The investigation of the signature of selection on genes associated with dyslexia of Chinese characters

西山　久美子

博士（理学）

総合研究大学院大学
先導科学研究科
生命共生体進化学専攻

令和２（２０２０）年度

The investigation of the signature of selection on genes associated with dysleхia of Chinese characters

Kumiko Nishiyama

Doctor of Philosophy

Department of Evolutionary Studies of Biosystems
School of Advanced Sciences
SOKENDAI (The Graduate University for Advanced Studies)

June 2020

Summary

Dyslexia, or reading disability, is found to have a genetic basis, and several related genes have been reported. Writing systems were developed just a few thousand years ago and used by only a limited number of people before modern times. So, in old times, dyslexic people would have lived without the reading difficulties or disadvantages that are present in modern society. Based on this viewpoint, alleles related to reading ability are expected to be under neutral evolution. Otherwise, if natural selection has acted on such alleles, its target would be traits other than reading ability itself. The question in the study is whether natural selection has acted on the alleles of single nucleotide polymorphisms (SNPs) that were reported to be risk/non-risk for reading ability. In this thesis, I focused on 15 SNPs that were found to be associated with dyslexia of Chinese characters in previous studies ("core SNPs", hereafter). Using publicly available databases, I applied two types of summary statistics (*nSL* and 2D SFS-based statistics) to SNP data of East Asian populations, in order to examine whether there is any sign of selective sweep. Because core SNPs themselves are not necessarily causal, and instead, the causal site may be a site that is tightly linked to a reported SNP, I also checked such linked SNPs in this study, considering that they also could be selection targets.

The findings of my study are shown in chapter 3. In my study, first, I conducted a brief study of principal component analysis (PCA) to confirm genetic background of the study populations. Second, as a neutrality test based on extended haplotype homozygosity, I applied *nSL* to the core SNPs. *nSL* did not detect any signatures of positive selection for any of the core SNPs. Third, as a neutrality test based on site

frequency spectrum (SFS), I conducted 2D SFS-based statistics, which are main analyses of my study.

2D SFS-based statistics measure the level of polymorphism within haplotypes carrying the derived (i.e., mutant) allele of a focal site. In the first section of 2D SFS-based statistics, I carried out screening of core SNPs using F_c from 2D SFS-based statistics. This examines whether a high linkage disequilibrium (LD) region containing a core SNP (core region) could be under selective sweep. I considered that if a linked SNP showed a similar number of derived alleles in EAS but it showed a different number from the core SNP when looking at the global population, the level of polymorphism should be different between the core SNP and such a linked SNP due to the difference in age. For such linked SNPs, it is inappropriate to apply the statistic to its core SNP. Thus, I checked the global derived allele count (number of derived alleles in populations worldwide) of every core SNP and its linked SNPs before applying F_c. I found that three core SNPs had some linked SNPs with global derived allele counts smaller than that of their core SNPs. In each of these three cases, I applied F_c statistic to a linked SNP that showed the smallest derived allele count as the "younger SNP", as well as core SNPs. Then, two SNPs remained after this screening ($p < 0.1$): rs17031962 on *GNPTAB* and rs3789228 (younger SNP for rs1091047) on *DCDC2*.

In the second section of 2D SFS-based statistics, I analyzed the two core regions that passed the screening, in order to search for the target site of natural selection. In this analysis, I compared the level of polymorphism around each of the candidate SNPs (core SNP and its linked SNPs) in the core region, using G_{c0} from 2D SFS-based statistics.

The first case identified was the core region of rs17031962 on *GNPTAB*. Notably, this region contains genes of *CHPT1* (partial) and *SYCP3*, in addition to *GNPTAB*. I searched for the target site and inferred that the target site could be a SNP (rs3751248), because the level of polymorphism around the SNP was significantly low, and it turns out to be located in an open chromatin region. The second case identified was the core region of rs3789228 on *DCDC2*. This SNP was distinguished as the younger SNP to the core SNP of rs1091047, based on global derived allele count. An SNP (rs12055879) is located in an enhancer region with CTCF binding site and showed significantly low level of polymorphism. Thus, I inferred that the target site could be this SNP. After 2D SFS-based statistics, I also investigated the phylogenetic relationship of haplotypes in global populations about the two cases.

In the general discussion of chapter 4, I considered the results of my study. For most of the core SNPs, both *nSL* and 2D SFS did not detect any signatures of selective sweep. Because most people were not engaged in reading and writing until recently, the genetic variations that my study focused on were unlikely to be maintained by natural selection, which is consistent with my results. Nevertheless, 2D SFS-based statistics suggested that two core regions could be under selective sweep. In both regions, I found candidate target sites which may have an effect on expression regulation. However, which genes these SNPs affect remain unknown; if the target sites have a functional effect not on *GNPTAB* or *DCDC2* but on other genes, my results suggest a possibility of genetic hitchhiking, whereby alleles of the reported SNPs may have increased in frequency together with the selected target, which could have functions for other genes and traits apart from reading ability. This possibility is also consisted with the reasoning that there has not been selective pressure on reading ability itself.

Now, dyslexia is considered as one of neurodevelopmental disorders. Although it is not directly related to the analyses in this study, through this study, I tried to think about the notion of disorders from the perspective of evolutionary studies. Because neurodevelopmental disorders have genetic basis, I think that the viewpoint of human evolutionary history would be meaningful for considering the notion of neurodevelopmental disorders. The focus of my study was polymorphisms that were reported to be associated with risk/non-risk for dyslexia. Modern society has introduced public education and demands universal literacy. So, primarily, the environment of the modern society likely determines which allele is "risk" or "non-risk" for dyslexia. Dyslexia should basically be a consequence of neutral variation. Even in the case where selection may have acted, the selected trait in human evolution should be different from reading ability itself. I hope that my study could provide an opportunity to be aware that modern society should be only a temporal environment in human evolutionary history.

Contents

Chapter 1

General introduction

1.1. Background

1.1.1. Overview of dyslexia

In modern society, people are required literacy for their lives, and learn how to read and write in school. Dyslexia, or reading disability, is generally defined as a difficulty in reading and writing despite normal intelligence and appropriate opportunity for education (Grigorenko 2001; Paracchini *et al.* 2007; Scerri and Schulte-Körne 2010). Reports of dyslexia appeared at first in the late 19th century in Germany and the UK (Kirby 2018), where people use European alphabets. Today, dyslexia is observed among various writing systems with, for example, Arabic, Chinse, and Indic scripts (Daniels and Share 2018).

Dyslexia is considered to be caused by neurobiological and neurocognitive differences (Peterson and Pennington 2012). This is supported by many neuroimaging studies using techniques based on magnetic resonance imaging (MRI) (Mascheretti *et al.* 2017). These studies have found in dyslexics, for example, different activations of language network in a left hemisphere, or a disconnection between posterior auditory processing areas and anterior motor planning areas (Peterson and Pennington 2015). The left occipitotemporal cortex is most consistently shown to contribute reading (Dehaene and Cohen 2007, 2011; Price and Devlin 2011; Martin *et al.* 2016; Protopapas and Parrila

2018). However, findings in MRI-based studies are heterogeneous, and so it is difficult to draw general conclusion (Mascheretti *et al.* 2017).

There is no universal criteria for diagnosis of dyslexia (Fisher and DeFries 2002; Paracchini *et al.* 2007), and the definition varies across writing systems (McBride *et al.* 2018). Now, dyslexia is usually diagnosed using psychometric measures of reading and writing, even though the measures are various (Fisher and DeFries 2002; Paracchini *et al.* 2007; Carrion-Castillo *et al.* 2013). Then, when an individual's score falls below a cutoff in the normal distribution, the individual is regarded as dyslexia (Peterson and Pennington 2015; Bishop 2015; Protopapas and Parrila 2018).

Continuously distributed traits, including reading ability, are considered to be polygenic traits (Plomin *et al.* 2009; Bishop 2015). Indeed, several genes have been reported to be related to dyslexia to date, for example, *DYX1C1*, *DCDC2*, *KIAA0319*, *ROBO1* (Paracchini *et al.* 2007; Scerri and Schulte-Körne 2010; Carrion-Castillo *et al.* 2013; Newbury *et al.* 2014; Kere 2014; Peterson and Pennington 2015). These genes could play roles in brain development of prenatal period, especially in neuronal migration, growth and function of cilia, or regulation of axonal and dendritic growth (Carrion-Castillo *et al.* 2013; Kere 2014; Peterson and Pennington 2015). Many studies have sought particular single nucleotide polymorphisms (SNPs) in dyslexia related genes, whose alleles show associations with some dimensions of reading/writing ability (i.e. risk alleles) (e.g., Taipale *et al.* 2003; Francks *et al.* 2004; Harold *et al.* 2006; Bates *et al.* 2010; Lind *et al.* 2010; Paracchini *et al.* 2011; Scerri *et al.* 2011).

1.1.2. Dyslexia in Chinese populations

Among various writing systems in the world, Chinese characters showed the earliest form around 1200 BCE, and have also been used at least once during history in other East Asian countries (e.g., Japan, Korea, and Vietnam), where spoken language systems are different from China (Hansell 2003). Chinese characters have distinct features: Most characters are visually complex because they are compound characters, it contains semantic radicals, and thousands of characters exist (McBride 2016; Daniels and Share 2018). Neurological studies showed that the brain areas involved in dyslexia are different between English and Chinese characters: reductions in gray matter volume in left temporoparietal and occipitotemporal regions was found for dyslexia of English; reductions in gray matter volume in left middle frontal gyrus region was found for dyslexia of Chinese characters (Siok *et al.* 2004, 2008; Hoeft *et al.* 2007).

While genetic research on dyslexia was initially conducted in populations that use alphabetic languages, genetic factors of dyslexia in Chinese populations have been investigated in the last decade (Su *et al.* 2015; Sun *et al.* 2017; Sharma and Sagar 2017). These studies found several SNPs with risk/non-risk alleles associated with some measures of reading (and writing) ability of Chinese characters (Table 1-1). In these SNPs, similar associations were found in preceding studies of populations using alphabetic languages (e.g., rs807724 on *DCDC2*), although alleles for risk or non-risk are not always the same between the populations studied, as found in rs4504469 on *KIAA0319* (Shao *et al.* 2016a, 2016b) and rs1091047 on *DCDC2* (Su *et al.* 2015). Among the reported SNPs, biological functions were experimentally investigated for rs3743205 on *DYX1C1* and rs1079727 on *DRD2* (Taipale *et al.* 2003; Kaalund *et al.* 2014). However, for most of the SNPs, their effects on biological function are unknown, and these SNPs themselves are not necessarily causal. Instead, the causal site may be a site that is tightly linked to a reported SNP (Balding 2006).

Table 1-1. The single nucleotide polymorphisms (SNPs) associated with dyslexia of Chinese characters in previous studies.

Gene	Core SNP	Chr.	Position	Risk Allele	Derived Allele Frequency		References
			(GRCh37/hg19)		EAS	(EAS and KPGP)	
KIAA0319L	rs28366021	1	36,022,859	Ancestral	0.234	(0.227)	(Shao *et al.* 2016a)
ROBO1	rs4535189	3	79,489,971	Derived	0.366	(0.373)	(Sun *et al.* 2017)
DCDC2	rs807724	6	24,278,869	Ancestral	0.957	(0.956)	(Zhang *et al.* 2016)
DCDC2	rs1091047	6	24,295,256	Ancestral	0.817	(0.823)	(Su *et al.* 2015)
KIAA0319	rs2760157	6	24,578,272	Ancestral	0.456	(0.470)	(Lim *et al.* 2014)
KIAA0319	rs807507	6	24,579,867	Derived	0.188	(0.187)	(Lim *et al.* 2014)
KIAA0319	rs4504469	6	24,588,884	Derived	0.112	(0.122)	(Shao *et al.* 2016a)
DOCK4	rs2074130	7	111,487,098	Derived	0.101	(0.115)	(Shao *et al.* 2016a)
DRD2	rs1079727	11	113,289,182	Derived	0.416	(0.420)	(Chen *et al.* 2014)
GNPTAB	rs17031962	12	102,146,558	Ancestral	0.294	(0.297)	(Chen *et al.* 2015)
DYX1C1	rs11629841	15	55,777,638	Derived	0.058	(0.056)	(Zhang *et al.* 2012)
DYX1C1	rs3743205	15	55,790,530	Derived	0.035	(0.037)	(Lim *et al.* 2011)
intergenic region	rs8049367	16	3,980,445	Derived	0.339	(0.340)	(Wang *et al.* 2017)
NAGPA	rs882294	16	5,092,118	Derived	0.189	(0.188)	(Chen *et al.* 2015)
DIP2A	rs2255526	21	47,971,539	Derived	0.264	(0.262)	(Kong *et al.* 2016)

1.1.3. Evolutionary perspective on reading/writing ability

From the perspective of human evolution, reading and writing are quite new activities, and have different histories to that of speaking. Writing systems were developed just a few thousand years ago and used by only a limited number of people before modern times; therefore, reading ability is unlikely to have been shaped by natural selection (Dalby 1986; Dehaene and Cohen 2007; Christiansen and Müller 2015; d'Errico and Colagè 2018). Dyslexia may be due to genetically based neurological variations that were not obstacles to humans until the introduction of public education in the 19th century (Dalby 1986); before this time, dyslexic people would have lived without the reading difficulties/disadvantages that are present in modern society (Protopapas and Parrila 2018, 2019). Based on this viewpoint, alleles related to reading ability are expected to be under neutral evolution. Otherwise, if natural selection has acted on such alleles, its target should be traits other than reading ability itself.

If natural selection has acted, at least two scenarios can be considered. The first scenario is proposed as the neuronal recycling hypothesis (Dehaene 2005; Dehaene and Cohen 2007) or cultural neural reuse (d'Errico and Colagè 2018; Colagè and D'Errico 2020). This is somewhat similar to the concept of exaptation, and explains the development of reading activity in humans as follows: An individual reuses a specific region of his/her brain, which functioned for something other than reading in the evolutionary past, such as face recognition (Dehaene 2005; Dehaene and Cohen 2007; d'Errico and Colagè 2018; Colagè and D'Errico 2020). Natural selection can act on such prior functions, and in this case, a non-risk allele for dyslexia is expected to be the allele selected for the prior functions. The second scenario is pleiotropy, whereby a gene is involved in more than one function (Stearns 2010; Paaby and Rockman 2013; Dediu and Christiansen 2016). Thus, a locus could be selected not for functions related to reading itself but for other functions (Mozzi *et al.* 2016); even alleles with risk for dyslexia could be selected if the risk alleles have an advantage for other functions.

Evolution of dyslexia-related genes has been investigated by comparing sequences of primates, which found a change in selective pressure on *ROBO1* after the divergence of the orangutan (Hannula-Jouppi *et al.* 2005) and signs of positive selection on *KIAA0319* in the human lineage (Mozzi *et al.* 2016). Some sites on *ROBO1*, *ROBO2*, and *CNTNAP2* showed signatures of selective sweeps among modern human populations, where the derived alleles significantly increased in frequency after the separation from archaic hominins, although they do not reach fixation (Mozzi *et al.* 2016). As mentioned above, several sites on dyslexia-related genes were found to have risk/non-risk alleles associated with reading ability, although evolutionary analyses in these previous studies (Hannula-Jouppi *et al.* 2005; Mozzi *et al.* 2016) did not focus on such alleles.

1.2. The aim of this study

The question in the present study is whether natural selection has acted on the alleles of SNPs that were reported to be risk/non-risk for reading ability. It is expected that there would not have been selective pressure on an individual's reading ability. Moreover, it is more unlikely that alleles of the SNPs related to the reading ability of a certain writing system were selected especially for features of the writing system; the time for adaptation to a writing system to occur is probably insufficient (Dalby 1986; Dehaene and Cohen 2007; Christiansen and Müller 2015; d'Errico and Colagè 2018). By examining East Asian populations, I investigated whether alleles of the SNPs associated with dyslexia of Chinese characters had evolved neutrally or not. Although my focus was the reading ability of Chinese characters, I also considered that genes associated with dyslexia of Chinese characters could be selected for their other functions as in pleiotropy.

For this aim, I performed neutrality tests on the SNPs associated with the reading/writing ability of Chinese characters (Table 1-1). Because each type of neutrality test would have its suitable time scale to detect the signature of selection (Sabeti 2006), I used two different types of summary statistics: Number of segregating sites by length (*nSL*) (Ferrer-Admetlla *et al.* 2014) and Two-dimensional site frequency spectrum (2D SFS) (Fujito *et al.* 2018b; Satta *et al.* 2019), which are based on extended haplotype homozygosity (EHH) and the site frequency spectrum (SFS), respectively. EHH-based statistics, such as *nSL*, are powerful at detecting signs of recent selective sweep, where linkage disequilibrium (LD) is expected to be relatively maintained (Sabeti 2006). Meanwhile, 2D SFS-based statistics can detect sweep signals in regions that have experienced recombination events over time and result in being with short LD. Both *nSL* and 2D SFS are developed basically to detect selection signal on derived alleles. So in this study, I focused on derived alleles of the SNPs regardless of whether they are risk or non-risk for reading ability.

I also considered SNPs that were tightly linked to the SNPs associated with reading/writing ability (Table 1-1), because they also could be causal for reading ability or have other functional effects, and therefore could be selection targets. In such cases, a reported SNP may be considered a hitchhiker of a tightly linked SNP that is under selection. To search for the selection targets, I analyzed in detail the LD regions that contain the candidate SNPs under selection. The attempt at searching the selection targets would be a characteristic of this thesis. I tried to infer the selection targets by comparing the level of polymorphism around each of the candidate SNPs in the LD region. This is based on the expectation that the level of polymorphism would be low around the selection target as a result of selective sweep and this level would increase with distance from the selection target. If a locus has been selected, the locus may have a biological function. I tried to search such crucial locus in human genome, by analyzing sequence data. I hope that this attempt could be contributed to the understanding of human evolution, and practically, could provide some assistance for experimental studies to confirm the functions.

1.3. The meaning of my study in social context

Although it is not directly related to the analyses in this thesis, the ultimate question of my thesis is what is a “disorder”. Through this thesis, I tried to think about the notion of disorders from the perspective of evolutionary studies.

About a certain trait of an organism, no difference in fitness may be observed among individuals with any phenotypes in some environments. When the environments change, a particular phenotype can be adaptive/maladaptive, and individuals with this phenotype can increase/decrease their fitness. This should be basic understanding in evolutionary studies. Any environments will change; modern society should be only a temporal environment in human evolutionary history.

Today, some phenotypes, such as autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD) and dyslexia, are categorized in neurodevelopmental disorders (American Psychiatric Association 2013; Thapar *et al.* 2017; Protopapas and Parrila 2018). According to the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) (American Psychiatric Association 2013), a person with neurodevelopmental disorders shows "developmental deficits that produce impairments of personal, social, academic, or occupational functioning". The disorders are caused by impairments or disruptions in the brain development (Meredith 2015; Thapar *et al.* 2017). Neurodevelopmental disorders is found to have a genetic basis (Meredith 2015; Thapar and Rutter 2015). It is also for dyslexia, and many researches have sought the risk alleles as mentioned above.

Neurodevelopmental disorders have been generally thought to be cured, eradicated, or prevented. Recently, this attitude have been challenged by the neurodiversity movement. The neurodiversity movement appeared in 1990's among autistic people, and it has been developed primarily as autism self-advocacy movement (Ortega 2009). It has expanded involving those diagnosed with other neurodevelopmental disorders including dyslexia (Griffin and Pollak 2009). In the concept of neurodiversity, a neurodevelopmental disorder such as autism is considered as "a natural variation among humans" (Jaarsma and Welin 2012), or "an example of diversity in the set of all possible brains" (Baron-Cohen 2017). The neurodiversity movement have demanded the rights, recognition and acceptance for certain neurological conditions, objecting to the notion of "to be cured" (Jaarsma and Welin 2012).

I mentioned above that neuroimaging studies have shown neural differences which are considered to cause dyslexia. Protopapas and Parrila (2018) questioned the interpretation of the results seen in neuroimaging studies. While the authors acknowledge these studies to find the details of the differences in brains between good and poor readers as scientific information, they contend that “differences” is not equal to “being wrong”. They argue that it is natural that different brain activities are observed if different reading performances are observed; it does not mean developmental failure in the brains of poor readers. Differences in brains will exist whenever differences in behavior exist; it should be also seen in singing, dancing, or playing chess, for example, although individuals with poor performance on them do not regard as disorders (Protopapas and Parrila 2018). They suggest that what modern society values should make some behavioral differences disorders. These authors are researchers in cognitive science field and do not refer to the term of neurodiversity in this article. However, their argument seems to be compatible with the notion of neurodiversity.

My study is also not related to the neurodiversity movement, but I also attempted to consider what is “neurodevelopmental disorders”. Because these disorders have genetic basis, I think that the viewpoint of human evolutionary history, beyond the context within present day, would be meaningful for considering the notion of neurodevelopmental disorders. In order to consider it, I decided to focus dyslexia in my thesis. Even if reading ability is valued in modern society, most people were not engaged in reading until recently. Therefore, it would be obvious that reading ability is unlikely to have been shaped by natural selection, described above. If so, is it appropriate to explain differences in reading/writing performances as a “disorder”? what are “risk/non-risk alleles" for dyslexia?

I know that it could be problematic to bring the perspective of evolutionary studies to human society, seen as eugenics. Researchers who use human genomic data are (should be) highly careful of how the technologies in their research area or their results could impact on or be interpreted by society (Juengst 2009; Weiss and Lambert 2011; Vitti *et al.* 2012; Yudell *et al.* 2016; Coller 2019). With recognizing such ethical issues, I would like to consider whether and how evolutionary studies could relate to human society. It should be difficult to answer to the question of what is a "disorder" only from this study, but I would like to struggle over this question through my thesis.

Chapter 2

Materials and Methods

2.1. Examined SNPs

I focused on 15 SNPs that were found to be associated with dyslexia of Chinese characters in previous studies (Table 1-1). Hereafter, these SNPs will be referred to as “core SNPs”.

2.2. Study populations

I examined East Asian populations, expecting if natural selection has acted on genes associated with dyslexia of Chinese characters, the signature would be seen in these populations. At present, publicly available data of these populations were East Asian populations (EAS) in the 1000 Genomes Project phase 3 (1 KG) (Auton *et al.* 2015) and the Korean population from The Personal Genome Project Korea (KPGP) (Kim *et al.* 2018, 2020). I used them as study populations to apply neutrality tests. I also used populations in 1 KG other than EAS for principal component analysis (PCA) and for the investigation of phylogenetic relationship of haplotypes in global populations.

I downloaded 1 KG and KPGP data from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/, and from ftp://biodisk.org/Release/KPGP/KPGP_Data_2017_Release_Candidate/WGS_VCF_89 _KOREAN_JOINT_CALL/, respectively. 1 KG was comprised of 2504 individuals from 26 global populations, and KPGP was comprised of 88 individuals (one sample of KPGP-00349 was removed because it was reported as a non-Korean sample on the ftp site). For KPGP, only SNP data with a filter status of "PASS" were used.

The unphased KPGP data was phased using Eagle2 (Loh *et al.* 2016). As the reference panel for phasing, I used 1 KG after excluding singleton and duplicated SNPs. The imputation of missing genotypes was not employed. For PCA and *nSL*, I merged 1 KG and KPGP after the phasing procedure. The merged data includes only sites that exist in both 1 KG and KPGP.

Study populations for nSL

From the merged data of 1 KG and KPGP, I extracted data of individuals in EAS and KPGP, to which *nSL* applied. The extracted data (EAS-KPGP, hereafter) was comprised of 592 individuals.

Study populations for 2D SFS-based statistics

2D SFS-based statistics (Fujito *et al.* 2018b; Satta *et al.* 2019) require plenty of phased SNPs and are sensitive to singletons. The phasing and merging procedures for EAS-KPGP described above led to a reduced number of SNPs in the data and are expected to be deficient in rare SNPs, because the procedures restricted the merged data to contain only sites existing in both 1 KG and KPGP. For this reason, EAS-KPGP would be inadequate for 2D SFS-based statistics. Therefore, I used only EAS (504 individuals) for the 2D SFS-based statistics. I used biallelic SNP data, with information of ancestral states and without missing genotypes.

2.3. Principal component analysis (PCA)

CDX (Chinese Dai in Xishuangbanna, China) of 1KG are geographically located in East Asia. However, the population historically has not used Chinese characters, and its writing system is related to the writing system of Thai (Davis 2003; Owen 2017) due to transmission of Buddhism (Cohen 2000). Because I consider the possibilities of selective pressure on traits other than reading ability of Chinese characters, it would be reasonable to include CDX into my analyses if it has similar genetic background to other East Asian populations irrespective of the type of writing systems. Then, I performed PCA to confirm it.

From the merged data of 1 KG and KPGP, I extracted data of individuals in 16 Eurasian populations: CEU (Utah Residents with Northern and Western Ancestry), FIN (Finnish in Finland), GBR (British in England and Scotland), IBS (Iberian Population in Spain), and TSI (Toscani in Italia) compose European populations (EUR) in 1 KG; BEB (Bengali from Bangladesh), GIH (Gujarati Indian from Houston, Texas), ITU (Indian Telugu from the UK), PJL (Punjabi from Lahore, Pakistan), and STU (Sri Lankan Tamil from the UK) compose South Asian populations (SAS) in 1 KG; CDX (Chinese Dai in Xishuangbanna, China), CHB (Han Chinese in Beijing, China), CHS (Southern Han Chinese), JPT (Japanese in Tokyo, Japan), and KHV (Kinh in Ho Chi Minh City, Vietnam) compose EAS in 1 KG. KPGP is also included.

PCA was run on genome-wide SNP data, using *smartpca* in EIGENSOFT version 7.2.1 (Patterson *et al.* 2006). I removed SNPs with missing genotype in ≥ 5 individuals as well as SNPs with $r^2 > 0.6$ in a window of 100kb, using *bcftools* (Li 2011). The total number of SNPs used was 2,145,530. From *smartpca,* I also obtained a PCA plot based on population; in this plot, the coordinates used for each population were medians of individual PC scores.

2.4. *nSL*

I used *nSL* (Ferrer-Admetlla *et al.* 2014) as a summary statistic for a neutrality test based on EHH. I applied *nSL* to EAS-KPGP, using the *selscan* program (Szpiech and Hernandez 2014). For calculation of the *nSL* values, only biallelic SNPs with a minor allele frequency ≥ 0.01 were retained. SNPs with missing genotypes and in the major histocompatibility complex (MHC) region (chr6: 28,477,797–33,448,354 of GRCh37) were not used. I referred to information in 1 KG for ancestral states of each SNP. The EHH decay cutoff was sufficiently extended by setting the program option of --max-extend-nsl as 1500, which allowed more accurate *nSL* computation than the default of 100. The total number of SNPs in the data was 6,143,039. *nSL* values were normalized in 100 frequency bins for minor allele, which is the default setting. One-tailed *p*-values were obtained to check neutrality on derived alleles.

2.5. 2D SFS-based statistics

2.5.1. Overview of 2D SFS-based statistics

In order to examine the neutrality of core SNPs and the surrounding regions, I conducted the 2D SFS-based statistics recently developed by Fujito *et al.* (2018) and Satta *et al.* (2019). These statistics measure the intra-allelic variability (IAV) (Slatkin and Rannala 1997, 2000), or the level of polymorphism within haplotypes carrying the derived allele of a focal site (core site). Among the several statistics related to 2D SFS, I used two for the present study: F_c and G_{c0}. The full derivation and equations are presented in Fujito *et al.* (2018) and Satta *et al.* (2019), and a general overview will be presented here.

I considered segregating sites in a region with high LD, which contains a core site. I assumed n chromosomes sampled from a single diploid population. The n samples are divided into two groups: The derived allele group (D group) that carries the derived allele of the core site, and the ancestral allele group (A group) that carries the ancestral allele. The size of the D group is $m\ (1 \leq m < n)$ and that of the A group is $n - m$. At a certain site other than the core site in the region, the number of derived alleles in the D group is described as $i\ (0 \leq i \leq m)$, and the number of derived alleles in the A group as $j\ (0 \leq j \leq n - m)$. Then, the 2D SFS of each site is represented as the matrix $\{\varphi_{i,j}\}$.

The SFS for the entire sample is expressed as:

$$\xi_k = \sum_{i=0}^{k} \varphi_{i,k-i} \text{ for } 1 \leq k < n, \text{ where } k = i + j, \qquad (1)$$

corresponding to Equation (1a) in Satta *et al.* (2019), and analogously, the SFS for the D group is expressed as:

$$\zeta_i = \sum_{j=0}^{n-m} \varphi_{i,j} \text{ for } 1 \leq i < m, \tag{2}$$

corresponding to Equation (1b) in Satta *et al.* (2019).

The statistics of F_c measure the ratio of the amount of mutations in the D group to that in the entire sample, using only mutations younger than the derived allele at the core site (Fujito *et al.* 2018b). The number of derived alleles at a site implies the age of the mutation: A large number (high derived allele frequency) is expected to be an old mutation whereas a small number (low derived allele frequency) suggests a young mutation (Kimura and Ohta 1973; Griffiths and Tavaré 1998; Slatkin and Rannala 2000; Fujito *et al.* 2018b). To exclude mutations older than the mutation on the core site, which should be shared by both the D and A group, the F_c statistic uses "frequency class(es)" based on the scaled mutation rate $\theta = 4N_e u$, where N_e is the effective population size and u is the mutation rate per region per generation. From $E\{\xi_k\} = \theta/k$ (Fu 1995), each frequency class is described as class 1 with $E\{\xi_1\} = \theta$, class 2 with $E\{\xi_2 + \xi_3\} = 5\theta/6$, class 3 with $E\{\sum_{k=4}^{9} \xi_k\} \approx \theta$, class 4 with $E\{\sum_{k=10}^{25} \xi_k\} \approx \theta$, class 5 with $E\{\sum_{k=26}^{68} \xi_k\} \approx \theta$, and so on. The F_c statistic is expressed as:

$$F_c = \frac{\Sigma i \varphi_{i,j}}{\Sigma (i+j) \varphi_{i,j}} \text{ for } i + j \leq k_m < m, \tag{3}$$

corresponding to Equation (4) in Fujito *et al.* (2018); Equation (2) in Satta *et al.* (2019), where k_m is the upper bound number of derived alleles of a frequency class that is one class lower (i.e., younger) than the class containing m.

The statistics of G_{c0} compute the average number of derived alleles per segregating site only observed in the D group, excluding polymorphisms caused by recombination between the D and A group (Satta *et al.* 2019). The G_{c0} statistic is expressed as:

$$G_{c0} = \frac{\sum_{i=1}^{m-1} i\varphi_{i,0}}{\sum_{i=1}^{m-1} \varphi_{i,0}}, \tag{4}$$

corresponding to Equation (7) in Satta *et al.* (2019).

In both statistics, the values are expected to be relatively small under selective sweep.

The time to the most recent common ancestor (TMRCA) of the D group is also estimated. The scaled TMRCA is obtained by

$$ut_D = \frac{\sum_{i=1}^{m-1} i\varphi_{i,0}}{m}, \tag{5}$$

corresponding to Equation (5a) in Satta *et al.* (2019). Then, the TMRCA is obtained by $ut_D/\mu l$, where μ is the mutation rate per site per year, assuming 0.5×10^{-9} (Scally and Durbin 2012), and l is the length of a region.

The variance is defined as

$$Vut_D = \frac{\sum_{i=1}^{m-1} i\varphi_{i,0}}{m} - \frac{m-1}{m}\frac{\theta_\varphi}{2}, \tag{6}$$

corresponding to Equation (5b) in Satta *et al.* (2019), where $\theta_\varphi = \frac{1}{\binom{m}{2}} \sum_{i=1}^{m-1} i(m-i)\varphi_{i,0}$.

2.5.2. Simulations

To obtain *p*-values of F_c and G_{c0}, I performed simulations by *ms* (Hudson 2002). I assumed neutrality without recombination and with the demographic model of (Schaffner 2005), following Fujito *et al.* (2018) and Satta *et al.* (2019). I sampled 30,000 replications, each of which contained a core site with a similar derived allele frequency to a focal SNP (e.g., core SNP). The derived allele frequency for core sites in simulations ranged within one standard deviation of a binomial distribution, as $fr \pm \sqrt{\frac{fr(1-fr)}{n}}$, where fr is m/n of a focal SNP. From the 30,000 replications, I described null distributions of F_c and G_{c0}, and obtained the p-values of F_c and G_{c0} of a focal SNP. I confirmed that 30,000 replications is large enough to obtain stable results.

2.5.3. Screening of the candidate core regions under selective sweep

Screening for further analysis was carried out to examine whether there is a sign of selective sweep in each high LD region containing a core SNP ("core region"). I collected neighboring SNPs that had r^2 with the core SNP $\geq$ 0.75 ("linked SNPs", hereafter) within a 0.5 Mb region in both directions of the core SNP. I then defined the boundaries of each core region by the linked SNPs that were located in the most upstream and downstream positions (Figure 2-1). Note that r^2 also becomes large when a derived allele at the core SNP is linked to ancestral alleles in the linked SNPs and vice versa (ancestral allele at core SNP linked to derived alleles in linked SNPs). I did not use SNPs that displayed this pattern for determining boundaries of the core regions. For each core SNP in its core region, I applied the F_c statistic, which detects the sweep signal by quantifying the amount of mutations in the D group after the emergence of a core SNP.

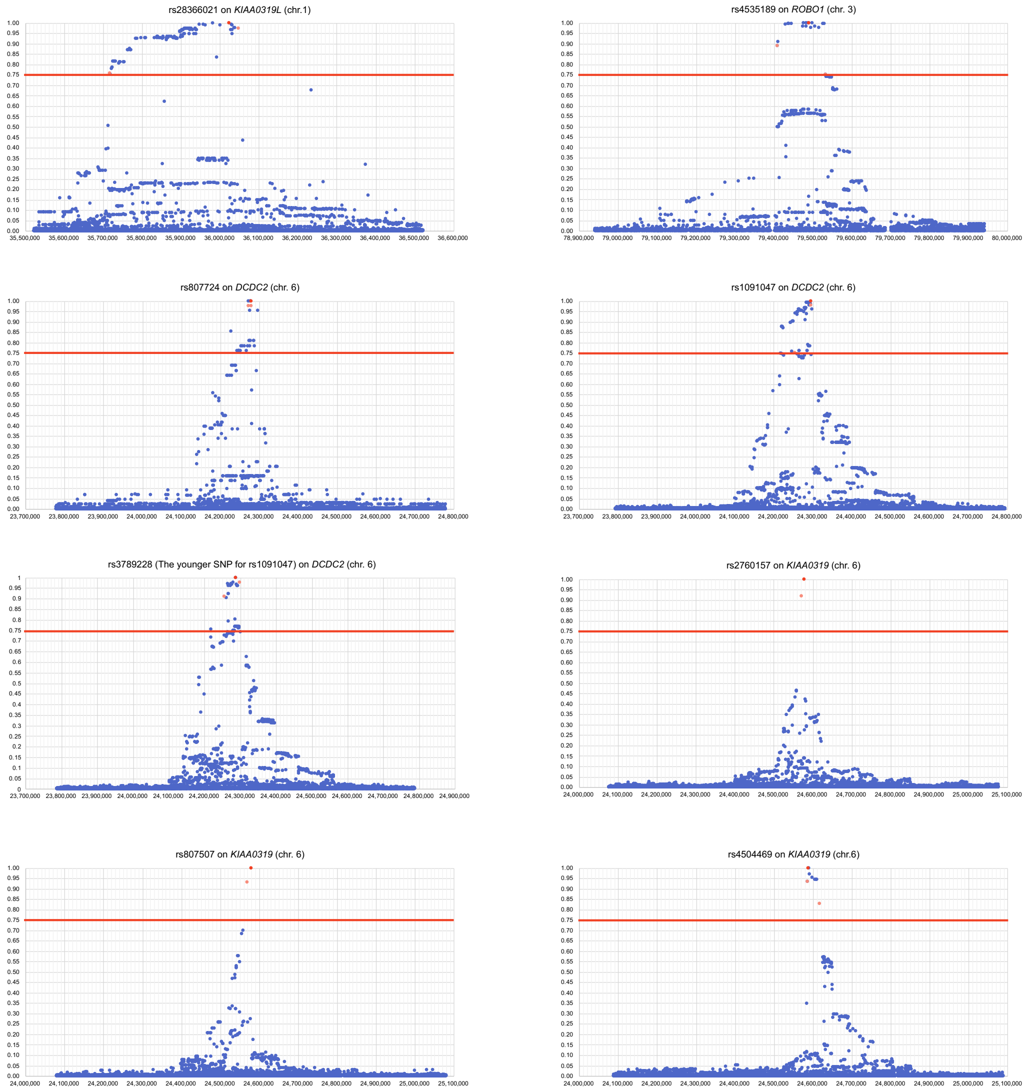

Figure 2-1 r^2 with the core SNP in a 1-Mb region, where the core SNP is centered (red dot). Pink dots represent SNPs that are boundaries of core regions. In some cases, core SNPs themselves are one of the boundaries. Red line represents $r^2 = 0.75$. The SNP is not taken as the boundary in the case when a derived allele at the core SNP is linked to ancestral alleles in the linked SNPs, and when ancestral allele at the core SNP is linked to derived alleles in linked SNPs, even if $r^2 > 0.75$. The surrounding region of rs2255526 on *DIP2A* is less than 1-Mb because the core SNP is located near the end of chromosome 21.

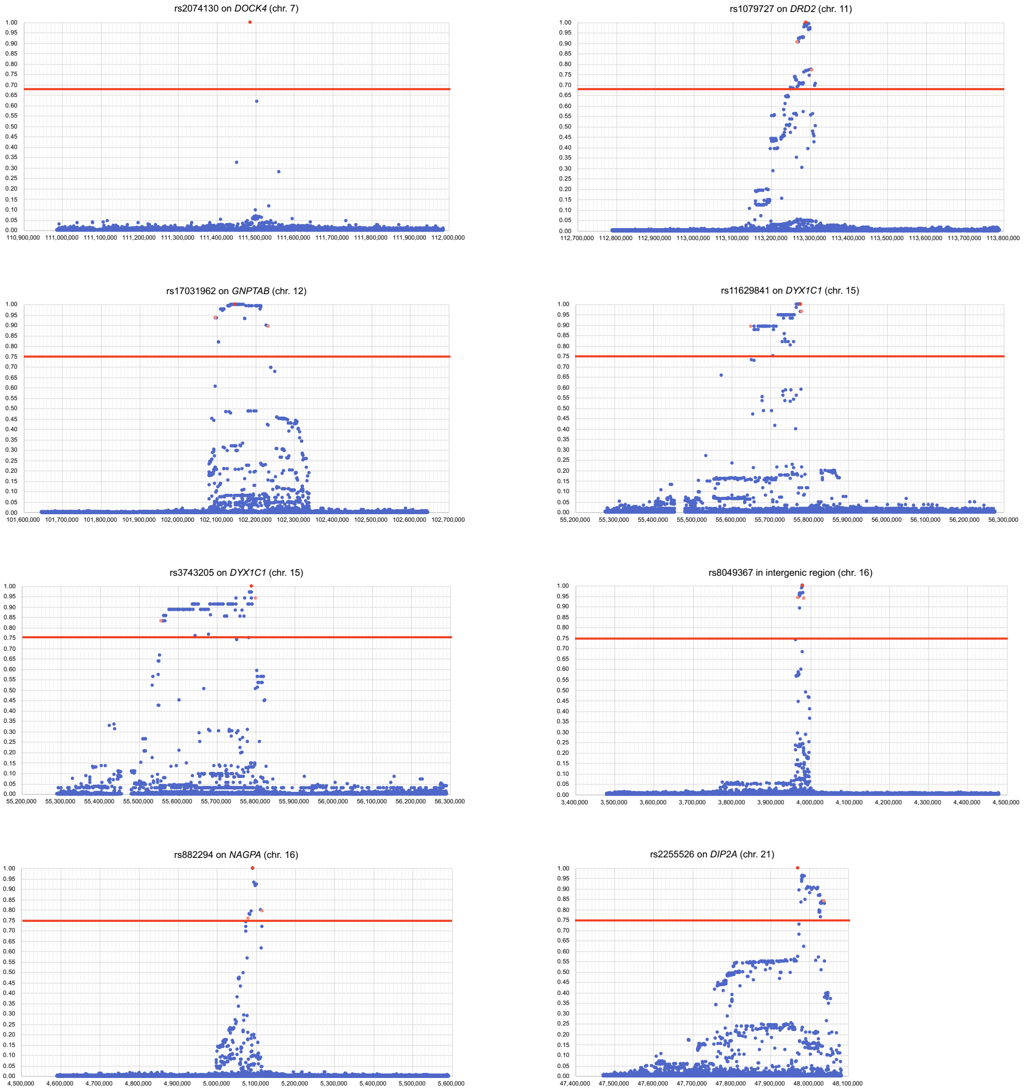

Figure 2-1 (Continued)

2.5.4. Searching for the target site of natural selection

After identifying candidate core regions under the selective sweep from screening (where the F_c value of the core SNP has $p < 0.1$), I further analyzed these regions in detail. Here, the aim was to search for the target site of natural selection ("target site") by comparing the level of polymorphism around each of the candidate SNPs (core

SNP and its linked SNPs) in the core region. It is expected that the level of polymorphism in the D group would be low around the target site due to selective sweep, and this level would increase with distance from the target site. Under this expectation, I used the G_{c0} statistic to examine the average amount of mutations within the D group of each candidate SNP in order to identify the target site.

In order to use the G_{c0} statistic, a surrounding region of each candidate SNP was defined by the following two steps. Firstly, within the core region, I calculated G_{c0} for all possible region lengths containing the specific candidate SNP. Next, I selected the region with the smallest G_{c0} value ("smallest region"). For statistical reliability, each region was set to contain at least 100 SNPs. If more than one region had the same smallest G_{c0} value, I selected the region containing the largest number of SNPs.

I applied this procedure to all candidate SNPs within the core region. The length of the smallest region varied among candidate SNPs, and because G_{c0} values were affected by the region length or the number of SNPs in the region, I could not directly compare the G_{c0} values of the smallest region of all candidate SNPs. Thus, I examined how unlikely the G_{c0} value of each candidate SNP was to be produced under neutrality, by converting the G_{c0} values into the p-values obtained from simulations. I compared these p-values with each other.

The procedures of 2D SFS-based statistics are described in Figure 2-2.

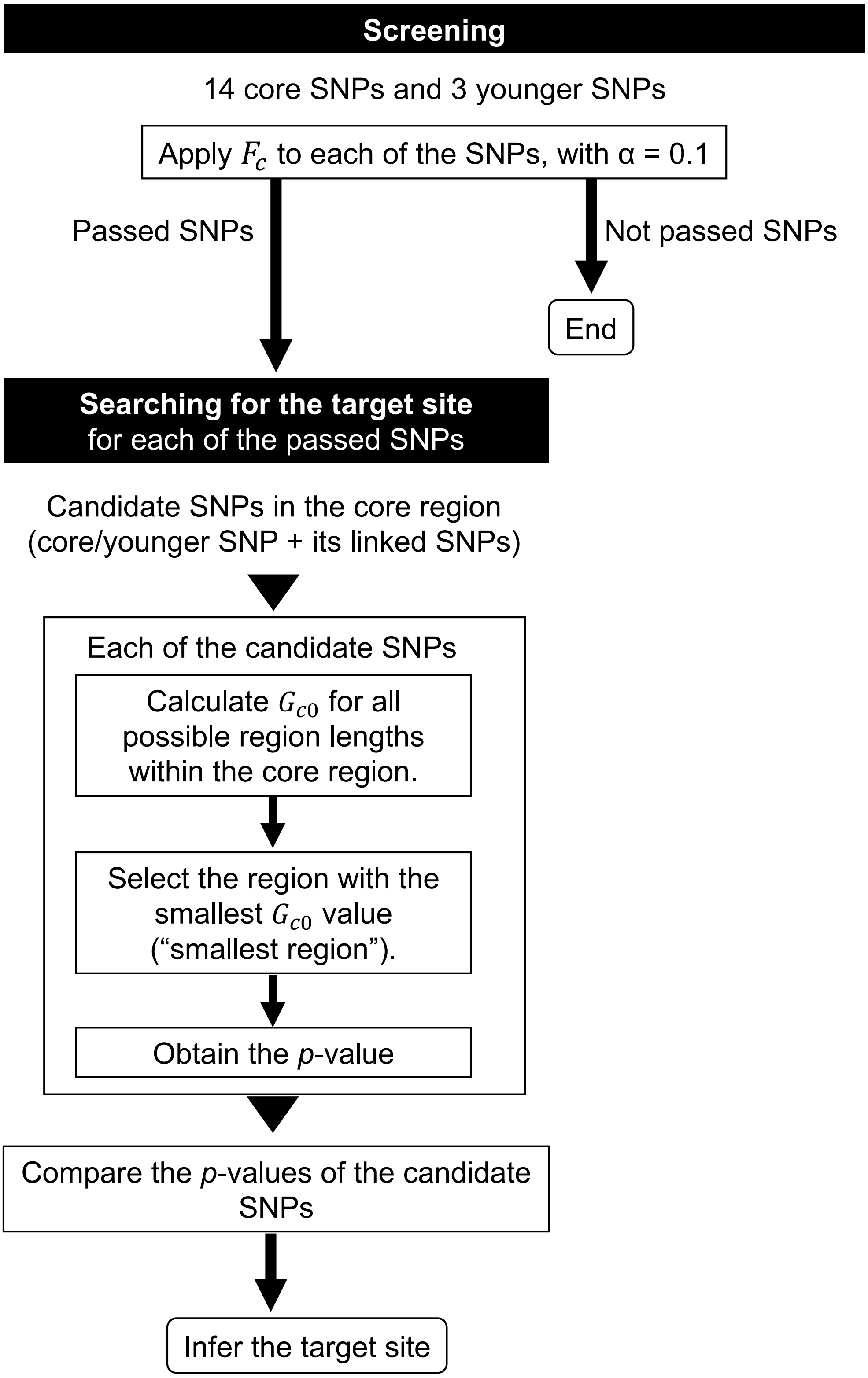
Screening
14 core SNPs and 3 younger SNPs
Apply F_c to each of the SNPs, with α = 0.1
Passed SNPs
Not passed SNPs
End
Searching for the target site
for each of the passed SNPs
Candidate SNPs in the core region
(core/younger SNP + its linked SNPs)
Each of the candidate SNPs
Calculate G_{c0} for all possible region lengths within the core region.
Select the region with the smallest G_{c0} value ("smallest region").
Obtain the *p*-value
Compare the *p*-values of the candidate SNPs
Infer the target site

Figure 2-2 The flowchart of 2D SFS-based statistics.

2.6 Constructing phylogenetic tree of haplotypes

About core regions that passed the screening, I constructed the phylogenetic tree of haplotypes in global populations, in order to investigate the phylogenetic relationship of the haplotypes. I used entire populations of 1KG: 2,504 individuals (5,008 sequences) from African population (AFR), EUR, SAS, EAS, and American population (AMR). Neighbor-joining (NJ) tree (Saitou and Nei 1987) of haplotypes in a core region was constructed by MEGA7 (Kumar *et al.* 2016). For the root of the NJ tree, I included an artificial ancestral sequence which consisted of ancestral state of each SNP informed in 1 KG.

Chapter 3

Results

3.1. The confirmation of genetic background of the study populations: PCA

CDX historically has not used Chinese characters; thus, in order to confirm whether CDX has similar genetic background to other East Asian populations, I performed PCA using data of 1,584 individuals from 16 populations in Eurasia (see section 2.3). PC1 and PC2 were plotted based on both individuals (Figure 3-1A) and populations (Figure 3-1B). CDX was found to share genetic background with other East Asian populations; thus, I included CDX for subsequent neutrality tests.

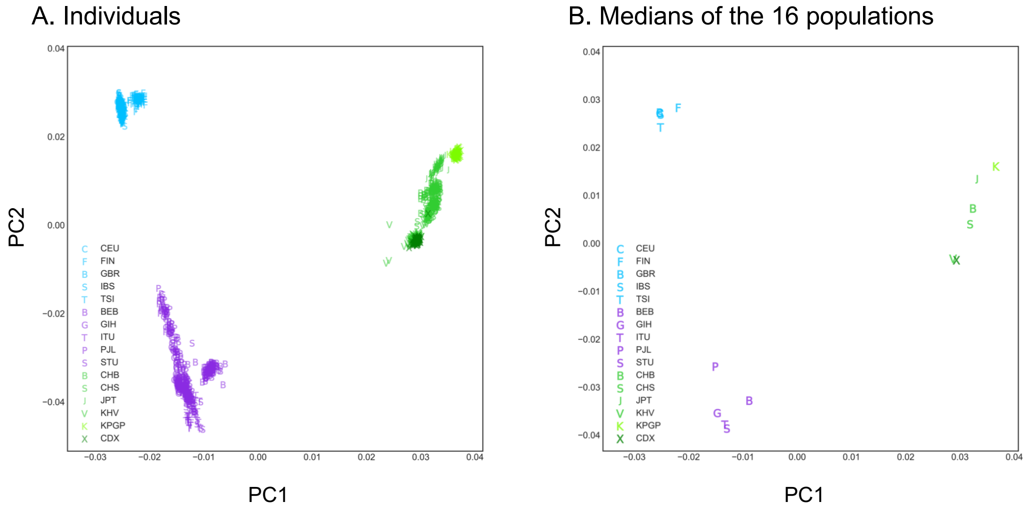

Figure 3-1 (**A**) PC1 and PC2 of 1584 individuals from 16 Eurasian populations. Each letter represents an individual, and corresponds to a population which the individual belongs to. (**B**) Medians of PC1 and PC2 of the 16 populations from (**A**). Each letter represents a population. Letters in blue, purple, and green indicate European, South Asian, and East Asian populations, respectively. In East Asian populations, letters in dark green indicate CDX and light green indicate KPGP.

3.2. Testing neutrality based on EHH: *nSL*

I removed SNPs containing missing genotypes because *selscan* required data without missing genotypes for *nSL*. I could not obtain *nSL* for the core SNP of rs28366021 on *KIAA0319L* because it contained 14 missing genotypes in KPGP. Instead of rs28366021, I examined a neighboring SNP (rs11264175) located 7.5 kb downstream from the core SNP. I used this neighboring SNP because the r^2 value of rs11264175 with rs28366021 was the highest (r^2 = 0.957) in the data when the 14 samples with missing genotypes were excluded.

Moreover, *nSL* could not be properly calculated for the core SNP of rs2255526 on *DIP2A*. This SNP was located at the edge of chromosome 21, and extended haplotypes reached the end of the chromosome before EHH decayed entirely.

I checked normalized *nSL* and their *p*-values of the core SNPs (rs11264175 representative of rs28366021 on *KIAA0319L*), except rs2255526 on *DIP2A*. For all 14 SNPs, normalized *nSL* values were not significant ($p \geq 0.01$ for all; Table 3-1). Therefore, *nSL* did not detect any signatures of positive selection for any of the core SNPs.

Table 3-1. The results of *nSL* for the core SNPs.

Gene	Core SNP	Normalized *nSL*	*p*-Value
KIAA0319L	rs11264175 [a]	0.0771	0.469
ROBO1	rs4535189	−0.1882	0.575
DCDC2	rs807724	1.1328	0.129
DCDC2	rs1091047	−0.5967	0.725
KIAA0319	rs2760157	−2.1853	0.986
KIAA0319	rs807507	0.7329	0.232
KIAA0319	rs4504469	0.7098	0.239
DOCK4	rs2074130	0.3068	0.379
DRD2	rs1079727	−0.1744	0.569
GNPTAB	rs17031962	1.2369	0.108
DYX1C1	rs11629841	−0.0922	0.537
DYX1C1	rs3743205	−0.1939	0.577
intergenic region	rs8049367	−0.4421	0.671
NAGPA	rs882294	0.2399	0.405
DIP2A	rs2255526	-	-

[a] Representative for rs28366021.

3.3. Testing neutrality based on SFS: 2D SFS-based statistics

For the 2D SFS-based statistics, I used two steps. First, I conducted screening using the F_c statistic to check whether a high LD region containing a core SNP (core region) could be under selective sweep. Second, I used the G_{c0} statistic to analyze the core regions that passed the screening, in order to search for the target site of natural selection.

3.3.1. Screening of the candidate core regions under selective sweep

To apply the F_c statistic to each core SNP, I needed to determine its core region.

To do so, I extracted its "linked SNPs" ($r^2 > 0.75$) (see section 2.5; Figure 2-1). However, I could not define the core region for rs2074130 on *DOCK4* because no linked SNPs were identified. This meant that F_c statistic could not be applied to this SNP. Thus, the SNP was omitted from subsequent analyses including the F_c statistic. Based on the absence of an LD region, I inferred that the derived allele of rs2074130 was not under positive selection, because if selection had acted, then the derived allele should at least have some extent of LD as a signature of the genetic hitchhiking.

At this stage of the screening, I could not determine whether the target site of selection was the core SNP or one of its linked SNPs. Thus, I considered both a core SNP and the linked SNPs in a core region as candidates for the target site. The number of derived alleles of linked SNPs should be similar to that of the core SNP, and therefore, the age of linked SNPs is expected to be similar to that of the core SNP. However, even if linked SNPs showed a similar number of derived alleles in a local population, such as EAS, they could show a different number from the core SNP when looking at the global population. The level of polymorphism should be different between the core SNP and such linked SNPs, due to the difference in age. So, for such linked SNPs, it is inappropriate to apply the statistic to its core SNP.

For this reason, I checked the global derived allele count (number of derived alleles in the entire population in 1 KG) of a core SNP and its linked SNPs, in addition to the count in EAS. Next, each SNP was classified into a "frequency class" (see the methods section). I found that three core SNPs (rs4535189 on *ROBO1*, rs1091047 on *DCDC2*, and rs3743205 on *DYX1C1*) had some linked SNPs with global derived allele counts smaller than that of their core SNPs, and that these linked SNPs were classified into lower (i.e., younger) frequency classes than their core SNPs. The global derived allele count of rs4535189 on *ROBO1* is 2280 and belonged to frequency class 9; in the core region, 16 of the 23 linked SNPs were classified into the same class 9 as the core SNP, but 7 linked SNPs were classified into class 8. Similarly, the global derived allele count of rs1091047 on *DCDC2* was 3871 and belonged to class 10; 7 of the 16 linked SNPs were also classified into class 10, but 9 were classified into class 9. Moreover, the global derived allele count of rs3743205 on *DYX1C1* was 517 and classified into class 8, whereas the classes of the 97 linked SNPs varied: 7 were classified into a class 7, 87 into class 6, and 3 into class 5. No linked SNPs were classified into the same class 8 as the core SNP.

In each of these three cases, in addition to the core SNPs, I analyzed one linked SNP in a younger class, because these linked SNPs should have different evolutionary depths and therefore different polymorphism levels from their core SNPs. For each of the three cases, among the several linked SNPs, I selected a linked SNP that showed the smallest derived allele count as the “younger SNP”: rs73129039 (global derived allele count = 1214 and frequency class 8) on *ROBO1*, rs3789228 (global derived allele count = 2583 and frequency class 9) on *DCDC2*, and rs79024225 (global derived allele count = 31 and frequency class 5) on *DYX1C1*.

I also found that some linked SNPs were classified into a globally older frequency class than their core SNP. However, I ignored such cases. The extent of polymorphism at an “older SNP” should be greater than that at a core SNP due to the difference in age. Although the F_c value is expected to be small under selective sweep, the F_c value at the "older SNP" cannot be smaller than that at the core SNP. Therefore, I did not examine older SNPs in subsequent analyses.

I screened core regions for detailed analysis. The F_c statistic was applied to the 14 core SNPs and the 3 younger SNPs to identify the regions suspected to have experienced selective sweep, using statistical significance of $\alpha = 0.1$. The *p*-values were obtained from simulations (Table 3-2), and two SNPs remained after this screening: rs17031962 on *GNPTAB* (p = 0.038) and rs3789228 (younger SNP for rs1091047) on *DCDC2* (p = 0.068).

Table 3-2. F_c statistic results for the core SNPs and three younger SNPs.

Gene	Core SNP	Number of Derived Alleles n = 1008	Length of the Core Region	Number of Segregating Sites	F_c	p-Value
KIAA0319L	rs28366021	236	330,223	2204	0.1476	0.718
ROBO1	rs4535189	369	124,626	866	0.1287	0.316
ROBO1	rs73129039 [a]	363	124,626	866	0.1232	0.303
DCDC2	rs807724	965	5910	53	0.6742	0.159
DCDC2	rs1091047	824	41,134	334	0.3044	0.111
DCDC2	rs3789228 [b]	782	41,134	334	0.2020	0.068 *
KIAA0319	rs2760157	460	7387	53	0.7765	0.939
KIAA0319	rs807507	189	11,475	81	0.0220	0.111
KIAA03219	rs4504469	113	32,025	241	0.0736	0.529
DOCK4	rs2074130	102	-	-	-	-
DRD2	rs1079727	419	38,525	372	0.1370	0.260
GNPTAB	rs17031962	296	136,804	868	0.0400	0.038 *
DYX1C1	rs11629841	58	130,280	1113	0.0589	0.769
DYX1C1	rs3743205	35	242,254	2024	0.0680	0.963
DYX1C1	rs79024225 [c]	31	242,254	2024	0.0308	0.758
intergenic region	rs8049367	342	14,513	177	0.1486	0.428
NAGPA	rs882294	191	34,706	339	0.2875	0.905
DIP2A	rs2255526	266	67,101	661	0.0899	0.361

* $p < 0.1$; [a] the younger SNP of rs4535189 on *ROBO1*; [b] the younger SNP of rs1091047 on *DCDC2*; [c] the younger SNP of rs3743205 on *DYX1C1*.

3.3.2. Searching for the target site of natural selection

On the two core regions that contained SNPs that passed screening (rs17031962 on *GNPTAB* and rs3789228 on *DCDC2*), I searched for the target site of natural selection using G_{c0} (see the methods section).

The core region of rs17031962 on GNPTAB

The derived allele count at rs17031962 is 296 out of 1008 chromosomes in EAS. The reported risk allele is the ancestral allele (Chen *et al.* 2015). The core region of rs17031962 is approximately 137 kb long (chr12: 102,096,776–102,233,579 of GRCh37) and contains two genes other than *GNPTAB*: *CHPT1* (partial) and *SYCP3*. *CHPT1* encodes cholinephosphotransferase (Henneberry *et al.* 2000), and *SYCP3* encodes a component of the synaptonemal complex, which is involved in the pairing and crossover of homologous chromosomes during meiosis (Yuan *et al.* 2000). I found that the core region contained 50 linked SNPs in the same global frequency class as rs17031962.

I found one possible phasing error for one of the SNPs (rs78494298). The derived allele count at rs78494298 was 15, and only 14 alleles were linked to the derived allele at rs17031962 (core SNP). For the calculation of 2D SFS, this was counted as $\varphi_{14,1}$. For sample HG00707, one of the two chromosomes carried the derived allele at rs17031962 and the ancestral allele at rs78494298. Conversely, the other chromosome carried the ancestral allele at rs17031962 and the derived allele at rs78494298; this is the cause of $\varphi_{14,1}$ at rs78494298. This pattern is not likely caused by recombination because surrounding SNPs did not display evidence of any cross-over event (Figure 3-2), and supports the possibility of a phasing error. Because the G_{c0} statistic counts only $\varphi_{i,0}$ (and therefore ignoring $\varphi_{14,1}$), the state at rs78494298 results in a smaller G_{c0} value and *p*-value than the case of $\varphi_{15,0}$, where the possible phasing error was corrected. I therefore altered the present (default) state of $\varphi_{i,j}$ at rs78494298 to $\varphi_{15,0}$. Regardless of whether this is a true phasing error or not, this manipulation provides an even more conservation approach, compared to the default state, for calculating G_{c0}.

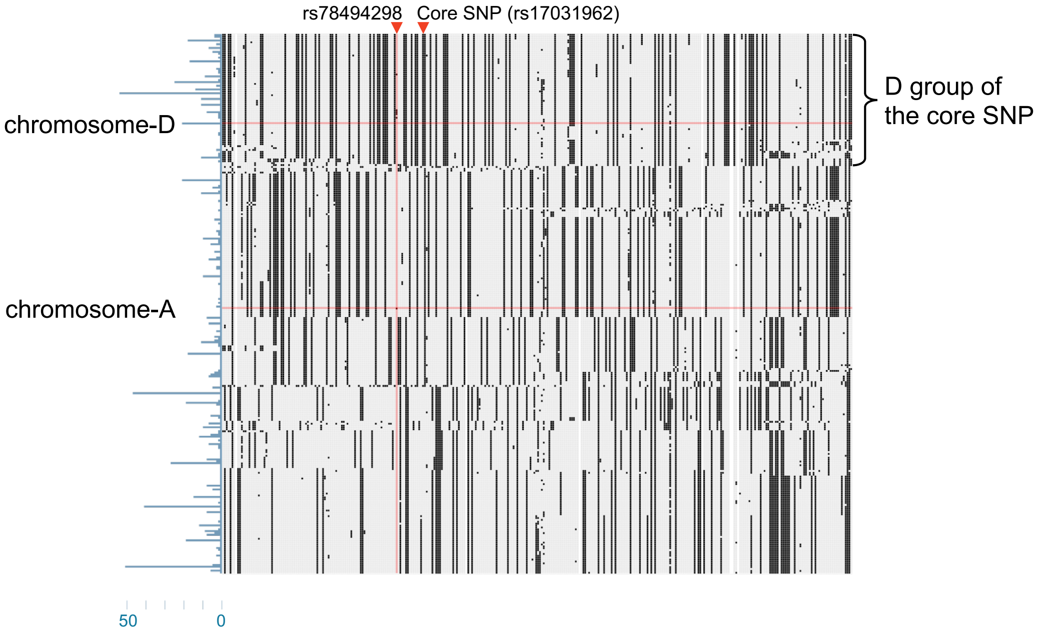

Figure 3-2 Haplotypes observed in EAS for the core region of rs17031962, using SNPs with minor allele frequency (MAF) in EAS ≥ 0.01. Columns represent genomic positions, and rows represent haplotypes. Haplotypes were sorted according to the Neighbor-joining (NJ) tree. For each haplotype, a black or grey cell represents derived or ancestral allele, respectively, at the position. The positions of rs17031962 and rs78494298 are indicated by red triangles. The vertical red line indicates the position of rs78494298. The lengths of blue bars on the left side display the counts of each haplotype. "chromosome-D" indicates the one chromosome of HG00707 carrying the derived allele of the core SNP, and "chromosome-A" indicates the other chromosome of HG00707 carrying the ancestral allele of the core SNP. Horizontal red lines represent haplotypes containing each of the two chromosomes of HG00707.

Among the linked SNPs, I found that three consecutive SNPs (rs557004549, rs183736467, rs188452374) showed different patterns from the other linked SNPs (Figure 3-3). These three SNPs are completely linked to each other; some of their D group seemed to be linked to the A group of the core SNP and other linked SNPs, and vice versa. If haplotypes with the derived allele at these three SNPs were selected for, then LD is not expected to break down immediately. Therefore, I removed these three SNPs from subsequent analysis, assuming that none of them would be the target site. Subsequently, the number of the candidate SNPs became 48 (the core SNP and 47 linked SNPs).

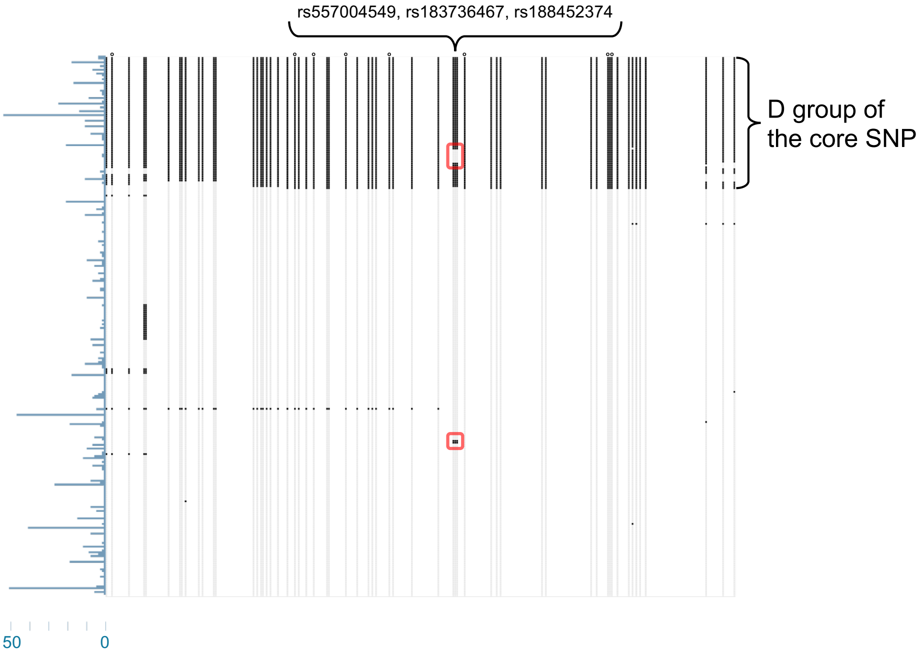

Figure 3-3 rs17031962 and its linked SNPs. SNPs with “o” above indicate linked SNPs in globally older frequency classes than rs17031962, which were ignored in the analysis. Three consecutive SNPs (rs557004549, rs183736467, rs188452374), which were excluded from the analysis, are noted. Different pattern from the other linked SNPs are marked with red rectangles: some of their derived alleles seemed to be linked to ancestral alleles of the other candidate SNPs, and vice versa.

For each of the 48 candidate SNPs in the core region of rs17031962, I selected the region with the smallest G_{c0} value, and obtained p-values from simulations (Figure 3-4A). Among them, 12 SNPs were statistically significant ($p < 0.01$; Figure 3-4A bottom). SNPs that overlapped in the same "smallest region" and shared the same p-values were grouped together into the same region. I identified five regions that contained significant SNPs; these regions were numbered according to the ascending order of p-values (Figure 3-4B).

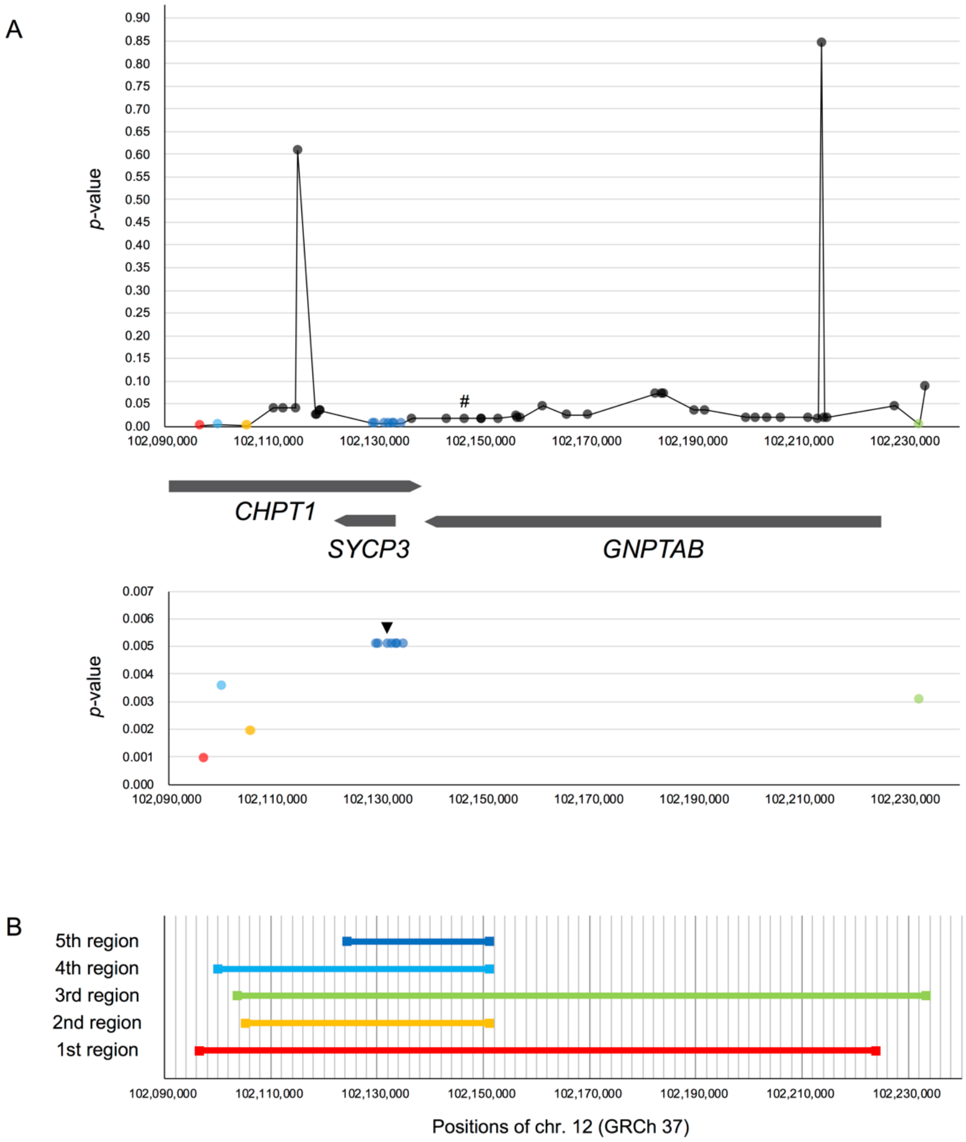

Figure 3-4. (**A**) Top: p-values of G_{c0} for 48 candidate SNPs in the core region of rs17031962 on *GNPTAB*. Each dot represents a candidate SNP. The core SNP is indicated by "#". Colored dots other than black indicate the 12 SNPs with $p < 0.01$. SNPs with the same p-value and smallest region are indicated in the same color. Positions of the three genes in the core region are illustrated as thick lines underneath. Bottom: The same plot showing only the SNPs with $p < 0.01$. The possible target site is indicated by a black arrow. (**B**) The lengths and positions of the smallest regions of the SNPs with $p < 0.01$. The regions are numbered according to the ascending order of the p-value. The color of the regions corresponds to the dot color in (**A**).

The range of the smallest region for each candidate SNP, as well as the p-value, may provide insights into the target site. The candidate SNP of the first region (p = 0.0009) was one of the core region boundaries, and that of the third region (p = 0.0031) was located close to the other core region boundary, where LD seemingly began to break down (r^2 values for the first and third region are 0.934 and 0.892, respectively; Figure 2-1). The first region covers almost the entire core region, where the average amount of mutations in the D group (i.e., G_{c0} value) was the smallest. Shorter regions with this candidate SNP had higher G_{c0} values (Figure 3-5), indicating that the average amount of mutations in the area around this candidate SNP is high; this contradicts the expectation that the level of polymorphism around the target site is small. Thus, I do not consider the candidate SNP of the first region to be the target site. This also applied for the candidate SNP of the third region. Furthermore, while the second and fourth regions overlapped with the fifth region (Figure 3-4B), when I investigated shorter regions that covered the candidate SNP in the second (or fourth) region but not that of the fifth region, I found higher G_{c0} values (Figure 3-6). From these observations, I considered that the fifth region may hold the target site, although the p-value of the SNPs in the fifth region (p = 0.0051) is the highest among the significant SNPs.

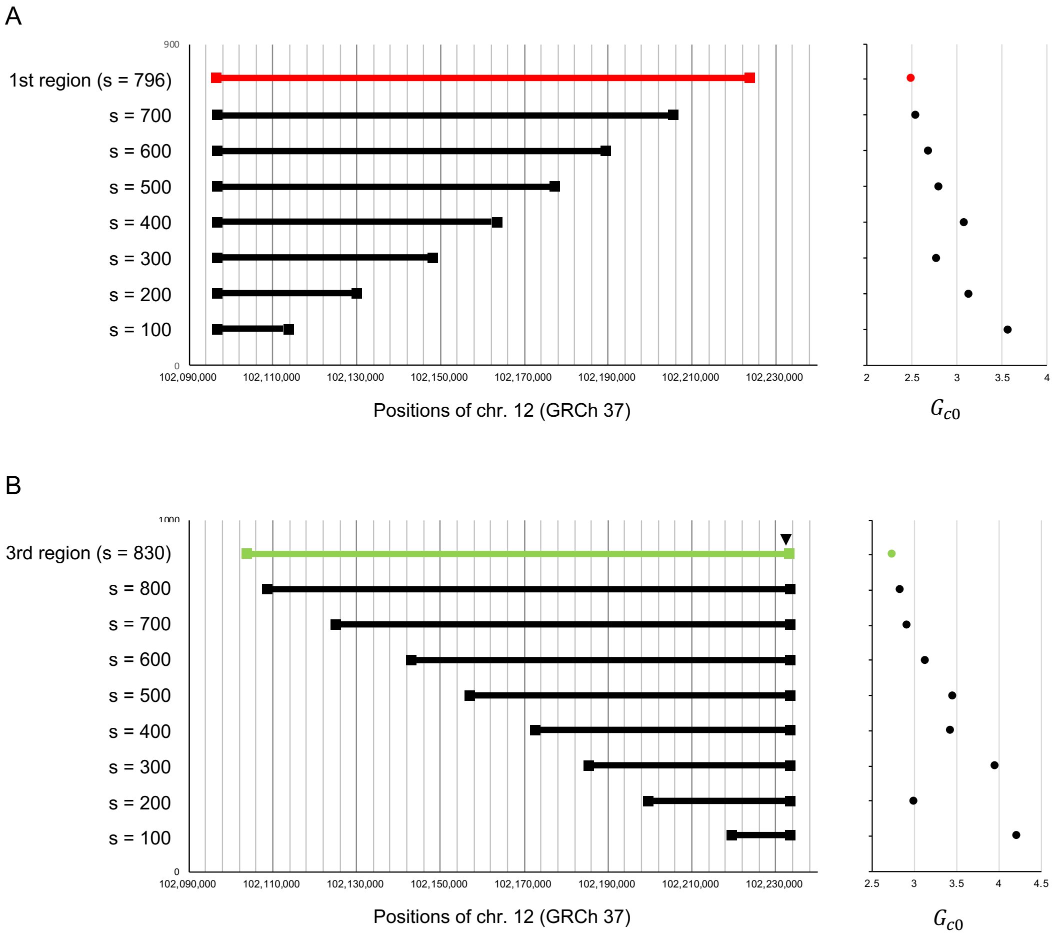

Figure 3-5 (**A**) **Left**: The lengths and the positions of the 1st region and seven kinds of shorter regions. The candidate SNP of the 1st region was the region boundary in upstream side. For the shorter regions, the region boundaries in upstream side were fixed. By change in the region boundary in downstream side, seven shorter regions were obtained according to the number of segregating sites ("s"). **Right**: G_{c0} values obtained from each of the regions. (**B**) **Left**: The lengths and the positions of the 3rd region and eight kinds of shorter regions. The candidate SNP of the 3rd region was located close to the region boundary in downstream side, indicated by a black arrow. For shorter regions, the region boundaries in downstream side were fixed. By change in the region boundary in upstream side, eight shorter regions were obtained according to the number of segregating sites ("s"). **Right**: G_{c0} values obtained from each of the regions.

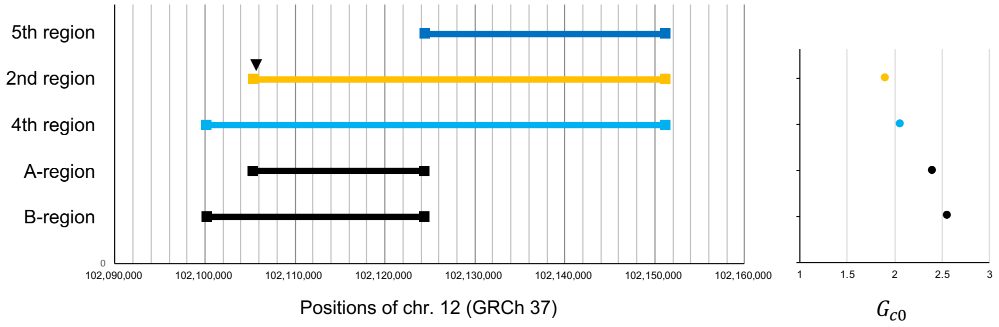

Figure 3-6 Left: The lengths and the positions of the 2nd, 4th and 5th regions, and two shorter regions (A-region and B-region). The candidate SNPs of the 2nd region were located close to the region boundary in upstream side, indicated by a black arrow (There were two candidate SNPs in the 2nd region, which were closely located and completely linked). The candidate SNP of the 4th region was the region boundary in upstream side. A-region was the shorter region of the 2nd region, where the region boundary in downstream was set not to overlap with the 5th region. B-region was the shorter region of the 4th region, where the region boundary in downstream was set not to overlap with the 5th region. **Right**: G_{c0} values of the 2nd and 4th regions, and the two shorter regions.

The candidate SNPs in the fifth region were located in *SYCP3* and its upstream region. To elucidate the possible biological trait under selection, I investigated the functional significance of the SNPs by checking the Ensembl Variant Effect Predictor (VEP) (McLaren *et al.* 2016) for GRCh37 (version 96). I found a candidate SNP in the fifth region (rs3751248) located on an open chromatin region; this SNP may have biological functions, possibly expression regulation, and the genotype difference may have different traits that affect individual fitness. Thus, I inferred that this SNP could be the target site.

I estimated the TMRCA of the D group of the target site in the fifth region

(102,124,515–102,151,310) as 15,434$\pm$4,390 years, from l =26,796, $ut_D = 0.207$, and $Vut_D = 0.0035$.

The core region of rs3789228 on DCDC2, as the younger SNP of rs1091047

rs3789228 is the "younger SNP" of rs1091047 on *DCDC2*. The number of derived alleles of rs3789228 (younger SNP) is 782 out of 1008 chromosomes in EAS, while that of rs1091047 (core SNP) is 824. The reported risk allele of the core SNP is the ancestral allele (Su *et al.* 2015). For this detailed analysis, I re-extracted linked SNPs of rs3789228. Almost all linked SNPs were clustered together. However, one linked SNP was located 38 kb from the cluster and thus removed from analysis as it is not likely to be the target site. Then, 20 linked SNPs in the same frequency class as rs3789228 (class 9) were collected. The core region of the younger SNP was ~43 kb long (chr6: 24,255,044–24,297,900 of GRCh37).

For each of the 21 candidate SNPs (the younger SNP and 20 linked SNPs) in the core region, I selected the region with the smallest G_{c0} and obtained the p-value for these G_{c0} values by simulations (Figure 3-7A). Among them, 10 SNPs were significant ($p <$ 0.01). The top SNP (p = 0.0003) and the second SNP (p = 0.0004) shared the same smallest region and were grouped together as the first to second region. I also grouped other SNPs together that were in the same smallest region and with the same p-value. In total, five regions were detected (Figure 3-7B), which I numbered according to the ascending order of the p-value.

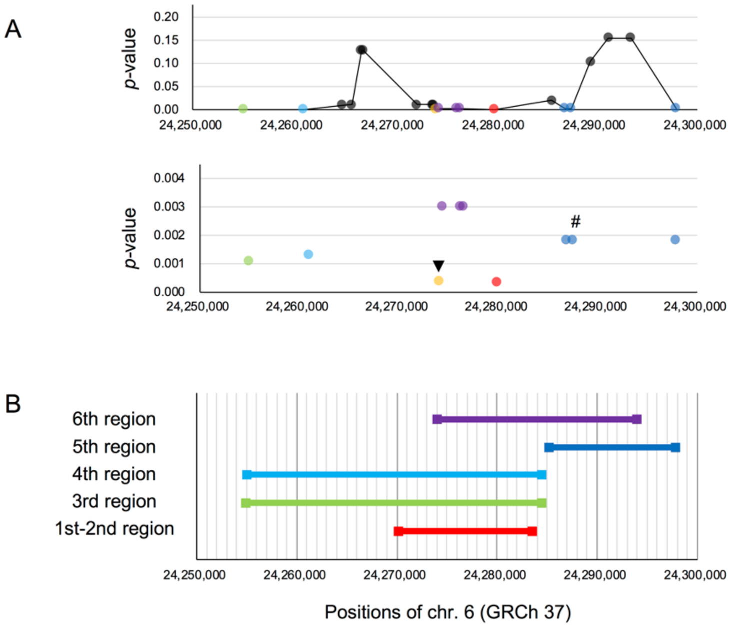

Figure 3-7. (**A**) Top: *p*-values of G_{c0} for 21 candidate SNPs in the core region of rs3789228 on *DCDC2*. Each dot represents a candidate SNP. Colored dots other than black indicate SNPs with $p < 0.01$. SNPs with the same *p*-value and smallest region are indicated in the same color. Bottom: The same plot showing only the SNPs with $p < 0.01$. The target site is indicated by a black arrow. The younger SNP is indicated by "#". (**B**) The lengths and the positions of the smallest regions of the SNPs with $p < 0.01$. The regions are numbered according to the ascending order of *p*-value. "1st-2nd region" indicates the smallest region containing both the top and the second SNP, shown as red and orange dots in (**A**), respectively, and overlapped in the same smallest region. The colors of the other regions correspond to the dot colors in (**A**).

Only the top and second SNPs showed $p < 0.001$. In a similar fashion to other case (rs17031962 on *GNPTAB*), the first to second region was overlapped with both the third and fourth regions, and partially overlapped with the sixth region. Based on this, I considered that either the top or second SNP may be the target site. On VEP (McLaren *et* *al.* 2016) for GRCh37 (version 96), I found that the second SNP (rs12055879) and a single SNP in the sixth region (rs807700) were in both the enhancer region and CTCF binding sites, which may affect expression regulation. Considering the p-value, I inferred that the target site could be the second SNP.

I estimated the TMRCA of the D group of the target site in the first to second region (24,270,213-24,283,618) as 22,236 ± 5,635 years, from $l =$ 13,406, $ut_D =$ 0.149, and $Vut_D =$ 0.0014.

3.4. Phylogenetic relationship of haplotypes in global populations

About the passed core regions of rs17031962 on *GNPTAB* and rs3789228 on *DCDC2* in 2D SFS-based statistics, I investigated the phylogenetic relationship of the haplotypes in global populations, using entire populations of 1 KG (2,504 individuals). Before that, I checked the distribution of the derived allele of the core SNPs and the younger SNPs among populations in the world, together with that of not-passed core regions (Figure 3-8). The derived allele distribution of the two target sites were also checked, i.e., rs3751248 in the core region of rs17031962 on *GNPTAB*, and rs12055879 in the core region of rs3789228 on *DCDC2*.

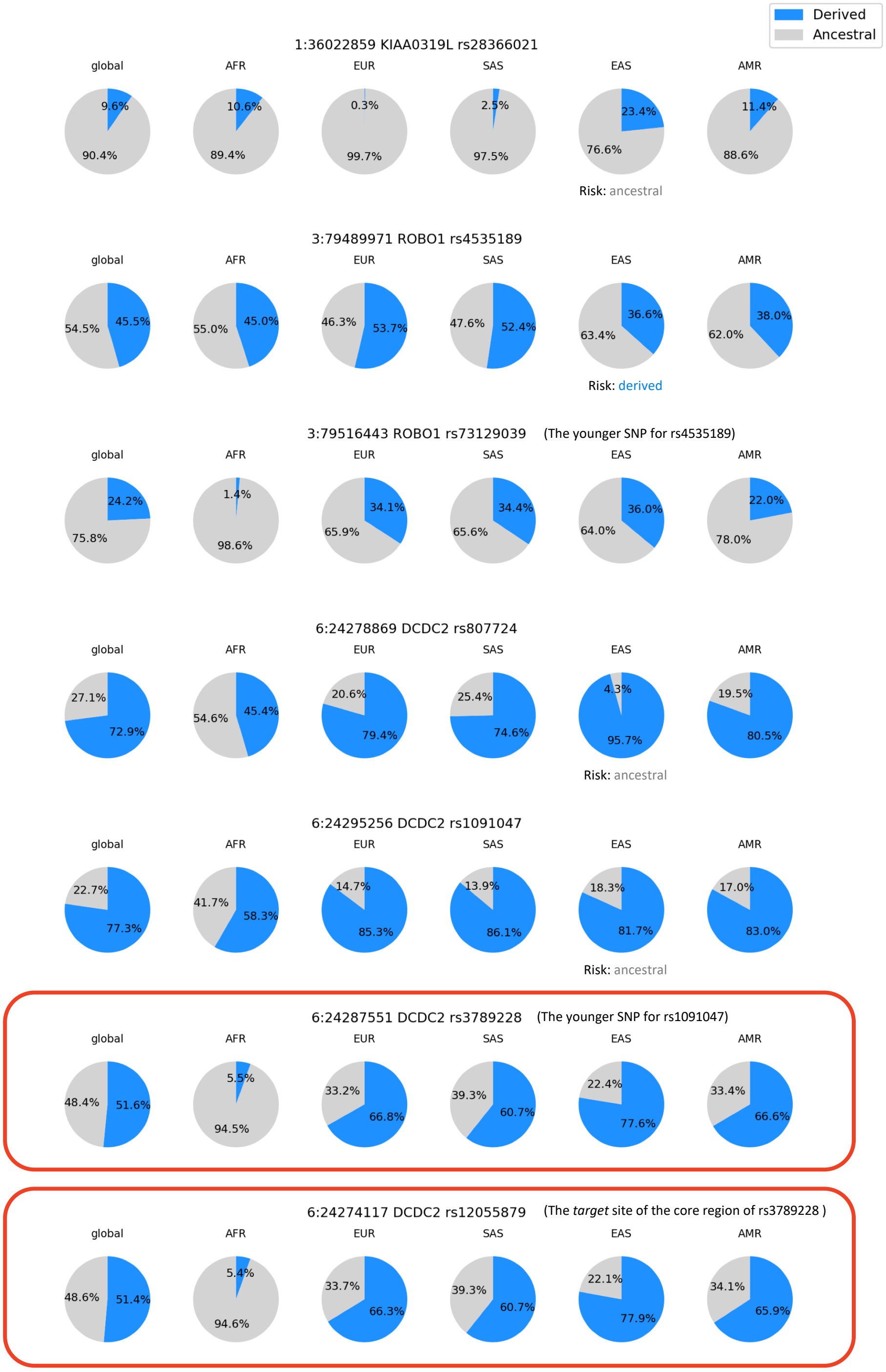

Figure 3-8 The derived allele frequencies of the 15 core SNPs in global, African (AFR), European (EUR), South Asian (SAS), East Asian (EAS), and American (AMR) populations of 1KG. Three younger SNPs and two target sites are also displayed after their core SNPs. For the 15 core SNPs, the risk alleles reported in the previous studies in Table 1 are noted below the pie charts of EAS.

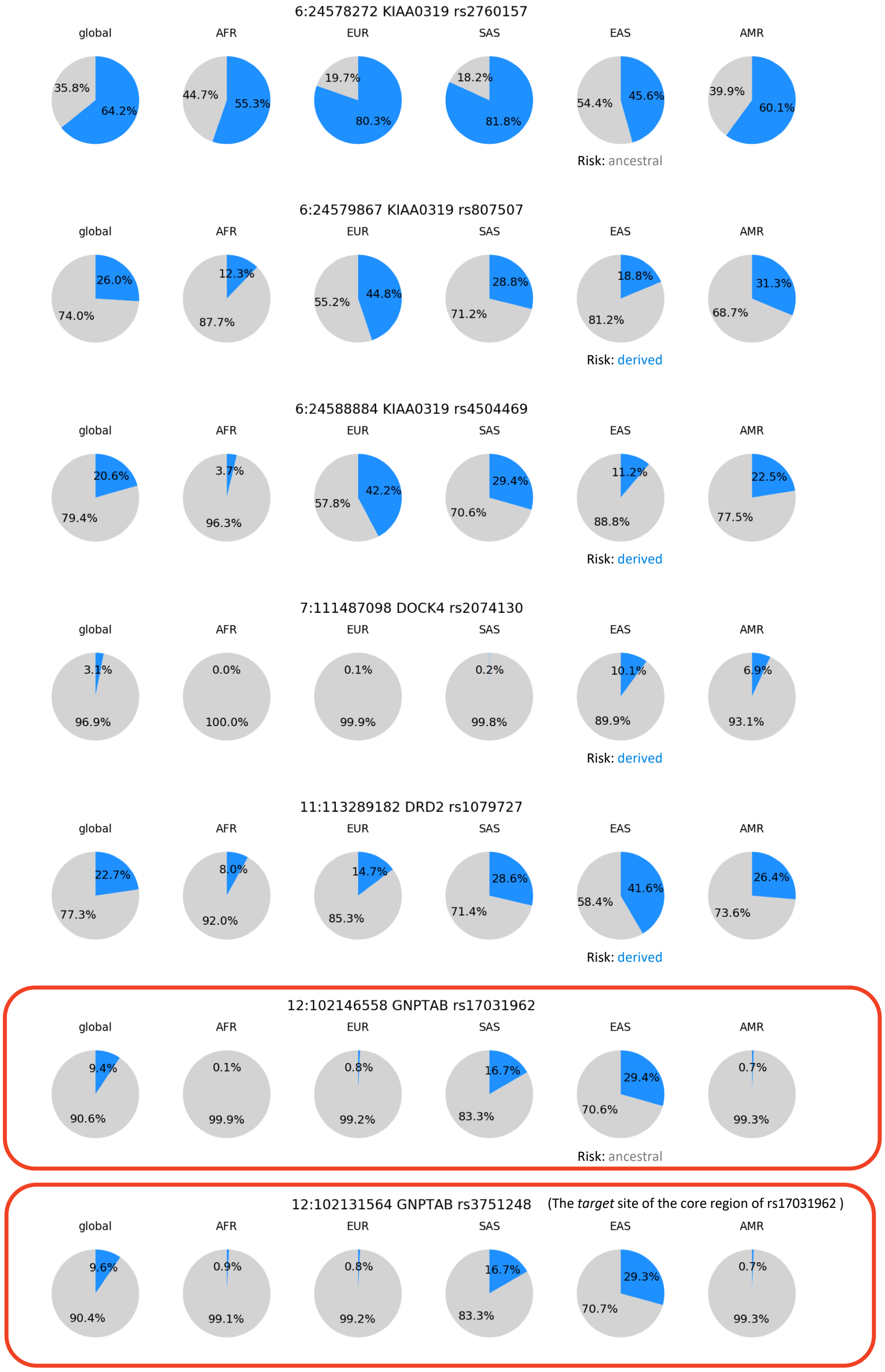

Figure 3-8 (Continued)

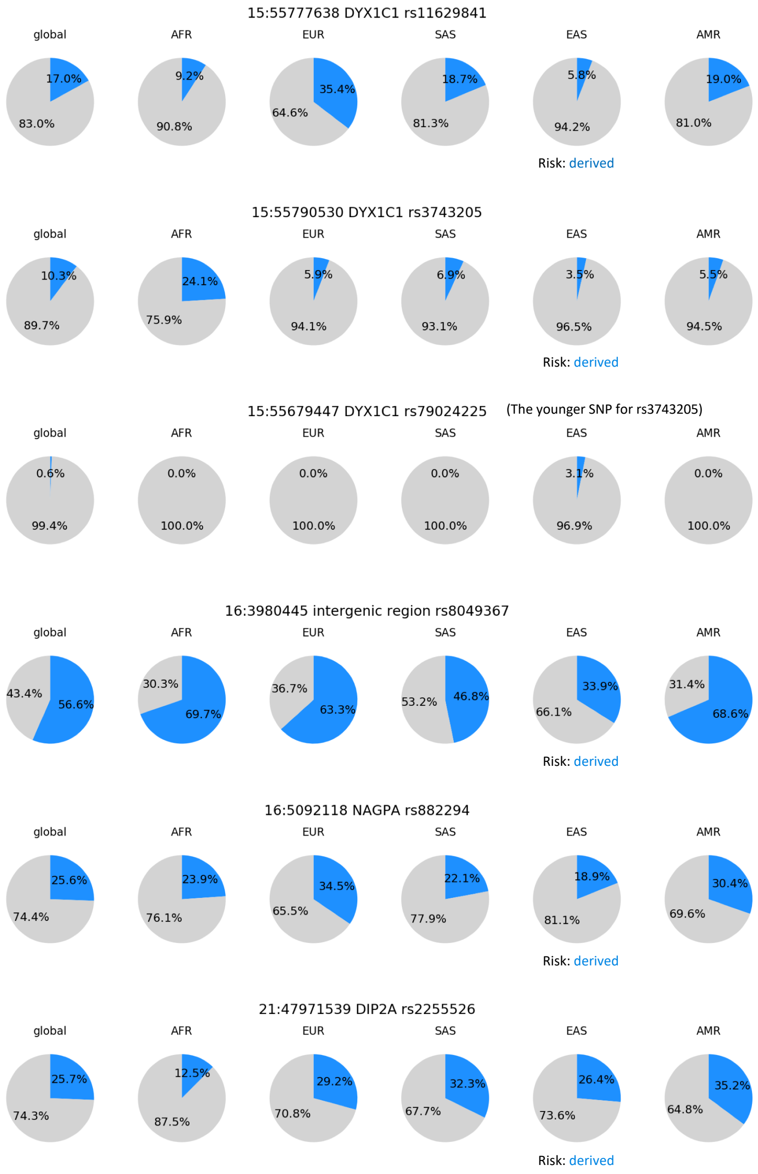

Figure 3-8 (Continued)

About the core SNP of rs17031962 on *GNPTAB*, I found that the derived allele are mainly observed in Asian populations (29.4% in EAS and 16.7% in SAS, and less than 1% in the other populations), where the derived allele is reported as non-risk allele (Chen *et al.* 2015). The target site (rs3751248) also shows similar derived allele distribution.

The core SNP of rs1091047 on *DCDC2* and the younger SNP (rs3789228) show the difference in derived allele frequency especially in African population: 58.3% for rs1091047, and 5.5% for the younger SNP (rs3789228). That for the target site (rs3751248) is similar to the younger SNP. The difference in frequency class between them (see section 3.2.1) should be caused by this difference in derived allele frequency in African population.

I constructed the NJ tree of the passed core regions. First, using entire populations of 1 KG, I constructed the NJ tree of the core region of rs17031962 on *GNPTAB* (Figure 3-9). The region is approximately 137 kb long (chr12: 102,096,776–102,233,579 of GRCh37). Because the core region was determined by r^2 with the core SNP $\geq$ 0.75, the region contains recombinants. SNPs with global MAF $\geq$ 0.01 were used, and 1,066 haplotypes were observed. I checked the D group of the target site (rs3751248; m =295). The D group from entire population is largely clustered in (1) in Figure 3-9, where 294 sequences of the D group from EAS are contained. The other one sequence of the D group from EAS is in (2).

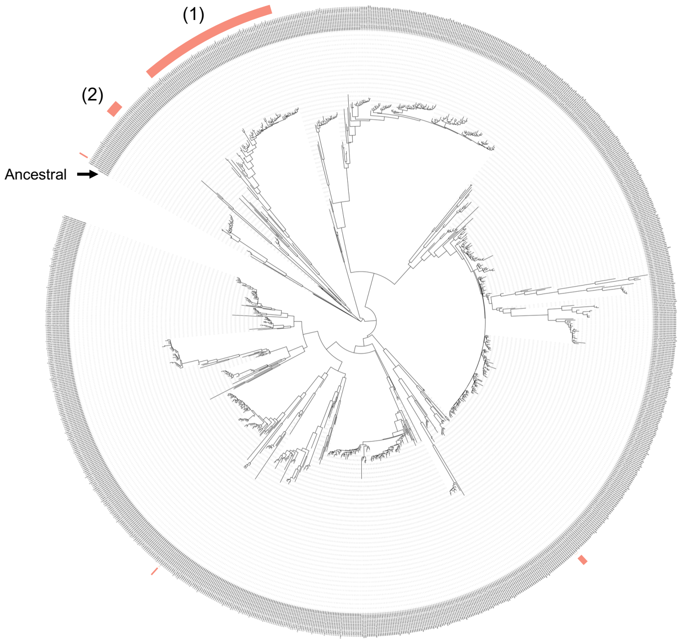

Figure 3-9 the NJ tree of the core region of rs17031962 on *GNPTAB*, using 1,066 haplotypes from entire population of 1 KG. Red lines indicate the D group of the target site. The ancestral sequence as the root is pointed by the arrow.

Next, I constructed the NJ tree of the core region of rs3789228 on *DCDC2* (Figure3-10). This SNP is the younger SNP of rs1091047. The core region of rs3789228 is approximately 43 kb long (chr6: 24,255,044–24,297,900 of GRCh37). I used SNPs with global MAF ≥ 0.005, and constructed the NJ tree of the 296 haplotypes. I checked the D group of the core SNP (rs1091047) and the target site (rs12055879). About the D group of the core SNP, some branches are overlapped by the D group of the target site. These overlapped branches are composed of haplotypes mainly from non-African populations. Haplotypes from African populations are largely observed in branches not overlapped by the D group of the target site. This region contains the core SNP and its 7 linked SNPs in frequency class 10, and the younger SNP and its 20 linked SNPs in frequency class 9 including the target site. In order to investigate the observation from the NJ tree in detail, I described which haplotypes carried the derived alleles of these SNPs (Figure 3-11). Derived alleles of the core SNP (rs1091047) and its linked SNPs in frequency class 10 were carried by various haplotypes containing sequences from both African and non-African populations. Meanwhile, derived alleles of the younger SNP (rs3789228) and its linked SNPs, including the target site (rs12055879), in frequency class 9 were carried by a small number of haplotypes mostly from non-African populations. When haplotypes not to be observed in EAS are excluded, similar patterns appeared among SNPs in frequency class 10 and SNPs in frequency class 9.

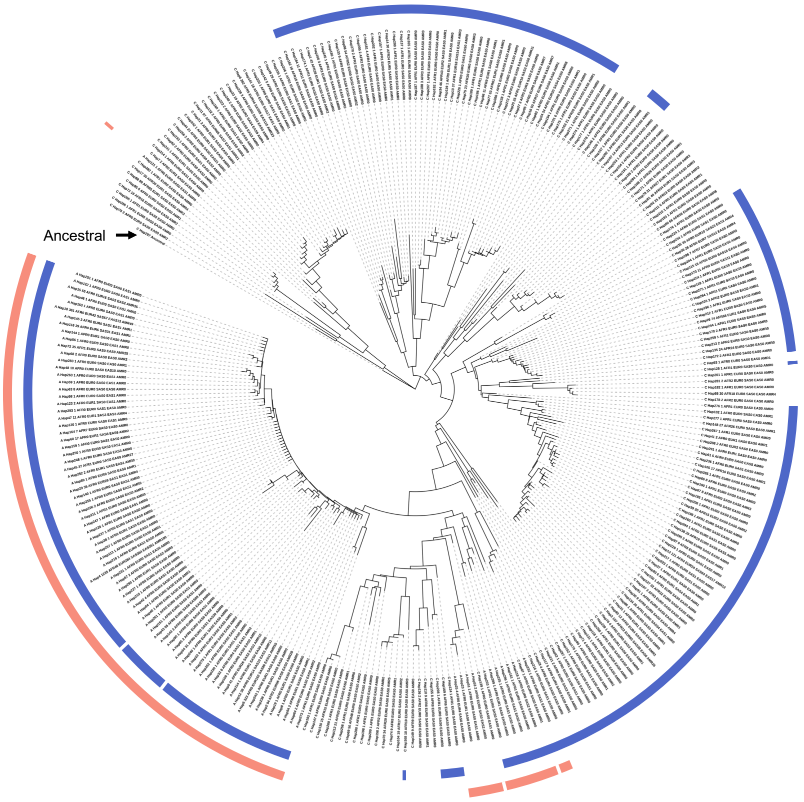

Figure 3-10 The NJ tree of the core region of rs3789228 on *DCDC2*, using 296 haplotypes from entire population of 1 KG. Blue lines indicate the D group of the core SNP. Red lines indicate the D group of the target site. The ancestral sequence as the root is pointed by the arrow.

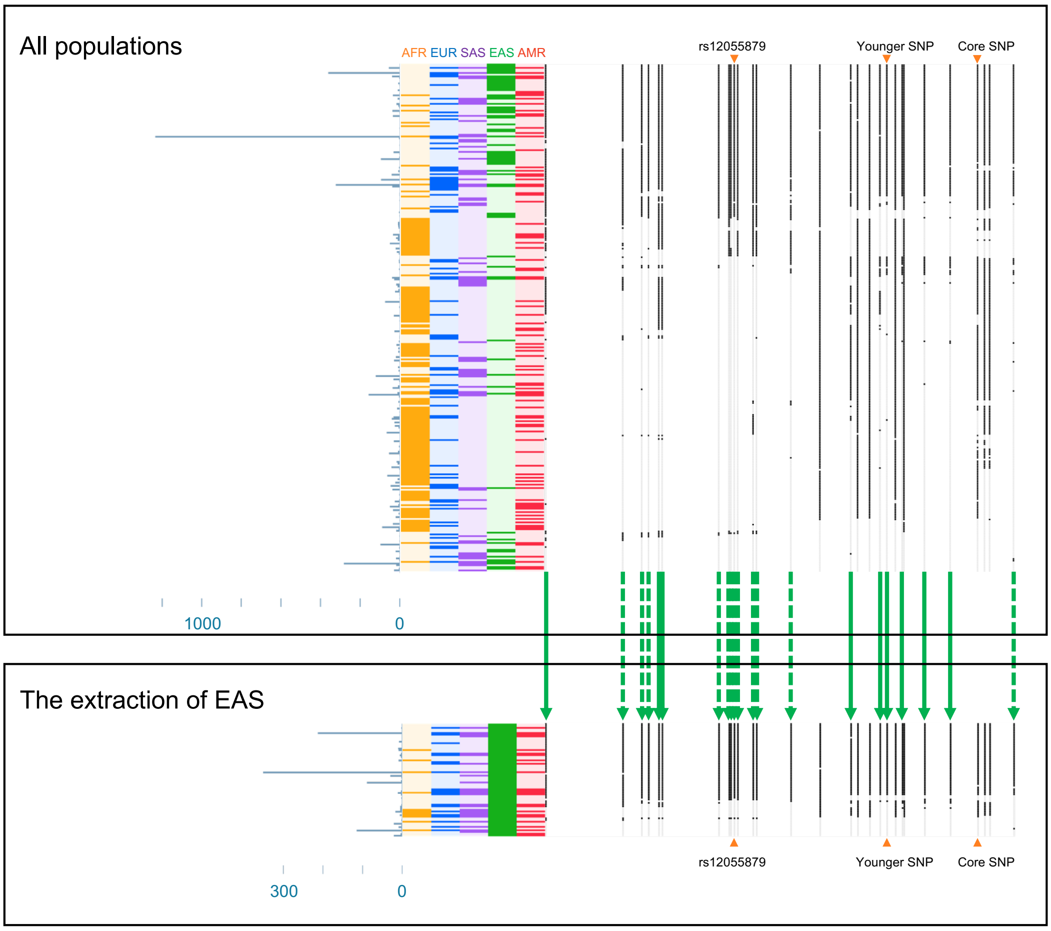

Figure 3-11 Top: haplotypes observed in all populations of 1KG for the core region of rs3789228, defined by SNPs with global MAF ≥ 0.005. Columns represent genomic positions, and rows represent haplotypes. Haplotypes were sorted according to the NJ tree. For each haplotype, a black or grey cell represents derived or ancestral allele, respectively, at the position. Only rs3789228 and rs1091047 and their linked SNPs are indicated. Bottom: haplotypes observed in EAS. SNPs with solid green arrows indicate SNPs linked to both rs3789228 and rs1091047 in global frequency class 9; SNPs with broken green arrows indicate SNPs linked only to rs3789228 in global frequency class 9; SNPs with no arrows indicate linked SNPs in global frequency class 10. The five colored columns represent presence/absence of sequences from the five populations of 1 KG for each haplotype. In the colored columns, dark tone indicates at least one sequence from the population is present in the haplotype while light color indicates no sequence from the population is observed in the haplotype. The blue bar lengths on the left side indicate the

counts of each haplotype.

Chapter 4

Discussion

4.1. General findings

In order to investigate whether natural selection has acted on the core SNPs of interest, I conducted two types of neutrality tests on the derived alleles: *nSL* (as an EHH-based test) and 2D SFS-based statistics. For most of the core SNPs, neither statistics detected any signatures of selective sweep, thus neutrality was not rejected. Previous studies found signs of natural selection on dyslexia-related genes by phylogenetic analyses (Hannula-Jouppi *et al.* 2005; Mozzi *et al.* 2016). A significant increase of derived allele frequencies were reported in some sites on dyslexia-related genes in modern human populations (Mozzi *et al.* 2016). While attempts to detect signatures of natural selection on dyslexia-related genes among modern human populations have been performed, my study focused on the SNPs that were reported to be associated with risk/non-risk for some traits related to an individual's reading ability in one of the writing systems. Because most people were not engaged in reading and writing until recently (Dalby 1986; Dehaene and Cohen 2007; Christiansen and Müller 2015), the genetic variations that my study focused on were unlikely to be maintained by natural selection, which is consistent with my results. Signs of acting natural selection were found on some alleles associated with autism spectrum disorder and schizophrenia (Polimanti and Gelernter 2017; Fujito *et al.* 2018a). Different from such traits, dyslexic traits should have been veiled until modern times. So, selective pressure on cognitive functions could be different between reading/writing and other traits. Nevertheless, the 2D SFS-based statistics suggested that two core regions could be under selective sweep. Because the selection target could be an SNP linked to a core SNP, I searched for the target site in these two exceptional cases.

4.2. The core region of rs17031962 on *GNPTAB*

The first case is the core region of rs17031962 on *GNPTAB*. The derived allele of this core SNP is the non-risk type (Chen *et al.* 2015). In addition to *GNPTAB*, this region also contains genes of *CHPT1* (partial) and *SYCP3*. I searched for the target site using the G_{c0} statistic and concluded that the target site could be an SNP (rs3751248) in one of the smallest regions with $p < 0.01$ (the fifth region), because it is located in an open chromatin region. However, even if this SNP has some biological function, it is still unknown which trait is affected. There are two possible scenarios where natural selection has acted on this SNP. The first scenario is the selection for the prior functions explained by the neuronal recycling hypothesis and cultural neural reuse (Dehaene 2005; Dehaene and Cohen 2007; d'Errico and Colagè 2018; Colagè and D'Errico 2020). In this scenario, the derived allele may have been selected for a prior function, and therefore, the derived allele was identified as the non-risk allele for the reading ability of Chinese characters. The second scenario is pleiotropy, which should also be considered. Although the core

SNP was associated with dyslexia of Chinese characters (Chen *et al.* 2015), *GNPTAB* has been found to be related to stuttering (Kang *et al.* 2010; Drayna and Kang 2011; Chen *et al.* 2015). Beyond functions related to language, this gene is involved in tagging for transport of lysosomal enzymes (Kang *et al.* 2010; Drayna and Kang 2011; Kang and Drayna 2012). In addition, Ebola virus was recently reported to utilize *GNPTAB* for efficient infection (Flint *et al.* 2019). If rs3751248, which I speculate to be the target site in this region, did not affect reading ability but instead some other function involving *GNPTAB*, then pleiotropy would explain this situation. However, it is unknown which gene is affected by a mutation on the target site (rs3751248). Because this SNP (rs3751248) is located in an open chromatin region, neither of the two scenarios can explain the case whereby the target site has a functional effect on genes other than *GNPTAB*. In such a case, my findings may be attributed to genetic hitchhiking, where alleles in dyslexia-related genes may increase their frequency together with the linked target site, which could have functions for other genes and traits other than reading ability. Thus, I consider this third scenario based on my results, and there may be other scenarios; however, it remains unclear which scenario actually occurred because of the current lack of understanding about the effect of mutations on the target site.

Although I focused on and analyzed only East Asian populations in this study with respect to applying neutrality test, it may be informative to look at the distribution of the derived allele among populations in the world. The derived alleles of the core SNP and its linked SNPs, such as rs3751248, are mainly observed in Asian populations (Figure 3-8), supporting the possibility of local adaptations (e.g., adaptations specific to Asian populations). Including the target site, the candidate SNPs in the fifth region were located in *SYCP3* and its upstream region. *SYCP3* is involved in the pairing and crossover of homologous chromosomes during meiosis (Yuan *et al.* 2000). Such a function should directly affect fitness, so a beneficial mutation in this gene could be selected for. Fundamentally, its effect on fitness should not only be for individuals in Asia but for individuals everywhere. Therefore, I consider that the trait under selection may not be related to meiosis, and that this gene region may be related to functions that are not yet elucidated.

4.3. The core region of rs3789228 on *DCDC2*

The second case is the core region of rs3789228 on *DCDC2*. This SNP was distinguished as the younger SNP to the core SNP of rs1091047, based on the global derived allele count. To date, there is no study investigating whether the derived allele of this younger SNP itself is risk or non-risk for dyslexia of Chinese characters, but the derived allele of the core SNP is a non-risk type (Su *et al.* 2015). Based on my analyses, the target site may be located in the first to second region, where both candidate SNPs showed $p < 0.001$. The second SNP (rs12055879) in this region is located in both the enhancer region and CTCF binding site; since this SNP may affect expression regulation, I speculate that rs12055879 is the target site in this core region. Like in the case of rs17031962 on *GNPTAB*, even if the target site has some biological function, it is unknown which gene is affected by mutations at this site and which trait is affected. Therefore, if natural selection has acted, any of the three scenarios mentioned above would also be possible for this case (i.e., pleiotropy, genetic hitchhiking, and selection for prior functions explained by neuronal recycling hypothesis and cultural neural reuse).

Looking at the distribution of the derived alleles among populations in the world (Figure 3-8) and the descriptions of haplotypes in the core region using samples from all populations (Figure 3-11), I found that derived alleles of the core SNP (rs1091047) and its linked SNPs in frequency class 10 were carried by various haplotypes containing sequences from both African and non-African populations. Meanwhile, derived alleles of the younger SNP (rs3789228) and its linked SNPs, including the target site (rs12055879), in frequency class 9 were carried by a small number of haplotypes predominantly from non-African populations. Therefore, these derived alleles may have spread after out of Africa migration. The derived allele frequency of the target site (rs12055879) seemed to be higher in East Asian populations than in other non-African populations (Figure 3-8). Interestingly, according to previous studies, the derived allele of the core SNP (rs1091047) was the non-risk type in the Chinese population, whereas the derived allele was the risk type in the European ancestry population where people use an alphabetic language (Lind *et al.* 2010; Su *et al.* 2015). However, I cannot infer whether the mutation on the target site itself has an effect on a certain prior function related to the reading ability of Chinese characters or not, because the effect of mutations on this target site has also not been explored.

Although the present study did not investigate the relationship between allele distribution and writing systems, there are cases showing a correlation between human genetic variation and certain features of the spoken language. The frequency of an allele group of the READ1 regulatory element in *DCDC2* was found to be positively correlated with the number of consonants (DeMille *et al.* 2018). Moreover, the frequency of particular haplotypes of *ASPM* and *Microcephalin* in populations was found to be correlated with use of linguistic tone (Dediu and Ladd 2007). *ASPM* and *Microcephalin* are genes related to brain size, and it is arguable whether they have or have not been under positive selection for brain growth (Evans *et al.* 2005; Mekel-Bobrov 2005; Currat *et al.* 2006; Yu *et al.* 2007).

4.4. The discrepancy in results between *nSL* and 2D SFS-based statistics

While 2D SFS-based statistics suggested that two core regions could be under selective sweep, this was not supported by the results of *nSL*. Several reasons could be considered for this discrepancy. One of the possibilities is recombination rate variation, which should affect the haplotype length (Sabeti 2006). In the core region of rs17031962 on *GNPTAB*, r^2 values with the core SNP sharply declined, especially in the upstream side (i.e., the region with a smaller genomic position number). This implies that the core region could be located very close to a recombination hotspot, which would weaken the signal of selective sweep detected using *nSL*. Additionally, the reduction of EHH in the downstream side seemed not to be caused by recombination. EHH of the D group of rs17031962 was dramatically broken down in the middle of the core region (Figure 4-1A), even though r^2 values with the core SNP were continue to be high (Figure 2-1) and therefore, linked SNPs were observed (Figure 4-1B). The reduction seemed to occur around rs11111017, rs7134161, rs80048426, and rs222511 (chr12: 102,166,802–

102,167,117 of GRCh37) (Figure 4-1A). By checking haplotypes observed in this core region, I found that derived alleles of these SNPs did not seem to be explained by mutations in single lineage. It seemed that mutations there frequently occur in parallel across haplotypes in any clades (Figure 4-1C). As a result, the EHH was broken down while LD is still observed beyond this area. Although mutations in this area, of course, also impacts on the A group (Figure 4-1A), the condition that the core region contains such area could affect the results of *nSL*. At present, I cannot determine what causes the observation on this area, but an explanation might be mutation hotspot. Different from *nSL*, 2D-SFS based statistics use pairwise r^2 with the core SNP, not extending homozygosity, and therefore the statistics would be less affected by the area.

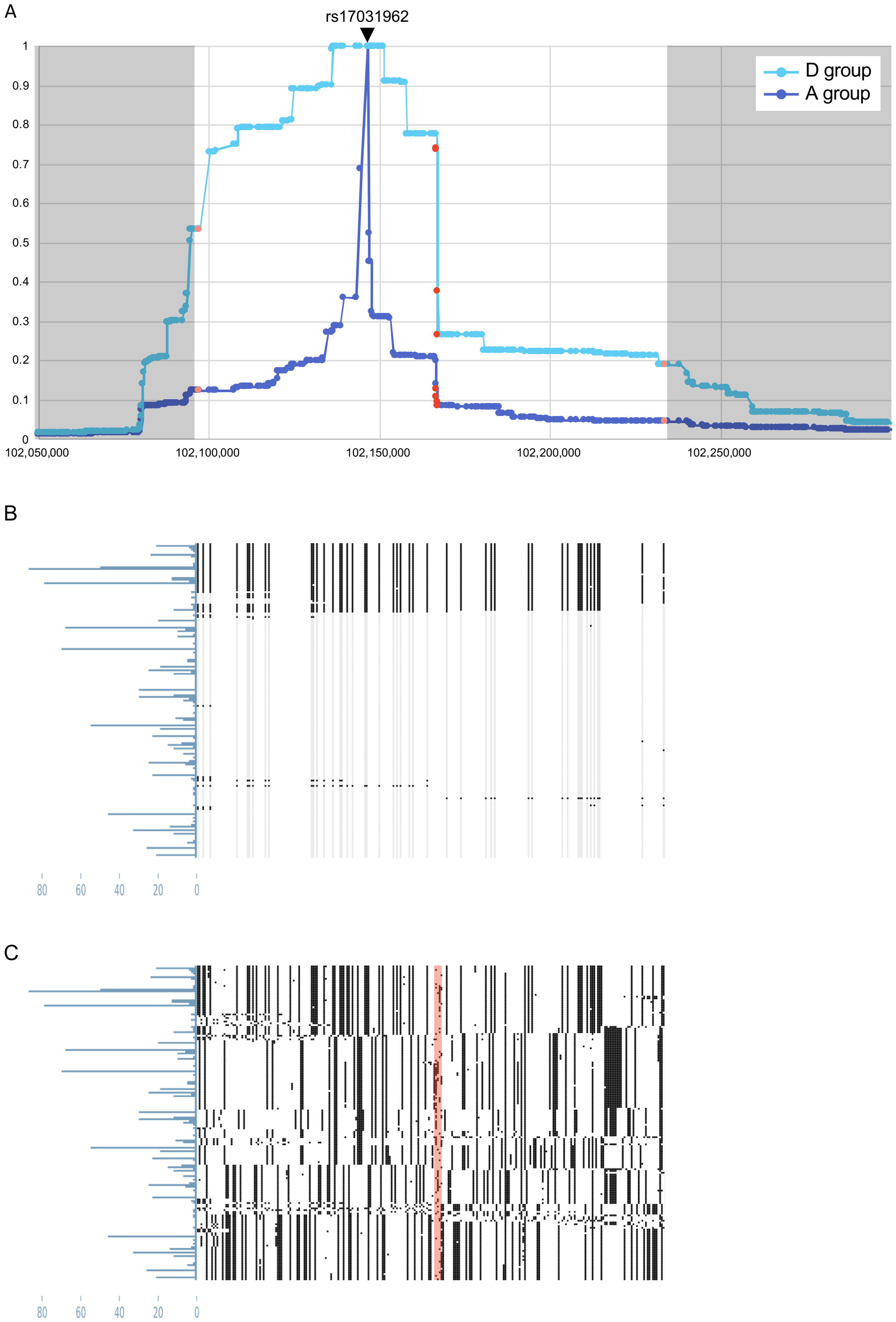

Figure 4-1 (**A**) EHH plot of D and A group of rs17031962. Pink dots indicate EHH value

of linked SNPs which are the boundaries of the core region. Outsides of the core region are shadowed. Red dots indicate EHH values of rs11111017, rs7134161, rs80048426, and rs222511. (**B**) Haplotypes observed in EAS-KPGP for the core region of rs17031962, using SNPs with MAF in EAS-KPGP ≥ 0.01. Columns represent genomic positions, and rows represent haplotypes. Haplotypes were sorted according to the NJ tree constructed. For each haplotype, a black or grey cell represents derived or ancestral allele, respectively, at the position. Only rs17031962 and its linked SNPs are indicated. The lengths of blue bars on the left side display the counts of each haplotype. (**C**) Haplotypes observed in EAS-KPGP for the core region of rs17031962, with all positions shown. Positions of rs11111017, rs7134161, rs80048426, and rs222511 are red-shadowed.

The other possibility for the discrepancy in results between *nSL* and 2D SFS-based statistics is that LD is broken down by recombination events over time, which renders it difficult to detect selection signals (Sabeti 2006). An SNP with a high derived allele frequency is assumed to have such a short LD. In addition, when the derived allele frequency is higher, the power of *nSL* declines in a subpopulation of structured populations (Vatsiou *et al.* 2016), such as populations in 1 KG. Although I only showed the results of the 15 core SNPs for *nSL*, I found that the result of *nSL* for rs3789228 (the younger SNP for the core SNP of rs1091047 on *DCDC2*) was also not significant (normalized *nSL* = 0.1851; *p* = 0.427). The derived allele frequency of rs3789228 in EAS is 77.6%, and therefore, this frequency could be relatively too high for *nSL* to detect sweep signals.

4.5. Conclusion

In this thesis, I investigated whether alleles of the SNPs associated with dyslexia of Chinese characters had evolved neutrally or not, applying two types of neutrality tests. While neutrality was not rejected for most of the core SNPs, the two exceptional cases were found. In particular, I searched for the target site in the two core regions, which could be under selective sweep. My study supported the possibility of genetic hitchhiking: The target sites could have functional effects on genes other than dyslexia-related genes, *GNPTAB* and *DCDC2*. The TMRCA of the D group of the target site of the two cases (the core region of rs17031962 on *GNPTAB* and the core region of rs3789228 on *DCDC2*) were estimated as 15,434±4,390 years and 22,236±5,635 years, respectively. Needless to say, the estimated TMRCA of the two cases are much older than the emergence of valuing literacy in modern times, and even older than the development of the earliest form of Chinese characters, mentioned in the General Introduction. This estimation of TMRCA would also support the reasoning that natural selection should not have acted on reading/writing ability itself.

I inferred the two of target sites because of the possibility of having functional effects. However, these effects are not biologically confirmed but were speculated based on annotation data. Future experiments are necessary to verify whether these target sites actually have a functional effect and which gene is affected. Conversely, I hope that the approach in this study could provide a priority for experimental studies to confirm the functions. The findings in my study should be the results seen only in my study populations, i.e., EAS in 1 KG. In order to check sampling effects, follow-up studies are required when other East Asian data become available. For further understanding the evolutionary history of the polymorphism examined in this study, it would be beneficial to apply neutrality tests to populations other than East Asian, or to analyze samples including archaic humans. In addition, although beyond my study, the validity of the association between core SNPs and reading ability in the previous studies would be uncertain and needs to be confirmed by replications.

With respect to the meaning of my study in social context, I attempted to consider the notion of disorders from the perspective of evolutionary studies through my thesis. In this study, for most of the core SNPs, *nSL* and 2D SFS did not detected any signatures of selection, which should be consistent with the reasoning that the genetic variations seen in dyslexia related genes were unlikely to be maintained by natural selection. Moreover, the two exceptional cases suggested the target sites could have a functional effect not on these dyslexia related genes but on other genes, and it also should be consistent with this reasoning. In this context, what are “risk/non-risk alleles” for dyslexia, one of neurodevelopmental disorders that is considered to be caused by impairments or disruptions in the brain development? Modern society has introduced public education and demands universal literacy (Dalby 1986; Protopapas and Parrila 2018). So, primarily, the environment of the modern society likely determines which allele is “risk” or “non-risk” for reading ability. Dyslexia should basically be a consequence of neutral variation. Even in the case where selection may have acted, the selected trait should be different from reading ability itself. Although my study suggested that “non-risk” alleles of the two case could have been selected, a study on ASD shows that “risk” alleles also could have been selected (Polimanti and Gelernter 2017). Modern society should be only a temporal environment in human evolutionary history, where some phenotypes or alleles are concerned. To be aware of it, I think evolutionary studies may be helpful.

The focus, results, and interpretation of my study would be compatible with the concept of neurodiversity, which I introduced in the Chapter 1. Through my study, I questioned the notion of neurodevelopmental disorders and the meaning of "risk/non-risk alleles" for one of the disorders (i.e., dyslexia), from the perspective of evolutionary studies. However, as I said in the Chapter 1, it could be problematic to bring the viewpoint of evolutionary studies to human society. The neurodiversity movement also has a political aspect, demanding the rights of certain neurological conditions (Jaarsma and Welin 2012). So, I think that an easy, positive conclusion should be inadequate. I will continue to think about the issue of science and society. I hope that, with recognizing ethical issues in evolutionary studies, this study could provide an example for thinking about some dimensions of human society, such as neurodevelopmental disorders, from the viewpoint of human evolutionary history.

Acknowledgments

I am grateful to many people for support of my PhD study. First of all, I would sincerely like to thank to my supervisor Dr. Jun Gojobori. When I started my PhD study, I had almost no knowledge of molecular evolution, population genetics, and even basic biology. He taught me knowledge of the research area from the basis and how interesting the research area was. He gave me deep knowledge and invaluable advice on my study. Without his support, guidance, and patience, my PhD study would not have been possible.

I would also like to thank to my supervisor Dr. Yoko Satta. She gave me essential advice, and supported me especially for conducting the analyses with 2D SFS-based statistics. The analyses would be the main part of my thesis. She helped me kindly anytime I asked, and her guidance was crucial for my study.

I would like to extend my thanks to Dr. Nozomi Mizushima. She is the supervisor of my sub-thesis. In addition to support for sub-thesis itself, her guidance gave me insight also for my PhD study because my PhD study was somewhat related to the topic in the sub-thesis. I would also like to express my thanks to Dr. Hisashi Ohtsuki. He gave me helpful advice for statistical analysis which I am weak. I am also grateful to Dr. Naoko Fujito. She helped me especially for analyses with 2D SFS-based statistics. I thank to Dr. Quintin Lau not only for correcting my English in my published paper, but also giving me helpful comments for PhD thesis.

I would like to express my thanks to Satta lab and ESB members for their support. In this department, people are interact beyond laboratories. So I could receive much help from various people in this department, for example, for learning programing. In ESB,

people take approaches to the evolution in various way, and the department also includes the research field of science and society. Such characteristic of this department broadened my horizons, and gave me meaningful insight.

References

American Psychiatric Association, 2013 *American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders Fifth Edition*.

Auton, A., G. R. Abecasis, D. M. Altshuler, R. M. Durbin, D. R. Bentley *et al.*, 2015 A global reference for human genetic variation. Nature 526: 68–74.

Balding, D. J., 2006 A tutorial on statistical methods for population association studies. Nat. Rev. Genet. 7: 781–791.

Baron-Cohen, S., 2017 Editorial Perspective: Neurodiversity – a revolutionary concept for autism and psychiatry. J. Child Psychol. Psychiatry Allied Discip. 58: 744–747.

Bates, T. C., P. A. Lind, M. Luciano, G. W. Montgomery, N. G. Martin *et al.*, 2010 Dyslexia and DYX1C1: Deficits in reading and spelling associated with a missense mutation. Mol. Psychiatry 15: 1190–1196.

Bishop, D. V. M., 2015 The interface between genetics and psychology: lessons from developmental dyslexia. Proc. R. Soc. B Biol. Sci. 282: 20143139.

Carrion-Castillo, A., B. Franke, and S. E. Fisher, 2013 Molecular Genetics of Dyslexia: An Overview. Dyslexia 19: 214–240.

Chen, H., G. Wang, J. Xia, Y. Zhou, Y. Gao *et al.*, 2014 Stuttering candidate genes DRD2 but not SLC6A3 is associated with developmental dyslexia in Chinese population. Behav. Brain Funct. 10: 29.

Chen, H., J. Xu, Y. Zhou, Y. Gao, G. Wang *et al.*, 2015 Association study of stuttering candidate genes GNPTAB, GNPTG and NAGPA with dyslexia in Chinese population. BMC Genet. 16: 7.

Christiansen, M. H., and R.-A. Müller, 2015 Cultural recycling of neural substrates during language evolution and development, pp. 675–682 in *The cognitive neurosciences V*, edited by M. S. Gazzaniga and G. R. Mangun. MIT Press, Cambridge, MA.

Cohen, P. T., 2000 A Buddha Kingdom in the Golden Triangle: Buddhist Revivalism and the Charismatic Monk Khruba Bunchum. Aust. J. Anthropol. 11: 141–154.

Colagè, I., and F. D'Errico, 2020 Culture: The Driving Force of Human Cognition. Top. Cogn. Sci. 12: 654–672.

Coller, B. S., 2019 Ethics of Human Genome Editing. Annu. Rev. Med. 70: 289–305.

Currat, M., L. Excoffier, W. Maddison, S. P. Otto, N. Ray *et al.*, 2006 Comment on "Ongoing Adaptive Evolution of ASPM, a Brain Size Determinant in Homo sapiens" and "Microcephalin, a Gene Regulating Brain Size, Continues to Evolve Adaptively in Humans." Science. 313: 172a-172a.

d'Errico, F., and I. Colagè, 2018 Cultural Exaptation and Cultural Neural Reuse: A Mechanism for the Emergence of Modern Culture and Behavior. Biol. Theory 13: 213–227.

Dalby, J. T., 1986 Note: An ultimate view of reading ability. Int. J. Neurosci. 30: 227–230.

Daniels, P. T., and D. L. Share, 2018 Writing System Variation and Its Consequences for Reading and Dyslexia. Sci. Stud. Read. 22: 101–116.

Davis, S., 2003 Premodern Flows in Postmodern China. Mod. China 29: 176–203.

Dediu, D., and M. H. Christiansen, 2016 Language Evolution: Constraints and

Opportunities From Modern Genetics. Top. Cogn. Sci. 8: 361–370.

Dediu, D., and D. R. Ladd, 2007 Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin. Proc. Natl. Acad. Sci. 104: 10944–10949.

Dehaene, S., 2005 Evolution of human cortical circuits for reading and arithmetic: The '"neuronal recycling"' hypothesis, pp. 133–157 in *From Monkey Brain to Human Brain*, edited by S. Dehaene, J. R. Duhamel, M. Hauser, and G. Rizzolatti. MIT Press, Cambridge, MA.

Dehaene, S., and L. Cohen, 2007 Cultural Recycling of Cortical Maps. Neuron 56: 384–398.

Dehaene, S., and L. Cohen, 2011 The unique role of the visual word form area in reading. Trends Cogn. Sci. 15: 254–262.

DeMille, M. M. C., K. Tang, C. M. Mehta, C. Geissler, J. G. Malins *et al.*, 2018 Worldwide distribution of the DCDC2 READ1 regulatory element and its relationship with phoneme variation across languages. Proc. Natl. Acad. Sci. 115: 4951–4956.

Drayna, D., and C. Kang, 2011 Genetic approaches to understanding the causes of stuttering. J. Neurodev. Disord. 3: 374–380.

Evans, P. D., S. L. Gilbert, N. Mekel-Bobrov, E. J. Vallender, J. R. Anderson *et al.*, 2005 Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. Science. 309: 1717–1720.

Ferrer-Admetlla, A., M. Liang, T. Korneliussen, and R. Nielsen, 2014 On Detecting

Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. Mol. Biol. Evol. 31: 1275–1291.

Fisher, S. E., and J. C. DeFries, 2002 Developmental dyslexia: genetic dissection of a complex cognitive trait. Nat. Rev. Neurosci. 3: 767–780.

Flint, M., P. Chatterjee, D. L. Lin, L. K. McMullan, P. Shrivastava-Ranjan *et al.*, 2019 A genome-wide CRISPR screen identifies N-acetylglucosamine-1-phosphate transferase as a potential antiviral target for Ebola virus. Nat. Commun. 10: 285.

Francks, C., S. Paracchini, S. D. Smith, A. J. Richardson, T. S. Scerri *et al.*, 2004 A 77-kilobase region of chromosome 6p22.2 is associated with dyslexia in families from the United Kingdom and from the United States. Am. J. Hum. Genet. 75: 1046–1058.

Fu, Y. X., 1995 Statistical Properties of Segregating Sites. Theor. Popul. Biol. 48: 172–197.

Fujito, N. T., Y. Satta, M. Hane, A. Matsui, K. Yashima *et al.*, 2018a Positive selection on schizophrenia-associated ST8SIA2 gene in post-glacial Asia (K. Iwamoto, Ed.). PLoS One 13: e0200278.

Fujito, N. T., Y. Satta, T. Hayakawa, and N. Takahata, 2018b A new inference method for detecting an ongoing selective sweep. Genes Genet. Syst. 93: 149–161.

Griffin, E., and D. Pollak, 2009 Student experiences of neurodiversity in higher education: insights from the BRAINHE project. Dyslexia 15: 23–41.

Griffiths, R. C., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Commun. Stat. Stoch. Model. 14: 273–295.

Grigorenko, E. L., 2001 Developmental dyslexia: An update on genes, brains, and environments. J. Child Psychol. Psychiatry Allied Discip. 42: 91–125.

Hannula-Jouppi, K., N. Kaminen-Ahola, M. Taipale, R. Eklund, J. Nopola-Hemmi *et al.*, 2005 The Axon Guidance Receptor Gene ROBO1 Is a Candidate Gene for Developmental Dyslexia. PLoS Genet. 1: e50.

Hansell, M., 2003 Chinese writing, pp. 156–165 in *The Sino-Tibetan languages*, edited by G. Thurgood and R. J. LaPolla. Routledge, London.

Harold, D., S. Paracchini, T. Scerri, M. Dennis, N. Cope *et al.*, 2006 Further evidence that the KIAA0319 gene confers susceptibility to developmental dyslexia. Mol. Psychiatry 11: 1085–1091.

Henneberry, A. L., G. Wistow, and C. R. McMaster, 2000 Cloning, Genomic Organization, and Characterization of a Human Cholinephosphotransferase. J. Biol. Chem. 275: 29808–29815.

Hoeft, F., A. Meyler, A. Hernandez, C. Juel, H. Taylor-Hill *et al.*, 2007 Functional and morphometric brain dissociation between dyslexia and reading ability. Proc. Natl. Acad. Sci. 104: 4234–4239.

Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.

Jaarsma, P., and S. Welin, 2012 Autism as a natural human variation: Reflections on the claims of the neurodiversity movement. Heal. Care Anal. 20: 20–30.

Juengst, E. T., 2009 *Population Genetic Research and Screening: Conceptual and Ethical Issues*. Oxford University Press.

Kaalund, S. S., E. N. Newburn, T. Ye, R. Tao, C. Li *et al.*, 2014 Contrasting changes in DRD1 and DRD2 splice variant expression in schizophrenia and affective disorders, and associations with SNPs in postmortem brain. Mol. Psychiatry 19: 1258–1266.

Kang, C., and D. Drayna, 2012 A role for inherited metabolic deficits in persistent developmental stuttering. Mol. Genet. Metab. 107: 276–280.

Kang, C., S. Riazuddin, J. Mundorff, D. Krasnewich, P. Friedman *et al.*, 2010 Mutations in the Lysosomal Enzyme–Targeting Pathway and Persistent Stuttering. N. Engl. J. Med. 362: 677–685.

Kere, J., 2014 The molecular genetics and neurobiology of developmental dyslexia as model of a complex phenotype. Biochem. Biophys. Res. Commun. 452: 236–243.

Kim, J., S. Jeon, J.-P. Choi, A. Blazyte, Y. Jeon *et al.*, 2020 The Origin and Composition of Korean Ethnicity Analyzed by Ancient and Present-Day Genome Sequences (N. Saitou, Ed.). Genome Biol. Evol. 12: 553–565.

Kim, J., J. A. Weber, S. Jho, J. Jang, J. Jun *et al.*, 2018 KoVariome: Korean National Standard Reference Variome database of whole genomes with comprehensive SNV, indel, CNV, and SV analyses. Sci. Rep. 8: 5677.

Kimura, M., and T. Ohta, 1973 The age of a neutral mutant persisting in a finite population. Genetics 75: 199–212.

Kirby, P., 2018 A brief history of dyslexia. Psychologist 31: 56–59.

Kong, R., S. Shao, J. Wang, X. Zhang, S. Guo *et al.*, 2016 Genetic variant in DIP2A gene is associated with developmental dyslexia in Chinese population. Am. J. Med.

Genet. Part B Neuropsychiatr. Genet. 171: 203–208.

Kumar, S., G. Stecher, and K. Tamura, 2016 MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. 33: 1870–1874.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27: 2987–2993.

Lim, C. K. P., C. S. H. Ho, C. H. N. Chou, and M. M. Y. Waye, 2011 Association of the rs3743205 variant of DYX1C1 with dyslexia in Chinese children. Behav. Brain Funct. 7: 16.

Lim, C. K.-P., A. M.-B. Wong, C. S.-H. Ho, and M. M.-Y. Waye, 2014 A common haplotype of KIAA0319 contributes to the phonological awareness skill in Chinese children. Behav. Brain Funct. 10: 23.

Lind, P. A., M. Luciano, M. J. Wright, G. W. Montgomery, N. G. Martin *et al.*, 2010 Dyslexia and DCDC2: normal variation in reading and spelling is associated with DCDC2 polymorphisms in an Australian population sample. Eur. J. Hum. Genet. 18: 668–673.

Loh, P.-R., P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A Reshef *et al.*, 2016 Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. 48: 1443–1448.

Martin, A., M. Kronbichler, and F. Richlan, 2016 Dyslexic brain activation abnormalities in deep and shallow orthographies: A meta-analysis of 28 functional neuroimaging studies. Hum. Brain Mapp. 37: 2676–2699.

Mascheretti, S., A. De Luca, V. Trezzi, D. Peruzzo, A. Nordio *et al.*, 2017 Neurogenetics of developmental dyslexia: From genes to behavior through brain neuroimaging and cognitive and sensorial mechanisms. Transl. Psychiatry 7: e987-15.

McBride, C. A., 2016 Is Chinese Special? Four Aspects of Chinese Literacy Acquisition that Might Distinguish Learning Chinese from Learning Alphabetic Orthographies. Educ. Psychol. Rev. 28: 523–549.

McBride, C., Y. Wang, and L. M.-L. Cheang, 2018 Dyslexia in Chinese. Curr. Dev. Disord. Reports 5: 217–225.

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie *et al.*, 2016 The Ensembl Variant Effect Predictor. Genome Biol. 17: 122.

Mekel-Bobrov, N., 2005 Ongoing Adaptive Evolution of ASPM, a Brain Size Determinant in Homo sapiens. Science. 309: 1720–1722.

Meredith, R. M., 2015 Sensitive and critical periods during neurotypical and aberrant neurodevelopment: A framework for neurodevelopmental disorders. Neurosci. Biobehav. Rev. 50: 180–188.

Mozzi, A., D. Forni, M. Clerici, U. Pozzoli, S. Mascheretti *et al.*, 2016 The evolutionary history of genes involved in spoken and written language: beyond FOXP2. Sci. Rep. 6: 22157.

Newbury, D., A. Monaco, and S. Paracchini, 2014 Reading and Language Disorders: The Importance of Both Quantity and Quality. Genes. 5: 285–309.

Ortega, F., 2009 The Cerebral Subject and the Challenge of Neurodiversity.

Biosocieties 4: 425–445.

Owen, R. W., 2017 A description and linguistic analysis of the tai khuen writing system. J. Southeast Asian Linguist. Soc. 10: 140–164.

Paaby, A. B., and M. V. Rockman, 2013 The many faces of pleiotropy. Trends Genet. 29: 66–73.

Paracchini, S., Q. W. Ang, F. J. Stanley, A. P. Monaco, C. E. Pennell *et al.*, 2011 Analysis of dyslexia candidate genes in the Raine cohort representing the general Australian population. Genes, Brain Behav. 10: 158–165.

Paracchini, S., T. Scerri, and A. P. Monaco, 2007 The Genetic Lexicon of Dyslexia. Annu. Rev. Genomics Hum. Genet. 8: 57–79.

Patterson, N., A. L. Price, and D. Reich, 2006 Population Structure and Eigenanalysis. PLoS Genet. 2: e190.

Peterson, R. L., and B. F. Pennington, 2012 Developmental dyslexia : The Lancet. Lancet 379: 1997–2007.

Peterson, R. L., and B. F. Pennington, 2015 Developmental Dyslexia. Annu. Rev. Clin. Psychol. 11: 283–307.

Plomin, R., C. M. A. Haworth, and O. S. P. Davis, 2009 Common disorders are quantitative traits. Nat. Rev. Genet. 10: 872–878.

Polimanti, R., and J. Gelernter, 2017 Widespread signatures of positive selection in common risk alleles associated to autism spectrum disorder (S. M. Williams, Ed.). PLOS Genet. 13: e1006618.

Price, C. J., and J. T. Devlin, 2011 The Interactive Account of ventral occipitotemporal contributions to reading. Trends Cogn. Sci. 15: 246–253.

Protopapas, A., and R. Parrila, 2019 Dyslexia: Still not a neurodevelopmental disorder. Brain Sci. 9: 9.

Protopapas, A., and R. Parrila, 2018 Is Dyslexia a Brain Disorder? Brain Sci. 8: 61.

Sabeti, P. C., 2006 Positive Natural Selection in the Human Lineage. Science. 312: 1614–1620.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406–25.

Satta, Y., W. Zheng, K. V. Nishiyama, R. L. Iwasaki, T. Hayakawa *et al.*, 2019 Two-dimensional site frequency spectrum for detecting, classifying and dating incomplete selective sweeps. Genes Genet. Syst. 94: 283–300.

Scally, A., and R. Durbin, 2012 Revising the human mutation rate: implications for understanding human evolution. Nat. Rev. Genet. 13: 745–753.

Scerri, T. S., A. P. Morris, L. L. Buckingham, D. F. Newbury, L. L. Miller *et al.*, 2011 DCDC2, KIAA0319 and CMIP are associated with reading-related traits. Biol. Psychiatry 70: 237–245.

Scerri, T. S., and G. Schulte-Körne, 2010 Genetics of developmental dyslexia. Eur. Child Adolesc. Psychiatry 19: 179–197.

Schaffner, S. F., 2005 Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15: 1576–1583.

Shao, S., R. Kong, L. Zou, R. Zhong, J. Lou *et al.*, 2016a The Roles of Genes in the Neuronal Migration and Neurite Outgrowth Network in Developmental Dyslexia:

Single- and Multiple-Risk Genetic Variants. Mol. Neurobiol. 53: 3967–3975.

Shao, S., Y. Niu, X. Zhang, R. Kong, J. Wang *et al.*, 2016b Opposite Associations between Individual KIAA0319 Polymorphisms and Developmental Dyslexia Risk across Populations: A Stratified Meta-Analysis by the Study Population. Sci. Rep. 6: 30454.

Sharma, P., and R. Sagar, 2017 Unfolding the genetic pathways of dyslexia in Asian population: A review. Asian J. Psychiatr. 30: 225–229.

Siok, W. T., Z. Niu, Z. Jin, C. A. Perfetti, and L. H. Tan, 2008 A structural-functional basis for dyslexia in the cortex of Chinese readers. Proc. Natl. Acad. Sci. 105: 5561–5566.

Siok, W. T., C. A. Perfetti, Z. Jin, and L. H. Tan, 2004 Biological abnormality of impaired reading is constrained by culture. Nature 431: 71–76.

Slatkin, M., and B. Rannala, 2000 Estimating Allele Age. Annu. Rev. Genomics Hum. Genet. 1: 225–249.

Slatkin, M., and B. Rannala, 1997 Estimating the age of alleles by use of intraallelic variability. Am. J. Hum. Genet. 60: 447–458.

Stearns, F. W., 2010 One Hundred Years of Pleiotropy: A Retrospective. Genetics 186: 767–773.

Su, M., J. Wang, U. Maurer, Y. Zhang, J. Li *et al.*, 2015 Gene–environment interaction on neural mechanisms of orthographic processing in Chinese children. J. Neurolinguistics 33: 172–186.

Sun, X., S. Song, X. Liang, Y. Xie, C. Zhao *et al.*, 2017 ROBO1 polymorphisms, callosal connectivity, and reading skills. Hum. Brain Mapp. 38: 2616–2626.

Szpiech, Z. A., and R. D. Hernandez, 2014 selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. Mol. Biol. Evol. 31: 2824–2827.

Taipale, M., N. Kaminen, J. Nopola-Hemmi, T. Haltia, B. Myllyluoma *et al.*, 2003 A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain. Proc. Natl. Acad. Sci. 100: 11553–11558.

Thapar, A., M. Cooper, and M. Rutter, 2017 Neurodevelopmental disorders. The Lancet Psychiatry 4: 339–346.

Thapar, A., and M. Rutter, 2015 Neurodevelopmental disorders, pp. 31–40 in *Rutter's Child and Adolescent Psychiatry*, John Wiley & Sons, Ltd, Chichester, UK.

Vatsiou, A. I., E. Bazin, and O. E. Gaggiotti, 2016 Detection of selective sweeps in structured populations: a comparison of recent methods. Mol. Ecol. 25: 89–103.

Vitti, J. J., M. K. Cho, S. A. Tishkoff, and P. C. Sabeti, 2012 Human evolutionary genomics: Ethical and interpretive issues. Trends Genet. 28: 137–145.

Wang, B., Y. Zhou, S. Leng, L. Zheng, H. Lv *et al.*, 2017 Genetic polymorphism of nonsyndromic cleft lip with or without cleft palate is associated with developmental dyslexia in Chinese school-aged populations. J. Hum. Genet. 62: 265–268.

Weiss, K. M., and B. W. Lambert, 2011 When the Time Seems Ripe: Eugenics, the Annals, and the Subtle Persistence of Typological Thinking. Ann. Hum. Genet. 75:

334–343.

Yu, F., R. S. Hill, S. F. Schaffner, P. C. Sabeti, E. T. Wang *et al.*, 2007 Comment on "Ongoing Adaptive Evolution of ASPM, a Brain Size Determinant in Homo sapiens." Science. 316: 370b-370b.

Yuan, L., J.-G. Liu, J. Zhao, E. Brundell, B. Daneholt *et al.*, 2000 The Murine SCP3 Gene Is Required for Synaptonemal Complex Assembly, Chromosome Synapsis, and Male Fertility. Mol. Cell 5: 73–83.

Yudell, B. M., D. Roberts, R. Desalle, and S. Tishkoff, 2016 Taking race out of human genetics. 16–18.

Zhang, Y., J. Li, S. Song, T. Tardif, M. Burmeister *et al.*, 2016 Association of DCDC2 Polymorphisms with Normal Variations in Reading Abilities in a Chinese Population (K. Paterson, Ed.). PLoS One 11: e0153603.

Zhang, Y., J. Li, T. Tardif, M. Burmeister, S. M. Villafuerte *et al.*, 2012 Association of the DYX1C1 Dyslexia Susceptibility Gene with Orthography in the Chinese Population (T. B. Penney, Ed.). PLoS One 7: e42969.