Supporting Online Information

Table of Contents

1. Exome resequencing

1.1 Study sample

The NHLBI Exome Sequencing Project (ESP) is a multi-center study to deeply sequence the exomes of individuals segregating a variety of heart, lung, and blood disorders[1]. The 6,515 individuals used in the analysis were generated from samples ascertained from 20 different cohorts (Supplementary Table 1; detailed information of cohorts can be found in Tennessen et al.[1]). Although these individuals are not a random sample, they were ascertained on a variety of distinct phenotypes such that cohort specific effects are not expected to bias patterns of SNVs. Indeed, detailed analyses of a large subset (n=2,440) of these 6,515 individuals found no systematic biases in patterns and characteristics of SNVs attributable to cohort or technical sources of variation[1]. All study participants in each of the component studies provided written informed consent for the use of their DNA in studies aimed at identifying genetic risk variants for disease and for broad data sharing. Institutional certification was obtained for each sample to allow deposition of phenotype and genotype data in dbGaP and BAM files in the short-read archive.

1.2 Exome resequencing, variant calling, and filtering

The processes of library construction, exome capture, sequencing, and mapping were performed as previously described[1]. SNVs were called using the UMAKE pipeline at University of Michigan, which allowed all samples to be analyzed simultaneously, both for variant calling and filtering. Briefly, we used BAM files summarizing BWA alignments generated at the University of Washington and the Broad Institute as input. These BAM files summarized alignments generated by BWA[2], refined by duplicate removal, recalibration, and indel re-alignment. We excluded all reads that were not confidently mapped (Phred-scaled mapping quality < 20) from further analysis. To avoid PCR artifacts, we clipped overlapping ends in paired reads. We then computed genotype likelihoods for exome targeted regions and 50 flanking bases, accounting for per base alignment quality (BAQ) using samtools[3]. Variable sites and their allele frequencies were identified using a maximum-likelihood model, implemented in glfMultiples[4]. These analyses assumed a uniform prior probability of polymorphism at each site. We used a support vector machine (SVM) classifier, which is a machine-learning algorithm, to separate likely true positive and false-positive variant sites. SVM filtering started by collecting a series of features related to quality of each SNV, including overall depth, fraction of samples with coverage, fraction of reference bases in heterozygous individuals (allele balance), correlation of alternative alleles with strand and read position (strand and cycle bias), and inbreeding coefficient for each variant. SNVs that deviated significantly from expected values in three or more categories were flagged as likely false positives when training the SVM filter. SNVs at HapMap polymorphic sites and Omni 2.5 array polymorphic sites in the 1000 Genomes project data were flagged as likely true

positives. After examining this training set, the SVM classifier was used to identify all likely false positive sites, which were excluded from downstream analyses. A total of 1,908,614 SNVs passed the SVM filter, with an overall transversion to transition ratio (Ts/Tv) of 2.84.

After the initial SNV calls were generated, we re-examined the VCF files and applied filters considering total read depth, the number of individuals with coverage at the site, the fraction of variant reads in each heterozygote, the ratio of forward and reverse strand reads for reads carrying reference and variant alleles, and the average position of variant alleles along a read. Next, the SNV call set included variants that were called with posterior probability >99% (glfMultiples SNP quality >20), were at least 5bp away from an indel detected in the 1000 Genomes Pilot Project, were targeted in at least 99% individuals, and had a total depth across samples between 6823 to 6823000 (~1-1000 reads per sample at average). Sites where the read depth of the variant allele was >65% in heterozygotes or where the absolute squared correlation between allele (variant or reference) and strand (forward or reverse) was >0.15 were excluded. In order to obtain genotypes with high accuracy suitable for population genetics analyses, we further set individual genotype to missing data if it had quality (GQ) less than 30 and/or filtered depth (DP) less than 10. After such filtering, variants with more than 10% of missing genotypes across individuals were excluded from further analysis.

1.3 Identification of related individuals and assignment of ancestry

In total, 6,823 exomes were obtained from individuals who self-identified as European American (EA, n=4,419), African American (AA, n=2,343) and others (including Asian, Hispanic and Native American). To remove related individuals, we performed a KING analysis[5] on the filtered data. Specifically, we performed LD pruning using PLINK (–indep-pairwise 50 5 0.5) to the variants with minor allele frequency (MAF) >5%. This resulted in 34,945 SNVs for the analysis. KING identifies kinship by pairwise comparisons across all individuals, and is robust to population structure. Using the authors' guidelines for a 3rd degree relationship (i.e., first cousins), we used a kinship coefficient threshold of 0.04419 (Supplementary Fig. 1). From this, we were able to form clusters of related individuals, with the majority of clusters consisting of two individuals. When all individuals were related to all other individuals in a cluster, we preferentially removed those with the greatest overall missingness. When these clusters had partial relationships (i.e., A is related to B and C but B and C are not related) then we preferentially removed those who would leave the largest number of samples. This resulted in the removal of 242 individuals. After removing these individuals, we repeated the KING analysis and found no kinships in the remaining data set.

Using the same filtered data set from the KING analysis, we performed a principal component analysis (PCA) to infer genetic ancestry (Supplementary Fig. 2). Asian, Hispanic, and Native American samples were removed from the analysis and

we used the dotted lines in Supplementary Fig. 2 as cut offs to assign individuals to EA and AA populations. This likely removed some self-identifying EAs and AAs, but resulted in two genetically identifiable groups for further analysis.

1.4 Data summary and annotation

We further removed 13 individuals because of mismatches between self-declared and genetically determined sex, and one individual with a call rate less than 80%. Thus, the final data set consisted of 4,298 EAs (1,879 males and 2,419 females) and 2,217 AAs (582 males and 1,635 females). Variants that deviated significantly from Hardy-Weinberg Equilibrium ($p<10^{-6}$) within populations were then filtered out.

A subset of the individuals (n = 2,440) contained in this data set were previously analyzed in Tennessen et al[1]. Extensive validation of SNVs was performed on this subset of 2,440 individuals[1]. Briefly, based on the analysis of 22 individuals sequenced in duplicate, a per-base error rate of $5.5x10^{-7}$ and a false discovery rate <0.2% were estimated. In addition, over 1,200 novel SNVs functionally annotated as a nonsynonymous or nonsense variant were randomly selected for validation by orthogonal experiment approaches, including 400 singletons, 768 variants with a MAF <10%, and all 52 variants with a MAF ≥10%. Approximately 98% of all variant sites that were tested were confirmed, including 98% of singletons, 98% of non-singleton SNV sites with a MAF < 10%, and 97% of SNV sites with a MAF ≥ 10%.

The ancestral allele for each variant was inferred from the six primate EPO alignments, which can be downloaded from the 1000 Genomes Project (GRCh37)[6]. Other information, such as site type and gene information, was annotated by SeattleSeq Annotation 134. Only SNVs that could be confidently polarized into derived and ancestral alleles were included in our analysis, resulting in a final data set of 1,146,401 (97.3%) autosomal SNVs with known ancestral states (Supplementary Fig. 3).

2. Estimates of allele age

2.1 Theory for estimates of allele age

Let $t_{n,b}$ denote the age of a mutant having b copies in a sample of n genes ($0<b<n$). Griffiths and Tavaré (1998)[7] showed that the mean and variance of $t_{n,b}$ can be obtained as,

$$E(t_{n,b}) = \frac{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1} E(S_k^2 - S_{k+1}^2)}{2\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1} E(T_k)} \quad (1)$$

$$Var(t_{n,b}) = E(t_{n,b}^2) - E(t_{n,b})^2 = \frac{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1} E(S_k^3 - S_{k+1}^3)}{3\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1} E(T_k)} - \left\{ \frac{\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1} E(S_k^2 - S_{k+1}^2)}{2\sum_{k=2}^{n} k(k-1)\binom{n-k}{b-1} E(T_k)} \right\}^2 \quad (2)$$

where $T_k (k = n,\ldots,2)$ is coalescent time which measures the time from k lineages to (k-1) lineages, and $S_k = T_n + T_{n-1} + \ldots + T_k$ (Supplementary Fig. 4).

In a constant population, they showed that

$$E(t_{n,b}) = 2\binom{n-1}{b}^{-1} \sum_{k=2}^{n} \binom{n-k}{b-1} \frac{n-k+1}{n(k-1)} \quad (3)$$

When $n \to \infty$, $b/n \to x$, we can obtain $E(t_x) = -\frac{2x}{1-x}\log x$, which is the well known formula calculating the expected age of a neutral mutation with frequency x in a stationary population derived by Kimura and Ohta (1973)[8].

2.2 A simulation approach to estimate allele age

A simulation approach was developed to estimate allele age. Specifically, a series of coalescent trees were simulated under a given demographic model (e.g. Supplementary Fig. 5) by the programs ms[9] (for neutral simulations) or msms[10] (for models with selection). According to these trees, $E(T_k)$ and $E(S_k^2 - S_{k+1}^2)$ can be calculated and plugged into formula (1) and (2) to calculate the mean and the variance of allele age for a given frequency. For all results reported in the manuscript, we assumed a generation time of 25 years.

2.3 Average allele age

We define the average allele age across SNVs as:

$$Exp(T) = \sum_{b=1}^{n-1} f_{n,b} t_{n,b} \quad (4)$$

where the weights are calculated according to the site-frequency-spectrum (SFS), $f_{n,b} = b/n$. In this case, the variance for the average allele age is:

$$Var(T) = \sum_{b=1}^{n-1} f_{n,b} Var(t_{n,b}) \quad (5)$$

A test statistic, similar to a t-test, was used to compare the average allele age from two different groups,

$$X_t = \frac{Exp(T)_1 - Exp(T)_2}{\sqrt{Var(T)_1 + Var(T)_2}} \quad (6)$$

The significance of the test was determined through permutations.

3. Coalescent simulations

3.1 Constant population

We first validated our simulation approach by comparing estimates from simulations with theoretical expectations in a constant population with size of 7,310. In brief, 4,000 diploid sequences with length 10kb were simulated, assuming the mutation rate was 1.5×10^{-8} per generation and with no recombination. Several factors, such as sample size ($2N$=100, 1000, 2000, 4000, and 8000), sequence length (5kb, 10kb and 50kb) and recombination rate (0, 1cM/Mb, and 10cM/Mb), were introduced to the basic model to explore the robustness of the estimates. In each scenario, 1,000 replicates (100 replicates when considering recombination) were simulated by ms. The mean and variance of allele age in neutral simulations were calculated according to formula (1) and (2) based on a series of coalescent trees, and compared with the theoretical expectation according to formula (3).

Compared with the theoretical estimates, allele age at a given derived allele frequency (DAF) can be accurately estimated by our simulation approach (Supplementary Fig. 6a), and the variance for the estimates becomes stable when the sample size is large enough (Supplementary Fig. 6b). Simulations also illustrated that the estimation of allele age is robust to different lengths of simulated sequence and recombination rates (Supplementary Fig. 6c -6f). Thus, we fixed the sequence length of 10kb and simulated without recombination in the following analyses to reduce the computational burden.

3.2 Population growth, migration, and selection

Effects of population growth, migration, and selection on the SFS and estimates of allele age were investigated in a more realistic demographic model for EAs and AAs (Supplementary Fig 5)[1]. In this model, explosive population growth beginning 5,115 years ago was introduced into the commonly used Out-of-African Model[11] to account for the abundance of rare variants observed by resequencing in large sample sizes. To match the empirical data, we simulated sequences for 4,298 EAs and 2,217 AAs, assuming the mutation rate was 1.5×10^{-8} per generation and no recombination. Simulations were conducted by the programs ms and msms (when considering selection). For each parameter value described below, 1,000 replicates were conducted.

When considering the influence of recent population growth, migration between populations and admixture for AAs were excluded. The change of population size before 5,115 years ago was fixed as shown in Supplementary Fig. 5, but different recent population growth rates (0, 0.5%, 1.0%, 1.5%, 2.0% and 2.5% per generation) were considered for EAs and AAs in the past 5,115 years. As expected, these simulations revealed that more variants accumulated for larger growth rates, especially for rare variants (Supplementary Fig. 7a and 7b). However, population growth has only modest effects on estimates of allele age for a given DAF. This observation was expected because the mutation rate remains constant, while the increasing population size generates more mutations per generation. In addition, we also estimated the expected SNV density per bp in a population without recent explosive growth based on this model.

When considering the influence of migration and admixture (20% contribution from European ancestry is assumed for AAs), the demographic model was fixed as shown in Supplementary Fig. 5, except that we varied the migration rate (0, 0.5×10^{-5}, 2.5×10^{-5}, 5×10^{-5}, 10^{-4} and 15×10^{-5} per chromosome per generation) during the past 5,115 years. These simulations revealed that migration has very modest effects on the SFS, although it results in EAs obtaining more ancient variants from the African population (Supplementary Fig. 7c and 7d).

Next, we considered the effects of purifying selection (s=0, 0.01%, 0.1%, 1% and 5%) acting on a single locus in the demographic model shown in Supplementary Fig.5. Both the SFS and estimates of allele age change imperceptibly (Supplementary Fig. 7e and 7f). A potential explanation for this observation is that a site under strong purifying selection will be removed quickly from the population, and therefore has limited effects on the structure of the coalescent tree; while for weak purifying selection, the stochastic effects inherent to the coalescent process will have a greater effect on gene genealogies relative to weak purifying selection, thus resulting in stable estimates for the expected allele age at given DAF. Therefore, estimation of allele age from the neutral demographic model with recent accelerated population growth is a reasonable proxy for more complicated and realistic evolutionary scenarios, even though the exome is expected to be subject to purifying selection[12,13].

3.3 Different reported demographic models

We investigated the robustness of allele age estimated across six reported demographic models for European populations[1,11,14-17]. The main parameters for these models are listed in Supplementary Table 2. Sequences for 4,298 European individuals were simulated by ms. For each model, 1,000 replicates were performed.

Several parameter values differ among these models, including growth rates, time of the most recent common ancestor of human population (T_{MRCA}), and time of the Out-of-Africa (T_{OOA}) dispersal. As a result of recent explosive population growth revealed by three studies that analyzed large sample sizes, an excess of rare variants was consistently observed (Supplementary Fig. 8a). The estimation of allele age is affected by different times of events, such as T_{MRCA} (425 thousand years ago (kya) v.s. 148 kya) and T_{OOA} (87.5 kya v.s. 51 kya), but such effects were generally modest and predominately confined to SNVs with a higher DAF. Thus, as the vast majority of variants in the real data are rare and low frequency SNVs, our results are robust to the demographic models considered here (Supplementary Fig. 8b and 8c).

3.4 Effects of errors on the estimation of allele age

To determine how robust our estimates of allele age are to various types of sequencing errors, we pursued three complimentary analysis strategies. All analyses were performed on simulated sequences from a demographic model with recent explosive population growth as shown in Supplementary Fig. 5[1]. First, errors were randomly introduced into the simulated sequences with a per-base error rate ranging from 5.5×10^{-8} - 5.5×10^{-6} (note, the estimated per base error rate in this data set is $5.5x10^{-7}$). In this model, errors occurring at invariant sites lead to an excess of rare variants, resulting in a slight downward bias of average allele age (Supplementary Fig. 9a). Second, we assessed how genotyping errors that modulate the allele frequency spectrum influences average allele age by randomly introducing errors at variant sites. In particular, we randomly changed 1%, 5%, or 10% of allele counts at variant sites by switching ancestral and derived status. This resulted in a slight upward bias of average allele age (Supplementary Fig. 9b). Third, to assess the influence of false negatives, we randomly removed 1%, 5% or 10% of variant sites. This resulted in very negligible changes to average allele age (Supplementary Fig. 9c). Thus, these results demonstrate that our estimates of allele age, and conclusions therein, are very robust to a variety of errors that could be present in the data.

4. Estimates of allele age in 6,515 exomes

4.1 Allele age for exomes

We quantitatively estimated the age of 1,146,401 autosomal SNVs identified from 6,515 exomes, based upon the demographic models with recent accelerated

population growth[1,16,17]. Average allele age was calculated according to equation (4), and allele age between different groups was compared using equation (6) with 100,000 permutations. Estimates of allele age were remarkably robust across different models (Supplementary Table 3); therefore, we presented the following results based on the models proposed by Tennessen et al. (2012)[1]. Moreover, this model was the only one to estimate parameter values for both the EA and AA populations. Finally, in contrast to the other two models[16,17], the time of the out of Africa dispersal was estimated from the data and not arbitrarily fixed[11,15].

In order to investigate the relationship between average allele age and selective constraint, a 1Mb sliding window analysis was performed. Average allele age was calculated for each window across exomes. These windows were merged into 10 bins according to the percentile of the corresponding average age. In each bin, the neutrality index[18] was calculated based on the number of polymorphic/divergent nonsynonymous and synonymous sites. Spearman's correlation test was used to show the correlation between the median of average allele age in each bin and its corresponding neutrality index.

4.2 Allele age for deleterious variants

To identify putatively deleterious variation, we used SIFT[19], PolyPhen2[20], a likelihood ratio test (LRT)[21], MutationTaster[22], GERP++[23] and PhyloP[24]. Thresholds in determining whether a given metric predicted a SNV to be deleterious were as follows: SIFT "Damage", PolyPhen2 "Probably Damaging", LRT "p<0.001", MutationTaster "disease causing automatic" or "disease causing", GERP++≥5, and PhyloP≥3.Characteristics of the average age of SNVs predicted to be deleterious by each method are shown in Supplementary Fig. 12. Note, the average age of nonsynonymous variants that were not predicted to be deleterious by any prediction method (42.1 kyr and 55.1 kyr in EAs and AAs, respectively) was nearly identical to unconstrained synonymous and noncoding SNVs (defined as GERP++≤5 and PhyloP≤3; 45.5 kyr in EAs and 58.9 kyr in AAs; t-test, $p>0.1$ by permutation for both EAs and AAs).

To investigate the features associated with nonsynonymous SNVs predicted to be deleterious, we classified amino acids into four categories (non-polar and neutral, polar and neutral, acidic and polar, and basic and polar), and compared the proportion of SNVs that do or do not result in amino acid changes between groups predicted to be deleterious. As expected, the proportion of deleterious variants is higher for SNVs resulting in changes between groups (Supplementary Fig. 10b and 10c). Moreover, for nonsense variants, changes to stop codons are significantly younger than those from stop codons (t-test; $p<10^{-5}$ by permutation) (Supplementary Fig. 10a). A significant negative correlation was observed between average allele age and the proportion of deleterious variants in different classes of amino acid changes by all methods (Spearman's correlation, $p<10^{-6}$), except for the LRT.

A majority rule approach was used to identify SNVs most likely to have functional consequences[1]. Specifically, nonsynonymous variants predicted to be functionally important by at least four methods and synonymous, splice, and noncoding variants predicted by two conservation-based methods (i.e., GERP++ and PhyloP) were designated as deleterious. We examined the relationship between the proportion of deleterious variants and corresponding allele age. The trend of the relationship was smoothed by the R function 'loess' with a smoothing parameter of 0.75 and degree of 2.

Note, although refer to these SNVs as "deleterious" variants to be consistent with the terminology in the literature, it is plausible that a subset are advantageous. However, among the set of SNVs predicted to be deleterious only < 0.02% (32/164688) have a derived frequency >90%, consistent with an advantageous allele sweeping to fixation. More generally, the fitness effects of this set of variants do not affect our analyses or conclusions.

4.3 Allele age of deleterious variants in disease genes

We classified genes into Mendelian disease genes, complex disease genes, essential genes and others. The list of Mendelian disease genes was obtained from a hand-curated OMIM database (referred as hOMIM)[25]. The list of complex disease genes was obtained from the Genetic Association Database (GAD, August 1, 2011 update)[26]. Essential genes are human orthologs where knockouts in mice result in lethality or sterility[27], and was obtained from Mouse Genome Informatics database (MGI). Note, we do not expect all essential genes in mice to be essential in humans, but this is a useful proxy and is likely enriched for genes that when disrupted have severe phenotypes.

We investigated the relationship between the proportion of deleterious variants and corresponding allele age. The trend of the relationship was smoothed by 'loess' with a smoothing parameter of 0.75 and degree of 2. Bootstrapping was used to calculate 95% confidence intervals, where 1,000 replicates were conducted by sampling sites with replacement from the group of interest.

In the EA samples, a peak in the proportion of deleterious variants was observed in Mendelian disease genes and essential genes for variants that arose before the out-of-Africa dispersal (Fig. 4a). We performed a forward simulation to examine the probability that a deleterious SNV survives to the present in a demographic model as illustrated in Supplementary Fig. 5 (for the EA population). Factors such as when the variant arose (i.e., before out-of-Africa, such as 150, 120, 100, 55 kya; in the process of Out-of-African (50 kya), and after the Out-of-Africa, such as 10 and 5 kya) and magnitude of purifying selection (i.e., s=0, 0.0001%, 0.001%, 0.01%, 0.1% and 1% for the derived allele) were considered. For each parameter value, 1,000,000 replicates were performed.

4.4 Allele age in KEGG pathways

Information of KEGG pathways were collected from the KEGG database (June, 2011)[28], and were annotated to the identified variants in this study. Average allele age for (deleterious) SNVs was calculated according to formula (4) for each pathway. The relationship between the proportion of deleterious variants and corresponding allele age was investigated. The trend of the relationship was smoothed by 'loess' with a smoothing parameter of 0.75 and degree of 2. Bootstrapping was used to calculate 95% confidence intervals, where 1,000 replicates were conducted by sampling sites with replacement from the group of interest.

Both the average allele age for deleterious variants (Kruskal-Wallis Rank Sum Test; $p=2.5\times10^{-3}$ and 1.08×10^{-6} for EAs and AAs, respectively; Fig. 4b) and the proportion of deleterious variants (Kruskal-Wallis Rank Sum Test; $p=2.3\times10^{-12}$ and 1.97×10^{-15} for EAs and AAs, respectively; Supplementary Fig. 15c) were significantly different across 235 KEGG pathways. Interestingly, metabolic pathways possess the highest proportion of deleterious variants (Fisher's exact test, $p= 1.55\times10^{-166}$ and 3.06×10^{-169} for EAs and AAs, respectively) and have an older average age (Mann-Whitney test, $p=1.11\times10^{-4}$ and 6.27×10^{-9} for EAs and AAs, respectively) compared with other pathways. The relationship between the proportion of deleterious variants and allele age suggests that metabolic pathways are under less functional constraint relative to other pathways, and hence deleterious alleles are purged more slowly (Supplementary Fig. 15a and 15b).

URLs

UMAKE http://genome.sph.umich.edu/wiki/UMAKE
1000 Genomes Project http://www.1000genomes.org/
ms http://home.uchicago.edu/~rhudson1/source/mksamples.html
msms http://www.mabs.at/ewing/msms/index.shtml
SIFT http://sift.jcvi.org/
PolyPhen2 http://genetics.bwh.harvard.edu/pph2/
MutationTaster http://www.mutationtaster.org/
GERP++ http://mendel.stanford.edu/SidowLab/downloads/gerp/
PhyloP (from UCSC database) http://genome.ucsc.edu/index.html?org=Human
R package http://www.r-project.org/
Genetic Association Database (GAD) http://geneticassociationdb.nih.gov/
Mouse Genome Informatics (MGI) http://www.informatics.jax.org/
KEGG database http://www.genome.jp/kegg/

Further Acknowledgements

HeartGO:

Atherosclerosis Risk in Communities (ARIC): NHLBI (N01 HC-55015, N01 HC-55016, N01HC-55017, N01 HC-55018, N01 HC-55019, N01 HC-55020, N01 HC-55021); **Cardiovascular Health Study (CHS)**: NHLBI (N01HC-85239, N01HC-85079 through N01HC-85086, N01HC-35129, N01 HC-15103, N01 HC-55222, N01HC-75150, N01HC-45133, and grant HL080295), with additional support from NINDS and from NIA (AG-023629, AG-15928, AG-20098, and AG-027058); **Coronary Artery Risk Development in Young Adults (CARDIA)**: NHLBI (N01HC95095, N01HC48047, N01HC48048, N01HC48049 and N01HC48050); **Framingham Heart Study (FHS)**: NHLBI (N01HC-25195 and grant R01 NS17950) with additional support from NIA (AG08122 and AG033193); **Jackson Heart Study (JHS)**: NHLBI and the National Institute on Minority Health and Health Disparities (N01 HC-95170, N01 HC-95171 and N01 HC-95172); **Multi-Ethnic Study of Atherosclerosis (MESA):** NHLBI (N01HC-95159 through N01HC-95169 and RR-024156).

Lung GO:

Cystic Fibrosis (CF): Cystic Fibrosis Foundation (GIBSON07K0, KNOWLE00A0, OBSERV04K0, RDP R026), the NHLBI (R01 HL-068890, R02 HL-095396), NIH National Center for Research Resources (UL1 RR-025014), and the National Human Genome Research Institute (NHGRI) (5R00 HG-004316). **Chronic Obstructive Pulmonary Disease (COPDGene):** NHLBI (U01 HL-089897, U01 HL-089856), and the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, and Sunovian. The COPDGene clinical centers and investigators are available at www.copdgene.org. **Acute Lung Injury (ALI)**: NHLBI (RC2 HL-101779). **Lung Health Study (LHS)**: NHLBI (RC2 HL-066583), the NHGRI (HG-004738), and the NHLBI Division of Lung Diseases (HR-46002). **Pulmonary Arterial Hypertension (PAH):** NIH (P50 HL-084946, K23 AR-52742), and the NHLBI (F32 HL-083714). **Asthma:** NHLBI (RC2 HL-101651), and the NIH (HL-077916, HL-69197, HL-76285, M01 RR-07122).

SWISS and ISGS:

Siblings with Ischemic Stroke Study (SWISS): National Institute of Neurological Disorders and Stroke (NINDS) (R01 NS039987); **Ischemic Stroke Genetics Study (ISGS):** NINDS (R01 NS042733)

WHISP:

Women's Health Initiative (WHI): The WHI Sequencing Project is funded by the NHLBI (HL-102924) as well as the National Institutes of Health (NIH), U.S. Department of Health and Human Services through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118-32119, 32122, 42107-26, 42129-32, and 44221. A full listing of WHI investigators can be found at: http://www.whiscience.org/publications/WHI_investigators_shortlist.pdf

NHLBI GO Exome Sequencing Project

BroadGO

Stacey B. Gabriel (Broad Institute)[4,5, 11, 16, 17], David M. Altshuler (Broad Institute, Harvard Medical School, Massachusetts General Hospital)[1, 5, 7, 17], Gonçalo R. Abecasis (University of Michigan)[3, 5, 9, 13, 15, 17], Hooman Allayee (University of Southern California)[5], Sharon Cresci (Washington University School of Medicine)[5], Mark J. Daly (Broad Institute, Massachusetts General Hospital), Paul I. W. de Bakker (Broad Institute, Harvard Medical School, University Medical Center Utrecht)[3, 15], Mark A. DePristo (Broad Institute)[4, 13, 15, 16], Ron Do (Broad Institute)[5, 9, 13, 15], Peter Donnelly (University of Oxford)[5], Deborah N. Farlow (Broad Institute)[3, 4, 5, 14, 12, 16, 17], Tim Fennell (Broad Institute), Kiran Garimella (University of Oxford)[4, 16], Stanley L. Hazen (Cleveland Clinic)[5], Youna Hu (University of Michigan)[3, 9, 15], Daniel M. Jordan (Harvard Medical School, Harvard University)[13], Goo Jun (University of Michigan)[13], Sekar Kathiresan (Broad Institute, Harvard Medical School, Massachusetts General Hospital)[5, 8, 9, 14, 12, 15, 17, 20], Hyun Min Kang (University of Michigan)[9, 13, 16], Adam Kiezun (Broad Institute)[5, 13, 15], Guillaume Lettre (Broad Institute, Montreal Heart Institute, Université de Montréal)[1, 2, 13, 15], Bingshan Li (University of Michigan)[3], Mingyao Li (University of Pennsylvania)[5], Christopher H. Newton-Cheh (Broad Institute, Massachusetts General Hospital, Harvard Medical School)[3, 8, 15], Sandosh Padmanabhan (University of Glasgow School of Medicine)[3, 12, 15], Gina Peloso (Broad Institute, Harvard Medical School, Massachusetts General Hospital)[5], Sara Pulit (Broad Institute)[3, 15], Daniel J. Rader (University of Pennsylvania)[5], David Reich (Broad Institute, Harvard Medical School)[15], Muredach P. Reilly (University of Pennsylvania)[5], Manuel A. Rivas (Broad Institute, Massachusetts General Hospital)[5], Steve Schwartz (Fred Hutchinson Cancer Research Center)[5, 12], Laura Scott (University of Michigan)[1], David S. Siscovick (University of Washington)[5, 1, 25], John A. Spertus (University of Missouri Kansas City)[5], Nathaniel O. Stitziel (Brigham and Women's Hospital)[5, 15], Nina Stoletzki (Brigham and Women's Hospital, Broad Institute, Harvard Medical School)[13], Shamil R. Sunyaev (Brigham and Women's Hospital, Broad Institute, Harvard Medical School)[1, 3, 5, 13, 15], Benjamin F. Voight (Broad Institute, Massachusetts General Hospital), Cristen J. Willer (University of Michigan)[1, 9, 13, 15]

HeartGO

Stephen S. Rich (University of Virginia)[2, 4, 7, 8, 9, 14, 11, 15, 17, 18, 31], Ermeg Akylbekova (Jackson State University, University of Mississippi Medical Center)[29], Larry D. Atwood (Boston University)[1, 11, 28], Christie M. Ballantyne (Baylor College of Medicine, Methodist DeBakey Heart Center)[9, 22], Maja Barbalic (University of Texas Health Science Center Houston)[9, 14, 15, 17, 22], R. Graham Barr (Columbia University Medical Center)[10, 31], Emelia J. Benjamin (Boston University)[14, 20, 28], Joshua Bis (University of Washington)[15, 23], Eric Boerwinkle (University of Texas Health Science Center Houston)[3, 5, 9, 13, 15, 17, 22], Donald W. Bowden (Wake Forest University)[1, 31], Jennifer Brody (University of Washington)[3, 5, 15, 23], Matthew Budoff (Harbor-UCLA Medical Center)[31], Greg Burke (Wake Forest University)[5, 31], Sarah Buxbaum (Jackson State University)[3, 13, 15, 29], Jeff Carr (Wake Forest University)[25, 29, 31], Donna T. Chen (University of Virginia)[6, 11], Ida Y. Chen (Cedars-Sinai Medical Center)[1, 31], Wei-Min Chen (University of Virginia)[13, 15, 18], Pat Concannon (University of Virginia)[11], Jacy Crosby (University of Texas Health Science Center Houston)[22], L. Adrienne Cupples (Boston University)[1, 3, 5, 9, 13, 15, 18, 28], Ralph D'Agostino (Boston University)[28], Anita L. DeStefano (Boston University)[13, 18, 28], Albert Dreisbach (University of Mississippi Medical Center)[3, 29], Josée Dupuis (Boston University)[1, 28], J. Peter Durda (University of Vermont)[15, 23], Jaclyn Ellis (University of North Carolina Chapel Hill)[1], Aaron R. Folsom (University of Minnesota)[5, 22], Myriam Fornage (University of Texas Health Science Center Houston)[3, 18, 25], Caroline S. Fox (National Heart, Lung, and Blood Institute)[1, 28], Ervin Fox (University of Mississippi Medical Center)[3, 9, 29], Vincent Funari (Cedars-Sinai Medical Center)[1, 11, 31], Santhi K. Ganesh (University of Michigan)[2, 22], Julius Gardin (Hackensack University Medical Center)[25], David Goff (Wake Forest University)[25], Ora Gordon (Cedars-Sinai Medical Center)[11, 31], Wayne Grody (University of California Los Angeles)[11, 31], Myron Gross (University of Minnesota)[1, 5, 14, 25], Xiuqing Guo (Cedars-Sinai Medical Center)[3, 15, 31], Ira M. Hall (University of Virginia), Nancy L. Heard-Costa (Boston University)[1, 11, 28], Susan R. Heckbert (University of Washington)[10, 14, 20, 23], Nicholas Heintz (University of Vermont), David M. Herrington (Wake Forest University)[5, 31], DeMarc Hickson (Jackson State University, University of Mississippi Medical Center)[29], Jie Huang (National Heart, Lung, and Blood Institute)[5, 28], Shih-Jen Hwang (Boston University, National Heart, Lung, and Blood Institute)[3, 28], David R. Jacobs (University of Minnesota)[25], Nancy S. Jenny (University of Vermont)[1, 2, 23], Andrew D. Johnson (National Heart, Lung, and Blood Institute)[2, 5, 11, 28], Craig W. Johnson (University of Washington)[15, 31], Steven Kawut (University of Pennsylvania)[10,31], Richard Kronmal (University of Washington)[31], Raluca Kurz (Cedars-Sinai Medical Center)[11, 31], Ethan M. Lange (University of North Carolina Chapel Hill)[3, 5, 9, 13, 34], Leslie A. Lange (University of North Carolina Chapel Hill)[1, 2, 3, 5, 9, 12, 13, 15, 17, 18, 20, 25, 34], Martin G. Larson (Boston University)[3, 15, 28], Mark Lawson (University of Virginia), Cora E. Lewis (University of Alabama at Birmingham)[25,34], Daniel Levy (National Heart, Lung, and Blood Institute)[3, 15, 17, 28], Dalin Li (Cedars-Sinai Medical Center)[11, 15, 31], Honghuang Lin (Boston University)[20, 28], Chunyu Liu (National Heart, Lung, and

Blood Institute)[3, 28], Jiankang Liu (University of Mississippi Medical Center)[1, 29], Kiang Liu (Northwestern University)[25], Xiaoming Liu (University of Texas Health Science Center Houston)[15, 22], Yongmei Liu (Wake Forest University)[2, 5, 31], William T. Longstreth (University of Washington)[18, 23], Cay Loria (National Heart, Lung, and Blood Institute)[25], Thomas Lumley (University of Auckland)[9, 23], Kathryn Lunetta (Boston University)[28], Aaron J. Mackey (University of Virginia)[16, 18], Rachel Mackey (University of Pittsburgh)[1, 23, 31], Ani Manichaikul (University of Virginia)[8, 15, 18, 31], Taylor Maxwell (University of Texas Health Science Center Houston)[22], Barbara McKnight (University of Washington)[15, 23], James B. Meigs (Brigham and Women's Hospital, Harvard Medical School, Massachusetts General Hospital)[1, 28], Alanna C. Morrison (University of Texas Health Science Center Houston)[3, 15, 17], Solomon K. Musani (University of Mississippi Medical Center)[3, 29], Josyf C. Mychaleckyj (University of Virginia)[13, 15, 31], Jennifer A. Nettleton (University of Texas Health Science Center Houston)[9, 22], Kari North (University of North Carolina Chapel Hill)[1, 3, 9, 10, 13, 15, 17, 34], Christopher J. O'Donnell (Massachusetts General Hospital, National Heart, Lung, and Blood Institute)[2, 5, 9, 14, 11, 12, 15, 17, 20, 28], Daniel O'Leary (Tufts University School of Medicine)[25, 31], Frank Ong (Cedars-Sinai Medical Center)[3, 11, 31], Walter Palmas (Columbia University)[3, 15, 31], James S. Pankow (University of Minnesota)[1, 22], Nathan D. Pankratz (Indiana University School of Medicine)[15, 25], Shom Paul (University of Virginia), Marco Perez (Stanford University School of Medicine), Sharina D. Person (University of Alabama at Birmingham, University of Alabama at Tuscaloosa)[25], Joseph Polak (Tufts University School of Medicine)[31], Wendy S. Post (Johns Hopkins University)[3, 9, 14, 11, 20, 31], Bruce M. Psaty (Group Health Research Institute, University of Washington)[3, 5, 9, 14, 11, 15, 23], Aaron R. Quinlan (University of Virginia)[18, 19], Leslie J. Raffel (Cedars-Sinai Medical Center)[6, 11, 31], Vasan S. Ramachandran (Boston University)[3, 28], Alexander P. Reiner (Fred Hutchinson Cancer Research Center, University of Washington)[1, 2, 3, 5, 9, 11, 12, 13, 14, 15, 20, 25, 34], Kenneth Rice (University of Washington)[15, 23], Jerome I. Rotter (Cedars-Sinai Medical Center)[1, 3, 6, 8, 11, 15, 31], Jill P. Sanders (University of Vermont)[23], Pamela Schreiner (University of Minnesota)[25], Sudha Seshadri (Boston University)[18, 28], Steve Shea (Brigham and Women's Hospital, Harvard University)[28], Stephen Sidney (Kaiser Permanente Division of Research, Oakland, CA)[25], Kevin Silverstein (University of Minnesota)[25], David S. Siscovick (University of Washington)[5, 1, 25], Nicholas L. Smith (University of Washington)[2, 15, 20, 23], Nona Sotoodehnia (University of Washington)[3, 15, 23], Asoke Srinivasan (Tougaloo College)[29], Herman A. Taylor (Jackson State University, Tougaloo College, University of Mississippi Medical Center)[5,29], Kent Taylor (Cedars-Sinai Medical Center)[31], Fridtjof Thomas (University of Texas Health Science Center Houston)[3, 22], Russell P. Tracy (University of Vermont)[5, 9, 14, 11, 12, 15, 17, 20, 23], Michael Y. Tsai (University of Minnesota)[9, 31], Kelly A. Volcik (University of Texas Health Science Center Houston)[22], Chrstina L Wassel (University of California San Diego)[9, 15, 31], Karol Watson (University of California Los Angeles)[31], Gina Wei (National Heart, Lung, and Blood Institute)[25], Wendy White (Tougaloo College)[29], Kerri L. Wiggins (University of Vermont)[23], Jemma B. Wilk (Boston University)[10, 28], O. Dale Williams

(Florida International University)[25], Gregory Wilson (Jackson State University)[29], James G. Wilson (University of Mississippi Medical Center)[1, 2, 5, 8, 9, 14, 11, 12, 17, 20, 29], Phillip Wolf (Boston University)[28], Neil A. Zakai (University of Vermont)[2, 23]

ISGS and SWISS

John Hardy (Reta Lila Weston Research Laboratories, Institute of Neurology, University College London)[18], James F. Meschia (Mayo Clinic)[18], Michael Nalls (National Institute on Aging)[2, 18], Stephen S. Rich (University of Virginia)[2, 4, 7, 8, 9, 14, 11, 15, 17, 18, 31], Andrew Singleton (National Institute on Aging)[18], Brad Worrall (University of Virginia)[18]

LungGO

Michael J. Bamshad (Seattle Children's Hospital, University of Washington)[4, 6, 7, 8, 10, 11, 13, 15, 17, 27], Kathleen C. Barnes (Johns Hopkins University)[2, 10, 14, 12, 15, 17, 20, 24, 30, 32], Ibrahim Abdulhamid (Children's Hospital of Michigan)[27], Frank Accurso (University of Colorado)[27], Ran Anbar (Upstate Medical University)[27], Terri Beaty (Johns Hopkins University)[24, 30], Abigail Bigham (University of Washington)[13, 15, 27], Phillip Black (Children's Mercy Hospital)[27], Eugene Bleecker (Wake Forest University)[33], Kati Buckingham (University of Washington)[27], Anne Marie Cairns (Maine Medical Center)[27], Wei-Min Chen (University of Virginia)[13, 15, 18], Daniel Caplan (Emory University)[27], Barbara Chatfield (University of Utah)[27], Aaron Chidekel (A.I. Dupont Institute Medical Center)[27], Michael Cho (Brigham and Women's Hospital, Harvard Medical School)[13, 15, 24], David C. Christiani (Massachusetts General Hospital)[21], James D. Crapo (National Jewish Health)[24, 30], Julia Crouch (Seattle Children's Hospital)6, Denise Daley (University of British Columbia)[30], Anthony Dang (University of North Carolina Chapel Hill)[26], Hong Dang (University of North Carolina Chapel Hill)[26], Alicia De Paula (Ochsner Health System)[27], Joan DeCelie-Germana (Schneider Children's Hospital)[27], Allen Dozor (New York Medical College, Westchester Medical Center)[27,] Mitch Drumm (University of North Carolina Chapel Hill)[26], Maynard Dyson (Cook Children's Med. Center)[27], Julia Emerson (Seattle Children's Hospital, University of Washington)[27], Mary J. Emond (University of Washington)[10, 13, 15, 17, 27], Thomas Ferkol (St. Louis Children's Hospital, Washington University School of Medicine)[27], Robert Fink (Children's Medical Center of Dayton)[27], Cassandra Foster (Johns Hopkins University)[30], Deborah Froh (University of Virginia)[27], Li Gao (Johns Hopkins University)[24, 30, 32], William Gershan (Children's Hospital of Wisconsin)[27], Ronald L. Gibson (Seattle Children's Hospital, University of Washington)[10, 27], Elizabeth Godwin (University of North Carolina Chapel Hill)[26], Magdalen Gondor (All Children's Hospital Cystic Fibrosis Center)[27], Hector Gutierrez (University of Alabama at Birmingham)[27], Nadia N.

Hansel (Johns Hopkins University, Johns Hopkins University School of Public Health)[10, 15, 30], Paul M. Hassoun (Johns Hopkins University)[10, 14, 32], Peter Hiatt (Texas Children's Hospital)[27], John E. Hokanson (University of Colorado)[24], Michelle Howenstine (Indiana University, Riley Hospital for Children)[27], Laura K. Hummer (Johns Hopkins University)[32], Jamshed Kanga (University of Kentucky)[27], Yoonhee Kim (National Human Genome Research Institute)[24, 32], Michael R. Knowles (University of North Carolina Chapel Hill)[10, 26], Michael Konstan (Rainbow Babies & Children's Hospital)[27], Thomas Lahiri (Vermont Children's Hospital at Fletcher Allen Health Care)[27], Nan Laird (Harvard School of Public Health)[24], Christoph Lange (Harvard School of Public Health)[24], Lin Lin (Harvard Medical School)[21], Xihong Lin (Harvard School of Public Health)[21], Tin L. Louie (University of Washington)[13, 15, 27], David Lynch (National Jewish Health)[24], Barry Make (National Jewish Health)[24], Thomas R. Martin (University of Washington, VA Puget Sound Medical Center)[10, 21], Steve C. Mathai (Johns Hopkins University)[32], Rasika A. Mathias (Johns Hopkins University)[10, 13, 15, 30, 32], John McNamara (Children's Hospitals and Clinics of Minnesota)[27], Sharon McNamara (Seattle Children's Hospital)[27], Deborah Meyers (Wake Forest University)[33], Susan Millard (DeVos Children's Butterworth Hospital, Spectrum Health Systems)[27], Peter Mogayzel (Johns Hopkins University)[27], Richard Moss (Stanford University)[27], Tanda Murray (Johns Hopkins University)[30], Dennis Nielson (University of California at San Francisco)[27], Blakeslee Noyes (Cardinal Glennon Children's Hospital)[27], Wanda O'Neal (University of North Carolina Chapel Hill)[26], David Orenstein (Children's Hospital of Pittsburgh)[27], Brian O'Sullivan (University of Massachusetts Memorial Health Care)[27], Rhonda Pace (University of North Carolina Chapel Hill)[26], Peter Pare (St. Paul's Hospital)[30], H. Worth Parker (Dartmouth-Hitchcock Medical Center, New Hampshire Cystic Fibrosis Center)[27], Mary Ann Passero (Rhode Island Hospital)[27], Elizabeth Perkett (Vanderbilt University)[27,] Adrienne Prestridge (Children's Memorial Hospital)[27], Nicholas M. Rafaels (Johns Hopkins University)[30], Bonnie Ramsey (Seattle Children's Hospital, University of Washington)[27], Elizabeth Regan (National Jewish Health)[24], Clement Ren (University of Rochester)[27], George Retsch-Bogart (University of North Carolina Chapel Hill)[27], Michael Rock (University of Wisconsin Hospital and Clinics)[27], Antony Rosen (Johns Hopkins University)[32], Margaret Rosenfeld (Seattle Children's Hospital, University of Washington)[27], Ingo Ruczinski (Johns Hopkins University School of Public Health)[13, 15, 30], Andrew Sanford (University of British Columbia)[30], David Schaeffer (Nemours Children's Clinic)[27], Cindy Sell (University of North Carolina Chapel Hill)[26], Daniel Sheehan (Children's Hospital of Buffalo)[27], Edwin K. Silverman (Brigham and Women's Hospital, Harvard Medical School)[24, 30], Don Sin (Children's Medical Center of Dayton)[30], Terry Spencer (Elliot Health System)[27], Jackie Stonebraker (University of North Carolina Chapel Hill)[26], Holly K. Tabor (Seattle Children's Hospital, University of Washington)[6, 10, 11, 17, 27], Laurie Varlotta (St. Christopher's Hospital for Children)[27], Candelaria I. Vergara (Johns Hopkins University)[30], Robert Weiss [30], Fred Wigley (Johns Hopkins University)[32], Robert A. Wise (Johns Hopkins University)[30], Fred A. Wright (University of North Carolina Chapel Hill)[26], Mark M. Wurfel (University of Washington)[10, 14, 21], Robert Zanni

(Monmouth Medical Center)[27], Fei Zou (University of North Carolina Chapel Hill)[26]

SeattleGO

Deborah A. Nickerson (University of Washington)[3, 4, 5, 7, 8, 9, 11, 15, 17, 18, 19], Mark J. Rieder (Adaptive Biotechnologies Corporation)[4, 11, 13, 15, 16, 17, 19], Phil Green (University of Washington), Jay Shendure (University of Washington)[1, 8, 14, 16, 17], Joshua M. Akey (University of Washington)[14, 13, 15], Michael J. Bamshad (Seattle Children's Hospital, University of Washington)[4, 6, 7, 8, 10, 11, 13, 15, 17, 27], Carlos D. Bustamante (Stanford University School of Medicine)[3, 13, 15], David R. Crosslin (University of Washington)[2, 9], Evan E. Eichler (University of Washington)[19], P. Keolu Fox[2], Wenqing Fu (University of Washington)[13], Adam Gordon (University of Washington)[11], Simon Gravel (Stanford University School of Medicine)[13, 15], Gail P. Jarvik (University of Washington)[9, 15], Jill M. Johnsen (Puget Sound Blood Center, University of Washington)[2], Mengyuan Kan (Baylor College of Medicine)[13], Eimear E. Kenny (Stanford University School of Medicine)[3, 13, 15], Jeffrey M. Kidd (Stanford University School of Medicine)[13, 15], Fremiet Lara-Garduno (Baylor College of Medicine)[15], Suzanne M. Leal (Baylor College of Medicine)[1, 13, 15, 16, 17, 19, 20], Dajiang J. Liu (Baylor College of Medicine)[13, 15], Sean McGee (University of Washington)[13, 15, 19], Timothy D. O'Connor (University of Washington)[13], Bryan Paeper (University of Washington)[16], Peggy D. Robertson (University of Washington)[4], Joshua D. Smith (University of Washington)[4, 16, 19], Jacob A. Tennessen (University of Washington)[13], Emily H. Turner (University of Washington)[4, 16], Gao Wang (Baylor College of Medicine)[1,13,20]

WHISP

Rebecca Jackson (Ohio State University)[1, 2, 4, 5, 8, 12, 14, 15, 17, 18, 20, 34], Kari North (University of North Carolina Chapel Hill)[1, 3, 9, 10, 13, 15, 17, 34], Ulrike Peters (Fred Hutchinson Cancer Research Center)[1, 3, 11, 12, 13, 15, 17, 18, 34], Christopher S. Carlson (Fred Hutchinson Cancer Research Center, University of Washington)[1, 2, 3, 5, 14, 12, 13, 15, 16, 17, 18, 19, 34], Garnet Anderson (Fred Hutchinson Cancer Research Center)[34], Hoda Anton-Culver (University of California at Irvine)[34], Themistocles L. Assimes (Stanford University School of Medicine)[5, 9, 11, 34], Paul L. Auer (Fred Hutchinson Cancer Research Center)[1, 2, 3, 5, 11, 12, 13, 15, 16, 18, 34], Shirley Beresford (Fred Hutchinson Cancer Research Center)[34], Chris Bizon (University of North Carolina Chapel Hill)[3, 9, 13, 15, 34], Henry Black (Rush Medical Center)[34], Robert Brunner (University of Nevada)[34], Robert Brzyski (University of Texas Health Science Center San Antonio)[34], Dale Burwen (National Heart, Lung, and Blood Institute WHI Project Office)[34], Bette Caan (Kaiser Permanente Division of Research, Oakland, CA)[34], Cara L. Carty (Fred Hutchinson Cancer Research Center)[18, 34], Rowan Chlebowski (Los

Angeles Biomedical Research Institute)[34], Steven Cummings (University of California at San Francisco)[34], J. David Curb* (University of Hawaii)[9, 18, 34], Charles B. Eaton (Brown University, Memorial Hospital of Rhode Island)[12, 34], Leslie Ford (National Heart, Lung, and Blood Institute, National Heart, Lung, and Blood Institute WHI Project Office)[34], Nora Franceschini (University of North Carolina Chapel Hill)[2, 3, 9, 10, 15, 34], Stephanie M. Fullerton (University of Washington)[6, 11, 34], Margery Gass (University of Cincinnati)[34], Nancy Geller (National Heart, Lung, and Blood Institute WHI Project Office)[34], Gerardo Heiss (University of North Carolina Chapel Hill)[5, 34], Barbara V. Howard (Howard University, MedStar Research Institute)[34], Li Hsu (Fred Hutchinson Cancer Research Center)[1, 13, 15, 18, 34], Carolyn M. Hutter (Fred Hutchinson Cancer Research Center)[13, 15, 18, 34], John Ioannidis (Stanford University School of Medicine)[11, 34], Shuo Jiao (Fred Hutchinson Cancer Research Center)[34], Karen C. Johnson (University of Tennessee Health Science Center)[3, 34], Charles Kooperberg (Fred Hutchinson Cancer Research Center)[1, 5, 9, 14, 13, 15, 17, 18, 34], Lewis Kuller (University of Pittsburgh)[34], Andrea LaCroix (Fred Hutchinson Cancer Research Center)[34], Kamakshi Lakshminarayan (University of Minnesota)[18, 34], Dorothy Lane (State University of New York at Stony Brook)[34], Ethan M. Lange (University of North Carolina Chapel Hill)[3, 5, 9, 13, 34], Leslie A. Lange (University of North Carolina Chapel Hill)[1, 2, 3, 5, 9, 12, 13, 15, 17, 18, 20, 25, 34], Norman Lasser (University of Medicine and Dentistry of New Jersey)[34], Erin LeBlanc (Kaiser Permanente Center for Health Research, Portland, OR)[34], Cora E. Lewis (University of Alabama at Birmingham)[25,34], Kuo-Ping Li (University of North Carolina Chapel Hill)[9, 34], Marian Limacher (University of Florida)[34], Dan-Yu Lin (University of North Carolina Chapel Hill)[1, 3, 9, 13, 15, 34], Benjamin A. Logsdon (Fred Hutchinson Cancer Research Center)[2, 34], Shari Ludlam (National Heart, Lung, and Blood Institute WHI Project Office)[34], JoAnn E. Manson (Brigham and Women's Hospital, Harvard School of Public Health)[34], Karen Margolis (University of Minnesota)[34], Lisa Martin (George Washington University Medical Center)[9, 34], Joan McGowan (National Heart, Lung, and Blood Institute WHI Project Office)[34], Keri L. Monda (Amgen, Inc.)[1,15, 34], Jane Morley Kotchen (Medical College of Wisconsin)[34], Lauren Nathan (University of California Los Angeles)[34], Judith Ockene (Fallon Clinic, University of Massachusetts)[34], Mary Jo O'Sullivan (University of Miami)[34], Lawrence S. Phillips (Emory University)[34], Ross L. Prentice (Fred Hutchinson Cancer Research Center)[34], Alexander P. Reiner (Fred Hutchinson Cancer Research Center, University of Washington)[1, 2, 3, 5, 9, 11, 12, 13, 14, 15, 20, 25, 34], John Robbins (University of California at Davis)[34], Jennifer G. Robinson (University of Iowa)[9, 11, 18, 34], Jacques E. Rossouw (National Heart, Lung, and Blood Institute, National Heart, Lung, and Blood Institute WHI Project Office)[5, 14, 17, 20, 34], Haleh Sangi-Haghpeykar (Baylor College of Medicine)[34], Gloria E. Sarto (University of Wisconsin)[34], Sally Shumaker (Wake Forest University)[34], Michael S. Simon (Wayne State University)[34], Marcia L. Stefanick (Stanford University School of Medicine)[34], Evan Stein (Medical Research Labs)[34], Hua Tang (Stanford University)[2, 34], Kira C. Taylor (University of Louisville)[1, 3, 13, 15, 20, 34], Cynthia A. Thomson (University of Arizona)[34], Timothy A. Thornton (University of Washington)[13, 15, 18, 34], Linda Van Horn (Northwestern University)[34], Mara Vitolins (Wake Forest University)[34], Jean

Wactawski-Wende (University of Buffalo)[34], Robert Wallace (University of Iowa)[2, 34], Sylvia Wassertheil-Smoller (Boston University)[18, 34], Donglin Zeng (University of North Carolina Chapel Hill)[9, 34]

NHLBI GO ESP Project Team

Deborah Applebaum-Bowden (National Heart, Lung, and Blood Institute)[4, 7, 12, 17], Michael Feolo (National Center for Biotechnology Information)[12], Weiniu Gan (National Heart, Lung, and Blood Institute)[7, 8, 16, 17], Dina N. Paltoo (National Heart, Lung, and Blood Institute)[4, 6, 11, 17], Jacques E. Rossouw (National Heart, Lung, and Blood Institute, National Heart, Lung, and Blood Institute WHI Project Office)[5, 14, 17, 20, 34], Phyliss Sholinsky (National Heart, Lung, and Blood Institute)[4, 12, 17], Anne Sturcke (National Center for Biotechnology Information)[12]

*deceased

ESP Groups

[1]Anthropometry Project Team, [2]Blood Count/Hematology Project Team, [3]Blood Pressure Project Team, [4]Data Flow Working Group, [5]Early MI Project Team, [6]ELSI Working Group, [7]Executive Committee, [8]Family Study Project Team, [9]Lipids Project Team, [10]Lung Project Team, [11]Personal Genomics Project Team, [12]Phenotype and Harmonization Working Group, [13]Population Genetics and Statistical Analysis Working Group, [14]Publications and Presentations Working Group, [15]Quantitative Analysis Ad Hoc Task Group, [16]Sequencing and Genotyping Working Group, [17]Steering Committee, [18]Stroke Project Team, [19]Structural Variation Working Group, [20]Subclinical/Quantitative Project Team

ESP Cohorts

[21]Acute Lung Injury (ALI), [22]Atherosclerosis Risk in Communities (ARIC), [23]Cardiovascular Health Study (CHS), [24]Chronic Obstructive Pulmonary Disease (COPDGene), [25]Coronary Artery Risk Development in Young Adults (CARDIA), [26]Cystic Fibrosis (CF), [27]Early Pseudomonas Infection Control (EPIC), [28]Framingham Heart Study (FHS), [29]Jackson Heart Study (JHS), [30]Lung Health Study (LHS), [31]Multi-Ethnic Study of Atherosclerosis (MESA), [32]Pulmonary Arterial Hypertension (PAH), [33]Severe Asthma Research Program (SARP), [34]Women's Health Initiative (WHI)

Supplementary Reference

1. Tennessen, J.A. et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *Science* **337**, 64-69 (2012).
2. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).
3. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
4. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-8 (2008).
5. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-73 (2010).
6. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).
7. Griffiths, R.C. & Tavaré, S. The age of a mutation in a general coalescent tree. *COMMUM. STATIST.-STOCHASTIC MODELS* **14**, 273-295 (1998).
8. Kimura, M. & Ota, T. The age of a neutral mutant persisting in a finite population. *Genetics* **75**, 199-212 (1973).
9. Hudson, R.R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-8 (2002).
10. Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**, 2064-5 (2010).
11. Gravel, S. et al. Demographic history and rare allele sharing among human populations. *Proc Natl Acad Sci U S A* **108**, 11983-11988 (2011).
12. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* **107**, 961-8 (2010).
13. Tennessen, J.A., O'Connor, T.D., Bamshad, M.J. & Akey, J.M. The promise and limitations of population exomics for human evolution studies. *Genome Biol* **12**, 127 (2011).
14. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H. & Bustamante, C.D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5**, e1000695 (2009).
15. Schaffner, S.F. et al. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* **15**, 1576-1583 (2005).
16. Coventry, A. et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun* **1**, 131 (2010).
17. Nelson, M.R. et al. An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *Science* **337**, 100-104 (2012).
18. Rand, D.M. & Kann, L.M. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. *Mol Biol Evol* **13**, 735-48 (1996).
19. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-1081 (2009).
20. Adzhubei, I.A. et al. A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249 (2010).
21. Chun, S. & Fay, J.C. Identification of deleterious mutations within three human genomes.

Genome Res **19**, 1553-1561 (2009).

22. Schwarz, J.M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575-576 (2010).
23. Davydov, E.V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6**, e1001025 (2010).
24. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121 (2009).
25. Blekhman, R. et al. Natural selection on genes that underlie human disease susceptibility. *Curr Biol* **18**, 883-889 (2008).
26. Becker, K.G., Barnes, K.C., Bright, T.J. & Wang, S.A. The genetic association database. *Nat Genet* **36**, 431-432 (2004).
27. Liao, B.Y., Scott, N.M. & Zhang, J. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol Biol Evol* **23**, 2072-2080 (2006).
28. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* **38**, D355-60 (2009).

Supplemental Tables

Table S1. Phenotypes associated with samples and cohort contributions.

Phenotype	Cohort*	Sample Size
Asthma (n=191)	SARP	191
Blind (n=9)	WHI	9
Blood pressure (n=778)	ARIC	123
	CARDIA	115
	CHS	8
	FHS	153
	JHS	32
	MESA	42
	WHI	305
Body mass index and type 2 diabetes (n=637)	JHS	110
	MESA	26
	WHI	501
Chronic obstructive pulmonary disease (n=620)	COPDGene	286
	LHS	334
Cystic Fibrosis (n=417)	CF	417
Direct peritoneal resuscitation (n=943)	ARIC	104
	CARDIA	53
	CHS	90
	FHS	61
	JHS	49
	MESA	206
	WHI	380
Earlyonset stroke (n=491)	ARIC	9
	CHS	1
	FHS	23
	ISGS	74
	MESA	12
	SWISS	49
	WHI	323
Early onset myocardial infarction (n=1636)	ARIC	277
	CHS	84
	Cleveland Clinic	39
	FHS	155
	HARPS	402
	JHS	104
	MGH-PCAD	142
	PennCATH	31
	TRIUMPH	97
	WHI	305
LDL (n=625)	ARIC	283
	CARDIA	31
	CHS	25

	FHS	21
	JHS	58
	MESA	115
	WHI	92
Pulmonary hypertension (n=79)	PAH	79
Ventilator free days (n=89)	ALI	89

*Acute Lung Injury (**ALI**); Atherosclerosis Risk in Communities (**ARIC**); Coronary Artery Risk Development in Young Adults (**CARDIA**); Cystic Fibrosis (**CF**); Cardiovascular Health Study (**CHS**); Cleveland Clinic Genebank (**Cleveland Clinic**); COPD Genetic Epidemiology (**COPDGene**); Framingham Heart Study (**FHS**); Heart Attack Risk in Puget Sound (**HARPS**); The Ischemic Stroke Genetic Study (**ISGS**); Jackson Heart Study (**JHS**); Lung Health Study (**LHS**); Multi-Ethnic Study of Atherosclerosis (**MESA**); Massachusetts General Hospital Premature Coronary Artery Disease Study (**MGH-PCAD**); Pulmonary Arterial Hypertension (**PAH**); A catheterization-lab based cohort study from the University of Pennsylvania Medical Center (**PennCATH**); Severe Asthma Research Project (**SARP**); Siblings with Ischemic Stroke Study (**SWISS**); Translational Research Underlying Disparities in Myocardial Infarction Patients' Health Status (**TRIUMPH**); Women's Health Initiative (**WHI**).

Table S2. Published demographic models in populations of Europeans ancestry.

Models	Sample Size	T_{MRCA} (kya)	T_{OOA} (kya)	T_{Growth} (kya)	Initial N_e	Growth rate (%)
Gutenkunst et al. (2009)[14]	22	220	140	21.2	1000	0.40
Gravel et al. (2011)[11]	60	148	51	23	1032	0.38
Tennessen et al (2012)[1]	1351	148	51	23* 5.1	1032	0.307 1.95
Schaffner et al. (2005)[15]	62	425	87.5	8.75	7700	0.73#
Coventry et al. (2010)[16]	10422	425	87.5	1.4	7700	9.4
Nelson et al. (2012)[17]	12514	425	87.5	9.2	7700	1.7

*Initial European expansion started at 23 kya with a growth rate of 0.307% per generation, followed by an accelerated population growth with the growth rate of 1.95% per generation at 5,115 years ago.

#Fixed parameters that were not estimated from the data. (Growth is instantaneous to a fixed value of N_e=100,000, which is approximately equivalent to exponential growth of 0.73% pergeneration.)

Table S3. Average allele age (±sd) under different demographic models considering recent population growth in European.

Models	All SNVs (kyr)	Sharing SNVs (kyr)	Specific SNVs (kyr)	Deleterious SNVs (kyr)
Tennessen et al (2012)[1]	34.2±1.0	104.4±1.7	5.4±0.3	5.2±0.3
Coventry et al. (2010)[16]	61.2±1.7	188.1±3.0	10.8±0.3	8.8±0.6
Nelson et al. (2012)[17]	65.9±1.0	200.0±1.0	9.0±0.6	10.8±0.3

Table S4. Genes with excess of deleterious variants arose after the onset of the recent accelerated population growth.

Gene	Deleterious Variants		Neutral Variants		*P**
	<5kyr	≥5kyr	<5kyr	≥5kyr	
EAs					
LAMC1	41	0	101	42	$<10^{-6}$
ITPR3	48	0	188	77	$<10^{-6}$
ZFHX3	89	5	152	57	$<10^{-6}$
AAs					
UBR4	61	10	184	179	$<10^{-6}$
ATP1A4	17	3	49	57	10^{-6}
PKP4	30	2	42	40	$<10^{-6}$
KALRN	34	4	121	94	10^{-6}
PDE6B	17	0	60	55	2×10^{-6}
CPE	15	0	8	15	$<10^{-6}$
CDH12	11	0	17	34	2×10^{-6}
PCDHA9	24	3	189	166	$<10^{-6}$
PCDHA10	24	3	154	147	$<10^{-6}$
PCDHA11	24	3	128	134	$<10^{-6}$
KIAA0196	18	0	37	38	$<10^{-6}$
PFKP	15	0	72	88	$<10^{-6}$
ITGA8	25	1	64	63	$<10^{-6}$
PDGFD	16	0	27	22	3×10^{-6}
LRP1	40	2	217	157	$<10^{-6}$
MYO1E	31	2	85	66	$<10^{-6}$
PI4KA	26	2	104	99	$<10^{-6}$
SCUBE1	21	1	50	56	$<10^{-6}$

*Empirical *p* value was calculated by adjusting exon length in genes by simulations. Fisher's exact test was used to compare the number of deleterious and neutral variants arose before (≥5kyr) or after (<5kyr) the onset of recent explosive population growth. The null distribution was produced by randomly selecting regions with corresponding exon length of gene from whole-exome resequencing data. Genes with *p* value passing Bonferroni correction were listed.

Supplemental Figures

Figure S1. The relationship between Kinship and IBS before filtering of related or duplicated individuals.

Figure S2. PCA plot for assignment of EA and AA ancestry.

Figure S3. The distribution of SNVs with known ancestral states across exomes.

Figure S4. An example of a coalescent tree with n=5.

Figure S5. A demographic model with recent explosive population growth.

Figure S6. Validation of allele age estimated by a simulation approach in a constant population.

Figure S7. Effects of population growth, migration, and selection on the SFS and estimates of allele age.

Figure S8. Comparison of the SFS and allele age under different demographic models for EAs.

Figure S9. Influence of errors in sequencing data on estimates of average allele age.

Figure S10. Relationship between average allele age and proportion of putatively deleterious variants predicted by different methods in different classes of amino acid changes.

Figure S11. Comparison of neutrality index among regions with different average age.

Figure S12. Relationship between average allele age and putatively deleterious variants predicted by different methods.

Figure S13. Relationship between allele age and the proportion of SNVs predicted to be deleterious by different methods.

Figure S14. Ratio of the probability that a deleterious variant survives to the present day versus that of a neutral variant as a function of age and strength of selection.

Figure S15. The proportion of deleterious SNVs across KEGG pathways.

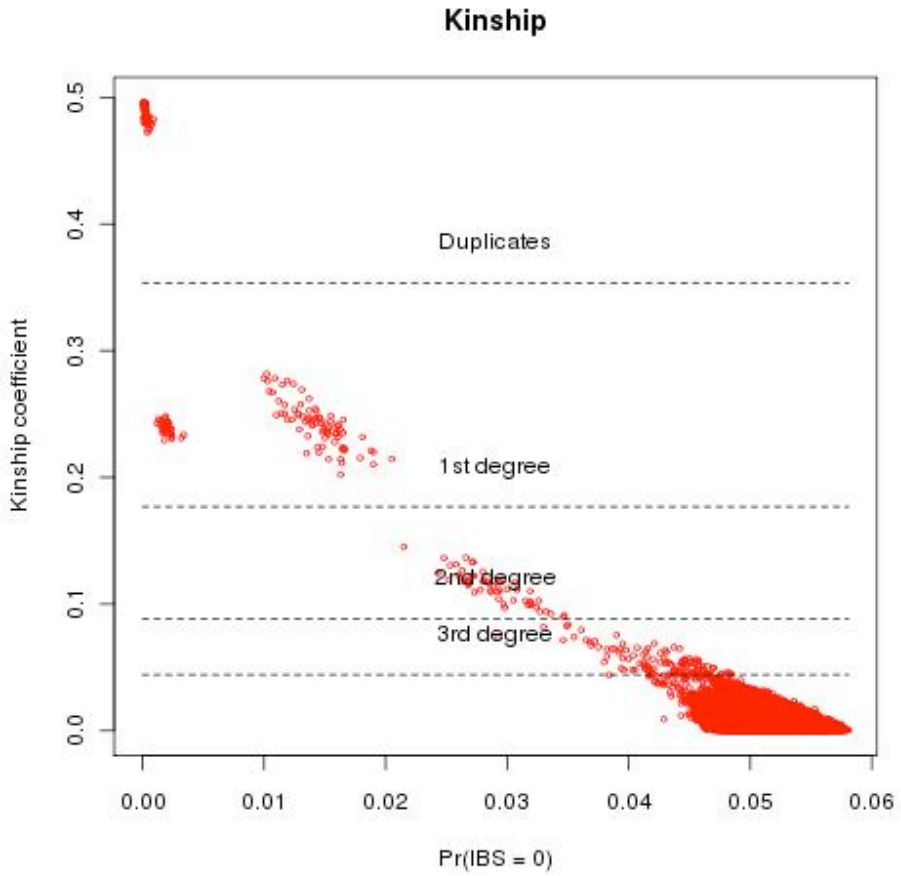

Figure S1. The relationship between Kinship and IBS before filtering of related or duplicated individuals.

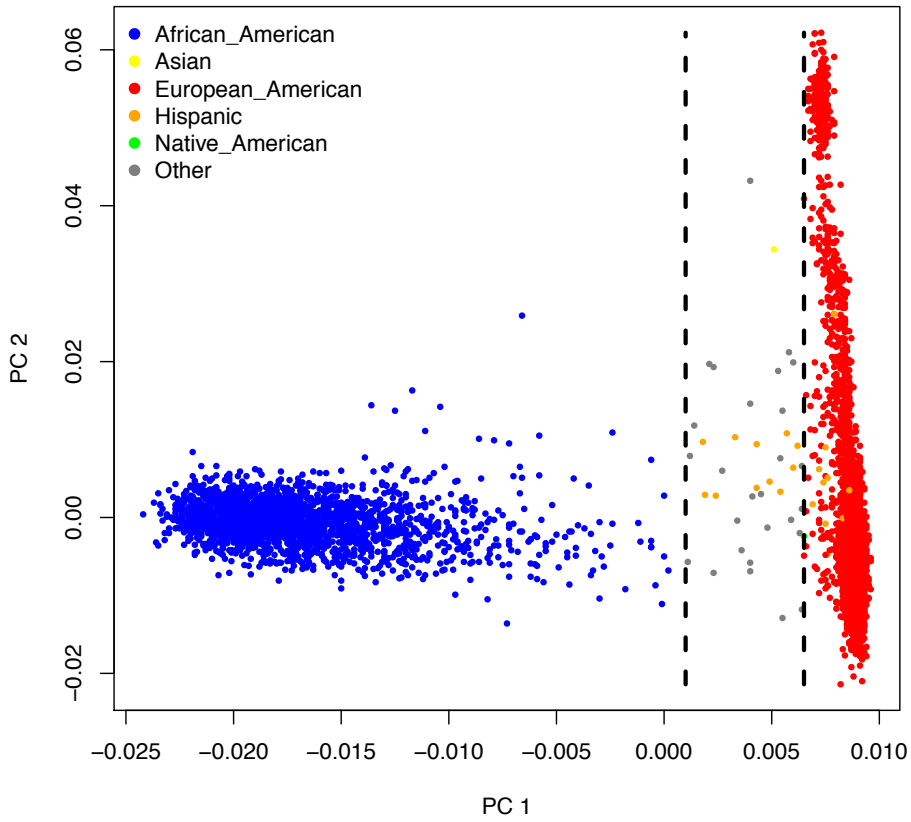

Figure S2. PCA plot for assignment of EA and AA ancestry.

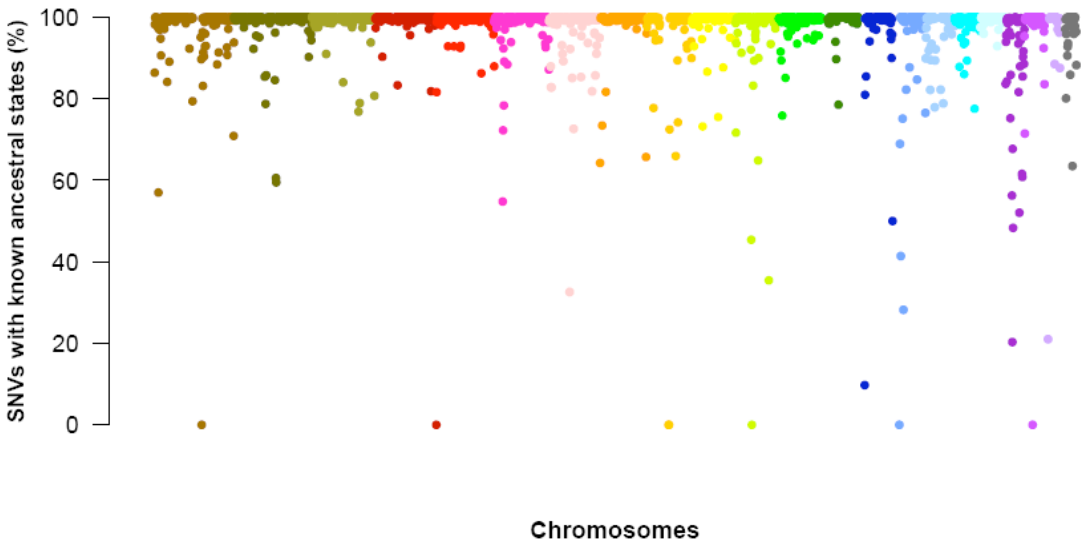

Figure S3. The distribution of SNVs with known ancestral states across exomes. In a 1Mb sliding window analysis, the proportion of SNVs with known ancestral states is high and uniformly distributed across the exomes, but decreases near the telomeric or centromeric regions.

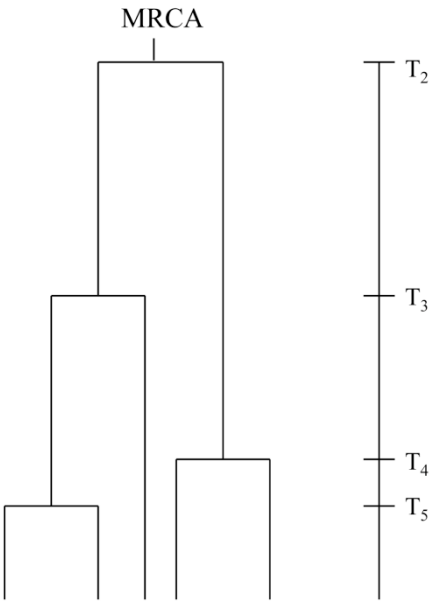

Figure S4. An example of a coalescent tree with n=5. T_k (k =5, 4, 3, 2) is coalescent time which measures the time from k lineages to (k-1) lineages. MRCA, most recent common ancestor.

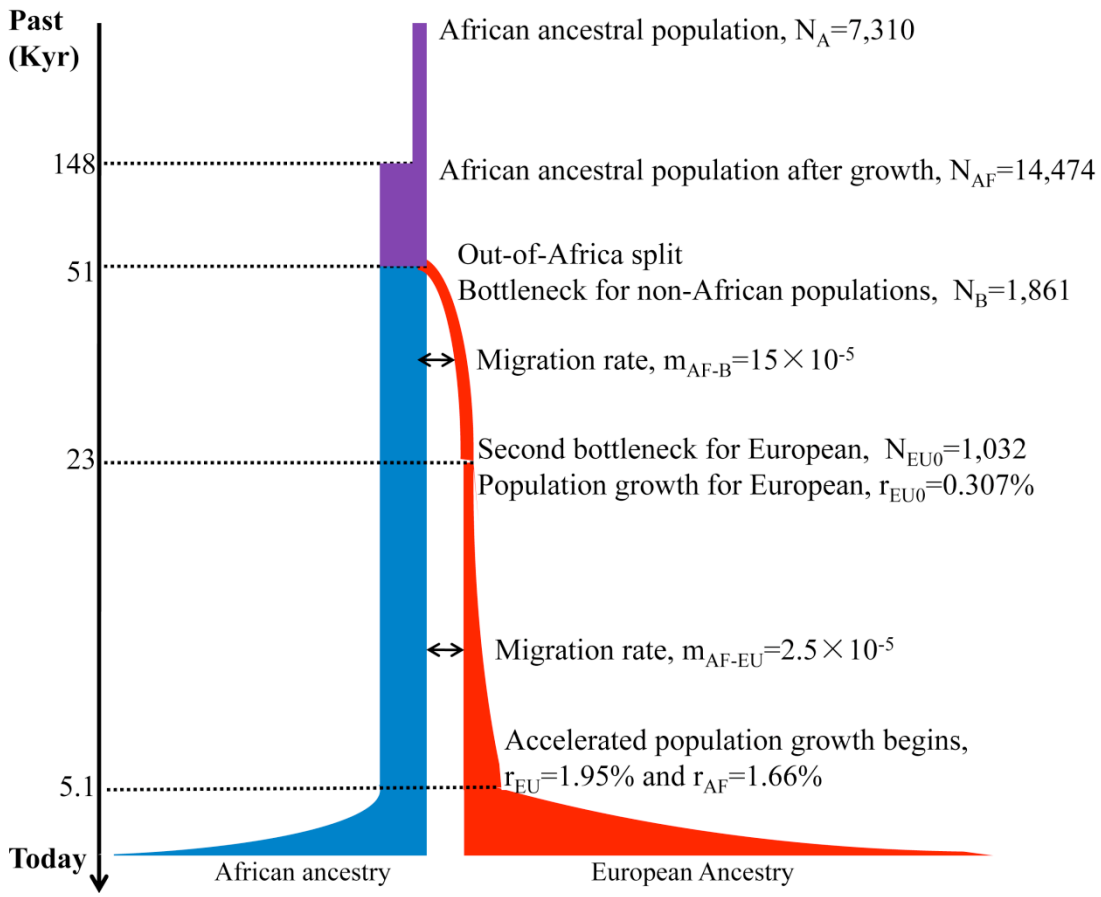

Figure S5. A demographic model with recent explosive population growth.

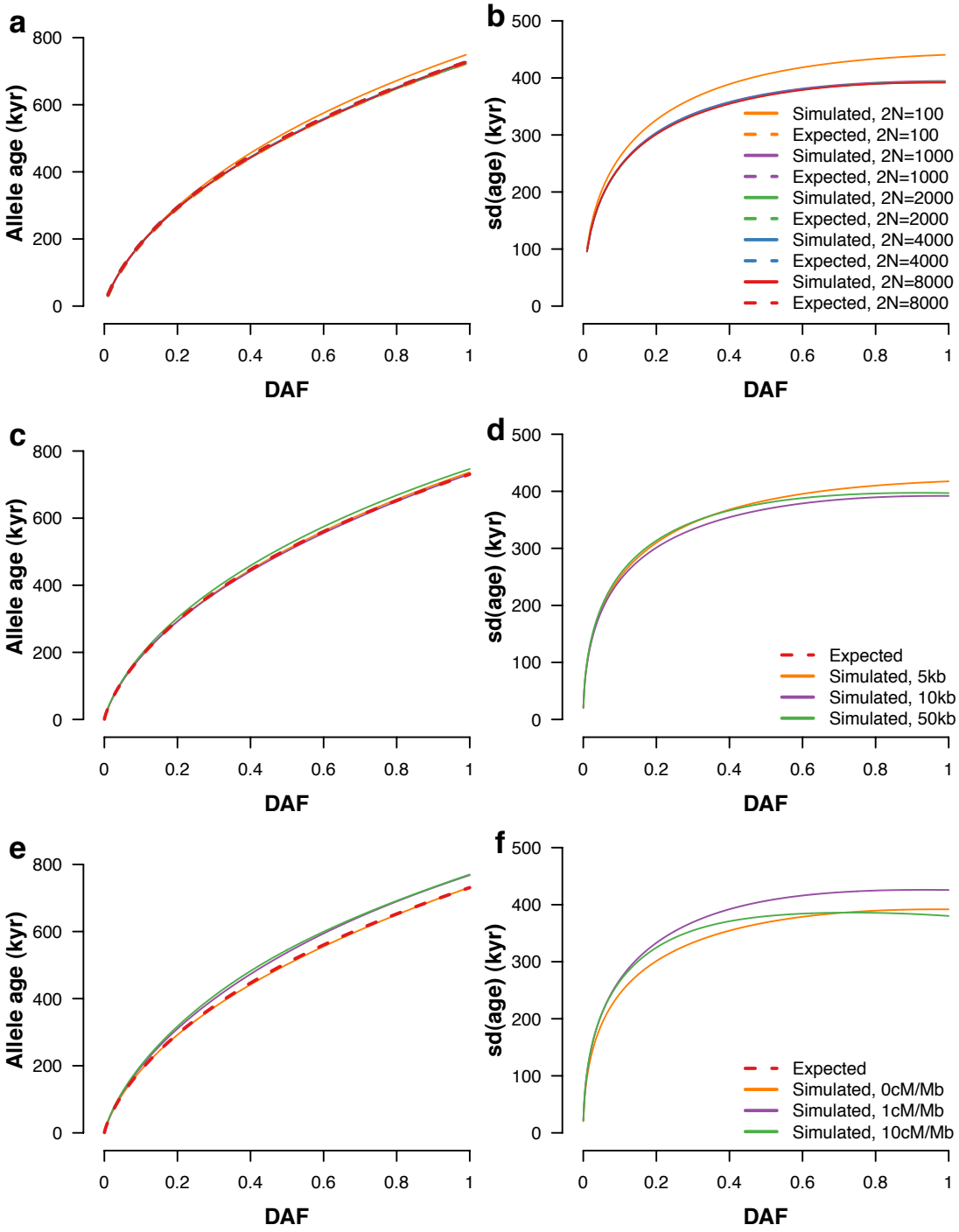

Figure S6. Validation of allele age estimated by a simulation approach in a constant population. **a** and **b** show the expected allele age and its standard deviation for different sample sizes; **c** and **d** show the expected allele age and its standard deviation for different sequence lengths; **e** and **f** show the expected allele age and its standard deviation for different recombination rates. **a-d** were calculated from 1,000 replicates, and **e-f** were calculated from 100 replicates.

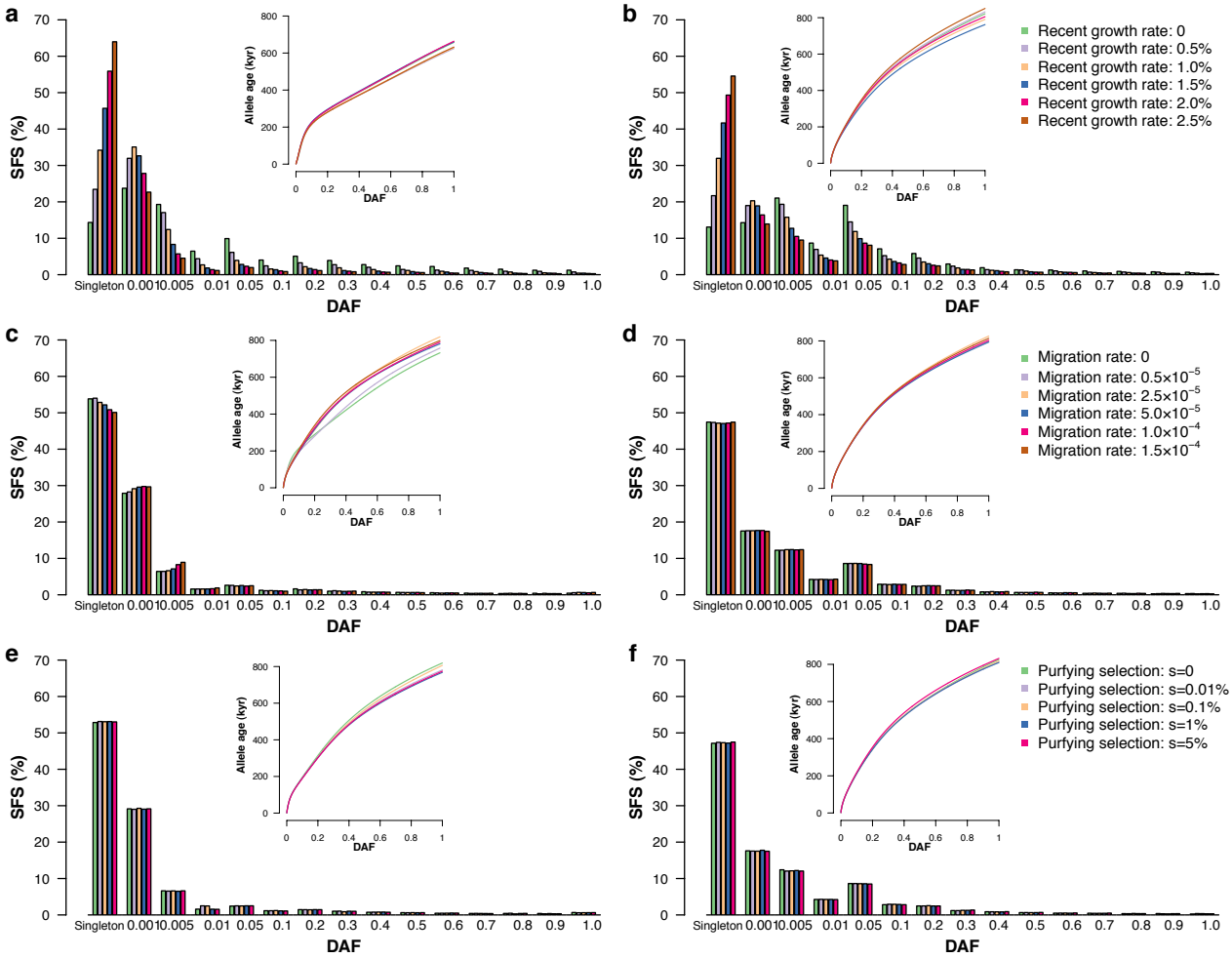

Figure S7. Effects of population growth, migration, and selection on the SFS and estimates of allele age. Simulations were performed based on a demographic model with recent accelerated population growth for EAs and AAs. **a** and **b** show the effect of different recent population growth rates (0, 0.5%, 1.0%, 1.5%, 2.0% and 2.5% per generation in the past 5,115 years) on the SFS and allele age; **c** and **d** show the effect of different migration rates between populations (0, 0.5×10^{-5}, 2.5×10^{-5}, 5×10^{-5},10^{-4} and 15×10^{-5} per chromosome per generation in the past 5,115 years) for the SFS and allele age; **e** and **f** show the effect of purifying selection (s=0, 0.01%, 0.1%, 1% and 5%) on the SFS and allele age. **a**, **c**, **e** for EAs and **b**,**d**,**f** for AAs. In each scenario, 1,000 replicates were conducted.

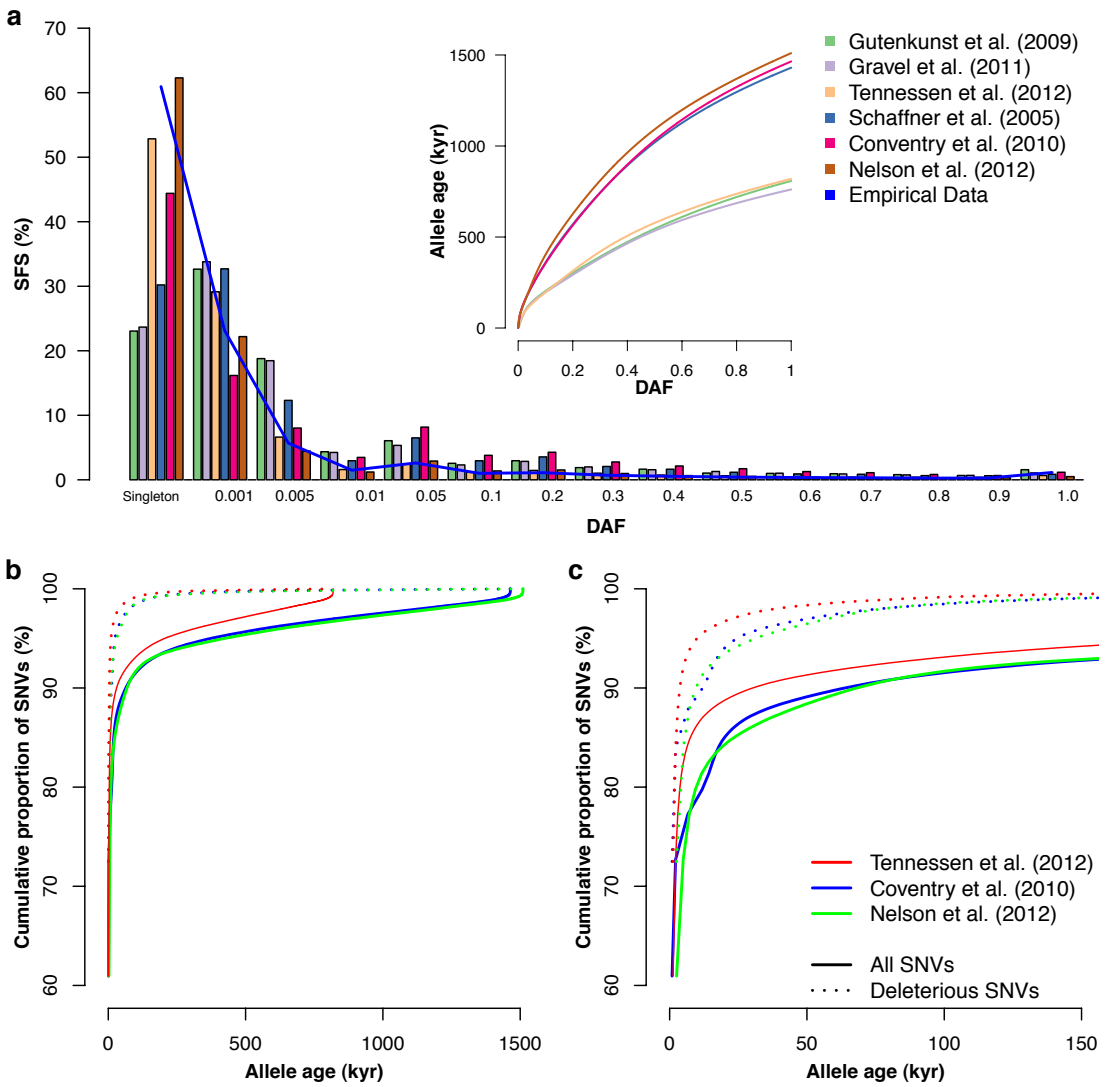

Figure S8. Comparison of the SFS and allele age under different demographic models for EAs. **a,** The SFS and estimates of allele age under different demographic models. **b,** The cumulative proportion of deleterious or all SNVs identified in 4,298 EAs as a function of allele age, according to three models considering explosive population growth. **c,** highlights the cumulative proportion of deleterious or all SNVs that are estimated to have arisen in the last 150 kyr.

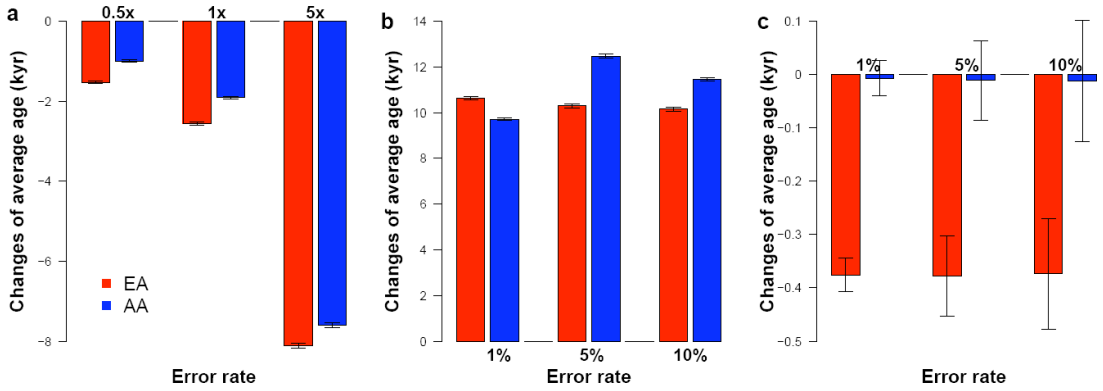

Figure S9. Influence of errors in sequencing data on estimates of average allele age. **a,** model of sequencing errors with a default per-base error rate of $x = 5.5\times10^{-7}$. **b,** model of genotyping errors with an error rate of 1%, 5%, or 10%. **c,** model of negative calls with a rate of 1%, 5%, or 10%. In each model, 1,000 replicates were conducted. Error bars denote standard deviations. The y-axis in each panel shows the change in average age (i.e., average age between models with minus without errors).

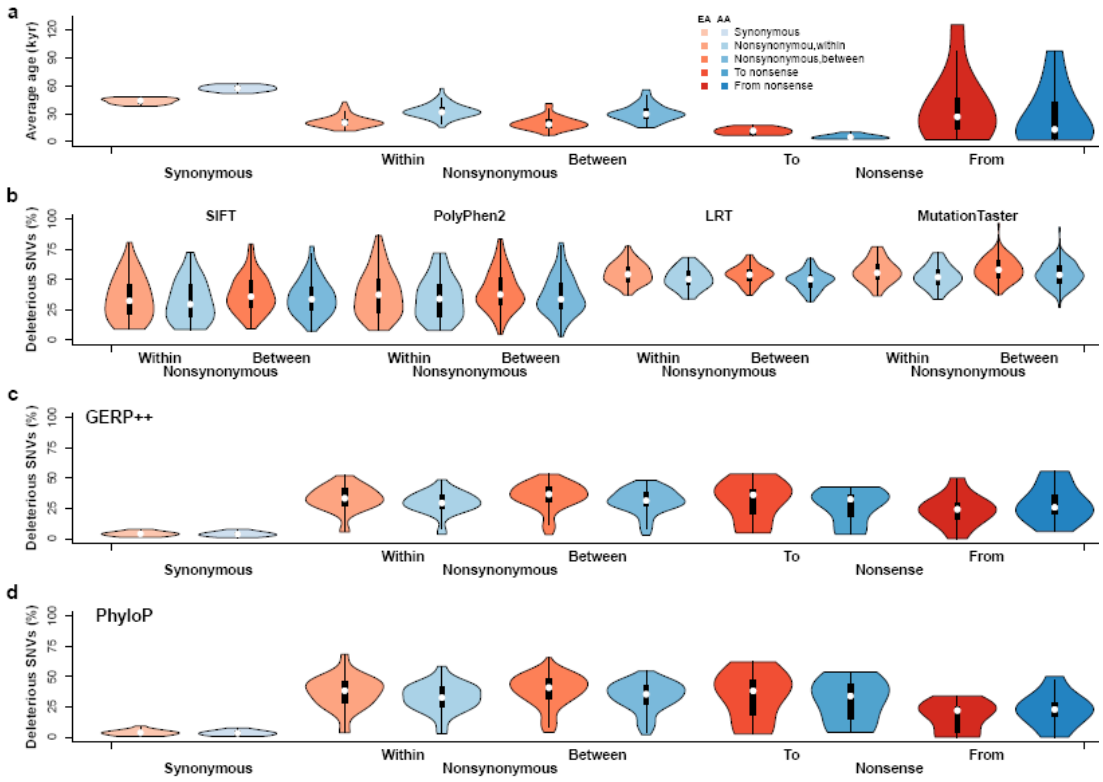

Figure S10. Relationship between average allele age and proportion of putatively deleterious variants predicted by different methods in different classes of amino acid changes. The changes of amino acid were classified as synonymous changes, nonsynonymous changes (i.e., changes within the same group and between groups), and changes from or to stop codons. **a,** Average allele age (kyr) for different classes of amino acid changes. **b,** The proportion of deleterious variants predicted by SIFT, PolyPhen2, LRT or MutationTaster in different classes of amino acid changes for nonsynonymous sites, **c** and **d,** The proportion of deleterious variants predicted by GERP++ and PhyloP respectively in different classes of amino acid changes for all coding sites. A significant negative relationship between average allele age and proportion of deleterious SNVs was observed in all prediction methods (i.e., SIFT, PolyPhen2, MutationTaster, GERP++ and PhyloP) (Spearman's correlation, $p<10^{-6}$), except for the LRT.

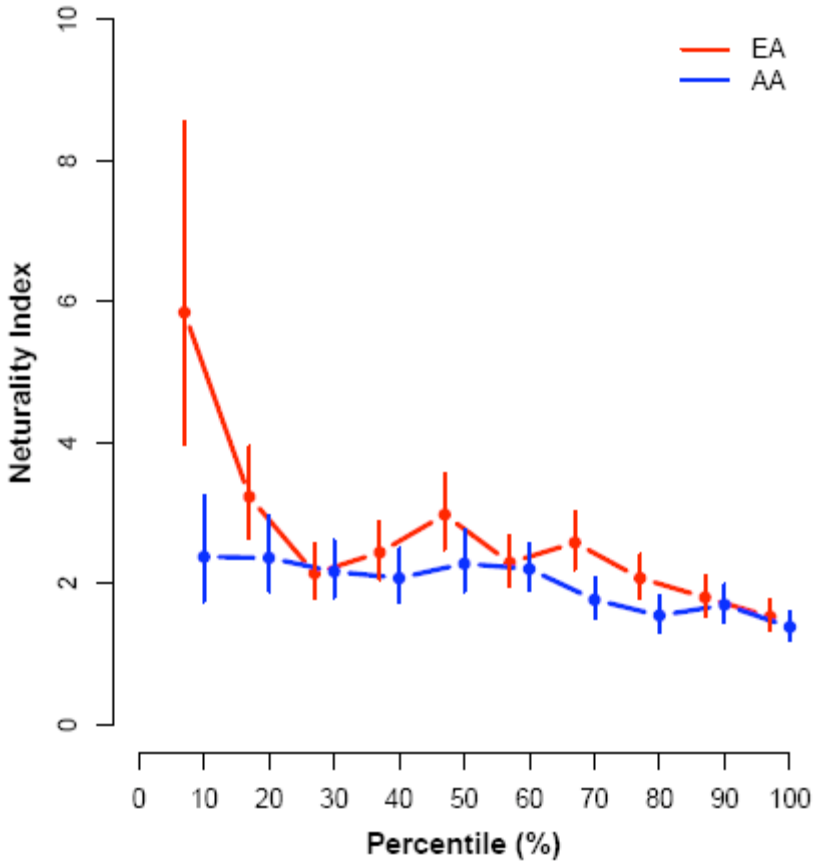

Figure S11. Comparison of neutrality index among regions with different average age. A 1Mb sliding window analysis was performed and windows were merged according to the percentile of their average ages. In each bin, neutrality index and its 95% confidence interval were calculated.

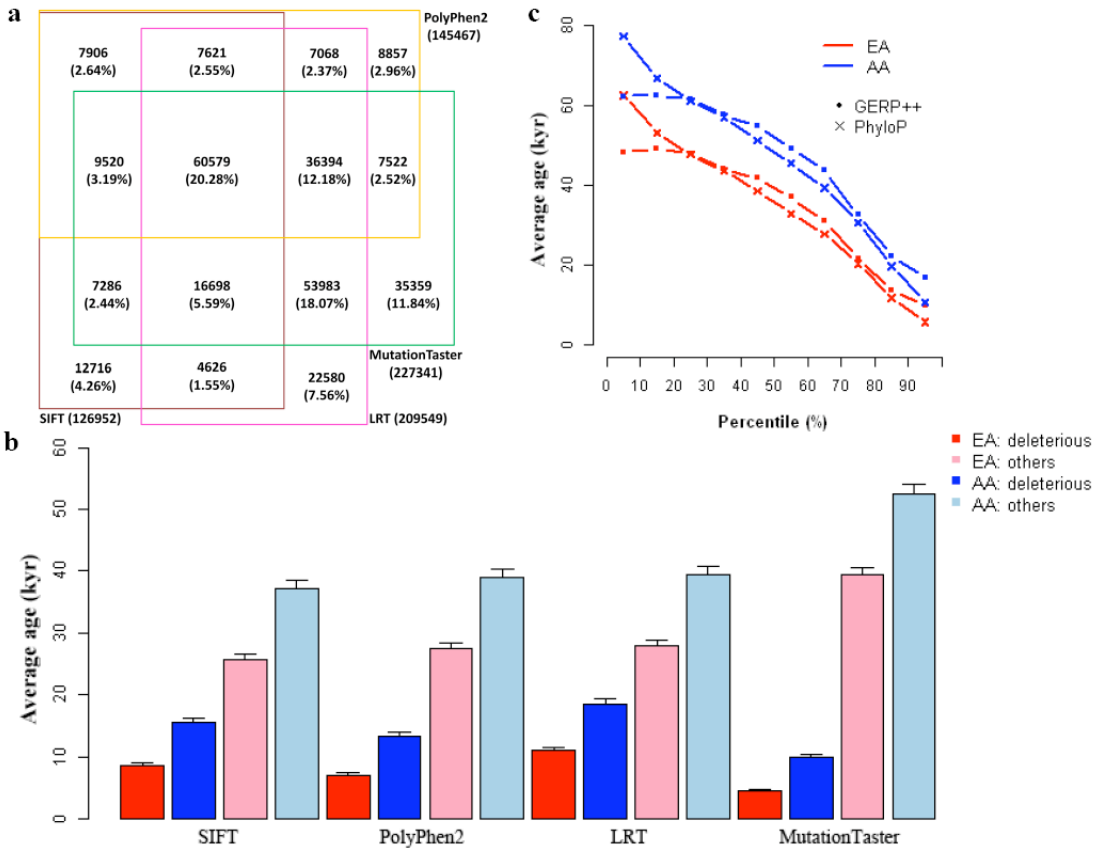

Figure S12. Relationship between average allele age and putatively deleterious variants predicted by different methods. For nonsynonymous variants, functionally damaging variants were predicted by four different methods: SIFT, PolyPhen2, a likelihood ratio test (LRT), and MutationTaster. **a,** The overlap of deleterious variants predicted by different methods. **b,** The average allele age for deleterious variants and other nonsynonymous variants. **c,** Variants were classified into 10 bins according to the percentile of conservation scores (i.e., GERP++ and PhyloP), and the average allele age for each bin is shown.

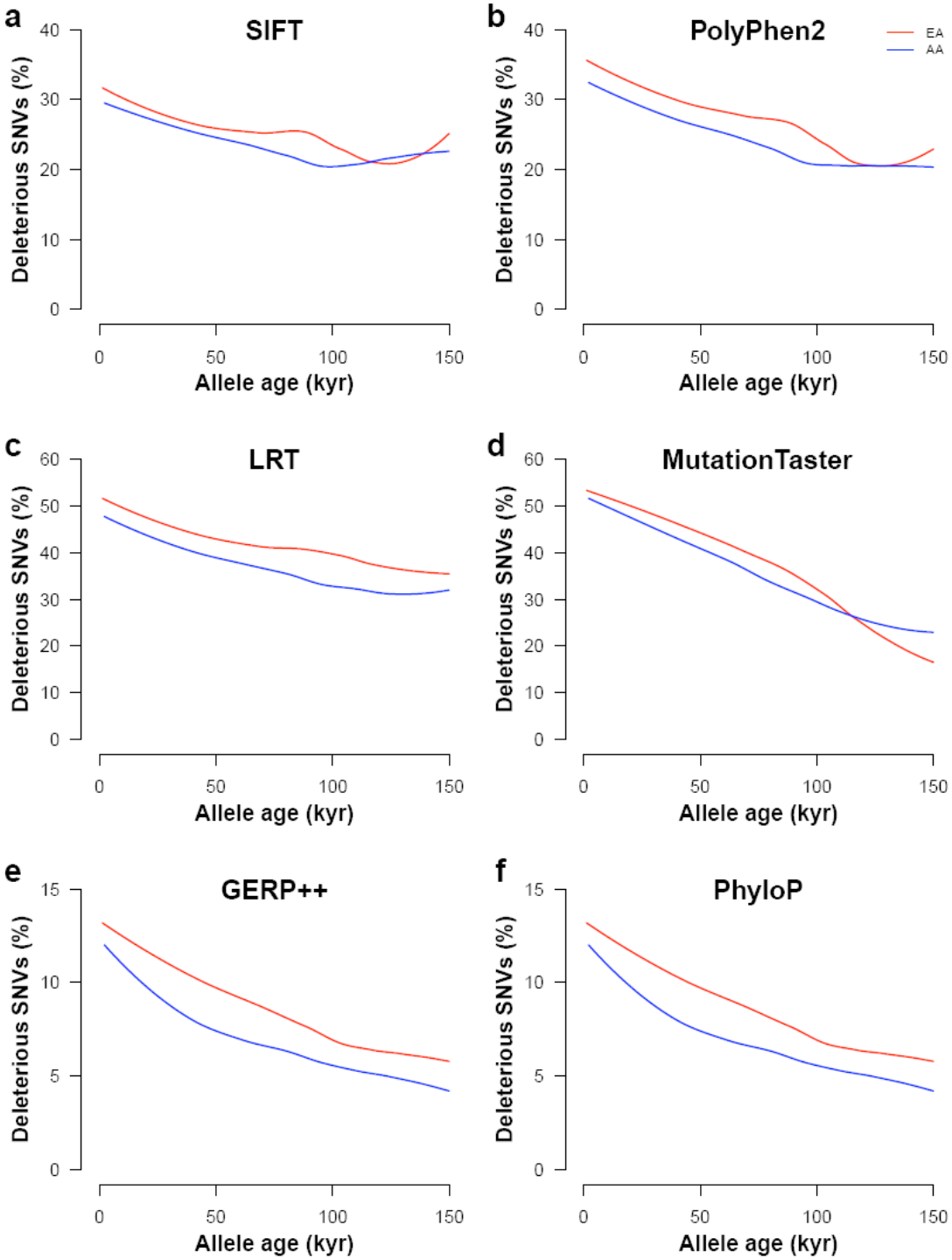

Figure S13. Relationship between allele age and the proportion of SNVs predicted to be deleterious by different methods. For nonsynonymous variants, deleterious variants were predicted by six different methods: SIFT, PolyPhen2, a likelihood ratio test (LRT), MutationTaster, GERP++ and PhyloP. For other variants, deleterious variants were predicted only by the latter two methods (i.e., GERP++ and PhyloP).

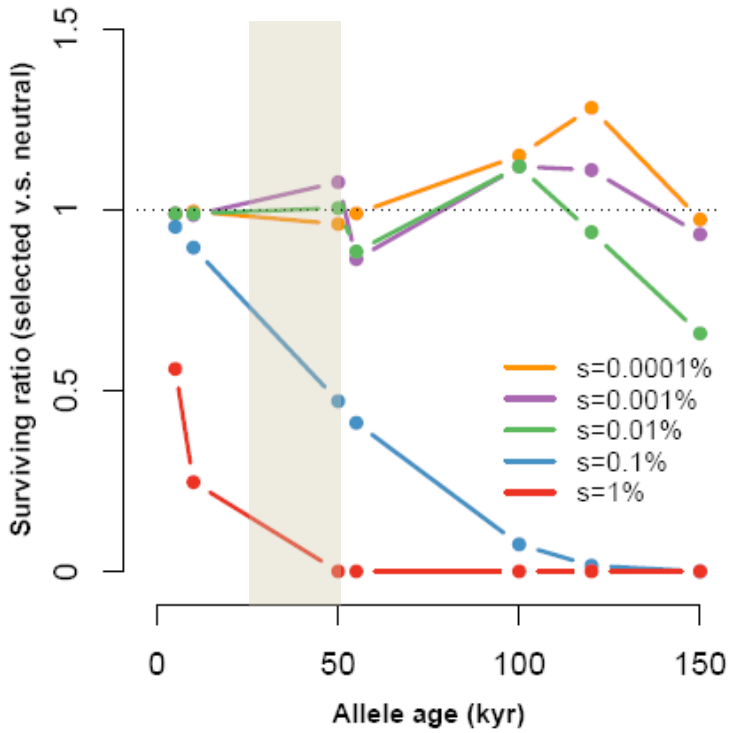

Figure S14. Ratio of the probability that a deleterious variant survives to the present day versus that of a neutral variant as a function of age and strength of selection. For each value, 1,000,000 replicates were performed. The shaded rectangle denotes the approximate time and duration of the population bottleneck, which began at 51 kyr and ended at 23 kyr.

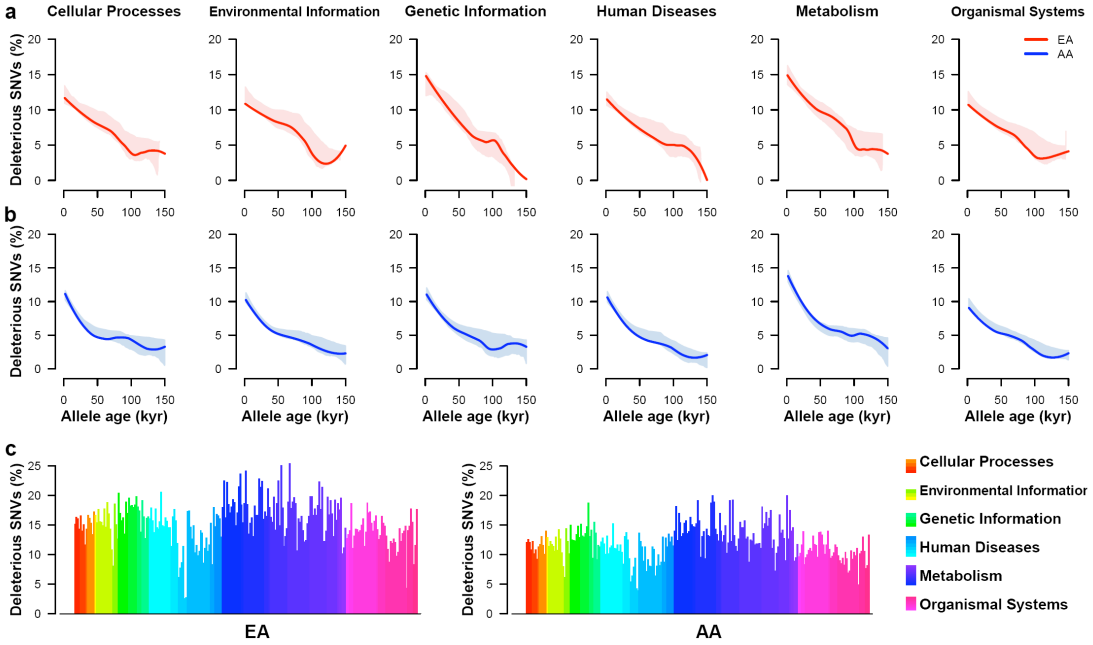

Figure S15. The proportion of deleterious SNVs across KEGG pathways. **a** and **b,** The proportion of deleterious SNVs for six broad classes of KEGG pathways in EAs and AAs respectively versus age in kyr. Shaded regions define 95% confidence intervals obtained by bootstrapping. **c**, The proportion of deleterious SNVs for 235 KEGG pathways.