Differences in disease burdens across human populations are governed more by neutral evolution than by natural selection

Ujani Hazra[1] and Joseph Lachance[1]*

[1] School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA

*Corresponding author: Joseph Lachance (joseph.lachance@biology.gatech.edu)

Keywords: evolutionary genetics, human genomics, natural selection, neutral evolution, polygenic risk scores

Abstract

The prevalence of most complex diseases varies across human populations, and a combination of socioeconomic and biological factors drives these differences. Likewise, divergent evolutionary histories can lead to different genetic architectures of disease, where allele frequencies and linkage disequilibrium patterns at disease-associated loci differ across global populations. However, it is presently unknown how much natural selection contributes to the health inequities of complex polygenic diseases. Here, we focus on ten hereditary diseases with the largest global disease burden in terms of mortality rates (e.g., coronary artery disease, stroke, type 2 diabetes, and lung cancer). Leveraging multiple GWAS and polygenic risk scores for each disease, we examine signatures of selection acting on sets of disease-associated variants. First, on a species level, we find that genomic regions associated with complex diseases are enriched for signatures of background selection. Second, tests of polygenic adaptation incorporating demographic histories of continental super-populations indicate that most complex diseases are primarily governed by neutral evolution. Third, we focus on a finer scale, testing for recent positive selection on a population level. We find that even though some disease-associated loci have undergone recent selection (extreme values of integrated haplotype scores), sets of disease-associated loci are not enriched for selection when compared to baseline distributions of control SNPs. Collectively, we find that recent natural selection has had a negligible role in driving differences in the genetic risk of complex diseases between human populations. These patterns are consistent with the late age of onset of many complex diseases.

Introduction

Disease risks have evolved substantially over recent human history (Crespi 2010; Quintana-Murci 2016). Increases in population size and changes in eating habits following the agricultural revolution have led to an increase in nutritional and infectious diseases and a decline in the overall health of many populations (Mummert, et al. 2011). While mortality from infectious diseases has decreased significantly in the 20th century (Armstrong, et al. 1999), the “transition to modernity” now puts the global population at a greater risk of non-communicable diseases (Corbett, et al. 2018). Indeed, the leading causes of death in sub-Saharan Africa have shifted from communicable diseases in children to non-communicable diseases in adults over the past three decades, with stroke, depression, diabetes, and ischemic heart disease dominating among middle-income countries (Bigna and Noubiap 2019).

Substantial heterogeneity in the mortality rates of non-communicable diseases exists across the globe (Warnecke, et al. 2008; Allen, et al. 2017). For example, disease burdens of stroke are high in Asia (Kim and Johnston 2011), and men of African descent suffer the highest mortality from prostate cancer (Rebbeck 2017). These and other health inequities arise from a complex combination of socioeconomic, demographic, environmental, and genetic causes. Socioeconomic factors like poverty and lack of access to quality treatment are known to increase chronic kidney disease risks (Nicholas, et al. 2015). Similarly, environmental factors like exposure to abandoned uranium mines have been reported to increase risks of hypertension, kidney disease, and cancer in some Native American populations (Lewis, et al. 2017). A population’s genetic makeup can also

impact disease susceptibility. For example, some women of Ashkenazi descent carry mutations in *BRCA1* and *BRCA2,* which subjects them to higher risks of breast cancer (Struewing, et al. 1997). We note that the narrow sense heritabilities of many complex diseases exceed 30%, i.e., a substantial proportion of the variance in disease risk is due to genetics (Visscher, et al. 2012).

The past decade has seen an upsurge in our collective understanding of the genetics of complex diseases. Genome-wide association studies (GWAS) have identified large numbers of disease-associated SNPs (Sollis, et al. 2023), and these SNPs can be used to generate polygenic predictions of disease risk (Lewis and Vassos 2020). One important lesson learned from GWAS is that most high-mortality non-communicable diseases are polygenic (Torkamani, et al. 2018), i.e., hereditary disease risks are due to the cumulative effects of many single nucleotide polymorphisms. Allele frequencies of disease-associated SNPs often vary among human populations, which in turn causes hereditary disease risks to vary across the globe (Adeyemo and Rotimi 2010). Multiple evolutionary phenomena contribute to population-level differences in allele frequencies, including natural selection (Lohmueller, et al. 2011) and stochastic processes like genetic drift and population bottlenecks (Tishkoff and Verrelli 2003; Chheda, et al. 2017). However, it is presently unknown how much natural selection, as opposed to neutral evolution, contributes to global health inequities.

Here, we focus on the ten hereditary diseases with the largest global disease burden in terms of mortality rates (Figure 1). Leveraging findings from multiple recent GWAS, we apply tests of natural selection to sets of disease-associated SNPs. We

address the following questions: 1) On a species level, have complex diseases experienced purifying selection? 2) To what extent are population-level differences in hereditary disease burdens due to polygenic adaptation and natural selection? 3) Are our findings robust to different ascertainment patterns of GWAS?

New Approaches

This paper examines whether sets of disease-associated SNPs are enriched for signatures of natural selection. As such, it focuses on signatures of selection acting on traits, as opposed to individual SNPs. Due to the highly polygenic nature of complex diseases, most individual SNPs have small effect sizes. However, significant evolutionary forces may be at play when multiple low-effect variants collectively contribute to disease susceptibility. Most existing tests of selection focus on individual SNPs or genes, including B-statistics, which identify loci under purifying selection (McVicker, et al. 2009), and integrative haplotype scores (iHS), which identify loci under recent positive selection (Johnson and Voight 2018). Recently, methods such as PolyGraph have been developed to identify selection acting on sets of SNPs (Racimo, et al. 2018). However, PolyGraph only focuses on adaptive evolution and does not leverage haplotype homozygosity information. Here, we adopt a polygenic framework that leverages B-statistics and iHS values to identify diseases that have been subject to purifying selection or recent positive selection.

Our approach consolidates SNP-level information to identify whether trait-associated SNPs are enriched for outlier values of test statistics compared to control

SNPs. Recognizing that each SNP does not contribute equally to disease risk, we account for their varying effects by weighting each data point by its effect size; outlier SNPs count more in our trait-level selection tests if they have large effect sizes. For each set of disease-associated SNPs, we obtained 1000 sets of matched control SNPs. These control SNPs are matched with respect to allele frequency, linkage disequilibrium (LD) patterns in the ascertained populations, gene density, and distance to the nearest gene. For each SNP set, we identify the proportion of SNPs, weighted by effect size, that exceeds an accepted outlier threshold ($B < 0.317$ for tests of background selection and $|iHS| > 1.96$ for tests of recent positive selection, see Methods). Enrichment tests involve comparing outlier proportions of disease-associated SNP sets to control sets to generate a percentile rank, with higher percentiles indicating greater trait-level signatures of selection (supplementary Fig. S1). Our approach differs from that of other research teams (Abraham, et al. 2022) in that we look for outlier enrichment, as opposed to trait averages, plus we weigh each SNP by effect size. Additional details can be found in the Methods section.

Results

Global differences in the mortality rates of polygenic diseases

Here, we focus on hereditary diseases that have the largest public health burden. Well-powered GWAS data exist for ten of the top twenty global causes of death, as reported by the WHO (World Health Organization 2020). These maladies are mostly comprised of cardiometabolic diseases, certain cancers, and neurological disorders (Table 1).

Although these diseases have the highest burden on a global scale, populations around the world differ significantly in their mortality rates, exceeding an order of magnitude in some cases. Focusing on nine countries that have comparable populations in the 1000 Genomes Project (1KGP) (1000 Genomes Project Consortium 2015), the heatmap in Figure 1 depicts mortality rates per 100,000 individuals for the ten polygenic diseases that have the largest global health burden. As seen in Figure 1, European countries have noticeably lower mortality rates of ischemic heart disease and stroke compared to other nations. By contrast, mortality rates of diabetes mellitus are considerably higher in South Asian and African countries. While socioeconomic and lifestyle factors play a considerable role in shaping mortality rates, these disparities can also be due to allele frequency differences at disease-associated loci.

To investigate natural selection acting on complex polygenic diseases, we compiled germline variants associated with the disease from publicly available GWAS data (Table 1). Using a pruning and thresholding approach, we obtained sets of independent SNPs associated for each disease. These sets of disease-associated SNPs were then used to test for polygenic signatures of background selection on a species-level, adaptation acting on continental scales, and recent positive selection in individual populations. Due to sample size and statistical power considerations, the main text of this paper primarily focuses on germline variants ascertained in European-ancestry GWAS. However, we later explore the impact of ascertainment bias and validate our results using germline variants ascertained in East Asian and multi-ancestry GWAS.

Evidence of background selection on a species level

Background selection (BGS) refers to reduced genetic diversity at a non-deleterious locus caused by negative selection against linked deleterious alleles. This term emphasizes that a neutral mutation's genomic environment or genetic background significantly influences whether it will be preserved or eliminated from a population. BGS has previously been shown to affect linkage disequilibrium patterns and the distribution of heritable variation across the genome (Gazal, et al. 2017; Zeng, et al. 2018; O'Connor, et al. 2019; Wendt, et al. 2021).

Given that BGS can influence the genetic architecture of complex traits, we tested whether SNPs that are associated with common polygenic diseases have undergone background or purifying selection. We used pre-computed B-statistics (McVicker, et al. 2009) to measure the impact of BGS near individual genomic loci. These statistics quantify the expected amount of genetic diversity flanking a given site in the genome. We extended the B-statistic framework to trait-level analyses by quantifying the extent that sets of disease-associated SNPs are enriched for outliers (see New Approaches and Methods).

SNPs that are associated with complex diseases are enriched for signatures of BGS. Figure 2 shows the percentile rank for each set of disease-associated SNPs compared to matched control sets. Percentile ranks range from 88.0 (colon cancer) to above 99.9 (chronic kidney disease and hypertensive heart disease), indicating that disease-associated SNPs are more likely to have outlier values of B-statistics. Overall, 8 out of 10 diseases had percentile ranks above 95, a fraction that was statistically significant (p-value = 1.605×10^{-9}, one-tailed binomial test). We note that these trait-level

signatures of BGS are not simply due to disease-associated SNPs being found in functional regions of the genome, as control sets are matched for distance to the nearest gene. Our background selection analyses focused on variation existing on a species-level. We next turn to signatures of selection acting on continental scales.

Minimal signatures of polygenic adaptation on a continental scale

Polygenic adaptation occurs through slight shifts in allele frequency at multiple loci (Barghi, et al. 2020). Although individual allele frequency changes may be small, their collective impact on the disease can be substantial. Disease-associated SNPs often vary in their allele frequencies across global populations (Kim, et al. 2018). Thus, we used PolyGraph (Racimo, et al. 2018) to quantify if such differences are driven by polygenic adaptation for the ten complex diseases. PolyGraph detects adaptation of polygenic traits due to allele frequency shifts at multiple loci using an admixture graph framework that considers the historical divergence of populations. It makes use of the ancestral and derived allele frequencies for each disease-associated loci at every population in the tree along with their effect sizes and compares them to a control distribution.

Tests of polygenic adaptation for the ten hereditary diseases with the largest public health burden are shown in Fig. 3. Although PolyGraph identifies weak signals of polygenic adaptation on some branches, FDR-adjusted q-values do not pass the threshold of statistical significance for most diseases. Branch-specific statistics from PolyGraph for each disease are listed in supplementary File S. Visually, this is illustrated by the preponderance of gray branches in Fig. 3. Although there are instances of branches with non-zero selection parameters (blue and red branches coloration in Fig.

3), these patterns were not replicated in PolyGraph analyses that used SNPs that were ascertained in other non-European GWAS (supplementary Figs. S2 and S3). Collectively, our PolyGraph analyses indicate that genetic drift is the primary cause of continental differences in allele frequencies for the diseases analyzed here. Subsequent tests of selection zoom in on individual populations.

Sparse signatures of recent positive selection on a local scale

To identify diseases under recent positive selection, we employ the integrated Haplotype Score (iHS), which can identify partial selective sweeps from stretches of extended haplotype homozygosity. iHS statistics are normalized based on a genome-wide empirical distribution, and extreme negative or positive iHS scores are considered potential indicators of recent positive selection ($|iHS| > 1.96$). Given iHS's emphasis on more recent selection, we narrowed our scope from major continental populations to 26 diverse populations from the 1KGP.

We performed an enrichment analysis to test if SNPs sets associated with each of the ten diseases are enriched for outlier iHS values when compared to controls. These analyses were repeated for all 26 populations in the 1KGP (Fig. 4). Higher percentiles in these polygenic tests are indicative of enrichment for outlier iHS values, i.e., recent positive selection. Notably, most diseases show low percentile values in all 26 populations, implying that the complex diseases analyzed in this study are not major targets of recent positive selection. Overall, only 6 out of 260 tests had percentile ranks above 95 when compared to controls (p-value = 0.9906, one-tailed binomial test).

Interestingly, ischemic heart disease shows some enrichment for outlier iHS values in South Asian populations, while hypertensive heart disease exhibits the most pronounced enrichment in genomes from Lima, Peru (PUR). The Peruvian population also demonstrates enrichment for other diseases when tested with SNP sets ascertained in non-European populations. Recent studies have shown evidence of associations between cardiovascular disease and adaptation to high altitude in Peruvian populations (Caro-Consuegra, et al. 2022; Hernandez-Vasquez, et al. 2022). These findings, along with our results, suggest that adaptive alleles may have pleiotropic effects with respect to disease risks. However, it is crucial to note that none of the observed percentile scores are high enough to withstand Bonferroni corrections.

Robustness of our findings to ascertainment bias

A major challenge when using GWAS data is ascertainment bias (Kim, et al. 2018). The ability to infer disease associations relies on allele frequencies being within an intermediate range in the discovery population, coupled with substantial effect sizes. This means that sets of disease-associated SNPs can differ across studies, particularly when the ancestries of study participants differ. This inherent variability in SNP sets and effect sizes can potentially yield varying outcomes in tests of polygenic selection. In this paper, we comprehensively address the issue of ascertainment bias by evaluating whether the conclusions of our polygenic tests of natural selection are similar for GWAS SNPs that were ascertained in different populations. When possible, we analyzed three different ascertainment schemes for each disease, i.e., SNP sets that were ascertained in European, East Asian, and multi-ancestry GWAS (Table 1).

Our tests of polygenic selection reveal consistent patterns regardless of the ancestry of the original source GWAS (Table 1). Although isolated exceptions exist, we found that disease-associated SNPs were strongly enriched for signatures of BGS regardless of whether the original GWAS was European, East Asian, or multi-ancestry (compare Fig. 2 and supplementary Figs. S4 and S5). Similarly, tests of positive selection acting on continental and local scales revealed that most differences in complex disease risks are not driven by natural selection. Although there were slightly stronger signatures of positive selection for SNPs that were ascertained in East Asian GWAS, PolyGraph results were largely robust to GWAS ancestry (compare Fig. 3 and supplementary Figs. S2 and S3). The haplotype homozygosity of disease-associated variants did not appreciably differ from that of control sets, and this pattern was consistent across ancestries (compare Fig. 4 and supplementary Figs. S6 and S7). Although the detectable genetic architectures of complex diseases may differ between populations, the genomic signatures of selection acting on these traits are largely robust to ascertainment bias.

Discussion

Focusing on the ten diseases with the largest global health burden, we tested whether sets of disease-associated SNPs are enriched for signatures of natural selection. B-statistics revealed that most complex diseases have been subject to purifying selection on a species-level. Results from Polygraph and iHS statistics were largely negative. This implies that recent positive selection has not been a major driver of population-level differences in the risks of polygenic diseases.

Complex disease risks appear to have evolved neutrally over recent human history. Although frequencies of disease-associated alleles differ between populations, these differences are largely due to genetic drift. Population genetics theory reveals that effects of genetic drift are inversely proportional to effective population size. Because of this, population bottlenecks and serial founder effects are likely to have had an outsized role in the divergence of hereditary disease risks across human populations (Keinan, et al. 2007). Our results are consistent with prior studies that that have found minimal evidence of selection in traits like type 2 diabetes in the Polynesians (Sun, et al. 2021). We note that our study focused on polygenic signatures of selection. Exceptions to this general pattern exist for a small subset of disease-associated loci, and future studies examining whether these exceptions are due to pleiotropy or genetic hitchhiking are likely to be fruitful.

Socioeconomic factors likely contribute more to differences in disease burden than genetic differences at trait-associated SNPs. Although many complex diseases have substantial heritabilities (Visscher, et al. 2012), these traits are highly polygenic and allele frequency differences at numerous loci of small effect loci can balance out. Other factors, like education, income, and access to health care, play a large role in determining mortality rates. Indeed, the Human Development Index (HDI) is correlated with many public health statistics. For example, mortality rates of colorectal cancer are high in countries that have a high HDI, while mortality rates of ischemic heart disease are high in countries that have a low HDI (UNDP 2022). An intriguing avenue of future research

279 involves quantifying how much genotype-environment interactions contribute to health
280 disparities (Rosenberg, et al. 2019).

281 One potential limitation of our study is that it relies on disease associations inferred
282 from GWAS. By necessity, GWAS hits are subject to ascertainment bias. However, our
283 findings are robust to differences in the ancestries of discovery cohorts. Furthermore, the
284 "known unknowns" (Kim, et al. 2018), i.e., alleles of small effect that have yet to be
285 implicated in a GWAS, are unlikely to change the conclusions of this paper. Each of these
286 as-yet-undiscovered disease associations makes only a small contribution to heritability
287 and their collective summary statistics are expected to resemble genome-wide baselines
288 (Carvalho, et al. 2022). Regardless, genetic differences in disease burdens across human
289 populations appear to be governed more by neutral evolution than by natural selection.

290

291 # Methods

292 ## Datasets

293 We conducted a comprehensive analysis of genome-wide association studies (GWAS)
294 encompassing ten diseases across three distinct ascertaining populations: European,
295 East Asian, and multi-ancestry (Table 1). Notably, due to an insufficient number of
296 significant associations identified for Alzheimer's Disease in East Asian and multi-
297 ancestry ascertained GWAS, we excluded this trait from ascertainment bias testing.
298 Significant SNPs with a p-value $< 5x10^{-5}$ were extracted from each GWAS. Subsequently,
299 LD pruning was performed to isolate independent associations with an $r^2 < 0.2$ within the
300 respective ascertained population, utilizing Plink 1.9 (Chang, et al. 2015) and 1KGP

phase 3 data (1000 Genomes Project Consortium 2015) as a reference. To ensure uniformity, the LiftOver tool (Hinrichs, et al. 2006) was employed to convert all coordinates of all GWAS SNPs to the hg19 build.

In all our analyses, control SNPs were obtained using SNPSnap (Pers, et al. 2015).Matching criteria included allele frequency, LD patterns, distance to gene, and gene density in the ascertained population. SNPs within the HLA region were removed. For European and East Asian ascertained GWAS, controls were matched within their respective populations from the 1KGP. In the case of multi-ancestry studies, controls were matched across pooled data from European, East Asian, and African populations to yield sets of SNPs.

Trait-level distributions of summary statistics

For the enrichment analyses, our focus is on assessing whether sets of disease-associated SNPs, considered collectively, have undergone selection. To integrate the SNP-level information from test statistics into a comprehensive trait-level distribution, we employ kernel density estimation (KDE). This method allows us to derive a probability distribution of the test statistic for each trait. Unlike traditional estimation techniques, KDE is a nonparametric approach that does not assume that the data follows a known distribution. Instead, nonparametric models determine the structure from the underlying data itself. In our implementation, we opt for a Gaussian kernel and conduct a five-fold cross-validation using GridSearchCV (Pedregosa 2011) to determine the optimal kernel bandwidth for the KDE. Since each associated SNP also has a strength of association to

the disease (beta or effect size), we also weigh the SNPs according to their absolute effect sizes while implementing KDE. The outcome of KDE is a probability density function (PDF) with the area under the curve standardized to one.

Outlier Enrichment: Background Selection

We use B-statistic as a measure of background selection. B indicates the expected fraction of neutral diversity present at a site, with values close to 0 representing near complete removal of diversity due to selection and values near 1 indicating little effect. Using BEDTools (Quinlan and Hall 2010),we extracted B values for SNPs from GWAS and their matched controls.

To check for background selection enrichment, we focus on lower B-values and calculate the probability of the trait having a B value less than 0.317 (area under the PDF from 0 to 0.317, $AUC_{0.317}$). Previous research suggests a B value of around 0.317 is a threshold for the lowest 5% of B values across the human genome (Torres, et al. 2018).We create PDFs for 1000 matched control sets using similar KDE steps described above. We estimate the probability of having a B-statistic of less than 0.317 in the control sets, where the SNPs are not linked to the disease but have similar allele frequencies and distances to genes. Comparing the $AUC_{0.317}$ of the trait to the 1000 control $AUC_{0.317}$ gives us a percentile rank for the trait. A high percentile rank indicates that trait-associated SNPs are enriched for outlier B-statistics (supplementary Fig. S1A).

Previous research has demonstrated that the B-statistic, while prone to potential misestimation and influenced by the assumptions of the underlying model, reliably

preserves the correct rank order of SNPs (Comeron 2014; Torres, et al. 2018).Thus, we expect McVicker et al.'s inference of B to provide good separation between the regions experiencing the weakest and strongest background selection effects at linked sites within the human genome. Nevertheless, to ensure the robustness of our findings, we conducted additional enrichment analyses using more stringent B-statistic thresholds (0.2 and 0.1) and obtained consistent results (supplementary Fig. S8).

Outlier Enrichment: Recent Positive Selection

We use an integrated Haplotype Score (iHS) to measure recent positive selection in 26 global populations from the 1KGP (Johnson and Voight 2018). iHS values are assigned to each SNP in the genome and are normalized, with negative values indicating selection of the derived allele and positive values indicating selection of the ancestral allele. Since the iHS value is normalized genome-wide, any SNP with a value two standard deviations away from the mean i.e., |iHS|>1.96, is operationally considered to be under selection (Voight, et al. 2006).

Following the method detailed earlier, we construct trait-associated and 1000 control set distributions using kernel density estimation (KDE). Subsequently, we calculate the probability of iHS values exceeding 1.96 or falling below -1.96 in both the trait and control distributions. We then derive a percentile rank for the trait AUC in comparison to the 1000 control sets. Higher percentile ranks signify that the trait exhibits more extreme iHS values compared to the controls (see supplementary Fig. S1B).

Polygenic Adaptation

To investigate signals of polygenic adaptation, we use PolyGraph (Racimo, et al. 2018), a Markov Chain Monte Carlo (MCMC) algorithm that utilizes admixture graph information to deduce traces of polygenic adaptation in populations. To detect selection on a trait PolyGraph requires a set of summary statistics from GWAS, neutral or control SNPs that are not associated with the trait, and an admixture graph of the representative populations. PolyGraph requires knowledge of the ancestral alleles of all GWAS hits to polarize effect sizes. Thus, only GWAS hits where ancestral allele information was available from the 1KGP dataset were used in our study.

The same set of control SNPs used for the enrichment analyses was used to build an admixture graph using MixMapper (Lipson, et al. 2014). We made scaffold trees with eight continental populations and added the population from Peru (PEL) as an admixed population (note that one branch leading to PEL represents Native American ancestry). We ran PolyGraph with its default parameters using 1,000,000 MCMC steps. PolyGraph reports a selection parameter alpha for each disease, a product of the selection coefficient for the advantageous allele and the duration of the selective process, and a p-value for selection on the entire admixture graph. To correct for multiple testing, we calculated FDR-adjusted q-values from the overall p-values of selection from PolyGraph (Table 1).

Supplementary Material

Supplementary material includes supplementary File S1 (.xslx) and a merged .pdf containing supplementary Figs. S1-S8.

Acknowledgments

We thank Rohini Janivara, Aaron Pfennig, Mimi Holness, and members of the Center for Integrative Genomics at Georgia Institute of Technology for their insight and helpful comments. This work was supported by an NIH MIRA grant (R35GM133727). The funders did not have any role in this article's design, analysis, or writing.

Author contributions

U.H and J.L. conceived this study and developed methodology. U.H. curated GWAS datasets, conducted polygenic tests of selection, and performed data visualization. J.L. supervised this research and provided funding. U.H. and J.L. wrote and edited this manuscript.

Conflict of interest statement: None declared.

Data Availability

The GWAS summary statistics used in this paper are publicly available. Details about specific studies can be found in Table 1.

References

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526:68.

Abraham A, LaBella AL, Capra JA, Rokas A. 2022. Mosaic patterns of selection in genomic regions associated with diverse human traits. PLoS Genet 18:e1010494.

412 Adeyemo A, Rotimi C. 2010. Genetic variants associated with complex human diseases
413 show wide variation across multiple populations. Public Health Genomics 13:72-79.

414 Allen L, Cobiac L, Townsend N. 2017. Quantifying the global distribution of premature
415 mortality from non-communicable diseases. J Public Health (Oxf) 39:698-703.

416 Aragam KG, Jiang T, Goel A, Kanoni S, Wolford BN, Atri DS, Weeks EM, Wang M,
417 Hindy G, Zhou W, et al. 2022. Discovery and systematic characterization of risk variants
418 and genes for coronary artery disease in over a million participants. Nat Genet 54:1803-
419 1815.

420 Armstrong GL, Conn LA, Pinner RW. 1999. Trends in infectious disease mortality in the
421 United States during the 20th century. JAMA 281:61-66.

422 Barghi N, Hermisson J, Schlotterer C. 2020. Polygenic adaptation: a unifying framework
423 to understand positive selection. Nat Rev Genet 21:769-781.

424 Bellenguez C, Kucukali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, Naj AC,
425 Campos-Martin R, Grenier-Boley B, Andrade V, et al. 2022. New insights into the
426 genetic etiology of Alzheimer's disease and related dementias. Nat Genet 54:412-436.

427 Bigna JJ, Noubiap JJ. 2019. The rising burden of non-communicable diseases in sub-
428 Saharan Africa. Lancet Glob Health 7:e1295-e1296.

429 Byun J, Han Y, Li Y, Xia J, Long E, Choi J, Xiao X, Zhu M, Zhou W, Sun R, et al. 2022.
430 Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls
431 identifies new susceptibility loci contributing to lung cancer. Nat Genet 54:1167-1177.

432 Cai L, Wheeler E, Kerrison ND, Luan J, Deloukas P, Franks PW, Amiano P, Ardanaz E,
433 Bonet C, Fagherazzi G, et al. 2020. Genome-wide association analysis of type 2
434 diabetes in the EPIC-InterAct study. Sci Data 7:393.

435 Caro-Consuegra R, Nieves-Colon MA, Rawls E, Rubin-de-Celis V, Lizarraga B,
436 Vidaurre T, Sandoval K, Fejerman L, Stone AC, Moreno-Estrada A, et al. 2022.

437 Uncovering Signals of Positive Selection in Peruvian Populations from Three Ecological
438 Regions. Mol Biol Evol 39.

439 Carvalho NRG, Harris AM, Lachance J. 2022. Different genetic architectures of complex
440 traits and their relevance to polygenic score performance.
441 bioRxiv:2022.2010.2029.514295.

442 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-
443 generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7.

444 Chheda H, Palta P, Pirinen M, McCarthy S, Walter K, Koskinen S, Salomaa V, Daly M,
445 Durbin R, Palotie A, et al. 2017. Whole-genome view of the consequences of a
446 population bottleneck using 2926 genome sequences from Finland and United
447 Kingdom. Eur J Hum Genet 25:477-484.

448 Comeron JM. 2014. Background selection as baseline for nucleotide variation across
449 the Drosophila genome. PLoS Genet 10:e1004434.

450 Corbett S, Courtiol A, Lummaa V, Moorad J, Stearns S. 2018. The transition to
451 modernity and chronic disease: mismatch and natural selection. Nat Rev Genet 19:419-
452 430.

453 Crespi BJ. 2010. The origins and evolution of genetic disease risk in modern humans.
454 Ann N Y Acad Sci 1206:80-109.

455 Fernandez-Rozadilla C, Timofeeva M, Chen Z, Law P, Thomas M, Schmit S, Diez-
456 Obrero V, Hsu L, Fernandez-Tajes J, Palles C, et al. 2023. Deciphering colorectal
457 cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of
458 European and east Asian ancestries. Nat Genet 55:89-99.

459 Gazal S, Finucane HK, Furlotte NA, Loh PR, Palamara PF, Liu X, Schoech A, Bulik-
460 Sullivan B, Neale BM, Gusev A, et al. 2017. Linkage disequilibrium-dependent

architecture of human complex traits shows action of negative selection. Nat Genet 49:1421-1427.

Giri A, Hellwege JN, Keaton JM, Park J, Qiu C, Warren HR, Torstenson ES, Kovesdy CP, Sun YV, Wilson OD, et al. 2019. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. Nat Genet 51:51-62.

Hernandez-Vasquez A, Vargas-Fernandez R, Chacon-Diaz M. 2022. Association between Altitude and the Framingham Risk Score: A Cross-Sectional Study in the Peruvian Adult Population. Int J Environ Res Public Health 19.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, et al. 2006. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res 34:D590-598.

Ishigaki K, Akiyama M, Kanai M, Takahashi A, Kawakami E, Sugishita H, Sakaue S, Matoba N, Low SK, Okada Y, et al. 2020. Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. Nat Genet 52:669-679.

Johnson KE, Voight BF. 2018. Patterns of shared signatures of recent positive selection across human populations. Nat Ecol Evol 2:713-720.

Kanai M, Akiyama M, Takahashi A, Matoba N, Momozawa Y, Ikeda M, Iwata N, Ikegawa S, Hirata M, Matsuda K, et al. 2018. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat Genet 50:390-400.

Keinan A, Mullikin JC, Patterson N, Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet 39:1251-1255.

485 Kim AS, Johnston SC. 2011. Global variation in the relative burden of stroke and
486 ischemic heart disease. Circulation 124:314-323.

487 Kim MS, Patel KP, Teng AK, Berens AJ, Lachance J. 2018. Genetic disease risks can
488 be misestimated across global populations. Genome Biol 19:179.

489 Law PJ, Timofeeva M, Fernandez-Rozadilla C, Broderick P, Studd J, Fernandez-Tajes
490 J, Farrington S, Svinti V, Palles C, Orlando G, et al. 2019. Association analyses identify
491 31 new risk loci for colorectal cancer susceptibility. Nat Commun 10:2154.

492 Lewis CM, Vassos E. 2020. Polygenic risk scores: from research tools to clinical
493 instruments. Genome Med 12:44.

494 Lewis J, Hoover J, MacKenzie D. 2017. Mining and Environmental Health Disparities in
495 Native American Communities. Curr Environ Health Rep 4:130-141.

496 Lipson M, Loh PR, Patterson N, Moorjani P, Ko YC, Stoneking M, Berger B, Reich D.
497 2014. Reconstructing Austronesian population history in Island Southeast Asia. Nat
498 Commun 5:4689.

499 Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, Tian G,
500 Huerta-Sanchez E, Feder AF, Grarup N, et al. 2011. Natural selection affects multiple
501 aspects of genetic variation at putatively neutral sites across the human genome. PLoS
502 Genet 7:e1002326.

503 Lu Y, Kweon SS, Cai Q, Tanikawa C, Shu XO, Jia WH, Xiang YB, Huyghe JR, Harrison
504 TA, Kim J, et al. 2020. Identification of Novel Loci and New Risk Variant in Known Loci
505 for Colorectal Cancer Risk in East Asians. Cancer Epidemiol Biomarkers Prev 29:477-
506 486.

507 Mahajan A, Spracklen CN, Zhang W, Ng MCY, Petty LE, Kitajima H, Yu GZ, Rueger S,
508 Speidel L, Kim YJ, et al. 2022. Multi-ancestry genetic study of type 2 diabetes highlights
509 the power of diverse populations for discovery and translation. Nat Genet 54:560-572.

Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, Rutten-Jacobs L, Giese AK, van der Laan SW, Gretarsdottir S, et al. 2018. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet 50:524-537.

Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, Tyrer JP, Chen TH, Wang Q, Bolla MK, et al. 2019. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet 104:21-34.

McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, Caporaso NE, Johansson M, Xiao X, Li Y, et al. 2017. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet 49:1126-1132.

McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet 5:e1000471.

Mishra A, Malik R, Hachiya T, Jurgenson T, Namba S, Posner DC, Kamanu FK, Koido M, Le Grand Q, Shi M, et al. 2022. Stroke genetics informs drug discovery and risk prediction across ancestries. Nature 611:115-123.

Mummert A, Esche E, Robinson J, Armelagos GJ. 2011. Stature and robusticity during the agricultural transition: evidence from the bioarchaeological record. Econ Hum Biol 9:284-301.

Nicholas SB, Kalantar-Zadeh K, Norris KC. 2015. Socioeconomic disparities in chronic kidney disease. Adv Chronic Kidney Dis 22:6-15.

O'Connor LJ, Schoech AP, Hormozdiari F, Gazal S, Patterson N, Price AL. 2019. Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. Am J Hum Genet 105:456-476.

534 Pedregosa F, Varoquaux, G. , Gramfort, A. , Michel, V. , Thirion, B. , Grisel, O. ,
535 Blondel, M. , Prettenhofer, P. , Weiss, R. , Dubourg, V. , Vanderplas, J. , Passos, A. ,
536 Cournapeau, D. a, Brucher, M. , Perrot, M. ,Duchesnay, E. 2011. Scikit-learn: Machine
537 Learning in Python. Journal of Machine Learning Research 12:2825--2830.

538 Pers TH, Timshel P, Hirschhorn JN. 2015. SNPsnap: a Web-based tool for identification
539 and annotation of matched SNPs. Bioinformatics 31:418-420.

540 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
541 features. Bioinformatics 26:841-842.

542 Quintana-Murci L. 2016. Understanding rare and common diseases in the context of
543 human evolution. Genome Biol 17:225.

544 Racimo F, Berg JJ, Pickrell JK. 2018. Detecting Polygenic Adaptation in Admixture
545 Graphs. Genetics 208:1565-1584.

546 Rebbeck TR. 2017. Prostate Cancer Genetics: Variation by Race, Ethnicity, and
547 Geography. Semin Radiat Oncol 27:3-10.

548 Rosenberg NA, Edge MD, Pritchard JK, Feldman MW. 2019. Interpreting polygenic
549 scores, polygenic adaptation, and human phenotypic differences. Evol Med Public
550 Health 2019:26-34.

551 Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA, Batini
552 C, Fawcett KA, Song K, Sakornsakolpat P, et al. 2019. New genetic signals for lung
553 function highlight pathways and chronic obstructive pulmonary disease associations
554 across multiple ancestries. Nat Genet 51:481-493.

555 Shrine N, Izquierdo AG, Chen J, Packer R, Hall RJ, Guyatt AL, Batini C, Thompson RJ,
556 Pavuluri C, Malik V, et al. 2023. Multi-ancestry genome-wide association analyses
557 improve resolution of genes and pathways influencing lung function and chronic
558 obstructive pulmonary disease risk. Nat Genet 55:410-422.

559 Shu X, Long J, Cai Q, Kweon SS, Choi JY, Kubo M, Park SK, Bolla MK, Dennis J,
560 Wang Q, et al. 2020. Identification of novel breast cancer susceptibility loci in meta-
561 analyses conducted among Asian and European descendants. Nat Commun 11:1217.

562 Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, Groza T, Gunes O, Hall P,
563 Hayhurst J, et al. 2023. The NHGRI-EBI GWAS Catalog: knowledgebase and
564 deposition resource. Nucleic Acids Res 51:D977-D985.

565 Spracklen CN, Horikoshi M, Kim YJ, Lin K, Bragg F, Moon S, Suzuki K, Tam CHT,
566 Tabara Y, Kwak SH, et al. 2020. Identification of type 2 diabetes loci in 433,540 East
567 Asian individuals. Nature 582:240-245.

568 Struewing JP, Hartge P, Wacholder S, Baker SM, Berlin M, McAdams M, Timmerman
569 MM, Brody LC, Tucker MA. 1997. The risk of cancer associated with specific mutations
570 of BRCA1 and BRCA2 among Ashkenazi Jews. N Engl J Med 336:1401-1408.

571 Sun H, Lin M, Russell EM, Minster RL, Chan TF, Dinh BL, Naseri T, Reupena MS, Lum-
572 Jones A, Samoan Obesity L, et al. 2021. The impact of global and local Polynesian
573 genetic ancestry on complex traits in Native Hawaiians. PLoS Genet 17:e1009273.

574 Surendran P, Feofanova EV, Lahrouchi N, Ntalla I, Karthikeyan S, Cook J, Chen L,
575 Mifsud B, Yao C, Kraja AT, et al. 2020. Discovery of rare variants associated with blood
576 pressure regulation through meta-analysis of 1.3 million individuals. Nat Genet 52:1314-
577 1332.

578 Tcheandjieu C, Zhu X, Hilliard AT, Clarke SL, Napolioni V, Ma S, Lee KM, Fang H,
579 Chen F, Lu Y, et al. 2022. Large-scale genome-wide association study of coronary
580 artery disease in genetically diverse populations. Nat Med 28:1679-1692.

581 Tishkoff SA, Verrelli BC. 2003. Patterns of human genetic diversity: implications for
582 human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293-340.

583 Torkamani A, Wineinger NE, Topol EJ. 2018. The personal and clinical utility of
584 polygenic risk scores. Nat Rev Genet 19:581-590.

585 Torres R, Szpiech ZA, Hernandez RD. 2018. Human demographic history has amplified
586 the effects of background selection across the genome. PLoS Genet 14:e1007387.

587 UNDP. 2022. Human Development Report 2021-22. UNDP (United Nations
588 Development Programme).

589 Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery.
590 Am J Hum Genet 90:7-24.

591 Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection
592 in the human genome. PLoS Biol 4:e72.

593 Warnecke RB, Oh A, Breen N, Gehlert S, Paskett E, Tucker KL, Lurie N, Rebbeck T,
594 Goodwin J, Flack J, et al. 2008. Approaching health disparities from a population
595 perspective: the National Institutes of Health Centers for Population Health and Health
596 Disparities. Am J Public Health 98:1608-1615.

597 Wendt FR, Pathak GA, Overstreet C, Tylee DS, Gelernter J, Atkinson EG, Polimanti R.
598 2021. Characterizing the effect of background selection on the polygenicity of brain-
599 related traits. Genomics 113:111-119.

600 World Health Organization. 2020. Global Health Estimates 2020: deaths by cause A,
601 sex, by country and by region, 2000–2019. WHO; 2020. In.

602 Wuttke M, Li Y, Li M, Sieber KB, Feitosa MF, Gorski M, Tin A, Wang L, Chu AY,
603 Hoppmann A, et al. 2019. A catalog of genetic loci associated with kidney function from
604 analyses of a million individuals. Nat Genet 51:957-972.

605 Zeng J, de Vlaming R, Wu Y, Robinson MR, Lloyd-Jones LR, Yengo L, Yap CX, Xue A,
606 Sidorenko J, McRae AF, et al. 2018. Signatures of negative selection in the genetic
607 architecture of human complex traits. Nat Genet 50:746-753.

608

609 # Tables

610

Trait	Ascertained Population	B-statistic %ile	PolyGraph q-value	Max iHS %ile (population)
Ischemic Heart Disease	European (Aragam, et al. 2022)	98.8	0.1689	98 (ITU)
	East Asian (Ishigaki, et al. 2020)	>99.9	5.21×10^{-11}	>99.9 (TSI)
	Multi-ancestry (Tcheandjieu, et al. 2022)	90.3	0.6884	86 (PEL)
Stroke	European (Malik, et al. 2018)	98.8	0.2951	88 (YRI)
	East Asian (Ishigaki, et al. 2020)	95.6	0.0941	84 (GWD)
	Multi-ancestry (Mishra, et al. 2022)	97.1	0.6884	74 (IBS)
COPD	European (Shrine, et al. 2019)	98.8	0.9245	93 (PUR)
	East Asian (Ishigaki, et al. 2020)	>99.9	0.7000	95 (JPT)
	Multi-ancestry (Shrine, et al. 2023)	99.8	0.6884	89 (GIH, PUR)
Lung Cancer	European (McKay, et al. 2017)	91.8	0.9245	92 (GBR)
	East Asian (Ishigaki, et al. 2020)	74.9	0.1144	92 (GBR)
	Multi-ancestry (Byun, et al. 2022)	39.8	4.37×10^{-09}	>99.99 (PEL)
Alzheimer's Disease	European (Bellenguez, et al. 2022)	96.1	0.8876	94.5 (TSI)
Type 2 Diabetes	European (Cai, et al. 2020)	97.7	0.4697	85 (JPT)
	East Asian (Spracklen, et al. 2020)	94.1	0.1777	98 (CLM)
	Multi-ancestry (Mahajan, et al. 2022)	99.9	0.6884	99 (PEL)
Chronic Kidney Disease	European (Wuttke, et al. 2019)	>99.9	0.0113	98 (CDX, PEL)
	East Asian (Kanai, et al. 2018)	>99.9	0.0010	57 (CHB)
	Multi-ancestry (Wuttke, et al. 2019)	>99.9	0.3186	70 (FIN, GWD)
Hypertensive Heart Disorder	European (Surendran, et al. 2020)	>99.9	0.0160	99 (PEL)
	East Asian (Kanai, et al. 2018)	>99.9	0.0100	96 (ESN)
	Multi-ancestry (Giri, et al. 2019)	99.9	0.6884	90 (CHS)
Colon Cancer	European (Law, et al. 2019)	88.0	0.9245	78 (GIH)
	East Asian (Lu, et al. 2020)	99.9	0.0061	72 (MSL)
	Multi-ancestry (Fernandez-Rozadilla, et al. 2023)	99.7	0.7225	62 (KHV)
Breast Cancer	European (Mavaddat, et al. 2019)	99.2	0.9245	89 (CHS)
	East Asian (Ishigaki, et al. 2020)	78.6	0.0265	81 (CDX)
	Multi-ancestry (Shu, et al. 2020)	98.9	0.9716	94 (CHS)

611

612 **Table 1.** Top ten hereditary diseases with the highest global mortality from the 2020 World
613 Health Organization Report. The second column list ancestries of each source GWAS
614 used in our study. The third column summarizes the enrichment for BGS on these

615 diseases, comparing results across three ascertainment schemes to 1000 control sets.
616 The fourth column provides insights into polygenic adaptation signals, presenting FDR-
617 adjusted q-values. Finally, the last column list the 1KGP population(s) exhibiting the
618 highest enrichment for extreme iHS values in comparison to 1000 control sets of SNPs.

Figures

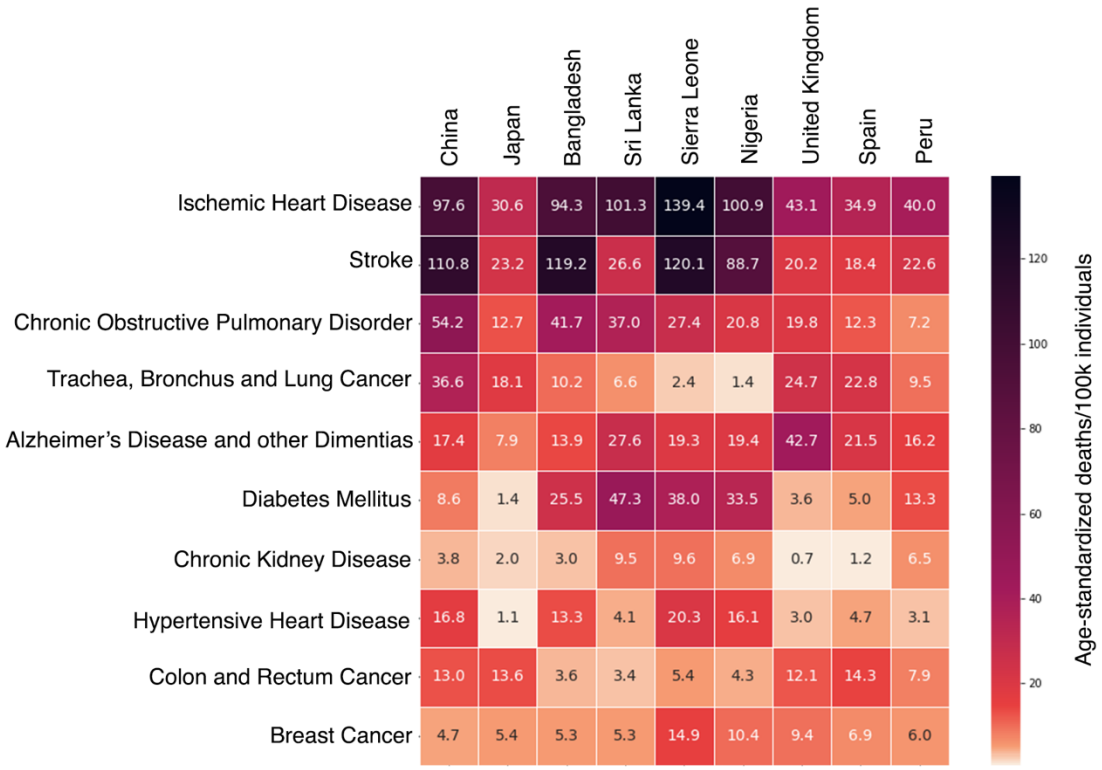

Fig. 1. Heatmap demonstrating the age-standardized mortality rates per 100,000 individuals for each disease in nine different countries (World Health Organization 2020). We observe heterogeneity in the mortality rates of each of these diseases. While some differences can be attributed to socioeconomic and lifestyle factors, this paper delves into the genetic contributors to each disease and tests if natural selection and a population's evolutionary history significantly contribute to such inequities.

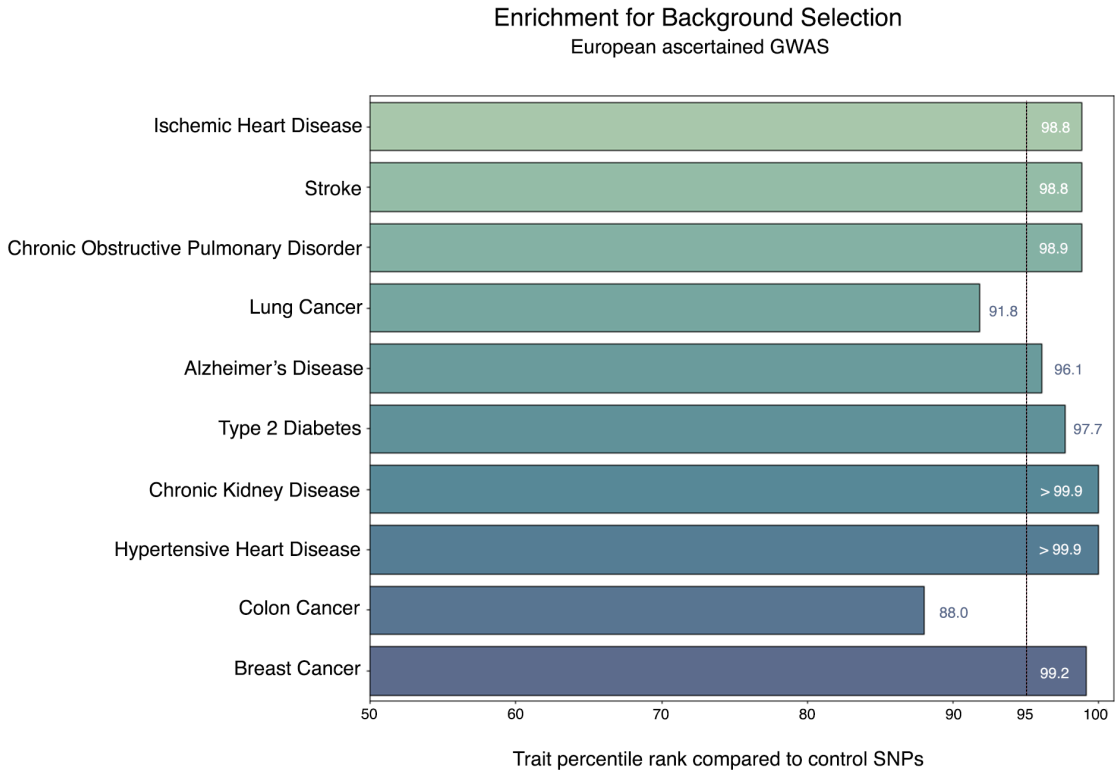

Fig. 2. Disease associated SNPs are enriched for signatures background selection. Plotted here are results from SNP sets that were ascertained in European ancestry GWAS. The percentile rank for each disease shows disease-associated SNPs are enriched for higher BGS compared to 1000 control sets before correcting for multiple testing, with a dotted line marks the 95th percentile of a control sets. SNP sets that were ascertained in East Asian and multi-ancestry GWAS yielded broadly similar patterns of BGS (supplementary Figs. S4 and S5). As per (Torres, et al. 2018), a B-statistic outlier threshold of 0.317 was used.

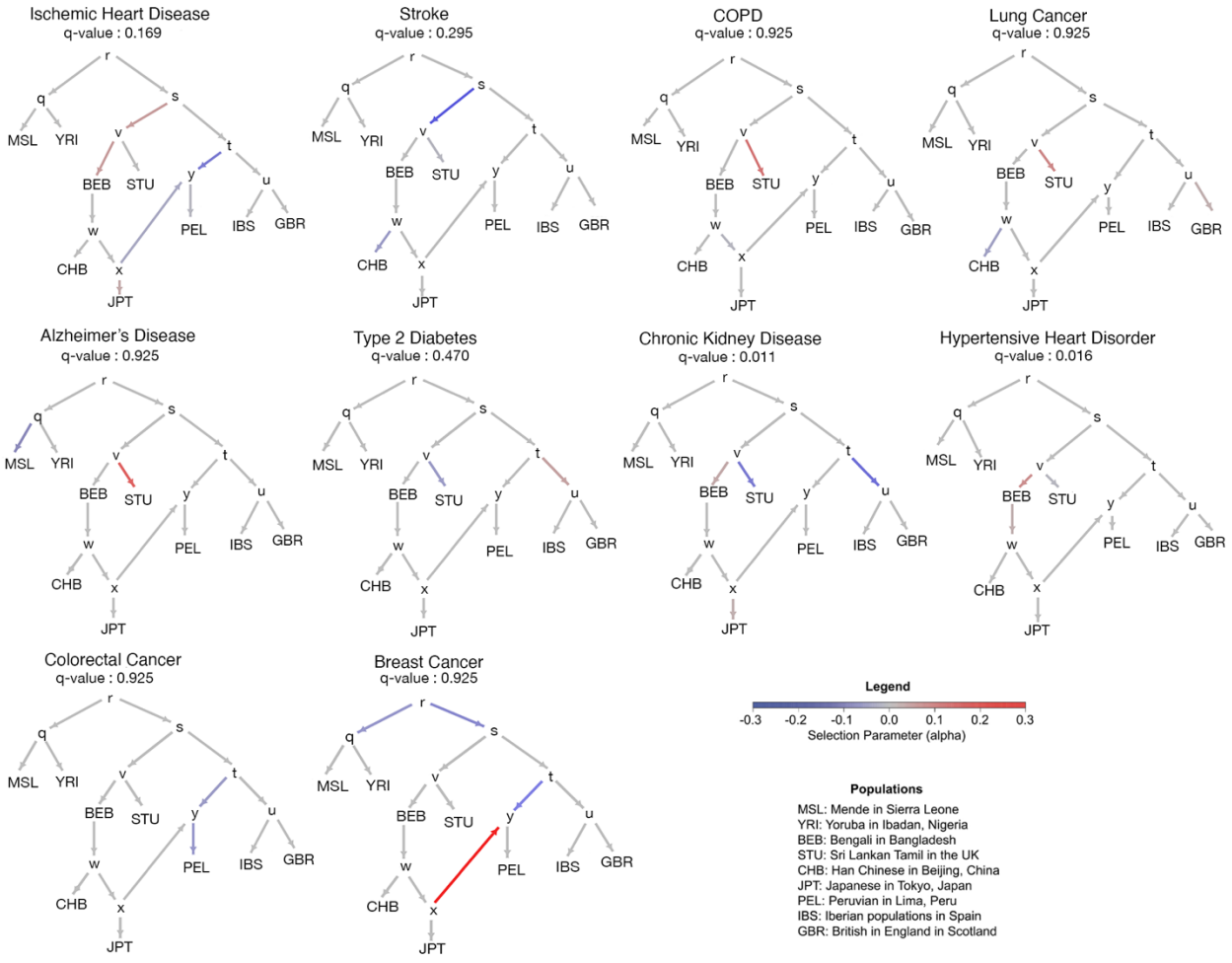

Fig. 3. Minimal evidence of polygenic adaptation acting on common diseases. Plotted here are results from SNPs sets that were ascertained in European ancestry GWAS. MixMapper was used to generate the admixture graph and PolyGraph was used to test for polygenic signatures of adaptation. FDR-adjusted q-values are above 0.05 for eight out of ten diseases. The selection parameter alpha reports a product of the selection coefficient for the advantageous alleles and the duration of the selective process. SNP sets that were ascertained in East Asian and multi-ancestry GWAS yielded broadly similar patterns of polygenic adaptation (supplementary Figs. S2 and S3).

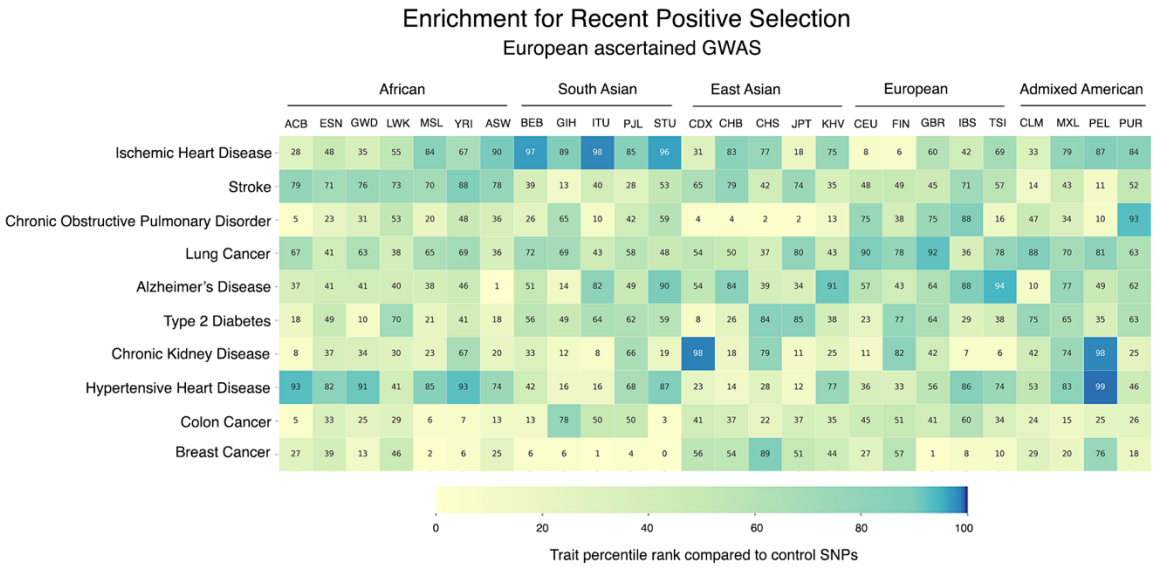

Fig. 4. Sparse signals of recent positive selection (partial sweeps) acting on complex diseases in 26 global populations from the 1KGP. Plotted here are results from SNPs sets that were ascertained in European ancestry GWAS. Percentile ranks quantify how much disease-associated loci are enriched for outlier iHS values compared to 1000 sets of control SNPs. Outlier threshold: |iHS|>1.96. Population acronyms are from the 1KGP. SNP sets that were ascertained in East Asian and multi-ancestry GWAS yielded broadly similar patterns (supplementary Figs. S6 and S7).