Genetic analysis of quantitative traits in medaka fish and humans

Ian Brettell

2022-06-29

Contents

About

Code to render PDF:

```
bookdown::render_book("book", bookdown::pdf_book())
```

Code to render PDF with different fonts:

First make the bookdown::pdf_book: entry in _output.yml first.

Then...

```
bookdown::render_book("book")
```

To clean up the references:

```
citr::tidy_bib_file(rmd_file = c("book/index.Rmd",
                                 "book/01-Introduction.Rmd",
                                 "book/02-MIKK_genome.Rmd",
                                 "book/03-Pilot.Rmd",
                                 "book/04-MIKK_F2.Rmd",
                                 "book/06-Fst.Rmd",
                                 "book/07-references.Rmd"),
                    messy_bibliography = "book/book.bib",
                    file = "book/tidy_references.bib")
```

Test to understand which commands are creating errors:

```
bookdown::render_book("book",
                      bookdown::pdf_book(pandoc_args = "--listings",
                                         latex_engine = "xelatex",
                                         biblio_style = "apalike",
                                         csl = "chicago-fullnote-bibliography.csl",
```

```
                                includes = rmarkdown::includes(in_h
                                documentclass = "book",
                                sansfont = "Georgia",
                                monofont = "Andale Mono",
                                highlight = "pygments"))
```

citation_package = "natbib" throws an ugly error, and ends up removing all references in the document.

Weirdly, fonts only change if they go in index.Rmd.

0.1 `bs4_book`

```
bookdown::render_book("book",
                      bookdown::bs4_book(css = "style.css",
                                         theme = list("primary" = "#52002E")
                                         repo = list("base" = "https://github

                                         includes = rmarkdown::includes(in_h
                                         footnotes_inline = T,
                                         bibliography = "book.bib"))
```

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports; for example, a math equation $a^2 + b^2 = c^2$.

0.2 Usage

Each **bookdown** chapter is an .Rmd file, and each .Rmd file can contain one (and only one) chapter. A chapter *must* start with a first-level heading: # A good chapter, and can contain one (and only one) first-level heading.

Use second-level and higher headings within chapters like: ## A short section or ### An even shorter section.

The index.Rmd file is required, and is also your first book chapter. It will be the homepage when you render the book.

0.3 Render book

You can render the HTML version of this example book without changing anything:

1. Find the **Build** pane in the RStudio IDE, and
2. Click on **Build Book**, then select your output format, or select "All formats" if you'd like to use multiple formats from the same book source files.

Or build the book from the R console:

```
bookdown::render_book()
```

To render this example to PDF as a bookdown::pdf_book, you'll need to install XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.org/tinytex/.

0.4 Preview book

As you work, you may start a local server to live preview this HTML book. This preview will update as you edit the book when you save individual .Rmd files. You can start the server in a work session by using the RStudio add-in "Preview book", or from the R console:

```
bookdown::serve_book()
```

Chapter 1

Introduction

1.1 A brief history of genetics

Humankind has long sought to understand the basis of biological variation. What gives rise to the wondrous variety of life forms on Earth? Why do individuals of a particular species differ from one another? How do children inherit traits that are similar to those of their parents, yet on the whole remain distinct from both their parents and their siblings? And are the traits we care about – our health, our intelligence, our ability to thrive in a changing world – pre-determined from birth, or continuously pliable throughout our lives?

1.1.1 Ancient Greece

Around 500 BC, the Ancient Grecian Pythagoras applied his understanding of triangles to this question, proposing the theory known as "spermism". He posited that hereditary information was passed down from parent to child via male sperm, with the female only providing the nutrients that would allow it to grow, and, like the theorem that bears his name, that these two sides of the "triangle" the length of the third side: the characteristics of the child (Mukherjee 2016).

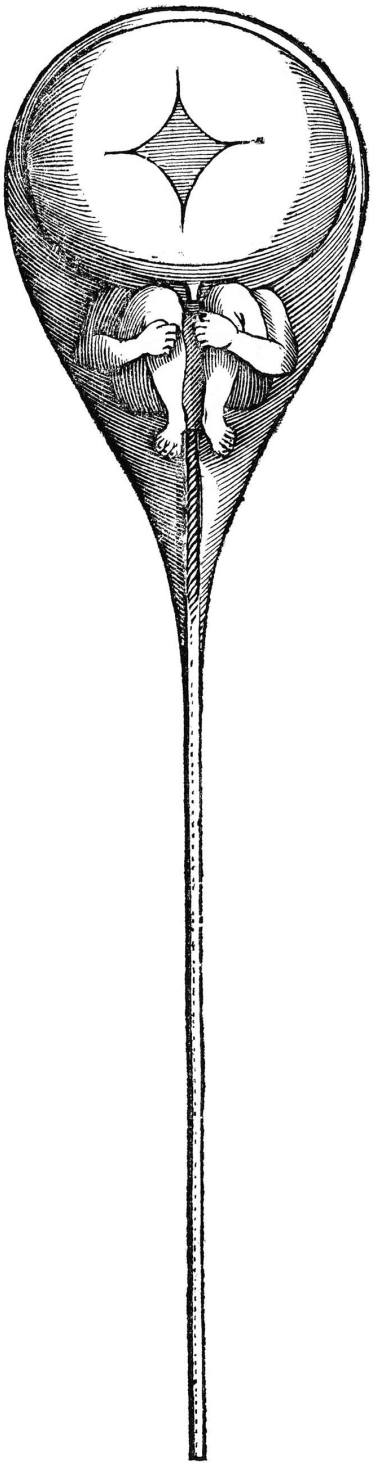
Over a century later, in 380 BC, Plato extended this metaphor in

The Republic to argue that this principle could be applied to perfect humanity, by breeding perfect combinations of parents.

Aristotle joined the discussion with his treatise *Generation of Animals*, where he noted cases where human skin colour and other traits could skip generations, and thus hereditary information must not only be transmitted through sperm. He suggested an idea of "movement" – the transmission of information – from the father's sperm, which sculpts the mother's menstrual blood in the same way a carpenter carves a piece of wood.(Mukherjee 2016)

1.1.2 Medieval times

In medieval times, the prevailing theory was that a tiny human – a homunculus – sat within the sperm, waiting to be inflated upon its introduction to a woman's uterus. However, this would require a homunculus to sit within another homunculus, *ad infinitum*, like Matryoshka dolls, all the way back to the Biblical first man, Adam. Even the inventor of the microscope, Nicolaas Hartsoeker, thought he saw one in a sperm he was studying.



1.1.3 Charles Darwin and Gregor Mendel

In 1831, a young Charles Darwin boarded the HMS *Beagle* to embark on an expedition to collect specimens from South America. After collecting a huge number of fossils from the along the eastern coast and shipping them back to England, the *Beagle* spent 5 weeks touring through the 18 volcanic islands of the Galàpagos, where Darwin collected

1.1.4 Breeding programs in agriculture

Chapter 2

Genomic variations in the MIKK panel

This project was carried out in collaboration with Felix Loosli's group at the Karlsruhe Institute of Technology (KIT), and Joachim Wittbrodt's group in the Centre for Organismal Studies (COS) at the University of Heidelberg.

This chapter sets out my contributions to the the following pair of papers published in the journal *Genome Biology*, on both of which I am joint-first author:

- Fitzgerald et al. (2022)
- Leger et al. (2022)

2.1 The Medaka Inbred Kiyosu-Karlruhe (MIKK) panel

Biological traits are the product of an interaction between an organism's genes and its environment, often described as the relationship

between "nature and nurture".(Plomin and Asbury 2005) This is especially true for complex traits such as behaviour, which I investigate in Chapters 3 and 4.

It is unfeasible to explore the relationship between genes and environment experimentally in humans due to the insufficient ability to manipulate either set of variables. Researchers accordingly resort to using model organisms, with which it is possible to control for both. The genetics of model organisms may be controlled to a degree by establishing inbred strains through the repeated mating of siblings over successive generations. Eventually, as the individuals within each line inherit the same same haplotype from their related parents, they become almost genetically identical to one another, with the added benefit that their genotypes can be replicated across time in subsequent generations. This utility has led to the establishment of "panels" of inbred strains for several model organisms including the thale cress (*Arabidopsis thaliana*),(Bergelson and Roux 2010) common bean (*Phaseolus vulgaris L*),(Johnson and Gepts 1999) tomato (*Lycopersicon esculentum*),(Saliba-Colombani et al. 2000) maize (*Zea mays*),(Limami et al. 2002) nematode (*Caenorhabditis elegans*),(Evans et al. 2021) fruit fly (*Drosophila melanogaster*) (Mackay and Huang 2018), and mouse (*Mus musculus*) (Saul et al. 2019).

Although the mouse is an appropriate model for humans due to their orthologous mammalian organ systems and cell types, inbred strains of this organism descend from individuals that had already been domesticated, and therefore do not represent the genetic variation present in wild populations. Furthermore, the large panels of inbred mice such as the Collaborative Cross (CC),(Threadgill et al. 2011) Diversity Outcross (DO)(Svenson et al. 2012) and B6-by-D2 (BXD)(Peirce et al. 2004) are derived from only a small number of individuals. As gene-environment studies seek to ultimately understand their effects on traits "in the wild" (such as with humans), there is accordingly a need for a panel of inbred vertebrates that represents the genetic variation present in natural populations.

The medaka fish (*Oryzias latipes*) has been studied as a model organism in Japan for over a century,(Wittbrodt, Shima, and Schartl 2002) and is gaining recognition elsewhere as a powerful genetic model for

vertebrates.(Spivakov et al. 2014) In addition to possessing a number of desirable traits that are characteristic of model organisms (including their small-size, short reproduction time, and high fertility), medaka are also – uniquely among vertebrates – resilient to inbreeding from the wild.

Since 2010, the Birney Group at EMBL-EBI, in collaboration with the Wittbrodt Group at COS, University of Heidelberg and the Loosli Group at the Karlsruhe Institute of Technology (KIT), have been working to establish the world's first panel of vertebrate inbred strains – now known as the Medaka Inbred Kiyosu-Karlsruhe Panel (**MIKK panel**). The MIKK Panel was bred from a wild population caught near Kiyosu in Southern Japan, and now comprises 80 inbred, near-isogenic "lines".(Fitzgerald et al. 2022)

The MIKK Panel was created to map genetic variants associated with quantitative traits at a high resolution, and to explore the interactions between those variants and any environmental variables of interest. The purpose of the companion papers Fitzgerald et al. (2022) and Leger et al. (2022) was to introduce the MIKK panel to the scientific community, and describe the genetic characteristics of the MIKK panel that would make it a useful resource for other researchers who wish to explore the genetics of quantitative traits in vertebrates. My contributions to these papers involved visualising the inbreeding trajectory of the panel (Chapter 2.2.2), exploring the evolutionary history of the MIKK panel's founding population (Chapter 2.2.3), measuring the levels of homozygosity across the panel (Chapter 2.2.4), assessing its allele-frequency distribution and rate of linkage disequilibrium (LD) decay (Chapter 2.2.5), and characterising the structural variants present in a smaller sample of lines using Oxford Nanopore long-read sequencing data (Chapter 2.3).

2.2 Genomic characterisation of the MIKK panel

2.2.1 MIKK panel DNA sequence dataset

For the preparation of Fitzgerald et al. (2022), 79 of the 80 extant MIKK panel lines – together with several wild Kiyosu samples and individuals from the established *iCab* medaka strain – had their DNA sequenced from brain samples using Illumina short-read sequencing technology. Tomas Fitzgerald from the Birney Group at EMBL-EBI then aligned these sequences to the *HdrR* medaka reference and called variants to produce the **MIKK Illumina call set** in the form of a .vcf file containing single nucleotide polymorphism (SNP) and small insertion-deletion (INDEL) calls for each line. To avoid allele frequency biases introduced by the 16 pairs/triplets of "sibling lines" (see 2.2.2), I removed each pair's arbitrarily-labelled second sibling line from the variant call set, leaving 63 MIKK panel lines (**MIKK non-sibling call set**), and used only those calls for the analyses in Chapters 2.2.4 and 2.2.5.

For the preparation of Leger et al. (2022), 12 MIKK panel lines had their DNA sequenced from brain samples using Oxford Nanopore Technologies (ONT) long-read sequencing technology. Adrien Leger from the Birney Group at EMBL-EBI then aligned these sequences to the *HdrR* medaka reference, and called variants to produce the **MIKK ONT call set** in the form of a .vcf file containing structural variants calls for each line with tags for insertions (INS), deletions (DEL), duplications (SUP), inversions (INV) and translocations (TRA). The work described below used these variant call sets as the primary datasets.

2.2.2 Assessing the inbreeding trajectory of the MIKK panel

The MIKK panel was bred from a wild population of medaka found in the Kiyosu area near Toyohashi, Aichi Prefecture, in southern

Japan.(Spivakov et al. 2014) From this wild population, the Loosli Group at KIT set up random crosses of single mating pairs to create 115 'founder families'. For each founder family, they then set up between two and five single full-sibling-pair inbreeding crosses, which resulted in 253 F1 lines. Lines derived from the same founder family are referred to as 'sibling lines'. Over the course of the next eight generations of inbreeding, they used only one mating pair per line. I generated **Fig. 2.1A** and **B** from the inbreeding data provided by the Loosli Group. **Fig. 2.1A** shows the number of lines that survived over the course of the first 14 generations of the inbreeding program, and the various causes for the termination of other lines. **Fig. 2.1B** shows the average fecundity levels of the surviving lines at generation F16. In addition, the Birney Group at EMBL-EBI generated morphometric data for the MIKK panel lines to demonstrate the distribution of physical phenotypes across the MIKK panel. I used this data on relative eye diameters to generate **Fig. 2.1C**.

2.2.3 Introgression with northern Japanese and Korean medaka populations

To explore the evolutionary history of the MIKK panel's founding population, we sought to determine whether there was evidence of introgression between that southern Japanese population, and northern Japanese and Korean medaka populations. To this end, I used the 50-fish multiple alignment from Ensembl release 102 to obtain the aligned genome sequences for the established medaka inbred lines *HdrR* (southern Japan), *HNI* (northern Japan), and *HSOK* (Korea), as well as the most recent common ancestor of all three strains.("Index of /Pub/Release-102/Emf/Ensembl-Compara/Multiple_alignments/50_fish.epo/" n.d.) Using the phylogenetic tree provided with the dataset, and the *ape* R package,(Paradis and Schliep 2019) I identified the most recent common ancestor of those three strains. For each locus with a non-missing base for *HdrR*, I assigned the allele in that ancestral sequence as the 'ancestral' allele, and the alternative allele as the 'derived' allele, and

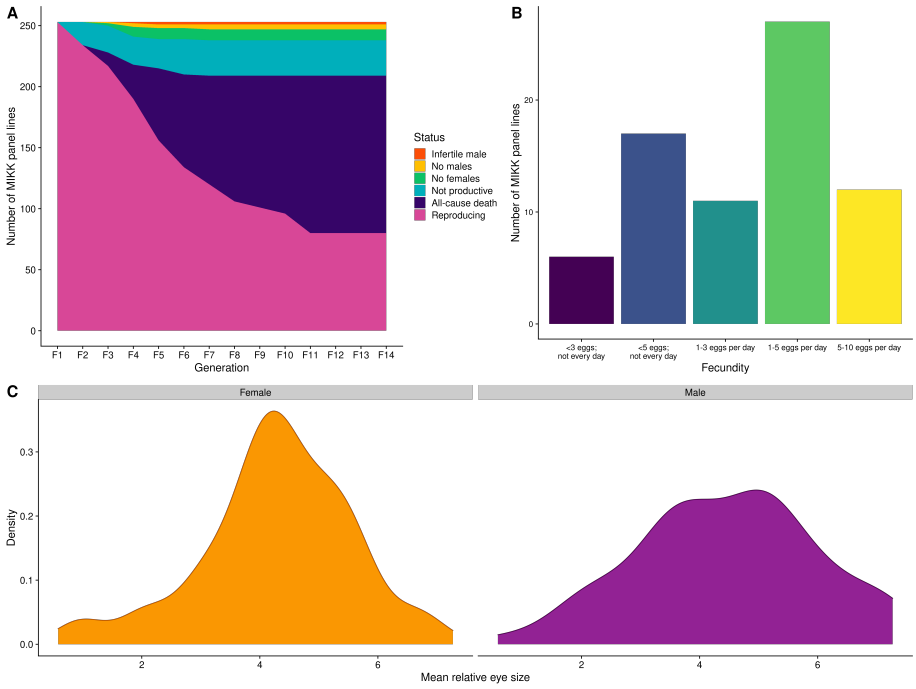


Figure 2.1: Inbreeding, fecundity and eye size in the MIKK panel lines. **A**: Status of all MIKK panel lines during the first 14 generations of inbreeding, showing cause of death for non-extant lines. **B**: Average fecundity of MIKK panel lines in generation F16, as measured during peak egg production in July 2020. **C**: Distribution of mean relative eye size for female and male medaka across all MIKK panel lines.

then combined that dataset with the MIKK Illumina call set and variant calls for the southern Japanese *iCab* strain (see 2.2.1).

I then carried out an ABBA BABA analysis to calculate a modified 'admixture proportion' statistic $\hat{f}_d$(S. H. Martin, Davey, and Jiggins 2015) as a measure of the proportion of shared genome in 500-kb sliding windows between the MIKK panel and either *iCab*, *HNI*, or *HSOK* (**Fig. 2.2**), using the scripts provided by the first author of S. H. Martin, Davey, and Jiggins (2015) on their GitHub page.(martin [2016] 2022)

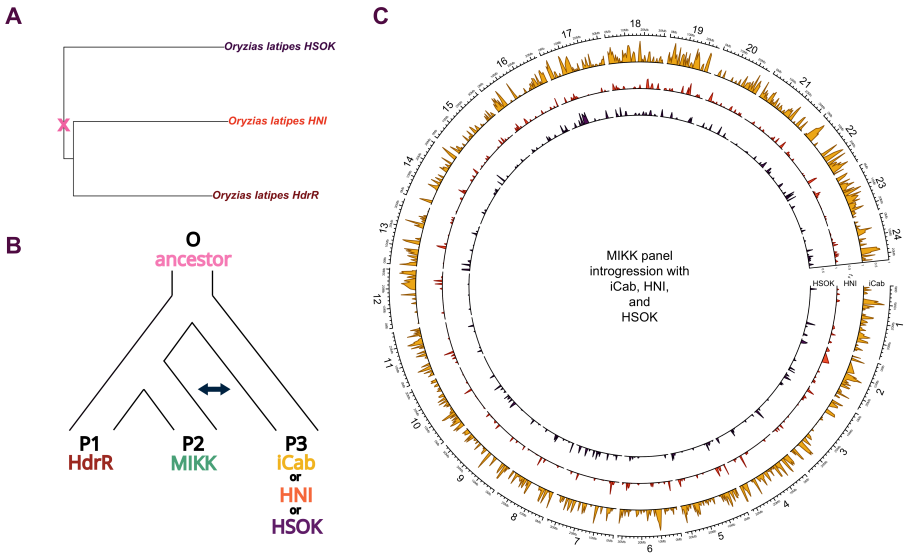


Figure 2.2: **Figure 2**: ABBA-BABA analysis. **A**. Phylogenetic tree generated from the Ensembl release 102 50-fish multiple alignment, showing only the medaka lines used in the ABBA-BABA analysis. **B**. Schema of the comparisons carried out in the ABBA-BABA analysis. **C**. Circos plot comparing introgression ($\hat{f}_d$) between the MIKK panel and either *iCab* (yellow), *HNI* (orange), or *HSOK* (purple), calculated within 500-kb sliding windows using a minimum of 250 SNPs per window.

Based on the genome-wide mean $\hat{f}_d$, the MIKK panel shares approximately 25% of its genome with *iCab*, 9% with *HNI*, and 12% with *HSOK*. These results provide evidence that the MIKK panel's originating population has more recently introgressed with medaka from Korea

than with medaka from northern Japan. This supports the findings in Spivakov et al. (2014), where the authors found little evidence of significant interbreeding between southern and northern Japanese medaka since the populations diverged. Although the proportional difference between *HNI* and *HSOK* is small, this further supports the general finding that northern and southern Japanese medaka strains show low levels of interbreeding that may be a result of geographical isolation or genome divergence.(Katsumura et al. 2019)

2.2.4 Nucleotide diversity

As a means of assessing genetic diversity in the MIKK panel, I calculated nucleotide diversity ($\hat{\pi}$) within 500-kb non-overlapping windows across the genome of the 63 lines in the MIKK non-sibling call set (see 2.2.1), and compared this to the nucleotide diversity in 7 wild medaka from the same Kiyosu population from which the MIKK panel was derived. Mean and median nucleotide diversity in both the MIKK panel and wild Kiyosu medaka were close to 0, and slightly higher in the MIKK panel (mean: MIKK = 0.0038, wild = 0.0037; median: MIKK = 0.0033, wild = 0.0031). The patterns of varying nucleotide diversity across the genome are shared between the MIKK panel and wild Kiyosu medaka, where regions with high levels of repeat content tend to have higher nucleotide diversity ($r = 0.386$, $p < 0.001$) (**Fig. 2.3**). I also calculated $\hat{\pi}$ for each line individually, and as expected, levels of $\hat{\pi}$ around the (XX/XY) sex determination region of 1:~16-17 Mb are elevated in all lines relative to the consistently low levels found in most other chromosomes.

The higher level of $\hat{\pi}$ observed within specific regions on several chromosomes – such as chromosomes 2, 11, and 18 – correspond closely to the regions we identified as containing large (>250 kb) inversions that appear to be shared across at least some of the MIKK panel (**Fig. 2.4**). These regions are also enriched for large deletions and duplications.(Leger et al. 2022) Inversions cause permanent heterozygosity (Hoffmann, Sgr'o, and Weeks 2004), and duplications and deletions may have increased the density of called SNPs in these regions (Fredman et al. 2004), so the observed depressions in ho-

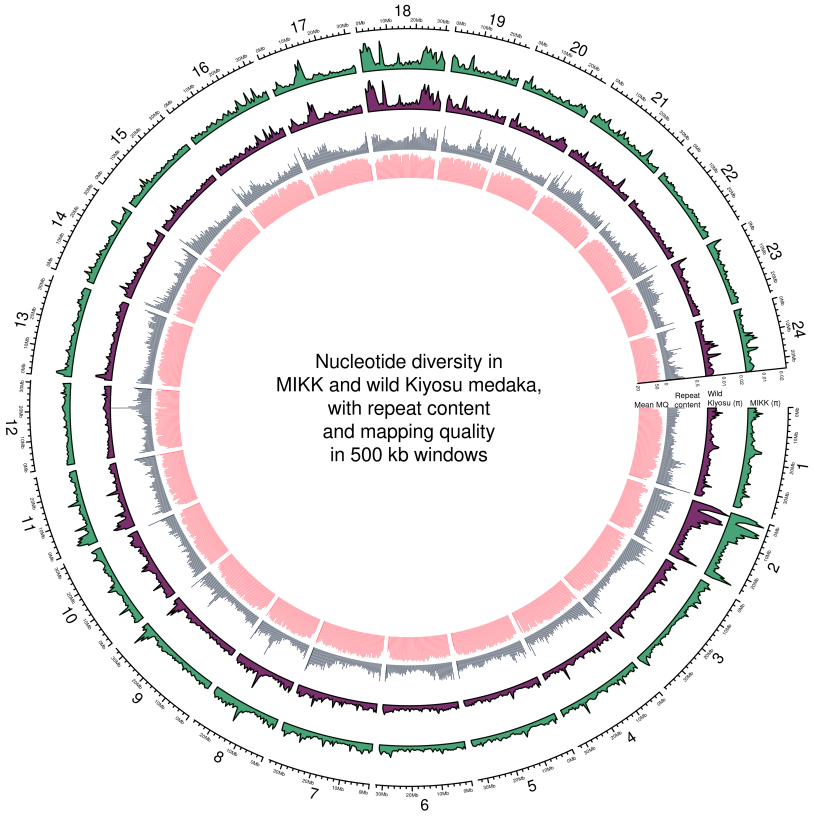


Figure 2.3: Circos plot with nucleotide diversity ($\hat{\pi}$) calculated within 500-kb non-overlapping windows for 63 non-sibling lines from the MIKK panel (*green*) and 7 wild Kiyosu medaka samples from the same originating population (*purple*); proportion of sequence classified as repeats by RepeatMasker (*blue*); and mean mapping quality (*pink*).

mozygosity at these loci may be the result of such large structural variants that are present in the MIKK panel's genomes.

Overall, this analysis confirms that the MIKK panel shows similar levels of homozygosity compared to classical laboratory inbred medaka strains, and possesses a strong increase in isogenic genotypes compared to wild medaka from the original wild population.

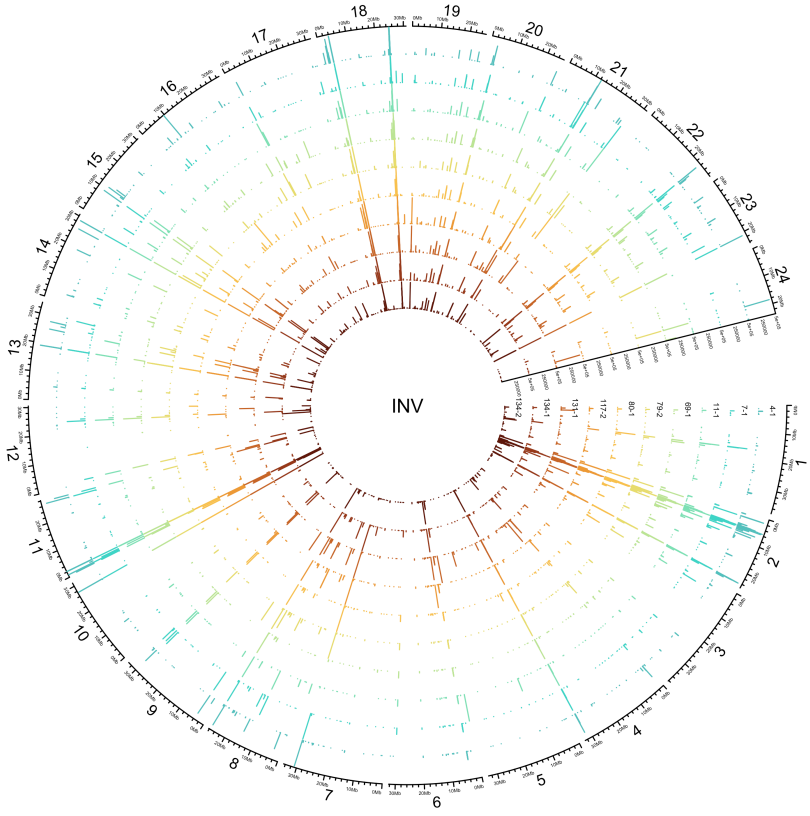


Figure 2.4: Inversions identified in 9 MIKK panel lines using a combination of Oxford Nanopore Technologies long-read and Illumina short-read sequences (see Chapter 2.3 below).

2.2.5 LD decay

I analysed the MIKK panel's allele frequency distribution and linkage disequilibrium (LD) structure to assess their likely effects on genetic mapping. To remove allele-frequency biases introduced by the presence of sibling lines in the MIKK panel, I used only the MIKK non-sibling call set (see Chapter 2.2.1).

To assess how accurately one may be able to map genetic variants using the MIKK panel relative to a human dataset, I compared the MIKK panel's minor allele frequency (MAF) distribution and LD structure against that of the 2,504 humans in the 1KG Phase 3 release.("A Global Reference for Human Genetic Variation" 2015) To prepare the **"1KG call set"**, I first downloaded the .vcf files for each autosome from the project's FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/), then merged them into a single VCF using GATK.(McKenna et al. 2010) I then used PLINK(Chang et al. 2015; Purcell and Chang, n.d.) to calculate the minor allele frequencies for all non-missing, biallelic SNPs in both the MIKK non-sibling and IKG call sets (*N* SNPs = 16,395,558 and 81,042,381 respectively) (**Fig. 2.5A**). As expected, the 1KG and MIKK panel calls are similarly enriched for low-frequency variants, albeit to a lesser extent in the MIKK panel, which is likely due to its smaller sample size.

To determine the rate of LD decay in the MIKK panel and compare it to that in the 1KG sample, for both the MIKK non-sibling and 1KG call sets, I used PLINK to compute r^2 on each autosome for all pairs of non-missing, biallelic SNPs with MAF $>$ 0.10 within 10 kb of one another (for 1KG and the MIKK panel respectively ~ 5.5M and ~ 3M SNPs, with a total number of pairwise r^2 observations of 204,152,922 and 146,785,673). I then grouped the r^2 observations for each pair of SNPs based on their distance from one another into non-overlapping bins of 100 bp in length, and calculated the mean r^2 in each of those bins to generate **Fig. 2.5B** using the mean r^2 and left boundary of each bin.

Based on the 1KG calls under these parameters, LD decays in humans to a mean r^2 of around 0.2-0.35 at a distance of 10 kb, whereas the

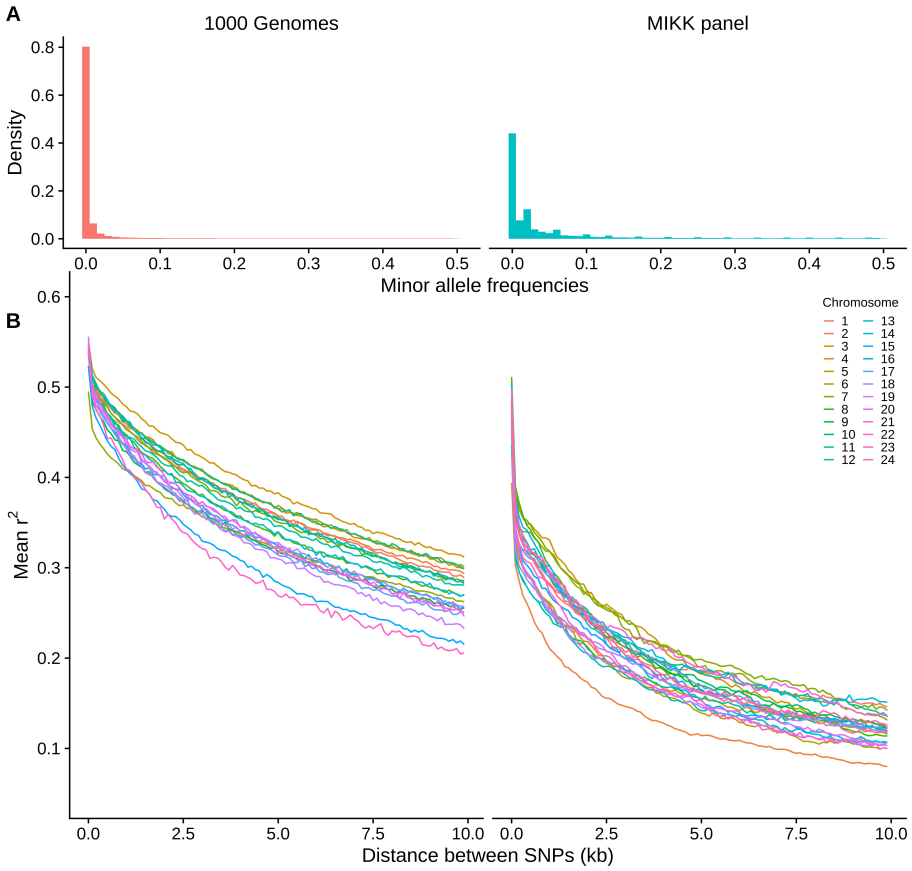


Figure 2.5: Minor allele frequency distributions and LD decay for biallelic, non-missing SNPs in the 1000 Genomes Phase 3 variant calls (N = 2,504) (1KG), and the MIKK panel Illumina-based calls excluding one of each pair of sibling lines (N = 63), across all autosomes (1KG: chrs 1-22; MIKK: chrs 1-24). **A**: Histogram of allele frequencies in the 1KG and MIKK panel calls. **B**: LD decay for each autosome, calculated by taking the mean r^2 of pairs of SNPs with MAF > 0.1 within non-overlapping 100 bp windows of distance from one another, up to a maximum of 10 kb. LD decays faster on chromosome 2 for the MIKK panel due to its higher recombination rate.

MIKK panel reaches this level within 1 kb, with a mean r^2 of 0.3-0.4 at a distance of ~100 bp. This implies that when a causal variant is present in at least two lines in the MIKK panel, one may be able to map causal variants at a higher resolution than in humans. We note that LD decays faster in chromosome 2 of the MIKK panel relative to the other chromosomes. This suggests that it has a much higher recombination rate, which is consistent with the linkage map described in Naruse et al. (2000), showing a higher genetic distance per Mb for this chromosome. This higher recombination rate in chromosome 2 may in turn be caused by its relatively high proportion of repeat content (**Fig. 2.6**).

2.3 Structural variation in the MIKK panel

As an alternative to the variation pangenome approach described in Leger et al. (2022), I explored the structural variants (SVs) present in 9 of the MIKK panel lines in a reference-anchored manner, similar to many human studies. Differences in SVs between panel lines is another important class of genetic variation that could cause or contribute to significant phenotypic differences. Here we used ONT data obtained for 9 of the 12 selected lines allowing us to characterise larger SVs in the MIKK panel and to create a more extensive picture of genomic rearrangements compared to available medaka reference genomes. Adrien Leger from the Birney Group at EMBL-EBI first called structural variants using only the ONT long reads, producing a set of structural variants classified into five types: deletions (DEL), insertions (INS), translocations (TRA), duplications (DUP) and inversions (INV). I then "polished" the called DEL and INS variants with Illumina short reads to improve their accuracy. The polishing process filtered out 7.4% of DEL and 12.8% of INS variants, and adjusted the breakpoints (i.e. start and end positions) for 75-77% of DEL and INS variants in each sample by a mean of 23 bp for the start position, and 33 bp for the end position. This process produced a total of 143,326 filtered SVs.

The 9 "polished" samples contained a mean per-sample count of ap-

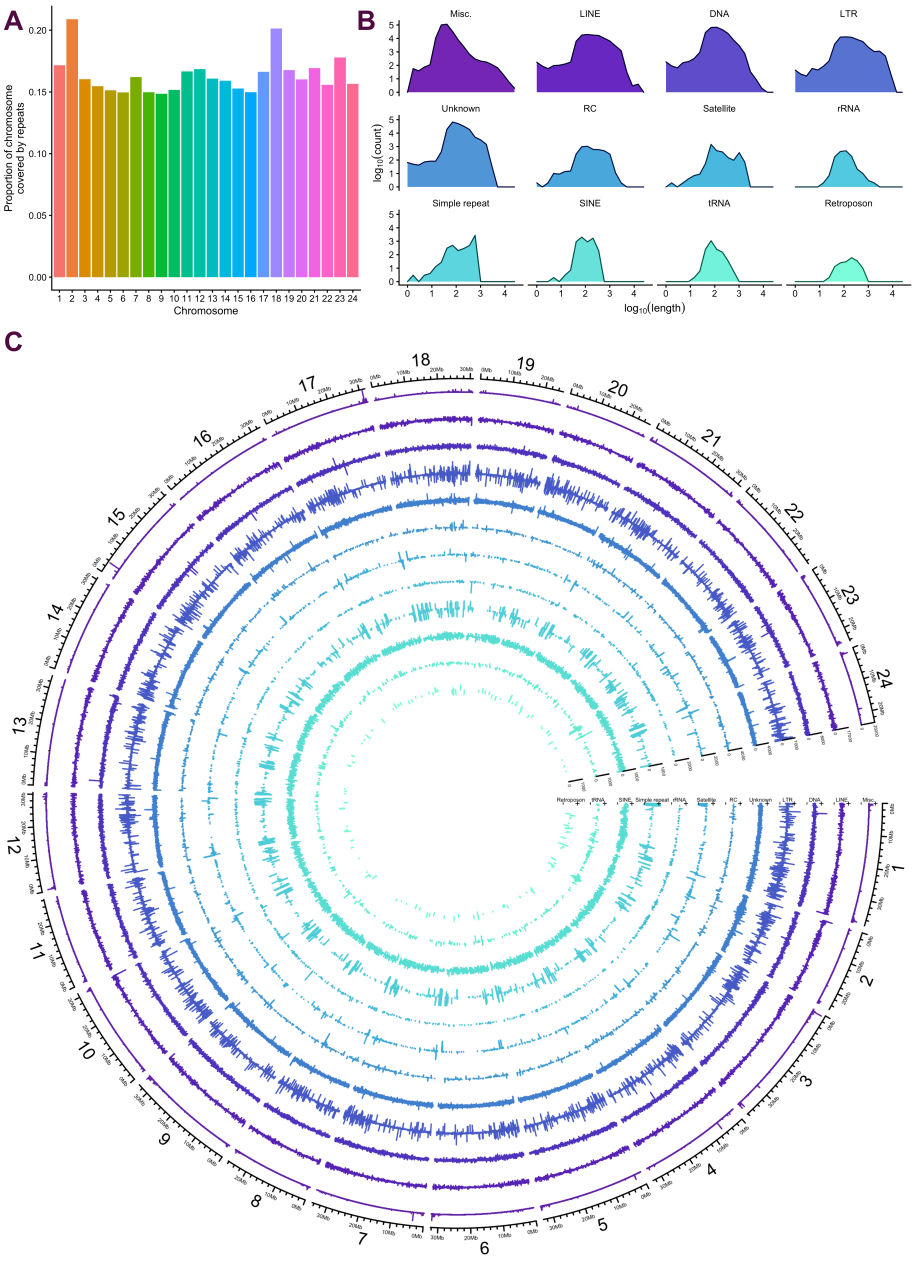


Figure 2.6: Repeat content in the *HdrR* genome based on RepeatMasker results obtained by Jack Monahan. **A**. Proportion of repeat content per-chromosome. **B**. log10 of repeat lengths and counts per repeat class. "Misc" includes all repeats assigned to their own specific class, for example "(GAG)n" or "(GATCCA)n". **C**. Circos plot showing repeat length (radial axes) by locus (angular axis) and repeat class (track).

proximately 37K DEL variants (12% singletons), 29.5K INS variants (14%), 3.5K TRA variants (9%), 2.5K DUP (7%) and 600 INV (7%) (**Fig. 2.7D**). DEL variants were up to 494 kb in length, with 90% of unique DEL variants shorter than 3.8 kb. INS variants were only up to 13.8 kb in length, with 90% of unique INS variants shorter than 2 kb. DUP and INV variants tended to be longer, with a mean length of 19 and 70.5 kb respectively (**Fig. 2.7A**). **Fig. 2.7E** shows the per-sample distribution of DEL variants across the genome. Most large DEL variants over 250 kb in length were common among the MIKK panel lines. A number of large DEL variants appear to have accumulated within the 0-10 Mb region of chromosome 2, which is enriched for repeats in the HdrR reference genome (**Fig. 2.6**).

SVs were generally enriched in regions covered by repeats. While only 16% of bases in the *HdrR* reference were classified as repeats (irrespective of strand), those bases overlapped with 72% of DEL, 63% of DUP, 81% of INV and 35% of TRA variant regions. However, repeat bases only overlapped with 21% of INS variants. We also assessed each SV's probability of being loss-of-function (pLI)(Lek et al. 2016) by calculating the logarithm of odds (LOD) for the pLI scores of all genes overlapping the variant (**Fig. 2.7B,C**). 30,357 out of 134,088 DEL, INS, DUP and INV variants overlapped at least one gene, and 9% of those had a score greater than 10, indicating a high probability that the SV would cause a loss of function. Two INS variants on chr2 had an outlying LOD score of 57 as a result of overlapping medaka gene ENSORLG00000003411, which has a pLI score of 1 – the highest intolerance to variants causing a loss of function. This gene is homologous with human genes *SCN1A*, *SCN2A* and *SCN3A*, which encode sodium channels and have been associated with neuronal and sleep disorders. We did not find evidence that longer SVs tended to have a higher probability of causing a loss of function (**Fig. 2.7B**).

We compared these polished INS and DEL calls with the high-quality graph-based alternative paths and large-scale deletions, respectively (see section titled *Novel genetic sequences and large-scale insertions and deletions in the MIKK panel* in Leger et al. (2022)). We found that 2 of the 19 regions covered by graph-based alternative paths, and 4 of the 16 regions covered by graph-based deletions, had no SVs that

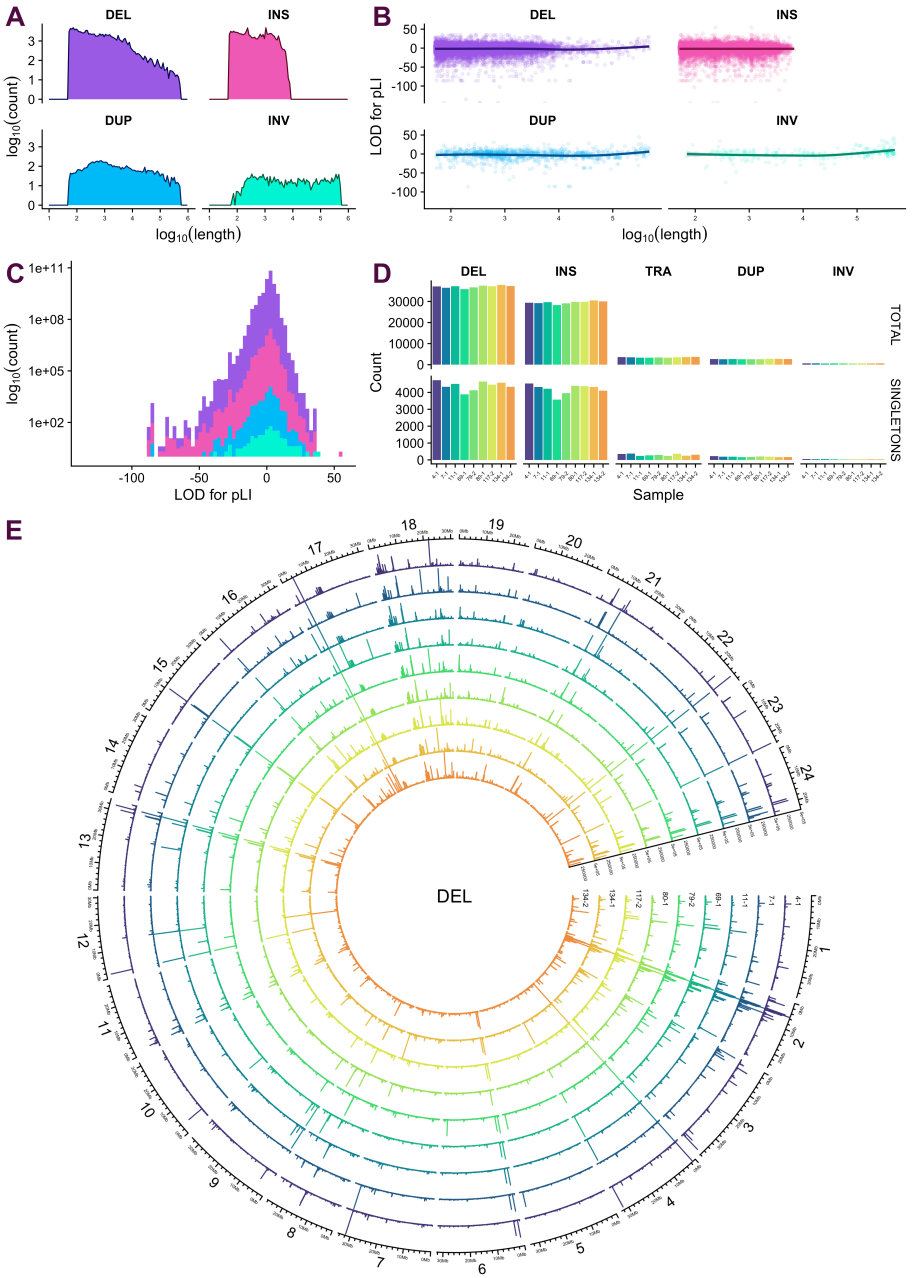


Figure 2.7: Polished SVs in 9 MIKK panel lines sequenced with ONT. DEL: deletion; INS: insertion; TRA: translocation; DUP: duplication; INV: inversion. **A**. Aggregate log10 counts and lengths of distinct SVs by type, excluding TRA. **B**. pLI LOD scores in distinct SVs by SV type. **C**. Histogram of LOD scores by SV type. **D**. Total and singleton counts of SV types per sample. **E**. Circos plot showing per-sample distribution and lengths of DEL variants across the genome.

overlapped those regions at all, which suggests they would have been missed entirely when using a reference-anchored approach alone.

With the exception of one alternative path on chromosome 20, the alternative paths were not captured by INS variants, which only covered up to 63% of the bases in each region, and in many cases substantially less. On the other hand, for 8 of the 16 graph-based deletions, the DEL variants covered at least 85% of the bases in those regions. The other 8 graph-based deletions were either not at all covered by DEL variants, or only slightly. This indicates that the reference-based approach is better at detecting large-scale deletions than alternative paths ("insertions"), but still misses around half of such variants relative to the graph-based approach.

2.4 Conclusions

Taken together, these analyses show that the MIKK panel is highly homozygous, with LD characteristics that will favour high-resolution genetic mapping relative to humans. In the future, the SV analysis performed on a subset of the MIKK panel will be expanded across the entire panel, which will permit the inclusion of both large- and small-scale variants in genetic linkage studies. I proceeded to use the MIKK panel to analyse bold/shy behaviours, as I describe in Chapter 4, with a view to carrying out an F2-cross linkage study to identify genetic variants associated with differences in the behaviours of an individual, and the extent to which they transmit those behaviours to their social companions. However, before carrying out this study, we first ran a "pilot" study on 5 previously-established inbred lines to validate our behavioural assay. This is the subject of the following chapter.

Chapter 3

Classification of bold/shy behaviours in *5* inbred medaka lines

Chapter 4

Bold/shy behaviours in the MIKK panel

Chapter 5

Variation in the frequency of trait-associated alleles across global human populations

5.1 Background

Humans have long sought to use genetic information to predict an individual's likely value for a given trait, in our own species and in other organisms (Chapter 1). As seen in previous chapters, an individual's phenotypic value at a given point in time is the product of complex interactions between their genome and their environment, beginning from embryonic development and continuing throughout their lifetimes. It is now clear that "complex" traits such as height, intelligence, and behaviour are highly polygenic, meaning that they are genetically influenced by hundreds or thousands of genetic variants, each exerting a small effect in one or the other direction along the trait's spectrum (Sella and Barton 2019).

A richer understanding of the cumulative effect of genetic variants

on any trait allows for the prediction of the value that an individual is most likely to have for that trait. Of all human traits, diseases are particularly salient; in 2018, the global healthcare industry was valued at US$8 trillion, and predicted to increase to US$12 trillion by 2022 ("The $11.9 Trillion Global Healthcare Market: Key Opportunities & Strategies (2014-2022) - ResearchAndMarkets.com" 2019). This strong financial imperative complements the moral imperative to reduce suffering, together driving the question of how to use genetic information to improve human health.

Recent technological developments have made it possible to sequence human genomes at scale, and it is thought that by combining detailed genetic information with with other environmental and phenotypic information (such as lifestyle or clinical factors), clinicians could move towards the practice of "precision medicine", where interventions could be tailored to their patients' unique risk profiles (Wray, Goddard, and Visscher 2007). The use of genetic information to predict individuals' values for a trait of interest entails the construction of metrics known as "polygenic scores" (**PGS**). When the trait is a disease, PGS is commonly known as polygenic risk scores (**PRS**), or genetic risk profiling, but I use the term PGS to encompass both disease and non-disease traits.

5.1.1 Polygenic Scores (PGS) and Genome-Wide Association Studies (GWAS)

5.1.1.1 PGS

PGS using genetic information alone show modest yet reliable accuracy for the prediction of complex traits (Alicia R. Martin et al. 2019): the correlations between PGS and the trait value as measured by R^2 have reached 0.24 for height (Yengo et al. 2018), and 0.12-0.16 for educational attainment (Okbay et al. 2022). PGS also improve predictions beyond non-genetic clinical models for a variety of health-related traits, including breast cancer (Maas et al. 2016), prostate cancer (Schumacher et al. 2018), and type I diabetes (Sharp et al. 2019). The predictive accuracy of PGS scores can be further improved by

combining genetic information with lifestyle and clinical factors, as seen with cardiovascular disease (Khera et al. 2018; Kullo et al. 2016; Natarajan et al. 2017; Paquette et al. 2017; Tikkanen et al. 2013; Sun et al. 2021).

5.1.1.2 GWAS

PGS are calculated for an individual by summing trait-associated alleles identified by genome-wide association studies (**GWAS**), as weighted by the alleles' effect sizes (Duncan et al. 2019). GWAS aim to identify genetic variants associated with traits by comparing the allele frequencies of individuals who share similar ancestries, but differ in values for the trait in question (Uffelmann et al. 2021). As of 2021, over 5,700 GWAS have been performed for more than 3,330 traits (Uffelmann et al. 2021).

However, most GWAS have been performed with individuals of European ancestry, despite only constituting around 16% of the present global population. Although the proportion of participants in GWAS from a non-European background increased from 4% in 2009 to 16% in 2016 (Popejoy and Fullerton 2016)), as of 2019, 79% of all GWAS participants recorded in the GWAS Catalog were of European ancestry, and the proportion of non-European individuals has remained the same or reduced since late 2014 (Alicia R. Martin et al. 2019). This bias extends to PGS studies, where as of 2019, only 67% of them included only participants of European ancestry, with another 19% including only East Asian ancestry participants, and only 3.8% with cohorts of African, Hispanic, or Indigenous ancestry (Duncan et al. 2019).

It is therefore unsurprising that PGS scores are far better at predicting disease risk in individuals of European ancestry than in those of non-European ancestry (Alicia R. Martin et al. 2017; Alicia R. Martin et al. 2019). Indeed, the predictive accuracy of PGS scores decays with genetic divergence of the GWAS "independent" or "test" sample, from the "discovery" or "training" sample, as established in both humans (Alicia R. Martin et al. 2017; Alicia R. Martin et al. 2019), and livestock (Clark et al. 2012; Habier et al. 2010; Pszczola et al. 2012).

Compared to PGS scores for those of European ancestry, PGS scores across multiple traits are ~64-78% less accurate for individuals of African ancestry, (Duncan et al. 2019; Alicia R. Martin et al. 2019), ~50% less accurate for individuals of East-Asian ancestry, and ~37% less accurate for individuals of South-Asian ancestry (Alicia R. Martin et al. 2019).

5.1.2 Contributors to PGS non-transferability

What explains this disparity in predictive value? A number of factors may be responsible, including:

1. The failure of GWAS to identify causal variants that either do not exist or are not identifiable within the "discovery" sample, for both technological and methodological reasons (Alicia R. Martin et al. 2019);
2. The sample populations may differ in linkage disequilibrium (**LD**) – the correlation structure of the genome – which would change the estimated effect sizes of the causal variants, even when the causal variants themselves are the same (Alicia R. Martin et al. 2019);
3. Allele frequencies of the causal variants, and the distribution of the effect sizes of the causal variants, may differ between populations (Alicia R. Martin et al. 2017; Scutari, Mackay, and Balding 2016); and
4. The environments and demographies of populations tend to differ. Such differences are often correlated with genetic divergence due to geography, making it difficult to determine whether the associations are driven by the differences between population in their genetics, or their environments (Alicia R. Martin et al. 2019; Kerminen et al. 2019).

The first three factors can degrade predictive performance even in the absence of biological and environmental differences. On the other hand, environmental and demographic differences can drive

forces of natural selection can in turn drive differences in causal genetic architecture (Alicia R. Martin et al. 2019).

I will discuss each of these factors in turn before addressing point (3) in this analysis.

5.1.2.1 Technological and methodological limitations of GWAS

The power to discover a causal variant through GWAS depends on the variant's effect size and frequency in the study population (Alicia R. Martin et al. 2019; Sham et al. 2000). That is to say, the stronger the variant's effect, or the more common it is, the more likely it is to be discovered. Rare variants tend to have stronger effect sizes (Watanabe et al. 2019), likely due to purifying selection (Park et al. 2011), and tend not to be shared across populations (Gravel et al. 2011; Consortium et al. 2015). This is particularly relevant for African populations, as they have a much greater level of genetic variance than other populations due to the human species having originated on that continent (Consortium et al. 2015). Therefore, if GWAS aren't performed on diverse populations, PGS can't take into account the rare variants present in non-European populations that are likely to exert stronger effects on the trait of interest. There are also several other issues that can affect the discoverability of causal variants through GWAS, including the technology used for genotyping, the selection of the cohort, and the necessary exclusion of genotypic outliers.

With respect to genotyping technologies, GWAS often use data from SNP microarrays. These do not sequence the whole genome, but rather a selection (from several hundred thousand to millions) of genetic markers intended to present *common* genetic variation (Porcu et al. 2013), which accordingly tend to neglect rare genetic variants (Uffelmann et al. 2021). To increase the density of genotypes, which would increase the likelihood of refining the association signal and identifying causal variants, researchers often "impute" variants that aren't sequenced directly (Porcu et al. 2013). The imputation process involves "phasing" the study genotypes onto the genotypes of a "reference panel" (McCarthy et al. 2016). However, if the reference panel

does not sufficiently represent the population in the study sample, they are likely to miss or incorrectly impute those genotypes (Alicia R. Martin et al. 2019). Again, this is particularly problematic for African populations.

The lack of representation of rare variants in SNP microarrays can be overcome by using next-generation sequencing technologies such as whole-genome sequencing (**WGS**) and whole-exome sequencing (**WES**). (The former seeks to sequence the full genome, and the latter of only targets the coding regions of the genome.) These methods are more expensive than SNP microarrays, which hinders their widespread use at scale, and although their costs are continuing to decrease rapidly, there is a question as to whether they return a proportionate benefit in all use cases (Schwarze et al. 2018).

A second limitation is the selection of GWAS cohorts, which can introduce selection and collider biases (Uffelmann et al. 2021). For instance, the UK Biobank, which contains genetic and phenotypic data on 500,000 participants who volunteered for inclusion between 2006 and 2010, tend to be older, female, healthier, and wealthier than non-participants (Fry et al. 2017). This creaties the possibility of confounding genetic associations with environmental factors, which I discuss further in 5.1.2.4 below.

A third limitation is the "quality control" step that is required during the GWAS process (Uffelmann et al. 2021). To avoid confounding from population stratification, which can lead to overestimated heritability and biased PGS, GWAS cohorts are filtered to include only those with similar ancestries – or relative genetic homogeneity – by clustering individuals through principal component analysis (**PCA**) of their genotypes, and excluding outliers. I elaborate on the issue of population stratification in section 5.1.2.4 below, but at present, a statistical model for GWAS that can include cohorts with diverse ancestries without the risk of serious confounding is yet to be developed (Jeremy J. Berg et al. 2019).

5.1.2.2 Differences in LD

Because GWAS SNP markers are often not the causal variants themselves, but merely in physical proximity to them, the estimated effect size of a SNP marker depends on the extent to which it is in LD with the causal variant (Mostafavi et al. 2020; Pritchard and Przeworski 2001). To illustrate the problem, if a SNP has an LD r^2 with a causal variant of 0.8 in the discovery population and 0.6 in the target population, it would explain 25% = (1 - 0.6/0.8) less trait variation in the target population, and would therefore be less predictive (Wang et al. 2020).

These differences in effect-size estimates may typically be small for most regions of the genome, but as PGS sum across all such effects, they aggregate these population differences (Alicia R. Martin et al. 2019; Jeremy J. Berg et al. 2019). Previous empirical and simulation studies have shown that accuracy of PGS scores decay with increased genetic differentiation (F_{ST} – described below in 5.1.3) and LD differences between populations (Habier et al. 2010; Pszczola et al. 2012; Scutari, Mackay, and Balding 2016; Wang et al. 2020). The issue may be addressed to a degree by using LD information from an external reference panel as a prior to infer the posterior mean effect size of a genetic variant – Vilhjálmsson et al. (2015) demonstrated through simulations that this could improve PGS predictive accuracy. Yet the most appropriate means of deal with differences between populations in LD remains an active area of research (Duncan et al. 2019).

5.1.2.3 Differences in allele frequencies

Causal variants can differ in both frequency and effect size between different ancestry groups, e.g. for lactase persistence (S'egurel and Bon 2017), or skin pigmentation (Adhikari et al. 2019). If a causal allele is rare in the GWAS discovery population, even if it is discovered (see 5.1.2.1), it is likely to have noisy effect size estimates, and therefore likely to inaccurately estimate its effect size in a different population where it exists at a higher frequency.

Differences in allele frequencies between populations can arise through random genetic drift, or be driven by selective pressures towards the trait optima for a given environment (Harpak and Przeworski 2021). However, evolutionary biologists have found that differences between populations in the mean values for traits tend to occur through small, coordinated shifts in their allele frequencies (Jeremy J. Berg et al. 2019; Edge and Coop 2019). In Chapter 5, I explore the differences in allele frequencies across populations for all polygenic traits in the GWAS Catalog, and confirm that with few exceptions – including skin pigmentation, and HIV viral load – the differences in allele frequencies between populations tends to be small.

5.1.2.4 Differences in environment

Genes continuously interact with each other (GxG, or "epistasis" (Gros, Le Nagard, and Tenaillon 2009)), the genes of one's parents ("genetic nurture", Kong et al. (2018)) or social companions ("social genetic effects") (Domingue et al. 2018; Baud et al. 2017),[1] and the wider non-genetic environment (GxE).

The respective contributions of genetics and environment to traits with social value – such as intelligence – is highly contentious, especially when there are apparent differences between populations in the mean values for those traits. PGS measure the proportion of variance within a population that is explained by genetics. Because PRS summarises a *proportion* of the total variance, when studying a population that is subject to greater environmental variation, the variance attributable to genetic factors will proportionately reduce. The corollary being that when studying a population where the environment is held constant, the proportion of variance for that trait that is explained by genetic factors will approach 1. Therefore, increases in the amount of environmental variance that a population is exposed to will reduce the accuracy of PGS predictions when applied to that population.

[1] As also explored in Chapters 3 and 4

Different environments are also often correlated with population structure (Jeremy J. Berg et al. 2019). For example, in East Asia, there is a greater proportion of individuals of East-Asian ancestry than there is of European ancestry, and *vice versa* in Europe. Those East-Asian individuals will therefore tend to share more of their genetic background with each other than with Europeans, and that population structure will be correlated with the different environments that exist in East Asia compared to Europe. This makes it difficult to determine whether it is the differences in their environments or the differences in their genetics that is driving the discrepancies between the mean values for traits between those populations. These complexities are unlikely to be resolved in the near future, which makes it attractive to turn to model organisms to address more basic biological questions regarding GxE in relation to complex traits (Andersson and Georges 2004), as we have done with respect to behaviour in Chapters 3 and 4.

5.1.3 F_{ST}

The widely-used fixation index (F_{ST}) was introduced independently by Sewall Wright (Wright 1949) and Gustave Malécot (Mal'ecot 1948) as a metric for measuring the genetic diversity between populations.[2] It quantifies the relative variance in allele frequency between groups compared to within groups, reflecting the combined effects of genetic drift, migration, mutation, and selection (Holsinger and Weir 2009). The metric ranges from 0 to 1, where loci with high F_{ST} values – that is, loci with a large relative between-group variance in allele frequencies – may have been subject to selection or different demographic processes (Holsinger and Weir 2009). The metric has customarily been used to identify regions of the genome have been subject to diversifying selection (Akey et al. 2002; Guo, Dey, and Holsinger 2009; Bruce S. Weir et al. 2005).

I first sought to explore the distribution of F_{ST} for SNPs across the human genome. As the reference for human genomic varia-

[2] In Wright's notation, F refers to "fixation" of an allele, and $_{ST}$ refers to "subpopulations within the total population".

tion across diverse populations, I used the New York Genome Center high-coverage, phased .vcf files("Index of /Voll/Ftp/Data_collections/1000g_2504_high_c n.d.) for the 2,504 individuals from 26 populations from across the globe, as described in the 1000 Genomes phase 3 release (Consortium et al. 2015). I then annotated those .vcf files with human SNP IDs from dbSNP release 9606 (Smigielski et al. 2000), and calculated F_{ST} for each of the ~69M SNPs in that dataset with PLINK 1.9 (Chang et al. 2015; Purcell and Chang, n.d.). Figure 5.1 shows the location and F_{ST} value for all SNPs in the 1000 Genomes dataset, and Figure 5.2 shows their distribution across the range of F_{ST} values.

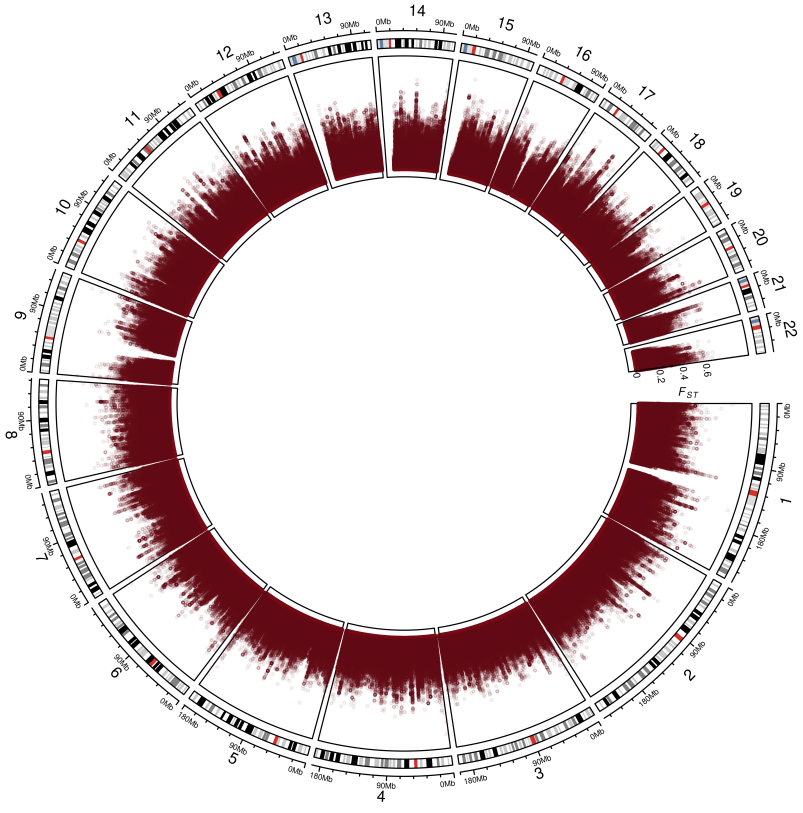


Figure 5.1: (ref:fst-circos)

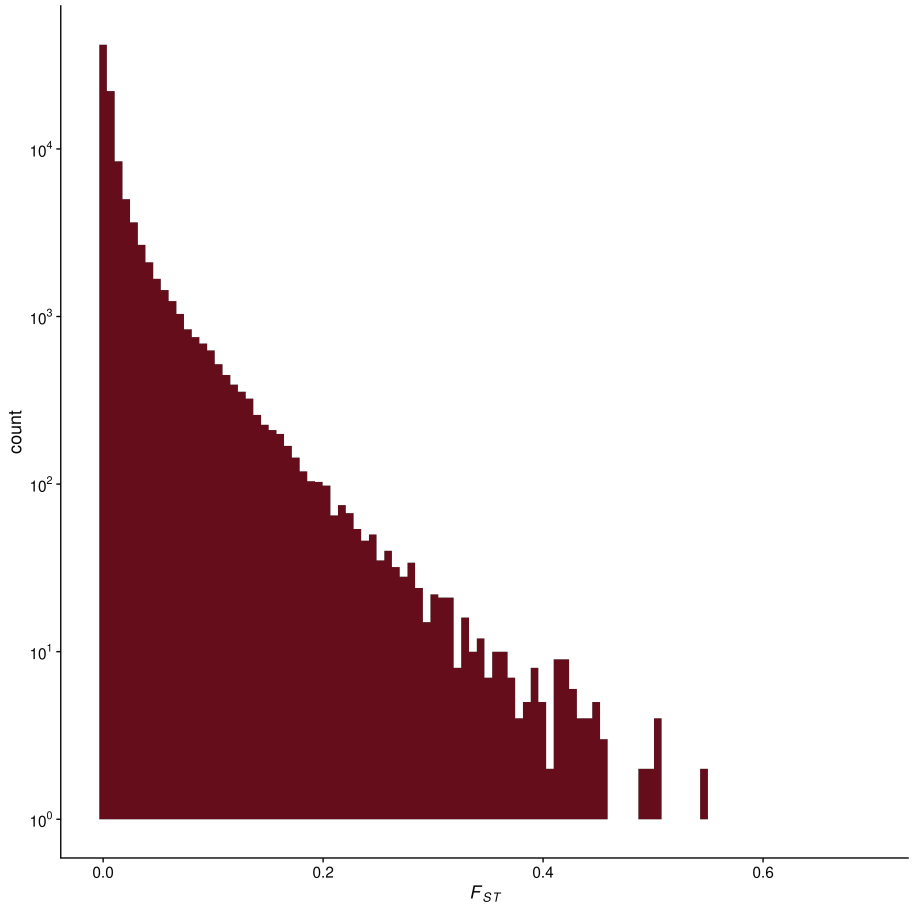


Figure 5.2: (ref:fst-histo)

Figure 5.1 shows that with very few exceptions, F_{ST} remains below 0.6, although specific regions do appear to be differentially selected, as indicated be the peaks on chromosomes 1, 2, and 17. Figure 5.2 makes it clear that the vast majority of SNPs have very low F_{ST}, with a genome-wide mean of 0.019. The question then is whether the SNPs that have a detectable effect on traits tend to have higher F_{ST} values – and therefore greater variation in allele frequencies between populations – than random SNPs with similar allele frequencies in European populations. If so, this would suggest that variation in allele frequencies might contribute significantly to the non-transferability of PRS scores derived from European-focused GWAS.

5.2 Analysis

In this analysis I explore the distribution of F_{ST} scores for loci associated with 587 traits, a subset of the GWAS Catalog that passed our criteria for suitable polygenic traits (see section 5.2). Using high-coverage sequence data for 2,504 individuals from the 1000 Genomes Project phase 3 release, for each trait in the GWAS Catalog I calculated the distribution of F_{ST} across all approximately-unlinked SNPs associated with it (**trait SNPs**), and compared these F_{ST} distributions with the F_{ST} distributions of random-selected SNPs that were matched to the trait SNPs by their allele frequencies in European populations (**control SNPs**).

5.2.0.1 GWAS Catalog

I used the R package *gwasrapidd* (Magno and Maia 2020) to query all traits in the GWAS Catalog (MacArthur et al. 2017) as of 9 August 2021 (N_{TRAITS} = 3,459). For 541 of these traits, no matching variant IDs could be pulled out from the 1000 Genomes VCFs, leaving N_{TRAITS} = 3,008.

5.2.1 Linkage disequilibrium

To obtain the "trait SNP" dataset, for each trait, I sought to isolate the SNP closest to each of its true causal variants, and exclude the SNPs in LD with them. To this end, I used PLINK 1.9 (Chang et al. 2015; Purcell and Chang, n.d.) to "clump" the SNPs associated with each of the remaining 3,008 traits, using an "index" SNP p-value threshold of 10^{-8} (Panagiotou, Ioannidis, and Genome-Wide Significance Project 2012), r^2 threshold of 0.1 (Hill and Robertson 1968), and base window size of 1 Mb. This process revealed 2,045 traits with at least one index SNP that met the p-value threshold. The index SNPs for each trait formed the set of trait SNPs, and **Figure 5.3** shows the counts of unique SNP IDs associated with each trait before and after clumping. In order to target relatively polygenic traits, I further filtered out traits with fewer than 10 trait SNPs, leaving N_{TRAITS} = 587.

5.2.2 Control SNPs

To obtain the "control SNP" dataset, I assigned each trait SNP to one of 20 bins based on its minor allele frequency in European populations (as provided in the original 1000 Genomes .vcf files under the column header 'INFO/AC_EUR'). For example, if a trait SNP had a minor allele frequency of 0.08 in European populations, it was assigned to the $(0.05, 0.1]$ bin. I did the same for all (un-associated) SNPs in the .vcf files, then paired each trait SNP with a random SNP from the .vcf file in the equivalent bin. These allele-frequency-paired random SNPs formed the set of "control SNPs", which I used to infer the F_{ST} distribution of a random set of SNPs with the same allele frequencies as the trait SNPs, and against which I could compare the F_{ST} distribution of the trait SNPs.

5.2.3 F_{ST} and ranking traits by signed Kolmogorov-Smirnov D statistic

I then calculated F_{ST} for each of the trait SNPs and their matched control SNPs using the Weir and Cockerham method (B. S. Weir and

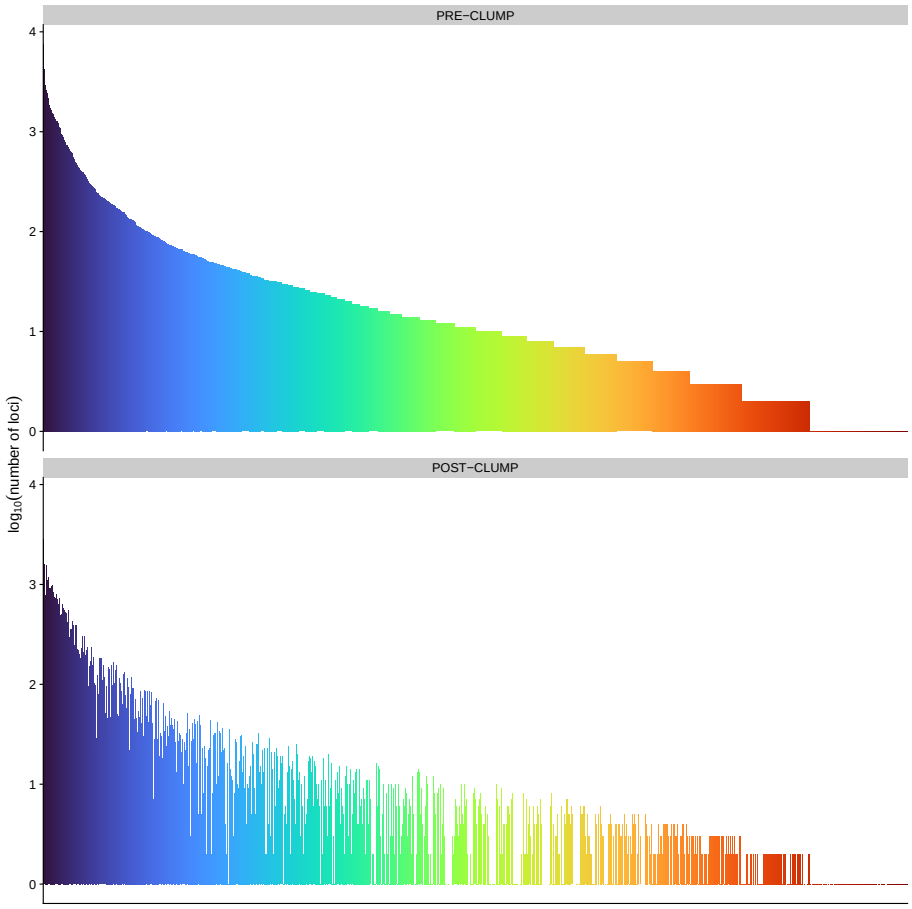


Figure 5.3: log10 counts of associated SNPs for each trait before and after the clumping process, which involved: a) excluding all SNPs with a p-value greater than 10^{-8}; and b) starting with the SNPs with the lowest p-values ("index" SNPs), excluding all other SNPs within a 1 Mb region of the index SNP with an LD r^2 of more than 0.1. [CHANGE COLOUR]

Cockerham 1984), as implemented in the R package *pegas* (Paradis 2010). I then sought to rank all traits based on the directional difference in F_{ST} distributions between trait and control SNPs. The Kolmogorov-Smirnov (KS) test is an appropriate test to measure the differences between distributions (Conover 1999), but the *p*-values are strongly affected by the number of samples in each distribution (here, the number of loci in each trait), and it does not measure whether one distribution is overall higher or lower than the other distribution. These limitations made it impossible to use the KS test *p*-values to compare the directional distances between trait and control SNPs across traits that have very different numbers of SNPs. I accordingly formulated the following method for ranking the traits based on the directional distance of the F_{ST} distributions of the trait SNPs compared to the control SNPs:

First, I ran three KS tests for each trait t with $x_t = F_{ST,traitSNPs}$ and $y_t = F_{ST,controlSNPs}$:

1. two-sided (D_t) ;
2. one-sided "greater" (D_t^+) ; and
3. one-sided "less" (D_t^-).

I note that $D_t = max(D_t^+, D_t^-)$, where D_t^+ is the greatest vertical distance attained by the eCDF of x_t over the eCDF of y_t, and D_t^- is the greatest vertical distance attained by the eCDF of y_t over the eCDF of x_t (Conover 1999; Durbin 1973). I then used a comparison of D_t^+ and D_t^- to formulate a "signed D statistic" (D_t^S), based on the logic that trait SNPs with a lower overall F_{ST} than control SNPs tend to have a higher D under the "greater" test than the "less" test, and vice versa.

Therefore, D_t^S:

$$\begin{aligned} D_t^- > D_t^+ : &\quad -D_t \\ D_t^- = D_t^+ : &\quad 0 \\ D_t^- < D_t^+ : &\quad D_t \end{aligned}$$

In **Figure 5.4** I present the F_{ST} distributions of trait SNPs for an illustrative subset of 28 human traits, ranked by D_t^S when compared with their matched control SNPs. **Figure 5.4A** shows the densities of SNPs as a function of F_{ST}, and **Figure 5.4B and C** show their empirical Cumulative Distribution Functions (eCDFs). **Figure 5.4B** includes the eCDFs of control SNPs in grey. eCDF figures for all 587 traits that passed our filters (Methods) are provided in the Appendix A.1.

These results show that for most traits, especially those that are highly polygenic, the F_{ST} of their associated alleles tend to be low, and to differ little from their matched control SNPs. This suggests that the causal variants of polygenic traits are generally shared across global populations at similar frequencies. Wang et al. (2020) determined through simulations that differences in LD and allele frequencies between populations can explain 70-80% of the loss of PGS relative accuracy for traits like body mass index and type 2 diabetes. This builds on the work of others [CITE PAPERS CITED BY BERG AND COOP AND ELABORATE]. As most of these variants have been discovered in European populations, the poor transferability of PGS between populations does not appear to be primarily driven by differences in the allele frequencies of these discovered variants, but perhaps rather by as-yet undiscovered variants of larger effect sizes in non-European populations, as well as differences between these populations in LD structure and environment.

There is a question as to whether the higher distributions of F_{ST} observed for hair shape, eye colour, skin pigmentation, and HIV infection, are due to those variants being discovered in diverse populations, such as those in Africa. For instance, if the alleles were predominantly discovered through GWAS on African populations, which contain an outsized proportion of genetic variation, those alleles would be more likely to show lower frequencies in other populations, and consequently higher F_{ST}. To examine the number of individuals from different ancestries that were used in these studies, I generated Figure 5.5. This figure shows that GWAS performed for the more comprehensively-studied phenotypes such as body height, BMI, and educational attainment, included individuals from many

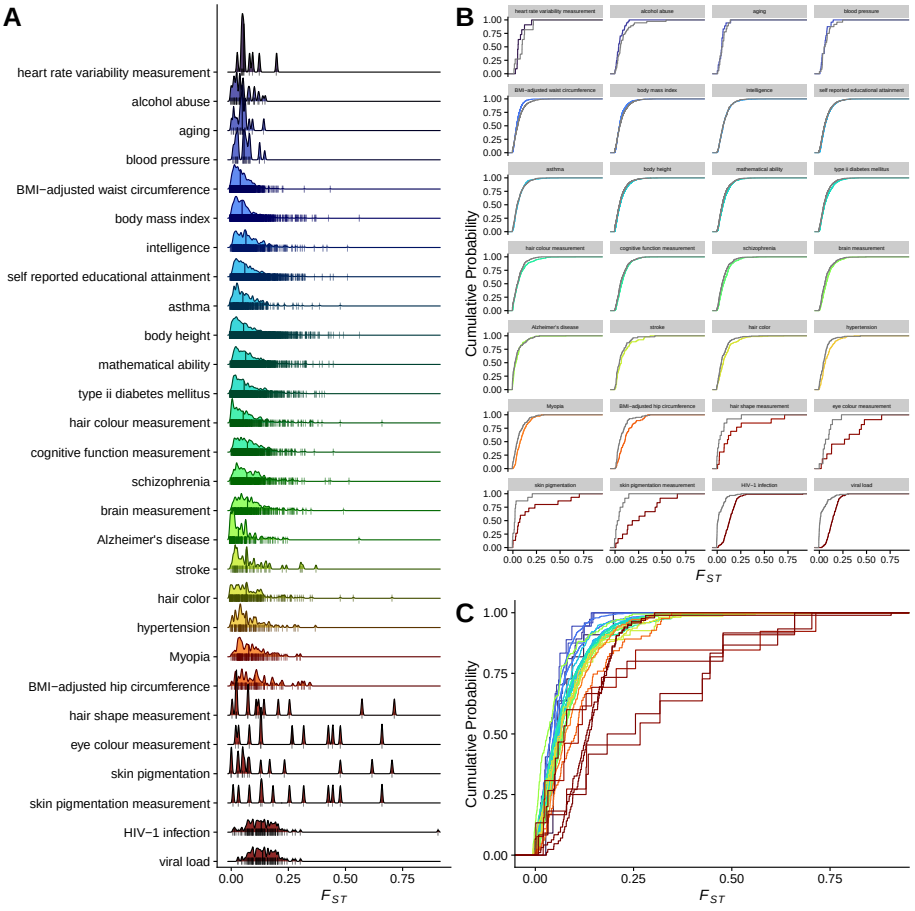


Figure 5.4: Distributions of F_{ST} across 28 illustrative human traits, ranked by signed-D (Kolmogorov-Smirnov test) comparing trait and control SNPs. **A**. F_{ST} density ridge plots with SNP markers. **B**. Empirical Cumulative Distribution Functions (eCDFs) of F_{ST} for trait-associated (colour) and random control (grey) SNPs, faceted by trait. **C**. Consolidated eCDFs of trait-associated SNPs from (B). eCDFs for all traits are included in the Appendix.

diverse populations, whereas the GWAS on pigmentation-related traits tended to focus on European or Latin American populations, as well as those of African ancestry in some cases. As expected, the GWAS on HIV-related traits included a higher proportion of individuals of African ancestry, which may explain the higher levels of F_{ST} observed there. I note that the GWAS Catalog data did not allow me to determine whether different studies on the same trait used the same cohorts (e.g. from UK Biobank), which would have the effect of inflating the counts shown in Figure 5.5.

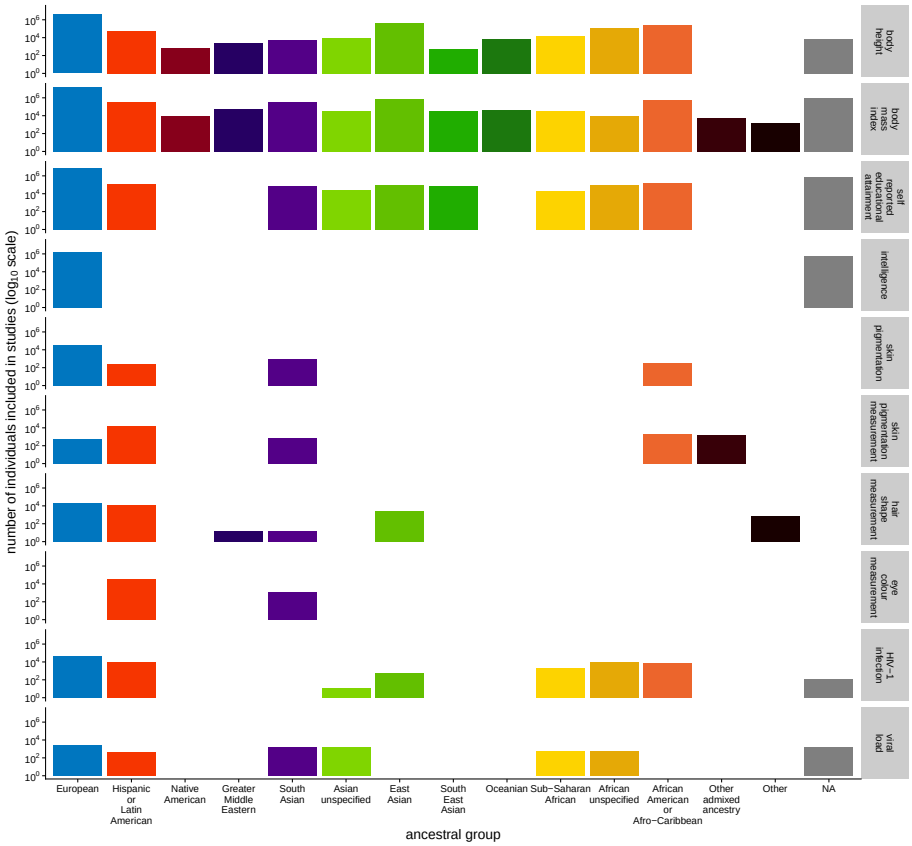


Figure 5.5: (ref:fst-n-indivs)

5.3 Implications

5.3.1 Limitation of F_{ST}

The first relevant limitation of F_{ST} is that it does not correspond to similarities or differences between populations in trait values. This is because F_{ST} does not take into account the effect size or the direction of effect of the trait-associated allele, so for highly-polygenic traits like the ones shown here, F_{ST} is almost entirely decoupled from the mean additive genetic value between populations (Jeremy J. Berg and Coop 2014a). This means that even when a trait's associated alleles have low F_{ST}, as we see for most traits here, it does not follow that the population mean values for that trait would be the same due to genetic similarities between those populations. As Kremer and Le Corre (2012) determined by through simulations, for highly polygenic traits, mean trait differences are driven by small, coordinated shifts in allele frequencies, which would not be detected by F_{ST} (Le Corre and Kremer 2012).

The second relevant limitation is that F_{ST} is not sufficiently powered to detect selection for polygenic traits (Jeremy J. Berg and Coop 2014b), which means that these results cannot be used to determine the extent to which alleles are under selection in a given population. This is apparent in our analysis, where in contrast with Racimo, Berg, and Pickrell (2018), height and BMI, which have been found to be under strong differential selection, appear similar to other polygenic traits.

5.3.2 The importance of environment, and a push for diversity in genetic studies

The obvious solution to this issue of non-transferability of PGS scores to diverse populations is to increase the representation of those populations in GWAS and PGS studies, as has been often proposed elsewhere (Alicia R. Martin et al. 2017; Alicia R. Martin

et al. 2018; Bien et al. 2019; Alicia R. Martin et al. 2019; Sirugo, Williams, and Tishkoff 2019; Wojcik et al. 2019). [ELABORATE]

References

"A Global Reference for Human Genetic Variation." 2015. *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Adhikari, Kaustubh, Javier Mendoza-Revilla, Anood Sohail, Macarena Fuentes-Guajardo, Jodie Lampert, Juan Camilo Chac'on-Duque, Malena Hurtado, et al. 2019. "A GWAS in Latin Americans Highlights the Convergent Evolution of Lighter Skin Pigmentation in Eurasia." *Nature Communications* 10 (1, 1): 358. https://doi.org/10.1038/s41467-018-08147-0.

Akey, Joshua M., Ge Zhang, Kun Zhang, Li Jin, and Mark D. Shriver. 2002. "Interrogating a High-Density SNP Map for Signatures of Natural Selection." *Genome Research* 12 (12): 1805–14. https://doi.org/10.1101/gr.631202.

Andersson, Leif, and Michel Georges. 2004. "Domestic-Animal Genomics: Deciphering the Genetics of Complex Traits." *Nature Reviews Genetics* 5 (3, 3): 202–12. https://doi.org/10.1038/nrg1294.

Baud, Amelie, Megan K. Mulligan, Francesco Paolo Casale, Jesse F. Ingels, Casey J. Bohl, Jacques Callebert, Jean-Marie Launay, et al. 2017. "Genetic Variation in the Social Environment Contributes to Health and Disease." *PLOS Genetics* 13 (1): e1006498. https://doi.org/10.1371/journal.pgen.1006498.

Berg, Jeremy J., and Graham Coop. 2014a. "A Population Genetic Signal of Polygenic Adaptation." *PLOS Genetics* 10 (8): e1004412. https://doi.org/10.1371/journal.pgen.1004412.

———. 2014b. "A Population Genetic Signal of Polygenic Adaptation." *PLOS Genetics* 10 (8): e1004412. https://doi.org/10.1371/journal.pgen.1004412.

Berg, Jeremy J, Arbel Harpak, Nasa Sinnott-Armstrong, Anja

Moltke Joergensen, Hakhamanesh Mostafavi, Yair Field, Evan August Boyle, et al. 2019. "Reduced Signal for Polygenic Adaptation of Height in UK Biobank." Edited by Magnus Nordborg, Mark I McCarthy, Magnus Nordborg, Nicholas H Barton, and Joachim Hermisson. *eLife* 8 (March): e39725. https://doi.org/10.7554/eLife.39725.

Bergelson, Joy, and Fabrice Roux. 2010. "Towards Identifying Genes Underlying Ecologically Relevant Traits in Arabidopsis Thaliana." *Nature Reviews Genetics* 11 (12, 12): 867–79. https://doi.org/10.1038/nrg2896.

Bien, Stephanie A., Genevieve L. Wojcik, Chani J. Hodonsky, Christopher R. Gignoux, Iona Cheng, Tara C. Matise, Ulrike Peters, Eimear E. Kenny, and Kari E. North. 2019. "The Future of Genomic Studies Must Be Globally Representative: Perspectives from PAGE." *Annual Review of Genomics and Human Genetics* 20 (1): 181–200. https://doi.org/10.1146/annurev-genom-091416-035517.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4 (1): 7. https://doi.org/10.1186/s13742-015-0047-8.

Clark, Samuel A., John M. Hickey, Hans D. Daetwyler, and Julius HJ van der Werf. 2012. "The Importance of Information on Relatives for the Prediction of Genomic Breeding Values and the Implications for the Makeup of Reference Data Sets in Livestock Breeding Schemes." *Genetics Selection Evolution* 44 (1): 4. https://doi.org/10.1186/1297-9686-44-4.

Conover, W. J. 1999. *Practical Nonparametric Statistics*. John Wiley & Sons. https://books.google.com?id=n_39DwAAQBAJ.

Consortium, 1000 Genomes Project et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68.

Domingue, Benjamin W., Daniel W. Belsky, Jason M. Fletcher, Dalton Conley, Jason D. Boardman, and Kathleen Mullan Harris. 2018. "The Social Genome of Friends and Schoolmates in the National Longitudinal Study of Adolescent to Adult Health." *Proceedings of the National Academy of Sciences* 115 (4): 702–7. https://doi.org/10.

1073/pnas.1711803115.

Duncan, L., H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson, and B. Domingue. 2019. "Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations." *Nature Communications* 10 (1, 1): 3328. https://doi.org/10.1038/s41467-019-11112-0.

Durbin, J. 1973. *Distribution Theory for Tests Based on Sample Distribution Function*. SIAM. https://books.google.com?id=zAryCrT1IUYC.

Edge, Michael D, and Graham Coop. 2019. "Reconstructing the History of Polygenic Scores Using Coalescent Trees." *Genetics* 211 (1): 235–62. https://doi.org/10.1534/genetics.118.301687.

Evans, Kathryn S., Marijke H. van Wijk, Patrick T. McGrath, Erik C. Andersen, and Mark G. Sterken. 2021. "From QTL to Gene: C. Elegans Facilitates Discoveries of the Genetic Mechanisms Underlying Natural Variation." *Trends in Genetics* 37 (10): 933–47. https://doi.org/10.1016/j.tig.2021.06.005.

Fitzgerald, Tomas, Ian Brettell, Adrien Leger, Nadeshda Wolf, Natalja Kusminski, Jack Monahan, Carl Barton, et al. 2022. "The Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel." *Genome Biology* 23 (1): 59. https://doi.org/10.1186/s13059-022-02623-z.

Fredman, David, Stefan J. White, Susanna Potter, Evan E. Eichler, Johan T. Den Dunnen, and Anthony J. Brookes. 2004. "Complex SNP-related Sequence Variation in Segmental Genome Duplications." *Nature Genetics* 36 (8, 8): 861–66. https://doi.org/10.1038/ng1401.

Fry, Anna, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. 2017. "Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population." *American Journal of Epidemiology* 186 (9): 1026–34. https://doi.org/10.1093/aje/kwx246.

Gravel, Simon, Brenna M Henn, Ryan N Gutenkunst, Amit R Indap, Gabor T Marth, Andrew G Clark, Fuli Yu, et al. 2011. "Demographic History and Rare Allele Sharing Among Human Populations." *Proceedings of the National Academy of Sciences* 108 (29): 11983–88. https://doi.org/10.1073/pnas.1019276108.

Gros, Pierre-Alexis, Herv'e Le Nagard, and Olivier Tenaillon. 2009. "The Evolution of Epistasis and Its Links With Genetic Robustness, Complexity and Drift in a Phenotypic Model of Adaptation." *Genetics* 182 (1): 277–93. https://doi.org/10.1534/genetics.108.099127.

Guo, Feng, Dipak K. Dey, and Kent E. Holsinger. 2009. "A Bayesian Hierarchical Model for Analysis of Single-Nucleotide Polymorphisms Diversity in Multilocus, Multipopulation Samples." *Journal of the American Statistical Association* 104 (485): 142–54. https://doi.org/10.1198/jasa.2009.0010.

Habier, David, Jens Tetens, Franz-Reinhold Seefried, Peter Lichtner, and Georg Thaller. 2010. "The Impact of Genetic Relationship Information on Genomic Breeding Values in German Holstein Cattle." *Genetics Selection Evolution* 42 (1): 5. https://doi.org/10.1186/1297-9686-42-5.

Harpak, Arbel, and Molly Przeworski. 2021. "The Evolution of Group Differences in Changing Environments." *PLOS Biology* 19 (1): e3001072. https://doi.org/10.1371/journal.pbio.3001072.

Hill, W. G., and Alan Robertson. 1968. "Linkage Disequilibrium in Finite Populations." *Theoretical and Applied Genetics* 38 (6): 226–31. https://doi.org/10.1007/BF01245622.

Hoffmann, Ary A., Carla M. Sgr'o, and Andrew R. Weeks. 2004. "Chromosomal Inversion Polymorphisms and Adaptation." *Trends in Ecology & Evolution* 19 (9): 482–88. https://doi.org/10.1016/j.tree.2004.06.013.

Holsinger, Kent E., and Bruce S. Weir. 2009. "Genetics in Geographically Structured Populations: Defining, Estimating and Interpreting FST." *Nature Reviews Genetics* 10 (9, 9): 639–50. https://doi.org/10.1038/nrg2611.

"Index of /Pub/Release-102/Emf/Ensembl-Compara/Multiple_alignments/50_fish.epo/." n.d. Accessed January 25, 2022. https://ftp.ensembl.org/pub/release-102/emf/ensembl-compara/multiple_alignments/50_fish.epo/.

"Index of /Voll/Ftp/Data_collections/1000g_2504_high_coverage/Working/20201028_3 n.d. Accessed March 24, 2022. http://ftp.1000genomes.ebi.ac.uk/voll/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/.

Johnson, William C., and Paul Gepts. 1999. "Segregation for Performance in Recombinant Inbred Populations Resulting from Inter-Gene Pool Crosses of Common Bean (Phaseolus Vulgaris L.)." *Euphytica* 106 (1): 45–56. https://doi.org/10.1023/A:1003541201923.

Katsumura, Takafumi, Shoji Oda, Hiroshi Mitani, and Hiroki Oota. 2019. "Medaka Population Genome Structure and Demographic History Described via Genotyping-by-Sequencing." *G3 Genes|Genomes|Genetics* 9 (1): 217–28. https://doi.org/10.1534/g3.118.200779.

Kerminen, Sini, Alicia R. Martin, Jukka Koskela, Sanni E. Ruotsalainen, Aki S. Havulinna, Ida Surakka, Aarno Palotie, et al. 2019. "Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland." *The American Journal of Human Genetics* 104 (6): 1169–81. https://doi.org/10.1016/j.ajhg.2019.05.001.

Khera, Amit V., Mark Chaffin, Krishna G. Aragam, Mary E. Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, et al. 2018. "Genome-Wide Polygenic Scores for Common Diseases Identify Individuals with Risk Equivalent to Monogenic Mutations." *Nature Genetics* 50 (9, 9): 1219–24. https://doi.org/10.1038/s41588-018-0183-z.

Kong, Augustine, Gudmar Thorleifsson, Michael L. Frigge, Bjarni J. Vilhjalmsson, Alexander I. Young, Thorgeir E. Thorgeirsson, Stefania Benonisdottir, et al. 2018. "The Nature of Nurture: Effects of Parental Genotypes." *Science* 359 (6374): 424–28. https://doi.org/10.1126/science.aan6877.

Kremer, A., and V. Le Corre. 2012. "Decoupling of Differentiation Between Traits and Their Underlying Genes in Response to Divergent Selection." *Heredity* 108 (4, 4): 375–85. https://doi.org/10.1038/hdy.2011.81.

Kullo, Iftikhar J., Hayan Jouni, Erin E. Austin, Sherry-Ann Brown, Teresa M. Kruisselbrink, Iyad N. Isseh, Raad A. Haddad, et al. 2016. "Incorporating a Genetic Risk Score Into Coronary Heart Disease Risk Estimates." *Circulation* 133 (12): 1181–88. https://doi.org/10.1161/CIRCULATIONAHA.115.020109.

Le Corre, Val'erie, and Antoine Kremer. 2012. "The Genetic Differentiation at Quantitative Trait Loci Under Local Adaptation."

Molecular Ecology 21 (7): 1548–66. https://doi.org/10.1111/j.1365-294X.2012.05479.x.

Leger, Adrien, Ian Brettell, Jack Monahan, Carl Barton, Nadeshda Wolf, Natalja Kusminski, Cathrin Herder, et al. 2022. "Genomic Variations and Epigenomic Landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel." *Genome Biology* 23 (1): 58. https://doi.org/10.1186/s13059-022-02602-4.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'NADonnell-Luria, et al. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." *Nature* 536 (7616, 7616): 285–91. https://doi.org/10.1038/nature19057.

Limami, Anis M., Clothilde Rouillon, Gaëlle Glevarec, Andr'e Gallais, and Bertrand Hirel. 2002. "Genetic and Physiological Analysis of Germination Efficiency in Maize in Relation to Nitrogen Metabolism Reveals the Importance of Cytosolic Glutamine Synthetase." *Plant Physiology* 130 (4): 1860–70. https://doi.org/10.1104/pp.009647.

Maas, Paige, Myrto Barrdahl, Amit D. Joshi, Paul L. Auer, Mia M. Gaudet, Roger L. Milne, Fredrick R. Schumacher, et al. 2016. "Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States." *JAMA Oncology* 2 (10): 1295–1302. https://doi.org/10.1001/jamaoncol.2016.1025.

MacArthur, Jacqueline, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, et al. 2017. "The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog)." *Nucleic Acids Research* 45 (D1): D896–901. https://doi.org/10.1093/nar/gkw1133.

Mackay, Trudy F. C., and Wen Huang. 2018. "Charting the Genotype–Phenotype Map: Lessons from the Drosophila Melanogaster Genetic Reference Panel." *WIREs Developmental Biology* 7 (1): e289. https://doi.org/10.1002/wdev.289.

Magno, Ramiro, and Ana-Teresa Maia. 2020. "Gwasrapidd: An R Package to Query, Download and Wrangle GWAS Catalog Data." *Bioinformatics* 36 (2): 649–50. https://doi.org/10.1093/bioinformatics/btz605.

Mal'ecot, Gustave. 1948. "Mathématiques de l'hérédité."

Martin, Alicia R, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. 2017. "Human Demographic History Impacts Genetic Risk Prediction Across Diverse Populations." *The American Journal of Human Genetics* 100 (4): 635–49. https://doi.org/10.1016/j.ajhg.2017.03.004.

Martin, Alicia R., Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. "Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities." *Nature Genetics* 51 (4, 4): 584–91. https://doi.org/10.1038/s41588-019-0379-x.

Martin, Alicia R, Solomon Teferra, Marlo Möller, Eileen G Hoal, and Mark J Daly. 2018. "The Critical Needs and Challenges for Genetic Architecture Studies in Africa." *Current Opinion in Genetics & Development*, Genetics of Human Origins, 53 (December): 113–20. https://doi.org/10.1016/j.gde.2018.08.005.

martin, Simon. (2016) 2022. *Simonhmartin/Genomics_general*. https://github.com/simonhmartin/genomics_general.

Martin, Simon H., John W. Davey, and Chris D. Jiggins. 2015. "Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci." *Molecular Biology and Evolution* 32 (1): 244–57. https://doi.org/10.1093/molbev/msu269.

McCarthy, Shane, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, et al. 2016. "A Reference Panel of 64,976 Haplotypes for Genotype Imputation." *Nature Genetics* 48 (10, 10): 1279–83. https://doi.org/10.1038/ng.3643.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

Mostafavi, Hakhamanesh, Arbel Harpak, Ipsita Agarwal, Dalton Conley, Jonathan K Pritchard, and Molly Przeworski. 2020. "Variable Prediction Accuracy of Polygenic Scores Within an Ancestry Group." Edited by Ruth Loos, Michael B Eisen, and Paul O'Reilly.

eLife 9 (January): e48376. https://doi.org/10.7554/eLife.48376.

Mukherjee, Siddhartha. 2016. *The Gene: An Intimate History*. Simon and Schuster. https://books.google.com?id=XvAsDAAAQBAJ.

Naruse, Kiyoshi, Shoji Fukamachi, Hiroshi Mitani, Mariko Kondo, Tomoko Matsuoka, Shu Kondo, Nana Hanamura, et al. 2000. "A Detailed Linkage Map of Medaka, Oryzias Latipes: Comparative Genomics and Genome Evolution." *Genetics* 154 (4): 1773–84. https://www.genetics.org/content/154/4/1773.

Natarajan, Pradeep, Robin Young, Nathan O. Stitziel, Sandosh Padmanabhan, Usman Baber, Roxana Mehran, Samantha Sartori, et al. 2017. "Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting." *Circulation* 135 (22): 2091–101. https://doi.org/10.1161/CIRCULATIONAHA.116.024436.

Okbay, Aysu, Yeda Wu, Nancy Wang, Hariharan Jayashankar, Michael Bennett, Seyed Moeen Nehzati, Julia Sidorenko, et al. 2022. "Polygenic Prediction of Educational Attainment Within and Between Families from Genome-Wide Association Analyses in 3 Million Individuals." *Nature Genetics*, March, 1–13. https://doi.org/10.1038/s41588-022-01016-z.

Panagiotou, Orestis A., John P. A. Ioannidis, and Genome-Wide Significance Project. 2012. "What Should the Genome-Wide Significance Threshold Be? Empirical Replication of Borderline Genetic Associations." *International Journal of Epidemiology* 41 (1): 273–86. https://doi.org/10.1093/ije/dyr178.

Paquette, Martine, Michael Chong, S'ebastien Th'eriault, Robert Dufour, Guillaume Par'e, and Alexis Baass. 2017. "Polygenic Risk Score Predicts Prevalence of Cardiovascular Disease in Patients with Familial Hypercholesterolemia." *Journal of Clinical Lipidology* 11 (3): 725–732.e5. https://doi.org/10.1016/j.jacl.2017.03.019.

Paradis, Emmanuel. 2010. "Pegas: An R Package for Population Genetics with an Integrated–Modular Approach." *Bioinformatics* 26 (3): 419–20. https://doi.org/10.1093/bioinformatics/btp696.

Paradis, Emmanuel, and Klaus Schliep. 2019. "Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R." *Bioinformatics* 35 (3): 526–28. https:

//doi.org/10.1093/bioinformatics/bty633.

Park, Ju-Hyun, Mitchell H. Gail, Clarice R. Weinberg, Raymond J. Carroll, Charles C. Chung, Zhaoming Wang, Stephen J. Chanock, Joseph F. Fraumeni, and Nilanjan Chatterjee. 2011. "Distribution of Allele Frequencies and Effect Sizes and Their Interrelationships for Common Genetic Susceptibility Variants." *Proceedings of the National Academy of Sciences* 108 (44): 18026–31. https://doi.org/10.1073/pnas.1114759108.

Peirce, Jeremy L., Lu Lu, Jing Gu, Lee M. Silver, and Robert W. Williams. 2004. "A New Set of BXD Recombinant Inbred Lines from Advanced Intercross Populations in Mice." *BMC Genetics 5* (1): 7. https://doi.org/10.1186/1471-2156-5-7.

Plomin, Robert, and Kathryn Asbury. 2005. "Nature and Nurture: Genetic and Environmental Influences on Behavior." *The ANNALS of the American Academy of Political and Social Science* 600 (1): 86–98. https://doi.org/10.1177/0002716205277184.

Popejoy, Alice B., and Stephanie M. Fullerton. 2016. "Genomics Is Failing on Diversity." *Nature 538* (7624, 7624): 161–64. https://doi.org/10.1038/538161a.

Porcu, Eleonora, Serena Sanna, Christian Fuchsberger, and Lars G. Fritsche. 2013. "Genotype Imputation in Genome-Wide Association Studies." *Current Protocols in Human Genetics* 78 (1): 1.25.1–14. https://doi.org/10.1002/0471142905.hg0125s78.

Pritchard, Jonathan K., and Molly Przeworski. 2001. "Linkage Disequilibrium in Humans: Models and Data." *The American Journal of Human Genetics* 69 (1): 1–14. https://doi.org/10.1086/321275.

Pszczola, M., T. Strabel, H. A. Mulder, and M. P. L. Calus. 2012. "Reliability of Direct Genomic Values for Animals with Different Relationships Within and to the Reference Population." *Journal of Dairy Science* 95 (1): 389–400. https://doi.org/10.3168/jds.2011-4338.

Purcell, Shaun M, and Christopher C Chang. n.d. *PLINK 1.9*. www.cog-genomics.org/plink/1.9/.

Racimo, Fernando, Jeremy J Berg, and Joseph K Pickrell. 2018. "Detecting Polygenic Adaptation in Admixture Graphs." *Genetics* 208 (4): 1565–84. https://doi.org/10.1534/genetics.117.300489.

S'egurel, Laure, and C'eline Bon. 2017. "On the Evolution of Lactase

Persistence in Humans." *Annual Review of Genomics and Human Genetics* 18 (August): 297–319. https://doi.org/10.1146/annurev-genom-091416-035340.

Saliba-Colombani, Vera, Mathilde Causse, Laurent Gervais, and Jacqueline Philouze. 2000. "Efficiency of RFLP, RAPD, and AFLP Markers for the Construction of an Intraspecific Map of the Tomato Genome." *Genome* 43 (1): 29–40. https://doi.org/10.1139/g99-096.

Saul, Michael C., Vivek M. Philip, Laura G. Reinholdt, and Elissa J. Chesler. 2019. "High-Diversity Mouse Populations for Complex Traits." *Trends in Genetics* 35 (7): 501–14. https://doi.org/10.1016/j.tig.2019.04.003.

Schumacher, Fredrick R., Ali Amin Al Olama, Sonja I. Berndt, Sara Benlloch, Mahbubl Ahmed, Edward J. Saunders, Tokhir Dadaev, et al. 2018. "Association Analyses of More Than 140,000 Men Identify 63 New Prostate Cancer Susceptibility Loci." *Nature Genetics* 50 (7, 7): 928–36. https://doi.org/10.1038/s41588-018-0142-8.

Schwarze, Katharina, James Buchanan, Jenny C. Taylor, and Sarah Wordsworth. 2018. "Are Whole-Exome and Whole-Genome Sequencing Approaches Cost-Effective? A Systematic Review of the Literature." *Genetics in Medicine* 20 (10): 1122–30. https://doi.org/10.1038/gim.2017.247.

Scutari, Marco, Ian Mackay, and David Balding. 2016. "Using Genetic Distance to Infer the Accuracy of Genomic Prediction." *PLOS Genetics* 12 (9): e1006288. https://doi.org/10.1371/journal.pgen.1006288.

Sella, G., and N. Barton. 2019. "Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies." *Annual Review of Genomics and Human Genetics*. https://doi.org/10.1146/annurev-genom-083115-022316.

Sham, P. C., S. S. Cherny, S. Purcell, and J. K. Hewitt. 2000. "Power of Linkage Versus Association Analysis of Quantitative Traits, by Use of Variance-Components Models, for Sibship Data." *The American Journal of Human Genetics* 66 (5): 1616–30. https://doi.org/10.1086/302891.

Sharp, Seth A., Stephen S. Rich, Andrew R. Wood, Samuel E. Jones,

Robin N. Beaumont, James W. Harrison, Darius A. Schneider, et al. 2019. "Development and Standardization of an Improved Type 1 Diabetes Genetic Risk Score for Use in Newborn Screening and Incident Diagnosis." *Diabetes Care* 42 (2): 200–207. https://doi.org/10.2337/dc18-1785.

Sirugo, Giorgio, Scott M. Williams, and Sarah A. Tishkoff. 2019. "The Missing Diversity in Human Genetic Studies." *Cell* 177 (1): 26–31. https://doi.org/10.1016/j.cell.2019.02.048.

Smigielski, Elizabeth M., Karl Sirotkin, Minghong Ward, and Stephen T. Sherry. 2000. "dbSNP: A Database of Single Nucleotide Polymorphisms." *Nucleic Acids Research* 28 (1): 352–55. https://doi.org/10.1093/nar/28.1.352.

Spivakov, Mikhail, Thomas O Auer, Ravindra Peravali, Ian Dunham, Dirk Dolle, Asao Fujiyama, Atsushi Toyoda, et al. 2014. "Genomic and Phenotypic Characterization of a Wild Medaka Population: Towards the Establishment of an Isogenic Population Genetic Resource in Fish." *G3 Genes|Genomes|Genetics* 4 (3): 433–45. https://doi.org/10.1534/g3.113.008722.

Sun, Luanluan, Lisa Pennells, Stephen Kaptoge, Christopher P. Nelson, Scott C. Ritchie, Gad Abraham, Matthew Arnold, et al. 2021. "Polygenic Risk Scores in Cardiovascular Risk Prediction: A Cohort Study and Modelling Analyses." *PLOS Medicine* 18 (1): e1003498. https://doi.org/10.1371/journal.pmed.1003498.

Svenson, Karen L, Daniel M Gatti, William Valdar, Catherine E Welsh, Riyan Cheng, Elissa J Chesler, Abraham A Palmer, Leonard McMillan, and Gary A Churchill. 2012. "High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population." *Genetics* 190 (2): 437–47. https://doi.org/10.1534/genetics.111.132597.

"The $11.9 Trillion Global Healthcare Market: Key Opportunities & Strategies (2014-2022) - ResearchAndMarkets.com." 2019. June 25, 2019. https://www.businesswire.com/news/home/20190625005862/en/The-11.9-Trillion-Global-Healthcare-Market-Key-Opportunities-Strategies-2014-2022---ResearchAndMarkets.com.

Threadgill, David W., Darla R. Miller, Gary A. Churchill, and Fernando Pardo-Manuel de Villena. 2011. "The Collaborative Cross:

A Recombinant Inbred Mouse Population for the Systems Genetic Era." *ILAR Journal* 52 (1): 24–31. https://doi.org/10.1093/ilar.52.1.24.

Tikkanen, Emmi, Aki S. Havulinna, Aarno Palotie, Veikko Salomaa, and Samuli Ripatti. 2013. "Genetic Risk Prediction and a 2-Stage Risk Screening Strategy for Coronary Heart Disease." *Arteriosclerosis, Thrombosis, and Vascular Biology* 33 (9): 2261–66. https://doi.org/10.1161/ATVBAHA.112.301120.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. "Genome-Wide Association Studies." *Nature Reviews Methods Primers* 1 (1, 1): 1–21. https://doi.org/10.1038/s43586-021-00056-9.

Vilhjálmsson, Bjarni J, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, et al. 2015. "Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores." *The American Journal of Human Genetics* 97 (4): 576–92.

Wang, Ying, Jing Guo, Guiyan Ni, Jian Yang, Peter M. Visscher, and Loic Yengo. 2020. "Theoretical and Empirical Quantification of the Accuracy of Polygenic Scores in Ancestry Divergent Populations." *Nature Communications* 11 (1, 1): 3865. https://doi.org/10.1038/s41467-020-17719-y.

Watanabe, Kyoko, Sven Stringer, Oleksandr Frei, Maša Umi'cevi'c Mirkov, Christiaan de Leeuw, Tinca J. C. Polderman, Sophie van der Sluis, Ole A. Andreassen, Benjamin M. Neale, and Danielle Posthuma. 2019. "A Global Overview of Pleiotropy and Genetic Architecture in Complex Traits." *Nature Genetics* 51 (9, 9): 1339–48. https://doi.org/10.1038/s41588-019-0481-0.

Weir, B. S., and C. Clark Cockerham. 1984. "Estimating F-Statistics for the Analysis of Population Structure." *Evolution* 38 (6): 1358–70. https://doi.org/10.2307/2408641.

Weir, Bruce S., Lon R. Cardon, Amy D. Anderson, Dahlia M. Nielsen, and William G. Hill. 2005. "Measures of Human Population Structure Show Heterogeneity Among Genomic Regions." *Genome Research* 15 (11): 1468–76. https://doi.org/10.1101/gr.4398405.

Wittbrodt, Joachim, Akihiro Shima, and Manfred Schartl. 2002. "Medaka — a Model Organism from the Far East." *Nature Reviews Genetics* 3 (1, 1): 53–64. https://doi.org/10.1038/nrg704.

Wojcik, Genevieve L., Mariaelisa Graff, Katherine K. Nishimura, Ran Tao, Jeffrey Haessler, Christopher R. Gignoux, Heather M. Highland, et al. 2019. "Genetic Analyses of Diverse Populations Improves Discovery for Complex Traits." *Nature* 570 (7762, 7762): 514–18. https://doi.org/10.1038/s41586-019-1310-4.

Wray, Naomi R., Michael E. Goddard, and Peter M. Visscher. 2007. "Prediction of Individual Genetic Risk to Disease from Genome-Wide Association Studies." *Genome Research* 17 (10): 1520–28. https://doi.org/10.1101/gr.6665407.

Wright, Sewall. 1949. "The Genetical Structure of Populations." *Annals of Eugenics* 15 (1): 323–54. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x.

Yengo, Loic, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon, Timothy M Frayling, et al. 2018. "Meta-Analysis of Genome-Wide Association Studies for Height and Body Mass Index in []700000 Individuals of European Ancestry." *Human Molecular Genetics* 27 (20): 3641–49. https://doi.org/10.1093/hmg/ddy271.

Appendix A

eCDF of all polygenic traits in the GWAS Catalog ranked by D_t^S

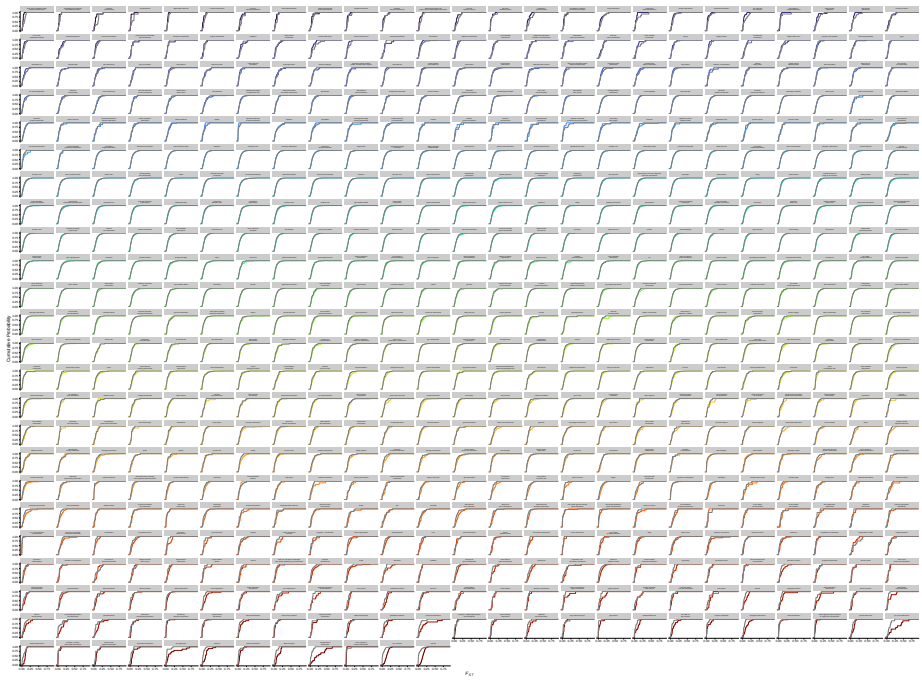


Figure A.1: 587 traits from the GWAS Catalog that passed our filters for polygenic traits, ranked by D_t^S.