**Methods (Matt’s part)**

**Gathering Allele Frequency Data**

We gathered allele frequency data from the gnomAD dataset [1], a consortium of several ongoing global research projects. Specifically, we used GnomAD v3.1.2, the latest version of their data which consists of 76,156 whole genome sequences mapped to the GRCh38 reference. This data is labeled by gnomAD and available for download with information about the sample demographics. The data is divided into ten ancestry groups: African/African American samples (abbreviated “afr” in the gnomAD files); Amish (“ami”); Latino/Admixed American (“amr”); Ashkenazi Jewish (“asj”); East Asian (“eas”); Finnish (“fin”); Non-Finnish European (“nfe”); Middle Eastern (“mid”); South Asian (“sas”); and samples not assigned to any population, designated Other (“oth”). The data is also separately divided into two chromosomal sexes, labeled XX and XY, based on coverage of X and Y chromosomes for the given individual. Aneuploid individuals (e.g., X or XXY) were removed from the dataset. For the purposes of this study, we equated XX with Female and XY with Male. Total numbers of individuals sampled can be found on the gnomAD website’s help page (https://gnomad.broadinstitute.org/help).

The gnomAD data is available as direct download in variant call format (.vcf) from their online browser (<https://gnomad.broadinstitute.org/downloads>). We downloaded files for all somatic chromosomes then used the command-line program VCFTools [2] to parse the information. The original gnomAD .vcf files contain many more INFO fields than were necessary for the scope of our study, so we selected only the following specific columns: CHROM and POS, the chromosomal location; AN, the Allele Number, the total number of alleles sampled for the given locus; AC, the Allele Count, the number of alternative alleles found in the sample; and AF, the Allele Frequency, the frequency of alternative alleles in the sample (i.e., AC/AN). All of these fields (except for CHROM and POS) are given for each ancestry group and sex and are labeled as such in the .vcf fields e.g., AN\_afr\_XX.

**Filtering gnomAD Data and Calculating Male-Female FST**

We then passed this data through several filtering steps to remove loci which would not be useful to the analysis. During the gnomAD data acquisition process, we only chose sites for which the alternative alleles were single nucleotide polymorphisms (SNPs) as opposed to structural mutations such as insertions or deletions. After selecting the specific columns of interest but before performing any other filtering, our whole-genome dataset consisted of 2,288,867 loci. Using R [3], we filtered the data by first simply removing any loci with missing data (labeled as a “?” in the .vcf). This removed 3,698 sites, bringing the total number down from 2,288,867 to 2,285,169 loci. In an effort to avoid confounding results that could arise from demography substructure, we split the data into the different ancestry groups labeled by gnomAD and worked with the data in each subpopulation separately from this point forward. We removed any loci where there were less than 1,000 alleles in the sample for that site in each specific ancestry group. The number of loci we removed at this step depended on the sample size of each group, as we removed fewer loci from large populations and removed more loci in small populations (**S Table 1**). Importantly, from this filtering step we completely removed the Amish and Middle Eastern populations because their low sample sizes meant that there were no loci which had more than 1,000 alleles in the sample. For this reason, we disregarded these populations for all future work.

We now moved to calculating Male-Female FST from the information gathered from the gnomAD data for every site remaining in our dataset. Following the method and model proposed in Cheng and Kirkpatrick [4], we estimated FST using eq(3). To get and , we could directly use the AF\_XX column for each subpopulation. To get , we use (AC\_XX + AC\_XY) **/** (AN\_XX + AN\_XY), such that is given as the number of alternate alleles in the sample divided by the total number of alleles in the sample. Following these steps, we could easily calculate this estimate of FST.

Our last step of preliminary filtering was to remove any loci which resulted in an NA FST indicating that AN\_XX + AN\_XY = 0, or in other words there were no males or female samples labeled for that given locus. This step resulted in a significant reduction in locus number by one or two orders of magnitude (**S Table 2**).

**Integrating with GWAS Data and Calculating VGxSex**

Having completed these preliminary filtering steps, we now integrated the gnomAD data with the GWAS data. Following a similar filtering process as described above, we continued to remove loci by only selecting the loci which were found in both the gnomAD dataset and the GWAS dataset. The GWAS data contained 9,607,691 SNPs, but as the GWAS loci were gathered from the UK Biobank dataset they did not completely overlap with the loci from our filtered gnomAD dataset. We performed this comparison by identifying sites with the exact same genomic location (represented as chromosome number and position separated by a colon e.g., chr1:9999). The chromosome number and position were obtained directly from the GWAS table and the gnomAD vcf. The number of loci removed through this step varied greatly across ancestry groups (**S Table 3**).

In order to investigate the effect of Sexually-Antagonistic Selection in causing divergence in allele frequencies based on biallelic additive effect size, we defined the relationship between FST and VGxS as the coefficient A as in eq(1), defined FST as in eq(3), and defined VGxS as the proportion of phenotypic variance contributed by GxSex as in eq(2).

The GWAS dataset contained information on the effect size of the alternative allele in males and females on a trait-by-trait basis for 27 different physiological or physical traits, and we used these effect size estimates to stand in for and respectively; was obtained from the same calculated in the previous section. We also got an estimate of the variance of from an equation by the following relationship by eq(25). Here, we can substitute and with the GWAS effect sizes as above, and and with the GWAS standard error (SE) estimates for each locus’ effect size per trait.

Our final step of filtering involved filtering loci by GWAS p-value. To investigate the contribution of GxSex on a trait, we focused our attention on only those loci which had a significant correlation with the trait of interest as indicated by p-value estimates for each locus per trait. We set three maximum p-value thresholds at 1e-3, 1e-5, and 1e-8. This filtering resulted in another significant reduction in site number (**S Table 4**). Specifically, we decided to focus on the 1e-5 threshold because of it striking a good middle ground between filtering out many points but still leaving many loci to use for analysis. We used this reduced and filtered subset of loci for each ancestry group to perform our calculations of FST vs. V.

**Calculating A by Weighted Linear Regression of FST vs. VGxSex**

As we defined , we can by extension define . Thus, we can estimate A as the slope of a linear regression line between FST and VGxS. To do this, we used paired FST and VGxS points for all loci which passed all previous stages of filtering.

One issue that arises is our expectation that many of the loci which were filtered through the p-value thresholds will be in strong linkage disequilibrium (LD) with each other by simple virtue of GWAS experimental design which relies on close physical linkage of loci to draw conclusions about correlation between allele and phenotype. Thus, we needed to reduce the effect of LD as much as possible. We achieved this by using LD blocks established in previous studies [6] to subdivide the genome into regions of linkage such that we expect sites from different LD blocks will have little to no LD between them. This method splits the genome into 1703 blocks of varying sizes across all chromosomes.

To perform this LD blocking we iterated over the 1703 blocks and identified all post-filtering loci which fell within each block; we then randomly sampled one of these loci within each LD block and used this sample to get our FST vs. V points. We then performed a weighted linear regression on these sampled points, using 1/var[V] as the weight where var[V] was calculated as in eq(4). The slope of this regression line was our estimate of A, and our SE of this estimate equaled the SE of the regression line slope. We then replicated this estimation process 100 times to generate 100 estimates of A. We could then use the estimates of A coupled with the SE of these estimates to generate 100 Z-scores for A, where Z-score ; in this case is our estimate of A, is zero because we expect there to be no correlation between FST and VGxS for most traits, and is the standard error of each estimate of A. We estimated the SE of our 100 Z-scores as . We computed this for all traits in the Ashkenazi Jewish, Finnish, and Non-Finnish European populations as we expected that these populations might be demographically most similar to the UK Biobank dataset used to generate the GWAS data.

**Generating Plots**

We generated figures using the ggplot2 package in R with touch up in Adobe Illustrator, and aligned figures using the ggarrange function in the ggpubr package in R.

|  |  |  |  |
| --- | --- | --- | --- |
| **Ancestry** | **# Loci Before Filtering by < 1,000 Alleles in Sample** | **# Loci After Filtering** | **# Loci Removed** |
| African/African American | 2,285,169 | 2,282,394 | 2,775 |
| Amish | 2,285,169 | 0 | 2,285,169 |
| Latino/American Admixed | 2,285,169 | 2,282,394 | 8,910 |
| Ashkenazi Jewish | 2,285,169 | 2,247,203 | 37,966 |
| East Asian | 2,285,169 | 2,255,778 | 29,391 |
| Finnish | 2,285,169 | 2,249,604 | 35,565 |
| Middle Eastern | 2,285,169 | 0 | 2,285,169 |
| Non-Finnish European | 2,285,169 | 2,283,952 | 1,217 |
| South Asian | 2,285,169 | 2,167,009 | 118,160 |
| Other | 2,285,169 | 2,080,336 | 204,833 |

**S Table 1. Results of filtering loci by removing any loci with less than 1,000 alleles in the sample.** We removed all loci with less than 1,000 alleles in the sample in order to narrow our research on loci with sufficiently large sample sizes. The number of sites removed during this step was highly dependent on the overall sample size for each ancestry. The Amish and Middle Eastern sub-populations, having the fewest samples overall, had no loci which passed through this round of filtering. All other ancestries retained a relatively high number of loci.

|  |  |  |  |
| --- | --- | --- | --- |
| **Ancestry** | **# Loci Before Filtering by NA FST Values** | **# Loci After Filtering** | **# Loci Removed** |
| African/African American | 2,282,394 | 924,946 | 1,357,448 |
| Latino/American Admixed | 2,282,394 | 519,496 | 1,756,763 |
| Ashkenazi Jewish | 2,247,203 | 161,271 | 2,085,932 |
| East Asian | 2,255,778 | 300,323 | 1,955,455 |
| Finnish | 2,249,604 | 207,162 | 2,042,442 |
| Non-Finnish European | 2,283,952 | 1,104,008 | 1,179,944 |
| South Asian | 2,167,009 | 330,518 | 1,836,491 |
| Other | 2,080,336 | 267,812 | 1,812,524 |

**S Table 2. Results of filtering loci by removing any loci with NA FST Values.** FST was calculated for all loci using the Allele Count and Allele Number values in the gnomAD data set. We then removed any loci which had an FST value of NA in any ancestry, as this would indicate that there were 0 samples of that locus in at least one sub-population.

|  |  |  |  |
| --- | --- | --- | --- |
| **Ancestry** | **# Loci Before Filtering by GWAS-gnomAD Overlap** | **# Loci After Filtering** | **# Loci Removed** |
| African/African American | 924,946 | 886,467 | 38,479 |
| Latino/American Admixed | 519,496 | 506,760 | 12,736 |
| Ashkenazi Jewish | 161,271 | 160,358 | 913 |
| East Asian | 300,323 | 295,801 | 4,522 |
| Finnish | 207,162 | 204,914 | 2,248 |
| Non-Finnish European | 1,104,008 | 1,051,015 | 52,993 |
| South Asian | 330,518 | 325,444 | 5,074 |
| Other | 267,812 | 264,682 | 3,130 |

**S Table 3. Results of filtering loci by keeping only overlapping loci between the gnomAD dataset and the GWAS dataset.** The gnomAD dataset contains SNP information from a consortium of several ongoing studies, while the GWAS data was gathered from studies using UK Biobank samples, so there was not complete overlap between the loci sampled. Therefore, any loci which did not occur in both the GWAS and gnomAD datasets were excluded.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ancestry** | **P-value Threshold** | **# Loci Before Filtering by NA FST Values** | **# Loci After Filtering** | **# Loci Removed** |
| Ashkenazi Jewish | 1e-3 | 160,358 | 2,749 | 157,609 |
| Ashkenazi Jewish | 1e-5 | 160,358 | 578 | 159,780 |
| Ashkenazi Jewish | 1e-8 | 160,358 | 298 | 160,060 |
| Finnish | 1e-3 | 204,914 | 3,975 | 200,939 |
| Finnish | 1e-5 | 204,914 | 852 | 204,062 |
| Finnish | 1e-8 | 204,914 | 373 | 204,541 |

**S Table 2. Representation of results of filtering by p-value threshold.** We established three thresholds of p-values to further restrict which loci were used for analysis by focusing on sites with strong correlation with the trait of interest. This was performed for all ancestries on all traits. A representative snapshot showing the number of SNPs removed at each threshold is shown here.

**References**

1. Karczewski, K.J., Francioli, L.C., Tiao, G. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581,** 434–443 (2020). <https://doi.org/10.1038/s41586-020-2308-7>
2. Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools, Bioinformatics, Volume 27, Issue 15, 1 August 2011, Pages 2156–2158, <https://doi.org/10.1093/bioinformatics/btr330>
3. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
4. Cheng, C., &amp; Kirkpatrick, M. (n.d.). Sex-specific selection and sex-biased gene expression in humans and flies. PLOS Genetics. Retrieved March 18, 2022, from <https://doi.org/10.1371/journal.pgen.1006170>
5. Wright S. The genetical structure of populations. Ann Eugen. 1951 Mar;15(4):323-54. doi: 10.1111/j.1469-1809.1949.tb02451.x. PMID: 24540312.
6. Tomaz Berisa, Joseph K. Pickrell, Approximately independent linkage disequilibrium blocks in human populations, Bioinformatics, Volume 32, Issue 2, 15 January 2016, Pages 283–285, <https://doi.org/10.1093/bioinformatics/btv546>