MONASH University

Identifying positively selected genetic variants in the Chinese, Malay and Indian populations in Southeast Asia

Fadilla Ramadhani Wahyudi
Master of Science (Research)

A thesis submitted for the degree of Master of Science (Research) at Monash University in 2020

Malaysia School of Science

COPYRIGHT NOTICE

ABSTRACT

The emergence of whole genome sequencing datasets has contributed to the advancement of unbiased detection of all kinds of genomic variations in human populations. Detecting these genetic variants aids in the understanding of human evolutionary adaptation and can provide insight into how our genomes have changed during interactions with pathogens, climate, and their diet. This research leverages on three publicly available whole genome sequencing projects of Chinese, Malay and Indian population groups in China and Singapore, which are underrepresented in positive selection studies. The aim was to use a statistical method called Fine-Mapping of Adaptation Variation (*FineMAV*) to prioritise candidate population-specific positively selected variants for functional validation. It does this by incorporating three metrics: population differentiation, derived allele frequency and functional annotation. This generates high *FineMAV* scores for variants that are high in frequency, population-specific and predicted to be deleterious. I was able to replicate well-known selection signals that were previously identified in East Asians, such as the missense variant rs3827760 in ectodysplasin A receptor (*EDAR*), and found novel variants like the missense rs79597880 in pre-rRNA-processing protein TSR1 homolog (*TSR1*) in Singaporean Malays. Mutations in *TSR1* have been linked with a rare heart condition called spontaneous coronary artery dissection. To make *FineMAV* more accessible for researchers, I developed a software program so that they can generate *FineMAV* scores for sequencing datasets of their interest and graphically visualise their genome-wide *FineMAV* scores on a human genome browser, like the web-based University of California, Santa Cruz (UCSC) Genome Browser or Ensembl Genome Browser. I also evaluated the performance of the software on a much larger whole genome sequencing dataset called the GenomeAsia 100K, comprising 1,428 individuals from Northeast Asian, South Asian, Southeast Asian and Oceanian populations, and ensured that it was built to be memory-efficient in anticipation for larger human genomic datasets.

DECLARATION

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Name: Fadilla Ramadhani Wahyudi

Date: 1st December 2020

LIST OF PUBLICATIONS AND PRESENTATIONS

Publications

***FineMAV*: A software for prioritising positively selected variants in whole-genome sequencing data**

In the process of submission

Fadilla Ramadhani Wahyudi, Jasbir Dhaliwal, Farhang Aghakhanian, Sadequr Rahman, Teo Yik Ying, Michał Szpak, Qasim Ayub

***FineMAV* identifies outliers associated with adaptation in GenomeAsia 100K dataset.**

Manuscript in preparation

Fadilla Ramadhani Wahyudi, Jasbir Dhaliwal, Qasim Ayub

Presentations at conferences

Poster presentation at the Human Evolution 2019

Held on 30th October to 1st November 2019 at the Wellcome Genome Campus Conference Centre, Hinxton, Cambridge, United Kingdom.

Oral presentation at the Taylor's University Graduate Research Symposium 2019

Held on 30th September 2019 at the Taylor's University Lakeside Campus, No 1. Jalan Taylor's, 47500 Subang Jaya Selangor Darul Ehsan, Malaysia.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my supervisor, Associate Professor Qasim Ayub, and co-supervisors, Dr. Jasbir Dhaliwal and Professor Sadequr Rahman, for giving me this exciting opportunity. I feel incredibly fortunate to have them as my mentors. I would like to acknowledge Jasbir for her help in the high-level conceptual ideas for building the software and to thank my collaborator Dr. Michał Szpak for his guidance on *FineMAV*. I would also like to thank the past and present staff at the Monash University Malaysia Genomics Facility, especially Zarul Hanifah and Dr. Farhang Aghakhanian, who ushered me into the bioinformatics world. Thank you to my review panel members Professor Sunil K Lal, Dr. Gavin Wee Wei Yee and Professor Maude Elvira Phipps for all their constructive feedback during my milestone presentations. Thank you to Monash University Malaysia for the financial support and for the opportunity to attend a conference overseas. I am grateful for the friends that I have made here, especially Nishat, Isra, Wana, Heerman and Rick, for the emotional and intellectual support. Thank you to my sister, Thara, who has been a great source of happiness for me. Lastly, I am forever grateful to my parents for their unwavering support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

1 INTRODUCTION

1.1 Positive selection

The concept of natural selection was originally conceived by Darwin and Wallace (1858). The essence of their publication details how there are heritable variations within a population of organisms and that those who are better adapted to their environment are more likely to survive and reproduce, thus passing their advantageous traits to future generations (Darwin and Wallace, 1858). The mechanisms by which natural selection occurred were an enigma at that time. It was only after the assimilation of Mendelian inheritance with natural selection that evolution could be viewed from a molecular perspective (Fisher, 1930), which laid the foundation for population genetics (Provine, 1971).

Further understanding of genetics led to a more precise understanding of natural selection. Theories about different modes of selection were formulated in which the basis of these modes are allele frequency changes that occur within a population over time (Nielsen, 2005). One type of selection acts in a directional manner (Nielsen, 2005; Vitti et al., 2013). As Nielsen (2005) reviews in his paper, new alleles are introduced into the population via mutations. These are known as derived alleles. They can be advantageous or deleterious, which in turn could affect the fitness of an organism or the ability of an organism to survive and reproduce. Derived alleles that are advantageous and confer higher fitness would increase in allele frequency. This is known as classical Darwinian or positive selection. On the other hand, negative selection, also known as purifying selection, occurs when derived alleles are deleterious and are selected against. However, the neutral theory of evolution states that the bulk of evolutionary changes occur because of random fluctuations in allele frequencies, termed as genetic drift of neutral alleles. These are alleles that do not affect reproductive fitness (Kimura, 1991; Nielsen, 2005). It should be noted that advantageous alleles can also increase in frequency due to genetic drift (Nielsen, 2005).

The study of positive selection has garnered the interests of researchers worldwide because identifying genetic variants that are positively selected can provide insights into new

molecular functions that come with adaptation (Pritchard et al., 2010). During the Out of Africa migration, where modern humans expanded from Africa to Eurasia and then migrated to the rest of the globe, they were subjected to diverse environments and diets, especially after the dawn of agriculture (Pritchard et al., 2010; Jeong and Di Rienzo, 2014; Jobling et al., 2014). These new selection pressures allowed for populations to amass locally adaptive features through positive selection (Pritchard et al., 2010; Jeong and Di Rienzo, 2014). Detecting genetic variants that have undergone positive selection aids in the understanding of human evolutionary adaptation. Such detection has provided genetic insight into how humans interact with pathogens (Hamblin and Di Rienzo, 2000; Sakagami et al., 2004), the climate (Lamason et al., 2005; Soejima and Koda, 2007; Hancock et al., 2008) and their diet (Tishkoff et al., 2007) and has also shed light on population disease susceptibilities (Ferrer-Admetlla et al., 2009).

1.1.1 Signatures of positive selection

When a positive selection pressure acts on an allele at a particular genetic locus, it leaves patterns within the genome, also known as “signatures”. The basis of these signatures is genetic hitchhiking which occurs when the selection of one allele increases the frequency of other neutral alleles that are in proximity to it on the same genomic segment in a population, a phenomenon termed as genetic linkage (Smith and Haigh, 1974). Over generations, a selective sweep can occur in which the frequency of the positively selected allele and the ‘hitchhiked’ alleles rises within the population causing variation at linked sites to be swept out (Smith and Haigh, 1974; Sabeti et al., 2006). Recombination, however, can eliminate these nearby alleles thus decreasing the size of the ‘hitchhiked’ region over time (Sabeti et al., 2006). Some positively selected alleles can reach fixation through hard selective sweep (Smith and Haigh, 1974; Pritchard et al., 2010). However, evidence has shown that this has been rare for the last ~250,000 years of human evolution (Hernandez et al., 2011). As described by Sabeti et al. (2006), the signatures that positive selection leaves behind are:

1. Long haplotypes, or group of alleles that are inherited together from a single parent, with low genetic diversity.

2. High frequency of derived alleles.

3. Population differentiation or differences in allele frequencies between spatially separated populations.

4. High frequency of rare alleles.

1.1.2 Statistical methods to detect positive selection

In population genomics, polymorphisms aid in detecting evolutionary selective events (Nielsen, 2005). There are many types of genetic variations, including single nucleotide polymorphisms (SNPs), insertions and deletions (indels), microsatellites and structural variations. Most positive selection studies in humans utilise SNPs as they are the largest source of genetic variation and are easily detectable (Nguyen et al., 2006; The 1000 Genomes Project Consortium, 2010). Additionally, it is difficult to identify the ancestral state for indels and structural variations, which is imperative in establishing the direction of change (Donald and Matthew, 2007; Kvikstad and Duret, 2014).

Table 1 is a general summary of the types of statistical methods that employ within-species polymorphisms to detect positive selection. Akey (2009) categorizes them into three main groups based on the signatures they detect:

1. **Site frequency spectrum**

   These tests examine the distribution of the allele frequencies in a given genomic region and are therefore able to detect high frequency of rare alleles.

2. **Linkage disequilibrium (LD)**

   LD is defined as the non-random association of alleles between different loci. These tests scan for high frequencies of long haplotypes that are left behind during an ongoing or incomplete selective sweep.

3. **Population differentiation**

   These tests leverage on the differences in allelic frequencies between populations.

In recent years, composite methods that combine signatures from these main groups have also been developed (Grossman et al., 2010) (Table 1).

Table 1: Summary of statistical methods used to detect positive selections as categorized by (Akey, 2009).

Classification	Statistical methods	References
Site frequency spectrum	Composite likelihood approaches	(Kim and Stephan, 2002; Zhu and Bustamante, 2005)
	F_S	(Fu, 1997)
	Fu and Li's D*	(Fu and Li, 1993)
	Maximum frequency of derived mutations (MFDM)	(Li, 2011)
	Tajima's D	(Tajima, 1989)
Linkage disequilibrium	Cross population extended haplotype homozygosity (XP-EHH)	(Sabeti et al., 2007)
	haploPS	(Liu et al., 2013)
	Haplotype homozygosity (H12) and haplotype homozygosity statistic (H2/H1)	(Garud et al., 2015)
	Haplotype similarity (HS)	(Hanchard et al., 2006)
	Integrated extended haplotype homozygosity of a SNP site (iES)	(Tang et al., 2007)
	Integrated haplotype score (iHS)	(Voight et al., 2006)
	Kim and Nielsen's method	(Kim and Nielsen, 2004)
	Linkage disequilibrium decay test (LDD)	(Wang et al., 2006)
	Number of segregating sites by length (nS_L)	(Ferrer-Admetlla et al., 2014)
	Relative extended haplotype homozygosity (relative EHH)	(Sabeti et al., 2002)
Population differentiation	Ancestral branch statistic (ABS)	(Cheng et al., 2017)
	BayEnv	(Coop et al., 2010)
	Beaumont and Balding's method	(Beaumont and Balding, 2004)
	Efficient mixed-model association eXpedited (EMMAX)	(Kang et al., 2010)
	F_{ST}	(Weir, 1996)
	Derived allele frequency differences (ΔDAF)	(Colonna et al., 2014)
	Locus-specific branch length (LSBL)	(Shriver et al., 2004)
	PCAdapt	(Duforet-Frebourg et al., 2014)
	p_{excess}	(Hästbacka et al., 1994)
	Population branch statistic (PBS)	(Yi et al., 2010b)
Composite methods	Composite of Multiple Signals (CMS)	(Grossman et al., 2010)

1.1.3 Fine-Mapping of Adaptation

The challenges that existing statistical methods face is that they are unable to distinguish between neutral, passenger variants and true positively selected variants that are identified by genome-wide scans of positive selection in humans. Only a handful of variants have been functionally validated and conclusively shown to be responsible for the underlying adaptation signal in humans, although thousands of such signals have been mapped (Szpak et al., 2019) Fine-Mapping of Adaptive Variation (*FineMAV*) was developed to overcome this hurdle and provide a way forward to select variants that could be modelled *in vitro* or *in vivo* model systems (Szpak et al., 2018). As the name suggests, it is a method that pinpoints the variant, within a putative locus, that is driven by positive selection. It does this by incorporating methods that detect regions showing signatures of positive selection (population differentiation and high frequency of derived alleles) and merges it with functional annotation under the assumption that it is unlikely for a deleterious or functional variant to reach high frequency in a given randomly mating population unless it confers some sort of an advantage (Szpak et al., 2018).

To measure population differentiation, *FineMAV* employs a derived allele purity (*DAP*) equation to describe the disparate spread of derived alleles across populations (Szpak et al., 2018). The derived allele frequency (*DAF*) equation is used to determine sites with high frequency of derived alleles (Szpak et al., 2018). Combining the two would identify positively selected genomic regions (Szpak et al., 2018). The addition of functional annotation is what makes *FineMAV* different from existing statistical methods. To annotate functionality, it uses the Combined Annotation-Dependent Depletion (CADD) method which takes into account multiple variant annotations and condenses it into a single score called the C score (Kircher et al., 2014). According to Kircher et al. (2014), the C scores predict whether a SNP or indel in the human genome is functional, deleterious and pathogenic. In a PHRED-like scaled C score, the scores are expressed as rankings relative to all possible substitutions of the human genome and range from 1 to 99. For example, a variant that scores more than 20 would be within the top 1% of deleterious substitutions. A score of 30 would indicate top 0.1% and 40 would be 0.01% and so on (Kircher et al., 2014). If an allele was predicted to be deleterious and its frequency was low, the allele would probably be harmful. If the allele was deleterious but its frequency was high, it

would not signify that the allele is harmful, but rather, it may be an adaptive allele (Szpak et al., 2018). Incorporating CADD scores, therefore, enables us to differentiate between neutral alleles, which are predicted as non-deleterious, and true positively selected alleles, which are predicted as effectively functional or deleterious (Szpak et al., 2018).

Integrating these three metrics allow *FineMAV* to prioritise candidate positively selected genetic variants for functional validation in a high-throughput fashion (Szpak et al., 2018).

1.2 Genomic scans of positive selection

Most genome-wide selection scans in humans have been based on populations of European origins, although some have used populations from mainland East Asia, Japan or Africa (Reviewed by Akey 2009). To address this discrepancy, recently, there have been growing efforts in East and Southeast Asia to develop whole genome sequencing datasets which can be used to capture genomic evidence for local adaptation.

1.2.1 Positive selection in East Asia

In recent years, many positive selection scans have been performed on East Asian populations, mostly Chinese and Japanese, and this is, in part, due to the inclusion of East Asians in international genome-wide datasets (Table 2). Two of the strongest selection signals observed in East Asians is in the alcohol dehydrogenase (*ADH*) gene cluster, a ~370kb segment on chromosome 4 that consists of seven *ADH* genes, and the aldehyde dehydrogenase 2 family member (*ALDH2*) on chromosome 12. (Han et al., 2007; Teo et al., 2009; Okada et al., 2018; Yasumizu et al., 2020). However, the positively selected causal variant(s) from these regions have not been elucidated. SNPs from these regions have been associated with alcohol dependence, assortative mating related to alcohol consumption and risky behaviour (Frank et al., 2012; Park et al., 2013; Lai et al., 2019; Linnér et al., 2019). In Japanese populations, variants in alcohol dehydrogenase 1B (*ADH1B*), located in the *ADH* gene cluster, and *ALDH2*, were found to be associated with all-cause mortality (Sakaue et al., 2020). There are several selection signals identified in East Asians that are less studied and it is unsure as to why these genes have

undergone positive selection. This includes melanoma-associated antigen E2 (*MAGEE2*) and centromere protein W (*CENPW*) (Cheng et al., 2017; Szpak et al., 2018; Wu et al., 2019).

Table 2: List of datasets that had genome-wide East and Southeast Asian positive selection scans performed on them.

Name	Type of data	Include East and Southeast Asians?	Reference(s)
1000 Genomes Project	Low-coverage whole genome sequencing	East Asians	(The 1000 Genomes Project Consortium, 2010, 2012, 2015)
Asian Diversity Project (ADP)	Genotype	East and Southeast Asians	(Liu et al., 2017)
Genotyping of indigenous ethnic groups in northern Borneo	Genotype	Southeast Asians	(Yew et al., 2018)
HUGO Pan-Asian SNP dataset	Genotype	East and Southeast Asians	(The HUGO Pan-Asian SNP Consortium, 2009)
Human Genome Diversity Project (HGDP)	Genotype [a]	East Asians	(Cann et al., 2002; Li et al., 2008)
International HapMap Project	Genotype	East Asians	(The International HapMap Consortium, 2005, 2007; The International HapMap 3 Consortium, 2010)
Perlegen dataset	Genotype	East Asians	(Hinds et al., 2005)
SG10K (from Singapore)	High-coverage whole genome sequencing	East and Southeast Asians	(Wu et al., 2019)
Singapore Genome Variation Project (SGVP)	Genotype	East and Southeast Asians	(Teo et al., 2009)

[a] The samples in the HGDP were also sequenced recently by Bergström et al. (2020), but no genome-wide positive selection scans have been reported, thus far, for this sequenced dataset.

Perhaps the most extensively studied positively selected genes in East Asia are genes related to pigmentation. East Asians, like Europeans, have lighter skin and studies were conducted to investigate whether they share genetic variants associated with depigmentation. Genes like KIT ligand (*KITLG*) have selection signals in both Europeans and East Asians, suggesting that there may have been a selective event prior to these two populations splitting (Izagirre et al., 2006; McEvoy et al., 2006; Voight et al., 2006; Lao et al., 2007; Pickrell et al., 2009). Selection signals in oculocutaneous albinism II (*OCA2*) were also observed in both populations. However,

the light skin alleles that had undergone sweep in these populations are different, suggesting this trait evolved independently (i.e. convergent evolution) (Lao et al., 2007; Edwards et al., 2010). Several variants in some pigmentation genes (e.g. *DCT*, *ADAM17*, *MFSD12*), were positively selected in East Asians but not in Europeans, which is further evidence that convergent evolution towards lighter skin pigmentation has taken place (Izagirre et al., 2006; McEvoy et al., 2006; Norton et al., 2006; Voight et al., 2006; Lao et al., 2007; Myles et al., 2007; Hider et al., 2013; Adhikari et al., 2019).

In terms of hair morphology, the missense mutations in ectodysplasin A receptor (*EDAR;* rs3827760) and serine protease 53 (*PRSS53*; rs11150606) were found to be positively selected in East Asians (Fujimoto et al., 2007; Sabeti et al., 2007; Fujimoto et al., 2008; Kamberov et al., 2013; Adhikari et al., 2016; Wu et al., 2016; Szpak et al., 2018). The *EDAR* variant has been consistently identified in genomic scans on East Asian populations and observed to have pleiotropic effects in which rs3827760 has also been associated with shovel-shaped incisors (Kimura et al., 2009; Park et al., 2012b; Tan et al., 2014), ear shape (Adhikari et al., 2015; Shaffer et al., 2017), increased density of sweat glands, reduced mammary fat pad and increased branching in the mammary ductal gland (Kamberov et al., 2013).

Populations that live at extremely high altitudes in the Himalayas, such as those from Bhutan, Nepal and Tibet, were found to have positively selected genes (e.g. *EPAS1*, *EGLN1*, *PPARA*) associated with the hypoxia response (Simonson et al., 2010; Yi et al., 2010a; Hackinger et al., 2016; Arciero et al., 2018). In particular, intronic variants in *EPAS1* stand out and have been shown to be a classical example of archaic introgression in humans, indicating interbreeding between Denisovans, an extinct human species, and modern Tibetans (Huerta-Sánchez et al., 2014).

1.2.2 Positive selection in Southeast Asia

There have been many studies conducted on Southeast Asian genomic data, especially the Malay population, that have examined the genetic diversity, admixture and migration of these populations (Hatin et al., 2011; Ismail et al., 2013; Hatin et al., 2014; Wan Juhari et al., 2014; Yahya et al., 2017; McColl et al., 2018). Most of them relied on genotype datasets such as

the Malaysian Node of the Human Variome Project (MyHVP) (Halim-Fikri et al., 2015). However, Southeast Asian populations are underrepresented in whole genome sequencing datasets. As of now, there are five datasets that incorporate Southeast Asian populations: the Singapore Sequencing Malay Project (SSMP) (Wong et al., 2013), the Singaporean SG10K (Wu et al., 2019), the Estonian Biocentre Human Genome Diversity Panel (Pagani et al., 2016), the GenomeAsia 100K (GenomeAsia100K Consortium, 2019) and the Indonesian Genome Diversity Project (Jacobs et al., 2019). Due to the lack of representation and access to Southeast Asian genomic sequences, there is a poor understanding of how local adaptation occurred in this region and it is difficult to pinpoint causal variants. Studies that examined positive selection in Southeast Asian populations mainly use five datasets (Table 2), of which three are genotype datasets based on SNP chips.

In Southeast Asia, there seems to be more interest in indigenous ethic groups compared to urban populations because they have lived longer in the region and have been exposed to more diverse and harsher environments and, therefore, would give better genetic insight into the local adaptation that has occurred there. For example, Southeast Asia has a long history of endemic malaria (Copeland, 1935; William et al., 2013). Genome-wide genotyping of native individuals from Peninsular and East Malaysia and Taiwan have identified several genes (e.g. *HBB*, *TSBP1*) that may have conferred malaria resistance and increased their survival (Deng et al., 2014; Liu et al., 2014; Liu et al., 2015; Hoh et al., 2020). Another example is the Bajau people, also known as the Sea Nomads. They have resided in the coastal areas of Southeast Asia for over 1,000 years and free dive to gather their food (Sather, 1997). Their lifestyle led to genetic adaptations (e.g. *PDE10A*, *BDKRB2*) which may have enhanced their abilities to hold their breath under water (Ilardo et al., 2018). Philippine Negritos may have undergone convergent evolution towards short stature as a result of adaptations to life in hot and dense tropical forests (Migliano et al., 2013). A hypothesis proposed by Migliano et al. (2007) suggests that perhaps having short stature is an evolutionary trade-off between growth and reproduction (i.e. attaining sexual maturation early, and therefore, early growth secession, would ensure reproduction in environments of high mortality). Other strong selection signals, like forkhead box Q1 (*FOXQ1*) and phosphoinositide-3-kinase regulatory subunit 3 (*PIK3R3*), were seen in Philippine Negritos (Qian et al., 2013). *FOXQ1* is associated with metastasis in humans and *PIK3R3* is known to regulate the activity of protein-

tyrosine kinase and is responsible for cell signalling (Mothe et al., 1997; Qiao et al., 2011). It is unknown as to what selection pressures were responsible for this.

1.2.3 Research gaps

There are two research gaps that this project aims to address. The first research gap is that most prior positive selection studies in East and Southeast Asia used SNP arrays to collect their data (Table 2), instead of sequencing, as the technology has been around longer. SNP arrays have a well-known ascertainment bias, being discovered mainly in European populations (Lachance and Tishkoff, 2013) and these do not capture the entire genome and therefore, there may be many genetic variants that have been unaccounted for in these positive selection studies. Secondly, as seen in Table 2, positive selection in Southeast Asians are less understood because they are heavily underrepresented in these datasets compared to East Asian populations. At the time of analysis, no positive selection scans were done on whole genome sequences of Southeast Asians and this was due to the late arrival of whole genome sequencing in the region. Since then, only one positive selection scan was done, and it was performed on a recently published dataset consisting of three ethnic groups in Singapore using a population differentiation-based statistical test called population branch statistics (PBS) (Wu et al., 2019). They found a total of 20 candidate loci for positive selection in the Chinese, Indian and Malay population groups. They identified several loci that have been known to be selected in Asians (e.g. *EDAR, PRSS53, OCA2*) as well as lesser-known ones (e.g. *CENPW, MAGEE2*) (Wu et al., 2019). Only one out of the 20 loci (*FAM178B*) were specific to Malays. Compared to my findings, where I investigated positive selection using *FineMAV* on Chinese, Malay and Indian population groups, only four of the 20 loci were identified.

1.3 Objectives of this study

To fill in the research gaps, this project leverages on publicly available, high-coverage whole genome sequencing datasets that have East and Southeast Asian populations from China and Singapore, and include populations that are genetically close to populations from Malaysia, to highlight highly differentiated candidate variants in these populations that are most likely to underlie positive selection signals, using an algorithm called *FineMAV*.

There were three objectives of this project:

1. Use *FineMAV* to identify the population-specific variants in whole genome sequences obtained from the Chinese, Malay and Indian population groups in Southeast Asia.
2. Display the genome-wide *FineMAV* statistics in a human genome browser, like the University of California Santa Cruz (UCSC) Genome Browser (Kent et al., 2002; Navarro Gonzalez et al., 2020), to enable visualisation of the associated genetic variant that appears to be under selection in these populations in their genomic context. This will facilitate our understanding and modelling of adaptations that were associated with human settlement in this part of Asia.
3. To make *FineMAV* more accessible for researchers by developing a command-line and graphical user interface software for researchers to calculate the *FineMAV* statistic from population datasets of their interest. This allows researchers to generate *FineMAV* scores for their sequencing data and display the output as a customized track in the human genome browser.

2 RESEARCH METHODOLOGY

2.1 Whole genome sequencing datasets

Five publicly available, whole genome sequencing datasets were used and were filtered differently depending on the nature of the datasets and how it was used (Table 3). No ethical approval was required to obtain the datasets as the researchers who developed them have already received the relevant institutional ethical approval to conduct their study and the participants are aware that their genomic data is made public and used worldwide by the research community.

China officially recognizes 56 ethnic groups which exclude unknown ethnic groups and foreigners carrying the Chinese citizenship (Guo, 2017). Most of China is predominately Han Chinese as they account for more than 90% of the population (Guo, 2017). They have migrated to a plethora of countries and large Han communities can be found in every continent (excluding Antarctica) (Minahan, 2014). The 90 Han Chinese genomes dataset (90HC) was used to represent these populations in this study (Lan et al., 2017).

Singapore is a city-state situated at the tip of Peninsular Malaysia in Southeast Asia and has a history of migration which has led to it being the melting pot of various ethnicities it is today. Three major groups present in Singapore include the Chinese, Malay and Indian ethnicities, that are also found in Malaysia (Saw, 2012). Besides Singapore and Malaysia, the Malay ethnic group can also be found in neighbouring regions like Brunei, Southern Thailand (Pattani), Indonesia, Southern Philippines and Sri Lanka (West, 2009; Hatin et al., 2014; Deng et al., 2015). During the initial stages of migration, there were intermarriages between the migrant Chinese and Indian men and the local Malay women (Mathews, 2018). However, these interracial unions declined after the large migration of women from China and India, as they preferred spouses with similar ethnic origin as themselves (Mathews, 2018). Individuals with a multi-ethnic background from these three groups were traditionally assigned the ethnicity of their father (Rocha, 2011). However, within the last decade, to acknowledge hybrid identities, children of mixed parentage can be registered as having a “double-barrelled race” (Rocha, 2011).

Table 3: List of whole genome sequencing datasets used for *FineMAV* analysis.

Genome project	URL	Number of individuals [a]	Population groups	Reference
Sequencing of 90 Han Chinese genomes (90HC)	https://www.ebi.ac.uk/ena/data/view/PRJEB20820	90	Han Chinese from China	(Lan et al., 2017)
Singapore Sequencing Indian Project (SSIP)	https://blog.nus.edu.sg/sshsphphg/singapore-sequencing-indian/	35	Singaporean Indian	(Wong et al., 2014)
Singapore Sequencing Malay Project (SSMP)	https://blog.nus.edu.sg/sshsphphg/singapore-sequencing-malay/	96	Singaporean Malay	(Wong et al., 2013)
1000 Genomes Project (Phase 3)	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/	1668 [b]	See Appendix C	(The 1000 Genomes Project Consortium, 2015)
GenomeAsia 100K	https://browser.genomeasia100k.org/#tid=download	1428	See Appendix D	(GenomeAsia100K Consortium, 2019)

[a] Refers to number of individuals that were included in the *FineMAV* analysis and not the total number of individuals in the dataset.
[b] Performance evaluation of the *FineMAV* software was performed only on African, European and East Asian populations and not the American and South Asian populations.

The majority of Singaporean residents are of Chinese ethnicity and as of 2017, they consist of 74.3% of the population (Singapore Department of Statistics, 2017). The term "Chinese" refers to those of broad Chinese origin and they are subcategorised into groups based on their dialect such as Hokkien, Teochew or Cantonese (Saw, 2012). They are mainly of Han Chinese ancestry (Minahan, 2014). The Malay make up 13.4% of the population (Singapore Department of Statistics, 2017) and refers to those of Malay or Indonesian origin. Due to their racial, cultural and religious similarities, the Indonesian immigrants assimilated with the Malays (Saw, 2012). Lastly, the Indian ethnic group sits at 9.0% (Singapore Department of Statistics, 2017) and refers to individuals whose origins lie in the Indian sub-continent such as India, Pakistan, Bangladesh and Sri Lanka (Saw, 2012)

For the first objective, the genome-wide analysis was performed on three high-coverage, whole genome sequencing datasets from China and Singapore: the sequencing of 90 Han Chinese

genomes dataset (90HC), 35 Singaporean Indians from the Singapore Sequencing Indian Project (SSIP) and 96 Singaporean Malays from the Singapore Sequencing Malay Project (SSMP) (Table 3). The Variant Call Format (VCF) files for each dataset, which stores genotype information of the individuals for each SNP (Danecek et al., 2011), were downloaded from the URLs listed in Table 3. Only the autosomal and the X chromosome VCF files were used in this project. At the time of analysis, no whole genome sequences for the Singaporean Chinese group were made publicly available, and, therefore, the publicly available Han Chinese dataset (90HC) was used as a proxy for the Singaporean Chinese population.

I also tested the software on the 1000 Genomes Project (Phase 3) (The 1000 Genomes Project Consortium, 2015) in order to replicate previous analysis performed by Szpak et al. (2018). Subsequently, I also used the recently published GenomeAsia 100K datasets (GenomeAsia100K Consortium, 2019). The GenomeAsia 100K includes 1,428 individuals from four continental regions: Northeast Asia, South Asia, Southeast Asia, and Oceania (Appendix D).

2.2 Filtering the whole genome sequencing datasets

Data filtering was performed to select high-quality biallelic SNPs (Figure 1). Filtering was done using a combination of software programmes: BCFtools v1.9 (Li et al., 2009b), PLINK v1.9 (Chang et al., 2015) and SnpSift v4.3.1 (Cingolani et al., 2012). To filter out highly related individuals, I opted for a PI_HAT threshold of 0.35. This was based on observing the pairwise PI_HAT values of the individuals from the datasets (Appendix E). 95 individuals were removed from the 1000 Genomes Project dataset, therefore leaving the total number of individuals in the dataset to be 2,409. One individual from the Singaporean Sequencing Indian Project (SSIP) was also removed, which resulted in 35 individuals used for downstream analysis.

As seen in Figure 1, the datasets underwent different filtering procedures depending on whether the data will be used for population structure analysis (Chapter 2.3) or for generating the genome-wide *FineMAV* scores. For population structure analysis, the Singaporean and Chinese datasets were compared to 26 worldwide populations from the 1000 Genomes Project

(Phase 3) (The 1000 Genomes Project Consortium, 2015) (Table 3). The populations originate from five continental regions: Africa, the Americas, Europe, East Asia and South Asia (Appendix

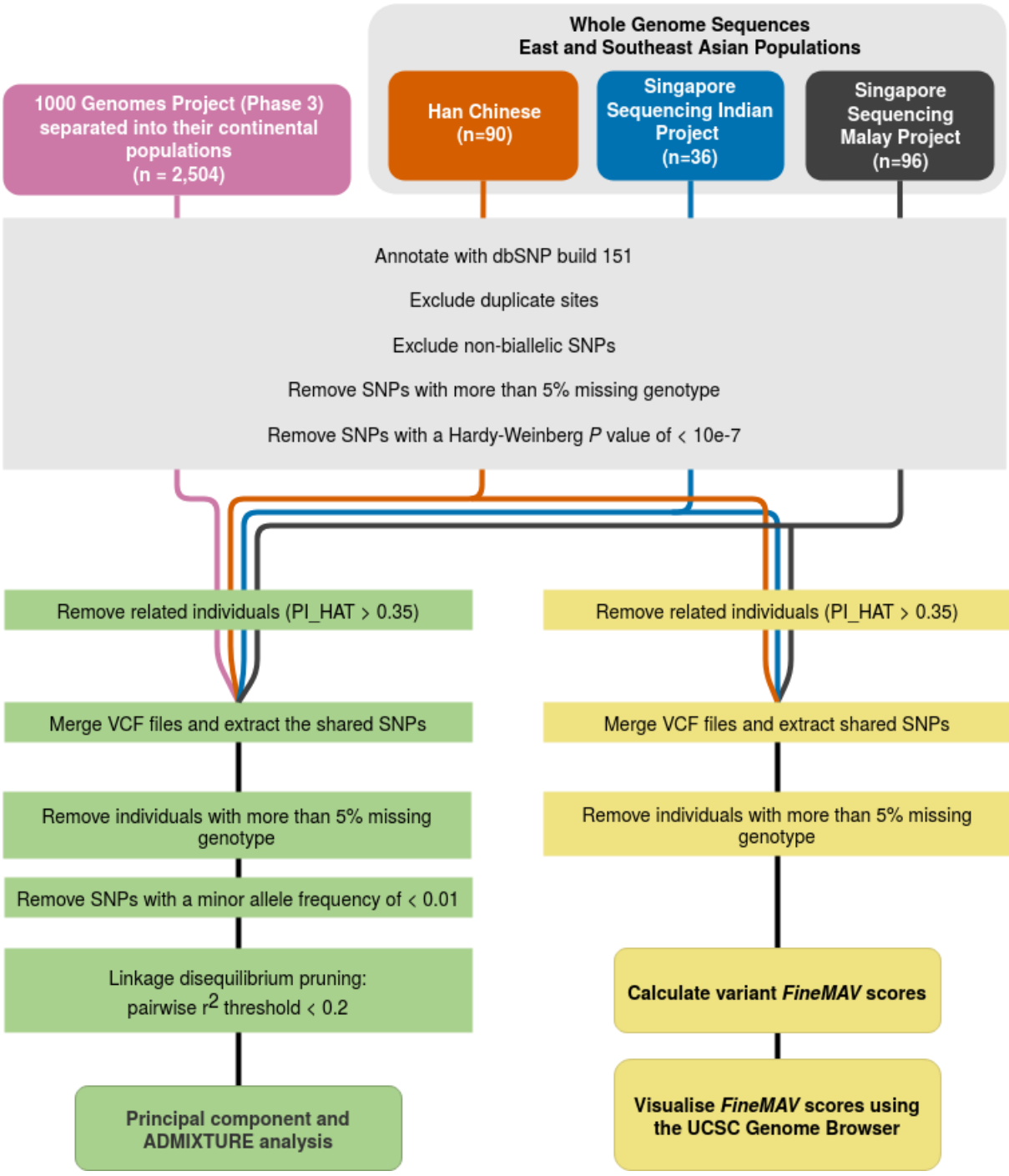


Figure 1: Dataset filtering workflow for population structure (principal component and ADMIXTURE analyses) and generation of *FineMAV* scores from whole genome sequences.

C). The 1000 Genomes Project required its own set of filtering procedures as illustrated in Figure 1.

2.3 Population structure analysis

To ensure that the datasets recapitulate well-established population genetic structure, I conducted principal component (PCA) and ADMIXTURE analyses of the filtered datasets, and compared them with the 26 worldwide populations from the 1000 Genomes Project Phase 3 dataset (The 1000 Genomes Project Consortium, 2015). The analyses were conducted on merged and pruned data (Figure 1) comprising 239,996 autosomal SNPs from 2,630 individuals. It should be noted that there are 83 Han Chinese individuals in the 90HC dataset that have also been sequenced by the 1000 Genomes Project. These were used for checking the quality of the high-coverage sequenced dataset. There was a 99.80% genotype concordance between the low-coverage 1000 Genomes Project samples and the 83 90HC samples. Subsequently, the corresponding low-coverage samples from the 1000 Genomes Project were removed from further downstream analysis.

The PCA and unsupervised ADMIXTURE analysis was carried out using PLINK v1.9 (Chang et al., 2015) and ADMIXTURE v1.3 (Alexander et al., 2009) respectively. For ADMIXTURE analyses, each ancestral component *K*, from 2 to 14, were repeated 10 times with different random seeds and the outcome with the highest likelihood was chosen. Five-fold cross-validation was used to determine the most optimum *K* value. After preliminary filtering and quality checks of the datasets, the population sub-structure was investigated to see whether the data agrees with the expected published population relationships. Since the population genetic relationships were as expected, the *FineMAV* scores for the datasets were subsequently generated.

2.4 *FineMAV* algorithm and analysis

For the third objective, which is to develop a software program for researchers to calculate *FineMAV* scores for datasets of their interest, I tested the software on the 1000 Genomes Project (Phase 3) (The 1000 Genomes Project Consortium, 2015) and the recently published GenomeAsia 100K dataset (GenomeAsia100K Consortium, 2019) (Table 3). The authors

who developed *FineMAV* applied the statistic on the African, European, and East Asian populations of the 1000 Genomes Project to assess whether it was able to pinpoint experimentally validated, positively selected variants and identify other novel variants for functional follow-up. To ensure that the software could correctly calculate the *FineMAV* scores, I compared the scores I generated to the ones the authors published. For this, the 1000 Genomes Project was filtered only to include biallelic SNPs from the autosomal and sex chromosomes.

The software was also tested using the GenomeAsia 100K to evaluate its performance on larger datasets The GenomeAsia 100K includes populations from four continental regions: Northeast Asia, South Asia, Southeast Asia, and Oceania (Appendix D). However, the complete VCF files containing the individual genotype information required approval from their data access committee. I opted to use composite VCF files which they made publicly available. These files contain the allele frequencies of each continental region from the autosomal chromosomes. Because the dataset had already merged the populations together and filtered them, no additional filtering from my side was performed. I tested the software only on biallelic SNPs.

The *FineMAV* score of the derived allele for each SNP was calculated by multiplying three metrics: derived allele purity (*DAP*), derived allele frequency (*DAF*) and the PHRED-like scaled CADD score (CADD_PHRED) (Equation 1) (Szpak et al., 2018).

Equation 1B was used to calculate *DAP*, a metric used to describe the disparate spread of the derived alleles across populations. *DAP* = 1, which is the maximum possible value, would signify that all the derived alleles are found in a single population. If the derived allele is shared between populations, this value is penalised. It relies on the penalty parameter (x) to maximise the magnitude between differentiated positively selected variants and less differentiated nearby neutral variants. It was determined empirically for different values of n populations (Table 4) (Szpak et al., 2018). For my analysis, I opted for the value of x that was recommended by Szpak et al. (2018) when evaluating between three populations, which is x = 3.50. The value of *DAP*, as per Equation 1B, is based on derived allele counts if the population sizes are equal. If the population sizes are different, which is true for my analysis, the derived allele frequency is used in lieu of the derived allele counts.

Equation 1: Equations involved in calculating *FineMAV*.(A) *FineMAV* scores were calculated for n populations. The derived allele purity (*DAP*), derived allele frequency for each population (*DAF*) and the PHRED-like scaled CADD scores (Kircher et al., 2014) are multiplied together. In this equation, $i \in \{1, 2, \ldots, n\}$. (B) *DAP* is computed per site across n populations. d_i represents the derived allele count in one population where $i \in \{1, 2, \ldots, n\}$.

(A)

$$FineMAV_i = DAP \times DAF_i \times CADD$$

(B)

$$d_N = \sum_{i=1}^{n} d_i$$

$$f_i = \frac{d_i}{d_N}$$

$$DAP = \sum_{i=1}^{n} f_i^x$$

Table 4: Recommended minimal value of the penalty parameter (x)rounded off to two decimal places, for a given n as determined by Szpak et al. (2018).

Number of populations (n)	**Penalty parameter (x)**
2	4.96
3	3.50
4	2.98
5	2.71
6	2.53
7	2.41

2.5 Pipeline for calculating the *FineMAV* scores

Following the filtering described in Figure 1, the *FineMAV* scores were calculated for the final, merged VCF dataset consisting of SNPs that were polymorphic in all three populations. This amounted to 5,774,118 SNPs from 90 Han Chinese, 35 Singaporean Indian and 96 Singaporean Malay individuals.

All three objectives were achieved using the bioinformatics pipeline illustrated in Figure 2. When I initially formulated the pipeline, it started off as deconstructed, where each step was done one by one, and it constituted multiple Python-based scripts and intermediate files. It was then optimised and automated to take the least amount of time and files. This section will elaborate on how the pipeline works. Chapter 3.2 of the Results and Discussion will describe the software from a user's perspective.

The *FineMAV* software requires the user to provide the following information from the VCF file in a tab-delimited file format, which is a simple text format in which the columns of the table are separated by a tab character (Table 5). In a VCF file, variations (whether it be SNPs or indels) are recorded based on their position on a standardised reference genome. This study used GRCh37 (hg19) as the human reference genome. The REF column (Table 5) refers to the reference allele that is found in the reference genome. The ALT column (Table 5) refers to the non-reference alleles.

If a VCF file contains all the necessary information listed in Table 5, then it can easily be extracted using software programmes that can manipulate VCF files, such as BCFtools (Li et al., 2009b). However, in most instances, the VCF file would not contain the required fields. In instances where the allele frequency (AF) for each population is not annotated in the VCF file, the AF can be calculated using a plugin on BCFtools called "fill-tags" when it is supplemented with a list of individuals from each population. Once calculated, it can be extracted to create the input file for the *FineMAV* software. If the ancestral allele and/or the CADD_PHRED is not available in the VCF file, the software allows the user to supplement this information using the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016), a well-known software that can import various annotations from different sources.

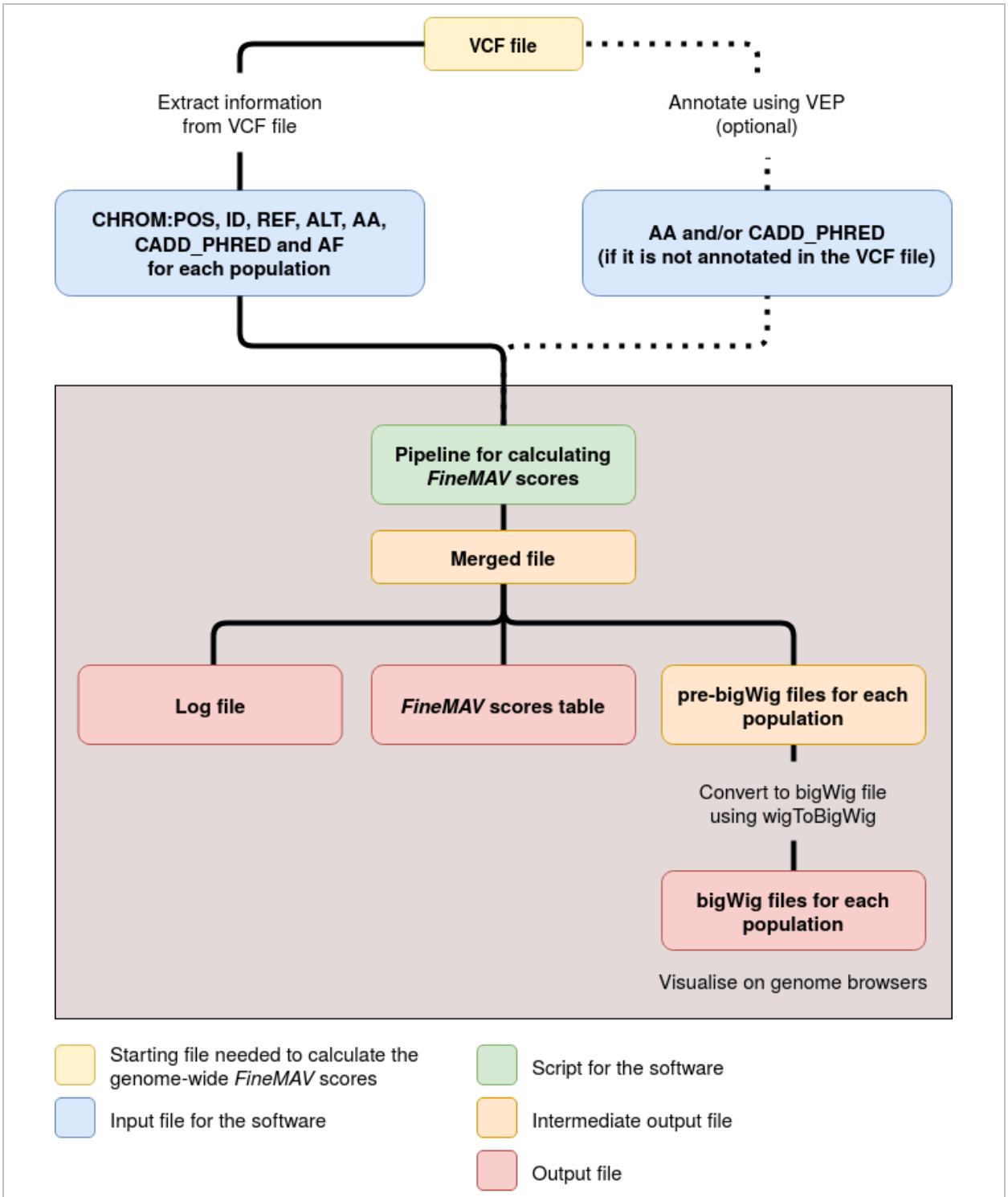


Figure 2: Pipeline for calculating the genome-wide *FineMAV* scores. The grey boxed region highlights the parts of the workflow that are automated by the software I developed. The intermediate output files are deleted when the pipeline is complete.
VCF: Variant Call Format, VEP: Variant Effect Predictor. The abbreviations for the column names can be found in Table 5.

For my analysis, the datasets did not have the updated ancestral allele nor the CADD_PHRED annotated scores. Therefore, the annotations had to be supplied using Ensembl VEP (McLaren et al., 2016). Ensembl VEP relies on its plugins to retrieve the CADD_PHRED score and ancestral allele data for each SNP when supplied with the appropriate files. The data file for the latest CADD_PHRED version (v1.4) for reference genome GRCh37/hg19 can be found here: https://krishna.gs.washington.edu/download/CADD/v1.4/ GRCh37/whole_genome_SNVs.tsv.gz (Kircher et al., 2014). The AncestralAllele plugin fetches ancestral allele information from FASTA files containing the ancestral sequences that were inferred from the multiple species alignment of six primates (Paten et al., 2008a; Paten et al., 2008b). The FASTA files were downloaded here: ftp://ftp.ensembl.org/pub/release75/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh37_e71.tar.bz2. This was a lengthy process as it took almost 12 hours to complete. After the two input files were generated by using BCFtools v1.9 and Ensembl VEP v98, they were fed into the software for the genome-wide *FineMAV* scores to be calculated.

Table 5: Information that can be extracted from the VCF file and provided in a tab-delimited format for the software to calculate the *FineMAV* scores.

Information needed from the VCF file	**Description**	**Mandatory VCF column**
CHROM:POS	Chromosome number and position	Yes
ID	Identifier(s), if available. It is usually the dbSNP ID number.	Yes
REF	Reference base	Yes
ALT	Alternative base(s)	Yes
AA	Ancestral allele	No
CADD_PHRED	PHRED-scaled Combined Annotation Dependent Depletion (CADD) score	No
AF	Allele frequency for each ALT allele. The AF should be reported for each population.	No

The *FineMAV* software generates three different types of output files (Figure 2). The first is a basic log file which records metadata such as the number of SNPs that were analysed and how many of them do not have ancestral allele information. The second file is a table containing the *FineMAV* scores for each population along with the intermediate calculations. This would include the derived allele frequency (*DAF*) for each population and the derived allele purity (*DAP*) for each derived allele. The third type of output file is a bigWig (Kent et al., 2010) which is a format that is commonly used for graphical visualisation on human genome browsers, such as that hosted online by the UCSC (Kent et al., 2002; Navarro Gonzalez et al., 2020) and the downloadable Integrative Genome Viewer (IGV) (Robinson et al., 2011). For this thesis, the bigWig graphical visualisation is showcased on IGV as it does not require the bigWig files to be made publicly available. The bigWig format is compressed, converted to binary and indexed. This makes it appropriate for viewing larger datasets because it allows the genome browser to access and load data that only pertains to the genomic region that is currently in view. The wigToBigWig utility mentioned in Figure 2 is used to convert the pre-bigWig output files produced by the *FineMAV* software into bigWig files. For the wigToBigWig to work, it requires a text file listing the name of the chromosomes and their corresponding size. This file was downloaded here: ftp://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.chrom.sizes.

2.6 Analysing the top *FineMAV* variants

In population genomics, it is assumed that only a number of SNPs would experience some form of selection and the majority would undergo genome-wide forces such as genetic drift (Black IV et al., 2001; Akey, 2009). It is common practice to set a threshold, often arbitrary, to decide which SNPs can be considered positively selected (Akey, 2009). In my case, I set it to the 99.99th percentile of the *FineMAV* score distribution.

For the 90HC, SSIP and SSMP, only SNPs that were polymorphic in the three datasets were included in the analysis. This would mean that known positively selected SNPs that are not polymorphic or missing in this dataset would be missing from my analysis (Table 6). This could result in highlighting adjoining population-specific loci that, if functionally relevant, could also result in high *FineMAV* scores because of the effect of genetic hitchhiking. To evaluate this, I

performed pairwise comparisons of linkage disequilibrium (LD) using Haploview (v 4.2) (Barrett et al., 2005) between the known SNPs and the SNPs of the same chromosome that are listed in the top 50 genome-wide *FineMAV* outliers across continental or regional populations. Pairwise LD was also performed between the top 50 *FineMAV* variants obtained from my analysis (90HC, SSIP, SSMP and the GenomeAsia 100K) with each other, and to the published *FineMAV* scores that were generated from the 1000 Genomes East and South Asian populations to determine if the high-scoring variants were close together in the same population. LD was measured using r^2 values.

Table 6: List of notable positively selected variants that have been identified in East and South Asia using *FineMAV*and whether these variants are polymorphic in all three datasets: the Han Chinese (90HC), the Singaporean Indian (SSIP) and the Singaporean Malay (SSMP). The most severe variant consequence according to Ensembl is included.

Gene	SNP ID	Consequence	Population	Polymorphic
EDAR	rs3827760:A>G	Missense (p.Val370Ala)	East Asian	Yes
ZAN	rs2293766:G>A	Stop gained (p.Trp1883Ter)	East Asian	Yes
MAGEE2	rs1343879:C>A	Stop gained (p.Glu120Ter)	East Asian	No
PRSS53	rs11150606:T>C	Missense (p.Gln30Arg)	East Asian	No
PRSS53	rs201075024:C>T	Missense (p.Gly34Ser)	South Asian	No

3 RESULTS AND DISCUSSION

3.1 Population structure analysis

The 90HC, SSIP and SSMP datasets were analysed together with the 1000 Genomes Project Phase 3 dataset, which contains 26 worldwide populations (Appendix C), to ensure that the dataset was representative of the region and genetic relationships among these populations were as expected.

In the PCA plot, based on 2,630 individuals, samples that cluster closer together are genetically similar. As expected, the Han Chinese (90HC) and the Singaporean Indian (SSIP) populations overlap with the East Asian and the South Asian populations from the 1000 Genomes dataset, respectively (Figure 3). The Singaporean Malay (SSMP) individuals are close to the Han Chinese and East Asians in the 1000 Genomes Project populations and cluster near Vietnamese individuals (KHV) in that dataset.

To further investigate and consolidate the PCA results, ADMIXTURE analysis was performed. Based on the cross-validation error graph in Figure 4A, the ADMIXTURE plot with *K*=10 is deemed to be the model that best represents the ancestry of the individuals in all four datasets. In an ADMIXTURE plot, each vertical bar represents an individual. The proportion of the different colours in each bar corresponds to the proportion of the estimated ancestry from the inferred ancestral populations. Therefore, individuals with similar coloured segments are genetically similar. As seen in Figure 4B where *K*=10, the major ancestral component in SSMP individuals (denoted in dark grey) can be found in moderate proportions in the Vietnamese individuals (KHV) and Chinese Dai in Xishuangbanna (CDX), a region in China that shares a border with Laos and Myanmar. The ancestral proportions for the 90HC, Han Chinese in Beijing, China (CHB) and Southern Han Chinese (CHS) populations are alike, with varying magnitudes of the grey-coloured component, which are considerably correlated with latitude. The SSIP individuals are genetically similar to other South Asian populations and, especially, to the Indian Telugu (ITU) and Sri Lankan Tamil (STU) populations in the 1000 Genomes Project dataset.

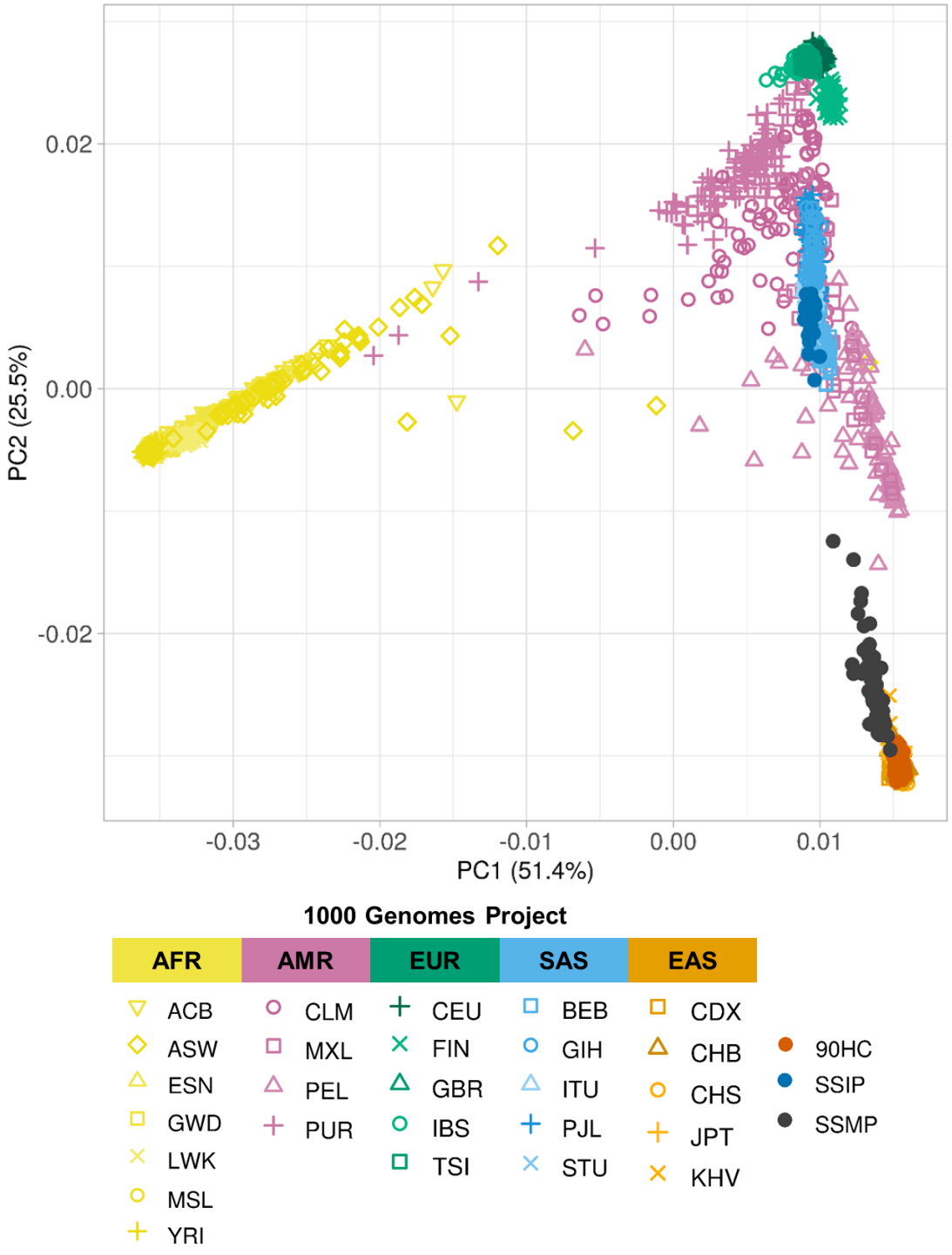


Figure 3: Principal component analysis (PCA) of the populations in the 1000 Genomes Project Phase 3 dataset with the Han Chinese (90HC), Singaporean Indian (SSIP) and Singaporean Malay (SSMP) populations. The percentage in the axis label indicates the proportion of the genotypic variance explained by each principal component. The population codes for the 1000 Genomes Project dataset can be found in Appendix C. This plot is based on 239,996 autosomal SNPs.

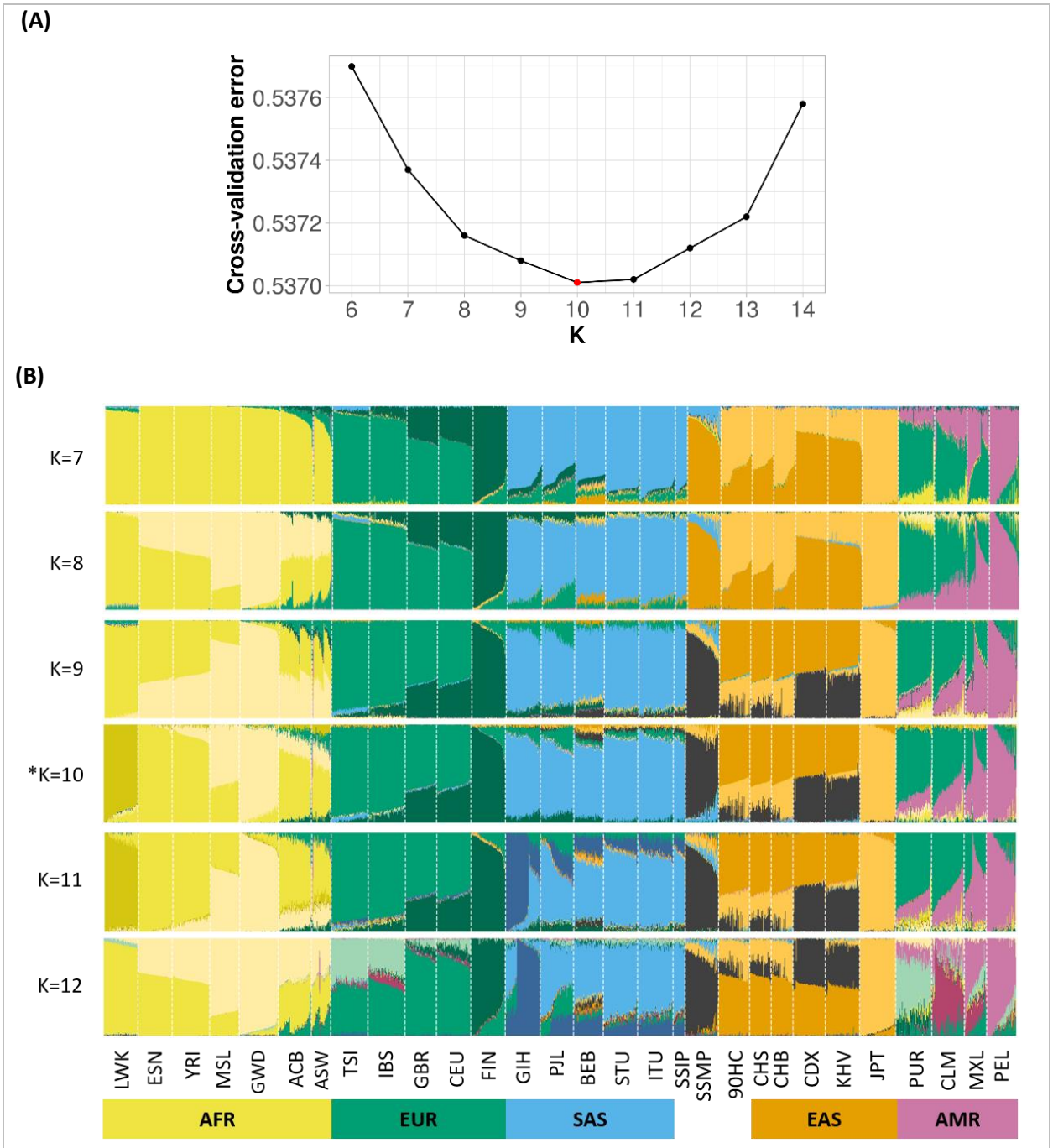


Figure 4: ADMIXTURE analysis of 2,409 individuals from the 1000 Genomes Project Phase 3 dataset, 90 Han Chinese individuals (90HC), 35 Singaporean Indian individuals (SSIP) and 96 Singaporean Malay individuals. **(A)** Five-fold cross-validation error for every *K* from 6 to 14. The point highlighted in red indicates the K with the lowest cross-validation error. **(B)** Ancestry proportions for *K* from 7 to 12. The asterisk at *K*=10 indicates with K with the lowest cross-validation error. The population codes for the 1000 Genomes Project dataset can be found in Appendix C. This plot is based on 239,996 autosomal SNPs.

The PCA and ADMIXTURE analysis are in agreement with one another and as genetic relationships are in line with the expectations for non-African populations. These plots (Figure 3 and Figure 4B) show that the Han Chinese are closely related with other East Asian populations and that Singaporean Indians share similar ancestry to other South Asian populations. Singaporean Malays, form a genetically distinct cluster, which is expected, as there are no Southeast Asian populations in the 1000 Genomes Project. It should be pointed out that they do share some ancestral component with individuals from Vietnam and Xishuangbanna, China, the two regions from mainland East Asia that are closest to Singapore.

3.2 Software usage and application

The Python-based *FineMAV* software works with sequencing data and relies on the information that can be extracted from VCF files (version 4.2 and above) (Table 5). To achieve a more complete scan, users are recommended to use jointly-called, multi-sample Genomic VCF (gVCF) files. Jointly-called variants would mean that the variants from all individuals were analysed and identified simultaneously as opposed to alone or separately in batches. To save time and computational storage, a typical VCF file would only record sites (SNPs and indels) that are different from the reference genome (i.e. record sites with variation). gVCF, on the other hand, is a type of VCF file that consists of every site in the genome regardless of whether they carry variation or not. Jointly-called gVCF files are preferable for *FineMAV* analysis because they make a clearer distinction between variants that are homozygous for the reference allele and have been sequenced, from those that have not been sequenced, or have missing data.

The software is available as a command-line interface (Figure 5A) and as a graphical user interface (GUI) (Figure 5B). To reduce the computational burden and optimise the random access memory (RAM) usage, the software performs these calculations by splitting the file(s) into smaller chunks and processing them chunk by chunk, as illustrated in Figure 6A and Figure 7A. The default size of a chunk is 200,000 lines. However, the user can specify the chunk size if they require (Figure 5). I also tested the performance of the chunk size option using the GenomeAsia 100K, a large dataset consisting of 66,236,516 biallelic SNPs across four population groups:

(**A**)

```
usage: finemav [-h] -i <file> -x <prefix> -r <hg19|hg38> [-v <file>]
               [-p <int|float>] [-c <int>]

Calculates the FineMAV scores

optional arguments:
  -h, --help            show this help message and exit

required flags:
  -i, --input-file <file>
                        file containing the CHROM:POS, REF, ALT and AF
                        (optional: AA, CADD_PHRED)
  -x, --prefix <prefix>
                        prefix assigned to the output files
  -r, --reference-genome <hg19|hg38>
                        hg19 or hg38

optional flags:
  -v, --vep-file <file>
                        file containing the LOCATION with the AA and/or
                        CADD_PHRED
  -p, --penalty <int|float>
                        penalty parameter
  -c, --chunksize <int>
                        number of lines per chunk (default=200000)
```

(**B**)

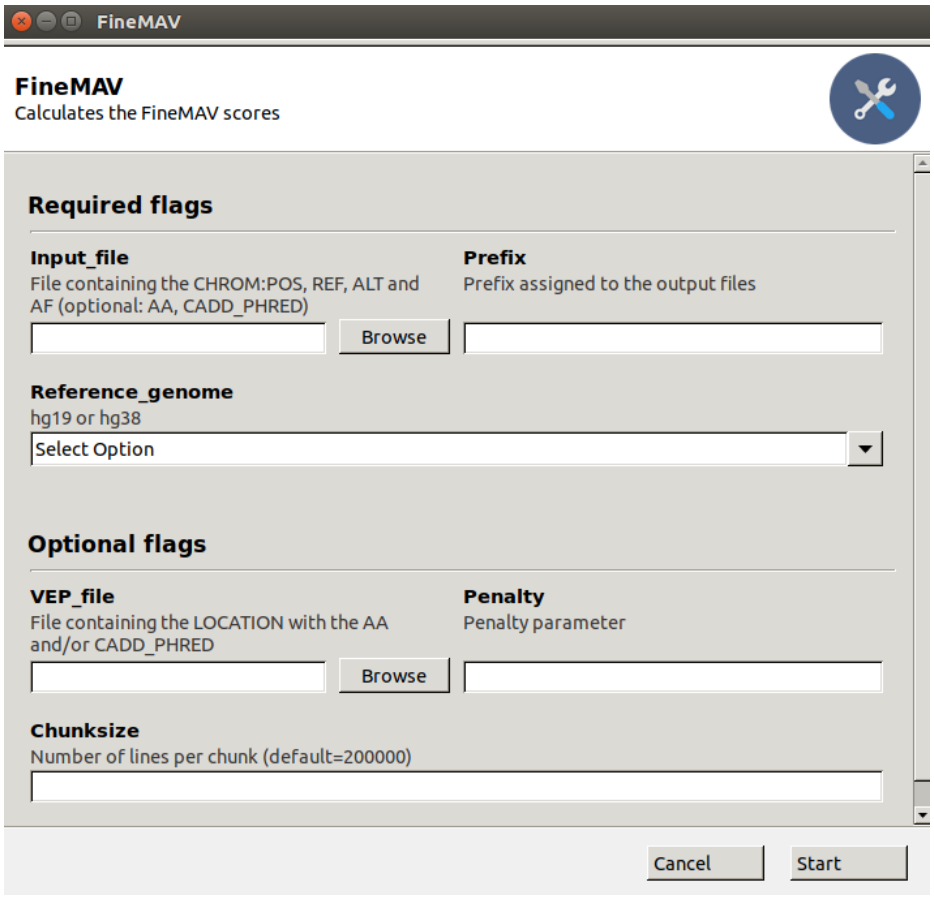
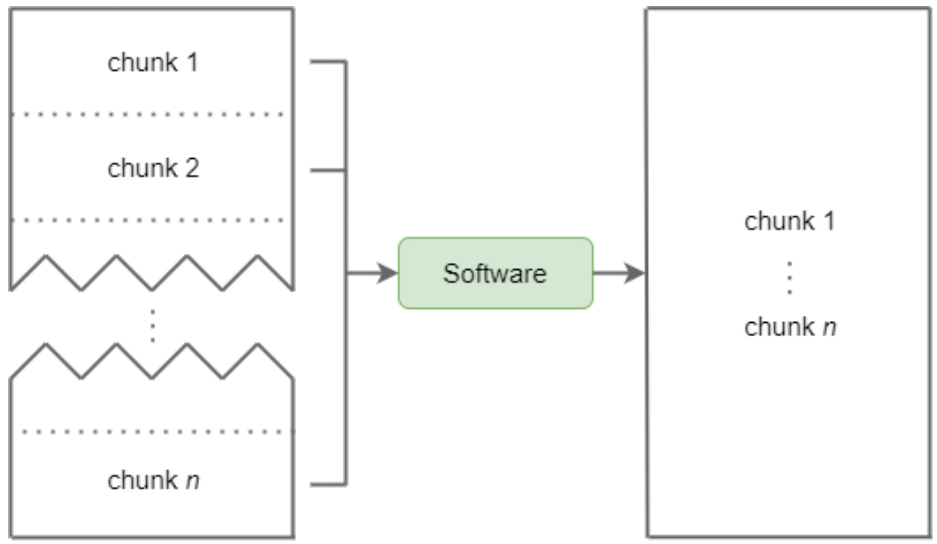


Figure 5: Screenshots of the *FineMAV* software as a **(A)** command line interface and a **(B)** graphical user interface.

(A)



(B)

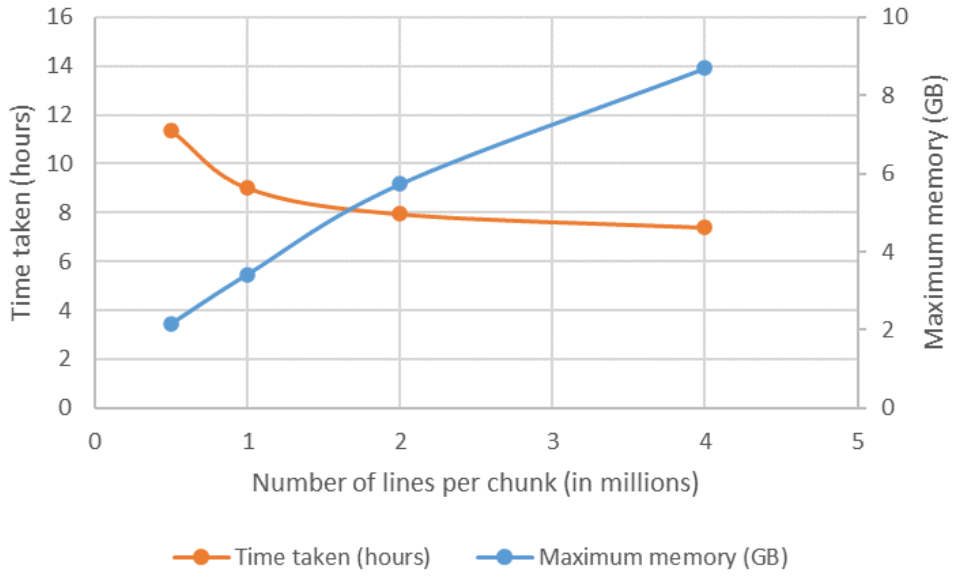


Figure 6: Utilising the chunk size option. **(A)** Diagram illustrating how the software separates the input files into chunks and iterates through them when performing the *FineMAV* calculations. It proceeds to merge them into one output file. **(B)** A graph that compares the time taken and the maximum random access memory (RAM) when different chunk sizes for the GenomeAsia 100K dataset, which consists of 66,236,516 biallelic SNPs.

(A)

```
Populations detected: '90HC', 'SSIP', 'SSMP'
Merging the two files
Total number of SNPs: 5774118
Calculating the FineMAV scores (chunk 1 of 29)
Exporting the FineMAV scores into a table (chunk 1 of 29)
Exporting the FineMAV scores into temporary .wig files (chunk 1 of 29)
Calculating the FineMAV scores (chunk 2 of 29)
Exporting the FineMAV scores into a table (chunk 2 of 29)
Exporting the FineMAV scores into temporary .wig files (chunk 2 of 29)
Calculating the FineMAV scores (chunk 3 of 29)
Exporting the FineMAV scores into a table (chunk 3 of 29)
Exporting the FineMAV scores into temporary .wig files (chunk 3 of 29)
Calculating the FineMAV scores (chunk 4 of 29)
Exporting the FineMAV scores into a table (chunk 4 of 29)
Exporting the FineMAV scores into temporary .wig files (chunk 4 of 29)
Calculating the FineMAV scores (chunk 5 of 29)
Exporting the FineMAV scores into a table (chunk 5 of 29)
Exporting the FineMAV scores into temporary .wig files (chunk 5 of 29)
```

(B)

```
Name of BCFtools file: calc_all_v4/90HCSSIPSSMP_withoutExtract_location_AN_AF.txt
Name of VEP file: calc_all_v4/merge_FM_v2_withoutExtract_CADD_AA.txt
Reference genome: hg19
Chunksize: 200000 lines per chunk

Penalty parameter: 3.5
Number of SNPs: 5774118

Number of SNPs where the ancestral allele...
does not exist: 117389
is the reference allele: 3773190
is the alternative allele: 1851472
is neither: 32067

Number of populations: 3
Population names: 90HC, SSIP, SSMP
Maximum FineMAV score for 90HC: 4.661128
Maximum FineMAV score for SSIP: 7.677376
Maximum FineMAV score for SSMP: 3.378257

Time taken: 2683.77 seconds
```

Figure 7: Screenshots of the *FineMAV* software's **(A)** progress being printed out on the command line as it is running and the **(B)** log file that was produced at the end.

Northeast Asians, South Asians, Southeast Asians and Oceanians. Computational experiments were run on Ubuntu 16.04 LTS with 3.60 GHz 8-core Intel Core i7-4790 processors with 31.3 GB RAM and 950.6 GB of hard disk memory. The size of the input file, which contains the data extracted from the VCF file, and the VEP-generated file were 2.0 GB and 2.1 GB respectively. Figure 6B illustrates the maximum RAM usage and the time taken when different chunk sizes are utilised. As expected, the larger the chunk size, the faster the run time, up to a certain point. The optimal chunk size would vary depending on the size of the input files and the computing power.

Another option that users can specify is the penalty parameter (x) (Figure 5). If a user does not type in an x value, the software can detect the number of populations. If the number of populations ranges from 2 to 7, default x values are assigned according to Table 4, which are based on published data (Szpak et al., 2018). However, should the user intend to analyse more than 7 populations or decides on another value for x, they are able to change it.

To ensure that the pipeline calculated the *FineMAV* scores correctly, I tested the pipeline using the 1000 Genomes Project African, East Asian, and European continental populations and obtained *FineMAV* scores that were highly correlated with the published data (Spearman's correlation was 0.9999 for all three continental populations) (Szpak et al., 2018). When comparing the top 100 *FineMAV* outliers across all three continental populations of the published data, only 5/300 variants did not overlap with the published results and all five of these variants were missing from the 1000 Genomes Project dataset, because they were not biallelic SNPs and the sex chromosomes were filtered differently.

To reiterate what was mentioned previously in Chapter 2.5, the software outputs three kinds of file: the log file (Figure 7B), a file containing the genome-wide *FineMAV* scores along with the intermediate calculations and the bigWig files. bigWig files (Kent et al., 2010) allow users to visualise the genome-wide *FineMAV* scores on genome browsers whether they be web-based, such as the UCSC Genome Browser (Kent et al., 2002), or a downloadable browser such as the Integrative Genomics Viewer (Robinson et al., 2011). An example of what this would look like with the Han Chinese, Singaporean Indian and Singaporean Malay *FineMAV* scores can be seen in Figure 8. I will discuss the genome-wide *FineMAV* scores for these populations in the next

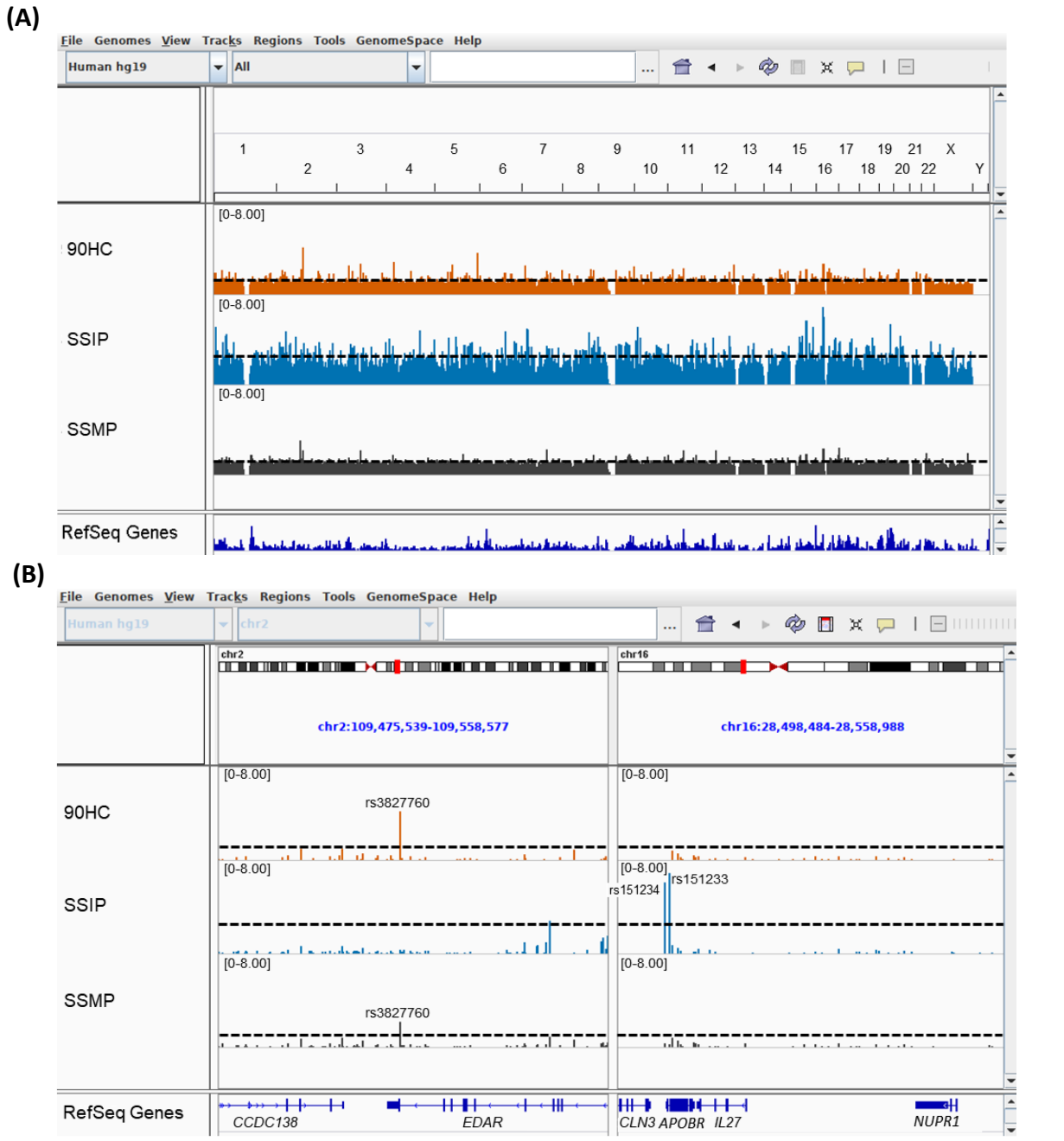


Figure 8: Annotated screenshots of the bigWig files of the genome-wide *FineMAV* scores. **(A)** *FineMAV* scores for Han Chinese (90HC, orange), Singaporean Indian (SSIP, blue) and Singaporean Malay (SSMP, grey) populations displayed on the Integrative Genomics Viewer (IGV).The genomic region on display are the autosomal and the X chromosomes. The dashed horizontal line represents the 99.99th percentile of the *FineMAV* score distribution in each population. **(B)** A multi-locus view of two regions where the left panel displays a locus with a well-known positively selected missense variant in *EDAR* (rs3827760) in East Asians that also stands out in the SSMP population. The right panel displays a novel locus with two high scoring variants in SSIP: rs151233, a synonymous variant in *APOBR* and rs151234, an intronic variant in *CLN3* that stand out in the SSIP.

section, Chapter 3.3. Within the genome browser, users could input a genomic position/region or gene name of their interest in a search bar and visualise its associated *FineMAV* statistic. They can navigate the genome by zooming and scrolling. Users are also able to add annotation tracks which enables users to make useful comparisons. The program is freely available on GitHub (https://github.com/fadilla-wahyudi/finemav), along with the documentation.

3.3 Genome-wide *FineMAV* scores in Chinese and Singaporean datasets

Figure 9 depicts the genome-wide *FineMAV* scores for the three population groups that were analysed and highlights a handful of novel outlier SNPs. The merged dataset consists of 5,774,118 SNPs of which 581 of the derived alleles passed the 99.99th percentile threshold. When comparing the shape of the Manhattan plots, it is apparent that the *FineMAV* scores for Singaporean Indians vary greatly and have more population-specific signals compared to Han Chinese and Singaporean Malays. For comparison, the highest score for Singaporean Indians is 7.68, but for Han Chinese and Singaporean Malays, it is 4.66 and 3.38 respectively (Appendix F). This is because Han Chinese and Singaporean Malays are genetically more closely related and *FineMAV* penalises allele sharing between populations, so as to highlight high frequency population-specific mutations.

Besides replicating known SNPs that were polymorphic in the East and South Asian populations, the study highlighted several interesting SNPs in the Singaporean Malays (Table 7), a population group that is not well-represented in genome-wide selection scans. As this study only included SNPs that were polymorphic in all three populations, there were several, previously reported, strong selection signals that were missed out by this whole genome positive selection scan. For example, selection signals in melanoma-associated gene (*MAGEE2*) and protease serine S1 family member 53 (*PRSS53*), which have been reported to be selected in East and South Asians (Yngvadottir et al., 2009; Szpak et al., 2018; Wu et al., 2019), are not reported here. This is because the positively selected variants in these genes were absent from the VCF files in at least one of the populations and were filtered out. Since access to the alignment (*.bam) files for these population samples was not available, I could not regenerate a jointly-called VCF file to address

this issue. In addition to high-scoring population-specific variants, like the ones found in *MAGEE2* and *PRSS53*, low-scoring variants, like the ones that would have been found in the counterpart

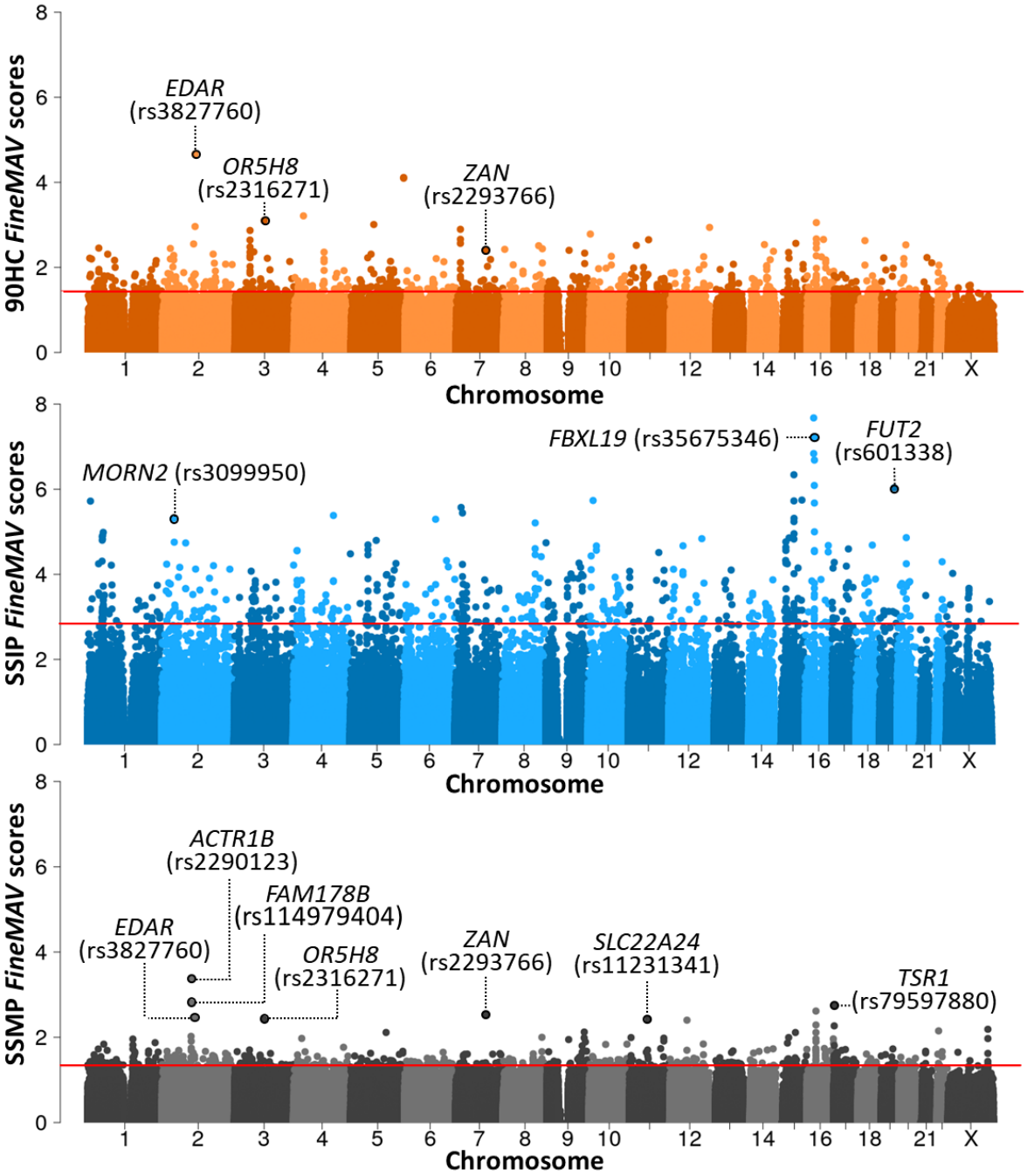


Figure 9: Manhattan plots of the Han Chinese (90HC, orange), Singaporean Indian (SSIP, blue) and Singaporean Malay (SSMP, grey) genome-wide *FineMAV* scores. Each circle on the Manhattan plot signifies a SNP plotted against its GRCh37 genomic coordinates along the x-axis. The red line represents the 99.99th percentile of the *FineMAV* score distribution in each population.

populations, would have been excluded from this analysis too. This would explain the *FineMAV* distribution seen in Figure 10 in which the first bin of the histogram (i.e. the bin containing the lowest *FineMAV* scores) is only approximately 60%. Had the three datasets been jointly-called, I would expect the first bin to be more than 90%, similar to the *FineMAV* distribution of the 1000 Genomes Project (Szpak et al., 2018).

Table 7: Top 10 *FineMAV* hits from the Singaporean Malay dataset (SSMP) with the chromosome (Chr), genomic position (position), the SNP ID according to the dbSNP build 151, most severe variant consequence according to Ensembl and whether it has been detected in previous positive selection scans. The derived allele frequencies (DAF) of the Han Chinese (90HC) and Singaporean Indian (SSIP) dataset are included for comparison.

Chr	Position	SNP ID	Gene	Consequence	DAF 90HC	DAF SSIP	DAF SSMP	*FineMAV*	Known or novel
2	98272491	rs2290123:A>G	*ACTR1B*	3 prime UTR	0.033	0.029	0.380	3.378	Known (Wu et al., 2019)
2	97613974	rs114979404:C>G	*FAM178B*	Intron	0.022	0.029	0.375	2.806	Known (Wu et al., 2019)
17	2238152	rs79597880:T>C	*TSR1*	Missense (p.Lys199Glu)	0.089	0.014	0.297	2.747	Novel
16	31088347	rs749671:G>A	*ZNF646*	Synonymous (p.Glu234=)	0.906	0.043	0.776	2.616	Known (Wu et al., 2019)
7	100371358	rs2293766:G>A	*ZAN*	Stop gained (p.Trp1883Ter)	0.528	0.257	0.557	2.531	Known (Szpak et al., 2018)
2	109513601	rs3827760:A>G	*EDAR*	Missense (p.Val370Ala)	0.922	0.029	0.490	2.474	Known (Sabeti et al., 2007; Szpak et al., 2018; Wu et al., 2019)
3	98031307	rs2316271:T>A	*OR5H8*	Stop gained (p.Leu184Ter)	0.767	0.314	0.599	2.424	Novel
11	62848487	rs11231341:A>C	*SLC22A24*	Stop gained (p.Tyr501Ter)	0.867	0.757	0.792	2.421	Novel
12	57865558	rs2229300:G>T	*GLI1*	Missense (p.Gly1012Val)	0.050	0.014	0.224	2.402	Novel
16	31075175	rs2303223:G>A	*ZNF668*	Synonymous (p.Gly225=)	0.911	0.043	0.781	2.290	Novel

As known positively selected SNPs were missing in this analysis, I expected that other high frequency functional SNPs within its vicinity would also generate high *FineMAV* scores because of the effect of genetic hitchhiking. In the Manhattan plots (Figure 9), I noticed a high-

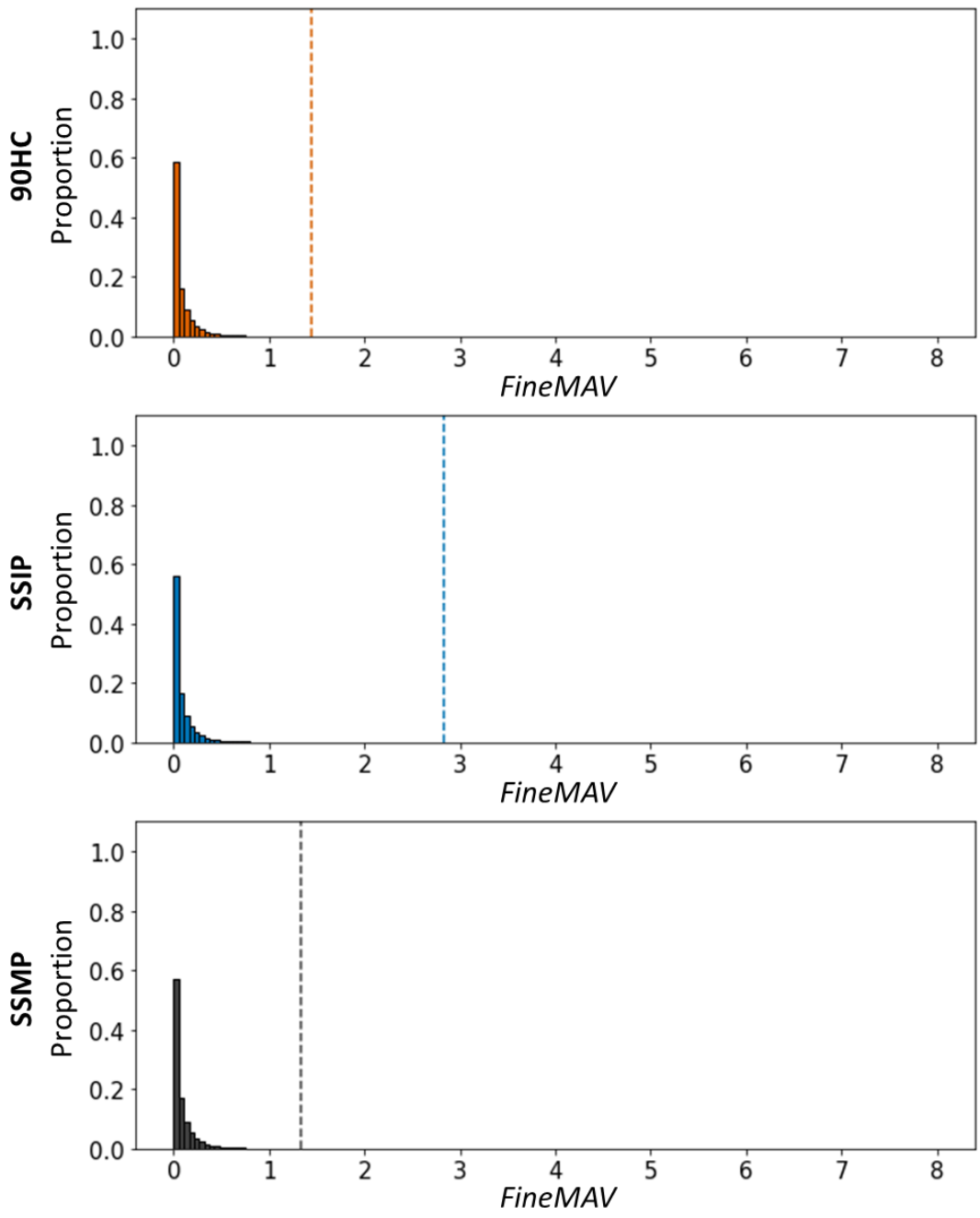


Figure 10: The frequency distribution of the genome-wide *FineMAV* scores for the Han Chinese (90HC), Singaporean Indian (SSIP) and Singaporean Malay (SSMP) populations. The vertical dashed line represents the 99.99th percentile.

scoring locus in chromosome 16 in all three populations (Figure 9). I suspected that *PRSS53* rs11150606 and rs201075024, that are known to be positively selected in East and South Asians, respectively (Szpak et al., 2018; Wu et al., 2019) may be responsible for this. To see if the *PRSS53* variants are in linkage disequilibrium (LD) with the high-scoring *FineMAV* variants, I performed pairwise comparisons of LD between them and other chromosome 16 outlier variants that are listed in the top 50 genome-wide *FineMAV* hits and found that they are in LD (average r^2 value for both SNPs is 0.32) with each other, suggesting that the other high-scoring loci may be neutral and tagging the *PRSS53* rs11150606 and rs201075024 variants (Figure 11). On the other hand, rs1343879 (*MAGEE2*), which is selected in East Asians (Yngvadottir et al., 2009; Szpak et al., 2018), did not produce any other nearby high-scoring locus in 90HC.

As the Han Chinese and Singaporean Malays are genetically related to each other, there were some SNPs that were positively selected in both. Examples of this are the derived alleles of rs3827760 in ectodysplasin A receptor (*EDAR*), rs2293766 in zonadhesin (*ZAN*) and rs2316271 in the olfactory receptor family 5 subfamily H member 8 (*OR5H8*) (Figure 9), in which the first two are established positively selected SNP that have been highlighted in several genomic scans for selection in East Asian populations (Sabeti et al., 2007; Szpak et al., 2018). Studies that have looked at the missense variant rs3827760 in *EDAR* have confirmed its pleiotropic effects. The non-synonymous mutation was found to be associated with hair thickness (Fujimoto et al., 2007; Fujimoto et al., 2008; Kamberov et al., 2013), shovel-shaped incisors (Kimura et al., 2009; Park et al., 2012a; Tan et al., 2014), ear morphology (Adhikari et al., 2015; Shaffer et al., 2017), increased density of eccrine sweat glands, reduced mammary fat pad and increased mammary ductal gland branching (Kamberov et al., 2013). Despite extensive studies, researchers still remain uncertain as to why this allele is positively selected. Some have theorised that increased sweat gland density results in better thermoregulation during warmer climates (Kamberov et al., 2013). Others hypothesise that male sexual preference may have played a role in its selection (Kamberov et al., 2013) while some have suggested that selection for greater mammary gland branching would lead to better mother-to-child nutrient transfer, especially for vitamin D, to prevent vitamin D deficiency in regions with lower ultraviolet (UV) levels (Hlusko et al., 2018).

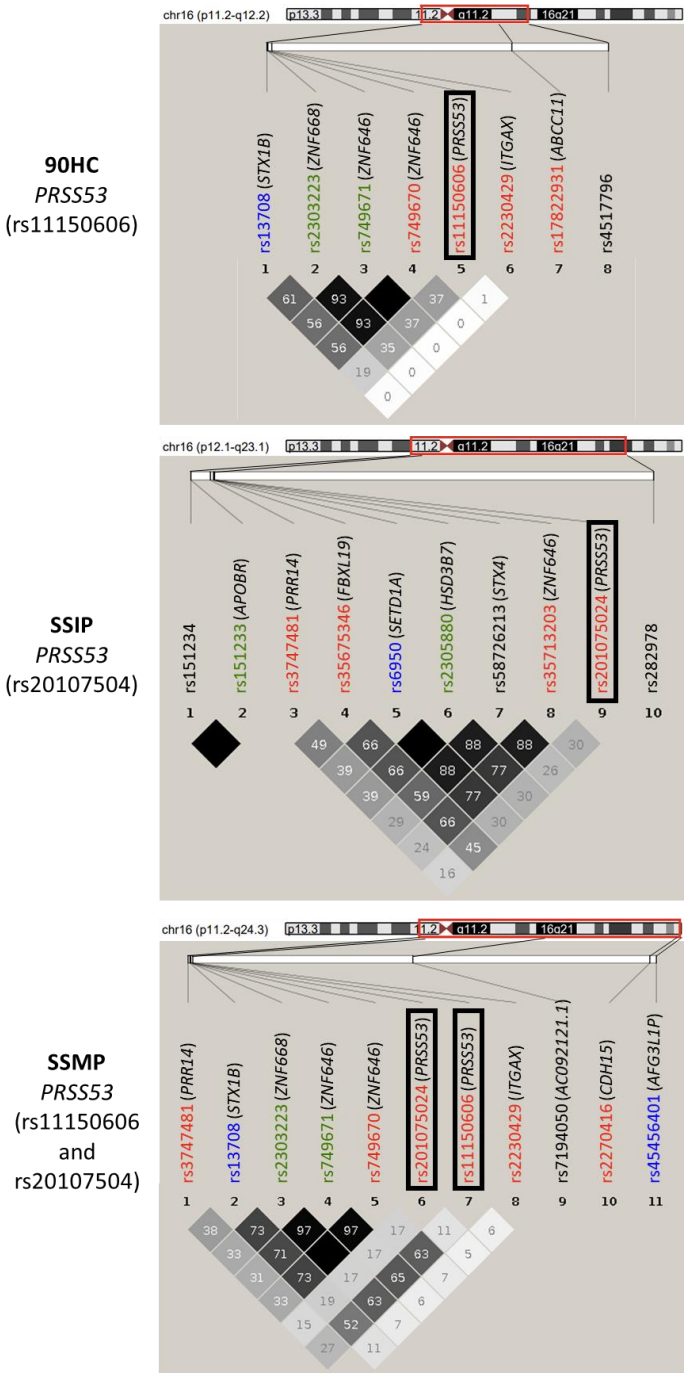


Figure 11: LD plots for the Han Chinese (90HC), Singaporean Indian (SSIP) and Singaporean Malay (SSMP) populations. The chromosome ideogram is presented above each plot with the red box indicating the area of interest with the SNP locations indicated on the white bar below. SNPs include protein-altering (*red*), synonymous (*green*), untranslated (*blue*) and non-transcribed variants (*black*). Pairwise plots of LD between known positively selected variants, *PRSS53*'s rs1150606 and rs201075024 in East and South Asians respectively (black rectangles), and the top 50 *FineMAV* SNPs from the same chromosome in each population. The colour of the squares signify the strength of LD (r^2 values) between a pair of SNPs where the darker the colour, the stronger the LD. The r^2 values, which are multiplied by 100, are shown in the squares. Figures were generated with Haploview (Barrett et al., 2005).

Another previously known signal that was replicated in East Asians occurred in *ZAN*, a gene that encodes an acrosomal protein in the sperm called zonadhesin. A study employing *Zan* knockout mice found that their sperms remained fertile and had increased adhesion to the jelly-like coating of the egg (zona pellucida) of other species like pig, cow and rabbit (Tardif et al., 2010). As *ZAN* is responsible for species-specific binding, it has been speculated that a truncation, as a result of this nonsense mutation (rs2293766), could have mediated interbreeding between archaic humans and modern humans in Asia (Skoglund and Jakobsson, 2011).

A novel signal that was picked up in both Han Chinese and Singaporean Malay populations was the nonsense mutation (rs2316271) in olfactory receptor family 5 subfamily H member 8 (*OR5H8*) (Figure 9), which is a pseudogene. There are studies that have shown that pseudogenes, when transcribed, can play a role in regulating gene expression (Rajkumar and Mark, 2008; An et al., 2017). However, in this instance, it is possible that the nonsense mutation has no phenotypic impact as the expression of *OR5H8* ranges from low to negligible in various tissues (Flegel et al., 2013; Papatheodorou et al., 2020).

Singaporean Indians have more population-specific signals as their population is genetically distinct in comparison with the Han Chinese and Singaporean Malays. Population-specific variants identified via *FineMAV* in Singapore Indians includes two missense variants; rs35675346 in F-box and leucine rich repeat protein 19 (*FBXL19*) and rs3099950 in MORN Repeat Containing 2 (*MORN2*) (Figure 9), in which the former is in moderate LD with rs201075024 (*PRSS53*). *FBXL19* has been linked to psoriasis susceptibility (Philip et al., 2010) and is associated with paradoxical adverse reactions to anti-tumour necrosis factor α (TNFα) drugs which are used to treat a specific type of psoriasis called plaque psoriasis (Cabaleiro et al., 2016). rs3099950 in *MORN2* was detected in a genome-wide association study (GWAS) of chronic peritonitis (Offenbacher et al., 2016) where it was specifically associated with *Porphyromonas gingivalis*-related inflammation. It cannot be verified as to whether South Asians have greater incidences of periodontal disease as epidemiological studies reported by the World Health Organization (WHO) are inconsistent. (World Health Organization, 2005). Furthermore, it is difficult to conclude whether incidences stem from genetic factors or other risk factors like chewing betel leaves, tobacco smoking or diabetes mellitus (Van Dyke and Dave, 2005).

A well-known stop-gain variant, rs601338 in fucosyltransferase 2 (*FUT2*), that is common in African and European populations (allele frequency is 0.49 and 0.44 respectively (The 1000 Genomes Project Consortium, 2015)) was also picked up the in Singaporean Indians. This was expected because the derived allele for this variant is rarely found in East Asians (The 1000 Genomes Project Consortium, 2015). The FUT2 enzyme is responsible for the secretion of ABO histo-blood group antigens and their expression in gastrointestinal tissues (Kelly et al., 1995). Individuals that are homozygous for the nonsense mutation are known as non-secretors (Kelly et al., 1995). Studies have shown that nonsense mutations in non-secretors confer protective effects from enteric pathogens such as rotavirus (Imbert-Marcille et al., 2014), norovirus (Thorven et al., 2005) and *Helicobacter pylori* (Ikehara et al., 2001), but increases risk of other diseases like Crohn's disease (McGovern et al., 2010) and type I diabetes (Smyth et al., 2011). Some of these findings were corroborated with knockout mice studies (Magalhães et al., 2009; Tong et al., 2014). Gut modifications were seen from both the host and the gut microbiota. For example, one study observed changes in the gastric mucosa that hindered *H. pylori* adhesion (Magalhães et al., 2009). Another study, using gut microbial metagenomics on humans and mice, revealed certain pathways being enriched, like carbohydrate and lipid metabolism, and depleted, like amino acid-related biosynthesis (Tong et al., 2014).

A genome-wide positive selection scan performed on the SG10K dataset, which comprises whole genome sequences from the Chinese, Malay and Indian populations in Singapore, identified several of the same loci that are top *FineMAV* hits from the Singaporean Malay dataset (SSMP). In the SG10K dataset, only one genomic region (chr2:97,477,374 – 98,332,858) was specific to Malays. The top two *FineMAV* hits in the Singaporean Malay dataset (SSMP), which are the derived alleles in rs2290123 in the 3' untranslated region (3'-UTR) of actin related protein 1B (*ACTR1B*) and rs114979404 in the intron of family with sequence similarity 178 member B (*FAM178B*), fall in this locus. According to the Genotype-Tissue Expression (GTEx) project, the derived allele in *FAM178B* (rs114979404) has been associated with a statistically significant increase in fumarylacetoacetate hydrolase domain containing 2C pseudogene (*FAHD2CP*) expression in the atrial appendage (The GTEx Consortium, 2017), which is a nearby gene within 1,000 kb of *FAM178B*. Interestingly, the *GPAT2-FAHD2CP* locus has been reported to be

associated with diastolic blood pressure (Warren et al., 2017). Both *ACTR1B* and *FAM178B* could be responsible for brain function. *ACTR1B* encodes a subunit in dynactin, which is a protein complex that plays a role in cell division and intracellular transport (Eckley et al., 1999). SNPs in *ACTR1B* have also been picked up in other genome wide association studies (GWAS) in which these SNPs were associated with alcohol consumption and smoking behaviour (Karlsson Linner et al., 2019; Liu et al., 2019). *FAM178B*, on the other hand, is found in a genetic locus that has an effect on both schizophrenia susceptibility and lithium treatment response for patients with bipolar affective disorder (Amare et al., 2018).

Examples of novel SNPs that were solely picked up in the Singaporean Malays are the missense variants rs2229300 in glioma-associated oncogene family zinc finger 1 (*GLI1*) and rs79597880 in pre-rRNA-processing protein TSR1 homolog (*TSR1*) (Figure 9). *GLI* is a well-established oncogene and its protein is a drug target for several anti-cancer medication (Palle et al., 2015). According to the Catalogue of Somatic Mutations in Cancer (COSMIC), 65.60% of mutations that are observed in *GLI1* are missense substitutions (Tate et al., 2019). However, there have not been any reports on rs2229300 (Tate et al., 2019) that is present at high frequency in Southeast Asians. With regards to *TSR1*, it was recently reported that rare (minor allele frequency < 1%) missense mutations of this gene may be associated with spontaneous coronary artery dissection (SCAD), a condition where the coronary artery tears resulting in two lumens: the true lumen and the false one (Sun et al., 2019b). *TSR1*, whose exact function is yet to be elucidated, plays a role in ribosome maturation (Urszula et al., 2016). Interestingly, the missense mutations they reported were all substitutions from arginine, a positively charged amino acid, to a neutral amino acid (Sun et al., 2019b). The researchers suspect that the positively charged clusters of arginine and lysine at the surface of the protein may be important to its functionality (Sun et al., 2019b). The missense mutation in rs79597880 is coincidently a substitution from a positively charged residue, lysine, to a negatively charged residue, glutamic acid and is predicted to be deleterious. There have yet to be any functional studies to confirm Sun et al.'s (2008) findings.

The stop-gain variant in solute carrier family 22 member 24 (*SLC22A24*; rs11231341) was the eighth highest *FineMAV* outlier in SSMP (Table 7). This mutation is common worldwide (global

derived allele frequency is 0.75) (The 1000 Genomes Project Consortium, 2015) and it should have been penalised by *DAP* but because this variant has a high CADD_PHRED score (47.00) and there are not many population-specific variants in SSMP due to its admixture, it was obtained in the top 10 *FineMAV* hits.

3.4 Genome-wide *FineMAV* scores in the GenomeAsia 100K dataset

The genome-wide *FineMAV* scores for the GenomeAsia 100K populations from Northeast Asia, South Asia, Southeast Asia and Oceania are displayed in Figure 12. The dataset consists of 66,236,516 SNPs of which 6,654 of the derived alleles passed the 99.99^{th} percentile threshold. The Manhattan plot for the Oceanian populations indicate that there are more population-specific signals compared to the other three continental regions, with Southeast Asian populations having the least population-specific signals. In Oceanian populations, the highest-scoring derived allele is 12.55 while the Northeast Asian, South Asian and Southeast Asian populations it is 5.86, 9.81 and 3.03 respectively (Appendix F).

The authors of this dataset performed a series of methods to infer the population structure, including PCA and ADMIXTURE analysis. Oceanian populations are represented by a single Melanesian ancestry group, with a handful of populations having some levels of Southeast Asian ancestry (GenomeAsia100K Consortium, 2019). As the continental region with the least admixture between the four regions, I correctly expected *FineMAV* to generate more high-scoring hits for Oceania, similar to the Singaporean Indian population in the earlier dataset. In contrast, Southeast Asian populations are highly admixed as they share genetic ancestry with the other three continental regions, and, therefore, have fewer population-specific variant outliers. Several Mainland Southeast Asian populations, such as Burmese, Thai, and Vietnamese, carry moderate levels of Northeast Asian ancestry (GenomeAsia100K Consortium, 2019). Indigenous Bruneian and Taiwanese populations as well as some Mainland Southeast Asians share genetic ancestry with many tribal groups living in India (GenomeAsia100K Consortium, 2019). Populations living in Flores, an island in Indonesia, carry varying degrees of Melanesian ancestry (GenomeAsia100K Consortium, 2019).

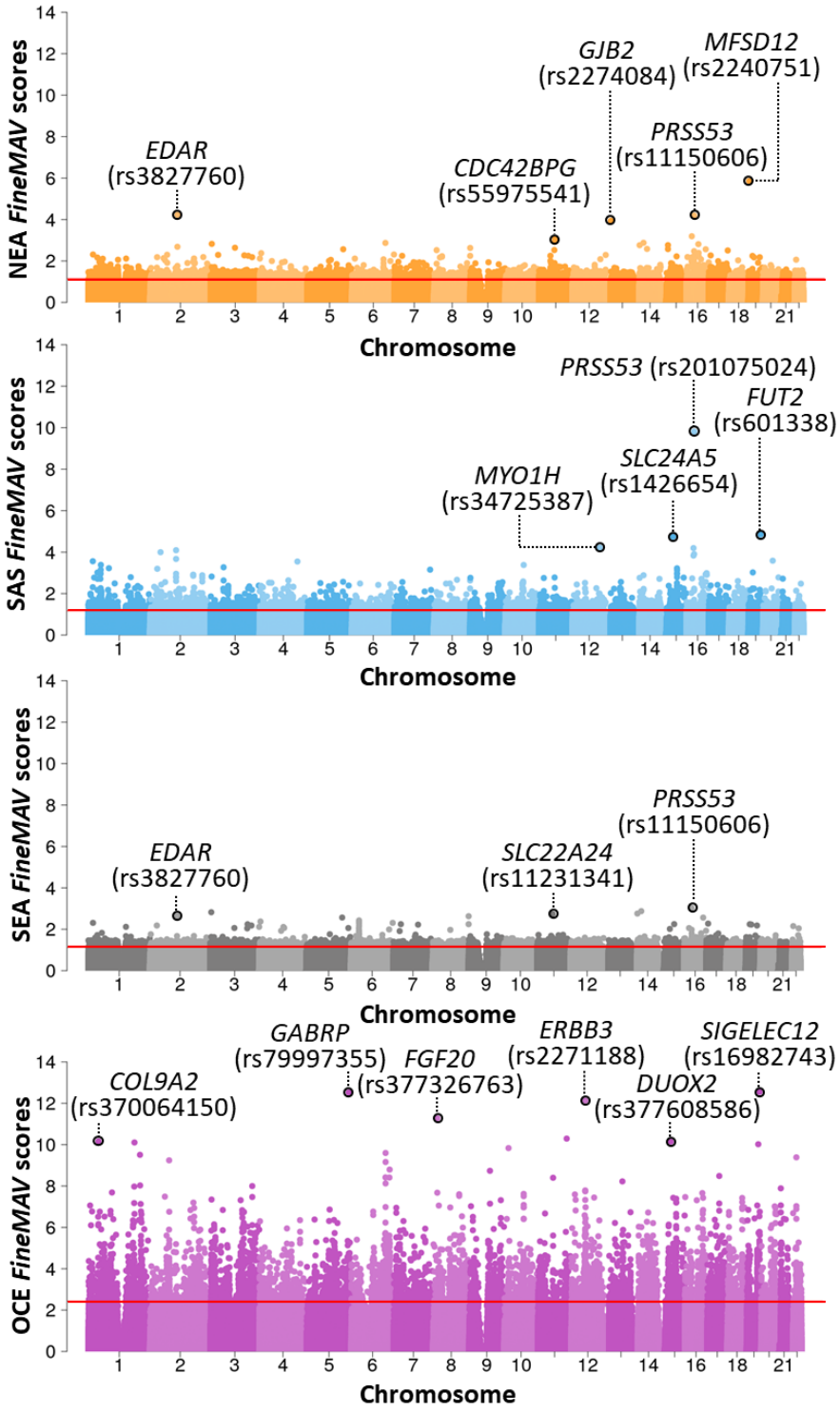


Figure 12: Manhattan plots of the GenomeAsia 100K Northeast Asian (NEA, orange), South Asian (SAS, blue), Southeast Asian (SEA, grey) and Oceanian (OCE, purple) genome-wide FineMAV scores. Each circle on the Manhattan plot signifies a SNP plotted against its GRCh37 genomic coordinates along the x-axis. The red line represents the 99.99th percentile of the *FineMAV* score distribution in each population.

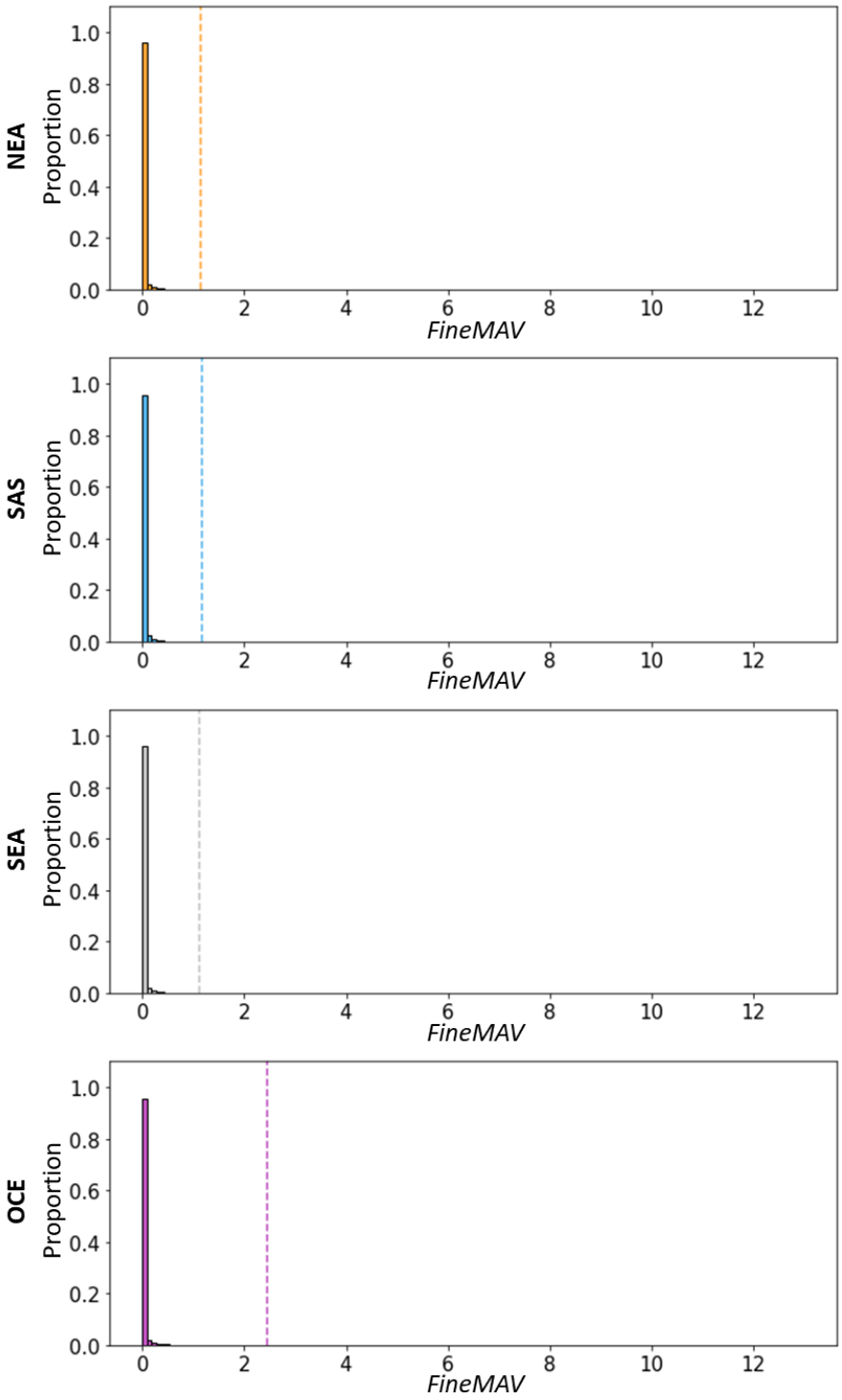


Figure 13: The frequency distribution of the genome-wide *FineMAV* scores for the GenomeAsia 100K Northeast Asian (NEA), South Asian (SAS), Southeast Asian (SEA) and Oceanian (OCE) populations. The vertical dashed line represents the 99.99th percentile.

As opposed to the merged Chinese and Singaporean datasets, in which only SNPs that were polymorphic in all populations were analysed, the VCF files from the GenomeAsia 100K were jointly-called. This is evident in the distribution graph for the genome-wide *FineMAV* scores (Figure 13) where the first bin of the histogram contains more than 90% of the whole genome derived alleles. This distribution is similar to the 1000 Genomes Project in which individuals from each continental population were also jointly-called (Szpak et al., 2018). When comparing between the distribution of *FineMAV* scores for a merged dataset (the 90HC, SSIP and SSMP) and a jointly-called dataset (GenomeAsia 100K and 1000 Genomes Project) (Figure 10 and Figure 13), we can see that there are many SNPs unaccounted for in merged datasets due to the fact that homozygous reference alleles are not called and cannot, therefore, be distinguished from missing SNPs in the analysis. This goes hand-in-hand with my recommendation for users who would like to use the *FineMAV* software, in that I suggest users to use jointly-called data to achieve a more complete scan.

In Northeast Asian populations, *FineMAV* was able to replicate missense variants in three genes that are known to be positively selected: *EDAR* (rs3827760), which was also picked up in 90HC, the major facilitator superfamily domain containing 12 (*MFSD12*; rs2240751) and *PRSS53* (rs11150606) (Figure 12) (Sabeti et al., 2007; Yngvadottir et al., 2009; Adhikari et al., 2016; Szpak et al., 2018; Adhikari et al., 2019; Sun et al., 2019a; Wu et al., 2019).

The rs2240751 (*MFSD12*) variant is the highest-scoring *FineMAV* variant in Northeast Asians. *MFSD12* plays a role in skin pigmentation processing. It encodes a transporter lysosomal protein and is expressed in melanocytes (Crawford et al., 2017; Adhikari et al., 2019). Downregulation of *MFSD12* was observed in melanocytes with darker pigmentation (Crawford et al., 2017) and significantly elevated expression was seen in melanoma tissue (Wei et al., 2019). The variant was first reported to be associated with lighter skin pigmentation in a GWAS in Latin Americans (Adhikari et al., 2019), and has since been associated with tanning ability in Japanese individuals (Shido et al., 2019) and facial pigmented spots in Koreans (Shin et al., 2020). This variant is found in East Asians but not in Europeans suggesting that it may have been positively selected in East Asians after splitting from Europeans, although the estimated selection coefficient is weaker in comparison to other known pigmentation genes (Adhikari et al., 2019). It

was theorised that *MFSD12* may be involved in the convergent evolution of lighter skin pigmentation in East Asians (Adhikari et al., 2019). UV radiation is considered a strong environmental selection pressure as skin colour has been correlated with solar radiation (Jablonski and Chaplin, 2000; Jablonski and Chaplin, 2010). Darker skin pigmentation can offer advantages in regions of higher UV radiation such as protection against skin cancer and prevent photolysis of folate, which could result in infertility (Branda and Eaton, 1978; Jablonski and Chaplin, 2000; Jablonski and Chaplin, 2010; Greaves, 2014). However, at higher latitudes in places with less UV radiation, this would be a disadvantage as melanin would block UV light and could hinder vitamin D biosynthesis, making lighter skin pigmentation a favourable trait (Jablonski and Chaplin, 2000; Jablonski and Chaplin, 2010).

The variant in *PRSS53* (rs11150606) is the third top-scoring *FineMAV* variant in Northeast Asians (Figure 12) and the sixth *FineMAV* hit in 1000 Genomes East Asian population (Szpak et al., 2018). rs11150606 was identified to be associated with hair shape in Latin Americans (Adhikari et al., 2016). Similar to rs3827760 (*EDAR*), the authors suggest that the *PRSS53* variant is likely to have been positively selected in East Asians and may have influenced their scalp hair shape (Adhikari et al., 2016). However, a Han Chinese genome-wide study disagreed with this conclusion and the authors affirm that *EDAR* predominantly affects straight hair in East Asians (Wu et al., 2016). *PRSS53* encodes a serine protease and is expressed in hair follicles. *In vitro* experiments confirm that that the variant, which results in a Q30R substitution, affects the processing and secretion of the protease (Adhikari et al., 2016).

Examples of novel SNPs that were picked up in Northeast Asians are missense mutations in rs2274084 in gap junction beta-2 protein (*GJB2*), which encodes the gap junction protein connexin 26, and rs55975541 in CDC42 binding protein kinase gamma (*CDC42BPG*). Multiple studies have identified rs2274084 (*GJB2*) in Asian and Hispanic patients with hereditary non-syndromic hearing loss, although this polymorphism is considered benign because it is also observed in healthy participants (Girish et al., 2007; Sung-Hee et al., 2008; Tekin et al., 2010; Wei et al., 2013; Zheng et al., 2015). It has been theorised that hearing loss risk increases when an individual carries two *GJB2* mutations: rs2274084, which results in a V27I amino acid substitution, and rs2274083, that results in E114G substitution (Tekin et al., 2010; Choi et al., 2011). There are

conflicting reports from *in vitro* experiments that attempted to study the effects of these two mutations. One report concluded that the double mutant alleles hindered the function of the gap function (Tekin et al., 2010) while the other study indicated that it may not be pathogenic and even suggests that p.V27I may compensate for the loss of hemichannel activity of p.E114G (Choi et al., 2011). A study has also associated rs2274084 (*GJB2*) with Epstein-Barr virus positive nasopharyngeal carcinoma, and concluded that the homozygous derived allele genotype, TT, may be a risk factor (Xiao et al., 2018). The other novel SNP, rs55975541 (*CDC42BPG*) has been associated, in Japanese and Koreans, with elevated serum uric acid levels, an indication of kidney disease (Yamada et al., 2017; Lee et al., 2018; Yasukochi et al., 2018).

The top *FineMAV* candidate in South Asians in this analysis is the missense rs201075024 in *PRSS53*. It was also the highest-scoring and only *FineMAV* variant highlighted in the 1000 Genomes South Asian population (Szpak et al., 2018). This variant lies 10 base pairs away from the aforementioned East Asian-specific *PRSS53* variant, rs11150606. The effects of rs201075024 on the *PRSS53* serine protease is still unknown, although it can be hypothesised that this variant, like the East Asian-specific variant, may influence hair shape in South Asians.

A novel SNP that was picked up in South Asians is a missense rs34725387 in myosin IH (*MYO1H*), which was the fourth highest-scoring *FineMAV* variant in the population. Zebrafish knockdown studies of *MYO1H* orthologs confirm that *MYO1H* is involved in craniofacial skeletal development (Sun et al., 2018). A few polymorphisms have been associated with malocclusion, especially with mandibular prognathism (Tassopoulou-Fishell et al., 2012; Cruz et al., 2017; Sun et al., 2018; Cunha et al., 2019), but rs34725387 was not reported to be causal in any of these studies.

There are several variants that have a high allele frequency in Africans and/or Europeans and generate high *FineMAV* scores in South Asians because of their absence in Asians and Oceanians. Had I included African and/or European populations in the analysis, these variants would be penalized by *FineMAV* because of allele sharing. These variants include the nonsense mutation in *FUT2* (rs601338), which was picked up in Singaporean Indians, and the nonsynonymous mutation in *SLC24A5* (rs1426654). rs1426654 (*SLC24A5*) is a well-known,

positively selected variant that has reached fixation in Europeans and has spread to neighbouring regions such as sub-Saharan Africa, West Eurasia and South Asia (Lamason et al., 2005; Mallick et al., 2013). This mutation was first reported to be responsible for lighter skin pigmentation and was functionally validated in zebrafish (Lamason et al., 2005). Following that, rs1426654 was also found to be associated with iris and hair colour in South Asians and Latin Americans (Edwards et al., 2016; Adhikari et al., 2019; Jonnalagadda et al., 2019).

Southeast Asians generated fewer population-specific signals (Figure 12), due to the extensive genetic admixture in the populations sampled from this region. However, *FineMAV* generated relatively high scores for variants that were common in other continental regions, such as the derived allele in rs11231341 (*SLC22A24*) (Figure 12) (The 1000 Genomes Project Consortium, 2015). rs11231341 was also picked up in SSMP, which is also a highly admixed population. Many of the top 50 variants in Southeast Asians have low *DAP* scores (32/50 have *DAP* scores that are less than 0.1) (Appendix F). The reason why these non-population-specific variants are in the top 50 is because they are highly deleterious and therefore, the CADD_PHRED scores are high (Appendix F). This may be a caveat that users of the *FineMAV* software need to keep in mind. Among the top *FineMAV* variants that produced high *DAP* scores, indicating that they were population-specific, were the missense rs11150606 (*PRSS53*) and rs3827760 (*EDAR*) mutations that were also identified as outliers in Northeast Asians, 90HC and SSMP.

In Oceania, the highest-scoring *FineMAV* variant is the stop-gain variant rs16982743 in sialic acid-binding immunoglobulin-like lectins 12 (*SIGLEC12*). Siglec-12 belongs to a family of transmembrane proteins that regulate immune response, called Siglecs (Varki and Angata, 2006). Siglecs can recognise sialic acids, which is a type of acidic sugar that is important in self-associated molecular patterns or also known as “signature of self” (Pillai et al., 2012; Matthew et al., 2014). Siglec-12 preferentially binds to Neu5Gc, which is a form of sialic acid (Angata et al., 2001; Angata, 2018). However, the gene that encodes the enzyme that forms Neu5Gc, has undergone pseudogenisation in non-human primates (Chou et al., 1998; Chou et al., 2002). In response to the loss of Neu5G on the cell surface membrane, a missense mutation (R122) in *SIGELEC12*, which results in the loss of recognition of Neu5Gc, spread and has even reached fixation in modern humans (Angata et al., 2001; Angata, 2018). It is likely that pathogens may have been the driving

selection pressure as it would offer protection against Neu5Gc-specific pathogens (Angata, 2018; Khan et al., 2020). The average allele frequency in humans for the stop-gain mutation in *SIGELEC12* is 0.18 (The 1000 Genomes Project Consortium, 2015). Oceanians have the highest frequency (0.78) (GenomeAsia100K Consortium, 2019) compared to other continental regions, with Africans having the second highest at 0.37 (The 1000 Genomes Project Consortium, 2015). It may be possible that the stop-gain mutation increased in frequency in Oceanians because of positive selection, although this seems unlikely because the R122C substitution would have already rendered Siglec-XII (Roman numerals are used when it can no longer recognise sialic acid) non-functional. Perhaps the high allele frequency in Oceanians can be attributed to archaic human introgression. The stop-gain mutation occurs in 60% and 50% of Neanderthal and Denisovan genomes, respectively (Khan et al., 2020), and Oceanian populations have greater archaic human admixture, especially with Denisovans, compared to other continental regions (David et al., 2010; Reich et al., 2011; Jacobs et al., 2019; Gokcumen, 2020). This stop-gain variant has been associated with adverse cardiovascular outcomes in hypertensive patients on antihypertension therapy (McDonough et al., 2013).

There are many missense mutations among the top *FineMAV* candidates in Oceania. Examples are rs79997355 in the gamma-aminobutyric acid type A receptor π subunit (*GABRP*), rs2271188 in erb-b2 receptor tyrosine kinase 3 (*ERBB3*), rs377326763 in fibroblast growth factor 20 (*FGF20*), rs370064150 in collagen type IX alpha 2 chain (*COL9A2*) and rs377608586 in dual oxidase 2 (*DUOX2*) (Figure 12). What these variants have in common is that they are present in Oceania and, to a small degree, in Southeast Asians, but are virtually absent in the rest of the world (The 1000 Genomes Project Consortium, 2015; Lek et al., 2016; GenomeAsia100K Consortium, 2019; Karczewski et al., 2020; Phan et al., 2020). Since Oceanian and Southeast Asian populations are underrepresented in genomic datasets, not much is known about these variants. These genes are responsible for a wide range of functions. Both *GABRP* and *ERRB3* have been associated with cancer (Sung et al., 2017; Jiang et al., 2019; Hafeez et al., 2020). With regards to *ERRB3*, conflicting conclusions have been made about whether it is a functional candidate gene in schizophrenia (Kanazawa et al., 2007; Li et al., 2009a). *COL9A2* encodes a component in collagen and mutations in this gene are associated with musculoskeletal disorders like multiple

epiphyseal dysplasia and intervertebral disc disease (Muragaki et al., 1996; Annunen et al., 1999; Seki et al., 2006). *FGF20* is expressed in dopaminergic neurons and has been linked to Parkinson's disease (Itoh and Ohta, 2013). Mutations in *DUOX2* have been linked with hyperthyroidism (Moreno and Visser, 2007; Kizys et al., 2017).

3.5 Comparing the top 50 *FineMAV* variants

I compared the top 50 *FineMAV* outlier variants I obtained from my analysis with each other, and to the published *FineMAV* scores Szpak et al. (2018) generated from the 1000 Genomes East and South Asian populations and found that there was a lack of overlap between the populations (Figure 14). I speculated that this may be the case because of the manner in which these call sets were generated resulting in missing data and that perhaps these high-scoring *FineMAV* variants are in LD with one another. Pairwise comparisons of LD between the top 50 variants from each population were performed (Figure 15 and Figure 16). Pairwise LD tests could not be performed on the GenomeAsia 100K dataset as their VCF files do not contain genotype information for each individual, which is required for LD tests, so I opted to conduct them using the 1000 Genomes Project East and South Asian populations.

Several regions of LD were identified among Northeast/East Asian populations (Figure 15) and among South Asian populations (Figure 16). An LD region located in chromosome 16 was observed in both continental regions (Figure 15B and Figure 16B) and I suspected that the East Asian-specific and South Asian-specific *PRSS53* variants, rs11150606 and rs201075024, respectively, may have genetically 'hitch-hiked' the neighbouring high-scoring SNPs. The LD block seen in chromosome 3 of East/Northeast Asian populations (from the 2nd to 11th SNP), for the most part, span across non-coding SNPs (intronic, upstream and intergenic) (Figure 15A). The 4th SNP (rs2072053) and the 9th SNP (rs2229647) of the block are missense mutations belonging to semaphorin 3F (*SEMA3F*) and interferon related developmental regulator 2 (*IFRD2*) respectively. This region has been picked up in the Singaporean SG10K dataset and hypothesised to be selected in the Chinese population either before or after their split from Malays (Wu et al., 2019). It is part of an introgressed segment that East Asians acquired from Neanderthals (Ding et al., 2014). The authors suspect the hyaluronidase (*HYAL*) genes may be selected in response to UV-

B irradiation (Ding et al., 2014). It is interesting to note that there are no high-scoring GenomeAsia 100K Northeast Asian SNPs in the LD block. This may be because the variant may be selected in the Southern parts of East Asia (e.g. Southern China, Mainland Southeast Asia) and not the Northern parts of Asia (e.g. Mongolia, Russia). In South Asian populations, I observed two regions of high LD in chromosome 15 (from 1st to 3rd SNP and from the 6th to 13th SNP) that also mostly spans across non-coding SNPs (intronic, 5' UTR, downstream and intergenic) (Figure 16A). The 13th SNP (rs61741344) is a synonymous variant located in RNA binding protein with multiple splicing 2 (*RBPMS2*). None of these variants have been identified in genome-wide association or selection studies. This could possibly mean that one of them may be potentially positively selected or that the causal variant may be within the LD block but is missing from the sequencing dataset.

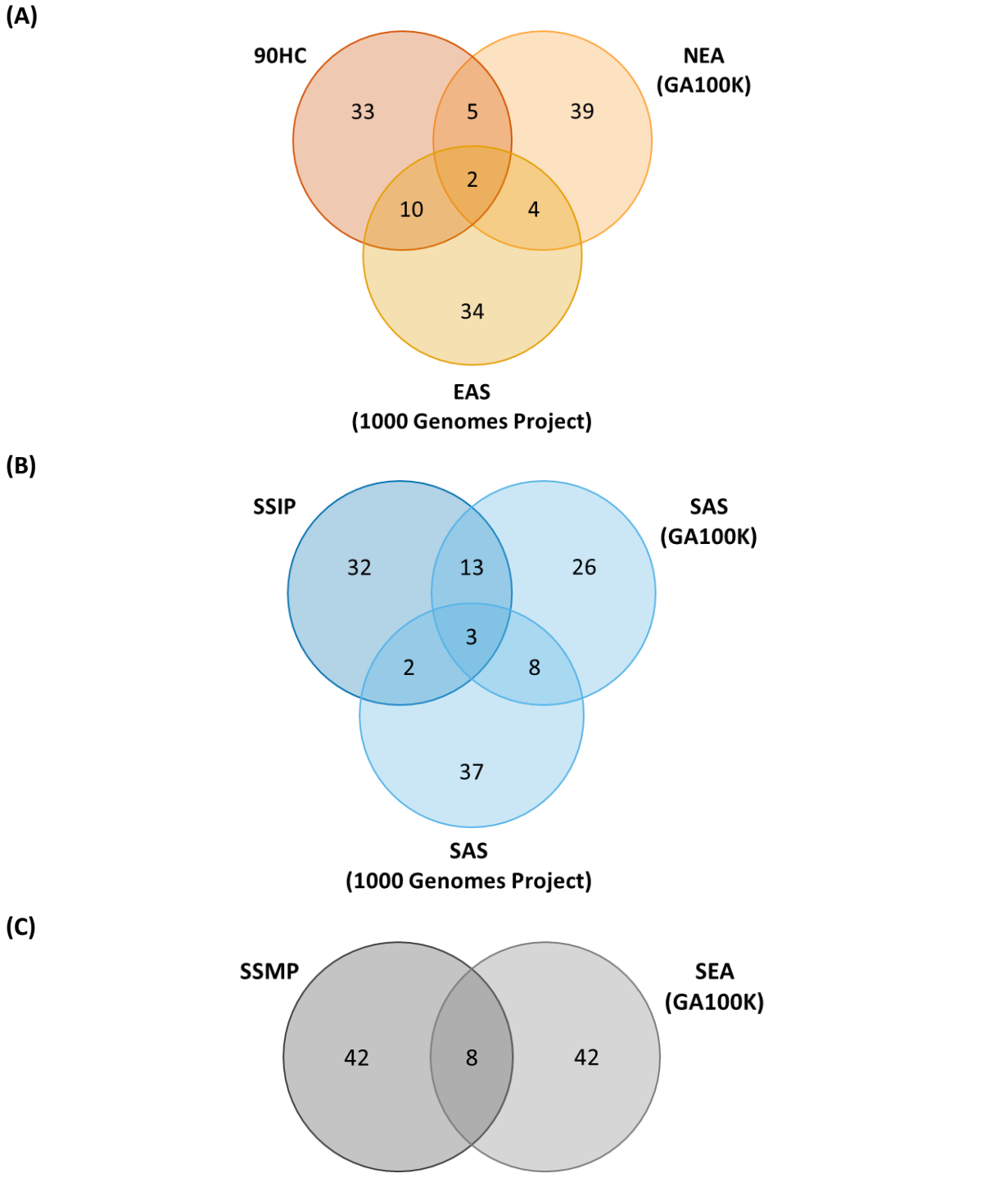


Figure 14: Venn diagram illustrating the overlap in the top 50 *FineMAV* hits between **(A)** the 90 Han Chinese (90HC), Northeast Asians (NEA) from the GenomeAsia 100K dataset (GA100K) and East Asians from the 1000 Genomes Project; **(B)** the Singaporean Indians (SSIP) and South Asians (SAS) from the GA100K and the 1000 Genomes Project and **(C)** Singaporean Malays (SSMP) and Southeast Asians (SEA) from GA100K.

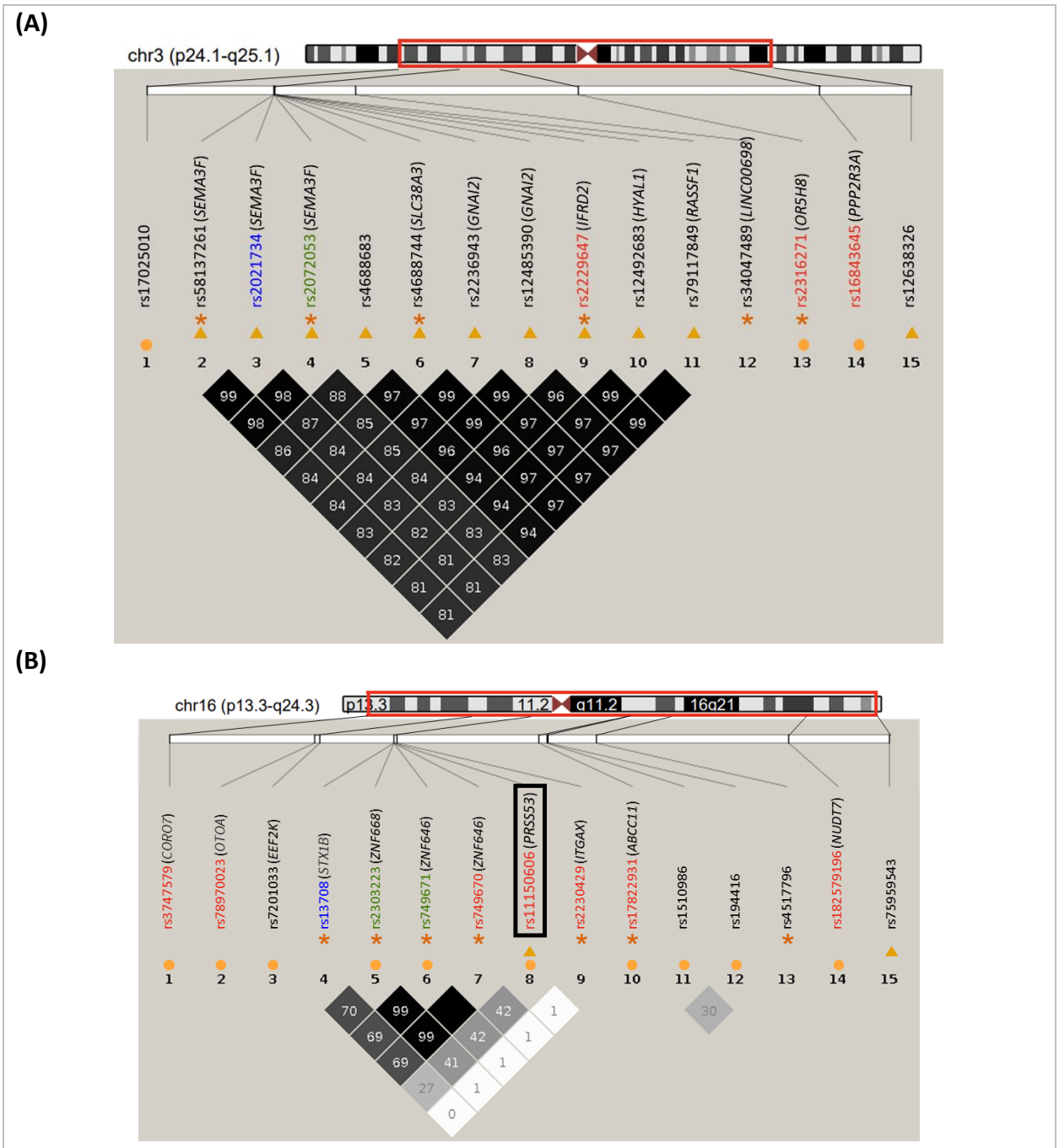


Figure 15: Pairwise LD plots for the top 50 *FineMAV* outliers from the Han Chinese (90HC, asterisk), 1000 Genomes East Asians (EAS, triangle) and the GenomeAsia 100K Northeast Asians (NEA, circle) in **(A)** chromosome 3 and **(B)** chromosome 16. The chromosome ideogram is presented above each plot with the red box indicating the area of interest with the SNP locations indicated on the white bar below. SNPs include protein-altering (*red*), synonymous (*green*), untranslated (*blue*) and non-transcribed variants (*black*). The black rectangle represents the East Asian-specific *PRSS53* variant. The colour of the squares signify the strength of LD (r^2 values) between a pair of SNPs where the darker the colour, the stronger the LD. The r^2 values, which are multiplied by 100, are shown in the squares. Figures were generated with Haploview (Barrett et al., 2005).

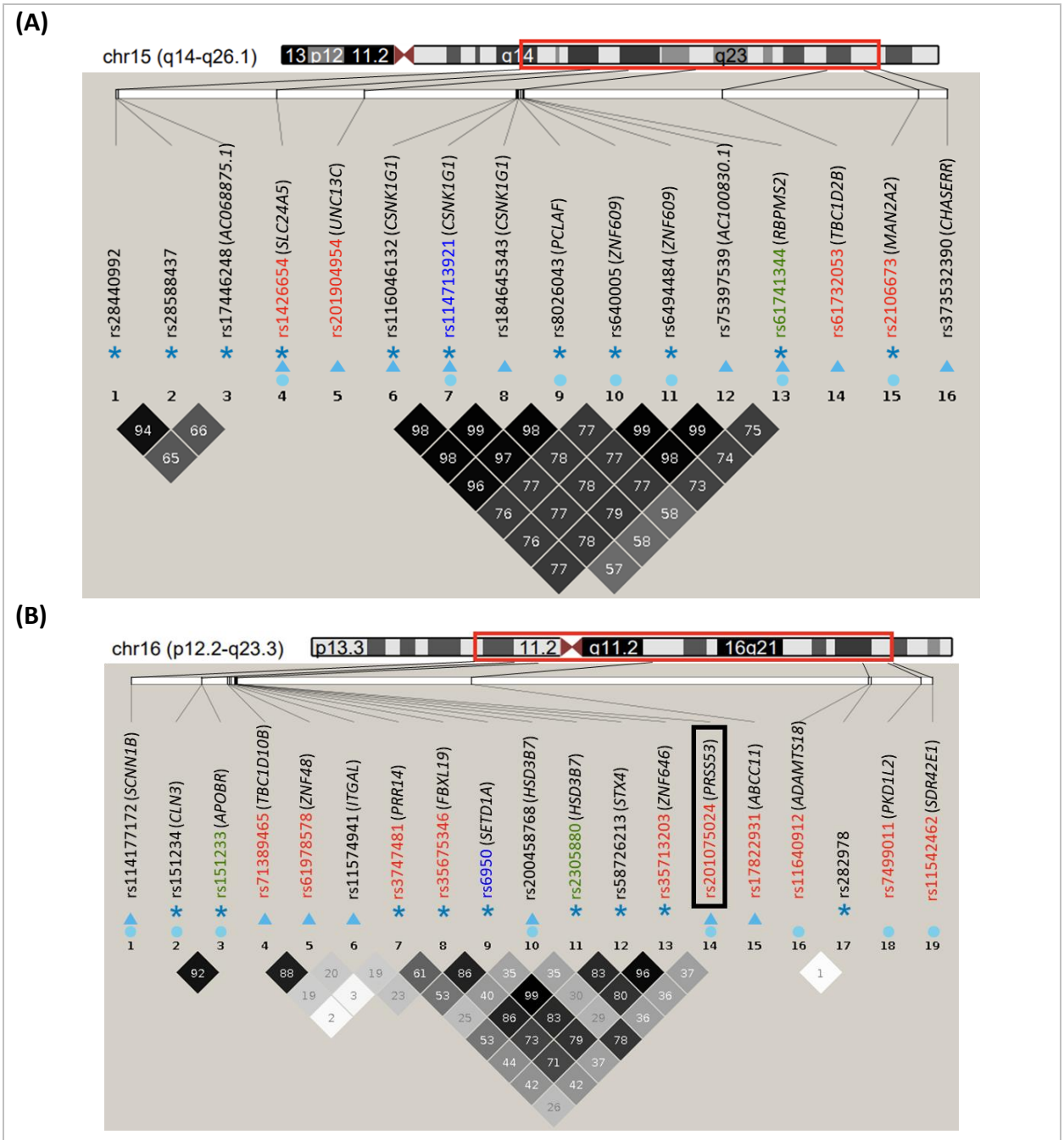


Figure 16: Pairwise LD plots for the top 50 *FineMAV* outliers from the Singaporean Indian (SSIP, asterisk), 1000 Genomes South Asians (SAS, triangle) and the GenomeAsia 100K South Asians (SAS, circle) in **(A)** chromosome 15 and **(B)** chromosome 16. The chromosome ideogram is presented above each plot with the red box indicating the area of interest with the SNP locations indicated on the white bar below. SNPs include protein-altering (*red*), synonymous (*green*), untranslated (*blue*) and non-transcribed variants (*black*). The black rectangle represents the South Asian-specific *PRSS53* variant. The colour of the squares signify the strength of LD (r^2 values) between a pair of SNPs where the darker the colour, the stronger the LD. The r^2 values, which are multiplied by 100, are shown in the squares. Figures were generated with Haploview (Barrett et al., 2005).

4 CONCLUSION

4.1 Summary

There were three objectives that have been addressed in this project. The first objective was to use *FineMAV*, which is a statistical method that was developed to prioritise candidate positively selected variants for functional follow-up. It was used to identify population-specific variants from whole genome sequences obtained from individuals across Southeast Asia. High-coverage whole genome sequences were obtained from Chinese, Indian and Malay groups China and Singapore, as well as larger continental regions like Northeast Asia, South Asia, Southeast Asia, and Oceania. Southeast Asia and Oceania are particularly interesting as they are underrepresented in genome-wide positive selection scans. I replicated well-established selection signals, such as the ones in *EDAR* and *PRSS53*, and found novel SNPs that may be potentially interesting for modelling and functional follow-up.

The second objective was to display the genome-wide *FineMAV* statistics in a human genome browser to enable graphical visualisation and genome annotations. This was achieved by creating bigWig files, containing the *FineMAV* scores, which were uploaded onto genome browsers such as the web-based UCSC Genome Browser (Kent et al., 2002; Navarro Gonzalez et al., 2020) and the downloadable IGV (Robinson et al., 2011).

The third objective was to make *FineMAV* more accessible to researchers by creating a software program which allows researchers to calculate the *FineMAV* statistic for datasets of their interest. This was achieved by creating a software program that exists as a command-line interface and a graphical user interface. The software can output bigWig files which users can use to visualise the genome-wide *FineMAV* scores. It was built to be memory-efficient in anticipation of larger whole genome sequencing datasets.

4.2 Future directions

After performing high-throughput *FineMAV* analysis, the next step would be to select variants for functional validation. A variant that would be useful to model *in vivo* would be the missense rs34725387 in *MYO1H*, which is a novel high-scoring SNP from South Asians. Knockdown experiments using *MYO1H* orthologs in zebrafish has determined that it plays a role in craniofacial skeletal development (Sun et al., 2018) and it would be interesting to investigate the effects of this missense mutation. Additionally, *FineMAV* analysis could be performed on newly released whole genome sequencing datasets such as the Genome Aggregation Database (gnomAD) and the recently sequenced HGDP as well as upcoming sequencing initiatives like the ones proposed in India, France and the United Kingdom to identify additional variants that can be prioritized for modelling by other researchers (Sudlow et al., 2015; Lévy, 2016; Department of Health and Social Care, 2018; Rajagopal, 2019; Bergström et al., 2020; Koch, 2020). It is just the beginning of the quest to understanding human evolutionary adaptations. The availability of millions of deeply phenotyped whole human genomes in the coming decade will provide unique opportunities to functionally validate some of the *FineMAV* outliers identified in this study and add to the growing catalogue of functionally validated variants driving population-specific selection in modern humans.

5 REFERENCES

Adhikari, K., et al. 2016. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. *Nature Communications*, 7, 10815.

Adhikari, K., et al. 2019. A GWAS in Latin Americans highlights the convergent evolution of lighter skin pigmentation in Eurasia. *Nature Communications*, 10.

Adhikari, K., et al. 2015. A genome-wide association study identifies multiple loci for variation in human ear morphology. *Nature Communications*, 6.

Akey, J. M. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Research*, 19, 711-722.

Alexander, D. H., et al. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19, 1655-1664.

Amare, A. T., et al. 2018. Association of polygenic score for schizophrenia and HLA antigen and inflammation genes with response to lithium in bipolar affective disorder: a genome-wide association study. *JAMA psychiatry*, 75, 65-74.

An, Y., et al. 2017, Pseudogenes regulate parental gene expression via ceRNA network.

Angata, T. 2018. Possible influences of endogenous and exogenous ligands on the evolution of human Siglecs. *Frontiers in Immunology*, 9, 2885.

Angata, T., et al. 2001. A second uniquely human mutation affecting sialic acid biology. *Journal of Biological Chemistry*, 276, 40282-40287.

Annunen, S., et al. 1999. An allele of COL9A2 associated with intervertebral disc disease. *Science*, 285, 409-412.

Arciero, E., et al. 2018. Demographic history and genetic adaptation in the Himalayan region inferred from genome-wide SNP genotypes of 49 populations. *Molecular Biology and Evolution*, 35, 1916-1933.

Barrett, J. C., et al. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21, 263-265.

Beaumont, M. A. & Balding, D. J. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, 13, 969-980.

Bergström, A., et al. 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367.

Black IV, W. C., et al. 2001. Population genomics: genome-wide sampling of insect populations. *Annual Review of Entomology*, 46, 441-469.

Branda, R. F. & Eaton, J. W. 1978. Skin color and nutrient photolysis: an evolutionary hypothesis. *Science*, 201, 625-626.

Cabaleiro, T., et al. 2016. Paradoxical psoriasiform reactions to anti-TNFα drugs are associated with genetic polymorphisms in patients with psoriasis. *The Pharmacogenomics Journal*, 16, 336–340.

Cann, H. M., et al. 2002. A human genome diversity cell line panel. *Science*, 296, 261-262.

Chang, C. C., et al. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4.

Cheng, X., et al. 2017. Fast and robust detection of ancestral selective sweeps. *Molecular Ecology*, 26, 6871-6891.

Choi, S.-Y., et al. 2011. Functional evaluation of GJB2 variants in nonsyndromic hearing loss. *Molecular Medicine*, 17, 550-556.
Chou, H.-H., et al. 2002. Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 11736-11741.
Chou, H.-H., et al. 1998. A mutation in human CMP-sialic scid hydroxylase occurred after the Homo-Pan divergence. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 11751-11756.
Cingolani, P., et al. 2012. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics*, 3.
Colonna, V., et al. 2014. Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biology*, 15, R88-R88.
Coop, G., et al. 2010. Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185, 1411-1423.
Copeland, A. J. 1935. The Muruts of North Borneo: Malaria and racial extinction. *The Lancet*, 225, 1233-1239.
Crawford, N. G., et al. 2017. Loci associated with skin pigmentation identified in African populations. *Science*, 358.
Cruz, C. V., et al. 2017. Genetic polymorphisms underlying the skeletal Class III phenotype. *American journal of orthodontics and dentofacial orthopedics*, 151, 700-707.
Cunha, A., et al. 2019. Genetic variants in ACTN3 and MYO1H are associated with sagittal and vertical craniofacial skeletal patterns. *Archives of Oral Biology*, 97, 85-90.
Danecek, P., et al. 2011. The variant call format and VCFtools. *Bioinformatics*, 27, 2156-2158.
Darwin, C. & Wallace, A. 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *Journal of the Proceedings of the Linnean Society of London. Zoology*, 3, 45-62.
David, R., et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468, 1053-1060.
Deng, L., et al. 2015. Dissecting the genetic structure and admixture of four geographical Malay populations. *Scientific Reports*, 5, 14375.
Deng, L., et al. 2014. The population genomic landscape of human genetic structure, admixture history and local adaptation in Peninsular Malaysia. *Human Genetics*, 133, 1169-1185.
Department of Health and Social Care. 2018. Matt Hancock announces ambition to map 5 million genomes.
Ding, Q., et al. 2014. Neanderthal Introgression at Chromosome 3p21.31 Was Under Positive Natural Selection in East Asians. *Molecular Biology and Evolution*, 31, 683-695.
Donald, F. C. & Matthew, E. H. 2007. The population genetics of structural variation. *Nature Genetics*, 39, S30-S36.
Duforet-Frebourg, N., et al. 2014. Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular Biology and Evolution*, 31, 2483-2495.
Eckley, D. M., et al. 1999. Analysis of dynactin subcomplexes reveals a novel actin-related protein associated with the Arpl minifilament pointed end. *The Journal of Cell Biology*, 147, 307-320.

Edwards, M., et al. 2010. Association of the OCA2 polymorphism His615Arg with melanin content in East Asian populations: further evidence of convergent evolution of skin pigmentation. *PLoS Genetics*, 6, e1000867-e1000867.
Edwards, M., et al. 2016. Iris pigmentation as a quantitative trait: variation in populations of European, East Asian and South Asian ancestry and association with candidate gene polymorphisms. *Pigment Cell & Melanoma Research*, 29, 141-162.
Ferrer-Admetlla, A., et al. 2014. On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. *Molecular Biology and Evolution*, 31, 1275.
Ferrer-Admetlla, A., et al. 2009. A natural history of FUT2 polymorphism in humans. *Molecular Biology and Evolution*, 26, 1993-2003.
Fisher, R. 1930, *The genetical theory of natural selection*, The Clarendon Press, Oxford.
Flegel, C., et al. 2013. Expression Profile of Ectopic Olfactory Receptors Determined by Deep Sequencing.(Research Article). *PLoS ONE*, 8, e55368.
Frank, J., et al. 2012. Genome-wide significant association between alcohol dependence and a variant in the ADH gene cluster. *Addiction Biology*, 17, 171-180.
Fu, Y. X. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, 147, 915–925.
Fu, Y. X. & Li, W. H. 1993. Statistical tests of neutrality of mutations. *Genetics*, 133, 693-709.
Fujimoto, A., et al. 2007. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Human Molecular Genetics*, 17, 835-843.
Fujimoto, A., et al. 2008. A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. *Human Genetics*, 124, 179-185.
Garud, N., et al. 2015. Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genetics*, 11, e1005004.
GenomeAsia100K Consortium 2019. The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*, 576, 106-111.
Girish, V. P., et al. 2007. A multicenter study of the frequency and distribution of GJB2 and GJB6 mutations in a large North American cohort. *Genetics in Medicine*, 9, 413-426.
Gokcumen, O. 2020. Archaic hominin introgression into modern human genomes. *American Journal of Physical Anthropology*, 171, 60-73.
Greaves, M. 2014. Was skin cancer a selective force for black pigmentation in early hominin evolution? *Proceedings of the Royal Society B: Biological Sciences*, 281, 20132955.
Grossman, S. R., et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, 327, 883-886.
Guo, R. 2017, *China ethnic statistical yearbook 2016*, Palgrave Macmillan.
Hackinger, S., et al. 2016. Wide distribution and altitude correlation of an archaic high-altitude-adaptive EPAS1 haplotype in the Himalayas. *Human Genetics*, 135, 393-402.
Hafeez, U., et al. 2020. New insights into ErbB3 function and therapeutic targeting in cancer. *Expert Review of Anticancer Therapy*, 1-18.
Halim-Fikri, H., et al. 2015. The first Malay database toward the ethnic-specific target molecular variation. *BMC Research Notes*, 8, 176.

Hamblin, M. T. & Di Rienzo, A. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *The American Journal of Human Genetics*, 66, 1669-1679.
Han, Y., et al. 2007. Evidence of positive selection on a class I ADH locus. *American Journal of Human Genetics*, 80, 441-456.
Hanchard, N. A., et al. 2006. Screening for recently selected alleles by analysis of human haplotype similarity. *The American Journal of Human Genetics*, 78, 153-159.
Hancock, A. M., et al. 2008. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics*, 4, e32-e32.
Hästbacka, J., et al. 1994. The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell*, 78, 1073-1087.
Hatin, W. I., et al. 2014. A genome wide pattern of population structure and admixture in peninsular Malaysia Malays. *The HUGO Journal*, 8, 5.
Hatin, W. I., et al. 2011. Population Genetic Structure of Peninsular Malaysia Malay Sub-Ethnic Groups. *PLOS ONE*, 6, e18312.
Hernandez, R. D., et al. 2011. Classic selective sweeps were rare in recent human evolution. *Science*, 331, 920-924.
Hider, J. L., et al. 2013. Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. *BMC Evolutionary Biology*, 13.
Hinds, D. A., et al. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science*, 307, 1072-1079.
Hlusko, L., et al. 2018. Environmental selection during the last ice age on the mother-to-infant transmission of vitamin D and fatty acids through breast milk. *Proceedings of the National Academy of Sciences of the United States of America*, 115, E4426.
Hoh, B.-P., et al. 2020. Shared Signature of Recent Positive Selection on the TSBP1 - BTNL2 - HLA-DRA Genes in Five Native Populations from North Borneo. *Genome Biology and Evolution*.
Huerta-Sánchez, E., et al. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512, 194-197.
Ikehara, Y., et al. 2001. Polymorphisms of two fucosyltransferase genes (Lewis and Secretor genes) involving type I Lewis antigens are associated with the presence of anti-Helicobacter pylori IgG antibody. *Cancer Epidemiology, Biomarkers & Prevention*, 10, 971-977.
Ilardo, M. A., et al. 2018. Physiological and Genetic Adaptations to Diving in Sea Nomads. *Cell*, 173, 569-580.e515.
Imbert-Marcille, B.-M., et al. 2014. A FUT2 gene common polymorphism determines resistance to rotavirus A of the P8 genotype. *The Journal of infectious diseases*, 209, 1227-1230.
Ismail, E., et al. 2013. Peninsular Malaysia's Negrito Orang Asli and Its Theory of African Origin. *Sains Malaysiana*, 42, 921–926.
Itoh, N. & Ohta, H. 2013. Roles of FGF20 in dopaminergic neurons and Parkinson's disease. *Frontiers in Molecular Neuroscience*, 6.
Izagirre, N., et al. 2006. A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Molecular Biology and Evolution*, 23, 1697-1706.

Jablonski, N. & Chaplin, G. 2010. Human skin pigmentation as an adaptation to UV radiation. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 8962-8968.
Jablonski, N. G. & Chaplin, G. 2000. The evolution of human skin coloration. *Journal of Human Evolution*, 39, 57-106.
Jacobs, G. S., et al. 2019. Multiple deeply divergent Denisovan ancestries in Papuans. *Cell*, 177, 1010-1021.e1032.
Jeong, C. & Di Rienzo, A. 2014. Adaptations to local environments in modern human populations. *Current Opinion in Genetics and Development*, 29, 1-8.
Jiang, S.-H., et al. 2019. GABRP regulates chemokine signalling, macrophage recruitment and tumour progression in pancreatic cancer through tuning KCNN4-mediated Ca2+ signalling in a GABA-independent manner. *Gut*, 68, 1994.
Jobling, M. A., et al. 2014, *Human evolutionary genetics*, Garland Science, New York.
Jonnalagadda, M., et al. 2019. A genome-wide association study of skin and iris pigmentation among individuals of South Asian ancestry. *Genome biology and evolution*, 11, 1066-1076.
Kamberov, G. K., et al. 2013. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell*, 152, 691-702.
Kanazawa, T., et al. 2007. Schizophrenia is not associated with the functional candidate gene ERBB3: results from a case-control study. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 144B, 113-116.
Kang, H. M., et al. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42, 348-354.
Karczewski, K. J., et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581, 434-443.
Karlsson Linner, R., et al. 2019. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics*, 51, 245-257.
Kelly, R. J., et al. 1995. Sequence and expression of a candidate for the human Secretor blood group alpha(1,2)fucosyltransferase gene (FUT2). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. *The Journal of Biological Chemistry*, 270, 4640-4649.
Kent, W. J., et al. 2002. The human genome browser at UCSC. *Genome Research*, 12, 996-1006.
Kent, W. J., et al. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26, 2204–2207.
Khan, N., et al. 2020. Multiple genomic events altering hominin SIGLEC biology and innate immunity predated the common ancestor of humans and archaic hominins. *Genome Biology and Evolution*, 12, 1040-1050.
Kim, Y. & Nielsen, R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167, 1513-1524.
Kim, Y. & Stephan, W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160, 765-777.
Kimura, M. 1991. The neutral theory of molecular evolution: a review of recent evidence. *The Japanese Journal of Genetics*, 66, 367-386.

Kimura, R., et al. 2009. A common variation in EDAR is a genetic determinant of shovel-shaped incisors. *American Journal of Human Genetics*, 85, 528.
Kircher, M., et al. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46, 310-315.
Kizys, M. M. L., et al. 2017. DUOX2 mutations are associated with congenital hypothyroidism with ectopic thyroid gland. *The Journal of Clinical Endocrinology & Metabolism*, 102, 4060-4071.
Koch, L. 2020. Exploring human genomic diversity with gnomAD. *Nature reviews. Genetics*, 21, 448.
Kvikstad, E. & Duret, L. 2014. Strong Heterogeneity in Mutation Rate Causes Misleading Hallmarks of Natural Selection on Indel Mutations in the Human Genome. *Molecular Biology and Evolution*, 31, 23-36.
Lachance, J. & Tishkoff, S. A. 2013. SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *Annals of Medicine and Surgery*, 35, 780-786.
Lai, D., et al. 2019. Genome-wide association studies of alcohol dependence, DSM-IV criterion count and individual criteria. *Genes, Brain and Behaviour*, 18, e12579.
Lamason, R. L., et al. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, 310, 1782-1786.
Lan, T., et al. 2017. Deep whole-genome sequencing of 90 Han Chinese genomes. *GigaScience*, 6, 1-7.
Lao, O., et al. 2007. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Annals of Human Genetics*, 71, 354-369.
Lee, J., et al. 2018. Genome-wide association analysis identifies multiple loci associated with kidney disease-related traits in Korean populations. *PLoS One*, 13, e0194044.
Lek, M., et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536, 285-291.
Lévy, Y. 2016. Genomic medicine 2025: France in the race for precision medicine. *The Lancet*, 388, 2872.
Li, D., et al. 2009a. Case–control study of association between the functional candidate gene ERBB3 and schizophrenia in Caucasian population. *The World Journal of Biological Psychiatry*, 10, 595-598.
Li, H. 2011. A New Test for Detecting Recent Positive Selection that is Free from the Confounding Impacts of Demography. *Molecular Biology and Evolution*, 28, 365-375.
Li, H., et al. 2009b. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, 2078-2079.
Li, J. Z., et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319, 1100-1104.
Linnér, K. R., et al. 2019. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics*, 51, 245-257.
Liu, M., et al. 2019. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use.(Report). *Nature Genetics*, 51, 237.

Liu, X., et al. 2017. Characterising private and shared signatures of positive selection in 37 Asian populations. *European Journal of Human Genetics*, 25, 499-508.
Liu, X., et al. 2013. Detecting and characterizing genomic signatures of positive selection in global populations. *American Journal of Human Genetics*, 92, 866-881.
Liu, X., et al. 2014. Evaluating the possibility of detecting evidence of positive selection across Asia with sparse genotype data from the HUGO Pan-Asian SNP Consortium. *BMC Genomics*, 15, 332-332.
Liu, X., et al. 2015. Differential positive selection of malaria resistance genes in three indigenous populations of Peninsular Malaysia. *Human Genetics*, 134, 375-392.
Magalhães, A., et al. 2009. Fut2-null mice display an altered glycosylation profile and impaired BabA-mediated Helicobacter pylori adhesion to gastric mucosa. *Glycobiology*, 19, 1525-1536.
Mallick, C. B., et al. 2013. The light skin sllele of SLC24A5 in South Asians and Europeans shares identity by descent. *PLoS Genetics*, 9, e1003912.
Mathews, M. 2018, *The Singapore Ethnic Mosaic: Many Cultures, One People*, vol. 2018, World Scientific Publishing, Singapore.
Matthew, S. M., et al. 2014. Siglec-mediated regulation of immune cell function in disease. *Nature Reviews Immunology*, 14, 653-666.
McColl, H., et al. 2018. The prehistoric peopling of Southeast Asia. *Science*, 361, 88-92.
McDonough, W. C., et al. 2013. Pharmacogenomic association of nonsynonymous SNPs in SIGLEC12, A1BG, and the Selectin region and cardiovascular outcomes. *Hypertension*, 62, 48-54.
McEvoy, B., et al. 2006. The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. *Human Molecular Genetics*, 15, R176-181.
McGovern, D. P. B., et al. 2010. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Human Molecular Genetics*, 19, 3468-3476.
McLaren, W., et al. 2016. The Ensembl Variant Effect Predictor. *Genome Biology*, 17, 122.
Migliano, A. B., et al. 2013. Evolution of the pygmy phenotype: evidence of positive selection fro genome-wide scans in African, Asian, and Melanesian pygmies. *Human Biology*, 85, 251-284.
Migliano, A. B., et al. 2007. Life history trade-offs explain the evolution of human pygmies. *Proceedings of the National Academy of Sciences*, 104, 20216-20219.
Minahan, J. B. 2014, *Ethnic groups of North, East, and Central Asia: an encyclopedia*, ABC-CLIO.
Moreno, J. C. & Visser, T. J. 2007. New phenotypes in thyroid dyshormonogenesis: hypothyroidism due to DUOX2 mutations. *Endocrine Development*, 10, 99-117.
Mothe, I., et al. 1997. Interaction of wild type and dominant-negative p55PIK regulatory subunit of phosphatidylinositol 3-kinase with insulin-like growth factor-1 signaling proteins. *Molecular Endocrinology*, 11, 1911-1923.
Muragaki, Y., et al. 1996. A mutation in the gene encoding the α2 chain of the fibril-associated collagen IX, COL9A2, causes multiple epiphyseal dysplasia (EDM2). *Nature Genetics*, 12, 103-105.
Myles, S., et al. 2007. Identifying genes underlying skin pigmentation differences among human populations. *Human Genetics*, 120, 613-621.

Navarro Gonzalez, J., et al. 2020. The UCSC Genome Browser database: 2021 update. *Nucleic Acids Research*.
Nguyen, D.-Q., et al. 2006. Bias of selection on human copy number variants. *PLoS Genetics*, 2, e20.
Nielsen, R. 2005. Molecular signatures of natural selection. *Annual Review of Genetics*, 39, 197-218.
Norton, H. L., et al. 2006. Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Molecular Biology and Evolution*, 24s, 710-722.
Offenbacher, S., et al. 2016. Genome-wide association study of biologically informed periodontal complex traits offers novel insights into the genetic basis of periodontal disease. *Human Molecular Genetics*, 25, 2113-2129.
Okada, Y., et al. 2018. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nature Communications*, 9, 1631.
Pagani, L., et al. 2016. Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*, 538, 238-242.
Palle, K., et al. 2015. Aberrant GLI1 activation in DNA damage response, carcinogenesis and chemoresistance. *Cancers*, 7, 2330-2351.
Papatheodorou, I., et al. 2020. Expression Atlas update: from tissues to single cells. *Nucleic Acids Research*, 48, D77.
Park, B. L., et al. 2013. Extended genetic effects of ADH cluster genes on the risk of alcohol dependence: from GWAS to replication. *Human Genetics*, 132, 657-668.
Park, J., et al. 2012a. Effects of an Asian-specific nonsynonymous EDAR variant on multiple dental traits. *J. Hum. Genet.*, 57, 508-514.
Park, J., et al. 2012b. Effects of an Asian-specific nonsynonymous EDAR variant on multiple dental traits. *Journal of Human Genetics*, 57, 508-514.
Paten, B., et al. 2008a. Enredo and pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*, 18, 1814-1828.
Paten, B., et al. 2008b. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Research*, 18, 1829-1843.
Phan, L., et al. 2020. *ALFA: Allele Frequency Aggregator* [Online]. National Center for Biotechnology Information, U. S. National Library of Medicine. Available: www.ncbi.nlm.nih.gov/snp/docs/gsr/alfa/ [Accessed November 15 2020].
Philip, E. S., et al. 2010. Genome-wide association analysis identifies three psoriasis susceptibility loci. *Nature Genetics*, 42, 1000.
Pickrell, J. K., et al. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Research*, 19, 826-837.
Pillai, S., et al. 2012. Siglecs and immune regulation. *Annual Review of Immunology*, 30, 357-392.
Pritchard, J. K., et al. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, 20, R208-R215.
Provine, W. B. 1971, *The origins of theoretical population genetics*, University of Chicago Press, Chicago.
Qian, W., et al. 2013. Genome-wide landscapes of human local adaptation in Asia. *PLoS One*, 8, e54224-e54224.

Qiao, Y., et al. 2011. FOXQ1 regulates epithelial-mesenchymal transition in human cancers. *Cancer Research*, 71, 3076-3086.
Rajagopal, D. 2019. India to launch its 1st human genome cataloguing project. *The Economic Times*
Rajkumar, S. & Mark, G. 2008. Genomics: Protein fossils live on as RNA. *Nature*, 453, 729.
Reich, D., et al. 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American Journal of Human Genetics*, 89, 516-528.
Robinson, J. T., et al. 2011. Integrative Genomics Viewer. *Nature Biotechnology*, 29, 24-26.
Rocha, Z. L. 2011. Multiplicity within singularity: racial categorization and recognizing "mixed race" in Singapore. *Journal of Current Southeast Asian Affairs*, 30, 95-131.
Sabeti, P. C., et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419, 832-837.
Sabeti, P. C., et al. 2006. Positive natural selection in the human lineage. *Science*, 312, 1614-1620.
Sabeti, P. C., et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449, 913-918.
Sakagami, T., et al. 2004. Local adaptation and population differentiation at the interleukin 13 and interleukin 4 loci. *Genes and Immunity*, 5, 389-397.
Sakaue, S., et al. 2020. Functional variants in ADH1B and ALDH2 are non-additively associated with all-cause mortality in Japanese population. *European Journal of Human Genetics*, 28, 378-382.
Sather, C. 1997, *The Bajau Laut: Adaptation, History, and Fate in a Maritime Fishing Society of South-Eastern Sabah*, Oxford University Press, New York.
Saw, S.-H. 2012, *The population of Singapore*, Institute of Southeast Asian Studies, Singapore.
Seki, S., et al. 2006. Association study of COL9A2 with lumbar disc disease in the Japanese population. *Journal of Human Genetics*, 51, 1063-1067.
Shaffer, J. R., et al. 2017. Multiethnic GWAS Reveals Polygenic Architecture of Earlobe Attachment. *The American Journal of Human Genetics*, 101, 913-924.
Shido, K., et al. 2019. Susceptibility loci for tanning ability in the Japanese population identified by a genome-wide association study from the Tohoku Medical Megabank Project cohort study. *Journal of Investigative Dermatology*, 139, 1605-1608.e1613.
Shin, J.-G., et al. 2020. Genome-wide association analysis of 17,019 Korean women identifies variants associated with facial pigmented spots. *The Journal of Investigative Dermatology*.
Shriver, M., et al. 2004. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human Genomics*, 1, 274-286.
Simonson, T. S., et al. 2010. Genetic evidence for high-altitude adaptation in Tibet. *Science*, 329, 72-25.
Singapore Department of Statistics, T. 2017, *Population trends, 2017*, Department of Statistics, Ministry of Trade & Industry, Singapore.
Skoglund, P. & Jakobsson, M. 2011. Archaic human ancestry in East Asia. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 18301-18306.
Smith, J. M. & Haigh, J. 1974. The hitch-hiking effect of a favourable gene. *Genetics Research*, 23, 23-35.

Smyth, D., et al. 2011. FUT2 nonsecretor status links type 1 diabetes susceptibility and resistance to infection. *Diabetes*, 60, 3081–3084.
Soejima, M. & Koda, Y. 2007. Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. *International Journal of Legal Medicine*, 121, 36-39.
Sudlow, C., et al. 2015. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. 12, e1001779.
Sun, H., et al. 2019a. Application of partial least squares in exploring the genome selection signatures between populations. *Heredity*, 122, 288-293.
Sun, R., et al. 2018. Identification and functional studies of MYO1H for mandibular prognathism. *Journal of Dental Research*, 97, 1501-1509.
Sun, Y., et al. 2019b. Association of TSR1 variants and spontaneous coronary artery dissection. *Journal of the American College of Cardiology*, 74, 167-176.
Sung-Hee, H., et al. 2008. Carrier frequency of GJB2 (connexin-26) mutations causing inherited deafness in the Korean population. *Journal of Human Genetics*, 53, 1022-1028.
Sung, H. Y., et al. 2017. Aberrant epigenetic regulation of GABRP associates with aggressive phenotype of ovarian cancer. *Experimental & Molecular Medicine*, 49, e335.
Szpak, M., et al. 2018. FineMAV: prioritizing candidate genetic variants driving local adaptations in human populations. *Genome Biology*, 19.
Szpak, M., et al. 2019. How well do we understand the basis of classic selective sweeps in humans? *FEBS Letters*, 593, 1431-1448.
Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123, 585-595.
Tan, J., et al. 2014. Characteristics of dental morphology in the Xinjiang Uyghurs and correlation with the EDARV370A variant. *Science China Life Sciences*, 57, 510-518.
Tang, K., et al. 2007. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome (Positive Selection in the Human Genome). *PLoS Biology*, 5, e171.
Tardif, S., et al. 2010. Zonadhesin is essential for species specificity of sperm adhesion to the egg zona pellucida. *The Journal of Biological Chemistry*, 285, 24863-24870.
Tassopoulou-Fishell, M., et al. 2012. Genetic variation in Myosin 1H contributes to mandibular prognathism. *American Journal of Orthodontics and Dentofacial Orthopedics*, 141, 51-59.
Tate, J. G., et al. 2019. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47, D941-D947.
Tekin, M., et al. 2010. GJB2 mutations in Mongolia: complex alleles, low frequency, and reduced fitness of the deaf. *Annals of Human Genetics*, 74, 155-164.
Teo, Y. Y., et al. 2009. Singapore genome variation project: a haplotype map of three South-East Asian populations. *Genome Research*, 19, 2154-2162.
The 1000 Genomes Project Consortium 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467, 1061-1073.
The 1000 Genomes Project Consortium 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491, 56.

The 1000 Genomes Project Consortium 2015. A global reference for human genetic variation. *Nature*, 526, 68-74.
The GTEx Consortium 2017. Genetic effects on gene expression across human tissues. *Nature*, 550, 204-213.
The HUGO Pan-Asian SNP Consortium 2009. Mapping human genetic diversity in Asia. *Science*, 326, 1541-1545.
The International HapMap 3 Consortium 2010. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467, 52-58.
The International HapMap Consortium 2005. A haplotype map of the human genome. *Nature*, 437, 1299-1320.
The International HapMap Consortium 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449, 851-861.
Thorven, M., et al. 2005. A homozygous nonsense mutation (428G->A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections. *The Journal of Virology*, 79, 15351–15355.
Tishkoff, S. A., et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, 39, 31-40.
Tong, M., et al. 2014. Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. *The ISME Journal*, 8, 2193–2206.
Urszula, M. M., et al. 2016. Pre-40S ribosome biogenesis factor Tsr1 is an inactive structural mimic of translational GTPases. *Nature Communications*, 7, 11789.
Van Dyke, T. & Dave, S. 2005. Risk factors for periodontitis. *Journal of the International Academy of Periodontology*, 7, 3-7.
Varki, A. & Angata, T. 2006. Siglecs - the major subfamily of I-type lectins. *Glycobiology*, 16, 1R-27R.
Vitti, J. J., et al. 2013. Detecting natural selection in genomic data. *Annual Review of Genetics*, 47, 97-120.
Voight, B. F., et al. 2006. A map of recent positive selection in the human genome. *PLoS Biology*, 4, e72-e72.
Wan Juhari, W. K., et al. 2014. A whole genome analyses of genetic variants in two Kelantan Malay individuals. *The HUGO Journal*, 8, 4.
Wang, E. T., et al. 2006. Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 135-140.
Warren, H. R., et al. 2017. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nature Genetics*, 49, 403-415.
Wei, C.-Y., et al. 2019. Bioinformatics-based analysis reveals elevated MFSD12 as a key promoter of cell proliferation and a potential therapeutic target in melanoma. *Oncogene*, 38, 1876-1891.
Wei, Q., et al. 2013. Genetic mutations of GJB2 and mitochondrial 12S rRNA in nonsyndromic hearing loss in Jiangsu Province of China. *Journal of Translational Medicine*, 11, 163.
Weir, B. S. 1996, *Genetic data analysis II: methods for discrete population genetic data*.
West, B. A. 2009, *Encyclopedia of the Peoples of Asia and Oceania*, Facts on File.

William, T., et al. 2013. Increasing incidence of Plasmodium knowlesi malaria following control of P. falciparum and P. vivax Malaria in Sabah, Malaysia. *PLoS Neglected Tropical Diseases*, 7, e2026.
Wong, L. P., et al. 2014. Insights into the genetic structure and diversity of 38 South Asian Indians from deep whole-genome sequencing. *PLoS Genetics*, 10, e1004377-e1004377.
Wong, L. P., et al. 2013. Deep whole-genome sequencing of 100 Southeast Asian Malays. *American Journal of Human Genetics*, 92, 52-66.
World Health Organization. 2005. *Periodontal country profiles* [Online]. Available: https://www5.dent.niigata-u.ac.jp/~prevent/perio/contents.html [Accessed December 31 2019].
Wu, D., et al. 2019. Large-scale whole-genome sequencing of three diverse asian populations in Singapore. *Cell*, 179, 736-749.e715.
Wu, S., et al. 2016. Genome-wide scans reveal variants at EDAR predominantly affecting hair straightness in Han Chinese and Uyghur populations. *Human Genetics*, 135, 1279-1286.
Xiao, H., et al. 2018. Single nucleotide polymorphism rs2274084 of gap junction protein beta 2 gene among Epstein-Barr virus-associated tumors. *Cancer Biomarkers*, 21, 499-504.
Yahya, P., et al. 2017. Analysis of the genetic structure of the Malay population: Ancestry-informative marker SNPs in the Malay of Peninsular Malaysia. *Forensic Science International: Genetics*, 30, 152-159.
Yamada, Y., et al. 2017. Identification of C21orf59 and ATG2A as novel determinants of renal function-related traits in Japanese by exome-wide association studies. *Oncotarget*, 8, 45259-452773.
Yasukochi, Y., et al. 2018. Identification of CDC42BPG as a novel susceptibility locus for hyperuricemia in a Japanese population. *Molecular Genetics and Genomics*, 293, 371-379.
Yasumizu, Y., et al. 2020. Genome-wide natural selection signatures are linked to genetic risk of modern phenotypes in the Japanese population. *Molecular Biology and Evolution*, 37, 1306-1316.
Yew, C. W., et al. 2018. Genetic relatedness of indigenous ethnic groups in northern Borneo to neighboring populations from Southeast Asia, as inferred from genome-wide SNP data. *Annals of Human Genetics*, 82, 216-226.
Yi, X., et al. 2010a. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329, 75-78.
Yi, X., et al. 2010b. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329, 75-78.
Yngvadottir, B., et al. 2009. A Genome-wide Survey of the Prevalence and Evolutionary Forces Acting on Human Nonsense SNPs. *The American Journal of Human Genetics*, 84, 224-234.
Zheng, J., et al. 2015. GJB2 mutation spectrum and genotype-phenotype correlation in 1067 Han Chinese subjects with non-syndromic hearing loss. *PLoS One*, 10, e0128691.
Zhu, L. & Bustamante, C. D. 2005. A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics*, 170, 1411-1421.

6 APPENDICES

6.1 Appendix A

Glossary

Allele	One of the alternative forms of a gene or any other locus on a chromosome.
Ancestral allele	An allele that is not derived.
Derived allele	An allele that arises due to mutation during the evolution of a species.
Fitness	The ability of an individual to survive and reproduce relative to the rest of the population.
Fixation	The change in a gene pool from a situation where there exists at least two alleles in a given population to a situation where only one of the alleles remain.
Genetic hitchhiking	When the selection of one allele increases the frequency of other neutral alleles in a population that are in proximity to it on the same genomic segment.
Haplotype	A group of genes within an organism that was inherited together from a single parent.
Hard selective sweep	A type of selective sweep in which a new advantageous mutation arises, and spreads quickly to fixation due to natural selection.
Incomplete selective sweep	A type of selective sweep in which an advantageous allele increases rapidly from low frequency, but has not yet reached fixation.
Linkage disequilibrium	Non-random association between alleles in a population due to their tendency to be co-inherited because of reduced recombination between them.
Neutral allele	An allele that does not affect the fitness of the carrier.
Population differentiation	The process by which allele frequencies in two or more populations diverge over time.
Positive selection	Any selection that acts upon new, favourable mutations.
Selective sweep	An event in which the frequency of an advantageous allele increases rapidly due to selection .

6.2 Appendix B

List of software and online resources that were used in this project

Software or resource	Reference	URL
Deposited data		
Deep whole-genome sequencing of 90 Han Chinese genomes	(Lan et al., 2017)	https://www.ebi.ac.uk/ena/data/view/PRJEB20820
Singapore Sequencing Indian Project	(Wong et al., 2014)	https://blog.nus.edu.sg/sshsphphg/singapore-sequencing-indian/
Singapore Sequencing Malay Project	(Wong et al., 2013)	https://blog.nus.edu.sg/sshsphphg/singapore-sequencing-malay/
1000 Genomes Project (Phase 3)	(The 1000 Genomes Project Consortium, 2015)	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
GenomeAsia 100K	(GenomeAsia100K Consortium, 2019)	https://browser.genomeasia100k.org/#tid=download
PHRED-like CADD scores (v1.4, GrCh37/hg19)	(Kircher et al., 2014)	https://krishna.gs.washington.edu/download/CADD/v1.4/GRCh37/whole_genome_SNVs.tsv.gz
FASTA files of the ancestral sequences for Homo sapiens (release 71)	(Paten et al., 2008a; Paten et al., 2008b)	ftp://ftp.ensembl.org/pub/release75/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh37_e71.tar.bz2
Software		
ADMIXTURE (v1.3)	(Alexander et al., 2009)	http://dalexander.github.io/admixture/download.html
BCFtools (v1.9)	(Li et al., 2009b)	https://samtools.github.io/bcftools/bcftools.html
Haploview (v4.2)	(Barrett et al., 2005)	https://www.broadinstitute.org/haploview/haploview
PLINK (v.1.9)	(Chang et al., 2015)	https://www.cog-genomics.org/plink2
SnpSift (v.4.3.1)	(Cingolani et al., 2012)	http://snpeff.sourceforge.net/SnpSift.html
Variant Effect Predictor (VEP) (v98)	(McLaren et al., 2016)	https://github.com/Ensembl/ensembl-vep.git

6.3 Appendix C

The population names and codes of the 1000 Genomes Project Phase 3 dataset (The 1000 Genomes Project Consortium, 2015).

Continental population and code	Population description	Population code
African (AFR)	Esan in Nigeria	ENS
	Gambian in Western Division, Mandinka	GWD
	Luhya in Webuye, Kenya	LWK
	Mende in Sierra Leone	MSL
	Yoruba in Ibadan, Nigeria	YRI
	African Caribbean in Barbados	ACB
	People with African Ancestry in Southwest USA	ASW
Admixed American (AMR)	Colombians in Medellin, Colombia	CLM
	People with Mexican Ancestry in Los Angeles, CA, USA	MXL
	Peruvians in Lima, Peru	PEL
	Puerto Ricans in Puerto Rico	PUR
European (EUR)	Utah residents (CEPH) with Northern and Western European ancestry	CEU
	British in England and Scotland	FBR
	Finnish in Finland	FIN
	Iberian Populations in Spain	IBS
	Toscani in Italia	TSI
East Asian (EAS)	Chinese Dai in Xishuangbanna, China	CDX
	Han Chinese in Beijing, China	CHB
	Southern Han Chinese	CHS
	Japanese in Tokyo, Japan	JPT
	Kinh in Ho Chi Minh City, Vietnam	KHV
South Asian (SAS)	Bengali in Bangladesh	BEB
	Gujarati Indians in Houston, TX, USA	GIH
	Indian Telugu in the UK	ITU
	Punjabi in Lahore, Pakistan	PJL
	Sri Lankan Tamil in the UK	STU

6.4 Appendix D

The populations from the GenomeAsia 100K dataset (GenomeAsia100K Consortium, 2019) that was used in this project.

Continental population and code	Number of individuals	Country of origin	Population group(s)
Northeast Asia (NEA)	346	China	Daur, Han, Hezhen, Mongola, Naxi, Oroqen, She, Tu, Tujia, Uygur, Xibo, Han Chinese in Beijing (CHB) and Southern Han Chinese (CHS) from the 1000 Genomes Project
		Japan	Japanese, Japanese (JPT) from the 1000 Genomes Project
		Korea	Korean
		Kyrgyzstan	Kyrgyz
		Mongolia	Buryat, Mongola, Xhalxh
		Russia	Aleut, Altaian, Chukchi, Eskimo Chaplin, Eskimo Naukan, Eskimo Sireniki, Even, Itelman, Mansi, Tlingit, Tubalar, Ulchi, Yakut
Oceania (OCE)	68	Australia	Australian
		New Zealand	Maori
		Papua New Guinea	Ata, Baining, Bougainville, Lavongai, Mamusi, Mussau, Nailik, Nakanai, Nakanai Bileki, Nakanai Loso, Papuan, Pasismanua
		United States	Hawaiian
South Asia (SAS)	681	Bangladesh	Bengali (BEB) from the 1000 Genomes Project
		India	Abujmaria, Agharia, Bagdi, Birhor, Bison Horn Maria, Birhor, Brahmin, Chakma, Chamar, Chanchu, Dhurwa, Dorla, Gaud, Halba, Hill Korwa, Indian Telugu, Indian, Irula, Iyangar, Iyer, Jamatia, Jarwa, Kamar, Kapu, Kaya Dora, Khatri, Khonda Dora, Konda Reddy, Kota, Lambada, Lodha, Madiga, Mahar, Mala, Manipuri, Mog, Munda, Muria, Nav Buddha, Nicobarese, Onge, Oraon, Paniya, Rana

			Tharu, Relli, Saryupari Brahmin, Sourasthra Brahmin, South Indian, Tanti, Toda, Toto, Urban Bangalore, Urban Chennai, West Bengal Brahmin, Yadava, Gujarati Indians in Houston, Texas, United States (GIH) from the 1000 Genomes Project
		Nepal	Kusunda
		Pakistan	Balochi, Brahui, Brusho, Burusho, Gujjar, Hazara, Kalash, Makrani, Parsi, Pathan, Punjabi, Rajput, Sindhi, Punjabi in Lahore (PJL) from the 1000 Genomes Project
		Sri Lanka	Sri Lankan Tamil in the United Kingdom (STU) from the 1000 Genomes Project
Southeast Asia (SEA)	333	Brunei	Dusun
		Cambodia	Cambodian
		China	Dai, Lahu, Miao, Yi, Chinese Dai in Xishuangbanna (CDX) from the 1000 Genomes Project
		Indonesia	Austronesian, Flores Bena, Flores Cibal, Flores Rampasasa
		Malaysia	Kenisu, Kintak, Malaysian, Senoi Che Wong, Senoi Semai, Senoi Smak Beri, Temuan
		Philippines	Aeta, Ati, Igorot
		Taiwan	Ami, Atayal
		Thailand	Thai
		Singapore	Burmese
		Vietnam	Kinh in Ho Chi Minh City (KHV) from the 1000 Genomes Project

6.5 Appendix E

Histograms of the pairwise PI_HAT values between the individuals in the 90 Han Chinese (90HC), Singaporean Indian (SSIP), Singaporean Malay (SSMP) and the 1000 Genomes Project datasets.

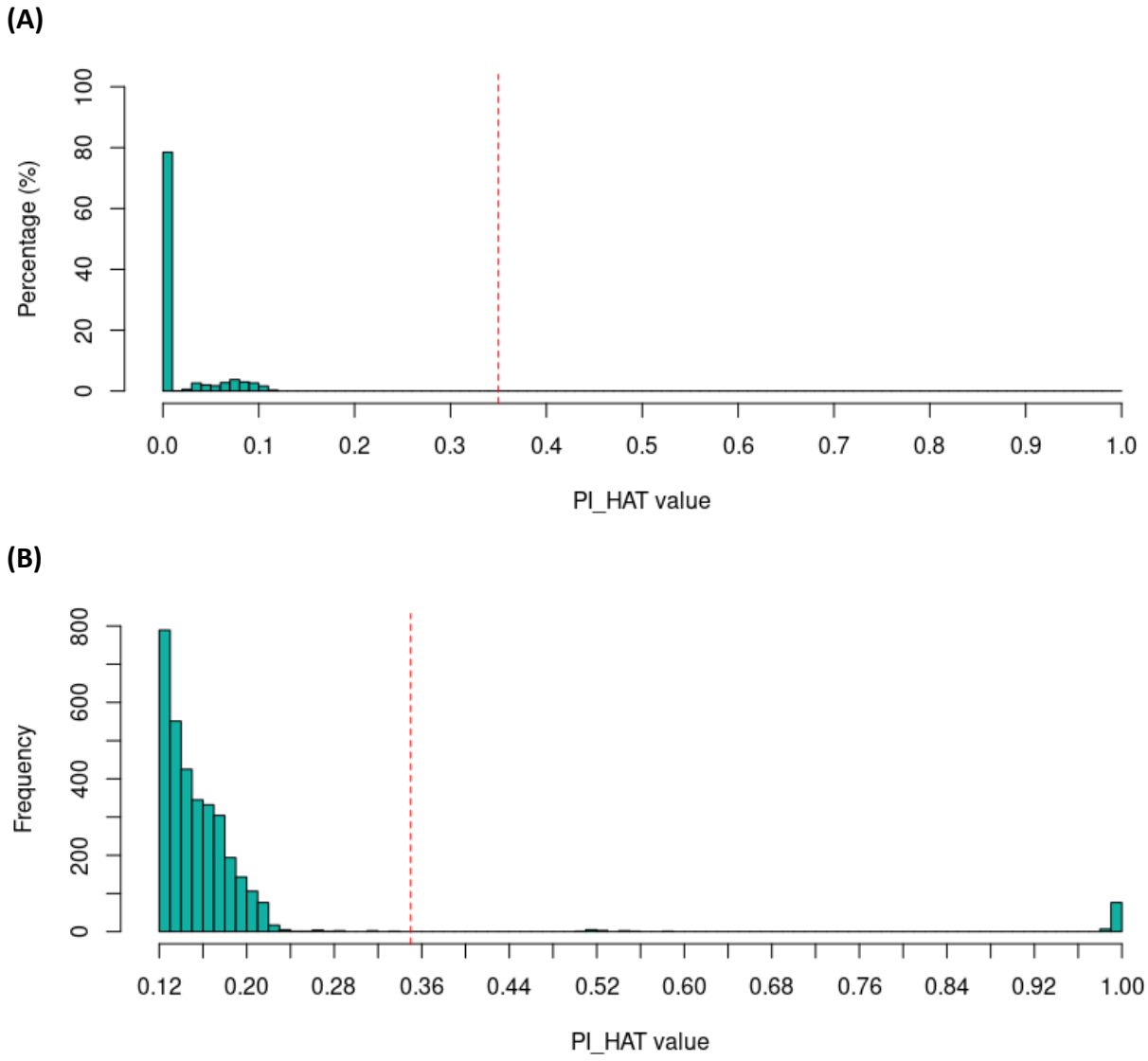


Figure A1. Panel A shows a histogram representing proportional distribution of the PI_HAT values ranging from 0.0 to 1.0. Panel B shows a histogram representing the frequency distribution of PI_HAT values ranging from 0.12 to 1.00. The red dotted line at PI_HAT = 0.35 represents the threshold that was used to filter the datasets. Setting a lower threshold would have excluded multiple samples, compromising allele frequency calculations in the individual populations.

6.6 Appendix F

Lists of top *FineMAV* candidates

The following pages contain the top 50 *FineMAV* hits from the Han Chinese (90HC), Singaporean Indian (SSIP) and Singaporean Malay (SSMP) datasets as well as the top 100 *FineMAV* hits from the GenomeAsia 100K Northeast Asian (NEA), South Asian (SAS), Southeast Asian (SEA) and Oceanian (OCE) populations. The values in the columns are rounded to two decimal places.

HGVS (hg19/GRCh37)	Genomic Human Genome Variation Society (HGVS) nomenclature using the hg19/GRCh37 reference genome
SNP ID	Single Nucleotide Polymorphism ID
DER	Derived allele
GENE	Gene name
CONSEQUENCE	Most severe variant consequence according to Ensembl (NC stands for non-coding; UTR stands for untranslated region)
CADD	PHRED-like scaled Combined Annotation-Dependent Depletion score
DAF	Derived allele frequency of the population
DAP	Derived allele purity
FineMAV	Fine-Mapping of Adaptive Variation score of the population

Top 50 *FineMAV* hits for the Han Chinese dataset (90HC).

HGVS (hg19/GRCh37)	SNP ID	DER	Gene	Consequence	CADD	DAF 90HC	DAF SSIP	DAF SSMP	DAP	*FineMAV* 90HC
NC_000002.11:g.109513601A>G	rs3827760	G	*EDAR*	Missense (p.Val370Ala)	21.70	0.92	0.03	0.49	0.23	4.66
NC_000005.9:g.176099727A>G	rs13186794	G	-	Intergenic	16.22	0.49	0.06	0.05	0.51	4.11
NC_000005.9:g.176099728A>G	rs13186795	G	-	Intergenic	17.15	0.49	0.06	0.06	0.48	4.10
NC_000004.11:g.31442427G>A	rs56345433	A	-	Intergenic	11.56	0.53	0.09	0.02	0.53	3.21
NC_000003.11:g.98031307T>A	rs2316271	A	*OR5H8*	Stop gained (p.Leu184Ter)	43.00	0.77	0.31	0.60	0.09	3.10
NC_000016.9:g.31088347G>A	rs749671	A	*ZNF646*	Synonymous (p.Glu234=)	20.30	0.91	0.04	0.78	0.17	3.05
NC_000005.9:g.76129053T>C	rs631465	T	*F2RL1*	Synonymous (p.Ile207=)	19.19	0.52	0.01	0.21	0.30	3.01
NC_000002.11:g.109451118A>G	rs72627476	G	*CCDC138*	Intron	13.82	0.92	0.03	0.48	0.23	2.96
NC_000012.11:g.132106717T>C	rs10794470	T	*AC117500.3*	Intron	11.54	0.27	0.00	0.01	0.94	2.94
NC_000007.13:g.14587199G>A	rs10236893	A	*DGKB*	Intron	19.92	0.42	0.03	0.12	0.35	2.89
NC_000003.11:g.50326020G>A	rs2229647	A	*IFRD2*	Synonymous (p.His446=)	19.84	0.67	0.01	0.40	0.22	2.87
NC_000010.10:g.3173092A>T	rs71502284	T	*PFKP*	Intron	10.61	0.32	0.01	0.01	0.81	2.78
NC_000016.9:g.31075175G>A	rs2303223	A	*ZNF668*	Synonymous (p.Gly225=)	17.64	0.91	0.04	0.78	0.17	2.67
NC_000007.13:g.14587021T>C	rs10252073	C	*DGKB*	Intron	17.43	0.43	0.03	0.12	0.36	2.66
NC_000016.9:g.55060635T>C	rs4517796	C	-	Intergenic	11.27	0.39	0.03	0.03	0.61	2.66
NC_000011.9:g.62848487A>C	rs11231341	C	*SLC22A24*	Stop gained (p.Tyr501Ter)	47.00	0.87	0.76	0.79	0.07	2.65
NC_000003.11:g.50187637T>C	rs58137261	C	*SEMA3F*	Upstream gene	17.75	0.69	0.01	0.41	0.22	2.64
NC_000018.9:g.22511045C>T	rs17188214	T	*AC018697.1*	Downstream gene	21.50	0.36	0.03	0.10	0.34	2.63
NC_000016.9:g.31374535C>G	rs2230429	G	*ITGAX*	Missense (p.Pro517Arg)	24.30	0.73	0.11	0.54	0.14	2.57
NC_000015.9:g.64759279C>T	rs35685348	T	*AC091231.1*	Upstream gene	22.00	0.78	0.09	0.64	0.15	2.56
NC_000007.13:g.14535608T>C	rs16878192	C	*DGKB*	Intron	16.85	0.54	0.07	0.17	0.28	2.56
NC_000002.11:g.104599371T>G	rs78407975	G	*LINC01965*	Intron	19.97	0.31	0.01	0.07	0.42	2.55
NC_000014.8:g.66950852A>G	rs28655067	G	*AL359232.1*	Intron	9.52	0.46	0.04	0.03	0.59	2.54
NC_000020.10:g.22384894G>T	rs12481108	T	*AL133464.1*	Intron	19.62	0.42	0.03	0.15	0.31	2.53
NC_000011.9:g.21745636C>T	rs12418851	T	-	Intergenic	15.16	0.28	0.00	0.05	0.59	2.52
NC_000008.10:g.117645029T>A	rs62510171	A	-	Intergenic	8.47	0.41	0.01	0.03	0.72	2.51
NC_000003.11:g.50198840G>C	rs74595980	C	*SEMA3F*	Intron	16.99	0.68	0.01	0.41	0.22	2.48
NC_000001.10:g.36225948T>C	rs7537203	C	*CLSPN*	Missense (p.Asn525Ser)	24.80	0.84	0.24	0.59	0.12	2.45
NC_000016.9:g.31088625A>G	rs749670	G	*ZNF646*	Missense (p.Glu327Gly)	16.32	0.91	0.04	0.78	0.17	2.45
NC_000002.11:g.26676395G>T	rs12623642	T	*DRC1*	Missense (p.Val633Phe)	26.10	0.64	0.14	0.37	0.15	2.45
NC_000008.10:g.129884159T>C	rs13276570	C	-	Intergenic	20.90	0.76	0.07	0.62	0.15	2.44
NC_000008.10:g.3920668A>C	rs77891957	C	*CSMD1*	Intron	5.25	0.63	0.03	0.03	0.73	2.42
NC_000009.11:g.73150984C>T	rs6560142	T	*TRPM3*	Missense (p.Arg1670Gln)	33.00	0.82	0.37	0.67	0.09	2.40
NC_000007.13:g.100371358G>A	rs2293766	A	*ZAN*	Stop gained (p.Trp1883Ter)	52.00	0.53	0.26	0.56	0.09	2.40
NC_000014.8:g.97272382T>G	rs2224442	G	*VRK1*	Intron	19.81	0.87	0.16	0.63	0.14	2.38
NC_000003.11:g.62986072A>G	rs34047489	G	*LINC00698*	Intron	15.03	0.33	0.01	0.06	0.48	2.37
NC_000004.11:g.100436354A>G	rs78349254	G	*C4orf17*	Intron	10.17	0.43	0.01	0.07	0.54	2.36
NC_000003.11:g.50197092C>T	rs2072053	T	*SEMA3F*	Synonymous (p.Leu13=)	16.15	0.68	0.01	0.41	0.22	2.36
NC_000015.9:g.40581543T>C	rs936212	C	*PLCB2*	Missense (p.Glu1095Gly)	20.30	0.62	0.04	0.40	0.19	2.35
NC_000009.11:g.125273435G>A	rs41277120	A	*OR1J2*	Missense (p.Ala119Thr)	23.00	0.37	0.09	0.08	0.27	2.33
NC_000002.11:g.170010985T>C	rs2075252	T	*LRP2*	Missense (p.Lys4094Glu)	22.60	0.55	0.09	0.28	0.19	2.31
NC_000005.9:g.66908751T>G	rs59491487	G	*AC112206.1*	Downstream gene	20.40	0.80	0.14	0.56	0.14	2.31
NC_000001.10:g.66036441A>G	rs1137100	G	*LEPR*	Missense (p.Lys109Arg)	18.75	0.87	0.14	0.64	0.14	2.31
NC_000002.11:g.37968684G>A	rs72802008	A	*AC006369.2*	Upstream gene	11.97	0.32	0.03	0.02	0.60	2.28
NC_000016.9:g.31000809G>A	rs13708	A	*STX1B*	3 prime UTR	18.00	0.92	0.14	0.80	0.14	2.27
NC_000003.11:g.50249500G>A	rs4688744	A	*SLC38A3*	Intron	16.01	0.67	0.01	0.41	0.21	2.27
NC_000010.10:g.73452238T>A	rs76555066	A	*CDH23*	Intron	18.68	0.34	0.03	0.09	0.36	2.26
NC_000012.11:g.112477055T>C	rs12231744	C	*NAA25*	Missense (p.Lys876Arg)	23.80	0.61	0.06	0.50	0.15	2.25
NC_000016.9:g.48258198C>T	rs17822931	T	*ABCC11*	Missense (p.Gly180Arg)	24.00	0.91	0.41	0.50	0.10	2.25
NC_000021.8:g.30331935G>A	rs57646126	A	*LTN1*	Missense (p.Ala859Val)	22.80	0.40	0.01	0.20	0.24	2.23

Top 50 *FineMAV* hits for the Singaporean Indian dataset (SSIP).

HGVS (hg19/GRCh37)	SNP ID	DER	GENE	CONSEQUENCE	CADD	DAF 90HC	DAF SSIP	DAF SSMP	DAP	*FineMAV* SSIP
NC_000016.9:g.28506428C>T	rs151233	T	*APOBR*	Synonymous (p.Leu22=)	16.22	0.01	0.57	0.03	0.83	7.68
NC_000016.9:g.30936081G>A	rs35675346	A	*FBXL19*	Missense (p.Glu10Lys)	23.10	0.06	0.80	0.19	0.39	7.21
NC_000016.9:g.28505660G>C	rs151234	C	*CLN3*	Intron	14.89	0.01	0.57	0.03	0.80	6.84
NC_000016.9:g.31044683A>G	rs58726213	G	*STX4*	Upstream gene	21.60	0.09	0.87	0.21	0.36	6.69
NC_000015.9:g.64592833T>C	rs114713921	C	*CSNK1G1*	5 prime UTR	17.45	0.01	0.49	0.04	0.75	6.34
NC_000016.9:g.30666367C>T	rs3747481	T	*PRR14*	Missense (p.Pro359Leu)	22.90	0.10	0.86	0.24	0.31	6.09
NC_000019.9:g.49206674G>A	rs601338	A	*FUT2*	Stop gained (p.Trp154Ter)	52.00	0.01	0.19	0.02	0.62	6.03
NC_000015.9:g.91452595A>G	rs2106673	A	*MAN2A2*	Missense (p.Gln412Arg)	18.43	0.02	0.51	0.06	0.61	5.75
NC_000010.10:g.17407147G>T	rs729170	T	*ST8SIA6*	Intron	18.64	0.01	0.34	0.01	0.90	5.74
NC_000015.9:g.64653984G>T	rs8026043	G	*PCLAF*	Downstream gene	15.76	0.01	0.49	0.04	0.75	5.73
NC_000001.10:g.10271688C>G	rs11121529	G	*KIF1B*	Intron	18.67	0.01	0.36	0.01	0.86	5.72
NC_000016.9:g.30999462T>C	rs2305880	T	*HSD3B7*	Synonymous (p.Arg356=)	17.74	0.08	0.86	0.20	0.37	5.68
NC_000007.13:g.21068814A>G	rs12665958	G	-	Intergenic	22.20	0.01	0.29	0.01	0.88	5.57
NC_000007.13:g.25696612T>C	rs11509164	C	*AC005165.1*	Intron	19.71	0.06	0.60	0.09	0.46	5.44
NC_000004.11:g.135297559C>T	rs1486995	C	-	Intergenic	19.94	0.01	0.40	0.04	0.67	5.38
NC_000015.9:g.65042560G>A	rs61741344	A	*RBPMS2*	Synonymous (p.Ile62=)	17.38	0.01	0.37	0.02	0.82	5.32
NC_000002.11:g.39109558G>A	rs3099950	A	*MORN2*	Missense (p.Glu48Lys)	25.50	0.01	0.26	0.01	0.81	5.31
NC_000006.11:g.106535936C>T	rs1340065	C	*PRDM1*	Intron	19.03	0.01	0.49	0.07	0.57	5.29
NC_000015.9:g.64792896G>A	rs640005	G	*ZNF609*	Intron	15.06	0.01	0.41	0.02	0.84	5.24
NC_000008.10:g.110547482A>G	rs72669129	G	*EBAG9*	Upstream gene	13.06	0.01	0.59	0.06	0.68	5.21
NC_000016.9:g.31090407G>C	rs35713203	C	*ZNF646*	Missense (p.Gly921Ala)	16.19	0.09	0.86	0.20	0.36	5.00
NC_000001.10:g.53153432T>C	rs443751	C	*COA7*	Missense (p.Lys219Arg)	21.70	0.02	0.40	0.05	0.57	4.99
NC_000015.9:g.64940203T>C	rs6494484	T	*ZNF609*	Intron	14.24	0.01	0.41	0.02	0.84	4.95
NC_000001.10:g.51121198T>C	rs11205753	C	*FAF1*	Splice region (p.Val220=)	14.72	0.01	0.41	0.02	0.81	4.91
NC_000001.10:g.49620445T>C	rs549430	C	*AGBL4*	Intron	19.16	0.04	0.44	0.03	0.58	4.89
NC_000020.10:g.30753270T>C	rs14316	T	*TM9SF4*	3 prime UTR	13.57	0.00	0.59	0.09	0.61	4.86
NC_000015.9:g.48426484A>G	rs1426654	A	*SLC24A5*	Missense (p.Thr111Ala)	19.66	0.01	0.43	0.07	0.58	4.86
NC_000012.11:g.111847740A>G	rs3803170	A	*SH2B3*	Intron	21.40	0.06	0.67	0.19	0.34	4.84
NC_000001.10:g.50576710T>A	rs4357572	A	*ELAVL4*	Intron	20.30	0.04	0.44	0.04	0.54	4.83
NC_000015.9:g.64513415A>G	rs116046132	G	*CSNK1G1*	Intron	13.81	0.01	0.47	0.04	0.74	4.83
NC_000001.10:g.49796157A>G	rs1494462	G	*AGBL4*	Intron	19.21	0.04	0.46	0.04	0.55	4.80
NC_000005.9:g.87929869T>C	rs10060622	C	*LINC00461*	Intron	14.38	0.02	0.46	0.02	0.73	4.80
NC_000015.9:g.37632513G>T	rs28588437	T	-	Intergenic	18.17	0.02	0.33	0.01	0.80	4.76
NC_000002.11:g.42181679A>T	rs6740960	A	*C2orf91*	Upstream gene	17.73	0.03	0.54	0.09	0.49	4.76
NC_000009.11:g.594262A>C	rs2641998	C	*KANK1*	Intron	20.80	0.08	0.59	0.10	0.39	4.74
NC_000002.11:g.80479986G>A	rs76873192	A	*CTNNA2*	Intron	14.81	0.01	0.39	0.02	0.83	4.74
NC_000005.9:g.60199363C>T	rs4647102	C	*ERCC8*	Intron	18.28	0.06	0.49	0.04	0.53	4.69
NC_000018.9:g.53963907A>G	rs1558536	G	-	Intergenic	21.90	0.04	0.50	0.10	0.43	4.69
NC_000016.9:g.77587338T>A	rs282978	T	-	Intergenic	18.67	0.04	0.46	0.05	0.55	4.68
NC_000012.11:g.48736985T>G	rs2732481	G	*ZNF641*	Missense (p.Gln363Pro)	22.50	0.01	0.47	0.11	0.44	4.67
NC_000010.10:g.28468459A>C	rs34772309	C	*MPP7*	Intron	11.98	0.01	0.49	0.03	0.80	4.67
NC_000015.9:g.37625116A>G	rs28440992	G	-	Intergenic	17.67	0.02	0.33	0.01	0.80	4.63
NC_000008.10:g.110283353T>G	rs34660136	G	*NUDCD1*	Missense (p.Asn394His)	23.50	0.01	0.33	0.05	0.60	4.60
NC_000017.10:g.54942723A>C	rs9911132	A	*DGKE*	3 prime UTR	13.26	0.03	0.63	0.08	0.55	4.60
NC_000005.9:g.60299072G>A	rs162242	G	*NDUFAF2*	Intron	17.25	0.05	0.49	0.04	0.55	4.58
NC_000010.10:g.28439185C>T	rs1781836	T	*MPP7*	Intron	11.72	0.01	0.49	0.03	0.80	4.57
NC_000015.9:g.37780648G>A	rs17446248	A	*AC068875.1*	Intron	17.29	0.02	0.36	0.02	0.74	4.56
NC_000004.11:g.13242982A>G	rs59531204	G	-	Intergenic	20.50	0.01	0.29	0.02	0.78	4.56
NC_000016.9:g.30995669T>C	rs6950	T	*SETD1A*	3 prime UTR	14.25	0.08	0.86	0.20	0.37	4.56
NC_000004.11:g.13174536A>G	rs28368703	G	-	Intergenic	17.22	0.01	0.34	0.02	0.77	4.55

Top 50 *FineMAV* hits for the Singaporean Malay dataset (SSMP).

HGVS (hg19/GRCh37)	SNP ID	DER	GENE	CONSEQUENCE	CADD	DAF 90HC	DAF SSIP	DAF SSMP	DAP	*FineMAV* SSMP
NC_000002.11:g.98272491A>G	rs2290123	G	*ACTR1B*	3 prime UTR	15.06	0.03	0.03	0.38	0.59	3.38
NC_000002.11:g.97613974C>G	rs114979404	G	*FAM178B*	Intron	11.67	0.02	0.03	0.38	0.64	2.81
NC_000017.10:g.2238152T>C	rs79597880	C	*TSR1*	Missense (p.Lys199Glu)	25.90	0.09	0.01	0.30	0.36	2.75
NC_000016.9:g.31088347G>A	rs749671	A	*ZNF646*	Synonymous (p.Glu234=)	20.30	0.91	0.04	0.78	0.17	2.62
NC_000007.13:g.100371358G>A	rs2293766	A	*ZAN*	Stop gained (p.Trp1883Ter)	52.00	0.53	0.26	0.56	0.09	2.53
NC_000002.11:g.109513601A>G	rs3827760	G	*EDAR*	Missense (p.Val370Ala)	21.70	0.92	0.03	0.49	0.23	2.47
NC_000003.11:g.98031307T>A	rs2316271	A	*OR5H8*	Stop gained (p.Leu184Ter)	43.00	0.77	0.31	0.60	0.09	2.42
NC_000011.9:g.62848487A>C	rs11231341	C	*SLC22A24*	Stop gained (p.Tyr501Ter)	47.00	0.87	0.76	0.79	0.07	2.42
NC_000012.11:g.57865558G>T	rs2229300	T	*GLI1*	Missense (p.Gly1012Val)	25.80	0.05	0.01	0.22	0.42	2.40
NC_000016.9:g.31075175G>A	rs2303223	A	*ZNF668*	Synonymous (p.Gly225=)	17.64	0.91	0.04	0.78	0.17	2.29
NC_000017.10:g.2116822G>A	rs17221357	A	*SMG6*	Intron	22.10	0.09	0.01	0.29	0.35	2.27
NC_000023.10:g.136126021G>A	rs7066345	A	-	Intergenic	16.87	0.01	0.01	0.20	0.66	2.19
NC_000022.10:g.22385399G>A	rs117052129	G	*IGLV4-69*	Stop gained (p.Trp3Ter)	34.00	0.94	0.97	0.98	0.06	2.15
NC_000009.11:g.125377734A>G	rs727913	G	*OR1Q1*	Missense (p.Thr240Ala)	25.90	0.26	0.10	0.49	0.17	2.12
NC_000016.9:g.31088625A>G	rs749670	G	*ZNF646*	Missense (p.Glu327Gly)	16.32	0.91	0.04	0.78	0.17	2.11
NC_000005.9:g.118811533G>A	rs25640	A	*HSD17B4*	Missense (p.Arg131His)	33.00	0.49	0.24	0.65	0.10	2.11
NC_000015.9:g.64759279C>T	rs35685348	T	*AC091231.1*	Upstream gene	22.00	0.78	0.09	0.64	0.15	2.11
NC_000002.11:g.97630870G>A	rs186840997	A	*FAM178B*	Intron	9.07	0.02	0.03	0.34	0.65	2.02
NC_000008.10:g.129884159T>C	rs13276570	C	-	Intergenic	20.90	0.76	0.07	0.62	0.15	2.00
NC_000015.9:g.42149506G>C	rs12442525	C	*SPTBN5*	Missense (p.Gln2851Glu)	22.50	0.97	0.33	0.87	0.10	1.99
NC_000011.9:g.5444136C>T	rs2647574	T	*OR51Q1*	Stop gained (p.Arg236Ter)	35.00	0.72	0.50	0.78	0.07	1.99
NC_000009.11:g.125014475C>T	rs1888218	T	*RBM18*	Intron	16.71	0.33	0.07	0.62	0.19	1.99
NC_000011.9:g.115254729T>A	rs75680309	A	*CADM1*	Intron	16.49	0.07	0.03	0.31	0.39	1.98
NC_000004.11:g.27219736A>G	rs13118735	G	*LINC02261*	Intron	21.00	0.41	0.01	0.53	0.18	1.97
NC_000023.10:g.136126139G>T	rs73567910	T	-	Intergenic	15.19	0.01	0.01	0.20	0.66	1.97
NC_000016.9:g.31000809G>A	rs13708	A	*STX1B*	3 prime UTR	18.00	0.92	0.14	0.80	0.14	1.96
NC_000016.9:g.89261482C>A	rs2270416	C	*CDH15*	Stop gained (p.Tyr788Ter)	36.00	0.81	0.94	0.83	0.07	1.96
NC_000001.10:g.152088040C>T	rs79969175	T	*TCHH*	Upstream gene	21.30	0.02	0.01	0.18	0.52	1.96
NC_000009.11:g.73150984C>T	rs6560142	T	*TRPM3*	Missense (p.Arg1670Gln)	33.00	0.82	0.37	0.67	0.09	1.95
NC_000011.9:g.115255842T>C	rs10488708	C	*CADM1*	Intron	17.22	0.07	0.03	0.30	0.37	1.91
NC_000002.11:g.98261636A>G	rs2071038	G	*COX5B*	Upstream gene	8.51	0.03	0.03	0.38	0.59	1.91
NC_000016.9:g.31374535C>G	rs2230429	G	*ITGAX*	Missense (p.Pro517Arg)	24.30	0.73	0.11	0.54	0.14	1.90
NC_000009.11:g.129253264A>G	rs10987302	G	*MVB12B*	Intron	22.10	0.54	0.07	0.59	0.14	1.89
NC_000001.10:g.226555302A>G	rs1136410	G	*PARP1*	Missense (p.Val762Ala)	28.10	0.44	0.01	0.40	0.17	1.87
NC_000015.9:g.38129204T>G	rs1502409	G	-	Intergenic	20.70	0.73	0.14	0.71	0.13	1.86
NC_000016.9:g.90048327G>A	rs45456401	A	*AFG3L1P*	Splice donor	26.70	0.28	0.01	0.39	0.18	1.86
NC_000012.11:g.112477055T>C	rs12231744	C	*NAA25*	Missense (p.Lys876Arg)	23.80	0.61	0.06	0.50	0.15	1.84
NC_000010.10:g.127271548C>A	rs7096815	A	*TEX36-AS1*	Downstream gene	15.93	0.17	0.03	0.43	0.27	1.84
NC_000001.10:g.54606804C>T	rs3766465	T	*CDCP2*	Missense (p.Gly244Arg)	31.00	0.84	0.83	0.92	0.06	1.84
NC_000010.10:g.55955444T>G	rs4935502	G	*PCDH15*	Missense (p.Asp435Ala)	28.40	0.84	0.43	0.78	0.08	1.83
NC_000001.10:g.152192813G>A	rs72477389	A	*HRNR*	Missense (p.Ser431Phe)	15.76	0.01	0.01	0.18	0.63	1.82
NC_000017.10:g.2090215G>A	rs78081565	A	*SMG6*	Intron	17.60	0.08	0.03	0.30	0.35	1.81
NC_000009.11:g.118378191G>T	rs10982846	G	-	Intergenic variant	20.80	0.58	0.14	0.69	0.13	1.80
NC_000004.11:g.167199148A>G	rs1675016	A	-	Intergenic variant	21.20	0.80	0.21	0.73	0.11	1.77
NC_000002.11:g.98270329A>G	rs78168940	G	*ACTR1B*	Downstream gene	8.26	0.03	0.03	0.35	0.60	1.75
NC_000016.9:g.30666367C>T	rs3747481	T	*PRR14*	Missense (p.Pro359Leu)	22.90	0.10	0.86	0.24	0.31	1.74
NC_000016.9:g.59197916A>G	rs7194050	G	*AC092121.1*	Downstream gene	20.70	0.71	0.19	0.72	0.12	1.74
NC_000011.9:g.104763117G>A	rs497116	A	*CASP12*	Stop gained (p.Arg125Ter)	27.00	0.99	0.96	1.00	0.06	1.73
NC_000014.8:g.97272382T>G	rs2224442	G	*VRK1*	Intron	19.81	0.87	0.16	0.63	0.14	1.73
NC_000019.9:g.39307103C>T	rs2229259	C	*ECH1*	Missense (p.Gly217Arg)	32.00	0.95	0.94	0.83	0.07	1.73

Top 50 *FineMAV* hits from the GenomeAsia 100K Northeast Asian (NEA) continental population.

HGVS (hg19/GRCh37)	SNP ID	DER	GENE	CONSEQUENCE	CADD	DAF NEA	DAF SAS	DAF SEA	DAF OCE	DAP	*FineMAV* NEA
NC_000019.9:g.3548231A>G	rs2240751	G	*MFSD12*	Missense (p.Tyr182His)	25.50	0.35	0.01	0.04	0.00	0.66	5.86
NC_000002.11:g.109513601A>G	rs3827760	G	*EDAR*	Missense (p.Val370Ala)	21.70	0.85	0.09	0.53	0.03	0.23	4.23
NC_000016.9:g.31099011T>C	rs11150606	C	*PRSS53*	Missense (p.Gln30Arg)	22.50	0.82	0.08	0.59	0.01	0.23	4.19
NC_000013.10:g.20763642C>T	rs2274084	T	*GJB2*	Missense (p.Val27Ile)	23.10	0.33	0.04	0.04	0.00	0.52	3.94
NC_000016.9:g.21689879T>A	rs78970023	A	*OTOA*	Missense (p.Phe15Tyr)	17.69	0.41	0.06	0.05	0.02	0.44	3.19
NC_000011.9:g.64597201G>A	rs55975541	A	*CDC42BPG*	Missense (p.Arg1237Trp)	32.00	0.20	0.01	0.04	0.01	0.48	3.07
NC_000011.9:g.62848487A>C	rs11231341	C	*SLC22A24*	Stop gained (p.Tyr501Ter)	47.00	0.85	0.81	0.80	0.45	0.07	2.93
NC_000014.8:g.37154111C>T	rs201299512	C	*SLC25A21*	Stop gained (p.Trp208Ter)	45.00	1.00	1.00	1.00	1.00	0.06	2.88
NC_000006.11:g.134385155C>G	rs78562617	G	-	Intergenic	18.66	0.19	0.01	0.01	0.00	0.79	2.87
NC_000003.11:g.4774832C>T	rs750361124	C	*ITPR1*	Stop gained (p.Arg1746Ter)	44.00	1.00	1.00	1.00	1.00	0.06	2.82
NC_000016.9:g.48258198C>T	rs17822931	T	*ABCC11*	Missense (p.Gly180Arg)	24.00	0.92	0.46	0.49	0.12	0.13	2.81
NC_000015.9:g.28228553C>T	rs74653330	T	*OCA2*	Missense (p.Ala481Thr)	24.70	0.12	0.00	0.00	0.00	0.93	2.76
NC_000014.8:g.21500218C>G	rs76101114	C	*TPPP2*	Stop gained (p.Tyr165Ter)	43.00	1.00	1.00	1.00	1.00	0.06	2.76
NC_000002.11:g.109451118A>G	rs72627476	G	*CCDC138*	Intron	13.82	0.84	0.08	0.53	0.03	0.23	2.68
NC_000003.11:g.98031307T>A	rs2316271	A	*OR5H8P*	Stop gained (p.Leu184Ter)	43.00	0.79	0.37	0.62	0.52	0.08	2.64
NC_000008.10:g.145736038T>A	rs143475431	T	*MFSD3*	Stop gained (p.Cys296Ter)	41.00	1.00	1.00	1.00	1.00	0.06	2.63
NC_000014.8:g.51219349G>A	rs61755995	A	*NIN*	Missense (p.Arg1613Cys)	32.00	0.14	0.00	0.03	0.00	0.57	2.58
NC_000015.9:g.28197037T>C	rs1800414	C	*OCA2*	Missense (p.His615Arg)	23.00	0.40	0.02	0.24	0.00	0.28	2.57
NC_000016.9:g.77756501G>T	rs182579196	G	*NUDT7*	Stop gained (p.Glu8Ter)	40.00	1.00	1.00	1.00	1.00	0.06	2.56
NC_000005.9:g.145176022G>A	rs375685870	G	*PRELID2*	Stop gained (p.Arg165Ter)	40.00	1.00	1.00	1.00	1.00	0.06	2.56
NC_000011.9:g.61664711A>T	rs76095489	T	*RAB3IL1*	Downstream gene	19.55	0.35	0.06	0.08	0.00	0.37	2.52
NC_000008.10:g.10555301G>T	rs77073793	T	*C8orf74*	Missense (p.Cys145Phe)	23.30	0.17	0.01	0.02	0.00	0.63	2.52
NC_000017.10:g.76121318G>A	rs12449858	A	*TMC6*	Missense (p.Leu153Phe)	25.10	0.36	0.04	0.08	0.07	0.28	2.50
NC_000016.9:g.31088347G>A	rs749671	A	*ZNF646*	Synonymous (p.Glu234=)	20.30	0.91	0.20	0.73	0.23	0.13	2.47
NC_000016.9:g.49164641T>C	rs1510986	C	-	Intergenic	18.64	0.63	0.24	0.22	0.03	0.20	2.39
NC_000004.11:g.5755658G>T	rs146232611	G	*EVC*	Stop gained (p.Glu488Ter)	37.00	1.00	1.00	1.00	1.00	0.06	2.37
NC_000001.10:g.18808668T>G	rs544927168	T	*KLHDC7A*	Stop gained (p.Leu398Ter)	36.00	1.00	1.00	1.00	1.00	0.06	2.31
NC_000006.11:g.168694840A>T	rs61674641	A	*DACT2*	Stop gained (p.Cys256Ter)	36.00	1.00	1.00	1.00	1.00	0.06	2.31
NC_000002.11:g.68019552C>T	rs6735221	T	*LINC01812*	Downstream gene	20.40	0.19	0.02	0.02	0.00	0.58	2.30
NC_000003.11:g.30209136A>G	rs17025010	G	-	Intergenic	21.20	0.21	0.01	0.03	0.01	0.53	2.29
NC_000012.11:g.133683020C>T	rs2229373	T	*ZNF140*	Missense (p.Ala386Val)	17.47	0.40	0.04	0.15	0.00	0.32	2.29
NC_000003.11:g.135745911G>A	rs16843645	A	*PPP2R3A*	Missense (p.Asp745Asn)	23.00	0.16	0.01	0.02	0.00	0.63	2.27
NC_000017.10:g.7386280G>A	rs7214088	A	*SLC35G6*	Stop gained (p.Trp326Ter)	38.00	0.91	0.75	0.91	0.97	0.07	2.27
NC_000011.9:g.45975130C>T	rs3736508	T	*PHF21A*	Missense (p.Arg347His)	22.50	0.61	0.12	0.41	0.08	0.16	2.25
NC_000003.11:g.151599300G>A	rs953734807	G	*SUCNR1*	Stop gained (p.Trp323Ter)	35.00	1.00	1.00	1.00	0.96	0.06	2.25
NC_000007.13:g.30806001C>T	rs766413713	C	*INMT-FAM188B*	Stop gained (p.Arg134Ter)	35.00	1.00	1.00	1.00	1.00	0.06	2.25
NC_000008.10:g.145640410C>T	rs561005562	C	*SLC39A4*	Stop gained (p.Trp251Ter)	35.00	1.00	1.00	1.00	1.00	0.06	2.25
NC_000019.9:g.16620328C>T	rs3826726	C	*C19orf44*	Stop gained (p.Gln390Ter)	35.00	1.00	1.00	1.00	1.00	0.06	2.24
NC_000016.9:g.31075175G>A	rs2303223	A	*ZNF668*	Synonymous (p.Gly225=)	17.64	0.91	0.20	0.73	0.20	0.14	2.23
NC_000008.10:g.32306158G>A	rs72612108	A	*NRG1*	Intron	17.31	0.62	0.14	0.35	0.00	0.21	2.23
NC_000016.9:g.49283154T>C	rs194416	C	-	Intergenic	20.10	0.65	0.35	0.23	0.02	0.17	2.22
NC_000015.9:g.62932556G>C	rs35757182	G	*AC100839.1*	Intron	35.00	0.99	0.93	1.00	1.00	0.06	2.22
NC_000008.10:g.32345444G>A	rs72612111	A	*NRG1*	Intron	18.14	0.59	0.14	0.34	0.00	0.21	2.20
NC_000016.9:g.4445327C>T	rs3747579	T	*CORO7*	Missense (p.Arg193Gln)	27.80	0.78	0.50	0.56	0.15	0.10	2.19
NC_000002.11:g.213682880C>G	rs76287803	G	*AC093865.1*	Intron	19.71	0.18	0.01	0.02	0.00	0.63	2.19
NC_000003.11:g.167183137G>T	rs1577176453	G	*SERPINI2*	Stop gained (p.Tyr241Ter)	34.00	1.00	1.00	1.00	0.99	0.06	2.18
NC_000002.11:g.27803325C>T	rs530926787	C	*C2orf16*	Stop gained (p.Arg1296Ter)	34.00	1.00	1.00	1.00	1.00	0.06	2.18
NC_000019.9:g.45821185G>C	rs554894916	G	*CKM*	Synonymous (p.Tyr82=)	34.00	1.00	1.00	1.00	1.00	0.06	2.18
NC_000016.9:g.22264427T>G	rs7201033	T	*EEF2K*	Intron	14.37	0.21	0.01	0.01	0.00	0.73	2.18
NC_000019.9:g.22940732C>T	rs148884059	C	*ZNF99*	Stop gained (p.Trp660Ter)	35.00	0.97	1.00	0.99	1.00	0.06	2.17

Top 50 *FineMAV* hits from the GenomeAsia 100K South Asian (SAS) continental population.

HGVS (hg19/GRCh37)	SNP ID	DER	GENE	CONSEQUENCE	CADD	DAF NEA	DAF SAS	DAF SEA	DAF OCE	DAP	*FineMAV* SAS
NC_000016.9:g.31099000C>T	rs201075024	T	*PRSS53*	Missense (p.Gly34Ser)	25.10	0.00	0.45	0.02	0.00	0.88	9.81
NC_000019.9:g.49206674G>A	rs601338	A	*FUT2*	Stop gained (p.Trp154Ter)	52.00	0.04	0.19	0.01	0.01	0.48	4.85
NC_000015.9:g.48426484A>G	rs1426654	A	*SLC24A5*	Missense (p.Thr111Ala)	19.66	0.11	0.49	0.03	0.00	0.49	4.74
NC_000012.11:g.109872909C>T	rs34725387	T	*MYO1H*	Missense (p.His695Tyr)	27.30	0.01	0.22	0.01	0.00	0.71	4.22
NC_000016.9:g.28506428C>T	rs151233	T	*APOBR*	Synonymous (p.Leu22=)	16.22	0.01	0.39	0.04	0.01	0.67	4.19
NC_000002.11:g.104817402A>G	rs4851673	G	-	Intergenic	20.30	0.06	0.37	0.02	0.00	0.54	4.09
NC_000002.11:g.42181679A>T	rs6740960	A	*C2orf91*	Upstream gene	17.73	0.05	0.46	0.05	0.02	0.49	4.00
NC_000016.9:g.30998152T>C	rs200458768	C	*HSD3B7*	Intron	10.65	0.00	0.42	0.02	0.00	0.88	3.90
NC_000016.9:g.28505660G>C	rs151234	C	*CLN3*	NC transcript exon	14.89	0.01	0.39	0.04	0.01	0.66	3.83
NC_000002.11:g.104830710A>T	rs34938541	T	-	Intergenic	18.37	0.06	0.37	0.03	0.00	0.54	3.67
NC_000020.10:g.38797747T>G	rs6071961	G	-	Intergenic	21.80	0.02	0.27	0.02	0.01	0.61	3.59
NC_000001.10:g.18809351G>C	rs137875112	C	*KLHDC7A*	Missense (p.Ala626Pro)	28.00	0.00	0.14	0.00	0.00	0.89	3.56
NC_000004.11:g.152147235C>T	rs371652018	T	*SH3D19*	5 prime UTR	19.35	0.00	0.20	0.01	0.00	0.90	3.55
NC_000001.10:g.51121198T>C	rs11205753	C	*FAF1*	Splice region (p.Val220=)	14.72	0.01	0.29	0.01	0.00	0.80	3.38
NC_000010.10:g.74020044T>C	rs10740396	T	-	Intergenic	22.30	0.05	0.35	0.04	0.03	0.43	3.38
NC_000003.11:g.52867718C>T	rs2276820	T	*TMEM110-MUSTN1*	Missense (p.Gly121Arg)	19.37	0.01	0.23	0.02	0.00	0.73	3.27
NC_000001.10:g.83484014A>G	rs2225576	G	*LINC01362*	Intron	18.48	0.02	0.24	0.01	0.00	0.72	3.22
NC_000015.9:g.64592833T>C	rs114713921	C	*CSNK1G1*	5 prime UTR	17.45	0.00	0.25	0.02	0.00	0.74	3.22
NC_000002.11:g.29152456A>G	rs1140697	G	*WDR43*	Synonymous (p.Glu439=)	19.26	0.12	0.50	0.07	0.03	0.33	3.20
NC_000007.13:g.143747870A>G	rs2961144	G	*OR2A5*	Missense (p.Ile126Val)	23.70	0.03	0.30	0.05	0.01	0.44	3.15
NC_000001.10:g.50576710T>A	rs4357572	A	*ELAVL4*	Intron	20.30	0.08	0.33	0.01	0.01	0.46	3.14
NC_000015.9:g.65042560G>A	rs61741344	A	*RBPMS2*	Synonymous (p.Ile62=)	17.38	0.00	0.22	0.01	0.00	0.81	3.12
NC_000019.9:g.17837512G>A	rs12983721	A	*MAP1S*	Missense (p.Cys440Tyr)	21.60	0.02	0.24	0.03	0.01	0.58	3.02
NC_000001.10:g.37560090G>A	rs11263973	A	-	Intergenic	21.00	0.06	0.35	0.06	0.00	0.41	3.01
NC_000001.10:g.49620445T>C	rs549430	C	*AGBL4*	Intron	19.16	0.08	0.34	0.01	0.01	0.46	2.99
NC_000016.9:g.81242198G>A	rs7499011	A	*PKD1L2*	Stop gained (p.Gln220Ter)	61.00	0.01	0.11	0.02	0.00	0.45	2.99
NC_000015.9:g.64792896G>A	rs640005	G	*ZNF609*	3 prime UTR	15.06	0.00	0.24	0.02	0.00	0.81	2.97
NC_000015.9:g.64653984G>T	rs8026043	G	*AC087632.1*	Intron	15.76	0.00	0.25	0.02	0.00	0.75	2.96
NC_000002.11:g.104788566A>G	rs35135256	G	-	Intergenic	15.65	0.07	0.36	0.02	0.00	0.52	2.96
NC_000001.10:g.49796157A>G	rs1494462	G	*AGBL4*	Intron	19.21	0.08	0.33	0.02	0.01	0.46	2.95
NC_000016.9:g.77359919A>T	rs11640912	A	*ADAMTS18*	Missense (p.Leu626Ile)	23.90	0.12	0.53	0.17	0.06	0.23	2.92
NC_000002.11:g.173846130A>C	rs16861119	C	*RAPGEF4*	Intron	21.70	0.00	0.17	0.01	0.00	0.79	2.91
NC_000016.9:g.82033810G>A	rs11542462	A	*SDR42E1*	Stop gained (p.Gln30Ter)	40.00	0.02	0.14	0.02	0.00	0.54	2.91
NC_000002.11:g.210291373A>G	rs6435527	G	*MAP2*	Intron	14.68	0.04	0.33	0.03	0.00	0.60	2.88
NC_000014.8:g.37154111C>T	rs201299512	C	*SLC25A21*	Stop gained (p.Trp208Ter)	45.00	1.00	1.00	1.00	1.00	0.06	2.88
NC_000002.11:g.173868790C>A	rs12053389	C	*RAPGEF4*	Intron	15.84	0.14	0.55	0.07	0.04	0.33	2.86
NC_000003.11:g.4774832C>T	rs750361124	C	*ITPR1*	Stop gained (p.Arg1746Ter)	44.00	1.00	1.00	1.00	1.00	0.06	2.82
NC_000002.11:g.104642666A>T	rs7568863	A	*LINC01965*	Intron	17.69	0.04	0.28	0.02	0.00	0.56	2.81
NC_000015.9:g.64940203T>C	rs6494484	T	*ZNF609*	Intron	14.24	0.00	0.24	0.02	0.00	0.81	2.79
NC_000001.10:g.171397015C>T	rs6671126	T	-	Intergenic	14.20	0.02	0.29	0.02	0.00	0.69	2.79
NC_000002.11:g.104763415A>G	rs4508618	G	*AC096554.1*	Intron	15.12	0.06	0.35	0.02	0.00	0.52	2.78
NC_000020.10:g.62119717C>T	rs1042796	T	*EEF1A2*	Synonymous (p.Glu442=)	15.56	0.00	0.19	0.00	0.00	0.95	2.78
NC_000011.9:g.62848487A>C	rs11231341	C	*SLC22A24*	Stop gained (p.Tyr501Ter)	47.00	0.85	0.81	0.80	0.45	0.07	2.78
NC_000001.10:g.24417415T>C	rs6700245	C	*MYOM3*	Missense (p.Gln435Arg)	18.40	0.02	0.27	0.02	0.01	0.55	2.77
NC_000015.9:g.91452595A>G	rs2106673	A	*MAN2A2*	Missense (p.Gln412Arg)	18.43	0.07	0.44	0.12	0.00	0.34	2.77
NC_000014.8:g.21500218C>G	rs76101114	C	*TPPP2*	Stop gained (p.Tyr165Ter)	43.00	1.00	1.00	1.00	1.00	0.06	2.76
NC_000009.11:g.594262A>C	rs2641998	C	*KANK1*	Intron	20.80	0.14	0.47	0.08	0.03	0.28	2.76
NC_000002.11:g.210261721G>A	rs11677857	A	-	Intergenic	15.80	0.03	0.28	0.02	0.00	0.62	2.76
NC_000011.9:g.19561284C>G	rs35070300	G	*NAV2*	Intron	18.65	0.00	0.19	0.02	0.00	0.79	2.75
NC_000016.9:g.23313882A>T	rs114177172	T	*SCNN1B*	Intron	18.40	0.00	0.16	0.00	0.00	0.95	2.73

Top 50 *FineMAV* hits from the GenomeAsia 100K Southeast Asian (SEA) continental population.

HGVS (hg19/GRCh37)	SNP ID	DER	GENE	CONSEQUENCE	CADD	DAF NEA	DAF SAS	DAF SEA	DAF OCE	DAP	*FineMAV* SEA
NC_000016.9:g.31099011T>C	rs11150606	C	*PRSS53*	Missense (p.Gln30Arg)	22.50	0.82	0.08	0.59	0.01	0.23	3.03
NC_000014.8:g.37154111C>T	rs201299512	C	*SLC25A21*	Stop gained (p.Trp208Ter)	45.00	1.00	1.00	1.00	1.00	0.06	2.88
NC_000003.11:g.4774832C>T	rs750361124	C	*ITPR1*	Missense (p.Arg1746Gly)	44.00	1.00	1.00	1.00	1.00	0.06	2.82
NC_000014.8:g.21500218C>G	rs76101114	C	*TPPP2*	Stop gained (p.Tyr165Ter)	43.00	1.00	1.00	1.00	1.00	0.06	2.76
NC_000011.9:g.62848487A>C	rs11231341	C	*SLC22A24*	Stop gained (p.Tyr501Ter)	47.00	0.85	0.81	0.80	0.45	0.07	2.74
NC_000002.11:g.109513601A>G	rs3827760	G	*EDAR*	Missense (p.Val370Ala)	21.70	0.85	0.09	0.53	0.03	0.23	2.63
NC_000008.10:g.145736038T>A	rs143475431	T	*MFSD3*	Stop gained (p.Cys296Ter)	41.00	1.00	1.00	1.00	1.00	0.06	2.63
NC_000005.9:g.145176022G>A	rs375685870	G	*PRELID2*	Stop gained (p.Arg165Ter)	40.00	1.00	1.00	1.00	1.00	0.06	2.57
NC_000016.9:g.77756501G>T	rs182579196	G	*NUDT7*	Stop gained (p.Glu8Ter)	40.00	1.00	1.00	1.00	1.00	0.06	2.56
NC_000006.11:g.32632638C>A	rs1130385	A	*HLA-DQB1*	Stop gained (Glu106Ter)	76.00	0.22	0.41	0.40	0.22	0.08	2.43
NC_000004.11:g.5755658G>T	rs146232611	G	*EVC*	Stop gained (p.Glu488Ter)	37.00	1.00	1.00	1.00	1.00	0.06	2.37
NC_000006.11:g.32660661G>A	rs150369468	A	-	Intergenic	15.87	0.02	0.02	0.24	0.01	0.61	2.36
NC_000001.10:g.18808668T>G	rs544927168	T	*KLHDC7A*	Stop gained (p.Leu398Ter)	36.00	1.00	1.00	1.00	1.00	0.06	2.31
NC_000006.11:g.168694840A>T	rs61674641	A	*DACT2*	Stop gained (p.Cys256Ter)	36.00	1.00	1.00	1.00	1.00	0.06	2.31
NC_000006.11:g.32660659C>A	rs142700936	A	-	Intergenic	15.35	0.02	0.02	0.24	0.01	0.61	2.29
NC_000017.10:g.7386280G>A	rs7214088	A	*SLC35G6*	Stop gained (p.Trp326Ter)	38.00	0.91	0.75	0.91	0.97	0.07	2.27
NC_000015.9:g.62932556G>C	rs35757182	G	*AC100839.1*	Intron	35.00	0.99	0.93	1.00	1.00	0.06	2.25
NC_000007.13:g.30806001C>T	rs766413713	C	*INMT-MINDY4*	Stop gained (p.Arg134Ter)	35.00	1.00	1.00	1.00	1.00	0.06	2.25
NC_000008.10:g.145640410C>T	rs561005562	C	*SLC39A4*	Stop gained (p.Trp251Ter)	35.00	1.00	1.00	1.00	1.00	0.06	2.25
NC_000019.9:g.16620328C>T	rs3826726	C	*C19orf44*	Stop gained (p.Gln390Ter)	35.00	1.00	1.00	1.00	1.00	0.06	2.24
NC_000003.11:g.151599300G>A	rs953734807	G	*SUCNR1*	Stop gained (p.Trp323Ter)	35.00	1.00	1.00	1.00	0.96	0.06	2.24
NC_000019.9:g.22940732C>T	rs148884059	C	*ZNF99*	Stop gained (p.Trp660Ter)	35.00	0.97	1.00	0.99	1.00	0.06	2.23
NC_000007.13:g.100371358G>A	rs2293766	A	*ZAN*	Stop gained (p.Trp1883Ter)	52.00	0.43	0.15	0.50	0.50	0.09	2.22
NC_000003.11:g.167183137G>T	rs1577176453	G	*SERPINI2*	Stop gained (p.Tyr241Ter)	34.00	1.00	1.00	1.00	0.99	0.06	2.18
NC_000019.9:g.45821185G>C	rs554894916	G	*CKM*	Stop gained (p.Tyr82Ter)	34.00	1.00	1.00	1.00	1.00	0.06	2.18
NC_000002.11:g.27803325C>T	rs530926787	C	*C2orf16*	Stop gained (p.Arg1296Ter)	34.00	1.00	1.00	1.00	1.00	0.06	2.18
NC_000006.11:g.32678064C>T	rs113556552	T	*MTCO3P1*	Upstream gene	14.62	0.02	0.02	0.24	0.01	0.61	2.17
NC_000022.10:g.22385399G>A	rs117052129	G	*IGLV4-69*	Stop gained (p.Trp3Ter)	34.00	0.94	0.98	0.99	1.00	0.06	2.15
NC_000001.10:g.152057802G>A	-	G	*TCHHL1*	Stop gained (p.Gln786Ter)	33.00	1.00	1.00	1.00	1.00	0.06	2.12
NC_000004.11:g.438045T>A	rs777532496	T	*ZNF721*	Stop gained (p.Lys71Ter)	33.00	1.00	1.00	1.00	1.00	0.06	2.12
NC_000004.11:g.98552931A>G	rs188128811	G	*STPG2*	Intron	20.50	0.00	0.00	0.17	0.03	0.59	2.11
NC_000004.11:g.101770683G>A	rs182265527	A	*EMCN*	Intron	17.18	0.00	0.00	0.16	0.01	0.75	2.06
NC_000003.11:g.98031307T>A	rs2316271	A	*OR5H8*	Stop gained (p.Leu184Ter)	43.00	0.79	0.37	0.62	0.52	0.08	2.06
NC_000005.9:g.170236616C>T	rs79997355	T	*GABRP*	Missense (p.Arg293Cys)	34.00	0.02	0.04	0.12	0.70	0.52	2.06
NC_000016.9:g.20499710C>T	rs79632868	C	*ENSG00000267824*	Downstream gene	33.00	1.00	0.96	0.97	0.95	0.06	2.06
NC_000006.11:g.32024552C>T	rs202211608	T	*TNXB*	Missense (p.Glu2652Lys)	23.20	0.00	0.01	0.12	0.01	0.73	2.05
NC_000003.11:g.150106188A>C	rs12638326	C	-	Intergenic	20.20	0.22	0.05	0.49	0.09	0.21	2.05
NC_000022.10:g.44495983A>G	rs11539650	G	*PARVB*	Missense (p.Lys118Glu)	23.60	0.01	0.01	0.18	0.03	0.47	2.04
NC_000015.9:g.45392075G>A	rs269868	A	*DUOX2*	Missense (p.Ser1067Leu)	24.10	0.85	0.96	0.91	0.22	0.09	2.02
NC_000006.11:g.154360569C>T	rs17174638	C	*OPRM1*	Stop gained (p.Gln57Ter)	31.00	1.00	1.00	1.00	0.93	0.06	1.99
NC_000016.9:g.31088347G>A	rs749671	A	*ZNF646*	Synonymous (p.Glu234=)	20.30	0.91	0.20	0.73	0.23	0.13	1.98
NC_000008.10:g.71646084C>T	rs115507207	T	*XKR9*	Stop gained (p.Gln183Ter)	44.00	0.01	0.02	0.09	0.00	0.50	1.97
NC_000016.9:g.89261482C>A	rs2270416	C	*CDH15*	Stop gained (p.Tyr788Ter)	36.00	0.76	0.95	0.82	0.92	0.07	1.93
NC_000006.11:g.32663243A>G	rs111940765	G	-	Intergenic	13.08	0.02	0.02	0.24	0.01	0.61	1.93
NC_000004.11:g.89363604C>T	rs4413373	C	*HERC6*	Stop gained (p.Gln1021Ter)	30.00	1.00	1.00	1.00	1.00	0.06	1.93
NC_000007.13:g.25924084T>C	rs17152880	C	-	Intergenic	20.80	0.25	0.06	0.42	0.01	0.22	1.91
NC_000006.11:g.32587530A>G	rs369095149	G	-	Intergenic	12.68	0.02	0.03	0.27	0.01	0.55	1.89
NC_000001.10:g.54606804C>T	rs3766465	T	*CDCP2*	Missense (p.Gly244Arg)	31.00	0.88	0.86	0.91	0.99	0.06	1.83
NC_000016.9:g.69354963A>G	rs1127231	A	*VPS4A*	Synonymous (p.Lys287=)	21.10	0.84	0.67	0.84	0.11	0.10	1.81
NC_000017.10:g.48557348T>C	rs2290861	T	*RSAD1*	Missense (p.Leu126Ser)	22.10	0.74	0.57	0.76	0.07	0.11	1.80

Top 50 *FineMAV* hits from the GenomeAsia 100K Oceanian (OCE) continental population.

HGVS (hg19/GRCh37)	SNP ID	DER	GENE	CONSEQUENCE	CADD	DAF NEA	DAF SAS	DAF SEA	DAF OCE	DAP	*FineMAV* OCE
NC_000019.9:g.52004903G>A	rs16982743	A	*SIGLEC12*	Stop gained (p.Gln29Ter)	35.00	0.03	0.10	0.10	0.78	0.46	12.55
NC_000005.9:g.170236616C>T	rs79997355	T	*GABRP*	Missense (p.Arg293Cys)	34.00	0.02	0.04	0.12	0.70	0.52	12.53
NC_000012.11:g.56495023G>A	rs2271188	A	*ERBB3*	Missense (p.Arg1127His)	29.10	0.01	0.00	0.03	0.53	0.79	12.13
NC_000008.10:g.16859307T>C	rs377326763	C	*FGF20*	Missense (p.Ile79Val)	25.60	0.00	0.01	0.05	0.58	0.76	11.28
NC_000011.9:g.116469150A>T	rs193134517	T	-	Intergenic	15.84	0.00	0.00	0.02	0.72	0.91	10.29
NC_000001.10:g.40776388G>A	rs370064150	A	*COL9A2*	Missense (p.Pro228Ser)	27.10	0.00	0.00	0.02	0.42	0.89	10.16
NC_000015.9:g.45402692G>A	rs377608586	A	*DUOX2*	Missense (p.Ser325Phe)	29.80	0.00	0.00	0.00	0.34	0.99	10.14
NC_000001.10:g.186140508G>A	rs150188026	A	*HMCN1*	Missense (p.Arg5205His)	19.06	0.01	0.01	0.04	0.68	0.78	10.10
NC_000019.9:g.46974003C>T	rs7248888	T	*PNMA8A*	Missense (p.Cys97Tyr)	22.00	0.01	0.01	0.03	0.59	0.77	10.02
NC_000010.10:g.17145204T>C	rs750735519	C	*CUBN*	Missense (p.Ser484Gly)	23.10	0.00	0.00	0.01	0.47	0.91	9.83
NC_000006.11:g.138196957A>C	rs141807543	C	*TNFAIP3*	Missense (p.Ile207Leu)	22.00	0.00	0.00	0.06	0.59	0.74	9.60
NC_000001.10:g.208397575T>C	rs17259450	C	*PLXNA2*	Intron	20.20	0.02	0.03	0.03	0.66	0.72	9.51
NC_000022.10:g.30642644G>A	rs201805161	A	*AC004264.1*	NC transcript exon	16.08	0.01	0.01	0.04	0.75	0.78	9.39
NC_000002.11:g.76927928A>G	rs72915629	G	-	Intergenic	20.80	0.00	0.00	0.03	0.54	0.82	9.24
NC_000006.11:g.138229771G>A	rs373854868	A	-	Intergenic	21.00	0.00	0.00	0.06	0.59	0.74	9.16
NC_000006.11:g.154479929C>A	rs951667384	A	*IPCEF1*	3 prime UTR	19.89	0.00	0.01	0.01	0.50	0.88	8.79
NC_000009.11:g.83255484G>T	rs913504	T	-	Intergenic	20.30	0.00	0.04	0.04	0.61	0.70	8.73
NC_000017.10:g.48658282A>G	rs198553	G	*CACNA1G*	Intron	16.14	0.03	0.04	0.05	0.81	0.65	8.48
NC_000006.11:g.137793302C>T	rs376613871	T	-	Intergenic	21.40	0.00	0.00	0.04	0.49	0.80	8.42
NC_000006.11:g.153453344T>A	rs372597711	A	*RGS17*	Upstream gene	20.50	0.00	0.00	0.02	0.47	0.87	8.41
NC_000011.9:g.61557979G>A	rs369511941	A	*TMEM258*	Synonymous (p.Thr33=)	17.25	0.00	0.00	0.02	0.55	0.88	8.40
NC_000013.10:g.72667695A>G	rs373350507	G	-	Intergenic	20.30	0.00	0.00	0.01	0.44	0.92	8.22
NC_000006.11:g.138226361A>C	rs376195905	C	-	Intergenic	20.00	0.00	0.00	0.06	0.55	0.74	8.12
NC_000001.10:g.208989631T>C	rs187254348	C	-	Intergenic	15.86	0.01	0.01	0.06	0.69	0.73	8.03
NC_000003.11:g.169378799G>A	rs73032054	A	*MECOM*	Intron	14.43	0.00	0.03	0.02	0.68	0.81	8.00
NC_000001.10:g.185255201T>G	rs374745720	G	*SWT1*	Intron	14.11	0.00	0.03	0.03	0.74	0.77	7.97
NC_000021.8:g.15942551T>A	rs377229395	A	*SAMSN1*	Intron	17.65	0.00	0.00	0.01	0.48	0.93	7.89
NC_000012.11:g.56538344T>C	rs58663297	C	*ESYT1*	3 prime UTR	18.25	0.01	0.00	0.03	0.53	0.80	7.79
NC_000012.11:g.56548150T>A	rs79606241	A	*MYL6B*	Intron	18.56	0.01	0.00	0.05	0.57	0.73	7.74
NC_000001.10:g.95392451G>C	rs144969776	C	*CNN3*	5 prime UTR	21.70	0.00	0.00	0.02	0.41	0.87	7.69
NC_000008.10:g.16861280T>C	rs140142364	C	*FGF20*	Upstream gene	18.51	0.00	0.01	0.05	0.55	0.75	7.68
NC_000016.9:g.70500801C>T	rs572099494	T	*FUK*	Missense (p.Pro143Leu)	31.00	0.00	0.00	0.01	0.26	0.94	7.66
NC_000010.10:g.78391460G>A	rs374777852	A	-	Intergenic	21.30	0.00	0.00	0.01	0.38	0.95	7.62
NC_000008.10:g.105266371G>T	rs57490000	T	*RIMS2*	3 prime UTR	19.64	0.02	0.01	0.05	0.57	0.68	7.60
NC_000012.11:g.14664250A>G	rs2287541	G	*PLBD1*	Missense (p.Val377Ala)	24.80	0.03	0.12	0.10	0.73	0.42	7.59
NC_000008.10:g.105360994C>A	rs61682032	A	*DCSTAMP*	Missense (p.Leu72Met)	21.40	0.02	0.02	0.04	0.53	0.66	7.52
NC_000010.10:g.78499696A>G	rs374415709	G	-	Intergenic	21.40	0.00	0.00	0.01	0.37	0.94	7.51
NC_000002.11:g.59688550T>G	rs189747995	G	*AC007179.2*	Intron	17.38	0.00	0.02	0.03	0.56	0.77	7.49
NC_000003.11:g.169500397T>C	rs35406871	C	*MYNN*	Synonymous (p.Ser455=)	17.80	0.01	0.02	0.06	0.62	0.68	7.47
NC_000016.9:g.1595600C>T	rs56342298	T	*TMEM204*	Intron	18.63	0.03	0.07	0.15	0.84	0.47	7.47
NC_000012.11:g.92397023G>A	rs11106391	A	*LINC01619*	Intron	17.11	0.00	0.01	0.03	0.53	0.81	7.44
NC_000016.9:g.73646124T>C	rs7197725	C	-	Intergenic	18.87	0.00	0.00	0.02	0.46	0.86	7.43
NC_000015.9:g.55722882C>A	rs57809907	A	*DNAAF4*	Stop gained (p.Glu417Ter)	43.00	0.01	0.07	0.03	0.38	0.46	7.43
NC_000018.9:g.60193406A>T	rs72941625	T	*ZCCHC2*	Intron	12.31	0.01	0.03	0.01	0.73	0.83	7.42
NC_000015.9:g.67692059T>A	rs77919550	A	*IQCH*	Intron	17.41	0.01	0.02	0.06	0.63	0.68	7.38
NC_000003.11:g.5020036C>T	rs367938373	T	*BHLHE40-AS1*	Intron	22.60	0.00	0.00	0.01	0.36	0.89	7.34
NC_000012.11:g.17653296T>G	rs140229689	G	-	Intergenic	17.22	0.00	0.00	0.04	0.55	0.78	7.33
NC_000012.11:g.56516569C>G	rs57280585	G	*AC034102.6*	Intron	17.17	0.01	0.00	0.03	0.53	0.80	7.33
NC_000009.11:g.127140816C>G	rs544483898	G	*PSMB7*	Intron	20.70	0.00	0.00	0.01	0.40	0.89	7.31
NC_000003.11:g.122474121G>C	rs61756481	C	*HSPBAP1*	Missense (p.Leu243Val)	24.10	0.00	0.01	0.01	0.35	0.86	7.31