Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

A probabilistic method for testing and estimating selection differences between populations

YUNGANG HE, Minxian Wang, Xin Huang, et al.

Genome Res. published online October 13, 2015
Access the most recent version at doi:10.1101/gr.192336.115

P<P Published online October 13, 2015 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

A probabilistic method for testing and estimating selection differences between populations

Yungang He,[1] Minxian Wang,[1] Xin Huang,[1] Ran Li,[1]

Hongyang Xu,[1] Shuhua Xu,[1] Li Jin[1,2]

[1] Chinese Academy of Sciences Key Laboratory of Computational Biology, Chinese Academy of Sciences-Max Planck Society Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

[2] State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai 200433, China

Corresponding author: lijin@fudan.edu.cn (L.J.), yunganghe@picb.ac.cn (Y.H.)

Running title: Investigating selection differences of populations

Key words: natural selection, population divergence.

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

Abstract

Human populations around the world encounter various environmental challenges, and consequently develop genetic adaptations to different selection forces. Identifying the differences in natural selection between populations is critical for understanding the roles of specific genetic variants in evolutionary adaptation. Although numerous methods have been developed to detect genetic loci under recent directional selection, a probabilistic solution for testing and quantifying selection differences between populations is lacking. Here we report the development of a probabilistic method for testing and estimating selection differences between populations. Using a probabilistic model of genetic drift and selection, we showed that logarithm odds ratios of allele frequencies provide estimates of the differences in selection coefficients between populations. The estimates approximate a normal distribution and variance can be estimated using genome-wide variants. This allows us to quantify differences in selection coefficients and to determine the confidence intervals of the estimate. Our work also revealed the link between genetic association testing and hypothesis testing of selection differences. It therefore supplies a solution for hypothesis testing of selection differences. This method was applied to a genome-wide data analysis of Han and Tibetan populations. The results confirmed that both *EPAS1* and *EGLN1* genes are under statistically different selection in Han and Tibetan populations. We further estimated differences in the selection coefficients for genetic variants involved in melanin formation and determined their confidence intervals between continental population groups. Application of the method to empirical data demonstrated the outstanding capability of this novel approach for testing and quantifying differences in natural selection.

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

Introduction

When anatomically modern humans emerged from Africa (Mellars 2006), and subsequently colonized throughout the world (Mellars et al. 2013; Hellenthal et al. 2008), they encountered many challenges, including essential environmental alterations, food resource shifts, and infectious diseases (Hancock et al. 2010, 2011; Leffler et al. 2013). The current huge size and wide distribution of modern human populations demonstrates the evolutionary success of human beings, which intrigues and attracts geneticists to investigate the natural selection and genetic adaptation of human populations. Studies of natural selection, especially directional selection, focus mainly on beneficial heritable traits and related genetic alterations (Fu and Akey 2013; Williams 2008). In recent years, genetic alterations under directional selection have attracted more attention than ever before. Consequently, some highly irregular genetic variants were discovered and further explored using various approaches (Grossman et al. 2010; Xu et al. 2011; Vitti et al. 2013; Bhatia et al. 2011; Sabeti et al. 2007; Xiang et al. 2013; Kamberov et al. 2013; Sabeti et al. 2006).

Directional selection usually involves genetic adaptation to local environments. Comparison of selection differences between populations is therefore important in genetic studies of directional selection. Differences in allele frequencies are indicators of possible selection differences between populations. As a measure of frequency difference, genetic distance, such as F_{ST}, is the most popular statistic in studies of natural selection (Akey et al. 2002; Lewontin and Krakauer 1973). The two-dimensional site frequency spectrum (2D-SFS) method was also designed to compare frequency differences between

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

populations and thus to identify selection differences (Nielsen et al. 2009). Selection differences can also be detected by comparing selective sweeps in different populations, such as cross-population extended haplotype homozygosity (XP-EHH) and cross-population composite likelihood ratio methods (XP-CLR) (Chen et al. 2010; Sabeti et al. 2007). Unfortunately, these methods lack efficient strategies to identify statistical outliers from the 'background noise' of genetic drift. Theoretical distributions of these statistics are not known in closed-form expressions. All the methods determine confidence levels based on empirical data distribution or computer simulation with limited prior knowledge of the demographic history (Akey et al. 2002; Nielsen et al. 2009; Chen et al. 2010; Sabeti et al. 2007). Although computer simulation can handle complicated genetic scenarios, it is unlikely that the "real" population genetic history can be accurately represented in computer simulations (Teshima et al. 2006). Furthermore, existing approaches do not provide effective solutions to quantify selection differences between populations.

In this report, we present a probabilistic method for estimating and testing selection differences between populations. The theoretical distribution of the involved statistics is well known and easy to compute. It enables us to conduct strict hypothesis testing without tedious computer simulation. Our approach supplies estimates and their confidence intervals for differences in selection coefficients. To demonstrate the capability of our approach, we conducted statistical hypothesis testing on a whole-genome dataset including samples of Han Chinese and Tibetan populations. The results regarding the *EPAS1* and *EGLN1* genes rejected the null hypothesis and confirmed their

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

significant differences in selection between the populations. We further estimated differences in the selection coefficients between continental populations for genetic variants involved in melanin formation.

Result

Model

In a scenario with two populations, we assumed that populations *A* and *B* have the same ancestral population *O*. For a given locus, we denoted the frequencies of mutated allele in the three populations as p_O^m, p_A^m, and p_B^m while frequencies of wild-type allele as p_O^w, p_A^w, and p_B^w, respectively. In a deterministic approximation with selection, the difference of logarithm ratio of frequencies was determined by divergence time *t* and selection coefficient *s* in the populations, say $\log(\frac{p_A^m}{p_A^w}) - \log(\frac{p_o^m}{p_o^w}) = s_A \times t$ and $\log(\frac{p_B^m}{p_B^w}) - \log(\frac{p_o^m}{p_o^w}) = s_B \times t$. Therefore, the difference of selection coefficients with uncertainty can be presented as

$$\Phi = s_A - s_B = [\log(\frac{p_A^m}{p_A^w}) - \log(\frac{p_B^m}{p_B^w})] / t + \Omega \ ,$$

where $\Omega = \frac{1}{t}\sum_{i=1}^{t}[(\omega_{A,i}^w - \omega_{A,i}^m) - (\omega_{B,i}^w - \omega_{B,i}^m)]$ indicates uncertainty due to genetic drift (see supplementary for details)

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

Estimating

Numbers of chromosomes sampling from populations *A* and *B* with mutated alleles are denoted C_A^m and C_B^m with those carrying wild-type alleles are denoted as C_A^w and C_B^w. When population divergence time *t* is large, the general effect of genetic drift Ω will approximate a normal distribution with mean zero following the central limit theorem (Feller 1968). The differences in the strength of natural selection between populations *A* and *B* can be estimated as

$$\hat{\Phi} = E(s_B - s_A) = \frac{\log(Odds)}{t}\ , \qquad \text{(Eq.1)}$$

where $Odds = (C_A^m C_B^w)/(C_A^w C_B^m)$. Variance of the estimation could be calculated as

$$Var(\hat{\Phi}) = Var[\log(Odds)]/t^2 + Var(\Omega). \qquad \text{(Eq.2)}$$

Consequently, 95% confidence interval of the estimation is determined as $\hat{\Phi} \pm 1.96 \cdot std(\hat{\Phi})$.

For a neural locus *i*, we have $\hat{\Phi}_i^2 = Var[\log(Odds_i)]/t^2 + Var(\Omega)$. Therefore, when a sample has *n* neural loci and the *n* is large, the general effect of genetic drift between population *A* and *B* can be estimated as

$$\hat{Var}(\Omega) = median\{\hat{\Phi}_i^2 / 0.455 - Var[\log(Odds_i)]/t^2, n \geq i \geq 1\},$$

where the variance of the log-odds ratio could be effectively approximated as $Var[\log(Odds)] = 1/C_A^m + 1/C_B^w + 1/C_A^w + 1/C_B^m$.

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

Testing

It is straightforward to propose a statistic for natural selection of a candidate locus, as follows:

$$\delta = \hat{\Phi}^2 / Var(\hat{\Phi}) \ . \qquad \text{(Eq.3)}$$

Under the null hypothesis that differences in natural selection are absent, the statistic δ follows a central chi-square distribution with a degree of freedom = 1. Under the alternative hypothesis with a selection difference, the statistic δ has a noncentral chi-square distribution with non-centrality parameter $\hat{\Phi}^2$ and a degree of freedom = 1.

The aforementioned statistical test for a single candidate locus could be generalized for a scenario with multiple linked loci to boost its power for detecting differences. We can rewrite the statistic as

$$\delta = X' \Sigma^{-1} X$$

where X is a vector with elements $\{\hat{\Phi}_1, \hat{\Phi}_2, ..., \hat{\Phi}_m\}$ and Σ is the covariance matrix of the vector with elements

$$Cov(\hat{\Phi}_i, \hat{\Phi}_j) = \begin{cases} Cov[\log(odd_i), \log(odd_j)] + Var(\Omega), i = j \\ Cov[\log(odd_i), \log(odd_j)], i \neq j \end{cases}.$$

The covariance of two correlated log-odds ratios is given as (Bagos 2012)

$$Cov[\log(odd_i), \log(odd_j)] = \sum_k \sum_l \sum_m (-1)^{l-m} \left(\frac{C_{klm}}{C_{kl+} C_{k+m}} \right) .$$

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

The notations for the covariance calculation are defined in Table 1. When testing for multiple linked loci, the statistic δ approximates a central chi-square distribution under the null hypothesis and the degree of freedom is the same as the number of involved loci (De Maesschalck et al. 2000).

Connection with case-control studies and its statistical power

The theoretical framework presented above bears an intrinsic conceptual and statistical connection with population-based association studies, as presented in Figure 1. The left panel illustrates the conceptual framework of the null hypothesis and alternative hypothesis in a genetic association study with the population stratification described by Devlin et al. (2001) (Devlin et al. 2001). Genetic association studies detect indirect associations between genetic markers (G) and phenotype (Y) that are mediated by correlations between the genetic cause (X) and phenotype (Y). A special approach, such as genomic control (GC), is capable of eliminating sporadic associations due to genetic confounding effects (C). The right panel presents the framework of the null hypothesis and alternative hypothesis in our method. The difference between the panels is that our method focuses on the difference in selection (D) but not phenotype (Y). Our method can distinguish selection differences from genetic background noise (B). GC controls type I errors using an inflation factor λ, while our method considers differences in the genetic background in the variance calculation by introducing Var(Ω) (Eq.2). The GC method

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

remains the same as a regular association test if $\lambda \approx 1$; our method also degenerates to a regular association test if $\mathrm{Var}(\Omega)$ approximates zero.

As the statistic of our method follows a chi-square distribution, the statistical power of the research design can be conveniently calculated. In this study, with a given difference in selection coefficients of 5.0×10^{-3} per generation, we show examples to demonstrate how sample size, genetic drift, and divergence time contribute to the statistical power of our method. Given a population divergence time of 300 generations, our calculation indicates that the statistical power effectively increases with an increase in the sample size (Figure 2 A). With a sample size of 500 chromosomes for each of the paired populations and genetic drift per generation $\mathrm{Var}(\Omega) = 1.0\times10^{-6}$, the statistical power of our method is as high as 0.98 (Figure 2 A). The power increase, however, is limited with an increase in the sample size when genetic drift is large. With genetic drift per generation $\mathrm{Var}(\Omega) = 5.0\times10^{-5}$, power is only ~0.20, even if we have a sample size as large as 500 chromosomes for each population (Figure 2A).

We also investigated the relationship between the population divergence time and statistical power. The increase in power with an increase in the sample size is prominent when the divergence time of involved populations is small (power curve marked by 'o' or '*', Figure 2B). When the divergence time is large, however, an increase in the sample size has only a minor effect on statistical power (power curve marked by '◊' or 'x', Figure 2B). This could be due to the fact that accumulated genetic drift contributes

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

significantly to the statistic's variance in this scenario. Because our method is based on allele frequencies of individual loci, but not a strict selective sweep, it is especially helpful for studying a 'soft sweep'. Other selection-sweep based methods, such as XP-EHH and XP-CLR, cannot work as well without significant linkage disequilibrium. We therefore strongly suggest that both our method and selection-sweep based methods should be applied as complementary methods to selection identification. Furthermore, our method supplies an estimate for differences in selection coefficients, whereas the others do not.

Testing selection differences between Tibetan and Han genomes

We applied our method of hypothesis testing on genotype data of Tibetan and Han Chinese. Because several genetic loci are reported to be involved in adaptation to high-attitude, most of these were not further verified by strict hypothesis testing but solely by inspection in simulation-based inference. A QQ plot of our single-variant testing showed that the obtained p-value was well fitted to the expectation (Figure 3), suggesting that our theoretical model handled genetic divergence of populations well, as least for this example. In particular, population divergence between Han and Tibetan populations did not lead to inflation of the type I error in our hypothesis testing.

The criterion to declare a genome-wide statistical significance is given by p-value $\leq 1.0\times10^{-8}$ in this study. Nineteen variants of the *EPAS1* gene have p-values that fit the criterion (Figure 4A). This observation agrees with previous reports suggesting that the *EPAS1* gene plays a major role in the high-attitude adaptation of Tibetan people (Xu et al. 2011; Simonson et al. 2010; Peng et al. 2011). We also conducted our aforementioned

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

multi-variant analysis on single nucleotide polymorphism (SNP) bins with different sizes. With a bin size of 5, 10, or 15 SNPs, SNP bins in the *EGLN1* gene region showed significant selection differences between the populations in our genome-wide hypothesis testing (Figure 4B). The power increase was consistent with previous reports that a multi-variant analysis could be more powerful than a single-variant approach in statistical tests of genetic data (Akey et al. 2001; He et al. 2011). These results support previous findings that both *EPAS1* and *EGLN1* genes are critical to high-attitude adaptation of the Tibetan population (Lorenzo et al. 2014). We obtained no positive findings in other gene regions, except for *EPAS1* and *EGLN1*. Other reported candidate genes should be further verified when more genetic data becomes available.

Estimating differences in selection

We estimated differences in the selection coefficients for several genetic variants of melanin formation between continental populations (Table 2). In this study, we assumed a simplified stepping stone model with five worldwide population groups (Figure 5). Selection differences were compared between neighboring groups while mutated alleles of the ancestral group served as a reference to determine the direction of the selection differences (Eq. 1).

Our estimations and their 95% confidence intervals suggested that most of the involved variants had similar selection coefficients in south and north Eurasian groups, except variant rs12913832 of the *OCA2* gene had an obvious difference $\hat{\Phi} = 4.87 \times 10^{-3}$ (Figure 6). For south and north Eurasian groups, 95% confidence intervals of the estimations were larger than those of population-group pairs involving both African and

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

non-African groups (Figure 6). This finding indicated that sampling variance contributed to the variance of the estimations of the south and north Eurasian groups. Therefore, the estimations could be further improved by increasing the sample sizes. Selection coefficients had only minor differences between Asians and Africans (Figure 6), suggesting that there are other genetic variants having a critical role in melanin formation in Asians (Edwards et al. 2010). The observed directions of the selection differences suggest that mutated alleles of the variants involved in melanin formation were more favorably selected in non-African populations (Wilde et al. 2014).

Discussion

We measured the differences in allele frequencies between populations using their logarithm odds ratios. Because genetic association studies usually present the effect size of risk alleles in odds ratios with estimated confidence intervals, this study revealed a statistical connection between our approach and classical genetic association studies. The close connection further allowed us the opportunity to explore natural selection in a genome-wide statistical test. There are other statistics with statistical properties better than logarithm odds ratio, especially when sample size is limited. As we present in this report, however, the logarithm odds ratio is an estimate of differences in the selection coefficient while the other statistics lack a direct connection with selection difference. Further, performance of logarithm odds ratio was acceptable in the presented case of the Han-Tibetan comparison, demonstrating the merits of logarithm odds ratio. When population divergence is small, variance of our estimate is due mainly to sampling

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

variance but not genetic drift (Eq.2, Figure 2). It is therefore possible to significantly improve the power of the statistical test by increasing the sample sizes. In this scenario, the benefit introduced by the large sample size is similar to that in genetic association studies. The statistical power of the hypothesis test using our method can be calculated for a specified study design. This provides a great advantage for determining the technical details of a research design, especially for determining sample sizes. When the evaluated locus is neutral in one of the two involved populations, our method provides estimation for selection coefficient in the rest population.

In our genetic model, the overall effects of demographic impact are summarized by variance in genetic drift (Eq.2). It is therefore unnecessary to separately consider the scale and duration of each demographic event in the analysis. The scales and durations of demographic events of the populations are often unknown, although some consensus has been reached in the research community. Tedious computer simulation is unnecessary in our approach while simulation is the only way to determine the confidence level in most previous reports. This simulation-free feature is a significant advantage for selection studies because “real” population history is unlikely to be accurately represented by computer simulation. It should be noted that our method of modeling genetic drift differs from the Wright-Fisher process. We use total variance $Var(\Omega)$ to capture the overall effect of genetic drift but not effective sample sizes.

There are other statistics that measure the differences in allele frequencies between populations, such as F_{ST} and ΔDAF. Both F_{ST} and ΔDAF have been applied to studies of natural selection (Akey et al. 2002). There is a close relationship between our logarithm

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

odds ratio with F_{ST} and ΔDAF. When F_{ST} or the absolute value of ΔDAF is larger, we generally have a larger positive or smaller negative logarithm odds ratio. Theoretical distributions of the F_{ST} and ΔDAF statistics, however, are not available in straightforward approaches. It therefore hinders their application in testing of selection difference. Furthermore, in the presence of population stratification, there is no convenient approach for quantifying the contribution of natural selection to F_{ST} and ΔDAF of individual variants. There lack a perfect quantitative correlation between the statistics.

In our genetic model, we considered only the mutations that occurred before the population stratification. This assumption holds for most genetic variants of the human genome, given its short evolutionary history. Our method is therefore applicable to populations with limited genetic divergence (Figure 2). When the frequency of an allele is low and the sample size is small, minor alleles may be missing from the samples. In these cases, we suggested a continuity correction in the calculation of the logarithm odds ratios and the variance (Friedrich et al. 2007). Consequently, differences in selection coefficients may be underestimated in this scenario. This potentially biased estimation could be partially improved in two ways. First, a larger sample size may be helpful for counting the minor allele; second, Bayesian estimation may be helpful for determining the frequency of the missing allele.

To summarize, we developed a probabilistic method for testing and estimating selection differences between populations. This method offers a statistical solution to study directional selection without tedious computer simulation. It is very powerful when the populations under investigation have close genetic connection. This method can be

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

used to quantify differences in selection coefficients, but not genotype fitness. Efficient estimation of genotype fitness remains a difficult task when no time-serial data is available.

Material and Methods

Data

Genotype data for 137 Han Chinese and 123 Tibetan unrelated individuals from three previous studies of human high-altitude adaptation were analyzed in this report (Wuren et al. 2014; Xu et al. 2011; Xing et al. 2013). All involved individuals were genotyped using Affymetrix Genome-Wide Human SNP Array 6.0. To investigate differences in the selection of genetic variants involved in melanin formation, genotype data of worldwide populations were downloaded from the website of the 1000 Genomes Project (1KG) (The 1000 Genomes Project Consortium 2010).

Computing

Haplotypes of the individuals were reconstructed using BEAGLE (Version 4.0) (Browning and Browning 2007). Other computing works of this report were conducted in R (version 2.14.2) (R Core Team 2015), a free software environment for statistical computing and graphics.

Acknowledgements

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

We thank three anonymous reviewers for their comments to improve this work. This work was supported by grants from National Natural Science Foundation of China (91331109 and 31171279 to Y.H.; 31271338 and 31330038 to L.J.; 91331204 and 31171218 to S.X.). S.X. was also supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040100). L.J. was also supported by Shanghai Leading Academic Discipline Project (B111) and the Center for Evolutionary Biology at Fudan University. Y.H. was also grateful for the supports of SA-SIBS scholarship program and the Youth Innovation Promotion Association of Chinese Academy of Science.

Competing interest statement

The authors have declared that no competing interests exist.

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure Legends

Figure 1 Conceptual framework of statistical tests for an association study and our method. Left panel shows conceptual framework of an association study in the presence of population stratification. Right panel shows the conceptual framework of our method in the same manner. Upper figures show conceptual frameworks for the null (H0) hypothesis; lower figures show conceptual frameworks of the alternative (H1) hypothesis.

Figure 2 Statistical power of our single-variant method increasing with an increase in the sample size. Statistical power is represented by the y-axis and sizes of the involved haplotypes are represented by the x-axis. Allele frequency of one population was given to be constant at 0.9, and frequency of the other population was determined by differences in selection coefficients of 5.0×10^{-3} per generation and divergence time. A. Power curve with a constant divergence time of 300 generations is marked by different symbols for different drift variances ('o' for Var(Ω) $=1.0\times10^{-6}$, '*' for Var(Ω) $=5.0\times10^{-6}$, '◊' for Var(Ω) $=1.0\times10^{-5}$, 'x' for Var(Ω) $=2.0\times10^{-5}$) B. Power curve with constant drift variance Var(Ω) $=5.0\times10^{-6}$ is marked by different symbols for different divergence times ('o' for t =100 generations, '*' for t =300 generations, '◊' for t =600 generations, 'x' for t =1000 generations)

Figure 3 QQ plot of single-variant analysis of Han-Tibetan data. Observed significance levels are represented by the y-axis on a scale of –log10(p-value). Expected quartile is represented by the x axis on the same scale.

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 4 Manhattan plots of significance levels for analysis of Han-Tibetan data. Chromosomes are shown on the x-axis; y-axis shows significance levels in –log10(P). A. Manhattan plot of single-variant analysis of all autosomes. B. Manhattan plots of single- and multi-variant analysis of chromosome 1 and 2. Bin sizes are shown under the x axis.

Figure 5 Simplified stepping-stone model of five population groups. Details of genetic demographic history were ignored in the model, such as backward gene flows and genetic admixture, etc. The five continental population groups are North Eurasian (NEU), South Eurasian (SEU), African (AFR), and Asian (ASN). Divergence of African and non-African groups was assumed to be 5000 generations. We further assumed that NEU and SEU have a divergence time of 400 generations.

Figure 6 Differences in selection coefficients between population groups. Estimated differences in selection coefficients are represented by the y-axis. Error bars indicate 95% confidence interval. Estimation for each neighboring group pair is marked by a group name and allele-frequency pie chart of the corresponding descendant group. Frequency of the derived allele is represented by the light color in the pie chart. North Eurasian population (NEU) is a combination of the 1000 genome population CEU, FIN, and GBR; South Eurasian population (SEU) is a combination of population IBS and TSI; African population (AFR) is combination of population YRI and LWK; Asian population (ASN) is a combination of population CHB, CHS, and JPT.

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

Table 1 Notations for the covariance calculation. C_{klm} is the haplotype count from population '*k*' which carries alleles in states '*l*' and '*m*' at locus 1 and 2, respectively.

		Locus 1			
		l=1		*l*=0	
		Locus 2			
		m=1	*m*=0	*m*=1	*m*=0
Pop A	*k*=1	C_{111}	C_{110}	C_{101}	C_{100}
Pop B	*k*=0	C_{011}	C_{010}	C_{001}	C_{000}

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

Table 2 Candidate sites involved in our estimation. 'AA' indicates ancestral allele and 'DA' indicates derived allele.

Gene	Chr	rs#	Coordinate	AA	DA	Reference
OCA2	15	rs12913832	28365618	A	G	(Eiberg et al. 2008; Sturm et al. 2008)
TYRP1	9	rs1408799	12672097	T	C	(Nan et al. 2009; Pośpiech et al. 2014)
TYR	11	rs1042602	88911696	C	A	(Durso et al. 2014; Pośpiech et al. 2014)
DCT	13	rs1407995	95096013	T	C	(Edwards et al. 2010; Zhu et al. 2007)
SLC24A5	15	rs1426654	48426484	G	A	(Tekola-Ayele et al. 2014; Basu Mallick et al. 2013; Durso et al. 2014)
SLC45A2	5	rs16891982	33951693	C	G	(Durso et al. 2014; Fernandez et al. 2008; Branicki et al. 2008)

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

References

Akey J, Jin L, Xiong M. 2001. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet EJHG* **9**: 291–300.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* **12**: 1805–1814.

Bagos PG. 2012. On the covariance of two correlated log-odds ratios. *Stat Med* **31**: 1418–1431.

Basu Mallick C, Iliescu FM, Möls M, Hill S, Tamang R, Chaubey G, Goto R, Ho SYW, Gallego Romero I, Crivellaro F, et al. 2013. The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. *PLoS Genet* **9**: e1003912.

Bhatia G, Patterson N, Pasaniuc B, Zaitlen N, Genovese G, Pollack S, Mallick S, Myers S, Tandon A, Spencer C, et al. 2011. Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. *Am J Hum Genet* **89**: 368–381.

Branicki W, Brudnik U, Draus-Barini J, Kupiec T, Wojas-Pelc A. 2008. Association of the SLC45A2 gene with physiological human hair colour variation. *J Hum Genet* **53**: 966–971.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**: 1084–1097.

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. *Genome Res* **20**: 393–402.

De Maesschalck R, Jouan-Rimbaud D, Massart DL. 2000. The Mahalanobis distance. *Chemom Intell Lab Syst* **50**: 1–18.

Devlin B, Roeder K, Wasserman L. 2001. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* **60**: 155–166.

Durso DF, Bydlowski SP, Hutz MH, Suarez-Kurtz G, Magalhães TR, Pena SDJ. 2014. Association of genetic variants with self-assessed color categories in Brazilians. *PloS One* **9**: e83926.

Edwards M, Bigham A, Tan J, Li S, Gozdzik A, Ross K, Jin L, Parra EJ. 2010. Association of the OCA2 polymorphism His615Arg with melanin content in east Asian populations: further evidence of convergent evolution of skin pigmentation. *PLoS Genet* **6**: e1000867.

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

Eiberg H, Troelsen J, Nielsen M, Mikkelsen A, Mengel-From J, Kjaer KW, Hansen L. 2008. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum Genet* **123**: 177–187.

Feller W. 1968. *An Introduction to Probability Theory and Its Applications, Vol. 1*. 3rd edition, pp244. Wiley, New York.

Fernandez LP, Milne RL, Pita G, Avilés JA, Lázaro P, Benítez J, Ribas G. 2008. SLC45A2: a novel malignant melanoma-associated gene. *Hum Mutat* **29**: 1161–1167.

Friedrich JO, Adhikari NKJ, Beyene J. 2007. Inclusion of zero total event trials in meta-analyses maintains analytic consistency and incorporates all available data. *BMC Med Res Methodol* **7**: 5.

Fu W, Akey JM. 2013. Selection and adaptation in the human genome. *Annu Rev Genomics Hum Genet* **14**: 467–489.

Grossman SR, Shlyakhter I, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, et al. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**: 883–886.

Hancock AM, Alkorta-Aranburu G, Witonsky DB, Di Rienzo A. 2010. Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Philos Trans R Soc Lond B Biol Sci* **365**: 2459–2468.

Hancock AM, Witonsky DB, Alkorta-Aranburu G, Beall CM, Gebremedhin A, Sukernik R, Utermann G, Pritchard JK, Coop G, Di Rienzo A. 2011. Adaptations to climate-mediated selective pressures in humans. *PLoS Genet* **7**: e1001375.

Hellenthal G, Auton A, Falush D. 2008. Inferring human colonization history using a copying model. *PLoS Genet* **4**: e1000078.

He Y, Li C, Amos CI, Xiong M, Ling H, Jin L. 2011. Accelerating haplotype-based genome-wide association study using perfect phylogeny and phase-known reference data. *PloS One* **6**: e22097.

Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H, et al. 2013. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* **152**: 691–702.

Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, Bowden R, Bontrop R, Wall JD, Sella G, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**: 1578–1582.

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

Lewontin RC, Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**: 175–195.

Lorenzo FR, Huff C, Myllymäki M, Olenchock B, Swierczek S, Tashi T, Gordeuk V, Wuren T, Ri-Li G, McClain DA, et al. 2014. A genetic mechanism for Tibetan high-altitude adaptation. *Nat Genet* **46**: 951–956.

Mellars P. 2006. Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc Natl Acad Sci U S A* **103**: 9381–9386.

Mellars P, Gori KC, Carr M, Soares PA, Richards MB. 2013. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. *Proc Natl Acad Sci U S A* **110**: 10699–10704.

Nan H, Kraft P, Hunter DJ, Han J. 2009. Genetic variants in pigmentation genes, pigmentary phenotypes, and risk of skin cancer in Caucasians. *Int J Cancer J Int Cancer* **125**: 909–917.

Nielsen R, Hubisz MJ, Hellmann I, Torgerson D, Andrés AM, Albrechtsen A, Gutenkunst R, Adams MD, Cargill M, Boyko A, et al. 2009. Darwinian and demographic forces affecting human protein coding genes. *Genome Res* **19**: 838–849.

Peng Y, Yang Z, Zhang H, Cui C, Qi X, Luo X, Tao X, Wu T, Ouzhuluobu null, Basang null, et al. 2011. Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Mol Biol Evol* **28**: 1075–1081.

Pośpiech E, Wojas-Pelc A, Walsh S, Liu F, Maeda H, Ishikawa T, Skowron M, Kayser M, Branicki W. 2014. The common occurrence of epistasis in the determination of human pigmentation and its impact on DNA-based pigmentation phenotype prediction. *Forensic Sci Int Genet* **11**: 64–72.

R Core Team. 2015. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES. 2006. Positive natural selection in the human lineage. *Science* **312**: 1614–1620.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**: 913–918.

Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Bai Z, Lorenzo FR, Xing J, Jorde LB, et al. 2010. Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**: 72–75.

Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

Sturm RA, Duffy DL, Zhao ZZ, Leite FPN, Stark MS, Hayward NK, Martin NG, Montgomery GW. 2008. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am J Hum Genet* **82**: 424–431.

Tekola-Ayele F, Adeyemo A, Chen G, Hailu E, Aseffa A, Davey G, Newport MJ, Rotimi CN. 2014. Novel genomic signals of recent selection in an Ethiopian population. *Eur J Hum Genet EJHG*.

Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Res* **16**: 702–712.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annu Rev Genet* **47**: 97–120.

Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, Hollfelder N, Potekhina ID, Schier W, Thomas MG, et al. 2014. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc Natl Acad Sci U S A* **111**: 4832–4837.

Williams GC. 2008. *Adaptation and natural selection: a critique of some current evolutionary thought*. Princeton University Press.

Wuren T, Simonson TS, Qin G, Xing J, Huff CD, Witherspoon DJ, Jorde LB, Ge R-L. 2014. Shared and unique signals of high-altitude adaptation in geographically distinct Tibetan populations. *PloS One* **9**: e88252.

Xiang K, Ouzhuluobu null, Peng Y, Yang Z, Zhang X, Cui C, Zhang H, Li M, Zhang Y, Bianba null, et al. 2013. Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. *Mol Biol Evol* **30**: 1889–1898.

Xing J, Wuren T, Simonson TS, Watkins WS, Witherspoon DJ, Wu W, Qin G, Huff CD, Jorde LB, Ge R-L. 2013. Genomic analysis of natural selection and phenotypic variation in high-altitude mongolians. *PLoS Genet* **9**: e1003634.

Xu S, Li S, Yang Y, Tan J, Lou H, Jin W, Yang L, Pan X, Wang J, Shen Y, et al. 2011. A genome-wide search for signals of high-altitude adaptation in Tibetans. *Mol Biol Evol* **28**: 1003–1011.

Zhu G, Montgomery GW, James MR, Trent JM, Hayward NK, Martin NG, Duffy DL. 2007. A genome-wide scan for naevus count: linkage to CDKN2A and to other chromosome regions. *Eur J Hum Genet EJHG* **15**: 94–102.

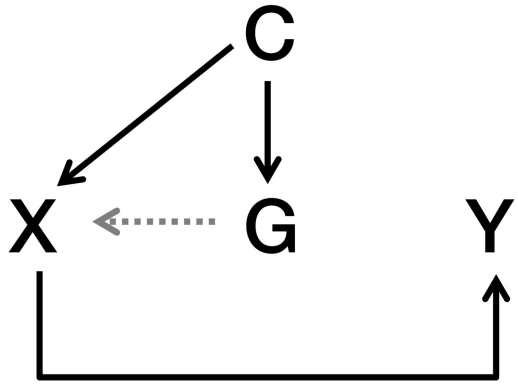
Downloaded from genome.cshlp.org on June 15, 2026 . Published by Cold Spring Harbor Laboratory Press

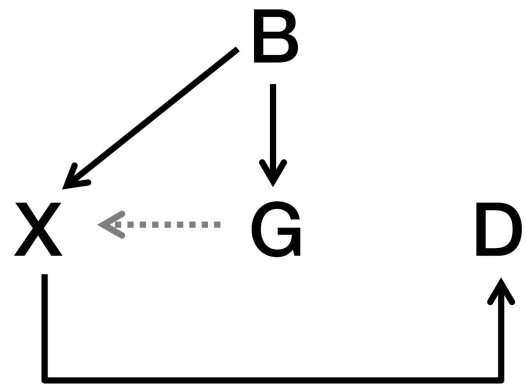
1

Test of genetic association

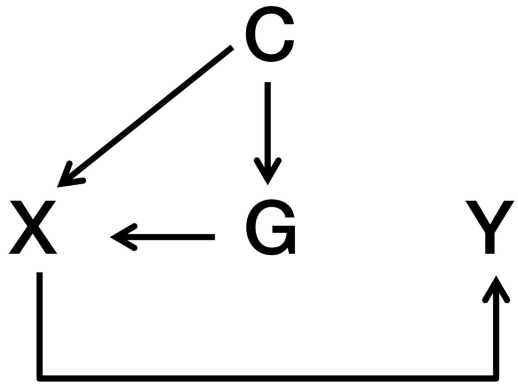
Test of selection difference

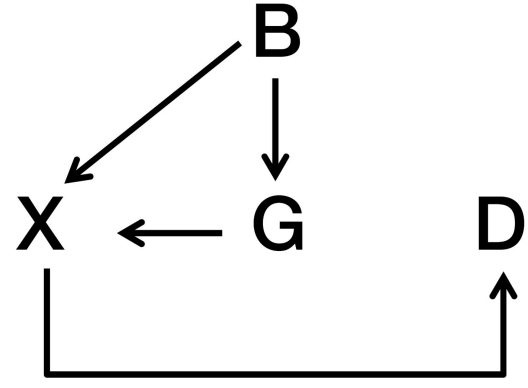
H0

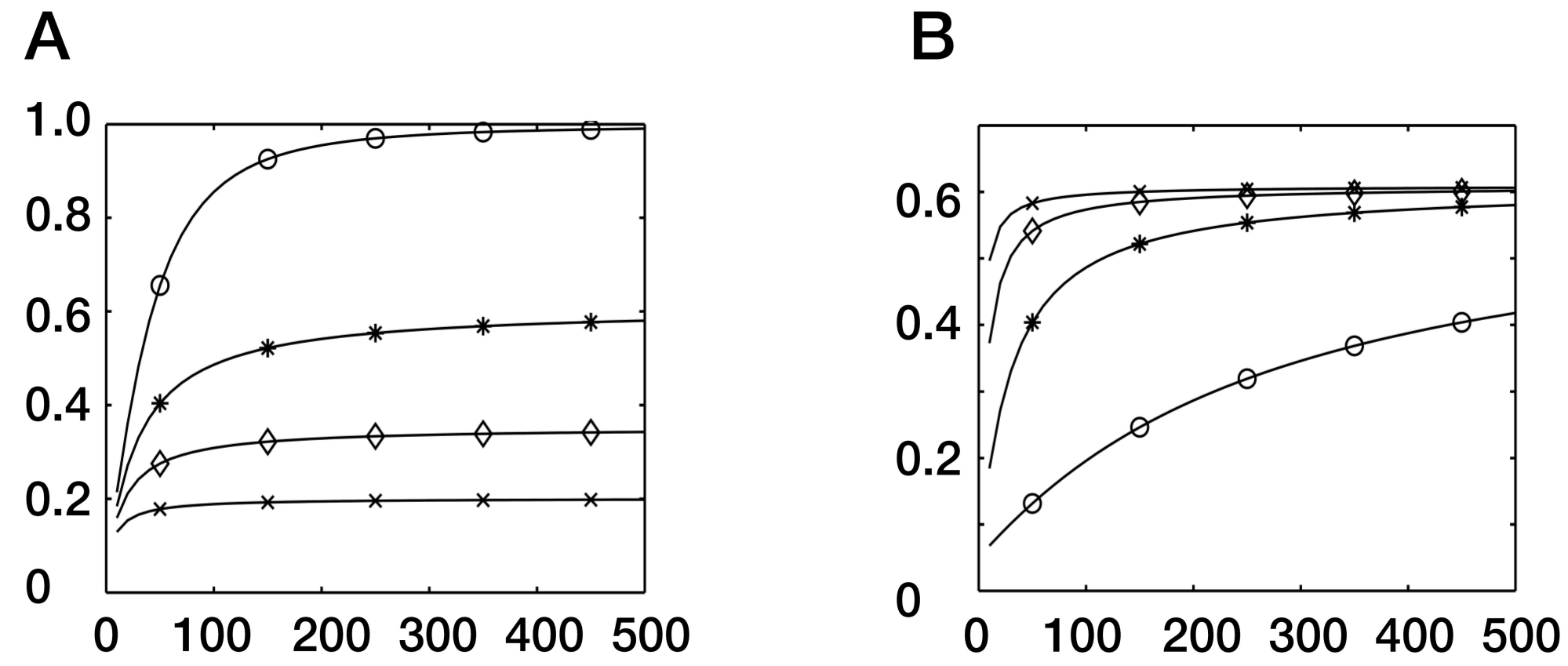


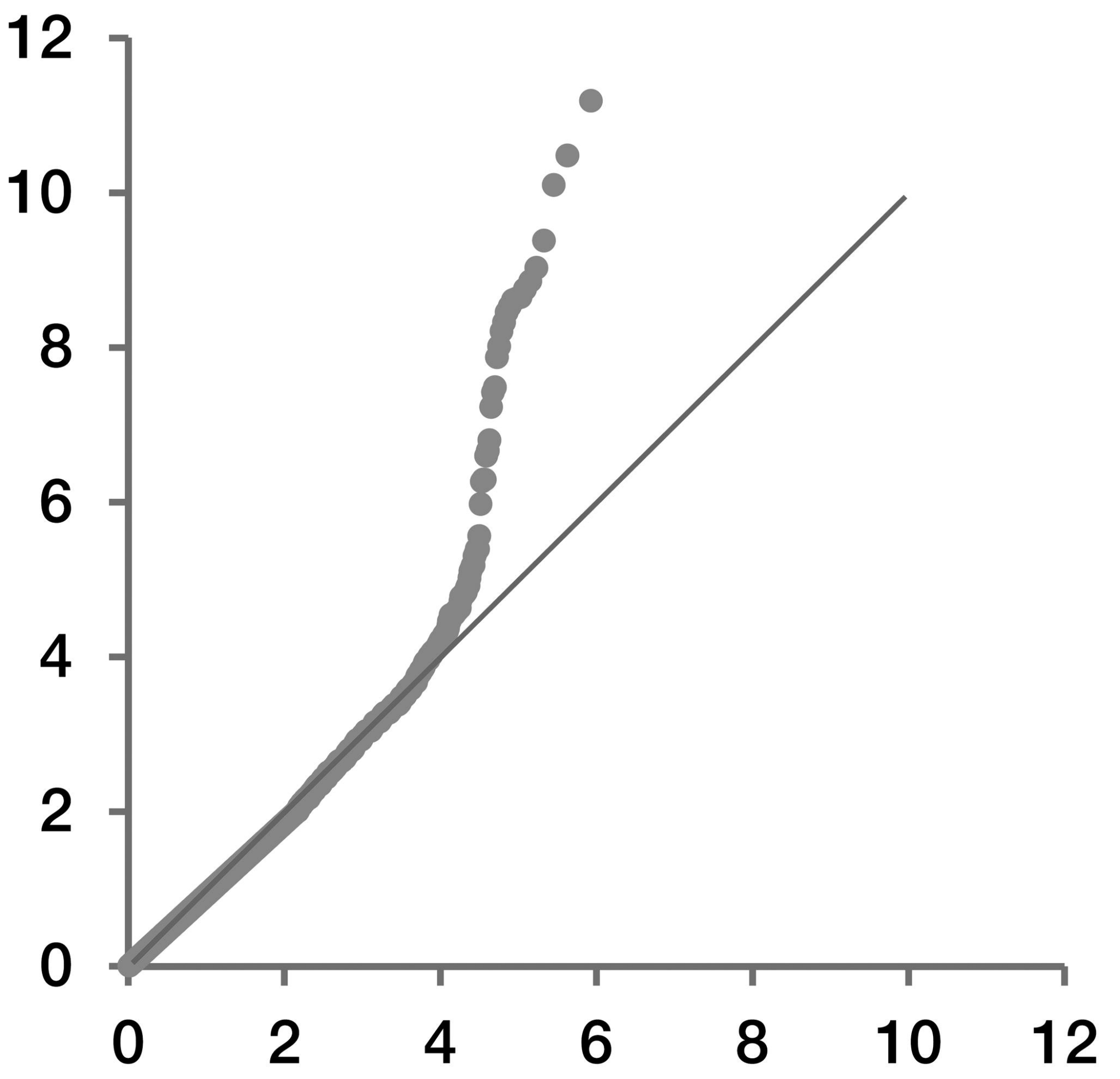


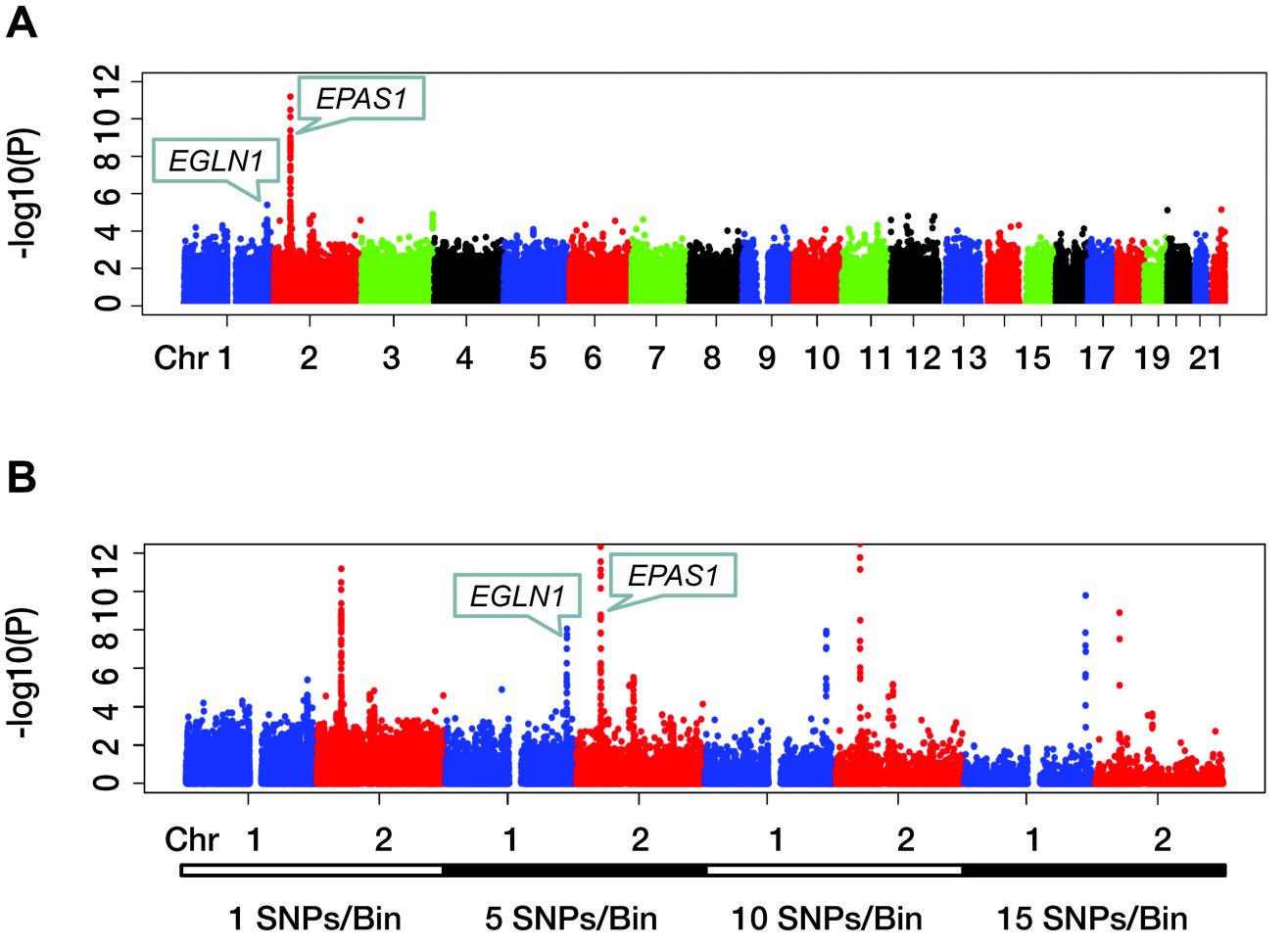
H1

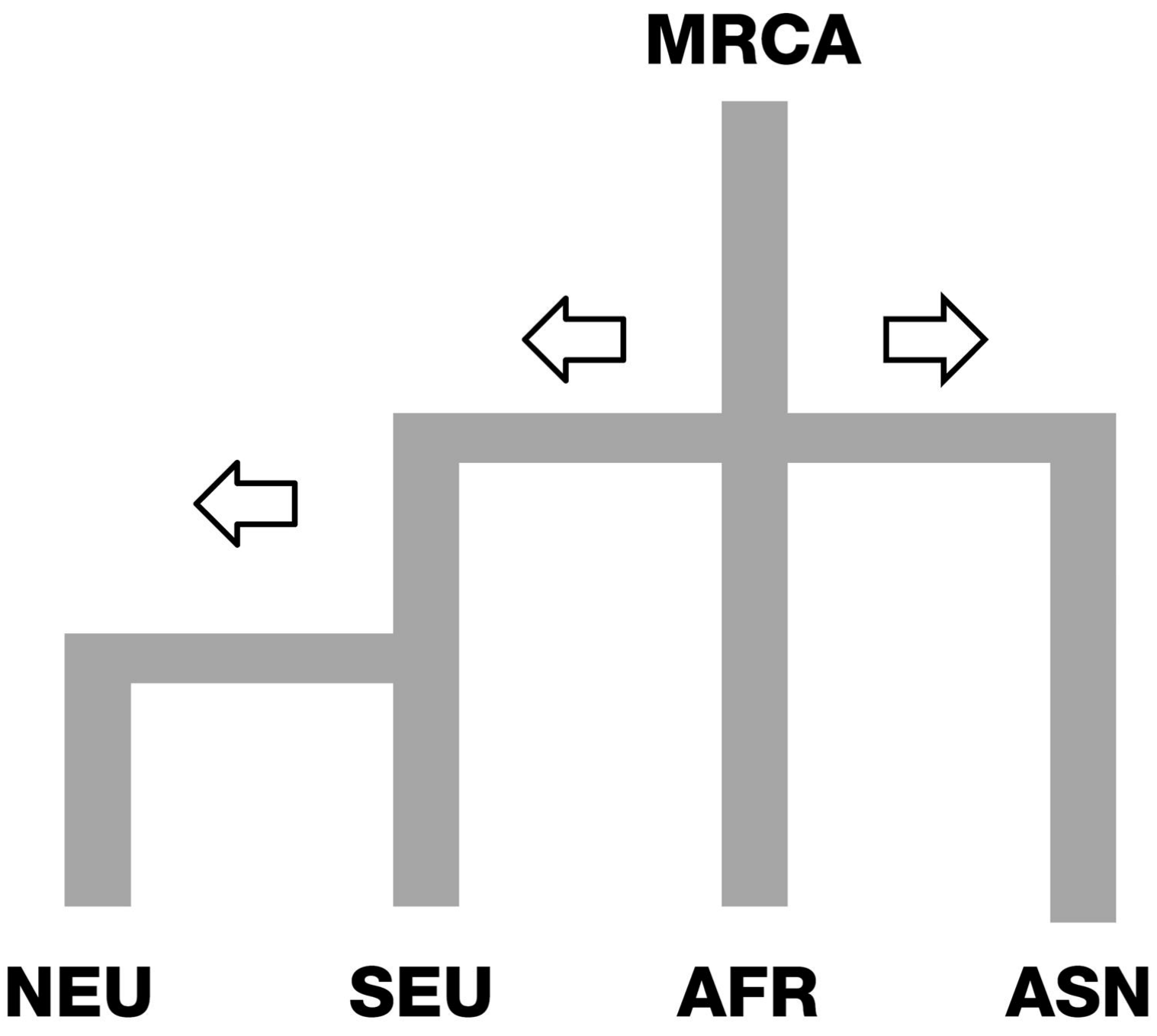


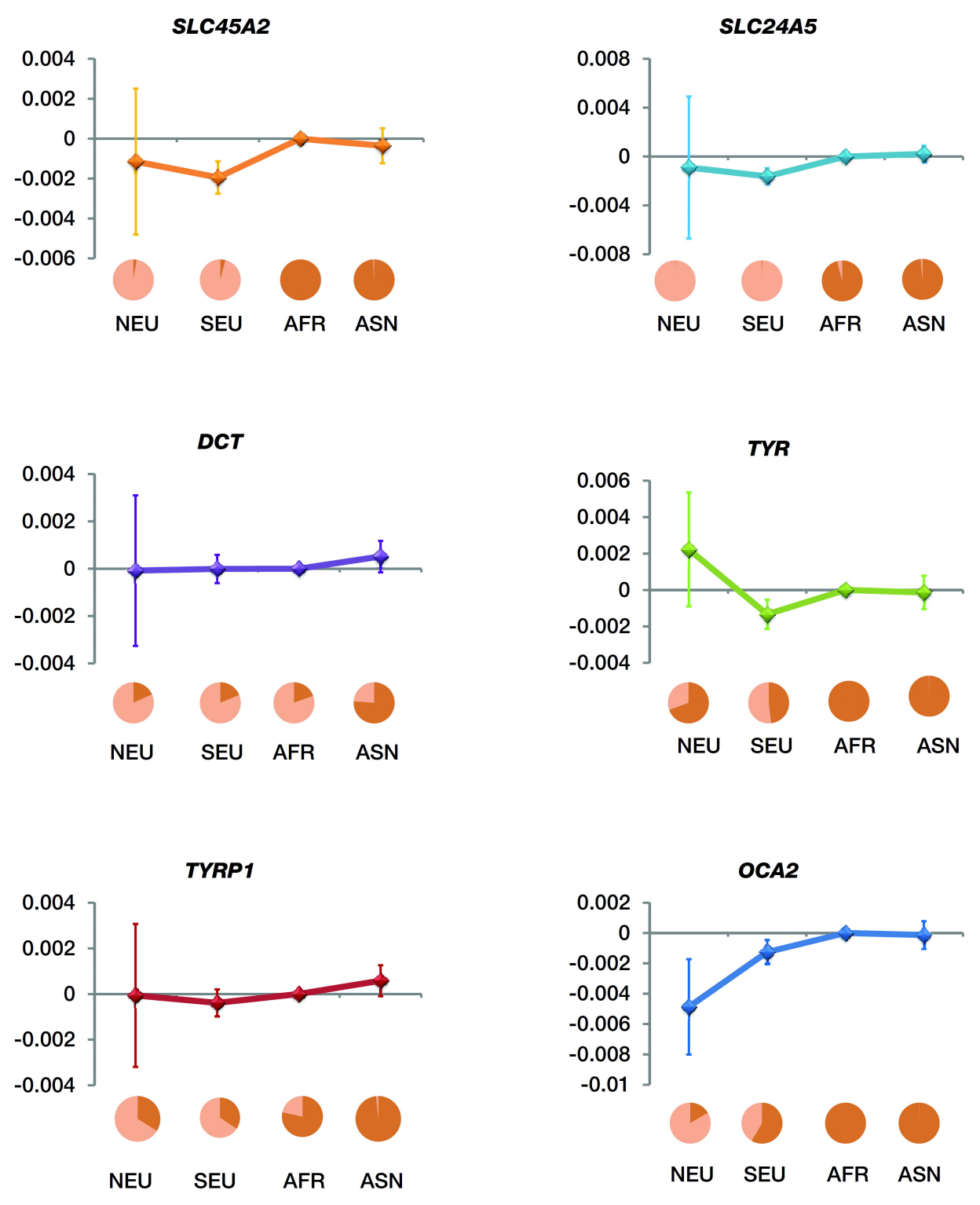



A
1.0
0.8
0.6
0.4
0.2
0
0
100
200
300
400
500
B
0.6
0.4
0.2
0
0
100
200
300
400
500


12
10
8
6
4
2
0
0
2
4
6
8
10
12


A
-log10(P)
0 2 4 6 8 10 12
EPAS1
EGLN1
Chr 1 2 3 4 5 6 7 8 9 10 11 12 13 15 17 19 21
B
-log10(P)
0 2 4 6 8 10 12
EGLN1
EPAS1
Chr 1 2 1 2 1 2 1 2
1 SNPs/Bin
5 SNPs/Bin
10 SNPs/Bin
15 SNPs/Bin


MRCA
NEU
SEU
AFR
ASN


SLC45A2
0.004
0.002
0
-0.002
-0.004
-0.006
NEU
SEU
AFR
ASN
SLC24A5
0.008
0.004
0
-0.004
-0.008
NEU
SEU
AFR
ASN
DCT
0.004
0.002
0
-0.002
-0.004
NEU
SEU
AFR
ASN
TYR
0.006
0.004
0.002
0
-0.002
-0.004
NEU
SEU
AFR
ASN
TYRP1
0.004
0.002
0
-0.002
-0.004
NEU
SEU
AFR
ASN
OCA2
0.002
0
-0.002
-0.004
-0.006
-0.008
-0.01
NEU
SEU
AFR
ASN