnology (BIOTEC).

The samples from the HapMap and SGVP were genotyped on both the Affymetrix Genome-wide Human SNP array 6.0 and the Illumina Human1M BeadChip. Genotyping of the Chinese and Korean samples were performed only on the Affymetrix 6.0 array. The Japanese samples were genotyped on different platforms, with Ohasama on Affymetrix SNP 5.0, Tokyo and Shimane on Illumina 550k, Okinawa on the Illumina Omni-Express ²³ and the rest of the populations on Illumina 2.5M. The three Sri Lankan populations were similarly genotyped on the Illumina 2.5M. Both sets of Thai samples were genotyped on the Illumina 610k while the Vietnamese Kinhs were genotyped on the Illumina 660W. All coordinates reported here have been referenced against Build 37 (hg19).

Datasets from the contributing sites have undergone quality checks, which have excluded problematic samples and SNPs. Specific details of the QC process can be found from the respective publications.

However, for the purpose of our intended analysis, we reiterate here that this QC process also excluded samples with a high level of identity-by-state (IBS) genotypes, where only the sample with the higher call rate was retained in each pair of individuals with excessive IBS.

Details of the 37 ADP populations, including the sample size in each population, can be found in **Table S1**.

Data availability

The post-QC genotype data of the 37 populations in the ADP is available upon request to statyy@nus.edu.sg. As the ADP includes datasets from public resources, only the post-QC data will be shared for these populations and requests for the raw dataset will be directed to the public resources. The raw genotype data for all other populations will however be available.

Population structure analyses

The analyses of the population structure for the 37 ADP populations were performed with four different strategies: (i) calculating the Weir and Hill F_{ST} between every pair of populations, defined as

$$F_{ST} = \frac{(MSP - MSG)}{(MSP + (n - 1)MSG)}$$

where $MSG = \frac{\sum_{i=1}^2 n_i \tilde{p}_i (1 - \tilde{p}_i)}{\sum_{i=1}^2 (n_i - 1)}$, $MSP = \sum_{i=1}^2 n_i (\tilde{p}_i - \bar{p})^2$, with n_i the sample size of the i^{th} population, n a constant determined by the number of samples and populations, and \tilde{p}_i and \bar{p} denote the average frequency of the allele in population i and both populations respectively; (ii) a nearest-neighbor phylogenetic tree constructed with a genetic distance matrix where the $(i, j)^{th}$ element of the matrix is represented by the F_{ST} for population i and population j ; (iii) a series of principal component analyses

(PCAs) with 30,927 SNPs that are present in all 37 ADP populations as well as CEU and YRI from HapMap3 using *smartPCA* in the EIGENSOFT package ²⁴; and (iv) a series of admixture analyses with ADMIXTURE version 1.22 ²⁵ with the same 30,927 SNPs where the number of ancestral populations (K) was allowed to vary from 2 to 12.

Haplotype phasing

The genotype data for the 33 ADP populations with at least 10 individuals were phased with BEAGLE version 3.3.2 ²⁶ to obtain the haplotype data necessary for the analysis of positive selection. Each population was phased independently of the others although internal validation indicated there were no significant differences in the phased haplotypes even if multiple populations were phased simultaneously (results not shown).

Locating genomic signatures of positive selection

The search for genomic signatures of positive selection was performed with the haplotype-based method haploPS ¹⁴, which was applied to the phased haplotype data from each of the 33 ADP populations and effectively relied on evidence of uncharacteristically long haplotypes when adjusted for background recombination.

HaploPS searches for the longest possible haplotype form that is present in the population with at least a pre-defined core haplotype frequency. At each of these frequencies, the identified haplotype form is measured by both the genetic distance spanned and the number of SNPs from the data that is present on the haplotype, where an adjusted haploPS score is subsequently derived from these two metrics after scaling against the number of such haplotypes present across the whole genome for the core frequency. In this study, to reduce the possibility that false positive signals of natural selection might confound the clustering analysis of the populations based on selection signals, we applied a

stringent threshold for identifying positive selection regions by haploPS, where a genomic region is only identified to present evidence of positive selection if the adjusted haploPS score is < 0.01 (see Discussion for sensitivity and specificity comparisons at this threshold). We conducted the search across a range of core haplotype frequencies from 5% to 95% in step-sizes of 5%. For a region identified by haploPS to be positively selected, combining the output across the range of core haplotype frequencies additionally allowed the putative frequency of the positively selected allele to be estimated. For a genomic region that is found to be positively selected in multiple populations, the individual selected haplotype forms can also be used to evaluate whether the emergence of the advantageous allele predates the separation of these populations, or is likely to have emerged independently in the different populations, also known as convergent evolution. The haploPS analysis used the software that is publicly available at <http://www.statgen.nus.edu.sg/~haplops/>.

HaploPS was chosen as it conferred higher statistical power and lower Type I error than conventional methods for detecting positive selection¹⁴. An equivalent analysis with two additional methods for detecting selection (iHS [12], XP-EHH [13]), along with a comparison of their results with those from haploPS can be found in the Supplementary Material.

Haplotype similarity index

Haplotype similarity index (HSI)¹⁴ is a metric to infer whether the selection signals shared by multiple populations originated from the same or different mutation events by quantifying the degree of similarity between the extended haplotype forms corresponding to a selection signal identified by haploPS that is concurrently present in multiple populations. A HSI larger or equal to 0.98 indicates high level of similarity and that the haplotype forms are likely to have descended from the same founder haplotype and thus indicative of a single-origin mutation event. HSI smaller than 0.9 indicates

different haplotype backgrounds exist due to convergent evolution. Details regarding the formulation and evaluation of the HSI and associated thresholds can be found in Liu et al¹⁴.

RESULTS

Population structure analyses

A total of 3,933 samples from 37 Asian populations remained after quality control, of which we identified a set of 30,927 SNPs distributed across the genome that were present in all 37 populations. This dataset is used for the analysis of population structure, together with the HapMap2 Europeans (CEU) and Africans (YRI) as outgroups. Using the F_{ST} values obtained from all possible pairwise comparisons of the 39 populations as inputs for a genetic distance matrix, we constructed a phylogenetic tree where YRI clearly segregated from the rest of the non-African populations, followed by CEU (**Figure 1B**). Amongst the Asian populations, the Kurdish and the Tajiks were genetically close to CEU ($F_{ST} = 1.4\%$ and 1.9% respectively), and all three populations were grouped together with South Asian populations (GIH, INS, Sinhalese, Indian Tamil, Sri Lankan Tamil) in the same clade in the phylogenetic tree. We thus defined this collection of populations as a population group corresponding to the South Region. Central Asian samples such as the Kyrgyzstani, Buryat and Mongolians were grouped together with Chinese ethnic minorities that are located in the same geographical region (Uyghur, Kazak, Kirgiz, Tibetan and Hui) in the Central Region. The classically defined East Asian samples from all eight Japanese populations, Korea, and the Han Chinese samples (CHB, CHD, CHS, Han) were grouped into the East Region, while the Southeast Asian populations from Thailand, Singapore, Vietnam and the ethnic minority populations of Dong, Li, Yao and Zhuang are grouped as the Southeast Region. However, it should be highlighted that the distinction between these regional groupings, in particular between East and Southeast Regions, are not as pronounced as those typically

observed when comparing between globally distributed ancestry groups, due to the complex nature of migration and admixture between Asian populations (see below and Supplementary Material).

A principal component analysis (PCA) performed with the same 30,927 SNPs yielded findings that concurred with the phylogenetic analyses, separating the YRI and CEU from the rest of the ADP populations (**Figure S1A**). The first two principal components from the PCA with only the 3,933 ADP samples exhibited remarkable correlation with geographical coordinates with the latitude and longitude of the populations (**Figures 1B, S1**, Pearson $r_{\text{longitude-PC1}} = 0.79$, $P\text{-value} < 4.9 \times 10^{-324}$; $r_{\text{latitude-PC1}} = 0.58$, $P\text{-value} < 4.9 \times 10^{-324}$; Pearson $r_{\text{longitude-PC2}} = 0.40$, $P\text{-value} = 9.9 \times 10^{-166}$; $r_{\text{latitude-PC2}} = 0.49$, $P\text{-value} = 32 \times 10^{-272}$). Owing to the ambiguity between the East Asian and Southeast Asian clades, we performed a separate admixture and PCA with the samples from Thailand, Vietnam, Singapore Malays and the Han Chinese populations (CHB, CHD, CHS). This indicated there is a considerable extent of genetic diversity in Thailand (**Figure S2**), where the Thai people have varying degree of Han Chinese admixture ranging from 10% to 80%²⁷ (**Figures 2 and S2**), with some that are virtually indistinguishable from the Malays despite the Malays and the Han Chinese being genetically distinct (**Figures 1B, S1**). We also observed that samples from Okinawa were genetically distinct from the other seven mainland Japanese groups^{23,28} (**Figures 1B, 2, S1**), consistent with the observation from the pairwise F_{ST} comparison (**Figure S3**) where the minimum F_{ST} between Okinawa and the seven Japanese populations was 0.55%, whereas the maximum F_{ST} between any pairs of the seven Japanese populations was only 0.04%.

Detecting genomic signatures of positive selection

To investigate whether local adaptation may have driven the genetic diversity between Asian populations, we selected 33 populations from ADP with more than 10 samples each and analyse each population for evidence of positive selection with haploPS as it exhibits higher statistical power at

locating genomic signatures of positive selection across different derived allele frequencies, while possessing lower false positive rates across the majority of the allelic spectrum, compared to iHS and XP-EHH ¹⁴ (see Discussion and Supplementary Material). Our analysis with haploPS identified 693 signals in total from the 33 populations which were the basis for all subsequent analyses.

Quantifying the distribution of selection signals in Asia

We observed that the 693 selection signals (**Table S3**) were located in 140 distinct genomic regions (**Table S4**), where the majority of the signals were actually present in multiple populations but located in the same genomic regions. We thus investigated the pattern of sharing of the 140 selected regions among the 33 populations, by calculating a similarity matrix for the 33 populations where each entry indicates the number of regions out of 140 that are positively selected in the respective population pair. When a PCA was performed on this similarity matrix, the first two leading principal components indicated that population groups that were geographically closer were more likely to share the same selection signals than population groups located further apart (**Figure 3**). We observed that populations in the South and Central Regions can similarly be grouped separately by the extent of sharing of selection signals, whereas there were considerable overlaps between the populations in the East and Southeast Regions.

Subsequently, we investigated whether sharing of selection signals was more likely between specific population groups. Each of the 140 regions was categorized according to whether it was uniquely present in one population group, or whether it was present in specific combinations of two or more population groups (**Figure 4A**). Of the 140 regions, 76 were only found in populations in one unique population group (**Table S4**). We observed that the greatest number of clade-specific signals was present in the South Region which predominantly comprised of South Asian populations. This concurred with the previous observations that the South Asian populations are most distinct from the

rest of the Asian populations. The least number of clade-specific signals were in the Southeast Region, comprising predominantly of Southeast Asian populations. It is likely due to the greater extent of admixture observed between the populations in East and Southeast regions.. For signals that were present in populations from three population groups, the majority were located outside the South Region, in agreement with earlier evidence that South Asian populations are considerably more distinct from the other Asian populations we have considered.

There were also five regions that were positively selected in all four population groups, which included two regions in the Major Histocompatibility Complex (MHC) on chromosome 6 that were similarly identified to be selected in CEU. The remaining three regions were only selected in our Asian populations, and none of the five regions were selected in YRI, highlighting the rarity for evolutionary signals to be shared between Africans and populations from all four Asian groups.

When haploPS identifies a selection signal, it also estimates the frequency of the (unknown) advantageous allele that is driving the selection. By grouping the selected signals according to the estimated frequencies into three categories (low: $< 30\%$; medium: $30\%–80\%$; and high: $\geq 80\%$), we observed that the majority of the signals that were present in all four ancestry clades were high frequency in nature (**Figure 4B**).

Sharing of selection signals: single-origin selection and convergent evolution across Asian populations

One feature of haploPS is the identification of the selected haplotype form, on which the beneficial allele is expected to reside. For a genomic region that is positively selected in multiple populations, haploPS compares theselected haplotype forms to deduce whether the shared selection signals were likely to have emerged from the same evolutionary event before the populations diverged. Should the selected haplotype forms be considerably different, then this suggests that separate mutation events

have occurred independently on different haplotype backgrounds in different populations, a phenomenon defined as convergent evolution. We utilized the haplotype similarity index (HSI) to infer whether the selected haplotype forms across multiple populations were likely to have descended from the same founder haplotype (defined as $HSI \geq 0.98$) and thus indicative of a single-origin mutation event, or whether fundamentally different haplotype backgrounds exist ($HSI < 0.90$) due to convergent evolution.

Of the 140 regions, 53 were positively selected in at least two populations from any of the four ancestry clades and have more than 20 SNPs in the overlapping regions. The majority (47/53) were inferred to have stemmed from the same evolutionary events given the HSIs for these regions were at least 0.98 (**Table S5**). These included the classical signals: (i) *EDAR* on chromosome 2 which was positively selected in 11 ADP populations, including the majority of Japanese and Han Chinese populations, and which exhibited perfectly concordant haplotype forms ($HSI = 1.00$) across all 11 populations (**Figure 5A**); and (ii) the skin pigmentation gene *KITLG* on chromosome 12 which was selected in 11 populations, and which similarly exhibited perfectly concordant haplotype forms ($HSI = 1.00$, **Figure 5B**). The *KEL* gene that encodes the KELL blood group antigens was also observed to be selected in 12 populations from East Asian, Central Asian and Southeast Asian, likely stemming from the same evolutionary event as well given a HSI of 1.00 (**Figure S4**).

Of the 53 regions that were positively selected in at least two populations, six regions demonstrated evidence of convergent evolution (**Table 1**), including a region spanning between 43.4Mb and 44.1Mb on chromosome 2 that encapsulated genes reported to be associated with lipid metabolism (*THADA*, *ABCG5*, *ABCG8*) and were positively selected in Japanese (Amagasaki, Ehime, Nagoya, Fukuoka) and Southeast Asians (Thailand, Vietnam).

Evidence of localised adaptation and ancestry-specific selection

In addition to the genomic regions that were positively selected in multiple ancestry clusters, there were also 76 regions that were uniquely selected in just one ancestry cluster in the ADP. *ABCC11* was identified to be positively selected in CHB and JPT. This gene carries a nonsynonymous SNP (hg19 chr16:g.48258198C>T,rs17822931) where homozygous carriers of the thymine allele possess dried earwax²⁹ and are more likely to be neutrally nonodorous³⁰, which has characteristically been found in East Asian populations. The thymine allele at rs17822931 was found on the selected haplotype form that was perfectly identical in the two populations (HSI = 1.00, **Figure 5C**).

Amongst the selection signals uniquely present in East Asian populations, the region containing the alcohol dehydrogenase (ADH) genes on chromosome 4 was selected in all the Japanese populations (JPT, Amagasaki, Ehime, Shimane, Nagoya, Okinawa, Fukuoka, Tokyo) and CHD. These ADH genes have been shown previously to protect against alcoholism³¹, and we observed that the selected haplotype forms from these populations were perfectly identical (HSI = 1.00).

DISCUSSION

This is the first Pan-Asian population genetics survey of 37 cosmopolitan, ethnic minority and indigenous groups in Asia, where at least half a million variants across the human genome have been genotyped. By defining the ADP populations into four population groups, we were able to evaluate how many of the identified genomic regions were positively selected in more than one population groups. Two lines of evidence have also emerged from our analyses to suggest that regions under positive selection in multiple ancestry clades were likely to be older and have emerged prior to the separation of the ancestry clades: (i) the advantageous alleles were estimated to be at higher frequencies; and (ii) the haplotype similarity indices indicated the background haplotype carrying the advantageous alleles were highly concordant even between populations of different ancestry clades.

Conversely, for regions that were positively selected uniquely in one ancestry clade, the advantageous alleles were estimated to be at lower frequencies, thus highly suggestive that the selection is very recent and confined to specific populations in the ancestry clade.

Such an unprecedented characterisation of positive selection in Asian populations relied on the confluence of three elements present in this study: (i) the deeper coverage of the genome provided by high-density genotyping arrays; (ii) 33 of the 37 populations surveyed at least 10 individuals each, thus providing greater confidence in estimating haplotype frequencies; and (iii) the availability of a highly sensitive algorithm (haploPS) to locate signatures of positive selection across almost the entire allele frequency spectrum. In addition, the ability to extract the specific haplotype form carrying the advantageous allele allows a direct comparison to be made between different populations that are similarly selected in the same genomic region, thus conferring the opportunity to determine whether the advantageous allele has emerged in a shared evolutionary event or in independent events due to convergent evolution. We highlight the contrast of the current findings with those from a previous analysis of the PASNP data, where only 59 regions were identified to be positively selected across 79 Asian populations due to the sparse data with only 50,000 markers ³. Here, we performed two separate analyses which leveraged on the density of the data: (i) the analysis of population structure relied on a common set of 30,927 SNPs across all the ADP populations; while (ii) the analysis of positive selection benefitted from the full set of SNP markers ranging from at least 500,000 SNPs up to 2.5 million in each population. The significantly higher SNP density has previously been shown to improve the detection of positive selection and to reduce false discoveries due to more accurate phasing ³.

The clustering of the ADP populations according to the PCA of the positive selection signals yielded a pattern that is remarkably similar to that obtained by the standard PCA of genome-wide SNP data. We

have attributed this to the sharing and differentiation of positive selection signals between the populations in the four ancestry clades. However, an alternative explanation is that haploPS may possess a high false positive rate that resulted in neutrally evolving loci accounting for the majority of the discovered sites under selection. This inevitably meant the discovered regions were but a random sampling of genome-wide data, thus reproducing what has been seen in the conventional PCA of genome-wide SNP data.

To investigate the extent of false positives produced by haploPS (as compared to well-known methods such as iHS, XP-EHH), we performed a series of simulations producing demographies representative of neutral loci under three varying degrees of bottlenecks in Asian populations, and examined how often each of the three methods erroneously report evidence of positive selection. These relied on the same thresholds used to compare the statistical power of the three methods for successfully discovering sites under selection ¹⁴. We observed that haploPS reported significantly lower false positive rates across all three scenarios (**Table S6**). In particular, at stronger bottlenecks corresponding to inbreeding coefficients of 0.2 and 0.3, haploPS reported false positive rates of 5.8% and 13.3% respectively, compared to those by iHS (17.2% and 55.3% respectively) and XP-EHH (4.4% and 8.7% respectively). Notably, the investigation into the sharing and differentiation of positive selection signals was only made possible with the advent of haploPS ^{3,14,32,33}, given the higher statistical power with accompanying lower Type I errors.

Our study also presented multiple lines of evidence to suggest that South Asian populations were considerably different from the rest of the Asian populations, which concurred with previous findings that South Asia is considerably more diverse than other Asian regions ^{32,34,35}. This is contrary to the observations between populations in the other three ancestry clades, where the admixture analyses and principal component analyses indicated there is a varying degree of mixing between these clades.

The distinction between East and Southeast Asian populations was particularly ambiguous, in part due to Southeast Asian populations such as the Vietnamese Kinh and the Thais that were genetically close to the Han Chinese from East Asia.

Our survey of positive selection in the ADP populations has essentially relied on the detection of classic selective sweeps that produce distinctive imprints of long haplotypes in the genome.

Hernandez and colleagues reported that such “hard” selective sweeps are likely to be rare in contrast to polygenic selection or selection on standing variation ³⁶, although we emphasize that our identification of 140 selected regions across 33 Asian populations is in line with their estimate that less than 10% of nonsynonymous variants are likely to be advantageous. Furthermore, we were able to consistently identify well-characterized signatures of positive selection, such as at *EDAR* which has been functionally established to influence physical traits predominantly found in East Asians ³⁷⁻³⁹, which is a valuable reassurance providing proof-of-concept for our findings.

The ADP is the second genomics research network to be developed in Asia following the PASNP, and the present coverage of Asia by the ADP is noticeably incomplete. Numerous populations in Asia are rapidly shrinking in size, where an entire community can possess less than 50 individuals found within a few consanguineous households. Global efforts studying the population genetics of early human populations have focused predominantly on African communities, and surveys of Asian populations, particularly the indigenous groups in many parts of Southeast Asia, remained sparse and fragmented. Given the significant variations of climate, culture and environment in Asia, the population genetics of Asian populations remained surprisingly unexplored. Valuable insights into human migration and possible links between Asian populations and distant hominin cousins such as the Neanderthals and Denisovans are starting to emerge ⁴⁰⁻⁴⁴, and this may present the much-needed motivation to coordinate efforts to study the population genetics of Asia.

REFERENCES

1. Abdulla MA, Ahmed I, Assawamakin A *et al*: Mapping human genetic diversity in Asia. *Science* 2009; **326**: 1541-1545.
2. Qian W, Deng L, Lu D, Xu S: Genome-wide landscapes of human local adaptation in Asia. *PLoS One* 2013; **8**: e54224.
3. Liu X, Saw WY, Ali M, Ong RT, Teo YY: Evaluating the possibility of detecting evidence of positive selection across Asia with sparse genotype data from the HUGO Pan-Asian SNP Consortium. *BMC Genomics* 2014; **15**: 332.
4. Qi X, Cui C, Peng Y *et al*: Genetic evidence of paleolithic colonization and neolithic expansion of modern humans on the tibetan plateau. *Mol Biol Evol* 2013; **30**: 1761-1778.
5. Hanaoka M, Droma Y, Basnyat B *et al*: Genetic variants in EPAS1 contribute to adaptation to high-altitude hypoxia in Sherpas. *PLoS One* 2012; **7**: e50566.
6. Simonson TS, McClain DA, Jorde LB, Prchal JT: Genetic determinants of Tibetan high-altitude adaptation. *Hum Genet* 2012; **131**: 527-533.
7. Peng Y, Yang Z, Zhang H *et al*: Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Mol Biol Evol* 2011; **28**: 1075-1081.
8. Simonson TS, Yang Y, Huff CD *et al*: Genetic evidence for high-altitude adaptation in Tibet. *Science* 2010; **329**: 72-75.
9. Suo C, Xu H, Khor CC *et al*: Natural positive selection and north-south genetic diversity in East Asia. *Eur J Hum Genet* 2012; **20**: 102-110.
10. Piel FB, Patil AP, Howes RE *et al*: Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nat Commun* 2010; **1**: 104.
11. Karlsson EK, Harris JB, Tabrizi S *et al*: Natural selection in a bangladeshi population from the cholera-endemic ganges river delta. *Sci Transl Med* 2013; **5**: 192ra186.
12. Voight BF, Kudaravalli S, Wen X, Pritchard JK: A map of recent positive selection in the human genome. *PLoS Biol* 2006; **4**: e72.

13. Sabeti PC, Varilly P, Fry B *et al*: Genome-wide detection and characterization of positive selection in human populations. *Nature* 2007; **449**: 913-918.
14. Liu X, Ong RT, Pillai EN *et al*: Detecting and Characterizing Genomic Signatures of Positive Selection in Global Populations. *Am J Hum Genet* 2013.
15. Frazer KA, Ballinger DG, Cox DR *et al*: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851-861.
16. Teo YY, Sim X, Ong RT *et al*: Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res* 2009; **19**: 2154-2162.
17. Xing J, Watkins WS, Shlien A *et al*: Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. *Genomics* 2010; **96**: 199-210.
18. Xing J, Watkins WS, Hu Y *et al*: Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biol* 2010; **11**: R113.
19. Altshuler DM, Gibbs RA, Peltonen L *et al*: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52-58.
20. Browning BL, Browning SR: A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 2009; **84**: 210-223.
21. Xu S, Li S, Yang Y *et al*: A genome-wide search for signals of high-altitude adaptation in Tibetans. *Mol Biol Evol* 2011; **28**: 1003-1011.
22. Lou H, Li S, Yang Y *et al*: A map of copy number variations in Chinese populations. *PLoS One* 2011; **6**: e27341.
23. Sato T, Nakagome S, Watanabe C *et al*: Genome-wide SNP analysis reveals population structure and demographic history of the ryukyu islanders in the southern part of the Japanese archipelago. *Mol Biol Evol* 2014; **31**: 2929-2940.
24. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904-909.
25. Alexander DH, Novembre J, Lange K: Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 2009; **19**: 1655-1664.
26. Browning SR, Browning BL: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007; **81**: 1084-1097.

27. Wangkumhang P, Shaw PJ, Chaichoompu K *et al*: Insight into the peopling of Mainland Southeast Asia from Thai population genetic structure. *PLoS One* 2013; **8**: e79522.
28. Yamaguchi-Kabata Y, Nakazono K, Takahashi A *et al*: Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am J Hum Genet* 2008; **83**: 445-456.
29. Yoshiura K, Kinoshita A, Ishida T *et al*: A SNP in the ABCC11 gene is the determinant of human earwax type. *Nat Genet* 2006; **38**: 324-330.
30. Rodriguez S, Steer CD, Farrow A, Golding J, Day IN: Dependence of deodorant usage on ABCC11 genotype: scope for personalized genetics in personal hygiene. *J Invest Dermatol* 2013; **133**: 1760-1767.
31. Chen CC, Lu RB, Chen YC *et al*: Interaction between the functional polymorphisms of the alcohol-metabolism genes in protection against alcoholism. *Am J Hum Genet* 1999; **65**: 795-807.
32. Ali M, Liu X, Pillai EN *et al*: Characterizing the genetic differences between two distinct migrant groups from Indo-European and Dravidian speaking populations in India. *BMC Genet* 2014; **15**: 86.
33. 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**: 375-392.
34. Reich D, Thangaraj K, Patterson N, Price AL, Singh L: Reconstructing Indian population history. *Nature* 2009; **461**: 489-494.
35. Wong LP, Lai JK, Saw WY *et al*: Insights into the genetic structure and diversity of 38 South Asian Indians from deep whole-genome sequencing. *PLoS Genet* 2014; **10**: e1004377.
36. Hernandez RD, Kelley JL, Elyashiv E *et al*: Classic selective sweeps were rare in recent human evolution. *Science* 2011; **331**: 920-924.
37. Kamberov YG, Wang S, Tan J *et al*: Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* 2013; **152**: 691-702.
38. Fujimoto A, Kimura R, Ohashi J *et al*: A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet* 2008; **17**: 835-843.
39. Kimura R, Yamaguchi T, Takeda M *et al*: A common variation in EDAR is a genetic determinant of shovel-shaped incisors. *Am J Hum Genet* 2009; **85**: 528-535.
40. Reich D, Green RE, Kircher M *et al*: Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 2010; **468**: 1053-1060.

41. Krause J, Fu Q, Good JM *et al*: The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* 2010; **464**: 894-897.
42. Lipson M, Loh PR, Patterson N *et al*: Reconstructing Austronesian population history in Island Southeast Asia. *Nat Commun* 2014; **5**: 4689.
43. Reich D, Patterson N, Kircher M *et al*: Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet* 2011; **89**: 516-528.
44. Shah AM, Tamang R, Moorjani P *et al*: Indian Siddis: African descendants with Indian admixture. *Am J Hum Genet* 2011; **89**: 154-161.

FIGURE LEGENDS

Figure 1. Coverage and population structure of Asian Diversity Project

(A) Illustration of the coverage of the Asian Diversity Project, where each circle highlights the location of origin for a specific population group with the assigned name, chosen using either existing naming conventions or by country name. The number accompanying each circle in the figure legend below indicates the number of individuals surveyed in that particular population. (B) A nearest-neighbor phylogenetic analysis of the 39 populations, where except for CEU and YRI, all the 37 ADP populations have been binned into four groups as shown by the colored segments. (C) Principal component analysis of the 37 ADP populations using a common set of 30,927 SNPs in the 39 populations.

Figure 2. Admixture analysis of ADP populations

Analysis of admixture with the program ADMIXTURE, where the number of ancestry ranges from $K = 2$ to $K = 12$. The populations are binned into four population groups, corresponding to the South, Central, East and Southeast Regions where each region comprises predominantly but not solely populations located in the corresponding geographical region in Asia.

Figure 3. Clustering of ADP populations by positive selection signals

A principal component analysis (PCA) on the extent of sharing of the 140 positively selected genomic regions across 33 ADP populations with at least 10 samples each. This effectively relies on the absence and presence of selection signals between the populations to place them in the biplot, where populations are found closer if they possess a greater degree of sharing of the 140 signals.

Figure 4. Sharing patterns of selection signals in the four population groups

(A) Of the 140 genomic regions that are found to be positively selected in the 33 ADP populations, 53 were present in at least two populations. This illustrates whether each genomic region is selected only in populations from the same group (red bars), or whether the region is selected in populations from two (yellow bars), three (green bars) or all four groups (blue bars). The numbers in brackets correspond to the total number of populations considered in that particular cluster. (B) This illustrates the population frequency spectrum of the advantageous alleles in each of the 140 genomic regions that were found to be positively selected in the 33 ADP populations was estimated, and binned according to whether the region was positively selected in populations from only one group, or from two, three or four groups. The frequency was also binned as low (<30%), medium (30% - 80%) and high ($\geq 80\%$).

Figure 5. Selected haplotype forms at *EDAR*, *KITLG* and *ABCC11*

HaploPS identified the background haplotypes that corresponded to the identification of positive selection at (A) *EDAR* in 17 populations, (B) *KITLG* in 19 populations and (C) *ABCC11* in 5 populations. For each of the three regions, the identified haplotypes were found to be identical (HSI of 1.00), thus indicating that there is a single shared advantageous variant that is likely to have arisen in the same mutation event prior to the splitting of the populations.

Figure 1.

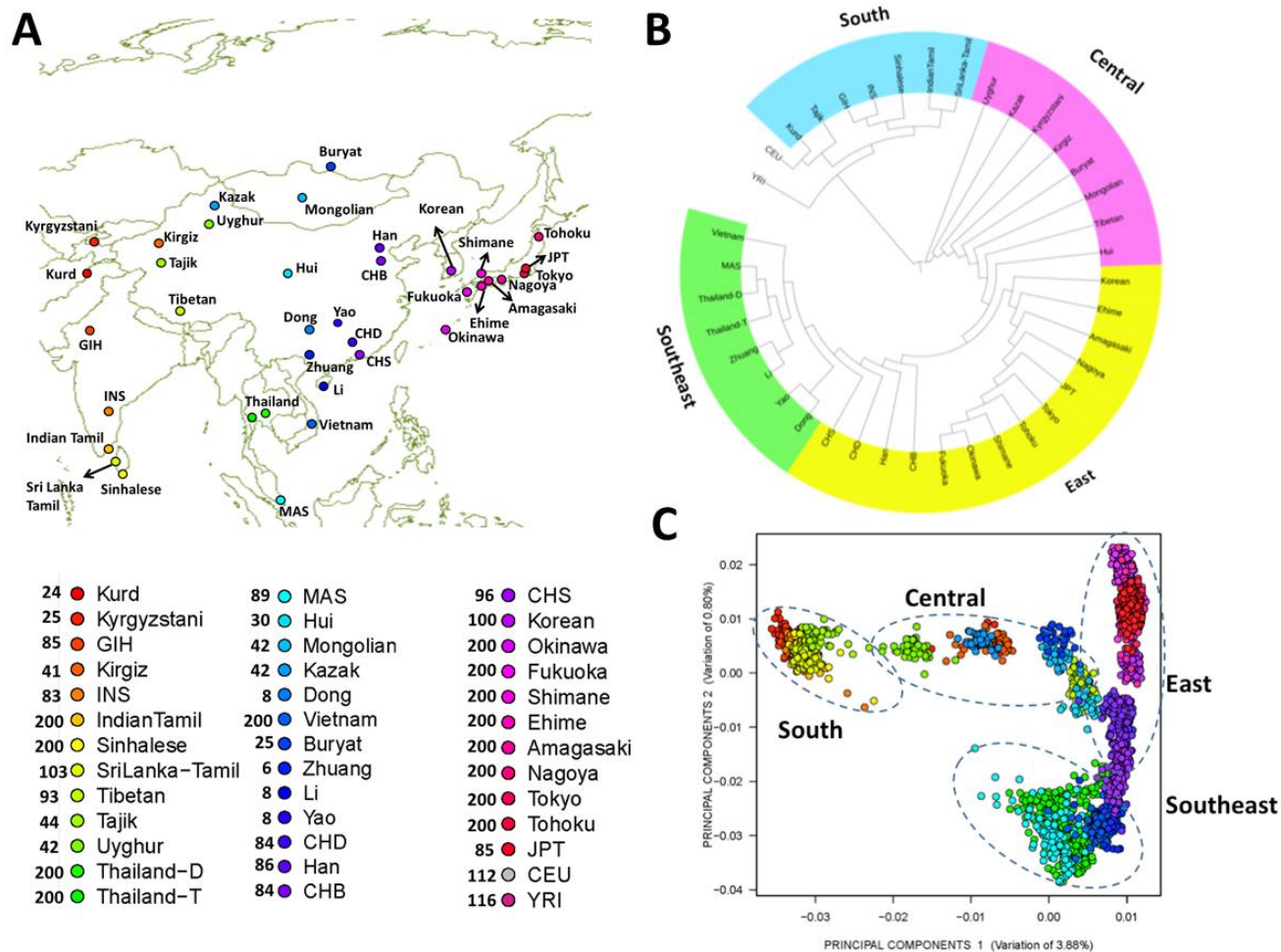


Figure 2.

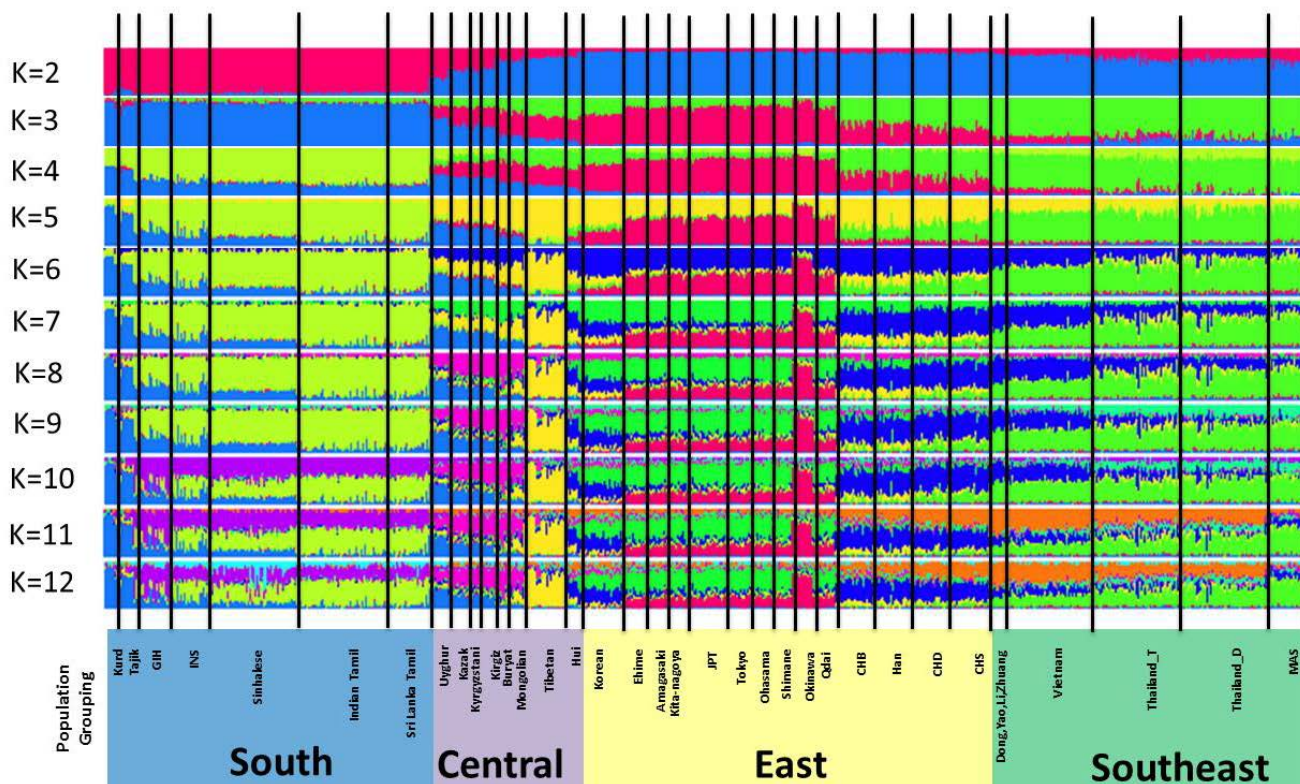


Figure 3.

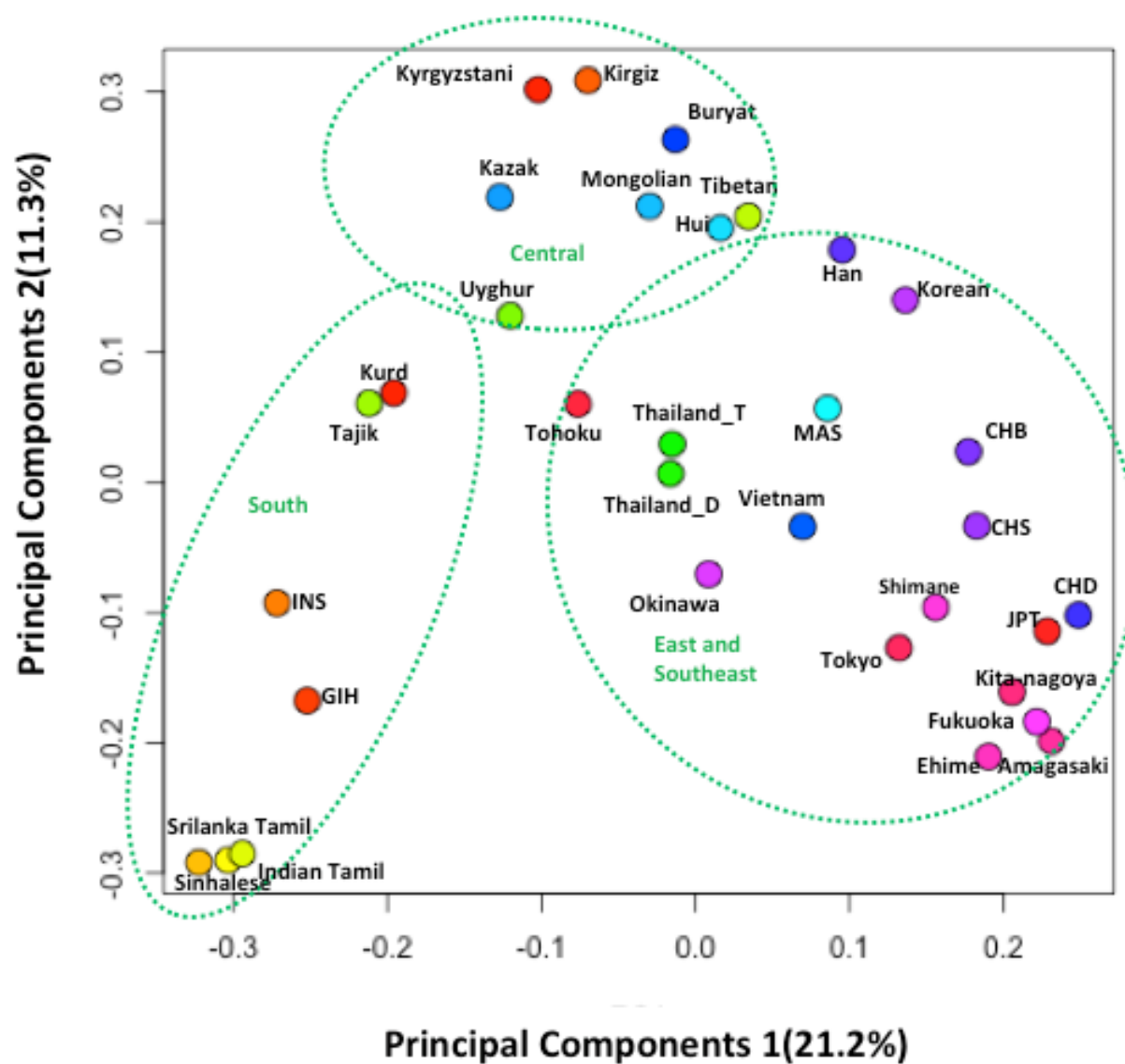
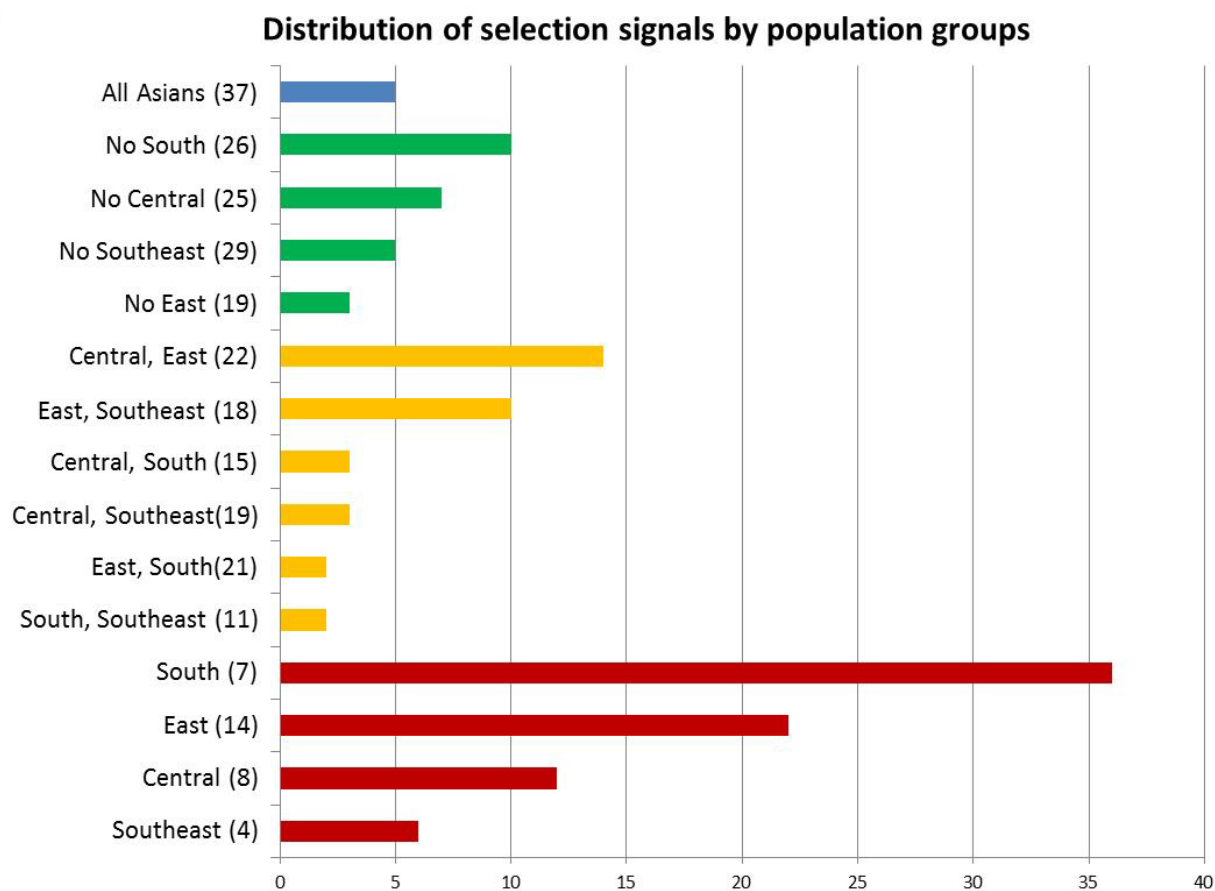


Figure 4.

A



B

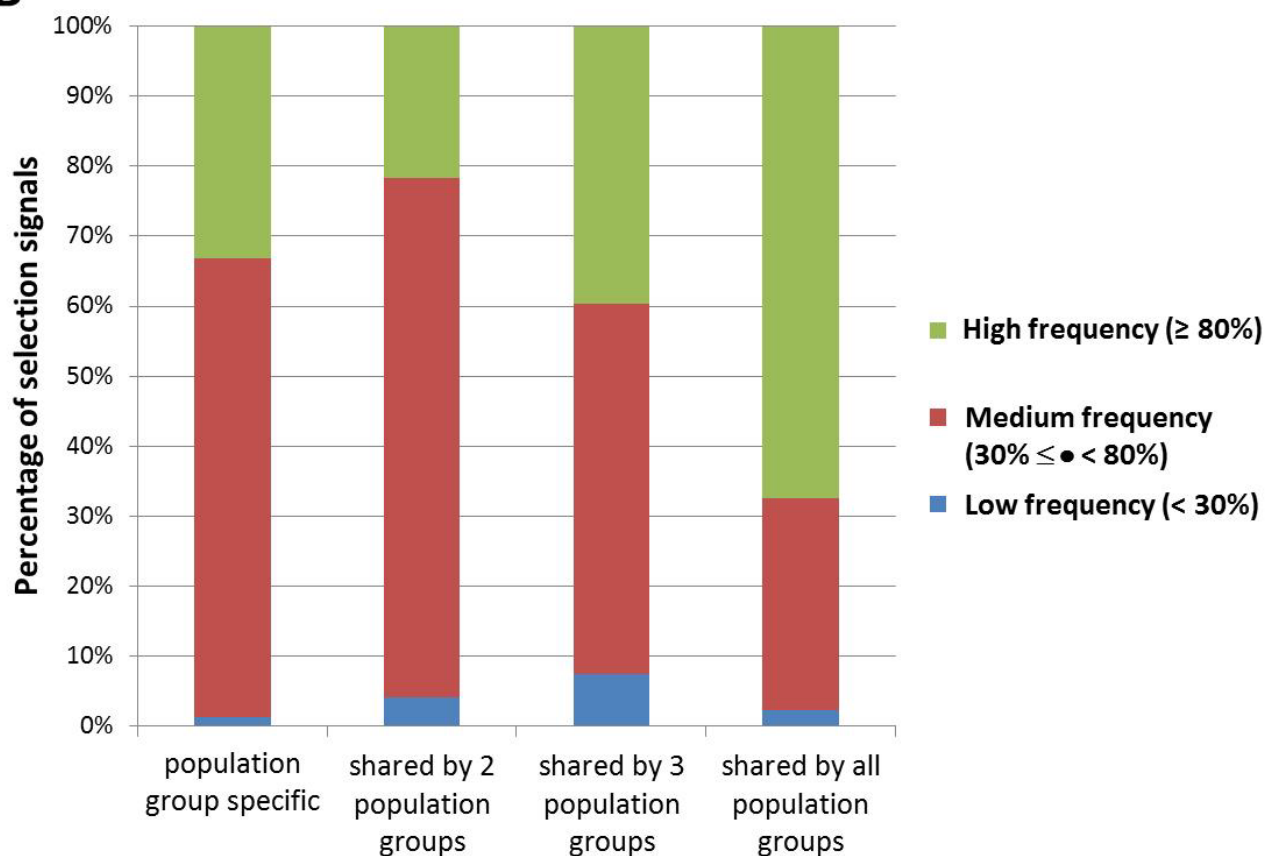
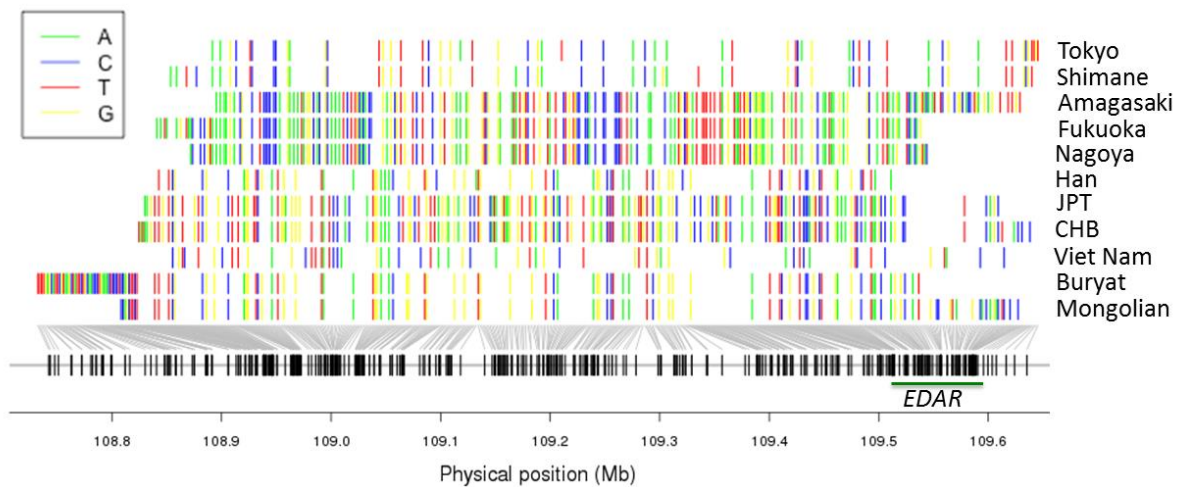


Figure 5.

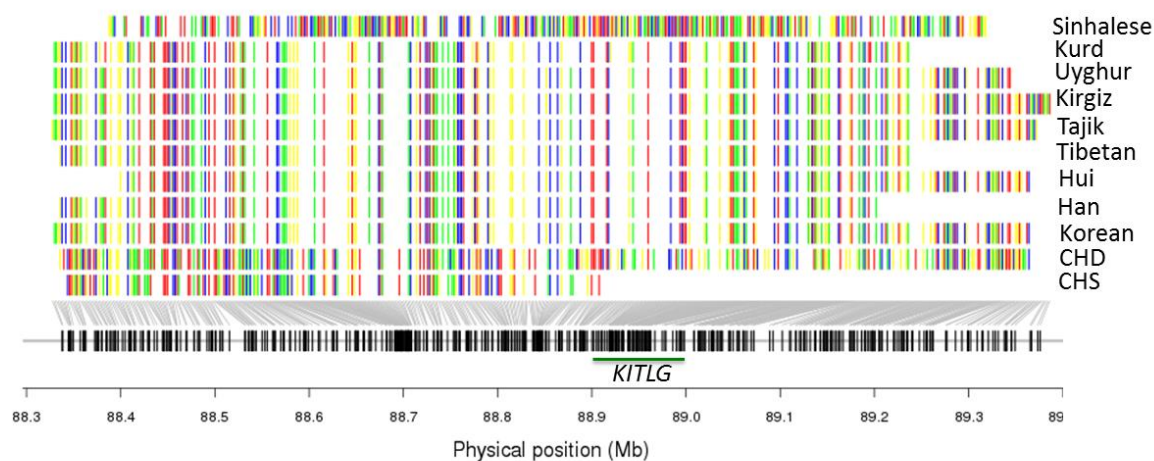
A

EDAR: chromosome 2



B

KITLG: chromosome 12



C

ABCC11: chromosome 16

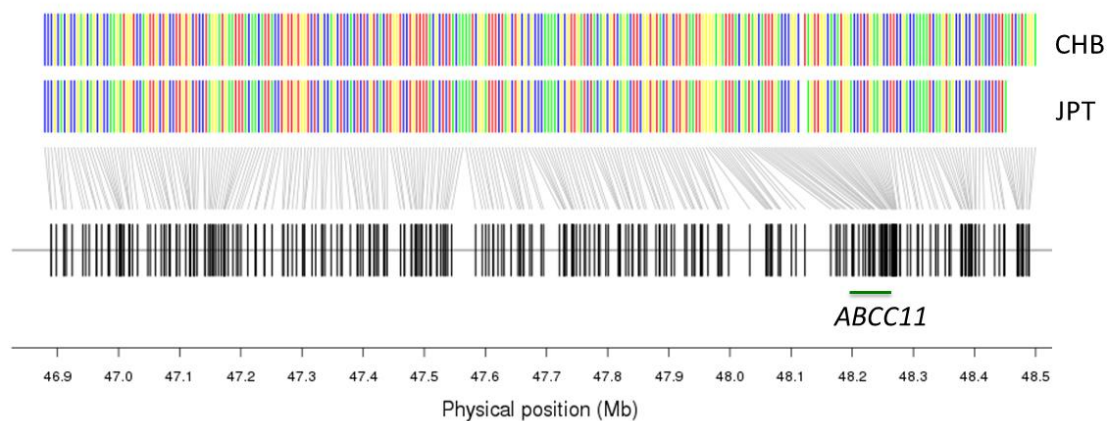


Table 1. Genomic regions identified by haploPS to be under convergent evolution in specific populations in the Asian Diversity Project. Genomic coordinates are in Human Genome version 19.

Chr	Start	End	Pop	#SNPs	HSI	Genes
1	154,811,127	155,860,091	JPT, Okinawa	73	0.50	<i>KCNN3, PMVK, PBXIP1, PYGO2, SHC1, CKS1B, MIR4258, FLAD1, LENE, ZBTB7B, DCST2, DCST1, ADAM15, EFNA4, EFNA3, EFNA1, SLC50A1, DPM3, KRTCAP2, TRIM46, MUC1, MIR92B, THBS3, MTX1, GBAP1, GBA, FAM189B, SCAMP3, CLK2, HCN3, PKLR, FDPS, RUSC1, RUSC1, ASH1L, MIR555, POU5F1P4, ASH1L, MSTO1, MSTO2P, YY1AP1, DAP3, GON4L, SYT11</i>
2	43,446,616	44,107,742	Vietnam, Thailand_T, Fukuoka, Kita-nagoya, Ehime, Amagasaki	27	0.67	<i>ZFP36L2, THADA, PLEKHH2, DYNC2LI1, ABCG5, ABCG8</i>
2	219,272,294	219,703,298	MAS, CHD, Ehime, Amagasaki	51	0.50	<i>VIL1, USP37, RQCD1, PLCD4, ZNF142, BCS1L, RNF25, STK36, TTLL4, CYP27A1, PRKAG3,</i>
3	44,092,609	45,132,637	Buryat, CHB, Amagasaki, Tibetan, Hui, Han	28	0.83	<i>MIR138, TOPAZ1, TCAIM, ZNF445, ZKSCAN7, ZNF660, ZNF197, ZNF35, ZNF502, ZNF501, KIAA1143, KIF15, MIR564, TMEM42, TGM4, ZDHHC3, EXOSC7, CLEC3B, CDCP1</i>
4	170,165,979	170,787,254	CHD, Tokyo, Amagasaki	34	0.67	<i>SH3RF1, NEK1, CLCN3, C4orf27,</i>
11	91,483,121	91,684,395	SriLanka_Tamil, Sinhalese, Amagasaki	260	0.67	NA