Detecting Genetic Ancestry and Adaptation in the Taiwanese Han People

Yun-Hua Lo,†,1 Hsueh-Chien Cheng,†,1 Chia-Ni Hsiung,2 Show-Ling Yang,2 Han-Yu Wang,1 Chia-Wei Peng,1 Chun-Yu Chen,1 Kung-Ping Lin,1 Mei-Ling Kang,1 Chien-Hsiun Chen,2 Hou-Wei Chu,1 Chiao-Feng Lin,3 Mei-Hsuan Lee,4 Quintin Liu,5 Yoko Satta,5 Cheng-Jui Lin,6 Marie Lin,6 Shu-Miaw Chaw,7 Jun-Hun Loo,6,6 Chen-Yang Shen,∗,2 and Wen-Ya Ko∗,1

1Faculty of Life Sciences and Institute of Genome Sciences, National Yang-Ming University, Taipei, Taiwan
2Institute of Biomedical Sciences, Academia Sinica, Taipei City, Taiwan
3Research Team, DNAnexus, Mountain View, CA
4Institute of Clinical Medicine, National Yang-Ming University, Taipei, Taiwan
5Department of Evolutionary Studies of Biosystems, SOKENDAI (The Graduate University for Advanced Studies), Hayama, Japan
6Molecular Anthropology and Transfusion Medicine Research Laboratory, Mackay Memorial Hospital, Taipei, Taiwan
7Biodiversity Research Center, Academia Sinica, Taipei City, Taiwan
†These authors contributed equally to this work.
*Corresponding authors: E-mails: loojunhun@gmail.com; bmcys@ibms.sinica.edu.tw; wenko@ym.edu.tw.
Associate editor: Michael Rosenberg

Abstract

The Taiwanese people are composed of diverse indigenous populations and the Taiwanese Han. About 95% of the Taiwanese identify themselves as Taiwanese Han, but this may not be a homogeneous population because they migrated to the island from various regions of continental East Asia over a period of 400 years. Little is known about the underlying patterns of genetic ancestry, population admixture, and evolutionary adaptation in the Taiwanese Han people. Here, we analyzed the whole-genome single-nucleotide polymorphism genotyping data from 14,401 individuals of Taiwanese Han collected by the Taiwan Biobank and the whole-genome sequencing data for a subset of 772 people. We detected four major genetic ancestries with distinct geographic distributions (i.e., Northern, Southeastern, Japonic, and Island Southeast Asian ancestries) and signatures of population mixture contributing to the genomes of Taiwanese Han. We further scanned for signatures of positive natural selection that caused unusually long-range haplotypes and elevations of hitchhiked variants. As a result, we identified 16 candidate loci in which selection signals can be unambiguously localized:

- CTNNA2
- LRP1B
- CSNK1G3
- ASTN2
- NEO1

Statistical associations were examined in 16 metabolic-related traits to further elucidate the functional effects of each candidate gene. All five genes appear to have pleiotropic connections to various types of disease susceptibility and significant associations with at least one metabolic-related trait. Together, our results provide critical insights for understanding the evolutionary history and adaption of the Taiwanese Han population.

Key words: ancestry, admixture, adaptation, natural selection, Taiwanese Han.

Introduction

Disease susceptibility differs greatly between populations and appears to be correlated with human population history (Chen et al. 2012; Corona et al. 2013). However, owing to the complex history of human migration, most contemporary populations are genetically admixed, which could complicate the efforts of genetic profiling for susceptibility to diseases (Gravel 2012; Kidd et al. 2012; Marnetto et al. 2020). Therefore, understanding the genetic ancestry, population substructure, and migration history of people who live in the same geographic region may allow us to better characterize the admixed ancestry for each individual genome, providing critical information to facilitate genome-wide association studies for mapping disease-causing variants. Disease susceptibility may also arise as side effects of evolutionary adaptation. Under a certain selection pressure (e.g., malaria), genetic adaptation could increase an individuals' fitness in terms of survival or reproductive success, but this could sometimes be accompanied with the cost of the carriers' health (Haldane 1932). Sickle-cell anemia, thalassemia, and APOL1-mediated kidney diseases are among the most noticeable examples in which carriers of the respective disease-causing variants confer protective effects against parasitic infection (Kwiatkowski 2005; Weatherall 2008; Genovese et al. 2010; Ko et al. 2012, 2013). Therefore, detection of genomic signatures of evolutionary adaptation provides an alternative approach to shed light on the biological
mechanisms underlying disease susceptibility (Lachance and Tishkoff 2013; Vasseur and Quintana-Murci 2013).

Taiwan is home to a diversity of human ethnic groups that can be roughly grouped into three major populations. Taiwanese Han people are the descendants of early immigrants (mainly Minnan and Hakka) who migrated from Southern China in the last 400 years and were recently joined by many immigrants from various geographic areas of China at the end of World War II in 1945 (Dittmer 2004). The second major group contains 16 officially recognized indigenous populations, representing 2.3% of the total population in Taiwan. These indigenous tribes harbor rich genetic diversity and have been considered as the ancestral lineages of Austronesian-speaking people (Trejaut et al. 2005; Soares et al. 2011; Ko et al. 2014; Lipson et al. 2014; Trejaut et al. 2014; Chang, Liu, et al. 2015; Soares et al. 2016). Finally, the third group, Taiwan plain aborigines (Pingpu), includes many tribes that previously inhabited plains across the island of Taiwan. Although they are thought to be descendants of Austronesian-speaking people, most of these tribes may have admixed with the Taiwanese Han people (Trejaut et al. 2005, 2014). However, the extent of contribution of genetic diversity from the Pingpu aborigines to the Taiwanese Han, as well as the degree of population mixture between the current Taiwanese Han and indigenous populations, is unclear.

In this study, we analyzed the Axiom Genome-Wide (whole-genome [WGI]) TWB genotyping array (650k single-nucleotide polymorphisms [SNPs]) in 14,401 individuals from the Taiwanese Han population and the WGS sequencing data for a subset of 772 people collected by the Taiwan Biobank. As a result, we detected four major genetic ancestries (with distinct geographic ranges) in the Taiwanese Han and revealed signatures of ancient population mixture before they migrated to Taiwan. We further scanned for genomic signatures of positive selection by summarizing the lengths of extended haplotype (using Integrated Haplotype Score [iHS]) and shapes of genealogy surrounding the selection-candidate loci (using iSAFE). Consequently, we identified 16 loci targeted by positive natural selection in which selection signatures can be localized unambiguously at five candidate genes. For each of the five candidate genes, we further performed multiple linear regression analyses with 16 metabolic-related traits and discussed the possible role of each gene in adaptive evolution and connections with disease susceptibility.

Results
Characterizing Genetic Structure and Ancestry in the Taiwanese Han and Neighboring Populations
ADMIXTURE analysis was conducted to characterize the patterns of genetic structure across 99 Asian populations by merging the Pan-Asia and Human Genome Diversity Project (HGDP) SNP genotyping data sets for a total of 19,290 intersected SNPs in 2,304 people (Li et al. 2008; Abdulla et al. 2009). We ran ADMIXTURE for $K = 2–30$ where $K$ is the number of ancestral populations assumed in the model and found that the $K$ value with the lowest cross-validation error (CVE) is 19. Because the inferred patterns of ancestral components (ACs) among the populations of interest are similar for $K \geq 13$, we chose the most parsimonious model $K = 13$ to summarize the results of Taiwanese populations, including Hakka (TW-HA), Minna (TW-HB), Ami (AX-AM), and Atayal (AX-AT), together with the other Sino-Tibetan speaking populations and several neighboring populations. We designated different colors to different ACs identified in our study (fig. 1A). In general, the patterns of genetic structure can be distinguished broadly into different language groups. For example, blue is the predominant AC for most of the Sino-Tibetan speaking populations, whereas yellow is predominant for the Turkic/Tungusic/Mongolic/Koreanic people, green for the Ryukyu and main-island Japanese (Japonic), and pink for the Austronesian speaking populations (e.g., Ami and Atayal). Particularly, the ancestries for the Taiwanese Hakka/Minna are mainly composed of these four aforementioned ACs with average proportions of 46%/45%, 23%/21%, 18%/21%, and 10%/10% for the blue, yellow, pink, and green ACs, respectively (fig. 1B). We confirmed these Taiwanese Hakka/Minna populations together with the Taiwanese Han since they are genetically very close to each other. In contrast, a very distinct pattern of genetic structure was observed for the other two Taiwanese indigenous populations (Ami and Atayal) who carry two major ancestries (pink and blue), with an average of 77% for the pink AC in both populations and 21% and 18% for the blue AC in Ami and Atayal, respectively (fig. 1B). Several neighboring populations within the Sino-Tibetan language group also displayed similar patterns of ancestry with the Taiwanese Han including Singapore Chinese (SG-CH), Chinese Cantonese (CN-GA), Chinese Han, and Tujia people (fig. 18). All of these populations live in Southeastern Asia (fig. 1C). The results for the remaining populations are provided in supplementary figure S1, Supplementary Material online.

We further investigated the geographic distribution across all populations with average ancestry proportion $\geq 3\%$ in each of the four ACs and detected distinct geographic distributions among them (see fig. 1C). We designated the yellow AC as the Northern ancestry since its proportion increases with latitude of population location ($\rho = 0.74, P = 1.4 \times 10^{-10}$ for Spearman’s rank correlation). The proportion of blue AC appeared to be significantly correlated with latitude ($\rho = 0.39, P = 0.0014$) and scatters around the region of Southeast Asia (referred as the Southeastern ancestry). The proportion of green AC (referred to as the Japonic ancestry) are significantly correlated with both longitude ($\rho = 0.58, P = 0.00057$) and latitude ($\rho = 0.56, P = 0.00091$), whereby the Ryukyu Japanese has the highest proportion (86%) among all populations studied, followed by the main-island Japanese (60%). Lastly, the proportion of pink ancestry is significantly correlated with latitude ($\rho = -0.63, P = 3.5 \times 10^{-8}$) and is high in many Austronesian-speaking populations living in Island Southeast Asia and Taiwan (referred to as the ISEA ancestry).
Identifying Admixed Ancestries in the Taiwanese Han People

Since the Taiwanese Han population carries a considerable proportion of Austronesian ancestry that is also high in many Austronesian-speaking populations, we attempted to detect signatures of population mixture by assuming the Ami (representing the Austronesian ancestry) as one of the two donor populations in the model of $F_3$ statistics and scanned for any other donor population that showed significant signatures of population mixture contributing to the genomes of Taiwanese Han (the recipient population). As a result, 23 out of the 99 populations in the data set showed significant negative $Z$ scores of $F_3$ after correcting for multiple tests using FDR ($Z$ scores ranged from $-9.1$ to $-2.2$). These 23 populations spread across a wide geographic range in East Asia (EA), from Siberia (Yakut people) to Singapore (SG-CH) in the South end of EA (fig. 2A). A similar pattern was also observed when the Taiwanese Han was replaced by the Chinese Han as the recipient population in $F_3$ ($Z$ scores ranged from $-8.1$ to $-2.5$). Similar outcomes were also obtained when the Atayal was used as the donor population instead of the Ami (see supplementary fig. S2, Supplementary Material online). We further applied $F_4$ which allows detection of recent gene flow between the ancestors of Taiwanese Han and indigenous Austronesian-speaking populations (represented by Ami) by testing $F_4$ (Yoruba, Ami; pop$_i$, Taiwanese Han) where pop$_i$ was selected from the 11 Sino-Tibetan speaking populations that are genetically close to Taiwanese Han (fig. 1B). Since Yoruba is the outgroup population in the test (assuming no admixture with pop$_i$ and Taiwanese Han), a significant positive $F_4$ value would suggest gene flow between the ancestors of Taiwanese Han (TWB) and Ami (AX-AM). Table 1 summarizes the results of the $F_4$ tests; all showed significant positive $F_4$ values ($Z = 3.2$–$35.3$) except for two outcomes when Singapore Chinese (SG-CH) and Chinese Cantonese (CN-GA) were used as pop$_i$, independently. These two populations appear to be genetically closest to the Taiwanese Han among all the Sino-Tibetan speaking populations (fig. 1B).

FIG. 1. Inferred genetic ancestries in the Sino-Tibetan people and their neighboring populations in East Asia. (A) Admixture results for the Sino-Tibetan and their neighboring populations. Each individual is indicated by a vertical line, which is subdivided into $K = 13$ colored segments, where $K$ is the number of ancestral populations assumed in the analysis. The $y$-axis represents the estimated ancestry proportions. Ethnicity names are labeled on the $x$-axis. The abbreviations of all populations are given in supplementary table S4, Supplementary Material online. The ADMIXTURE analysis was performed across 99 Asian populations for a total of 19,290 SNPs in 2,304 individuals, but only the Sino-Tibetan, and several neighboring populations from Altaic, Turkic, Tungusic, Mongolic, Koreanic, and Japonic linguistic groups as well as two Taiwanese Austronesian populations—Ami (AX-AM) and Atayal (AX-AT) are shown. (B) Average proportions of ancestry of these populations. (C) Geographic distributions of the four major ancestries of the Taiwanese Han are shown for the populations with average proportions $\geq 0.35$ in each ancestry. The genetic ancestries for the remaining populations are provided in supplementary figure S1, Supplementary Material online.
In addition, signatures of population mixture were also detected in the Chinese Han when the Hmong people were used (as the donor population) to represent the Southeastern ancestry (blue) using $F_3$, since they carry the highest blue ancestry proportions among all populations (0.89 and 0.82 for the Thailand and Chinese Hmong, respectively). Consequently, 13 populations were identified with significant negative values of $Z$ scores, ranging from $-2.1$ to $-7.9$ (fig. 2B). Ten of these populations speak languages belonging to the Altaic language groups (i.e., Mongolic/Tungusic/Turkic/Koreanic/Japonic) and live at relatively high latitudes (also see fig. 1C).

We characterized admixed ancestries for each individual genome across 14,401 people collected by the Taiwan Biobank (https://www.twbiobank.org.tw/new_web_en/) using ADMIXTURE together with the SNP data sets of the HGDP and eight populations from the Southeast Asia data set published by Mörseburg et al. (2016). The Pan-Asia data set was excluded from this analysis to increase the number of analyzed SNPs from 19,290 to 101,955. ADMIXTURE was run by assuming different $K$ values. The ADMIXTURE results for $K=3–9$ are provided in supplementary figure S3, Supplementary Material online, in which $K=9$ appears to fit best to the data set with the lowest CVE value. Figure 2C

Table 1. $F_4$ Test of Population Mixture for the Sino-Tibetan Speaking Populations.

| Pop1 (A) | Pop2 (B) | Pop3 (C) | Pop4 (D) | $F_4$ | Z   | P     | $P_{FDR}$ |
|----------|----------|----------|----------|-------|------|-------|-----------|
| Yoruba   | AX-AM    | SG-CH    | TWB      | 0     | 0.49 | 0.49  | 0.49      |
| Yoruba   | AX-AM    | CN-GA    | TWB      | 0.0011| 0.22 | 0.22  | 0.24      |
| Yoruba   | AX-AM    | Tujia    | TWB      | 0.0062| 3.2  | 0.00078| 0.0010   |
| Yoruba   | AX-AM    | Han      | TWB      | 0.0066| 5.3  | $7.0 \times 10^{-8}$ | $1.0 \times 10^{-7}$ |
| Yoruba   | AX-AM    | Lahu     | TWB      | 0.0164| 5.8  | $4.2 \times 10^{-9}$ | $7.3 \times 10^{-9}$ |
| Yoruba   | AX-AM    | CHB      | TWB      | 0.01  | 7.7  | $6.8 \times 10^{-15}$ | $1.4 \times 10^{-14}$ |
| Yoruba   | AX-AM    | CN-SH    | TWB      | 0.015 | 8.8  | $6.8 \times 10^{-15}$ | $<1.4 \times 10^{-14}$ |
| Yoruba   | AX-AM    | Yizu     | TWB      | 0.020 | 9.3  | $6.8 \times 10^{-15}$ | $<1.4 \times 10^{-14}$ |
| Yoruba   | AX-AM    | Naxi     | TWB      | 0.023 | 9.3  | $6.8 \times 10^{-15}$ | $<1.4 \times 10^{-14}$ |
| Yoruba   | AX-AM    | TH-KA    | TWB      | 0.026 | 11.7 | $6.8 \times 10^{-15}$ | $<1.4 \times 10^{-14}$ |
| Yoruba   | AX-AM    | IN-TB    | TWB      | 0.0823| 35.3 | $6.8 \times 10^{-15}$ | $<1.4 \times 10^{-14}$ |

**Note.**—$F_4$ was conducted by assuming $F_4(A; B; C; D)$ where the four populations are related by the unrooted population tree ((A, B), (C, D)). Population abbreviations are: AX-AM, Ami; SG-CH, Singapore Chinese; CN-GA, Chinese Cantonese; CHB, Chinese Han in Beijing; CN-SH, Chinese Han in Shanghai; and TH-KA, Thailand Karen.
illustrates the inferred ancestral proportions for each Taiwanese-Han individual from the Taiwan Biobank under the assumption of nine ancestral populations. Three major ACs were identified in the Taiwanese Han people. The Southeastern ancestry (blue) constitutes the highest percentage for most Taiwanese Han people (average 59%, range 24–64%), followed by the Northern ancestry (yellow) with 24% on average (range 8–41%). The ISEA ancestry (pink) constitutes 15% of the genome on average but varies considerably from 0.1% to 62% (fig. 2C). Since we excluded the Pan-Asia data set, the Ryukyu Japanese population (JP-RK), which contains the highest proportion of green AC (86%), was not included in this analysis and, consequently, the ancestry proportion previously attributed to the green AC can no longer be separated from the blue and yellow ACs.

We also applied fineSTRUCTURE developed by Lawson et al. (2012) to this data set (after exclusion of the Western and Southern Asian populations) for detecting any subtle population substructure within the Taiwanese Han. Due to computational limitation, only 854 individuals were included for this analysis (including 500 Han Taiwanese). Although based on the coancestry matrix, we did not observe any apparent pattern of population substructure within the Taiwan Han (see supplementary fig. S4A, Supplementary Material online), the population tree separated the Han Taiwanese into three main groups. Most of the Taiwanese individuals (472 of 500) were clustered together forming a subtree together with a few individuals of Chinese Han and the She population (group 1). Another 27 individuals (group 2) were grouped into the neighboring subtree that also contains several Northern Asian populations (e.g., Mongolia, Xibo, Hezhen, Tu, and Japanese). Finally, one Taiwanese-Han individual (group 3) was placed closer to the Dusun population (who live in Northern Borneo) on a relatively distantly related subtree that also includes several Austronesian speaking populations (supplementary fig. S4B, Supplementary Material online). These three groups differ considerably in their estimated proportions of the three major ancestries inferred by ADMIXTURE. Although the 472 Taiwanese Han people in group 1 showed similar estimates in average proportion of each ancestry, the 27 individuals in group 2 showed a significant increase in proportion of Northern ancestry (average proportion = 29%), but a decrease in proportion of ISEA ancestry (average proportion = 10%). For the last individual (group 3), the proportion of ISEA ancestry is as high as 54%, but only 10% and 34% proportion of Northern and Southeastern ancestries, respectively (supplementary fig. S4C, Supplementary Material online).

Identifying Candidate Loci Targeted by Positive Natural Selection

To detect for genome-wide signatures of positive selection in the Taiwanese Han, we applied Voight et al.’s (2016) iHS to scan for unusually long extended haplotypes using the WG SNP genotyping data of Taiwan Biobank. iHS was computed for every SNP with assured ancestral/derived information and with minor allele frequency >0.01 (Voight et al. 2006). Subsequently, iHS scores were obtained for a total of 562,983 SNPs in 14,401 individuals. The empirical distribution of |iHS| in our results is approximate to a folded standard normal distribution with top 1% value ≥2.66 (supplementary fig. S5, Supplementary Material online). Since it is well known that the EDAR gene has experienced recent positive selection in the Han population (Sabeti et al. 2007; Grossman et al. 2010; Kamberov et al. 2013), we used the observed selection signatures of EDAR in our result as the threshold for identifying other selection-candidate loci. Therefore, a SNP cluster would be considered as a selection-candidate locus if it contains ≥3 SNPs higher than 4.18 (the highest |iHS| value of EDAR) and ≥10 SNPs above the top-1% |iHS| cutoff-score (2.66) within the 500-kb range nearby the core SNP of the highest |iHS| score. As a result, selection signatures were identified in 16 genomic loci; the highest |iHS| score is rs10483453 (9.56) located at chromosome 14:35.6–36.0 Mb, encompassing multiple genes within this region, followed by rs9262558 (|iHS| = 7.5) located at the region of 28.5–33.1 Mb on chromosome 6, representing the HLA gene family (fig. 3). The number of SNPs above the cutoff of top-1% |iHS| is 25 for the former region and 325 for the HLA gene family (table 2). In addition, because the sample size is considerably large in our data set (14,401), we were able to identify candidate core SNPs from a wide range of allele frequencies. The allele frequency for the core SNP of EDAR (rs17034770) is as high as 0.89, whereas the frequency for the core SNP of LRP1B on chromosome 2 is as low as 0.05 (table 2).

A recent positive selection event not only causes unusually long blocks of shared haplotypes in the genome but also produces a “star-like” genealogy surrounding the selection-favored mutation due to an excess of newly arisen mutations (Hudson 1990). To further confirm these 16 candidate loci detected by iHS, we applied iSAFE to analyze the WG sequencing data from 772 individuals (a subset of the 14,401 individuals). The iSAFE statistic is designed to characterize the shape of genealogy for a given candidate region and to provide better resolution for identifying the underlying candidate gene/variant targeted by selection (Akbari et al. 2018). As a result, we identified five genes (out of 16 loci) with the peak of iSAFE signals pointed to a single gene. These five genes include CTNN2 (Catenin Alpha 2) and LRP1B (LDL Receptor Related Protein 1B) at chromosome 2, CSNK1G3 (Casein Kinase 1 Gamma 3) at chromosome 5, ASTN2 (Astrotactin 2) at chromosome 9, and NEO1 (Neogenin 1) at chromosome 15. Figure 4 displays the localized |iHS| and iSAFE plots for each of the five genes. An elevation of linkage disequilibrium (LD) is also noticeable underneath each candidate region. In addition, we incorporated the combined annotation-dependent depletion (CADD) score to examine the functional importance for each SNP. Based on the iSAFE scores, the top 20 candidate SNPs for each gene and their CADD scores are provided in supplementary table S1, Supplementary Material online. Of the remaining candidate loci (excluding EDAR), five appear to harbor multiple genes underneath the peak signals of iSAFE and, therefore, the selection-targeted gene cannot be determined, whereas three loci appear to be situated at the intergenic regions. Finally, selection signatures for the other two loci are less evident.
because their iSAFE peak scores are only marginally significant (≈0.1, the suggestive significance value by Akbari et al. [2018]). Table 2 summarizes the peak iSAFE scores for each candidate region and the identified candidate gene accordingly. The plots of |iHS|, iSAFE, and LD heat map for these loci are provided in supplementary figure S6, Supplementary Material online.

Analysis of Associations between Selection-Candidate Genes and Metabolic-Related Traits
For each selection-candidate gene, we performed multiple linear regression analyses in 16 metabolic-related traits by incorporating sex, age, body mass index, and PC1–8 in principal component analysis as covariates. These trait measurements were collected by the Taiwan Biobank from a series of physical/blood/urine examinations and can be broadly categorized into three functional classes including kidney/diabetic, cardiovascular, and liver functions. The mean and standard deviation of each trait are listed in supplementary table S2, Supplementary Material online. For each gene, the linear regression results were only considered for the SNPs located at the peak iSAFE-score region (“peak region” is defined by the SNPs of iSAFE scores ≥ 0.1 with an additional 50-kb extension at both ends). We next applied a LD-based...
The clumping procedure first takes SNPs that are significant at \( P / C \leq 10^{-4} \) as index SNPs. The threshold was set to correct for the number of identified candidate loci (5) multiplied by the number of traits (16). A clump was formed by including all other “clumped” SNPs that passed the second significance threshold \( P / C \leq 0.01 \) within a 250-kb distance from the index SNP and are in LD with the index SNP \( (r^2 \geq 0.5) \) (Purcell et al. 2007). Table 3 summarizes the results of the detected traits.

**Fig. 4.** Plots of \( \text{iHS} \) and \( \text{iSAFE} \) scores and LD heat maps of five selection-targeted genes in the Taiwanese Han population. The \( \text{iHS} \) and \( \text{iSAFE} \) scores were plotted against each of the five selection-candidate loci where the selection-targeted gene can be unambiguously identified (based on the \( \text{iSAFE} \) signals). In each \( \text{iSAFE} \) plot, the point size and color gradient represent \( C \) scores that were estimated to profile the degree of functional importance (deleteriousness) according to Kircher et al. (2014) and Rentzsch et al. (2019). The heat map demonstrates the pairwise estimates of LD. Each pixel represents a pairwise LD estimate using the squared correlation coefficient scaled by allele frequency \( (r^2) \). All possible pairs of polymorphic sites were measured. Levels of LD ranging from 0 to 1 are illustrated according to a white to red color gradient. The physical position of each polymorphic site is marked by a black line segment above the heat map, which is aligned with the plot of gene structures (based on the GRCh37/hg19, UCSC genome browser). The plots for the remaining candidate loci are presented in supplementary figure S6, Supplementary Material online.
Table 3. List of SNPs and Associated Metabolic-Related Traits in the Five Genes Targeted by Positive Natural Selection.

| Gene   | SNP          | Chr | Position | Trait      | Reg. Cof. | P          | ISAFE | Imp. |
|--------|--------------|-----|----------|------------|-----------|------------|-------|------|
| CTNNA2 | 2_80464202   | 2   | 80464202 | Albumin    | -1.25     | 9.3 x 10^{-5} | —     | 0.54 |
|        | rs554504577  | 2   | 80362623 | Albumin    | -0.91     | 0.0026    | —     | 0.51 |
|        | rs17018689   | 2   | 80373740 | Albumin    | 0.043     | 0.071     | 0.19  | 0.99 |
| LRP1B  | rs186045033  | 2   | 141638598| LDLC       | 0.20      | 3.3 x 10^{-5} | —     | 0.87 |
|        | rs185095358  | 2   | 141631133| SGOT       | -0.69     | 5.2 x 10^{-5} | —     | 0.91 |
|        | rs144646457  | 2   | 141580213| SGOT       | -0.69     | 5.2 x 10^{-5} | —     | 0.91 |
| CSNK1G3| 5_123001857  | 5   | 123001857| HbA1C      | -1.09     | 3.0 x 10^{-5} | —     | 0.81 |
|        | 5_122978454  | 5   | 122978454| HbA1C      | -1.09     | 3.0 x 10^{-5} | —     | 0.89 |
|        | rs79451111   | 5   | 122983696| TG         | 1.19      | 3.8 x 10^{-5} | —     | 0.52 |
|        | rs6868518    | 5   | 122837666| BUN        | 0.021     | 0.069     | 0.31  | 1.00 |
| ASN2   | rs564508867  | 9   | 119135159| FBG        | -0.69     | 1.0 x 10^{-4} | —     | 0.73 |
|        | rs888401     | 9   | 119207606| Albumin    | -0.028    | 0.028     | 0.19  | 1.00 |
|        | rs564508867  | 9   | 119135159| TG         | 0.20      | 1.0 x 10^{-4} | —     | 0.73 |
|        | rs888401     | 9   | 119207606| T-BIL      | 0.024     | 0.054     | —     | —    |
| NEO1   | 15_73481424  | 15  | 73481424 | SGOT       | 0.58      | 1.2 x 10^{-5} | —     | 0.85 |
|        | rs146077526  | 15  | 73424172 | SGOT       | 0.32      | 2.4 x 10^{-5} | —     | 0.93 |
|        | 15_73587033  | 15  | 73587033 | Creatinine | 0.72      | 8.6 x 10^{-5} | —     | 0.69 |
|        | rs8039418    | 15  | 73441432 | BUN        | 0.035     | 0.0023    | 0.41  | 0.99 |
|        | rs8039418    | 15  | 73441432 | UA         | 0.019     | 0.053     | —     | —    |

**NOTE.**—Multiple linear regressions were conducted for the iSAFE peak region (iSAFE ≥ 0.1) in each of the five genes across 16 metabolic-related traits. The SNP with an iSAFE score is the SNP of the highest iSAFE for a given candidate gene. The listed traits are albumin (g/dl), low-density lipoprotein cholesterol (LDLC, g/dl), serum level of aspartate aminotransferase (SGOT, U/l), hemoglobin A1c (HbA1c, %), triglyceride (TG, mg/dl), blood urea nitrogen (BUN, mg/dl), fasting blood glucose (FBG, mg/dl), total bilirubin (T-BIL, mg/dl), creatinine (mg/dl), and uric acid (UA, mg/dl). "Chr" represents chromosome. "Reg. cof." represents regression coefficient. "Imp." represents imputation posterior probability.

**FIG. 5.** Multiple linear regression analyses of metabolic-related traits for five selection-candidate genes in the Taiwanese Han population. For each trait, significant variants (highlighted in red) were identified based on a LD-based clumping method (Purcell et al. 2007) within the selection-targeted region (colored in dark gray). The plots for the remaining metabolic traits are presented in supplementary figure S7, Supplementary Material online.

and associated SNPs within the selection-targeted region (i.e., the peak iSAFE-score region) in each of the five genes as also shown in figure 5. CTNNA2 was found to be associated with serum albumin ($P = 9.3 \times 10^{-5}$), whereas LRP1B appears to be associated with levels of low-density lipoprotein cholesterol (LDLC) and serum level of glutamic-oxaloacetic
transaminase (SGOT) \((P = 3.3 \times 10^{-5}\) and \(5.2 \times 10^{-5}\), respectively). We also found SNPs at the iSAFE-score peak region of CSNK1G1 that are associated with hemoglobin A1c (HbA1C) and triglyceride (TG) \((P = 3.0 \times 10^{-5}\) and \(3.8 \times 10^{-5}\), respectively). Moreover, ASTN2 was found to be associated with fasting blood glucose (FBG) \((P = 1.0 \times 10^{-6}\)\), whereas NEO1 was associated with SGOT and serum creatinine \((P = 1.2 \times 10^{-5}\) and \(8.6 \times 10^{-5}\), respectively). However, we did not detect strong associations for any of the SNPs within the top-iSAFE scores in each gene. Table 3 also presents the results for the single top-iSAFE SNP in each gene for the traits that show weak associations. The detailed results of all tested phenotypes and SNPs are provided in supplementary table S3, Supplementary Material online.

**Discussion**

**Admixed Genetic Ancestries of the Taiwanese Han**

Population-specific genetic diversity accumulated along human migration trajectories could shape the genetic basis of diseases differently among populations (Chen et al. 2012; Corona et al. 2013; Wall et al. 2019). Although the genetic structure of the Han people in China has been investigated extensively in recent years (Wen et al. 2004; Xue et al. 2008; Chen et al. 2009; Xu et al. 2009; Zhao et al. 2015; Chiang et al. 2018), studies focusing on genetic ancestry of the Han populations outside China and the level of admixture with other ethnic groups, particularly on the island of Taiwan, are limited (Chen et al. 2016). In the present study, we first characterized the genetic ancestry of individual genomes and identified four major ancestries as well as subtle genetic structure within the Taiwanese Han. Our results are consistent with the findings of Chen et al. (2016), who utilized a smaller number of populations to identify four major ancestries and suggested that 80% of Taiwanese Han people are genetically closer to the Southern Han Chinese than to the Northern Han Chinese. However, the geographic patterns of these ancestries were not thoroughly discussed in their analysis. Although our inferred pattern of ancestries is also in good agreement with the previous studies that analyzed the Pan-Asia and HGDP data sets separately (Li et al. 2008; Abdulla et al. 2009), by analyzing the combined data, we were able to gain a better overview of the geographic distributions of these ancestries; consequently, they can be referred to as the Southeastern (blue), Northern (yellow), Island Southeast Asian (ISEA; pink), and Japonic (green) ancestries. Notably, we identified considerable proportions of ISEA ancestry (also carried by many Austronesian-speaking populations in high proportions) in most individuals of Taiwanese Han (average 15%, range 0.1–62%). The mixed ancestries observed in the Taiwanese Han could be attributed to either population mixture or shared ancestry before the divergence of descendent populations. We therefore applied the \(F_3\) tests to detect signatures of population mixture. Consequently, our results showed that the ISEA ancestry in the Taiwanese Han was the outcome of population mixtures rather than shared ancestry, and the admixture event likely occurred before the Taiwanese Han ancestors migrated to Taiwan (fig. 2A). If the admixture occurred only after the Han people migrated to Taiwan, then the observed results would only be seen in the Taiwanese Han. However, similar \(F_3\) outcomes were found in the Chinese Han (supplementary fig. S2, Supplementary Material online), supporting that admixture occurred prior to migration to Taiwan. Moreover, signatures of population admixture were also detected between the ancestors of Taiwanese Han and the Ami Austronesian-speaking population using the \(F_4\) test; significant positive \(F_4\) values were observed when most Sino-Tibetan speaking populations were individually included in the analysis, except for the Chinese Singapore and Chinese Cantonese (table 1). These two populations appear to be genetically closest to the Taiwanese Han among all other Sino-Tibetan speaking populations (fig. 18), which is consistent with the hypothesis of population mixture before the ancestors of Taiwanese Han migrated to Taiwan.

Using \(F_p\), Chiang et al. (2018) also identified significant signatures of population mixture between the Sichuan and Guangdong people (who live in Southwestern and Southeastern China, respectively) with the Ami and Atayal populations of Taiwan, which is in concordance with our hypothesis. Our results are also in a good agreement with the findings of McColl et al. (2018). By analyzing ancient human genomes, they revealed evidence of admixture and suggested that, during the demographic expansion from EA into Southeast Asia about 4,000 years ago, the EA framers did not simply replace the previous occupants. However, our results do not reject the possibility of recent admixture between the Taiwanese Han and indigenous populations on the island of Taiwan. Indeed, the wide range of individual variations in the proportion of Austronesian ancestry (0.1–62%) observed in the Taiwanese Han may be better attributed to recent admixture (McVean 2009). In the population tree inferred by fineSTRUCTURE, we observed 1 (from 500 individuals) of the Taiwanese Han grouped closer to the Dusun population, who are genetically closer to the indigenous populations of Taiwan than to the Sino-Tibetan populations (Mörseburg et al. 2016; Yew et al. 2018).

We also tested for signatures of population mixture between the Southeastern (blue) and Northern (yellow) ancestries in the Chinese Han by using the Hmong people to represent the blue ancestry and detected significant results in many Northern populations who carry high proportions of yellow ancestry (fig. 2A). A North-to-South cline of genetic structure has been well documented in the Chinese Han people (Wen et al. 2004; Zhang et al. 2007; Xue et al. 2008; Chen et al. 2009; Xu et al. 2009; Zhao et al. 2015; Chiang et al. 2018). Our results suggest that ancient population admixtures between the Northern and Southern populations in China may have played a role in shaping the North-to-South cline in the Han population.

Within the island of Taiwan, the genetic structure differs greatly between the Taiwanese Han and the two indigenous populations (Ami and Atayal), who carry very high proportions of ISEA ancestry (pink). These differences have been previously shown by Abdulla et al. (2009). The distinct patterns of genetic structure between the Taiwanese Han and indigenous populations imply that the genetic basis
underlying disease susceptibility could vary between them. The current WG genotyping bead-arrays used in the Taiwan Biobank are mainly customized for genotyping individuals of Han ancestry. It is therefore of great importance to incorporate genetic diversity of all Taiwanese indigenous populations when designing SNP arrays to uncover genetic variants that underlie disease susceptibility in the Taiwanese indigenous people.

Candidate Genes Targeted by Natural Selection
We searched for signatures of selection in the Taiwanese Han population by scanning for genetic loci that displayed unusually long haplotype lengths using iHS and identified 16 SNP clusters (including EDAR whose iHS score was used as the cutoff value; see fig. 3). Among them, only EDAR, the HLA gene family, and ASTN2 were previously reported as candidate genes favored by selection in human populations (Sabeti et al. 2007; Scheinfeldt et al. 2012). The rest of the SNP clusters were novel findings from our study. Since the sample size of our data is large ($n = 14,401$), the statistical power of iHS was substantially enhanced for detecting selection signatures. We next applied iSAFE statistics to verify these candidate loci. Consequently, we were able to link selection signals to five particular genes and conducted association analyses to further examine their possible effects on phenotypes.

Although metabolic-related traits and genotyping data of the Taiwanese Han curated by the Taiwan Biobank have been applied to conduct various association analyses, most have focused on identifying variants associated with certain disease cohorts while using the Taiwan Biobank samples as a control (Chung et al. 2017; Nfor et al. 2018; Lin et al. 2019). In the present study, association analyses were conducted only based on the Taiwan Biobank healthy individuals. However, we did not detect any significant association between the 16 metabolic-related traits and the selection-favored haplotypes (represented by the top-iSAFE SNPs). Kudaravalli et al. (2008) also found no significant association between expression quantitative trait loci (eQTLs) and the selection-favored allele that underlies recent positive selection at the lactase gene (Bersaglieri et al. 2004; Tishkoff et al. 2007). They observed significant association between eQTLs and $i$H$S$ scores in the Yoruba population but not for the non-African populations. Ambiguous evidence of linking selection signals with QTLs possibly reflects the complex nature of QTLs that are governed jointly by genetic and environmental factors (Mackay et al. 2009). Nonetheless, we attempted to explore the possible functional effects of each selection-candidate gene by examining significant associations (among the 16 metabolic-related traits) with the SNPs located within the selection-targeted region (defined by the peak of iSAFE signals, see fig. 5 and table 3). The significant association results in our analyses should not be treated as a direct consequence of selection because these identified SNPs are imputed and not in strong LD with the selected haplotypes. Rather, it can be treated as the possible effects of each candidate gene on the individuals’ phenotypes.

We identified five candidate genes targeted by natural selection. CTNNA2 encodes $\alpha$-N-catenin, a cytoskeleton protein that links cadherin adhesion receptor with actin cytoskeleton and plays an important role in the stability of dendritic spines. In the absence of $\alpha$-N-catenin, spine heads are abnormally motile (Abe et al. 2004). Several studies have shown that CTNNA2 variants are associated with schizophrenia and pachygyria (Mexal et al. 2008; Chu and Liu 2010; Schaeffer et al. 2018) as well as excitement seeking and impulsive behaviors (Terracciano et al. 2011; Ehlers et al. 2016). Although highly expressed in the brain, CTNNA2 is also expressed considerably in the testis (Fagerberg et al. 2014). However, little is known about its function in the testis. Since the actin cytoskeleton is important for the regulation of sperm motility, sperm capacitation, and acrosome reaction (Breitbart et al. 2005; Breitbart and Finkelstein 2018; Gervasi et al. 2018), it is possible that natural selection acted on CTNNA2 to increase male reproductive success and may be accompanied with negative side effects such as increased susceptibility to neurological disorders. In addition, some studies have suggested that serum albumin contributes to the production of seminal plasma albumin for maintaining sperm quality and morphology (Orlando et al. 1988; Elzanaty et al. 2007; Moura and Memili 2016). In the present study, the significant association identified between the CTNNA2 selection-targeted region and serum albumin level (table 3) supports the possible role of CTNNA2 in male reproduction.

We further retrieved data of correlation estimates between the top-ranked SNPs in iSAFE scores and tissue-specific gene expression levels from the database of the Genotype-Tissue Expression (GTEx) project (release V8). The top-iSAFE SNP ($r_{17018689}$) of CTNNA2 appears to have a significant effect on gene expression in the thyroid ($m = 1$ where $m$ is defined as the posterior probability that the effect exists in each study; Han and Eskin 2012). Its possible role in disease susceptibility requires future investigation (supplementary fig. S8A, Supplementary Material online; also see supplementary table S5, Supplementary Material online, for the remaining top-ranked iSAFE SNPs).

LRP1B encodes a member of the low-density lipoprotein receptor family, which is also a large family of cell-surface receptors. The function of LRP1B is related to LDL particle receptor activity and calcium ion binding (Liu et al. 2000). Diseases or complex traits associated with LRP1B variants include childhood obesity, Alzheimer’s disease, various types of cancer, and age of menarche (Speliotes et al. 2010; Chen et al. 2019; Kichaev et al. 2019; Lee 2019). Poledne et al. (2016) identified a positive correlation between non-high-density lipoprotein cholesterol concentrations and proportions of phagocytic macrophages in adipose tissue and hypothesized that the observed pattern is the consequence of evolutionary adaptation. They suggested that macrophage polarization in human visceral adipose tissue is related to fatty acid metabolism and that individuals with higher phagocytic activity of macrophages could provide selection advantages for survival against infectious diseases (Poledne et al. 2016, 2019; Poledne and Zicha 2018). From our results, we found that the selection-targeted region of LRP1B was indeed associated with serum LDL levels (table 3). Our findings are supportive of the hypothesis proposed by Poledne et al.; moreover,
LRP1B was also found to be associated with the age of menarche (Kichaev et al. 2019). Gluckman and Hanson (2006a, 2006b) suggested that, in a stressful environment, female individuals who mature early have higher reproductive success than individuals who mature later, whereas the advantages are reversed in a stable environment, whereby late maturation could result in better health, longer reproductive life, and potentially more offspring. Therefore, selection signatures identified in LRP1B may also be the result of selection for reproductive success in females.

CSNK1G3 encodes a serine/threonine kinase that phosphorylates caseins and other acidic proteins. This gene was found to be associated with bone density, leukocyte count, LDL cholesterol, and diastolic blood pressure (Giri et al. 2019; Kichaev et al. 2019; Morris et al. 2019). Signatures of artificial selection on CSNK1G3 have been identified in Jersey cattle (Kim et al. 2015). A subsequent study identified a significant association between CSNK1G3 variants and the content of major proteins in bovine milk including four casein proteins (Buitenhuis et al. 2016). Therefore, it is possible that natural selection also acted on CSNK1G3 in humans for altering the protein content in breast milk. In our study, we detected significant associations with serum triglyceride concentration and hemoglobin A1c percentage at the selection-targeted region of CSNK1G3, implying the possible effects of selection on glucose and lipid metabolism, which also affects bone metabolism (Cipriani et al. 2020). The eQTL results from the GTEx database showed significant associations between the top-iSAFE SNP (rs6868518) of CSNK1G3 and gene expression in the mammary tissue of breast (m = 0.95) and in many other tissues such as the ovary (m = 0.93), several cerebral tissues (m > 0.9), liver (m = 1.0), and pancreas (m = 0.99; see supplementary fig. S8B, Supplementary Material online).

ASTN2 is known to be highly expressed in the adult brain and involved in glial-guided neuronal migration at developmental stages (Wilson et al. 2010). ASTN2 variants were found to be associated with various neurological disorders including Alzheimer’s disease, autism-spectrum disorders (ASD), schizophrenia, bipolar, intellectual disability, and attention-deficit/hyperactivity disorder (Lesch et al. 2008; Vrijenhoek et al. 2008; Glessner et al. 2009; Lionel et al. 2014; Wang et al. 2016). Signatures of adaptive evolution at ASTN2 have been previously identified in South Asians, the Khomani San hunter-gatherers of southern Africa, and three Ethiopian populations (Scheinfeld et al. 2012; Tekola-Ayele et al. 2015; Racimo et al. 2017). However, its biological role underlying adaptive evolution remains unclear. Although the function of ASTN2 has been well characterized in the brain, this gene is also highly expressed in the prostate and testis (Fagerberg et al. 2014). From the GTEx eQTL data, the top-iSAFE SNPs of ASTN2 indeed showed a significant effect on gene expression in the testis (P < 5.4 × 10^{-8}). In our study, we detected a significant association between the selection-targeted region of ASTN2 and the level of FBG, supporting its possible role in glucose metabolism.

Finally, NEO1 (a member of the immunoglobulin gene superfamily) encodes neogenin, a multi-functional membrane receptor that regulates cell adhesion in diverse developmental processes including cortical interneuron migration and axon guidance (Matsunaga and Chédotal 2004; Matsunaga et al. 2006; Hagiura et al. 2011). Recessive functional polymorphisms in NEO1 were found to be associated with cardiac disease and ASD (Mcllnnes et al. 2010; Siu et al. 2016; Nolte et al. 2017; van Esch et al. 2018). Moreover, Polimanti and Gelemtier (2017) noted that many ASD common risk alleles were enriched for genomic signatures of positive selection due to enhanced cognitive ability. Therefore, selection may have acted on NEO1 for the same reason. In addition, several studies have reported abnormal regulations in SGOT and creatine among ASD children (Giulivi et al. 2010; Schulze et al. 2016). In our study, we indeed identified significant associations between the selection-targeted region of NEO1 with the levels of SGOT and creatine, supporting its possible role in ASD susceptibility. The eQTL data of GTEx further revealed correlations between the top-iSAFE SNP (rs8039418) and gene expression in sun-exposed skin (m = 0.94) and muscularis esophagus (m = 1.0; see supplementary fig. S8C, Supplementary Material online).

In summary, all five candidate genes identified in our study appear to have pleiotropic effects and connections to various disease susceptibilities. Each selection-targeted region also showed significant associations with at least one metabolism-related trait, suggesting that evolutionary adaptation could have a profound impact on human health. One evident example is the HLA-B*5801 allele, which carries the top candidate SNP (rs9262558), identified in this study (|iHS| = 7.51; see table 2). This allele has been reported to be significantly associated with increased risk for nasopharyngeal carcinoma in the Taiwanese Han (Hildesheim et al. 2002). In future studies, it would be intriguing to design and conduct experiments to identify each causal variant targeted by selection and the underlying molecular mechanism based on the list of top candidate variants provided in our study (supplementary table S1, Supplementary Material online).

Materials and Methods

Taiwan Biobank: WG Genotyping and Sequencing Data for the Taiwanese Han People

The Taiwan Biobank (https://www.twbiobank.org.tw/new_web_en/index.php) is a nationwide research database that collects genomic/epigenomic data together with various phenotypic/clinical profiles for each participant (Chen et al. 2016). We obtained the WG SNP genotyping data of 15,990 Taiwanese Han people from this database. All individuals were self-reported as Han people based on their parents’ ancestries. Each individual was genotyped using the customized Affymetrix Axiom genotyping array plate (TWB chip) with a total of 653,291 SNPs. In addition, we retrieved the WG sequencing data from the Taiwan Biobank for a subset of individuals (n = 791) whose genomes were both genotyped and sequenced. The WG sequencing data was generated based on the Illumina HiSeq platform with an average coverage of 30×. Approval of this research project was received from institutional review boards of both the Ethics and
Merging with Three Public Data Sets

In order to maximize the number of Asian ethnic groups that could be analyzed, three additional public data sets were retrieved for merging with the WG SNP genotyping data from the Taiwan Biobank: 1) HGDP containing 1,043 individuals from 51 worldwide populations and yielding ~650k SNPs (Li et al. 2008); 2) HUGO Pan-Asia Consortium containing 1,928 individuals of 71 Asian populations from China, India, Indonesia, Japan, Malaysia, the Philippines, Singapore, South Korea, Taiwan, and Thailand and yielding a considerably small number of SNPs (~55k SNPs) (Abdulla et al. 2009); and 3) Southeast Asia data set containing 700k SNPs from a total of 130 individuals from Burmese, Vietnamese, and six Austronesian populations (a partial data set from Mörseburg et al. [2016]).

Quality Control for the WG Genotyping and Sequencing Data

We performed several quality control (QC) steps to detect problematic individuals and SNPs for each of the data sets according to Anderson et al. (2010). First, we inspected the homozygosity rate of the X chromosome in each individual to check for discordance with the ascertained sex. Second, as low DNA quality or concentration affects call rate and genotype accuracy, we assessed missing call rate per individual by counting the number of missing SNPs for each individual. We also checked genome-wide heterozygosity rate per individual because excessive or reduced heterozygosity rate may be due to sample contamination or inbreeding, respectively. Individuals who had a missing call rate over 0.03 or heterozygosity rate deviating ±3 standard deviations from the population mean were removed. Lastly, we detected and removed closely related individuals by estimating pairwise identical-by-descent for all pairs of individuals (identical-by-descent > 0.1875, within third-degree relatives). Particularly, for the WG genotyping data, we also calculated the discordance rates of the polymorphic sites that were both sequenced and genotyped and subsequently filtered out individuals with discordance rate larger than 0.1% (Rasmussen-Torvik et al. 2017; Adelson et al. 2019). For per-SNP QC, we discarded SNPs with genotyping missing rates higher than 0.03. To identify SNPs caused by genotyping error, we also tested Hardy–Weinberg equilibrium (HWE) and excluded SNPs with significant departures from HWE at \( P \leq 10^{-50} \). A small number of SNPs with very strong departure from HWE are more likely caused by genotyping errors rather than any evolutionary forces (e.g., selection, see Anderson et al. 2010; Laurie et al. 2010). All the QC steps were conducted using PLINK v1.9 (Purcell et al. 2007; Chang, Chow, et al. 2015).

Inferring Genetic Ancestries

We employed a model-based approach to infer ancestries of individual genomes using ADMIXTURE (version 1.23). ADMIXTURE takes a maximum likelihood approach to estimate allele frequencies of SNPs in \( K \) ancestral populations and ancestry proportions for all individual genomes (\( Q \)) (Alexander et al. 2009). A CVE was applied to determine the optimum number of hypothetical ancestral populations (\( K \)) that gives the lowest prediction error among all \( K \) values. ADMIXTURE was first performed for a combined data set by merging the Pan-Asia and HGDP data sets for a total of 19,290 intersected SNPs in 2,304 people from 99 Asian populations after exclusion of non-Asian populations. We also performed several additional runs of ADMIXTURE for inferring ancestry proportions of all 14,401 individual genomes from the Taiwan Biobank by excluding the Pan-Asia data set from the analysis but including the Southeast Asia data set. This was done in order to increase the number of SNPs to be analyzed while retaining some Maritime Southeast Asia populations for inferring Austronesian ancestry. As a result, the combined data set contains 1,120 individuals from 56 populations for a total of 101,955 SNPs (in addition to the 14,401 samples from the Taiwan Biobank). Before running ADMIXTURE, we pruned out SNPs with strong LD (\( r^2 > 0.8 \)) using sliding window analysis (window size 50 SNPs and step size of ten SNPs).

Detecting Signatures of Admixture Using the \( F_3 \) and \( F_4 \) Population Tests

We performed multiple runs of the \( F_3 \) and \( F_4 \) statistics to test whether the Taiwanese Han population was admixed with the Austronesian ancestry. For a given locus, the \( F_3 \) statistic, \( F_3(X; Y, W) \), is defined as the product of allele-frequency differences between population \( X \) and \( Y \) and between population \( X \) and \( W \) and is scaled by binomial variance in allele frequency of \( X \) where \( X \) is the recipient population (Taiwanese Han) and \( Y \) (Ami or Atayal) and \( W \) (pop) are the two donor populations. Ami and Atayal represent the Austronesian speaking populations and pop, was selected from the remaining Asian populations of our combined data set with Pan-Asia and HGDP. In the case of no population mixture, the expected value of \( F_3 \) is positive, whereas a significant negative value for \( F_3(X; Y, W) \) indicates that the ancestors of population \( X \) experienced a history of population mixture with the populations close to \( Y \) and \( W \) (Reich et al. 2009). We also repeated the \( F_3 \) tests by assuming \( F_3(\text{Chinese Han}; \text{Ami}, \text{pop}) \) and \( F_3(\text{Chinese Han}; \text{Chinese Hmong}, \text{pop}) \) where Chinese Hmong (CN-HM) represents the Southeastern ancestry. The \( F_3 \) statistic is also defined in terms of correlations of allele-frequency (\( p \)) differences, but involving four populations (\( A, B, C, \text{and} D \)) as \( F_3(A, B, C, D) = E[p_{ABC} - p_{BD}] \) where the four populations are related by the unrooted population tree (\( (A, B), (C, D) \)) with the expected value = 0. By assigning a divergent outgroup population as \( A \) (i.e., no admixture into \( C \) or \( D \)), a significant negative \( F_3 \) value implies gene flow between \( B \) and \( D \), whereas gene flow between \( B \) and \( C \) could result in a positive \( F_3 \) value (Reich et al. 2009; Patterson et al. 2012). The \( F_3 \) test was conducted by assuming \( F_3(\text{Yoruba, Ami}; \text{pop}; \text{Taiwanese Han}) \) where pop, was taken from the other Sino-Tibetan speaking populations. Each test statistic was averaged over all SNPs and the variance was measured using a Block Jackknife. The \( F_3 \) and \( F_4 \) tests were performed by running “qpop3pop” and “qpopstat,” respectively.
implemented in ADMIXTOOLS version 5.1 (Patterson et al. 2012).

Detecting Fine-Scale Genetic Structure within the Taiwanese Han Population

We further applied fineSTRTURE (version 4.1.0) to investigate subtle genetic structure within the Taiwanese Han and related populations by analyzing the combined data set that includes 500 individuals from the Taiwan Biobank (WG genotyping data), 17 Eastern Asian populations from the HGDP, and eight Southeast populations from the data set of Mörgenow et al. (2016). Since linked SNPs on a given genomic region shared the same genealogy, patterns of LD are expected to reflect a shared history between closely related populations. FineSTRTURE exploits LD information between close markers and is proven to be useful for identifying subtle population structure (Lawson et al. 2012). The program first constructed a pairwise coancestry matrix between all sampled individuals. The matrix stores frequencies of DNA segments that are shared between individuals. This “chromosome painting” step took a Hidden Markov Model to identify the recombination breakpoints and the individuals (donors) for which each chunk has the most recent common ancestor. The constructed coancestry matrix was further used to infer population structure and the assignment of individuals (ancestors) to each population based on a likelihood approach via the Markov chain Monte Carlo algorithm for searching optima to each population based on a likelihood approach via (donors) for which each chunk has the most recent common ancestor. The constructed coancestry matrix was further used to infer population structure and the assignment of individuals to each population based on a likelihood approach via the Markov chain Monte Carlo algorithm for searching optimal parameter values. The required parameter settings for expectation maximization (EM) process were based on the default settings including: number of EM iterations = 10, minimum number of SNPs for EM estimation = 10,000, and fraction of genome = fraction of individuals (to use for EM estimation) = 0.1, while setting the starting value for \( N_e \) (effective population size) as 5 (e.g., "slarg5:in: -iM -emfile: -n 5").

Detecting Genomic Signatures of Positive Selection by iHS

To detect genomic signatures of recent positive selection in the Taiwanese Han population, we used the iHS, a LD-based method that can capture candidate loci of unusually long blocks of shared haplotypes across the genome (Voight et al. 2006). iHS summarizes the differences in extended haplotype homozygosity between ancestral and derived alleles. Under adaptive evolution, a selection-favored allele would increase its frequency in a population within a shorter time than a neutral allele of the same frequency. Consequently, selected alleles and their neighboring SNPs would remain linked and leave a longer block of haplotype homozygosity than neutral alleles, since recombination events do not have enough time to break down LD (Sabeti et al. 2002; Voight et al. 2006). iHS was calculated using the rehh v2.0 package (Gautier and Vitalis 2012) implemented in R. The ancestral and derived alleles for each SNP were determined parsimoniously by comparing with their orthologous sites in chimpanzee, gorilla, and orangutan, if the genetic information was available in the UCSC Genome Browser (https://genome.ucsc.edu/index.html). The inference algorithm was written in PERL according to Fitch (1971). Some SNPs were excluded from the analysis if their ancestral states could not be determined or if their allele frequencies are \( \leq 0.01 \) (Voight et al. 2006). Haplotype information was inferred by Eagle (v2.3.2) (Loh et al. 2016) for the WG SNP genotyping data of all 14,401 individuals from the Taiwan Biobank. The genetic map used for calculating iHS was downloaded from the International HapMap Project—phase 3 (https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html). A genetic region was considered as a candidate SNP region targeted by selection if at least three extreme markers were identified with their iHS values higher than the threshold and \( \geq 10 \) SNPs of top 1% iHS values within a 500-kb region. The threshold was set at the highest iHS score of the EDAR gene, a well-studied gene targeted by recent positive selection in the Han population (Sabeti et al. 2007; Grossman et al. 2010; Kamberov et al. 2013).

Characterizing Gene Genealogies for Identifying Selection-Favored Genes/Variants

To further confirm the candidate region detected by iHS, we also employed iSAFE (integrated selection of allele favored by evolution) statistic to characterize the shape of genealogy for each candidate region and pinpoint the possible gene/variant targeted by selection (Akbari et al. 2018). iSAFE is a sliding-window-based SAFE statistic, which is defined as:

\[
\text{SAFE}(e) = \frac{\phi - \kappa}{\sqrt{f(1-f)}}
\]

for a given mutation \( e \), where \( \phi \) is the fraction of derived-allele counts for all the carriers of mutation \( e \) relative to all derived-allele counts (including carriers and noncarriers), \( \kappa \) is the fraction of distinct haplotypes among the carriers relative to the total number of distinct haplotypes, and \( f(e) \) is the allele frequency of mutation \( e \). Carriers of the selection-favored mutation should have a high \( \phi \) and low \( \kappa \) (fewer distinct haplotypes) compared with noncarriers, resulting in a higher SAFE score compared with the SAFE score for a neutral mutation. For a given candidate region targeted by selection, the entire region was further divided into multiple windows and the SAFE score was computed for each SNP in a given window. Next, the SAFE scores of all variants over all windows were weighted and combined to assign an iSAFE score to each variant in the large region. The variants that ranked highest in the iSAFE scores were then considered as top candidates of the causal mutation favored by selection (Akbari et al. 2018). The iSAFE analysis was conducted for each candidate region identified by iHS where the sequences were extracted from the WG sequencing data for a total of 772 individuals from the Taiwan Biobank. The ancestral and derived states for each polymorphic site were determined parsimoniously as described above. The pairwise estimates of LD based on the squared correlation coefficient (\( r^2 \)) were conducted using PLINK and the maps of LD were plotted using the "LDheatmap" package (Shin et al. 2006) implemented in R.
Association Analyses for Detecting Functional Effects of the Identified Loci Targeted by Selection

We performed statistical association analyses to detect any functional effects for a given identified candidate loci targeted by selection across 16 metabolic-related traits from the Taiwan Biobank (https://www.twbiobank.org.tw/new_web_en/). These traits can be broadly categorized into three classes: 1) kidney/diabetic: blood urine nitrogen (BUN), creatinine (CREA), serum uric acid (UA), fasting glucose (FG), and HbA1c; 2) cardiovascular: high-density lipoprotein cholesterol, TG, total cholesterol (TC), TG, diastolic blood pressure (DBP), and systolic blood pressure (SBP, cardiac); and 3) liver function: serum albumin level (ALB), total bilirubin (tBIL), gamma glutamyl transpeptidase (γ-GT), serum level of glutamic-pyruvic transaminase (SGPT), and SGOT. Since these measurements are continuous variables, we assumed a multiple linear regression model where the independent variable is the genotype, coded as 0, 1, or 2, corresponding to the number of copies of minor allele carried by an individual, and the dependent variable is each of the 16 traits. The covariates included age, sex, body mass index, as well as the results of PC1–8 in the principal component analysis across all individuals to account for any hidden genetic structure. All results of PC1–8 in the principal component analysis across the panel (phase 3) (Howie et al. 2009; Delaneau et al. 2013; O’Connell et al. 2014). We also applied a LD-based clumping procedure to report significant SNPs for a given candidate region targeted by selection. The clumping procedure first takes SNPs that are significant at \( P \leq 10^{-8} \) as index SNPs. The threshold was set to correct for the number of identified candidate loci multiplied by the number of traits. A clump was formed by including all other “clumped” SNPs that passed the second significance threshold \( (P \leq 0.01) \) within a 250-kb distance from the index SNP and are in LD with the index SNP \((r^2 \geq 0.5, 10\times [-\log 0.01] = 20).\)

Data Availability

Raw data were generated at Taiwan Biobank (https://www.twbiobank.org.tw/new_web_en/about-export.php). Derived data supporting the findings of this study are available from the corresponding author on request.

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments

We are grateful to Dr Chia-Wei Chen and Dr Hsin-Chou Yang for their assistance with statistical analysis and deeply thankful to Dr Danial Lawson for his great help with running fineSTRUCTURE. We also thank Dr Hsiao-Hui Lee and two anonymous reviewers for their insightful comments. We greatly appreciate the technical services and support provided by Dr Tze-Tze Liu from the Genomics Center for Clinical and Biotechnological Applications of Cancer Progression Research Center, National Yang-Ming University. We also acknowledge the support from the National Core Facility for Biopharmaceuticals (NCFB, MOST 108-2319-B-492-001) and National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources. This research was funded by the Taiwan Ministry of Science and Technology (MOST) (Grant Nos. 104-2311-B-010-001-MY2, 105-2311-B-010-004-MY3, and 109-2311-B-010-006 to W-Y.K.).

References

Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal CC, Chaurasia A, Chen C-H, Chen J, Chen Y-T, et al. 2009. Mapping human genetic diversity in Asia. Science. 326(5959):1541–1545.

Abe K, Chisaka O, van Roy F, Takeichi M. 2004. Stability of dendritic spines and synaptic contacts is controlled by z-N-catenin. Nat Neurosci. 7(4):357–363.

Adelson RP, Renton AE, Li W, Barzilai N, Atzmon G, Goate AM, Davies P, Freudenberg-Hua Y. 2019. Empirical design of a variant quality control pipeline for whole genome sequencing data using replicate discordance. Sci Rep. 9(1):16156.

Akbari A, Vitti JJ, Iranmehr A, Bakhtiar M, Sabeti PC, Mirarab S, Bafna V. 2018. Identifying the favored mutation in a positive selective sweep. Nat Methods. 15(4):279–282.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Open Bioinf J. 19(9):1655–1664.

Anderson CA, Petterson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. 2010. Data quality control in genetic case–control association studies. Nat Protoc. 5(9):1564–1573.

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.
Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative frequency of human genetic variants. Nat Genet. 46(3):310–315.

Ko AM-S, Chen C-Y, Fu Q, Delfin F, Li M, Chiu H-L, Stoneking M, Ko Y-C. 2014. Early Austronesians: into and out of Taiwan. Am J Hum Genet. 94(3):426–436.

Ko W-Y, Gomez F, Tishkoff SA. 2012. Evolution of human erythrocyte-specific genes involved in malaria susceptibility. In: Singh RS, Xu J, Kulathinal RJ, editors. Rapidly evolving genes and genetic systems. Oxford: Oxford University Press. p. 223–234.

Ko W-Y, Rajan P, Gomez F, Scheinfeldt L, An P, Winkler CA, Froment A, Nyambo TB, Omar SA, Wambbebe C, et al. 2013. Identifying Darwinian selection acting on different human APOL1 variants among diverse African populations. Am J Hum Genet. 93(1):54–66.

Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK. 2008. Gene expression levels are a target of recent natural selection in the human genome. Mol Biol Evol. 26(3):649–658.

Kwitowski DP. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet. 77(2):171–192.

Lachance J, Tishkoff SA. 2013. Population genomics of human adaptation. Annu Rev Ecol Evol Syst. 44(1):123–143.

Laure C, Doherty KF, Wiens DB, Rugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornells MC, Edenberg HJ, et al. 2010. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol. 34(6):591–602.

Lawson DJ, Helliwell G, Myers S, Falush D. 2012. Inference of population structure and quality assurance in genotypic data for genome-wide association studies. Mol Biol Evol. 26(3):649–658.

Mackay TFC, Stone EA, Ayroles JF. 2009. The genetics of quantitative association and extended pedigree linkage studies. Nat Rev Genet. 10(8):565–577.

Matsunaga E, Nakamura H, Chédratol A. 2006. Repulsive guidance molecule plays multiple roles in neuronal differentiation and axon guidance. J Neurosci. 26(22):6082–6088.

McColl H, Racimo F, Vinner L, Demeter F, Gakuhari T, Moreno-Mayar JV, van Driem G, Gram Wilken U, Seguin-Orlando A, Castro Cdlf, et al. 2018. The prehistoric peopling of Southeast Asia. Science 361(6397):88–91.

Mclnsa LA, Nakamine A, Pihorge M, Brandt T, Jimenez Gonzalez P, Fallas M, Manghi ER, Edelmann L, Glessner J, Hakonarson H, et al. 2010. A large-scale survey of the novel 15q24 microdeletion syndrome in autism spectrum disorders identifies an atypical deletion that narrows the critical region. Mol Autism. 1(1):5.

McVean G. 2009. A genealogical interpretation of principal components analysis. PLoS Genet. 5(10):e1000686.

Mchok S, Berger R, Pearce L, Barton A, Logel J, Adams CE, Ross RG, Freedman R, Leonard S. 2008. Regulation of a novel α2δ-catenin splice variant in schizophrenic smokers. Am J Med Genet. 147B(6):759–768.

Mora JS, Kemp JP, Youten SE, Laurent L, Logan JG, Chai RC, Vulpescu NA, Forretta V, Kleinman A, Mohanty ST, et al. 2018. An atlas of genetic influences on osteoporosis in humans and mice. Nat Genet. 51(2):258–266.

Mörseburg A, Paganj L, Ricatou F-X, Yngvadottir B, Harney E, Castillo C, Hoogervorst T, Antão T, Kusuma P, Brucato N, et al. 2016. Multilayered population structure in Island Southeast Asian. Eur J Hum Genet. 24(11):1605–1611.

Moura AA, Memel E. 2016. Functional aspects of seminal plasma and sperm proteins and their potential as molecular markers of fertility. Ani Reprod. 13(3):191–199.

Nfor ON, Wu M-F, Lee C-T, Wang L, Liu W-H, Tantho DM, Hsu S-Y, Lee K-J, Ho C-C, Debnath T, et al. 2018. Body mass index modulates the association between CDKAL1 rs10946398 variant and type 2 diabetes among Taiwanese women. Sci Rep. 8:13235.

Nolte IM, Munoz ML, Tragante V, Amare AT, Jansen R, Vaez A, von der Heyde B, Avery CL, Bis JC, Dierckx B, et al. 2017. Genetic loci associated with heart rate variability and their effects on cardiac disease risk. Nat Commun. 8(1):15805.

O’Connell J, Gurdasani D, Delanauve O, Pirastu N, Ulivi S, Cocco M, Traglia M, Huang J, Huffman JE, Rudan I, et al. 2014. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 10(4):e1004234.

Orlando C, Casano R, Forti G, Barni T, Vannelli GB, Balboni GC, Serio M. 2018. The prehistoric peopling of Southeast Asia. Science 361(6397):88–91.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, N fouORN, WuM-F, LeeC-T, WangL, LiuW-H, TantohDM, HuSS-Y, Lee K, Ho C-C, Debnath T, et al. 2018. Body mass index modulates the association between CDKAL1 rs10946398 variant and type 2 diabetes among Taiwanese women. Sci Rep. 8:13235.

Peterson RA, Cauanagh JE. 2020. Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. J Appl Stat. 47:2312–2327.

Poland R, Kralova Lesna I, Kralova A, Fronek J, Cejkova S. 2016. The relationship between non-HDL cholesterol and macrophage phenotypes in human adipose tissue. J Lipid Res. 57(10):1899–1905.

Poland R, Malinska H, Kubatova H, Fronek J, Thieme F, Kauerova S, Petridou R, Malinska H, Kubatova H, Fronek J, Thieme F, Kauerova S, Petridou R. 2018. Polymorphisms in human adipose tissue is related to the fatty acid spectrum in membrane phospholipids. Nutrients 12(1):8–13.

Poland R, Zicha J. 2018. Human genome evolution and development of cardiovascular risk factors through natural selection. Physiol Res 67(2):155–163.

Polimanti R, Gelernter J. 2017. Widespread signatures of positive selection in common risk alleles associated to autism spectrum disorder. PLoS Genet. 13(2):e1006618.

Porcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PJW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81(3):559–575.
