Japanese courage: a genetic analysis of complex traits in medaka fish and humans

Ian Brettell

2022-08-20

Contents

About

0.1 Summary

Japanese courage: a genetic analysis of complex traits in medaka fish and humans

This thesis primarily explores how an individual's genes interact with the genes of their social companions to create differences in behaviour, using the Japanese medaka fish as a model organism. Chapter 1 sets out the introduction to the diverse topics covered in this thesis.

Chapter 2 describes several genomic characteristics of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel, which comprises 80 inbred lines of medaka that were bred from a wild population residing in Kiyosu, southern Japan. In this chapter I plot the inbreeding trajectory of the MIKK panel, and analyse its evolutionary relationship with other previously established inbred medaka strains; degree of homozygosity; rate of linkage disequilibrium decay; repeat content; and structural variation, all which relate to its utility for the genetic mapping of complex traits.

In Chapter 3, I use a custom behavioural assay to characterise and classify bold-shy behaviours in 5 previously established inbred medaka lines. Here I describe the assay, assess its robustness against confounding factors, and apply a hidden markov model (HMM) to classify the fishes' behaviours across a spectrum of boldness-shyness based on their distance and angle of travel between time points. I describe how the different lines differ in their behaviours over the

course of the assay (a direct genetic effect) and how the behaviour of a single "reference" line (*iCab*) differs in the presence of different lines (a social genetic effect).

In Chapter 4, I explain how I applied this behavioural assay to the MIKK panel in order to identify lines that diverge in both their own bold-shy behaviours (the direct genetic effect) and the extent to which they transmit those behaviours onto their tank partners (the social genetic effect). I then describe how we used those divergent lines as the parental lines in a multi-way F2 cross in an attempt to isolate the genetic variants that are associated with both direct and social genetic effects.

In Chapter *5* I describe the bioinformatic processes and genetic association models used to map the variants associated with differences in the period of somite development, based on a separate F2 cross between the southern Japanese *iCab* strain, and the northern Japanese *Kaga* strain.

Finally, in Chapter 6, I compare and rank all complex traits in the GWAS Catalog based on the extent to which their associated alleles vary across global human populations, using the Fixation Index (Fst) as a metric and the 1000 Genomes dataset as a sample of global genetic variation. In this chapter I set out the bioinformatic pipelines used to process the data, present the distributions of Fst for trait-associated alleles across the genome, and use the Kolmogorov-Smirnov test to compare the distributions of Fst across different traits.

Altogether, this thesis describes some of the genomic characteristics of both medaka fish and humans, and how those variations relate to differences in complex traits, with a particular focus on the genetic causes of adaptive behaviours and the transmission of those behaviours onto one's social companions.

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

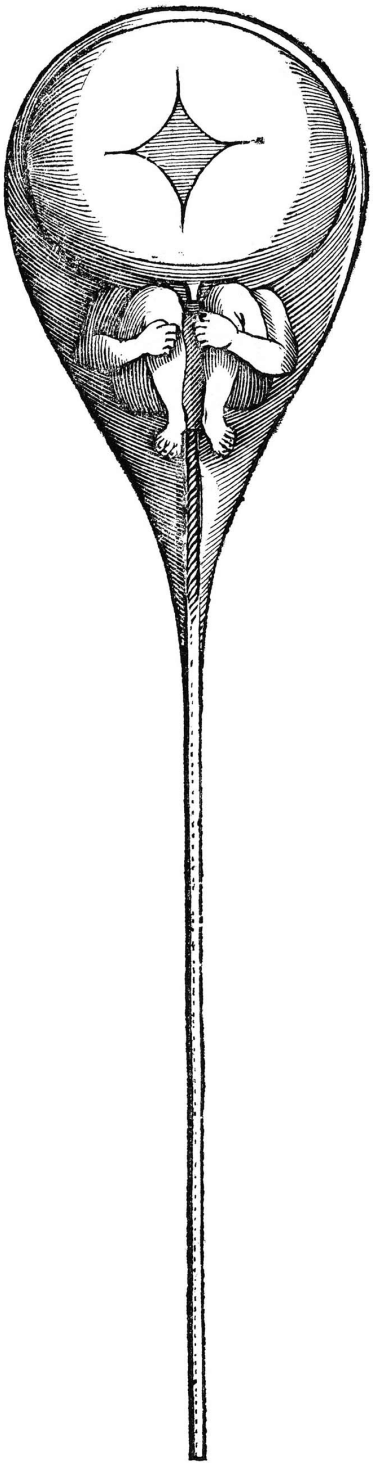
Around 500 BC, the Ancient Grecian Pythagoras applied his understanding of triangles to this question, proposing the theory known as "spermism". He posited that hereditary information was passed down from parent to child via male sperm, with the female only providing the nutrients that would allow it to grow, and, like the theorem that bears his name, that these two sides of the "triangle" would determine 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 in order to perfect humanity, by breeding perfect combinations of parents at perfect times.

Aristotle joined the discussion with his treatise *Generation of Animals*, where he noted that children inherited features from their mothers as well as their fathers, and raised 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). Aristotle, however, could not deduce the form in which the information was conveyed.

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. But what then triggered the expansion of the human form, involving the development of new parts from embryo to fetus? The answer could only have been some instruction, blueprint, or code, but any specifics were out of reach.



1.1.3 Darwin

In 1831, a 22-year-old English Clergyman named Charles Darwin boarded the HMS Beagle to commence what would turn out to be a XX-year-long voyage around PLACES. He had previously studied medicine and theology, although he was drawn to study the natural world, and had apprenticed with his fellow clergyman John Henslow, a botanist and geologist who curated the Cambridge Botanic Garden. At the time, natural historians were subject to their enquiries being restricted by the prevailing doctrine of the time, namely that of Creationism. A mechanistic description of how species – and individuals within the same species – differed from one another was a dangerous idea, as it was thought to threaten the doctrine of creation.

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

Two months after Darwin graduated from Cambridge, he received a letter from Henslow suggesting that he join the *Beagle*'s exploratory survey of South America as the "gentleman scientist" they were seeking to assist with the collection of specimens. As the *Beagle* travelled down the eastern coast of South America, Darwin collected a vast amount of living specimens and ancient fossils, including large extinct mammals such as the megatherium, near Punta Alta.

The *Beagle* eventually reached the Galápagos Islands on the coast of Peru, and archipelago of 18 islands formed from volcanic lava. They stayed there for five weeks, during which period Darwin collected carcasses of birds, lizards, and plants. Upon his return to England, Darwin was hailed as a minor celebrity among natural historians due to the collections of specimens he had gathered and shipped back. John Gould

Figure 1.1: Megatherium fossil collected by Charles Darwin, housed in the Natural History Museum, London. Photograph from "What Was Megatherium?" (n.d.).

1.1.4 Charles Darwin and Gregor Mendel

1.1.5 Breeding programs in agriculture

1.2 Mixed models for genetic association analysis

"The distinction between Mendelian loci and QTLs is artificial, as the same mapping techniques can be applied to both. In fact, the classification of genetic (and allelic) effects should be considered as a continuum. At one end of the spectrum is the dichotomous Mendelian trait with only two detectable and distinct phenotypes, which are governed by a single gene. At the other end are traits, such as growth, which are likely to be affected by many genes that each contribute a small portion to the overall phenotype." (Members of the Complex Trait Consortium 2003)

Since it became possible to sequence the genotypes of individuals at scale, it has been an ongoing point of debate as to how best to model the effects that genetic variants have on a trait of interest.

Population structure and unequal relatedness among individuals in a given cohort can lead to false discoveries (Ewens and Spielman 1995; Members of the Complex Trait Consortium 2003). This is because individuals who share common ancestries will share both variants that do affect the trait of interest, and variants that do not, and these variants will be correlated with one other due to that shared ancestry. Therefore, if an association is found between the causal variants and a trait of interest, the non-causal variants that are correlated with the causal variants will also be found to be statistically associated with the trait.

How then does one control for population structure?

Three methods were formerly used (Zhang et al. 2010):

1. Structured association (Jonathan K. Pritchard, Stephens, and Donnelly 2000)

2. Genomic control (Devlin and Roeder 1999)
3. Family-based tests of association (Abecasis, Cardon, and Cookson 2000)

"Ronald fished introduced random effects models to study the correlations of trait values between relatives" (Fisher 1919).

1.3 Somites

Vertebral number is precisely defined for a given species (Gomez et al. 2008), but is the most variable physical trait across species (Kimura, Shinya, and Naruse 2012); for example, frogs have 6-10, humans have 33, and snakes can have more than 300 (Gomez et al. 2008; Kimura, Shinya, and Naruse 2012). The total number of vertebrae is a function of both how long the somite segmentation process persists for, and the

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

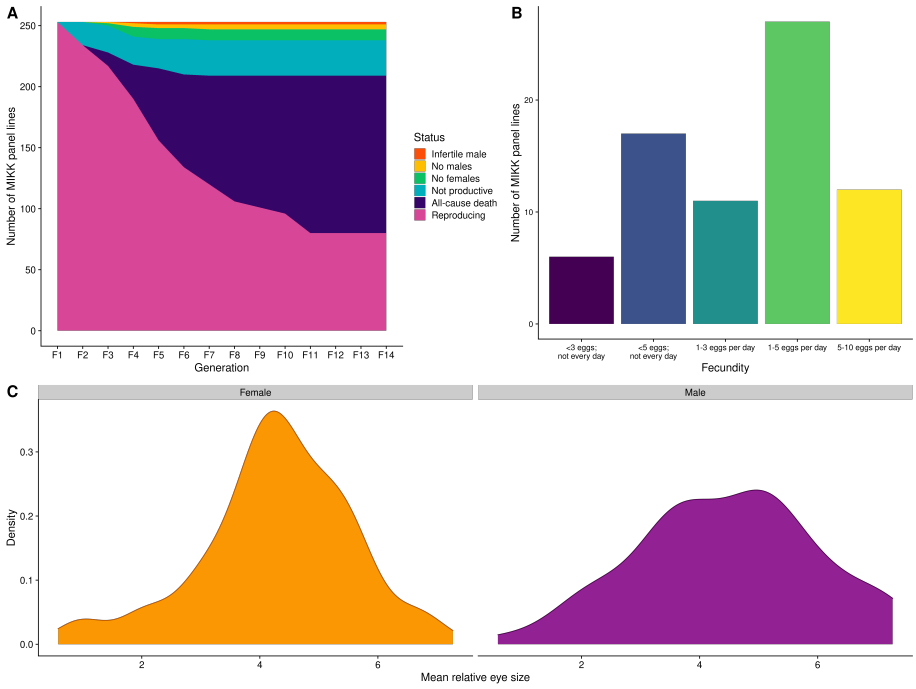


Figure 2.1: Inbreeding, fecundity and eye size in the MIKK panel lines. **A**: Status of all MIKK panel lines during the first 14 generations of inbreeding, showing cause of death for non-extant lines. **B**: Average fecundity of MIKK panel lines in generation F16, as measured during peak egg production 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)

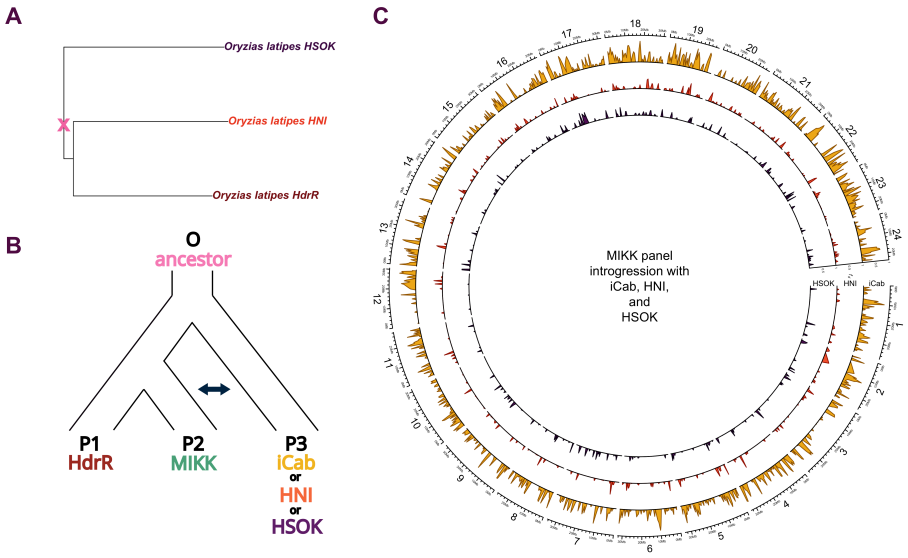


Figure 2.2: **Figure 2**: ABBA-BABA analysis. **A**. Phylogenetic tree generated from the Ensembl release 102 50-fish multiple alignment, showing only the medaka lines used in the ABBA-BABA analysis. **B**. Schema of the comparisons carried out in the ABBA-BABA analysis. **C**. Circos plot comparing 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-

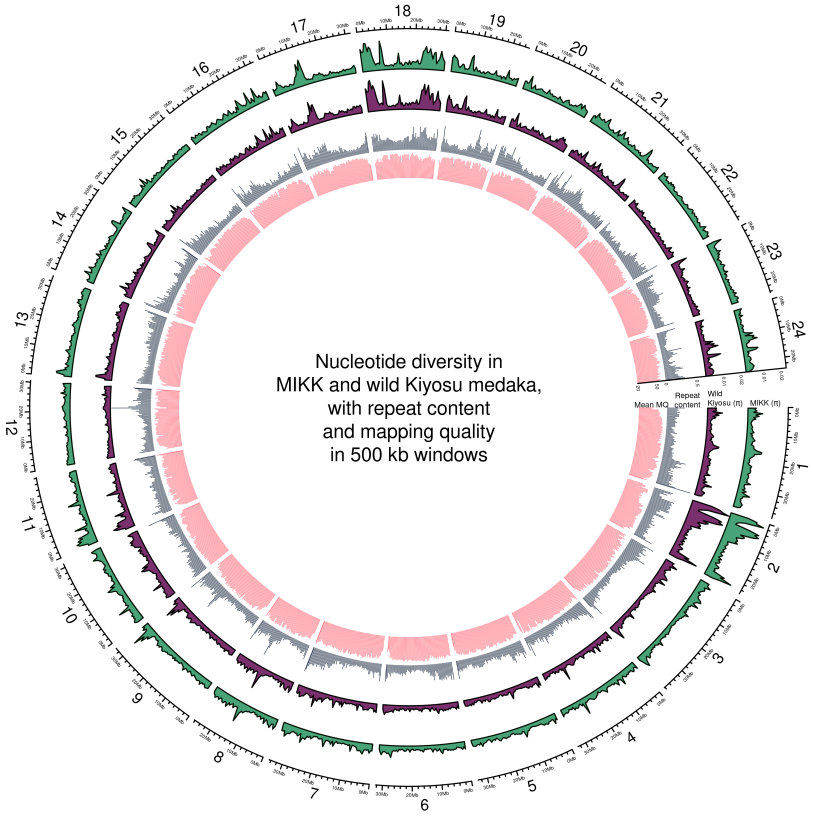


Figure 2.3: Circos plot with nucleotide diversity ($\hat{\pi}$) calculated within 500-kb non-overlapping windows for 63 non-sibling lines from the MIKK panel (*green*) and 7 wild Kiyosu medaka samples from the same originating population (*purple*); proportion of sequence classified as repeats by RepeatMasker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

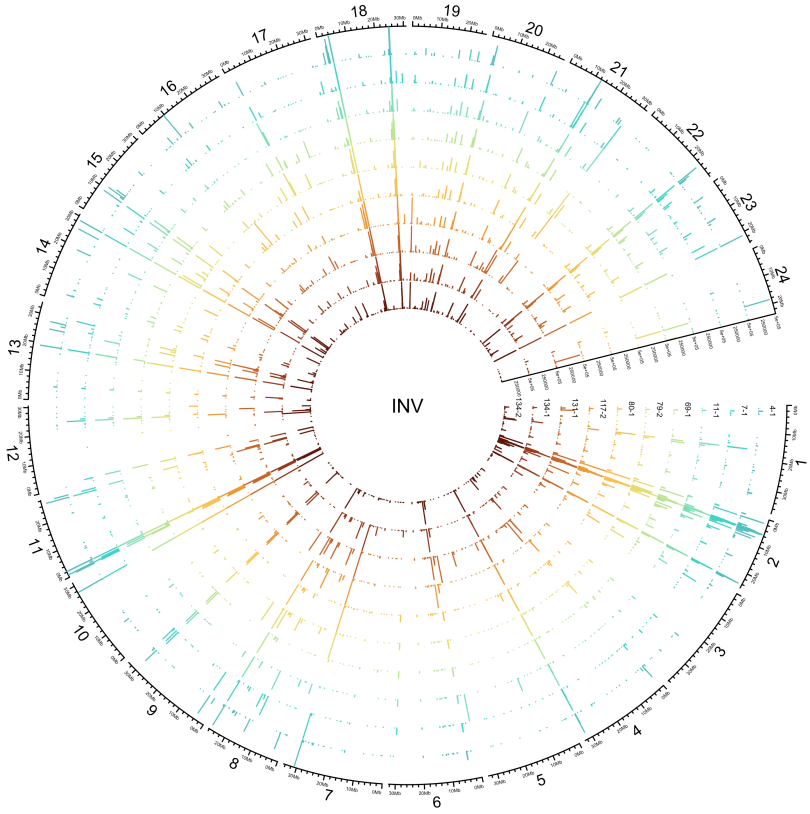


Figure 2.4: Inversions identified in 9 MIKK panel lines using a combination of Oxford Nanopore Technologies long-read and Illumina short-read sequences (see Chapter 2.3 below).

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

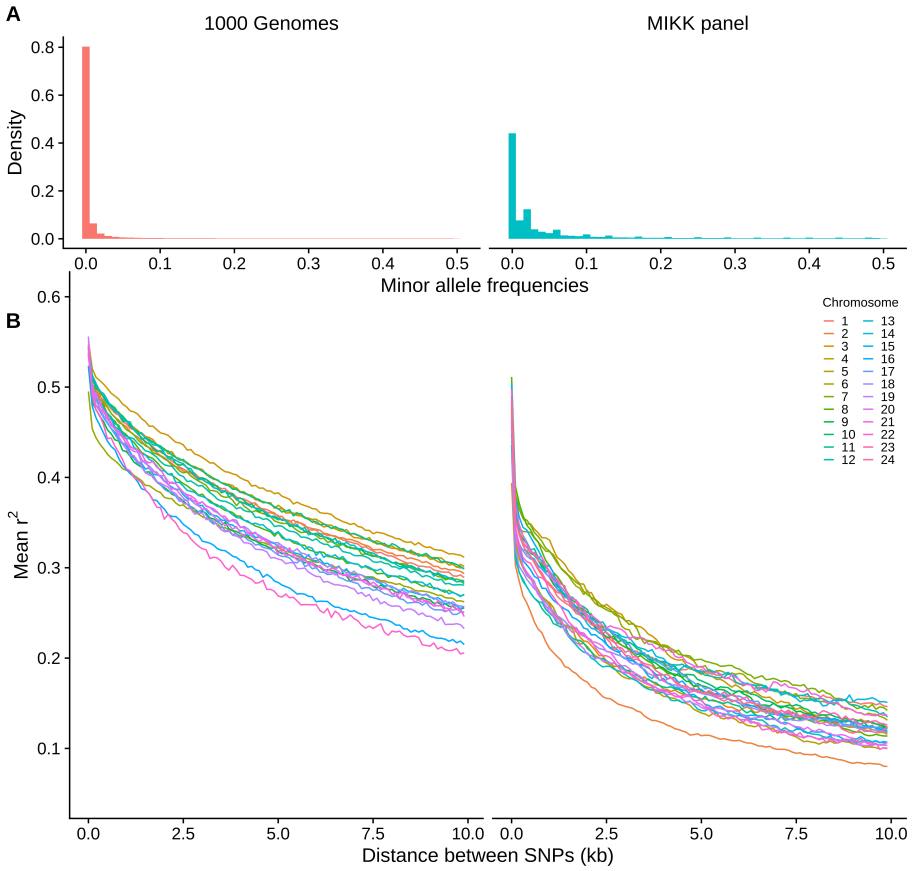


Figure 2.5: Minor allele frequency distributions and LD decay for biallelic, non-missing SNPs in the 1000 Genomes Phase 3 variant calls (N = 2,504) (1KG), and the MIKK panel Illumina-based calls excluding one of each pair of sibling lines (N = 63), across all autosomes (1KG: chrs 1-22; MIKK: chrs 1-24). **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-

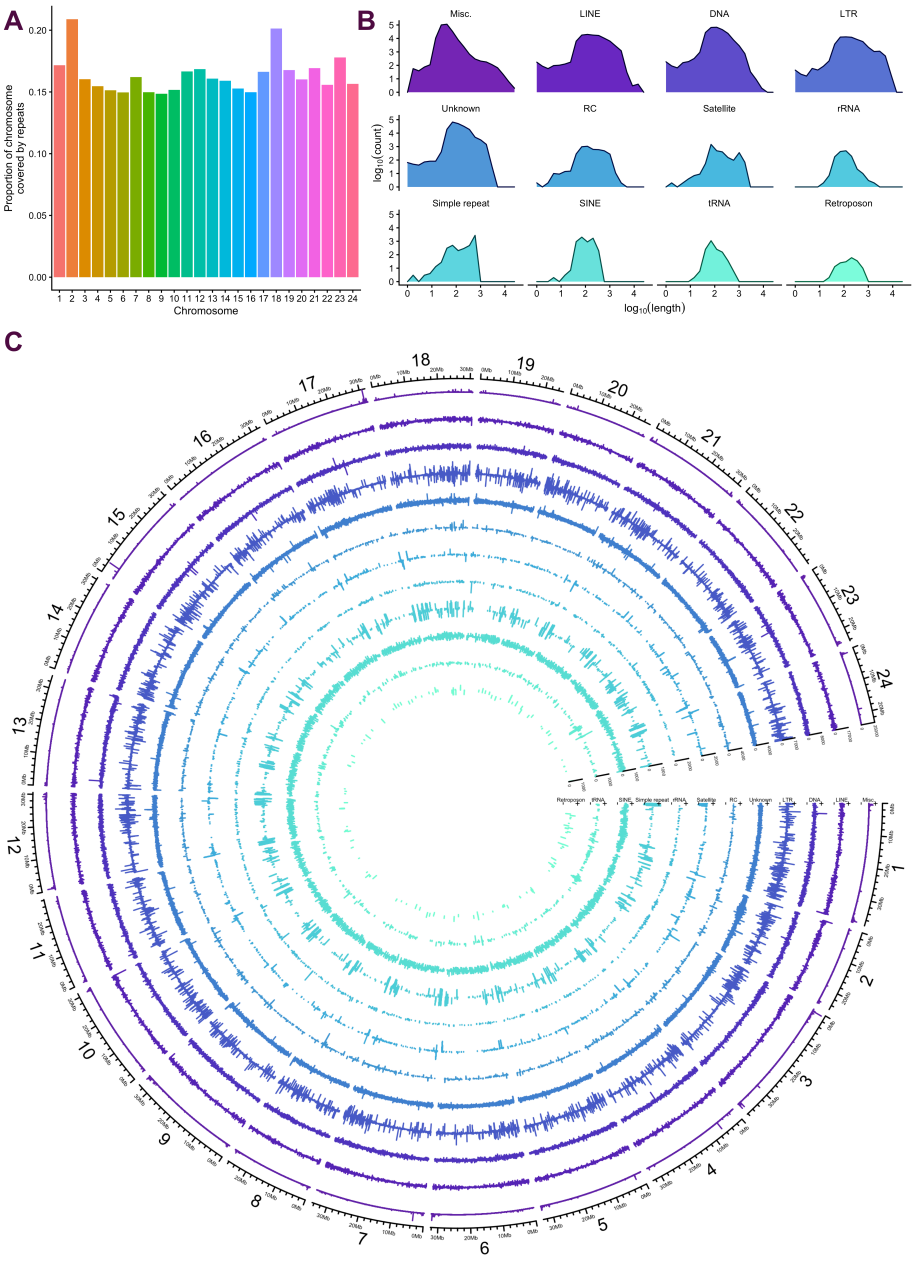


Figure 2.6: Repeat content in the *HdrR* genome based on RepeatMasker results obtained by Jack Monahan. **A**. Proportion of repeat content per-chromosome. **B**. log10 of repeat lengths and counts per repeat class. "Misc" includes all repeats assigned to their own specific class, for example "(GAG)n" 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

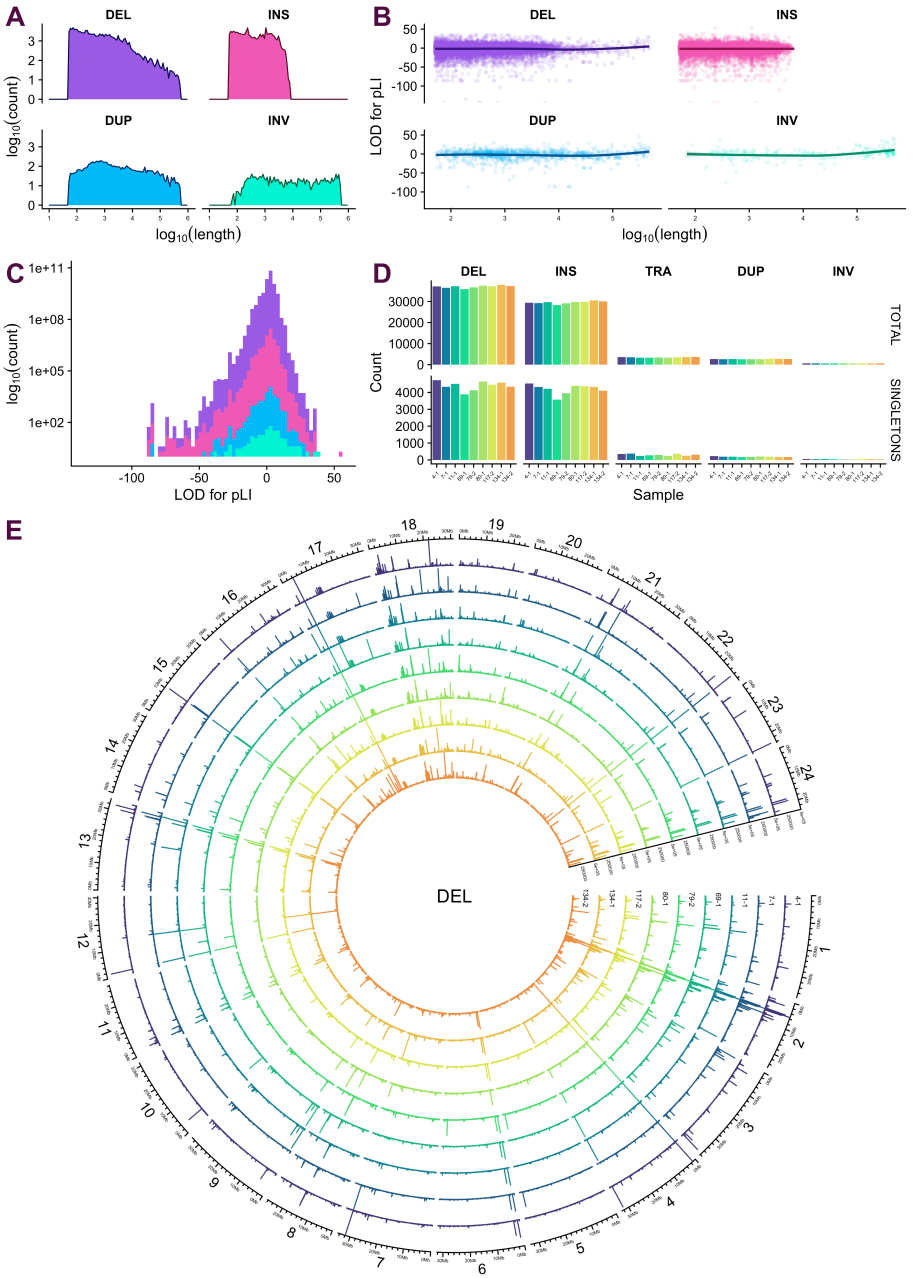


Figure 2.7: Polished SVs in 9 MIKK panel lines sequenced with ONT. DEL: deletion; INS: insertion; TRA: translocation; DUP: duplication; INV: inversion. **A**. Aggregate log10 counts and lengths of distinct SVs by type, excluding TRA. **B**. pLI LOD scores in distinct SVs by SV type. **C**. Histogram 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

3.1 Introduction

Humankind has long sought to understand what causes differences between the minds of individuals, and the behaviours they manifest. The origin of these differences resides in our genomes, as genetic differences between individuals are what set them on their unique developmental trajectories (Mitchell 2007). During the course of an organism's development, and ongoing throughout its lifespan, their genes interact with each other and the environment, leading to the creation of measurable traits, or phenotypes (Plomin and Asbury 2005). A core domain of biological inquiry concerns the extent to which an individual's value for a trait is determined by their genes (**G**), their environment (**E**), or an interaction between the two (**GxE**), commonly referred to as the interplay between 'nature and nurture' (Galton and Okamoto 1874; Plomin and Asbury 2005).

Experimental investigations of GxE in humans are hampered by the difficulties in sufficiently controlling for either genes or en-

vironment. To explore this fundamental question, researchers accordingly turned to model organisms, including the nematode (*Caenorhabditis elegans*) (Snoek et al. 2019), fruit fly (*Drosophila melanogaster*) (Schneider, Atallah, and Levine 2017), mouse (*Mus musculus*) (Baud et al. 2017), and teleost species such as zebrafish (*Danio rerio*) (Raterman et al. 2020) and medaka (*Oryzias latipes*). Medaka is an established model organism which has been studied in Japan for over a century(Wittbrodt, Shima, and Schartl 2002). In addition to possessing many favourable traits including ease of handling, a rapid reproduction cycle, a short lifespan, and a well-annotated genome, medaka has a trait that is singular among vertebrates – they have a high tolerance to inbreeding from the wild, allowing one to establish near-isogenic inbred strains with relative ease (Fitzgerald et al. 2022; Kirchmaier et al. 2015).

Inbred strains permit experimental quantification of complex trait parameters with sufficient replication under controlled assay conditions. By comparing individuals from different strains, one can measure the strength of genetic effects on a trait of interest. Furthermore, the low degree of genetic variance between individuals of an inbred strain allows one to measure the strength of environmental effects on a given trait by controlled variation of the environment.

3.2 Boldness-shyness

Behaviour is a complex trait that is affected by both genes and environment, and for social animals such as humans and medaka, one's social environment is considered likely to constitute a large component of the environmental effect (Ruzzante and Doyle 1990; Young 2008). Apart from social aspects, an organism must face many "hostile forces of nature" throughout its life (Buss 1991; Darwin 1859), such as food shortages, predation, harsh climate, and diseases. Adaptive behaviours allow individuals to navigate such dangers and maximise the likelihood of their survival at both the individual and population level (Lima and Dill 1990).

Boldness-shyness is thought to be a fundamental axis of behavioural

variation in many species, with an obvious causal relationship to an individual's likelihood of survival, and consequently with natural selection at the population level (Sloan Wilson et al. 1994). It represents an evolutionary trade-off between acquiring benefits (in terms of food or mates) and avoiding harms (in terms of predators or conspecific competitors), with each situation accompanied by its own optimal degree of risk (Lima and Dill 1990). It is both heritable (Svartberg 2002; Culum Brown, Burgess, and Braithwaite 2007), and subject to change following different life experiences or under different environmental conditions (Culum Brown, Burgess, and Braithwaite 2007). Boldness-shyness has been studied extensively with fish. Shy individuals tend to react to novelty by reducing their activity and becoming more vigilant, whereas bold individuals show higher levels of activity and exploratory behaviour (C. Brown, Jones, and Braithwaite 2007). One assay commonly used to measure this behavioural domain is referred to as the 'open field' assay, where fishes are observed while swimming freely in an experimental setting (C. Brown, Jones, and Braithwaite 2007; Laland, Krause, and Brown 2011; Lucon-Xiccato et al. 2022, 2020; Lucon-Xiccato and Bisazza 2017; Matsunaga and Watanabe 2010). Another is the 'novel object' assay, where a novel object is introduced to the fishes' environment to simulate a threat (C. Brown, Jones, and Braithwaite 2007; Schjolden, Stoskhus, and Winberg 2005; Wilson et al. 1993; Dominic Wright, Butlin, and Carlborg 2006; D. Wright et al. 2003). Where both assays were performed on the same fish, the behaviours exhibited were found to be correlated across assays, indicating that both were measuring the same boldness-shyness axis (C. Brown, Jones, and Braithwaite 2007). [**there is also an aspect of habituation in both parts of the test that should be mentioned**]

3.3 Social genetic effects

The specific environmental factor is the so-called indirect or social genetic effect, which describes the indirect interactions between an individual's genes and the genes of their companions (Baud et al.

2017). Social genetic effects have been shown to exert influence on various traits in mice including anxiety, wound healing, immune function, and body weight (Baud et al. 2017), and various traits related to development and survival in many species of livestock (Ellen et al. 2014). However, in those studies the social interactions were maintained throughout development, and it is unclear whether social genetic effects can still exert influence on adaptive behaviours during discrete, time-limited interactions.

Taking advantage of the rich genetic resources offered by medaka we established an OFT to examine the performance of isogenic inbred strains for an inter-strain comparative analysis of the boldness-shyness behaviour of medaka. We used a combined open-field and novel-object assay involving assaying pairs for fish, to determine: a) whether there were consistent differences in bold-shy behaviours exhibited by *5* established inbred strains of medaka fish (*iCab*, *HdrR*, and *HO5* from southern Japan, and *Kaga* and *HNI* from northern Japan); and b) whether there were consistent differences in bold-shy behaviours exhibited by a given strain (*iCab*) dependent on the strain that it was partnered with. The former was intended to measure the effect of an individual's own genes on its behaviour (a direct genetic effect), and the latter to measure the effect of the genes of the focal fish's tank partner on the behaviour of the focal fish (social genetic effect). Our experimental design allowed us to assess both direct and indirect effects on behaviour simultaneously, and to infer the degree to which variation in bold-shy behaviours is attributable to the differences in an individual's own genetics, the differences in the genetics of their social companions, and stochastic variation. A schema for the experimental plan is shown in **Figure 3.1**.

[Wrap up the findings in one or two sentences: We find inter-strain differences that are statistically significant Social genetic effects that influence behaviour of a strain depending on the partner (or genetic configuration of the test pair, not sure whether this sounds ok)]

To detect

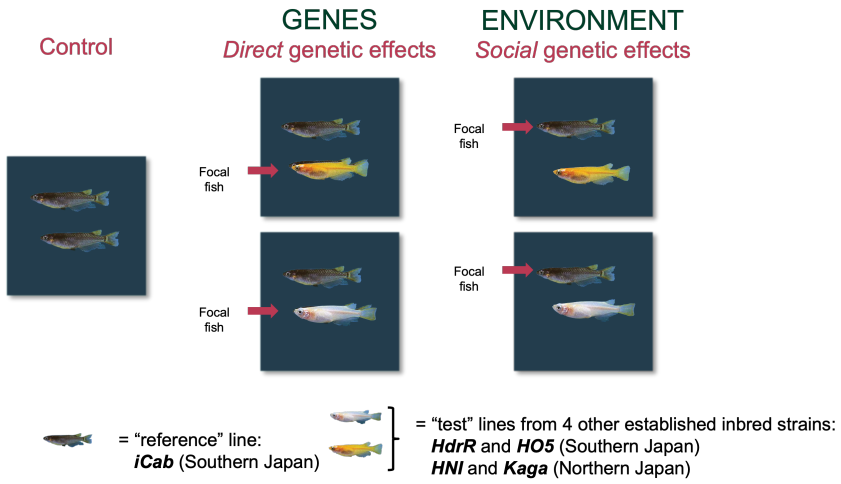


Figure 3.1: Schema for experimental plan. *iCab-iCab* pairings are the control condition. To explore direct genetic effects on behaviour, we compare the behaviours of test fishes from different lines, and infer that the differences between lines are caused by the differences in their genetics (direct genetic effects). To explore social genetic effects, we use the same data but turn our focus to the reference fish, and infer that the differences we observe between their behaviours are caused by the differences in their social environments, which are in turn driven by the different genetics in the test fish lines (social genetic effects).

3.4 Results

3.4.1 Data collection

Our behavioural assay is 20 minutes long, comprising two consecutively-run 10-minute components: a) an 'open field' component, where the fishes are introduced to the test tank and left to swim around freely; and b) a 'novel object' component, where a small black plastic cylinder is added to the tank at the beginning of the second 10-minute period, after which the fishes are again left to swim around freely. The assay is run on pairs of fish. Medaka is a seasonal breeder in which photoperiod has a strong effect on physiology and behaviour (L'opez-Olmeda et al. 2021). In this study we tested fish that were acclimated to summer conditions. To avoid confounding mating behaviours between males and females and associated aggressive interactions between males we used only female fish in all experiments. To increase the throughput of the assay, the test tank was divided into four quadrants with barriers, allowing us to run the assay on four pairs of fish simultaneously. Two test tanks situated side-by-side were used, allowing to run 8 concurrent assays. The experimental setup used is shown in Figure 3.2.

Between 11 and 16 June 2019, I assayed a total of 307 pairs of fish, comprising the following counts for each strain pairing: 68 *iCab/iCab*, 60 *iCab/HdrR*, 76 *iCab/HNI*, 47 *iCab/Kaga*, and 56 *iCab/HO5*. The locations of the lines' originating populations are shown in **Figure 3.3**. The fish from the iCab strain was denoted as the "reference fish", and was introduced to the test tank first. The "test fish" which was either another iCab fish (for the control condition), or a fish from one of the other four strains that were assayed in this experiment (*HdrR*, *HNI*, *Kaga*, *HO5*). The order in which the strains were assayed across the six days was randomly determined prior to the collection of the data. The test tanks were also rinsed between runs to remove any substances released by subjects during previous runs that could influence the behaviour of the subjects that followed.

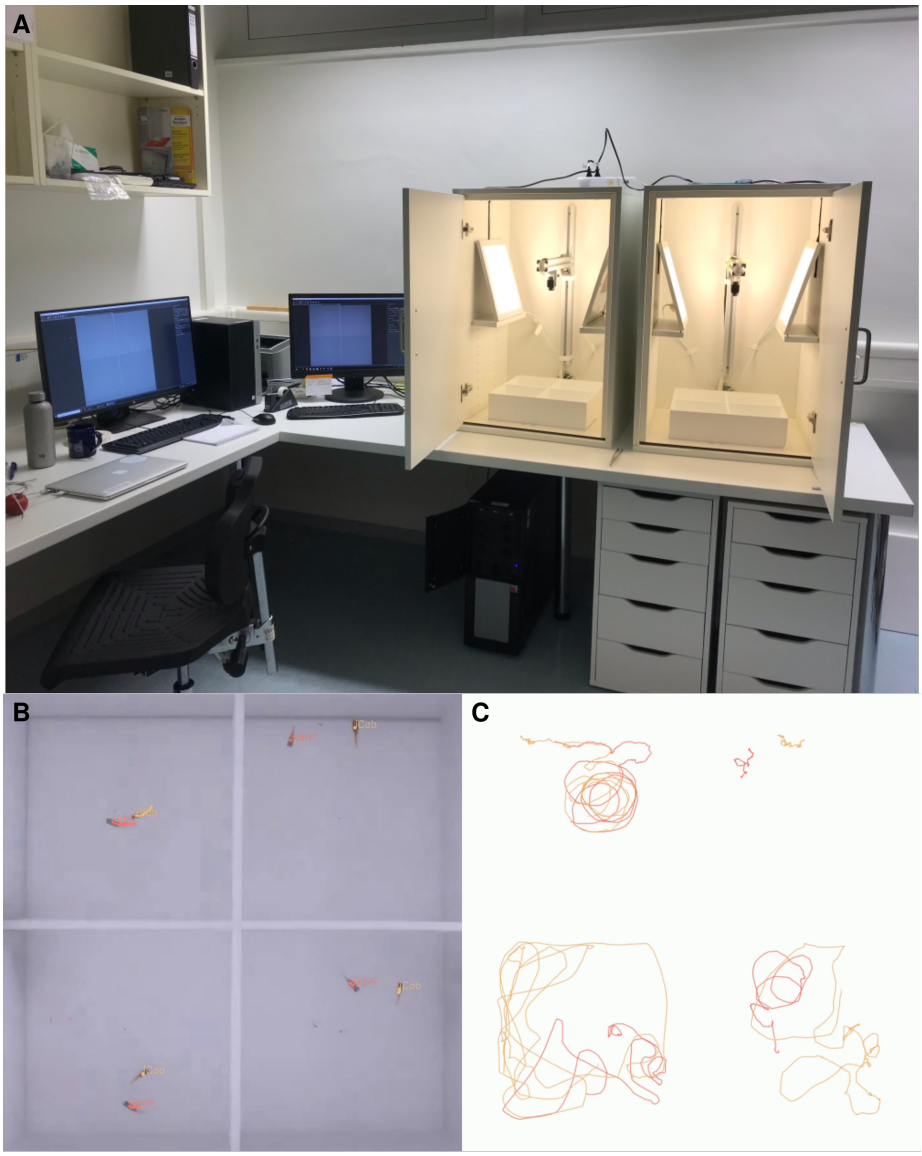


Figure 3.2: **A**: Experimental setup, with two test boxes side-by-side (denoted as "L" for left and "R" for right). Each test box contains one test tank, separated by removable barriers into quadrants, allowing for the simultaneous assay of four pairs of fish per test tank. The interior of the box is ambiently illuminated by LED lights, and a camera is suspended over the centre of each test tank to record the videos. **B**: Four pairs of fishes in a test tank with labelled quadrants (I, II, III, IV) and strains (*iCab* or *HdrR*). C: Paths of *iCab* reference fish and *HdrR* test fish from the video at panel (B) at 110 seconds.

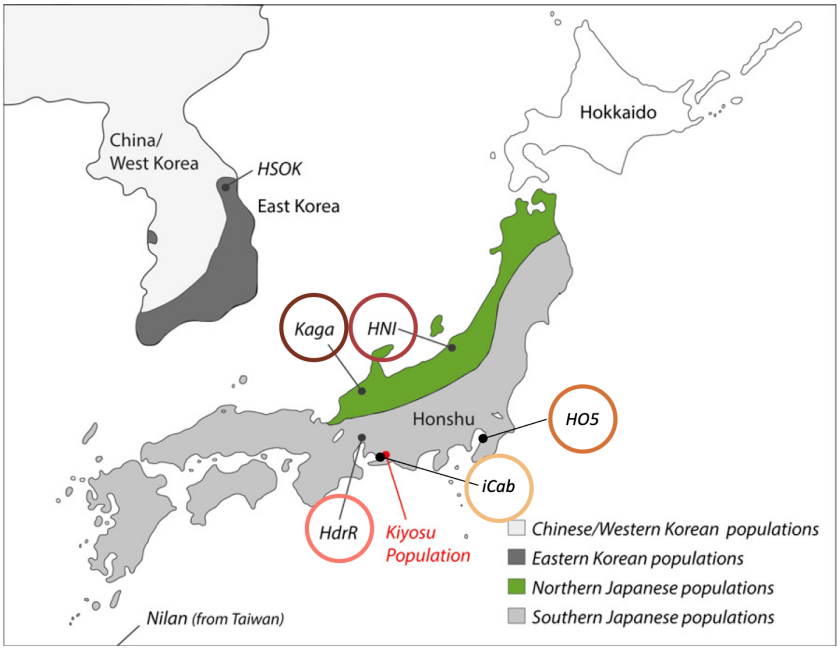


Figure 3.3: Image adapted from (Spivakov et al. 2014), showing the locations of the originating populations of the 5 inbred medaka lines used in this study.

3.4.2 Tracking

We split the videos by quadrant and assay component, and tracked the movement of the fishes with the open-access software package *idtrackerai* (Romero-Ferrero et al. 2019). Each individual fish was tracked across at least 85% of frames in each video, with 78% of videos tracked for over 99% of frames. The failure to track both fishes across all frames was usually caused by instances where a fish was situated behind the novel object or quadrant divider, rendering it invisible to the camera. A random selection of 20 videos was reviewed to search for instances of mislabeled fishes due to software errors. We found none, which is consistent with the reliability of the software as reported by its authors. We therefore concluded that such instances would be absent or very rare and thus negligible.

To identify which fish was the reference or test fish, we could generally distinguish between the strains based on their colour, as *iCab* pigmentation is light-brown, *HO5* are light-brown with an orange hue, *HdrR* females are white, and *HNI* and *Kaga* are grey. Where the reference fish was indistinguishable in colour from the test fish, we identified the fish that was introduced to the test tank first, and followed it by eye through to the first frame of the assay. In cases of *iCab-iCab* pairings, we randomly assigned the individuals to either the "reference" or "test" fish.

We then used the coordinates of the fishes in each frame to calculate the distance and angle travelled by each fish between a fixed time interval, for example every 0.05 seconds. Distance was calculated between points B [] C, and angle was calculated between points A [] B [] C. We used these distance and angle variables to train a HMM with the *hmmlearn* Python package (*Hmmlearn/Hmmlearn* [2014] 2022) in order to classify the movements as discrete behaviours or "modes of movement".

3.4.3 Effect of covariates

We examined the effects of several covariates, including date of assay, time of assay, tank quadrant, and tank side. To achieve this, we

calculated the mean speed of individuals in *iCab-iCab* pairings (N = 136) over the course of the entire 20-minute video (including both open field and novel object assays), and ran a multi-way ANOVA with all covariates (**Figure 3.4**). We found significant differences for date of assay and and tank quadrant (p = 0.0139 and 0.0108), but not for time of assay or tank side. This may have been driven by a difference in the way the assay was performed on the first day of the experiment (11 June 2019), where we used a thick fabric sheet to cover the front of the box rather than the wooden doors shown in **Figure 3.2**, as they were only installed the following day. The greater level of external light and sound permeating through the fabric may have caused the fishes to exhibit slower movement on that first day, and when the data from that day is excluded, the p-values for date of assay and tank quadrant increase to 0.0477 and 0.0469 respectively.

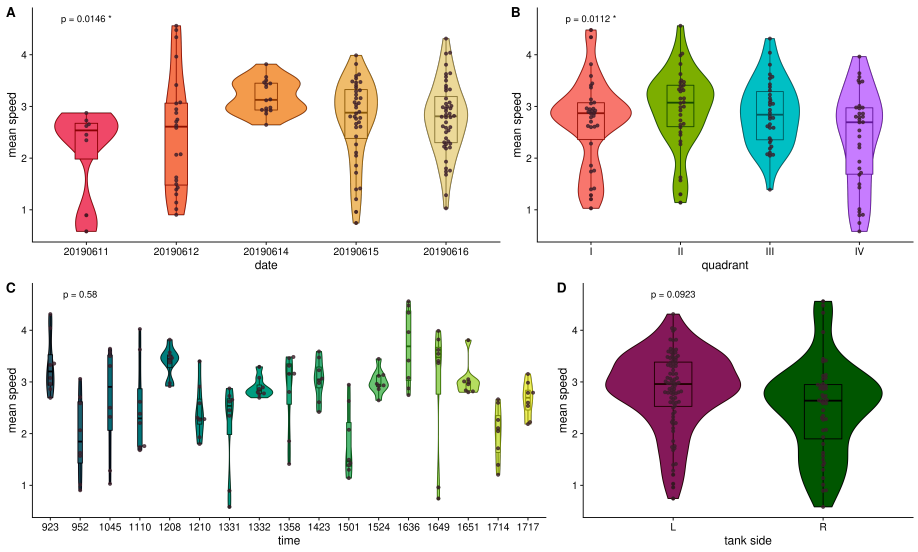


Figure 3.4: Effect of covariates on mean speed of iCab individuals when paired with another iCab over the course of the full video (including both open field and novel object assays). *P*-values were calculated from a multi-way ANOVA with all four covariates included as terms, and adjusted for false discovery rate.

3.4.4 Choice of time interval and number of HMM states

HMMs are well-suited for classifying the hidden states that generate stochastic, sequential observations. To determine the optimal parameters for the HMM's classification of behaviours, we sought to reduce overfitting (which tends to favour using a lower number of HMM states) while maximising the ability to distinguish between strains based on the relative time they spent in each HMM state (which tends to favour a higher number of HMM states). We additionally considered the time interval between which the distance and angle of travel was measured.

Figure 3.5 sets out the comparison of HMM parameters on two measures designed to quantify, respectively, the level of overfitting ('mean concordance between cross-validated HMM states'), and the quantification of differences between strains ('summed Kruskal-Wallis statistic comparing frequency of time spent in each HMM state across medaka strains) (Methods).

Based on these results and a visualisation of the polar plots for each combination of state number and time interval, we excluded combinations with 15 or more states due to an asymmetry across states that would create difficulties for interpreting their biological meaning (Supplementary material). For the downstream analysis we selected the combination of 14 states with a 0.08-second interval between time points, because out of the remaining combinations it appeared to optimally balance the level of overfitting and detection of differences between strains.

The distances and angles of travel for the 14 states are shown in Figure 4, generated from 10,000 randomly-selected data points (i.e. movements). Each point represents the radial distance in log10(pixels) that a fish travelled from its previous location (B [] C), and the polar angle travelled from A [] B [] C, where A [] B is always aligned vertically along the 0-180° axis. For example, a point at 45° far from the pole represents a fast, forward movement to the right, and a point at 260° close to the pole represents a slow, backward

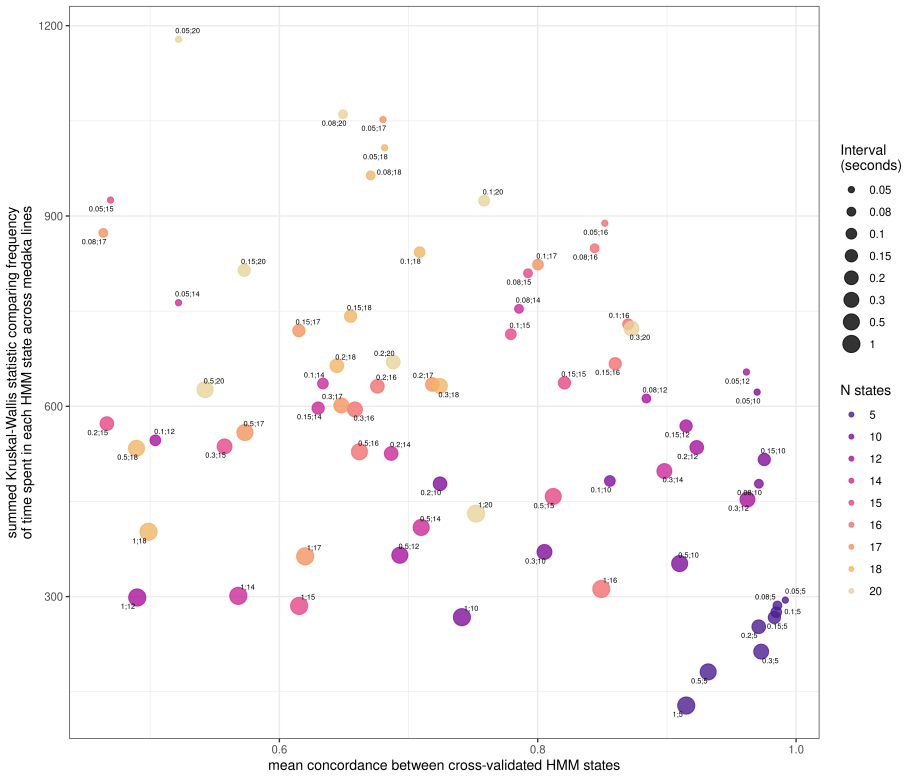


Figure 3.5: Comparison between HMM parameters. Horizontal axis: Mean concordance between states assigned by HMMs through a 2-fold cross-validation process. Vertical axis: Kruskal-Wallis statistic comparing strains based on the proportion of time spent in each HMM state, summed across all states. Size of points correspond to the interval, in seconds, between which the distance and angle of travel was calculated (Methods).

movement to the left.

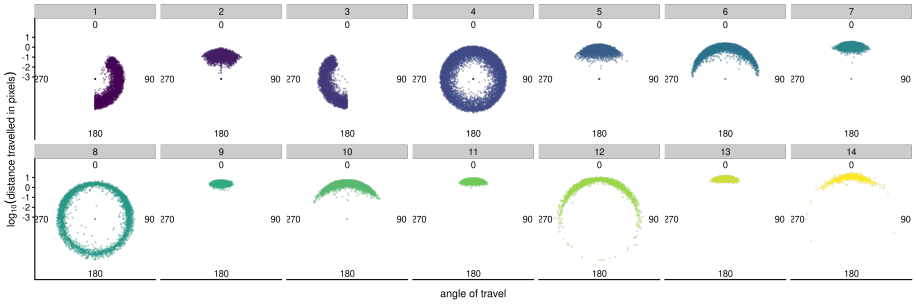


Figure 3.6: Classification of movements by the 14-state HMM, based on distance (in log10(pixels)) and angle of travel between a time interval of 0.08 seconds from points B [] C for distance, and points A [] B [] C for angle. Each point represents the distance and angle at point C, and A [] B is aligned vertically along the 0-180° radial axis. The figure shows an illustrative 10,000 data points (movements), randomly selected from the full dataset. States are sorted in ascending order by mean distance.

We next sought to determine whether there was a difference in the proportion of time spent in each state between the *5* inbred strains (implying the presence of direct genetic effects), and between the *iCab* reference fishes depending on the inbred strains of their tank partner (implying the presence of social genetic effects).

3.4.5 Direct and social genetic effects on bold/shy behaviour

3.4.5.1 Direct genetic effects

To determine whether the test fish strains differed in the proportions of time they spent in each state, we ran separate multi-way ANOVAs

for each combination of assay component (open field or novel object) and state (states 1 to 14):

$$frequency_{assaycomponent,state} = \beta_1(testfishstrain)+\beta_2(date)+\beta_3(time)+\beta_4(quadrar$$

The proportion of time each individual spent within a state was first inverse-normalised within each combination of assay and state, and the date of assay, time of assay, tank quadrant, and tank side were included as covariates. P-values were then adjusted for the False Discovery Rate (**FDR**).

The states that showed a significant difference are set out in **Table 3.1**. The test fish strains differed significantly in the proportion of time spent in a given state ($p < 0.05$, FDR-adjusted) for 11 out of 14 states in the open field assay (4.45x10-24 < p < 1.06x10-2), and 7 out of 14 states for the novel object assay (5.53x10-16 < p < 4.2x10-3), with the strain of the test fish explaining up to ~28% of the variance in the proportion of time spent in a given state. For some states, there was also a significant difference between quadrants, date of assay, and tank side (open field: 1.03e-09 < p < 1.18x10-2; novel object: 1.75x10-8 < p < 4.9x10-2). Full tables for all states and variables are provided in the Supplementary Material.

Figure 3.7 depicts the time dependence of HMM states over the course of the video, and the regions of the test tank that were most frequently occupied by the different strains. Figures A and H show the HMM states with panels coloured red to indicate the states that were found to significantly differ between test fish strains. **Figures 3.7B** and **E** show how each individual test fish, grouped by inbred strain, moved through HMM states throughout the course of the video, with each tile coloured by the state most frequently-occupied by the individual within 2-second intervals. Figures **3.7C** and **F** show the same data, recalculated as densities within each strain, with only the states that showed significant differences between strains in colour, and the remaining states consolidated and coloured grey. There is an especially clear difference between strains in the proportions of time spent in the slowest-moving states (i.e. states

Table 3.1: Significant differences in the proportion of time spent in each HMM state across test fish strains for the open field and novel object assay components.

Assay	State	Variance explained (%)	p-value (FDR-adjusted)
open field	1	26.62	3.60e-22
open field	2	21.94	6.18e-18
open field	3	28.21	4.45e-24
open field	4	17.73	2.12e-15
open field	5	7.48	2.78e-05
open field	6	5.60	2.48e-04
open field	7	5.46	3.11e-04
open field	8	5.79	1.08e-04
open field	9	7.36	8.96e-06
open field	10	3.78	1.06e-02
open field	13	4.66	3.42e-03
novel object	1	17.19	7.28e-14
novel object	2	14.28	2.92e-11
novel object	3	19.30	5.53e-16
novel object	4	13.12	1.26e-10
novel object	6	4.12	4.20e-03
novel object	7	8.93	1.84e-06
novel object	9	10.68	9.87e-08

1 to 3) at the beginning of each assay component, with an increase across all strains in the novel object assay, likely as a consequence of having less room to move, as well as the fear response that the novel object was designed to elicit. The differences are clearest when comparing the southern Japanese medaka strains (*iCab*, *HdrR*, and *HO5*) against the northern Japanese strains (*Kaga* and *HNI*). In addition, we note that *Kaga* tends to spend more time in the fast- and forward-moving state 13 at the beginning of the assay than at other times, which suggests that for that strain, such movements may be the manifestation of a stress response. **Figures 3.7D** and **G** show, as densities, the regions of the tank that were most frequently occupied by each strain. Although the northern Japanese strains *Kaga* and *HNI* were similarly fast-moving relative to the southern Japanese strains, in the open field assay component they appear to favour different regions of the tank – where *HNI* occupied the central regions of the tank with more frequency, *Kaga* tended to prefer swimming along the boundaries of the tank.

3.4.5.2 Social genetic effects

To determine whether the *iCab* reference fishes altered their behaviour depending on the inbred strain of their tank partner, we carried out the same analysis and model as above using only data from the *iCab* reference fishes. The states that showed a significant difference are set out in **Table 3.2**. The *iCab* reference fishes differed significantly in the proportion of time they spent in a given state depending on the strain of their tank partner ($p < 0.05$, FDR-adjusted) for 5 out of 14 states in the open field assay (2.04x10-6 < p < 2.11x10-2), and 7 out of 14 states for the novel object assay (4.76x10-7 < p < 4.06x10-2). The strain of the tank partner explained up to ~9% of the variance in the proportion of time the *iCab* reference spent in a given state. Full tables for all states and variables are provided in the Supplementary Material.

We observe that the proportions of time the *iCab* reference fishes spent in different states do differ based on the strains of their tank partner, and that their behavioural patterns tend to reflect those of

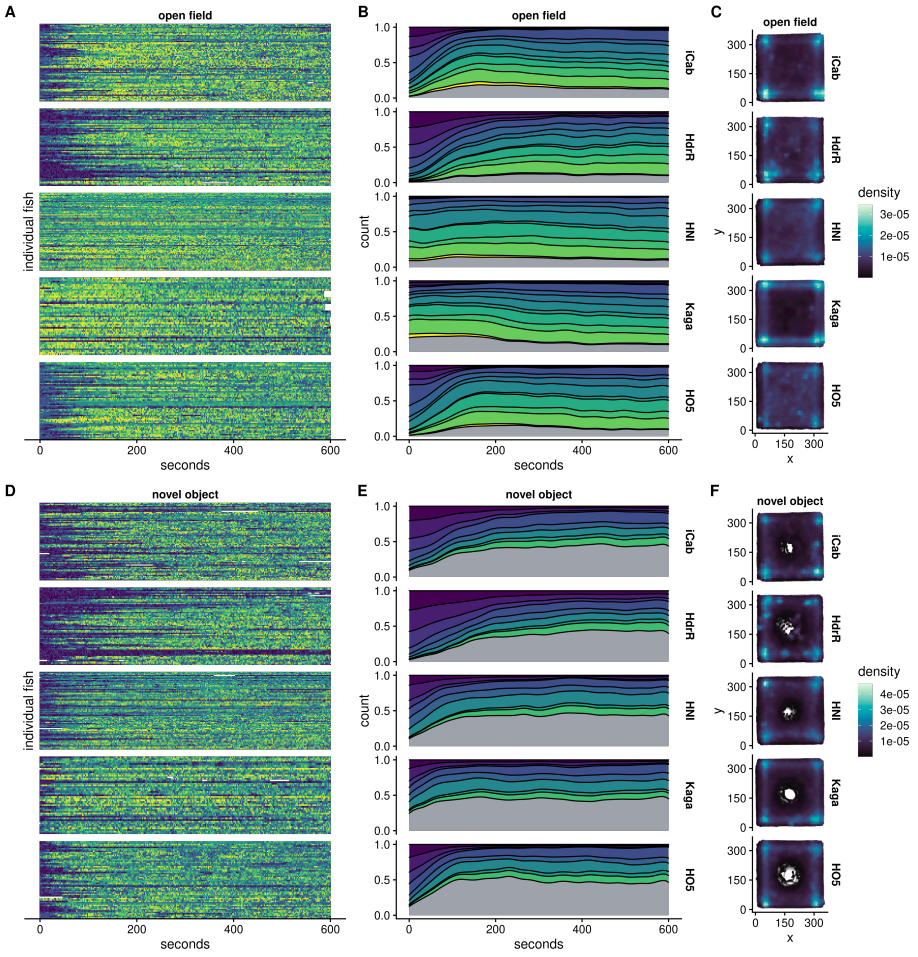


Figure 3.7: Differences between test fish strains in the HMM states they occupied during the open field (top) and novel object (bottom) assay components. **A and H**: 14 HMM states with panels coloured red to indicate significant differences between test fish strains in the proportion of time spent in those states during the separate assay components. **B and E**: Transitions between HMM states across time for each individual test fish, grouped by strain. Tiles are coloured by the state most frequently occupied by each fish within 2-second intervals. **C and F**: Densities within each strain for the occupation of states that significantly differed between strains (colour), with other states consolidated (grey). **D and G**: Densities of the test tank locations occupied by each strain, calculated within 900 grid points (30x30).

Table 3.2: Significant differences in the proportion of time spent in each HMM state by iCab reference fishes depending on the strain of their tank partner during the open field and novel object assay components.

Assay	State	Variance explained (%)	p-value (FDR-adjusted)
open field	1	8.52	5.88e-06
open field	2	9.09	2.04e-06
open field	3	7.97	1.77e-05
open field	5	3.92	4.29e-03
open field	12	2.97	2.11e-02
novel object	1	7.94	8.73e-06
novel object	2	9.38	4.76e-07
novel object	3	8.73	1.58e-06
novel object	6	4.76	1.23e-03
novel object	8	2.80	4.06e-02
novel object	10	3.81	8.43e-03

the test fish strains they are paired with (**Figure 3.8**). Thus, the *iCab* reference fishes spend less time in the slower-moving states when in the presence of the faster-moving northern Japanese strains *Kaga* and *HNI*, which in turn shows that the test fishes are transmitting their behaviours to some degree to their *iCab* tank partners.The differences in the behaviours of *iCab* reference fishes based on the strain of their tank partners are not as large as those between the test fish strains. However, unlike what was observed with the test fishes, here there is a greater number of states that show significant differences during the novel object component of the assay compared to the open field component. This suggests that when movement is restricted, or when in the presence of a potential threat, the behaviour of the test fish tank partner has more of an influence on the *iCab* reference fish's behaviour than otherwise.

Finally, to develop a metric to quantify the degree to which the *iCab* reference fishes' behaviour is influenced by the strain of its tank partner, we calculated the proportions of time that the pairs of fish spent simultaneously occupying the same HMM state. **Figure 3.9** shows that when paired with the slowest-moving *HdrR* strain, the *iCab* reference fishes spent more time co-occupying the slow- and

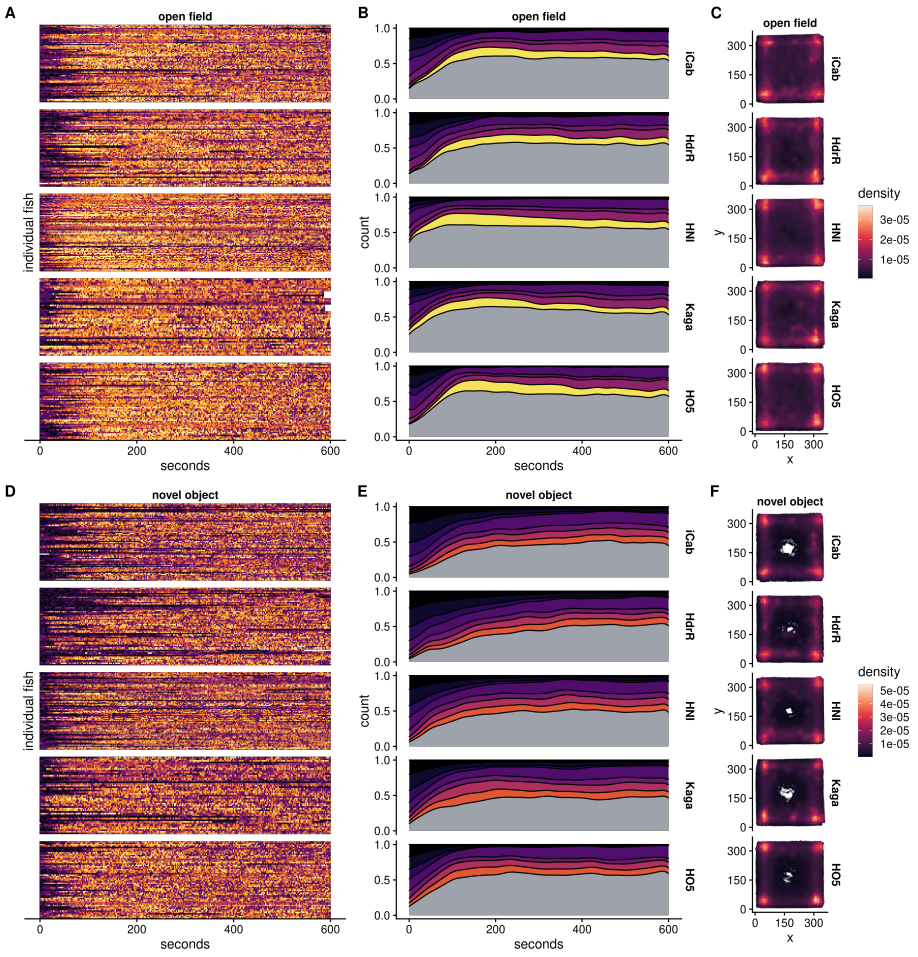


Figure 3.8: Differences between HMM states occupied by the reference fish when paired with different test fish strains during the open field (top) and novel object (bottom) assay components. **A and H**: 14 HMM states with panels coloured red to indicate significant differences between the reference fishes under different strain-pairings in the proportion of time spent in those states during the separate assay components. The HMM states are the same as those in **Figure 3.7**, but coloured with a different palette. **B and E**: Transitions between HMM states across time for each individual *iCab* reference fish, grouped by the strain of its tank partner. Tiles are coloured by the state most frequently occupied by each fish within 2-second intervals. **C and F**: Densities within each strain-pairing for the occupation of states that significantly differed between strain-pairings (colour), with other states consolidated (grey). **D and G**: Densities of the test tank locations occupied by the iCab reference fishes when paired with different strain, calculated within 900 grid points (30x30).

forward-moving state 2, whereas they tended to co-occupy the fast- and forward-moving state 11 when paired with the faster-moving northern *Kaga* and *HNI* strains. When paired with another *iCab*, they preferred to co-occupy the slow-moving and pan-directional state 4. For each combination of assay component and state, we then ran a Kruskal-Wallis test to determine whether there were differences in the frequencies of state co-occupancy under different strain pairings (p < 0.05, FDR-adjusted), and found significant differences for states 1 to 6 for the open field component (1.1x10-8 < p < 4.1x10-2), and states 1 to 4 and 7 for the novel object component (1.6x10-6 < p < 2.5x10-2). The heatmaps show that *iCab* tends to co-occupy state 4 at the highest frequency when paired with another *iCab*, which suggests that this pan-directional, slow-moving state is a more comfortable or natural state for *iCab* to occupy, particularly when under stress.

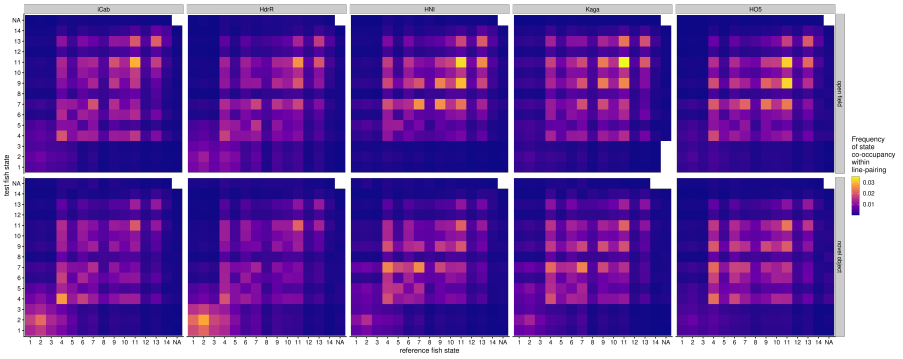


Figure 3.9: Frequency of HMM state co-occupancy between pairs of fish, calculated across all all videos per strain-pairing.

3.5 Discussion

In this chapter have described an assay for measuring bold-shy behaviours in medaka fish that can reliably detect differences: (a) between individuals from different inbred strains – allowing for the quantification of direct genetic effects on behaviour; and (b) between the behaviour of individuals when paired with tank partners from a

different strain – allowing for the quantification of social genetic effects on behaviour.

I found that behaviours exhibited during the assay show minimal variance across the covariates of date of assay, time of assay, tank quadrant, and tank side. This creates confidence that the genetic effects that we observed – both direct and indirect – are not unduly affected by these variables.

I have shown that a Hidden Markov Model (HMM) allows an adequate classification of movements based on the direction and angle of travel within set time intervals. The HMM method has previously been used to classify maternal behaviour patterns in mice (Carola, Mirabeau, and Gross 2011), and I have extended the application here to bold-shy behaviours in medaka fish. In principle, this method can be expanded to include additional behavioural features such as proximity to the wall or other objects, inwards/outwards orientation, proximity to the tank partner, and other metrics related to leader-follower dynamics, a possibility that will be tested in future studies.

The HMM classification of behaviours also facilitated the ranking of behaviours by speed and angle of movement, which may be further interpreted in terms of biology and adaptation styles. However, even though the HMM classifies the slowest-moving states as separate modes of movement, one should be cautious when attributing biological meaning to their differences, because even when a fish is almost completely still, the fish object's centroid (upon which the variables of distance and angle of travel are calculated) will tend to move by one or several pixels through minor changes in the segmentation of the object, and are thus subject to greater noise than the faster-moving states.

I have found that inbred strains of medaka fish can be distinguished by the proportion of time they spend in certain states. The slowest states, capturing no or minimal movement, were the states that most clearly separated the strains, and these differences appeared most clearly at the beginning of the open field and novel object assays. This suggests that for southern Japanese strains such as *iCab*, *HdrR*, and *HO5*, the "freeze" reflex is caused by anxiety, which even-

tually dissipates over time indicating habituation. Similar observations showing initial anxiety as shown by freezing behaviour and subsequent habituation have been made in other studies using open field assays with *iCab* fish (Lucon-Xiccato et al. 2022).

On the other hand, the northern Japanese strains *Kaga* and *HNI* spent little time in the slow-moving states at the beginning of the video, which indicates either that their habituation sets in earlier, or that their stress and anxiety is expressed to a lesser degree as freezing behaviour. The former appears to be more likely for *Kaga*, as in the open field assay, it spends more time in the faster-moving states at the beginning of the video and then slows down thereafter, which suggests that its higher level of movement may be induced by stress. It is interesting then that once the novel object is introduced, *Kaga* tends to move slowly like the other strains. Obviously introduction to a novel environment and exposure to a potentially dangerous object elicits different behavioural patterns in this strain. Taken together with the overt thigmotaxis behaviour (i.e. moving along the sides of the tank during the open field assay component) relative to other strains, its behaviour may suggest an "escape-seeking" response. It will be interesting to carry out further experiments to determine under which fear-inducing conditions *Kaga* reacts with this more frantic style of movement rather than the freeze response that it displays in the novel object component in common with the other strains, and to explore its physiological basis.

With respect to social genetic effects, we have found that a fish's behaviour is affected by the differential behaviour of its tank partner, although the effect is less powerful than the direct genetic effect on a fish's own biology. These social genetic effects are detectable when observing the proportions of time the reference fishes spend in certain states over the course of the video, and can be quantified by measuring the frequency of state co-occupancy with their tank partners.

It is also interesting to find that *iCab* prefers to co-occupy the slow, pan-directional state 4 when paired with another individual from the same strain. As the strains are raised exclusively in the company of individuals from the same strain, that type of behaviour may represent a more comfortable – yet still cautious – mode of movement,

or a type of mimicry that tends to occur when in the company of individuals that are more genetically or socially familiar.

In this study we only used one strain as the reference fish, but future experiments can expand on this analysis by using different strains as the reference, thereby exploring how a strain's genetics influence the degree of their behavioural plasticity, and how the distinct behaviours of strains may interact in intriguing ways.

In summary we show that the OFT experimental setup in combination with the HMM analysis can reliably detect both direct and social genetic effects. This creates the opportunity to carry out a similar study on a larger panel of inbred strains, such as the MIKK panel (Fitzgerald et al. 2022), with a view to identifying the genetic variants associated with not only these differences in behaviour, but also the differences in the degree to which an individual transmits their behaviour onto their social companions. This in turn will shed light on longstanding biological questions concerning the direct and indirect influences on behaviour, their physiological bases, and their adaptive purposes. This is the subject of the following chapter.

Chapter 4

Genetic linkage study of bold/shy behaviours in the MIKK panel

The purpose of the study described in this chapter was to run the behavioural analysis described in Chapter 3 over the MIKK panel described in Chapter 2, identify the lines that diverged in both their behaviour, and the level of transmission of their behaviour onto their *iCab* reference tank partner, and then use them as the parental strains in an F2 cross to attempt to identify the genetic variants associated with those differences.

4.1 The F2 cross experimental setup

The F2 cross is a traditional method for mapping genetic variants associated with traits of interest. A schema for the method is presented in **Figure 4.1**. In essence, it involves starting with two inbred strains that diverge for the trait of interest (the 'parental strains', or F0 generation F0). One then crosses the parental strains to create a generation of F1 hybrid individuals who each possess, for every pair

of their chromosomes, one chromosome from each of their parents. The individuals in this F1 generation are genetically identical to their parents with respect to their germ line. Finally, one inter-crosses the F1 generation to create a set of F2 individuals that share unique combinations of the original F0 strains' genotypes, and tend to display values for the trait of interest that span across the spectrum between the extreme values of their parents.

4.2 Data collection - F0 generation

In November 2019 I traveled to the fish facility managed by our collaborator, Felix Loosli at KIT in Karlsruhe, and over the course of 11 days from 11 to 21 November 2019, I ran the behavioural assay described in Chapter 3.4.1 another 206 times. I again used the *iCab* strain as the reference fish, and for the test fish I used either an individual from one of the MIKK panel lines, individuals captured from the same Kiyosu population as the MIKK panel but permitted to breed freely within a separate tank in the facility ('Kiyosu closed-capture', or 'Kiyosu CC'), or individuals from a related but different species of medaka from the Philippines, *Oryzias luzonensis*. I ensured that I performed at least 2 assay runs for each MIKK panel line that was available, generating a minimum of 8 test fish replicates per line. As there were four pairs of fish in the test tank during each run, the complete dataset comprises 824 videos of pairs of fish, which I further divided by assay component (open field and novel object) to create 1648 videos.

I again used the software *idtrackerai* (Romero-Ferrero et al. 2019) to track the movement of the fishes across frames of each video. After adjusting the software parameters for each video to maximise the number of frames that were successfully tracked, I was left with 1610 out of the 1648 videos (~97.7%) where both fishes were tracked over at least 85% of frames, and I only included these 1610 videos in the downstream analysis. The first question to address was whether the MIKK panel lines differed in their behaviours. I therefore computed each individual fish's mean speed (measured as the distance traveled

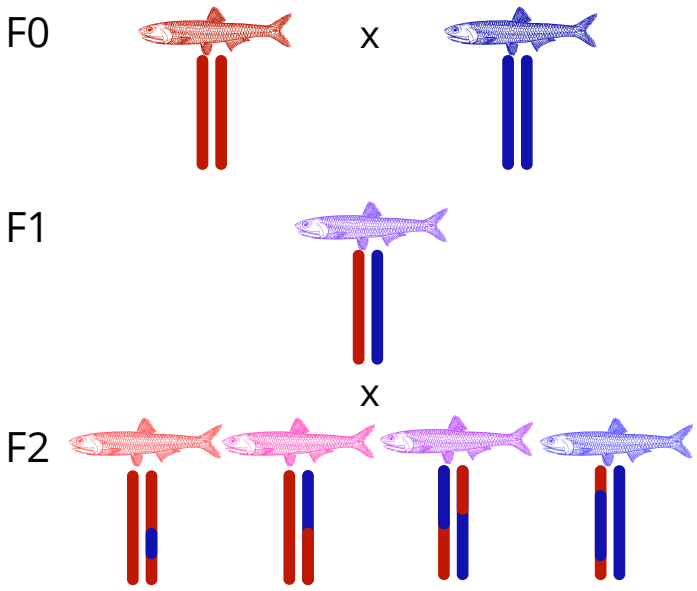


Figure 4.1: Schema of the F2-cross experimental setup. The F0 generation comprises two medaka strains that have extreme, opposing values for a trait of interest, represented by the colours red and blue. Below them is an illustrative single pair of chromosomes. The chromosomes within each pair are depicted as the same colour, as the strains are homozygous through successive generations of inbreeding. Their F1 offspring is heterozygous for each pair of their 24 chromosomes, and all F1 individuals are therefore almost genetically identical to one another (with the exception of somatic mutations and the regions of the genome that were not homozygous in the parental generations). The F1 individuals are then inter-crossed with one another to produce the F2 generation, which, due to recombination events during gamete formation, have unique combinations of the parental strains' genotypes, and tend to span the phenotypic spectrum between the extremes of their F0 parental strains, represented by their colours.

in pixels per 0.08 seconds) over the course of the full 20-minute video, grouped them by line, and plotted the results presented in **Figure 4.2**. I continue to use the same order and colour palette for the MIKK panel lines as in this Figure throughout the rest of this Chapter.

This figure shows that there are clear differences between some MIKK panel lines at the extremes, and that the lines differ in the amount of within-line variance observed. This figure acted as a guide to determine which lines to select as the parental strains in the F2 cross. To identify genetic variants directly associated with bold-shy behaviours, I sought to select lines that showed either high or low levels of movement, and preferably low within-line variance.

4.3 HMM states

To examine these behaviours at a finer resolution, as for the pilot study described in 3, I again applied a hidden markov model (**HMM**) to classify the fishes' movements based on their distance and angle of travel between time intervals. I used the same method to select the best choice of time interval and number of states (**Figure 4.4**). Here I observed the same phenomenon where the parameter combinations that performed the best showed an asymmetry between some states that would make interpretation difficult. For example, a time interval of 0.08 seconds combined with a state space of 17 caused state 4 to appear to get carved out of state 3 (**Figure 4.5**).

The best combination of parameters without this asymmetry was a time interval of 0.05 seconds with a state space of 15 (see the polar plots for the states in **Appendix @ref(fig:hmm-states-0.05)**). However, due to a glitch in the video recording software, several videos recorded on 13 November 2019 were incorrectly recorded with a frame rate of 14 instead of the desired 30. The insufficient number of frames for those videos meant that it was impossible to measure the distance and angle of travel between a time interval as low as 0.05 seconds. So that these videos could be included in the dataset, I accordingly selected the combination of 15 states and a 0.08-second

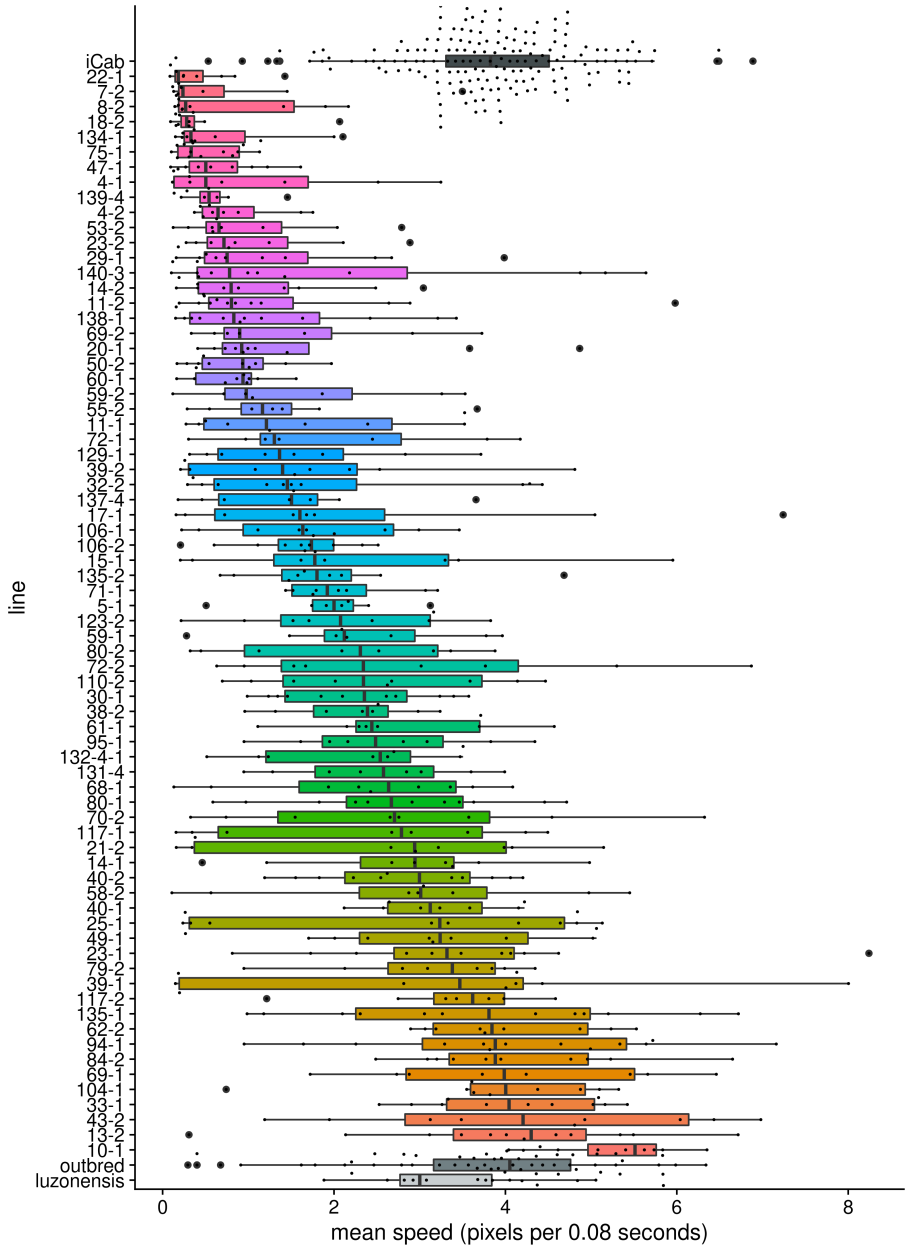


Figure 4.2: Mean speed of the MIKK panel and other strains over the course of the entire 20-minute video (measured as the distance traveled in pixels per 0.05 seconds). *iCab* fishes in the *iCab*-*iCab* control condition are at the top, the MIKK panel lines are sorted by their group median, and the Kiyosu closed capture and *O. luzonensis* fishes are at the bottom.

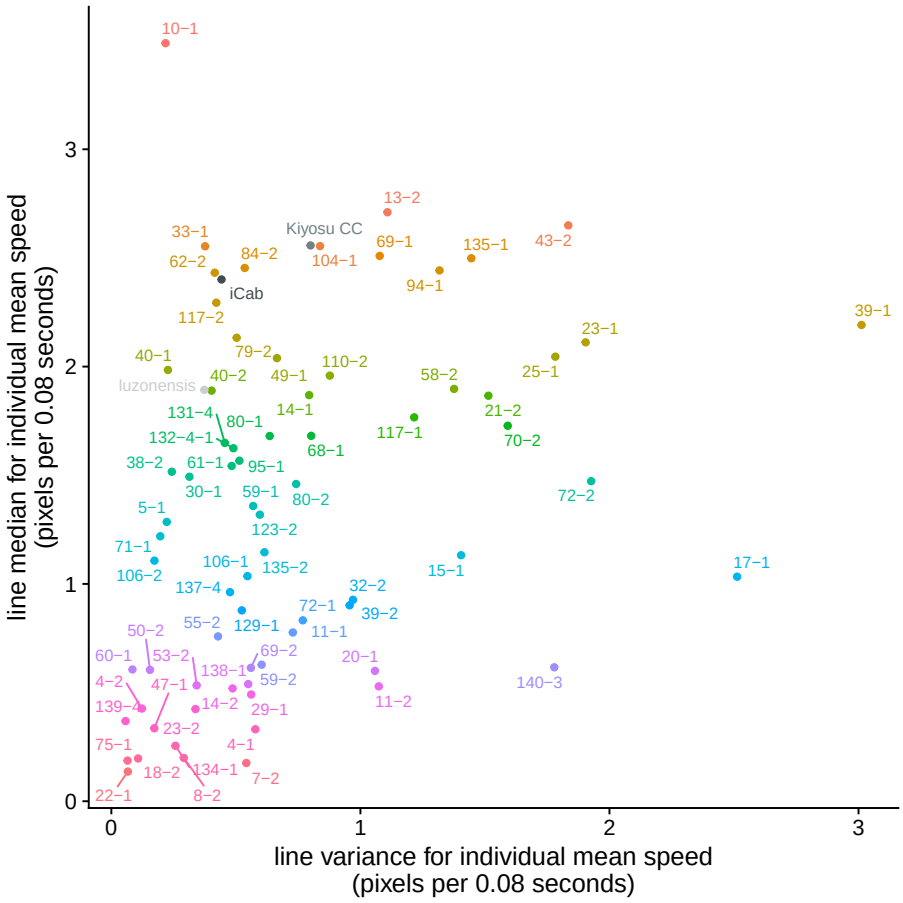


Figure 4.3: Line median (vertical axis) and line variance (horizontal axis) for individual mean speed across the full 20-minute video (i.e. both the open field and novel object assay components).

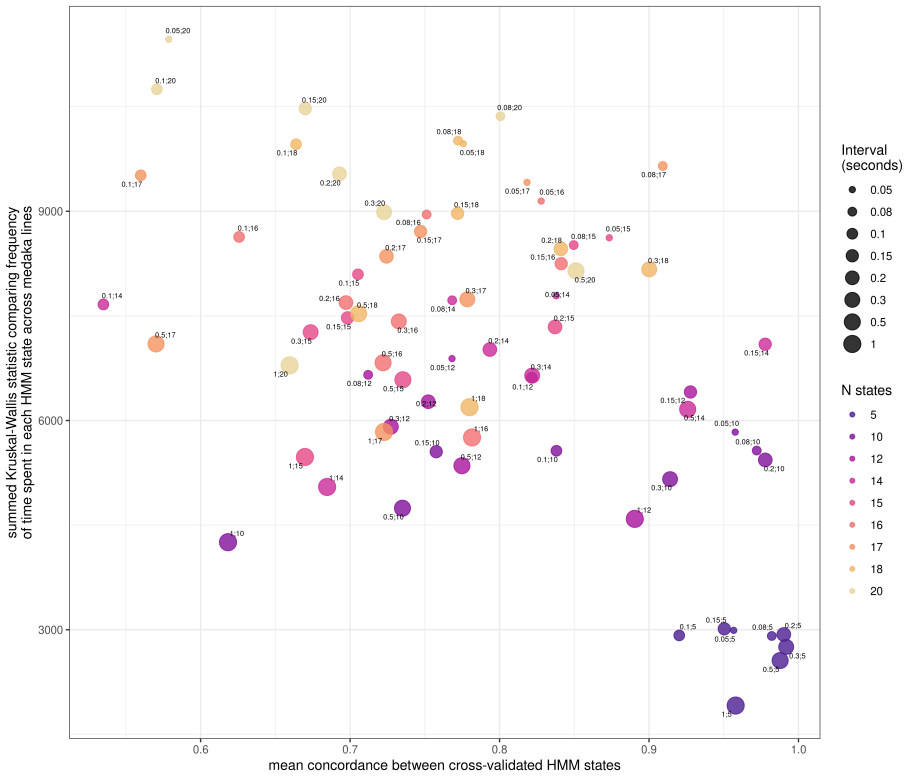


Figure 4.4: Comparison between HMM parameters. Horizontal axis: Mean concordance between states assigned by HMMs through a 2-fold cross-validation process. Vertical axis: Kruskal-Wallis statistic comparing strains based on the proportion of time spent in each HMM state, summed across all states. Size of points correspond to the interval, in seconds, between which the distance and angle of travel was calculated (Methods).

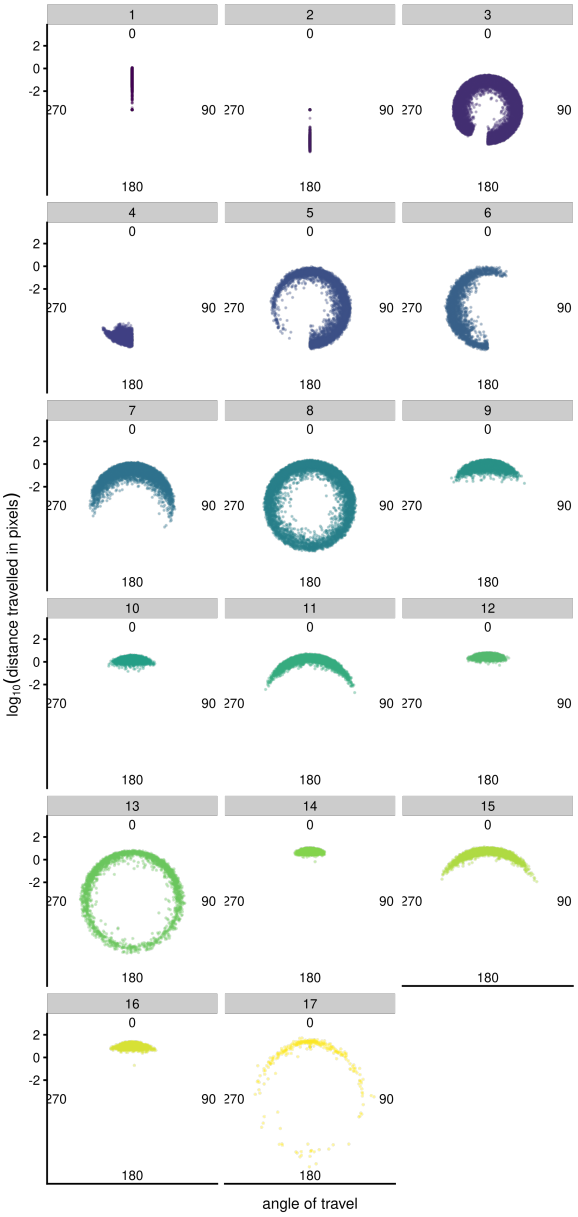


Figure 4.5: The best apparent combination of parameters (0.08 time interval with a 17-state space) created an asymmetry between states 3 and 4, which would causes difficulties in interpreting their biological relevance.

interval for all downstream analyses.

4.4 Social genetic effects

As discussed above, in this project our traits of interest include not only direct genetic behaviours, but also social genetic behaviours. We therefore sought to identify the MIKK panel lines that transmitted their behaviour onto the reference *iCab* tank partners either to the largest or smallest degrees. I formulated two methods to measure this, referred to as a) HMM state co-occupancy; and b) reference deviation. The first, HMM state co-occupancy, measured the proportions of time that the *iCab* reference fish spent in the same HMM state as its tank partner. The second, deviation of the reference fishes' deviation from the behaviour exhibited in the control condition, seeks to quantify the extent to which the *iCab*'s behaviour changes when partnered with

4.4.1 State co-occupancy

Figure 4.7 sets out the proportions of total time for each assay subcomponent that each pair of individual fish spent in the same HMM state, grouped by line and ranked in the same order as their group median for individual mean speed as shown above in **Figure 4.2**. **4.7A** shows the data as boxplots, with p-values from the Kruskal-Wallis test comparing all groups. **Figure 4.7B** shows the same data but with each group's median represented by columns to make it easier to compare group medians. Of the slower-moving lines, 8-2 and 18-2 tend to show relatively higher state co-occupancy in the open field component, but during the novel object component, the slow-moving line 139-4 has the highest median co-occupancy of all lines. Of the faster-moving lines, 43-2 and 13-2 showed the highest state co-occupancy during the open field assay component. However, the moderate-to-fast line 21-2 had relatively high state co-occupancy during both assay components.

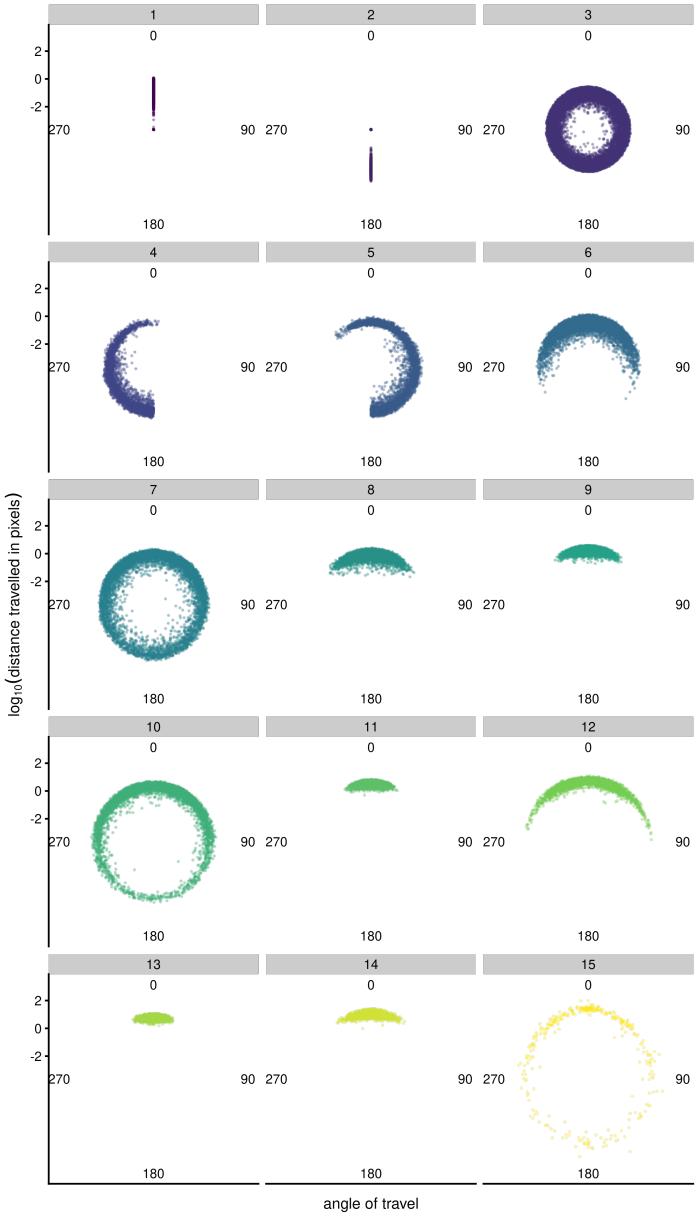


Figure 4.6: The HMM states used for the downstream analysis, with the model classified based on the distance of travel ($\log_{10}$ pixels, radial axis) and angle of travel (angle). A straight forward movement would sit around 0°, a left movement around 270°, and a right movement around 90°.

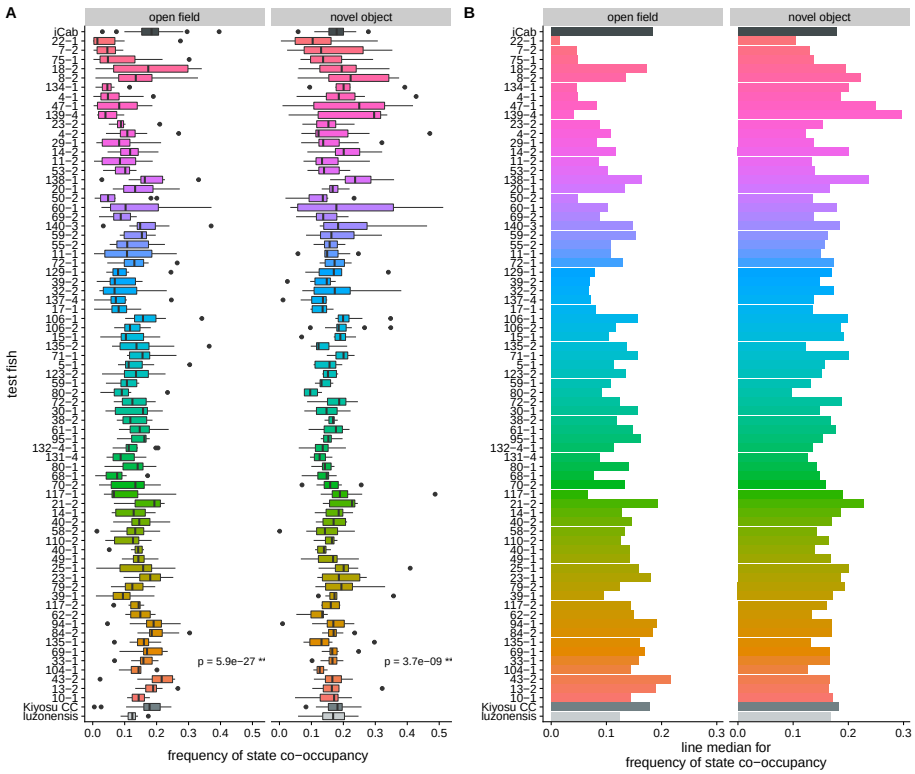


Figure 4.7: Frequency of HMM state co-occupancy

To visualise which states are driving the higher co-occupancy measures, for a selection of lines I generated a heatmap of the states occupied simultaneously by the *iCab* reference and MIKK test fishes, combining the observations for all individuals within each test fish group (**Figure 4.8**). When *iCab* is paired with 18-2 or 8-2, the fishes most frequently occupy states 3 or 1 in both open field and novel object components. In pairings with line 139-4, while the test fishes remain in the still states 1 or 3 during the open field assay, *iCab* tends to be moving much faster in states 11 and 13. However, during the novel object component, they co-occupy state 3 more than in any other combination. This general pattern is observed with line 14-2 as well, albeit to a lesser extent. For 38-2, the fishes tend not to show a strong preference for co-occupying a particular state for either assay component, but the diagonal spread indicates that they tend to move at similar speeds. When paired with the faster moving 21-2, the novel object component appears to accentuate the co-occupancy of state 3 that is also observed in the open field component. Finally, when paired with line 40-1, in both assay components, both fishes show a strong preference for the faster-moving states.

4.4.2 Deviation of *iCab* from its control condition

The second method for quantifying the level of behavioural transmission from test fish to *iCab* reference fish was to determine the proportion of time that the *iCab* spent in a particular state when paired with another *iCab*, and then quantify the degree to which those proportions change when in the presence of a fish from another line (**Figure 4.9**). **Figure 4.9A** presents boxplots for state frequencies for all *iCab* individuals in *iCab*-*iCab* pairings. I further calculated the state frequencies for all *iCab* reference fishes for all other MIKK line pairings. For each combination of assay component, line-pairing, and HMM state, I then ran Welch's t-test (Ruxton 2006) comparing the proportions of time the *iCab* individuals spent in that state when paired with another *iCab*, against the proportions of time the *iCab* reference individuals spent in that state when paired with a different MIKK line. I then summed the t-statistics across states to generate

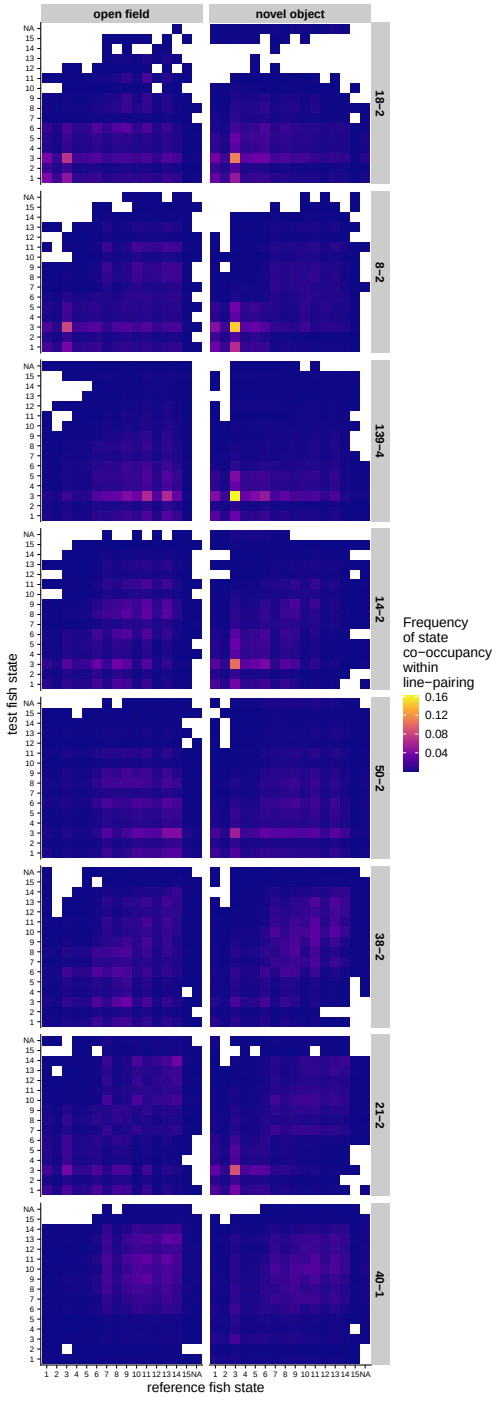


Figure 4.8: Heatmaps for a selection of MIKK panel lines (including those ultimately selected as the parental strains in the F2 cross) showing the frequency of HMM states simultaneously occupied by the reference (x-axis) and test (y-axis) fishes, aggregated over all replicates in each line.

a single metric for each combination of line and assay component, and plotted the results in **Figure 4.9B**.

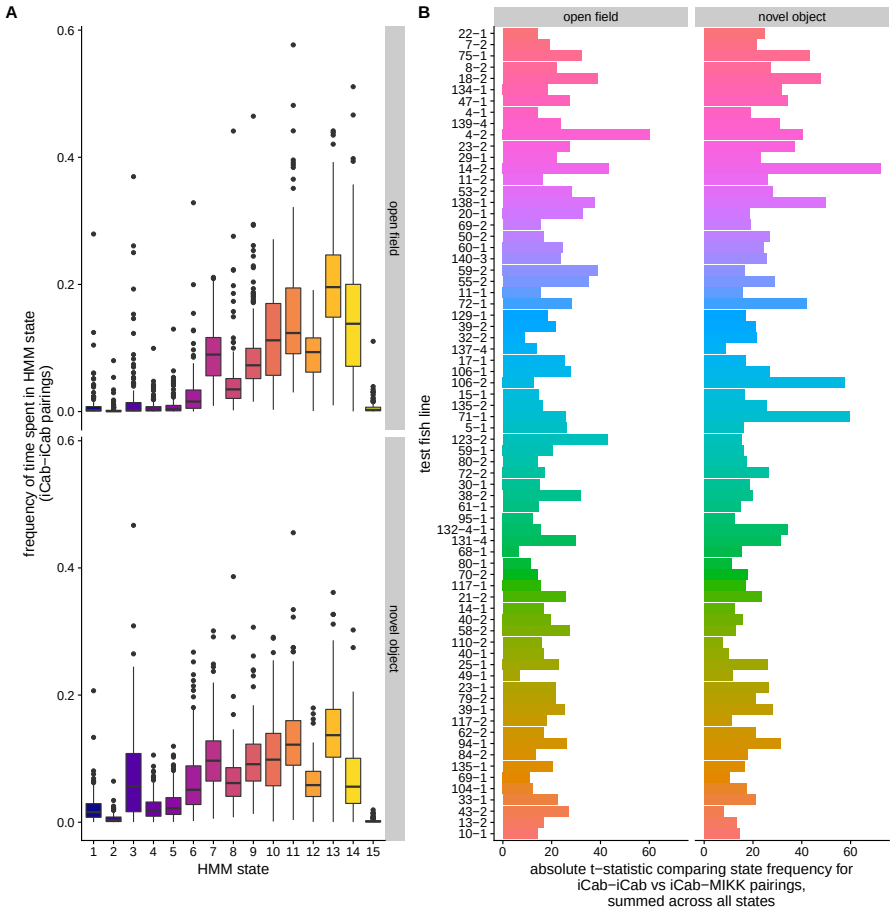


Figure 4.9: Deviation of state frequency for *iCab* reference fishes when paired with MIKK panel lines relative to when paired with another *iCab*. **A**: Boxplots of HMM state frequency for *iCab* individuals when paired with another *iCab*. **B**:

The first thing to note in this Figure is that *iCab*'s deviations from its behaviour exhibited in the control condition tend to be smaller when paired with faster-moving fishes. This was expected given *iCab* is also a faster-moving fish, but it strengthens the observation that when paired with slower-moving fishes, those tank partners are causing the *iCab* reference fish to move slower than they would

otherwise. Of the slower-moving MIKK lines, 4-2 and 14-2 causing the largest deviations of the *iCab* reference fishes' behaviours from their control condition behaviour in the open field and novel object components respectively. I also observe large deviations for the moderately-moving 106-2 and 71-1 during the novel object components. There are no clear outliers for either assay component for any of the faster moving lines.

4.5 Selection of lines for the F2 cross

On the basis of the above findings, I selected 6 MIKK panel lines as the parental lines for the F2 cross (**Figure @ref(F0-line-mean-speed-select**). Conceptually, I sought to select lines that diverged on two measures: a) bold-shy behaviours; and b) the extent to which the lines transmitted their behaviours onto their tank partners. As *iCab* is a fast-moving line, it was more difficult to detect instances where fast-moving MIKK panel lines were influencing its behaviour. I was therefore more confident of identifying slow-movement/high-charisma lines, and so to increase the likelihood of identifying genetic variants that are responsible for a stronger transmission of slow-moving behaviours, I chose two slow-movement/high-charisma lines for the F2 cross: 18-2 and 8-2. In the event that both lines possessed the same genetic variants that influence this behavioural transmission trait, it would vastly increase the power of detecting it during the genetic linkage analysis. Both these lines were one of the most slow-moving lines, had high levels of state co-occupancy during both assay components, and 18-2 caused a high level of deviation from the *iCab* control condition.

For the slow-movement/low-charisma line, I selected 50-2 which exhibited a moderately slow level of movement, low within-line variance for mean speed, and low measures for both state co-occupancy and *iCab* deviation. For the high-movement/high-charisma line, I selected 21-2. Despite its high within line variance (**Figure 4.12**), this potentially made it easier to detect its social genetic effects, as the slower-moving individuals appeared to transmit those behaviours

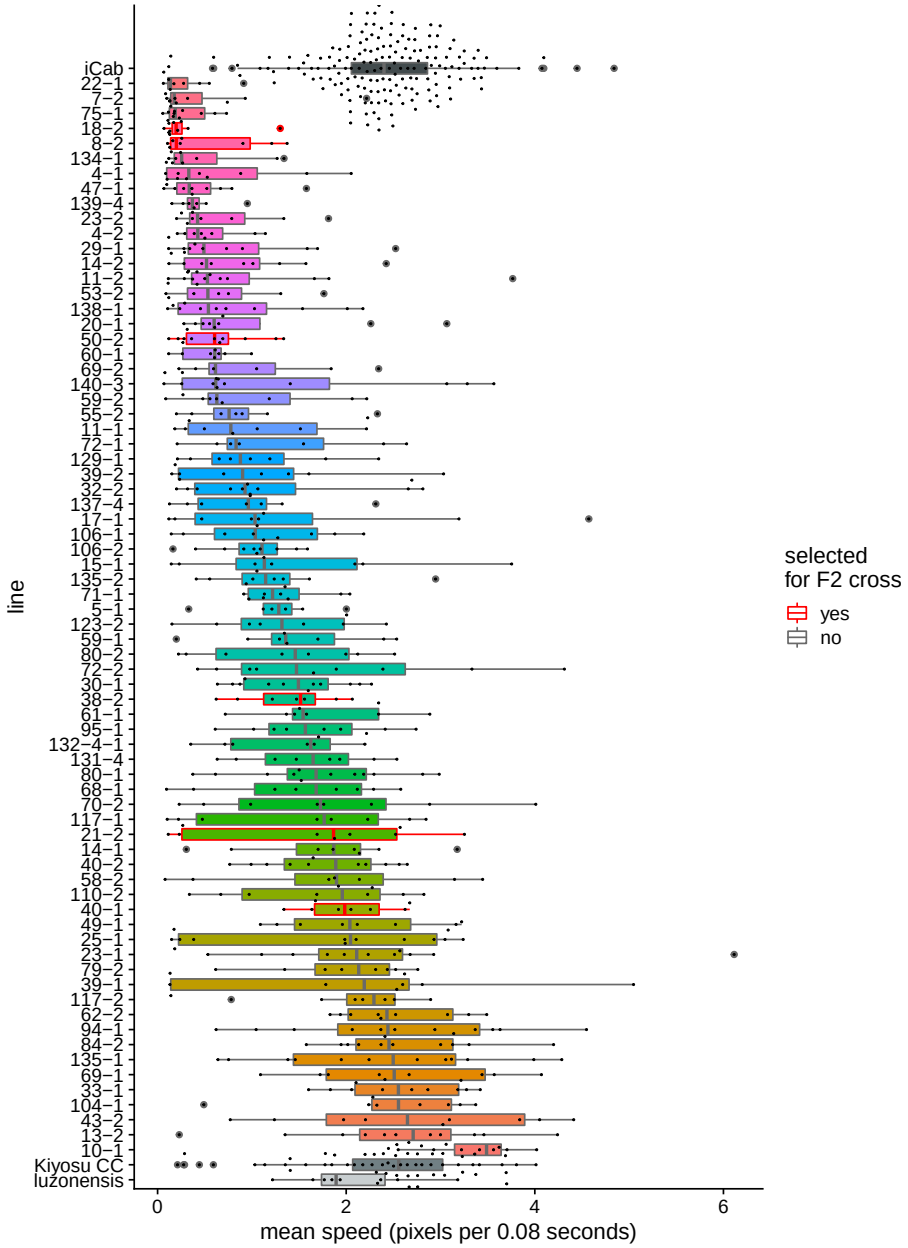


Figure 4.10: Mean speed of the MIKK panel and other strains over the course of the entire 20-minute video (measured as the distance traveled in pixels per 0.05 seconds), as shown above in Figure 4.2 but now higlighting the MIKK lines that were selected for the F2 cross.

strongly to their *iCab* tank partners, giving it a high score among fast-moving lines for state co-occupancy across both assay components. For the high-movement/low-charisma line I selected 40-1, as it has low within-line variance, and low-to-moderate metrics for state co-occupancy and *iCab* deviation. However, I note that these measures may be confounded by the possibility that 40-1 behaves in a similar way to *iCab*, which would be difficult to determine whether it was behaving differently.

In addition to these extreme lines, I also selected a line that was intermediate for both traits in an attempt to avoid breeding incompatibilities that might arise from attempting to cross lines with such divergent behavioural traits. For this purpose I selected line 38-2 for its intermediate speed, low within-line variance for mean speed, and intermediate measures for HMM state co-occupancy and *iCab* deviation.

4.6 Direct genetic effects

With these lines selected, I ran a similar analysis to what I described in 3, where I ran multi-way ANOVAs to determine whether certain lines differed in the proportions of time they spent in these HMM states, while including the date of assay, time of assay, tank quadrant and tank side as covariates. **Table 4.1** sets out the states which showed a significant difference between these 6 lines, with p-values adjusted for the False Discovery Rate (**FDR**).

Figures 4.12 and **4.13** highlight the states that showed significant differences in the proportions of time that these lines spent in those states for the open field and novel object assay components respectively. In both figures, **A** highlights the significant states, **B** shows how the individuals moved through those states over the course of the 10-minute assay component, and **C** shows the densities of the significant states within each line. For the open field component, although the tile plots show a notable level of variance within each line, the density plots clarify how the three slow-moving lines (8-2, 18-2 and 50-2) spent little time in the fast-moving states 10, 12 and 14

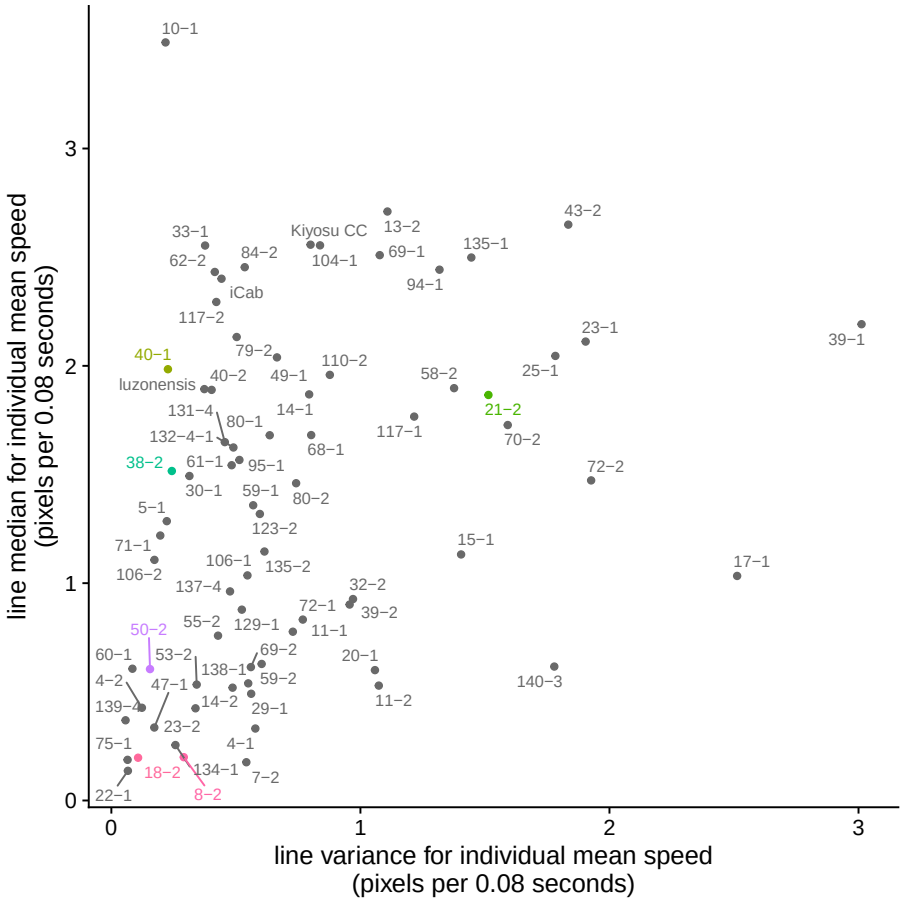


Figure 4.11: Line median (vertical axis) and line variance (horizontal axis) for individual mean speed across the full 20-minute video (i.e. both the open field and novel object assay components) as shown above in Figure 4.3, now coloured only for the lines selected as the parental strains for the F2 cross.

Table 4.1: Significant differences in the proportion of time spent in each HMM state across test fish lines selected for the F2 cross for the open field and novel object assay components.

Assay	State	Variance explained (%)	p-value (FDR-adjusted)
open field	3	21.53	2.67e-03
open field	4	23.28	1.60e-02
open field	5	20.61	3.15e-02
open field	10	29.06	1.41e-03
open field	12	24.91	1.45e-02
open field	14	29.52	1.29e-03
novel object	1	16.21	1.90e-02
novel object	2	15.14	4.86e-02
novel object	4	26.90	5.21e-03
novel object	5	27.84	4.27e-03
novel object	10	23.73	2.81e-02

relative to the fast-moving lines. Interestingly, the intermediate line 38-2 tended to transition into these states around the middle of the video, presumably a period of habituating to the new environment.

During the novel object component the HMM states that showed significant differences between lines were mostly restricted to the slow-moving states with the exception of state 10, which most clearly distinguishes the slow-moving lines from the intermediate- and fast-moving lines. Again, the intermediate line 38-2 shows a sharp drop in the occupation of the slow moving states after a period of habituation.

4.6.1 Social genetic effects

To confirm whether the *iCab* reference fishes altered their behaviour depending on the MIKK F0 line of their tank partner, we carried out the same analysis and model as above using only data from the *iCab* reference fishes. The states that showed a significant difference across any of the variables included in the model are set out in **Table 4.2**. The iCab reference fishes differed significantly in the proportion of time they spent in a given state depending on the strain of their tank partner ($p < 0.05$, FDR-adjusted) for 4 out of 15 states in

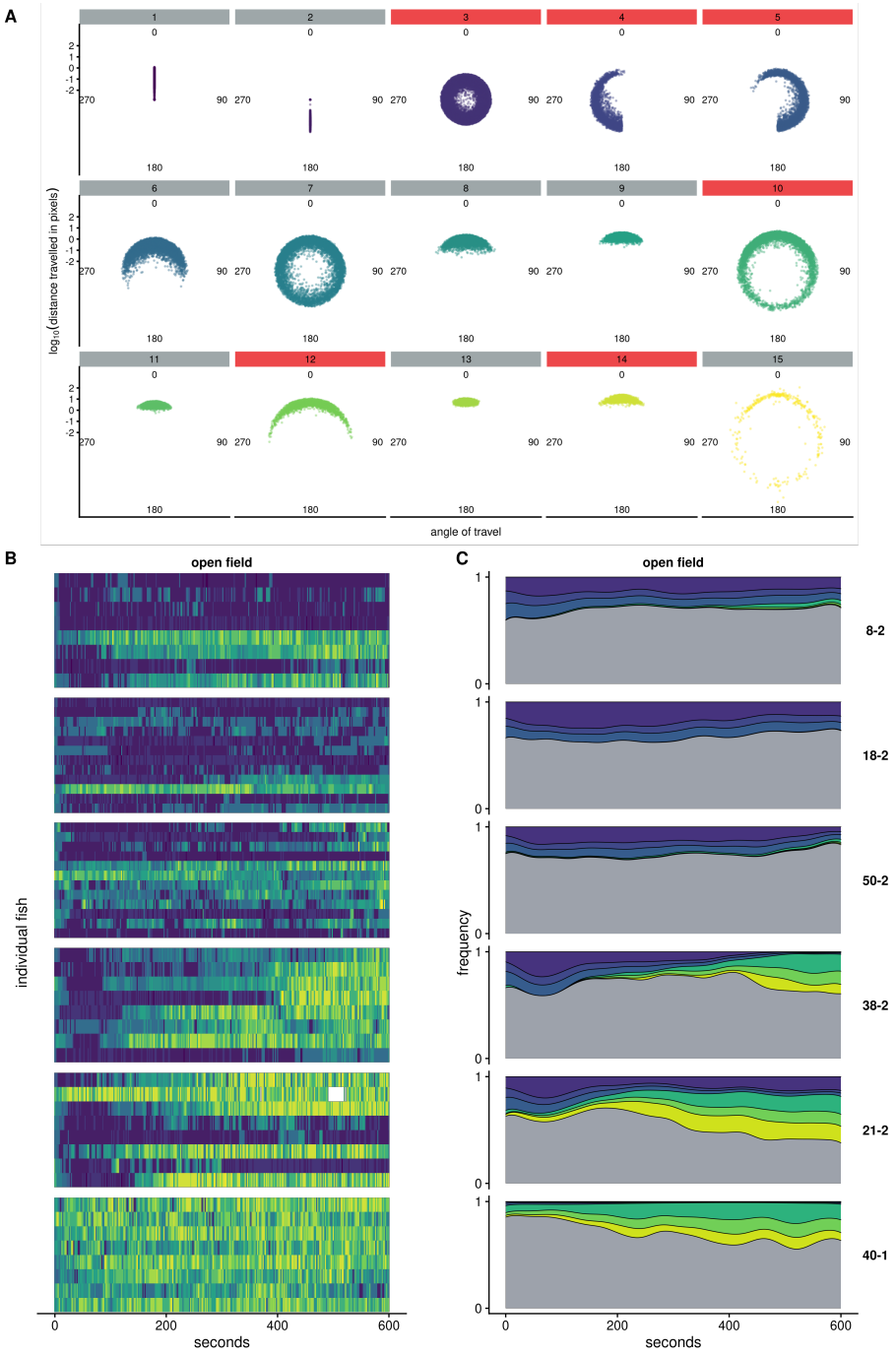


Figure 4.12: Differences between MIKK F0 lines in the HMM states they occupied during the open field assay component. **A**: 15 HMM states with panels coloured red to indicate significant differences between MIKK F0 lines in the proportion of time spent in those states. **B**: Transitions between HMM states across time for each individual test fish, grouped by MIKK line Tiles are coloured by the state most frequently occupied by each fish within 2-second intervals. **C**: Densities within each MIKK line for the occupation of states that significantly differed between strains (colour), with other states consolidated (grey).

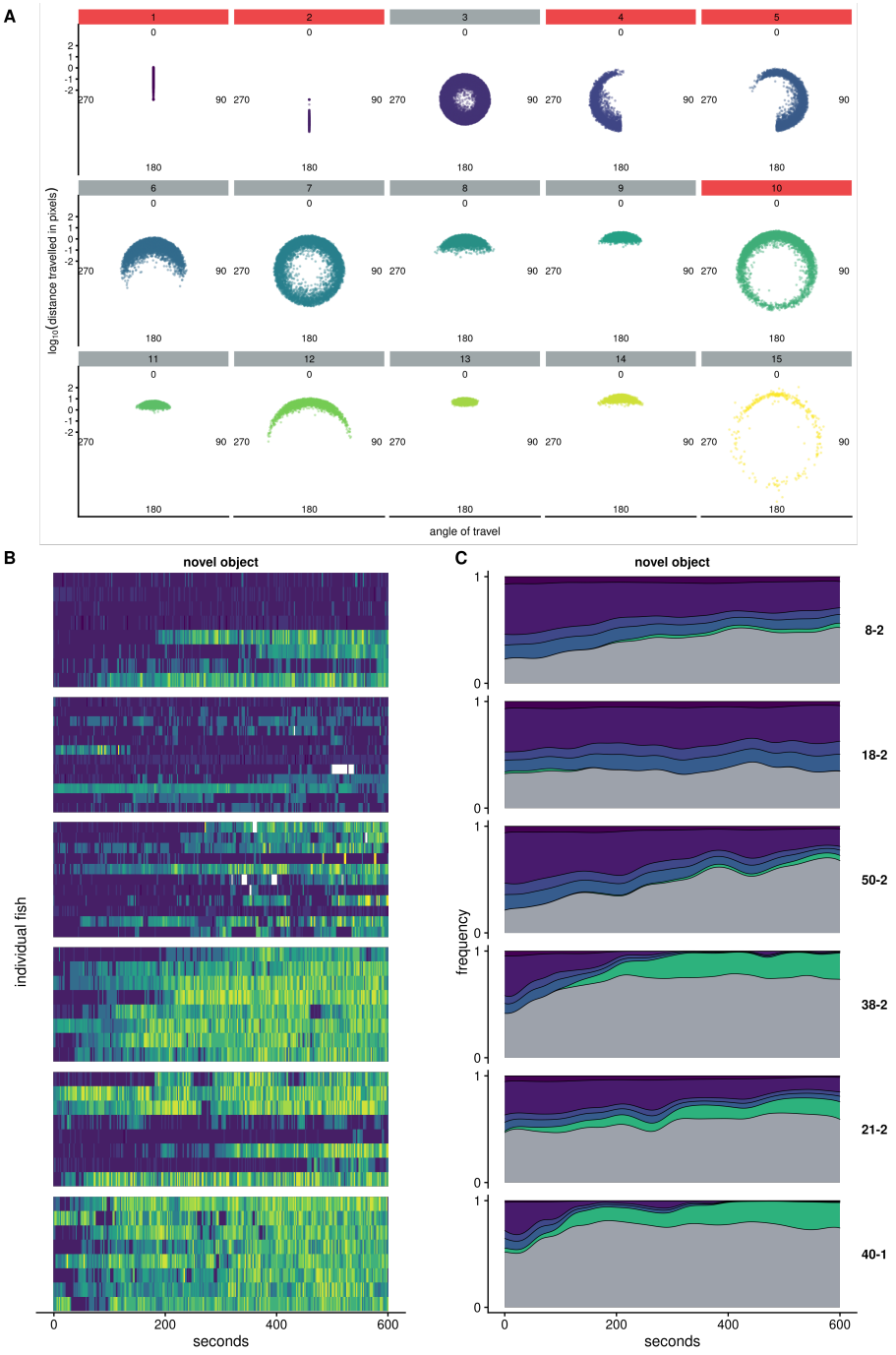


Figure 4.13: Differences between MIKK F0 lines in the HMM states they occupied during the novel object assay component. **A**: 15 HMM states with panels coloured red to indicate significant differences between MIKK F0 lines in the proportion of time spent in those states. **B**: Transitions between HMM states across time for each individual test fish, grouped by MIKK line Tiles are coloured by the state most frequently occupied by each fish within 2-second intervals. **C**: Densities within each MIKK line for the occupation of states that significantly differed between strains (colour), with other states consolidated (grey).

Table 4.2: Significant differences in the proportion of time spent in each HMM state by iCab reference fishes depending on the MIKK F0 line of their tank partner during the open field and novel object assay components.

Assay	State	Variance explained (%)	p-value (FDR-adjusted)
open field	3	14.93	2.10e-02
open field	4	13.60	2.28e-02
open field	5	15.33	1.26e-02
open field	7	14.26	4.00e-02
novel object	4	14.33	1.23e-02
novel object	5	14.38	1.15e-02
novel object	9	24.61	1.34e-02
novel object	11	28.77	7.71e-03

the open field assay ($1.26\text{x}10^{-2} < p < 4\text{x}10^{-2}$), and 4 out of 15 states for the novel object assay ($7.71\text{x}10^{-3} < p < 1.34\text{x}10^{-2}$). The line of the tank partner explained up to ~29% of the variance in the proportion of time the *iCab* reference spent in a given state. Full tables for all states and variables are provided in the Supplementary Material.

4.7 F2 generation

4.7.1 Behavioural data collection

Around August 2019, our collaborators in the Loosli Group at KIT commenced the breeding program for this experiment with the 6 MIKK panel lines I had selected above.[1] From 17 November 2021 to 5 May 2022, a Research Assistant in Prof. Loosli's lab, Alicia Günthel, performed the assay 69 time with F2 individuals from the 12 crosses they had bred, producing a total of 271 videos of pairs of fish. The

[1]Although the breeding program began in 2019, I understand from our collaborators that from around mid-2020 the F1 generation was producing poorly. After lengthy investigations, they discovered that the Covid pandemic had caused disruptions to the supply chains of their fish food manufacturer, which had compelled the manufacturer to modify the recipe. This change in the fish food recipe was the cause of the poor breeding. While the issue has since been resolved, it has resulted in a much smaller number of F2 individuals to be produced than was originally anticipated.

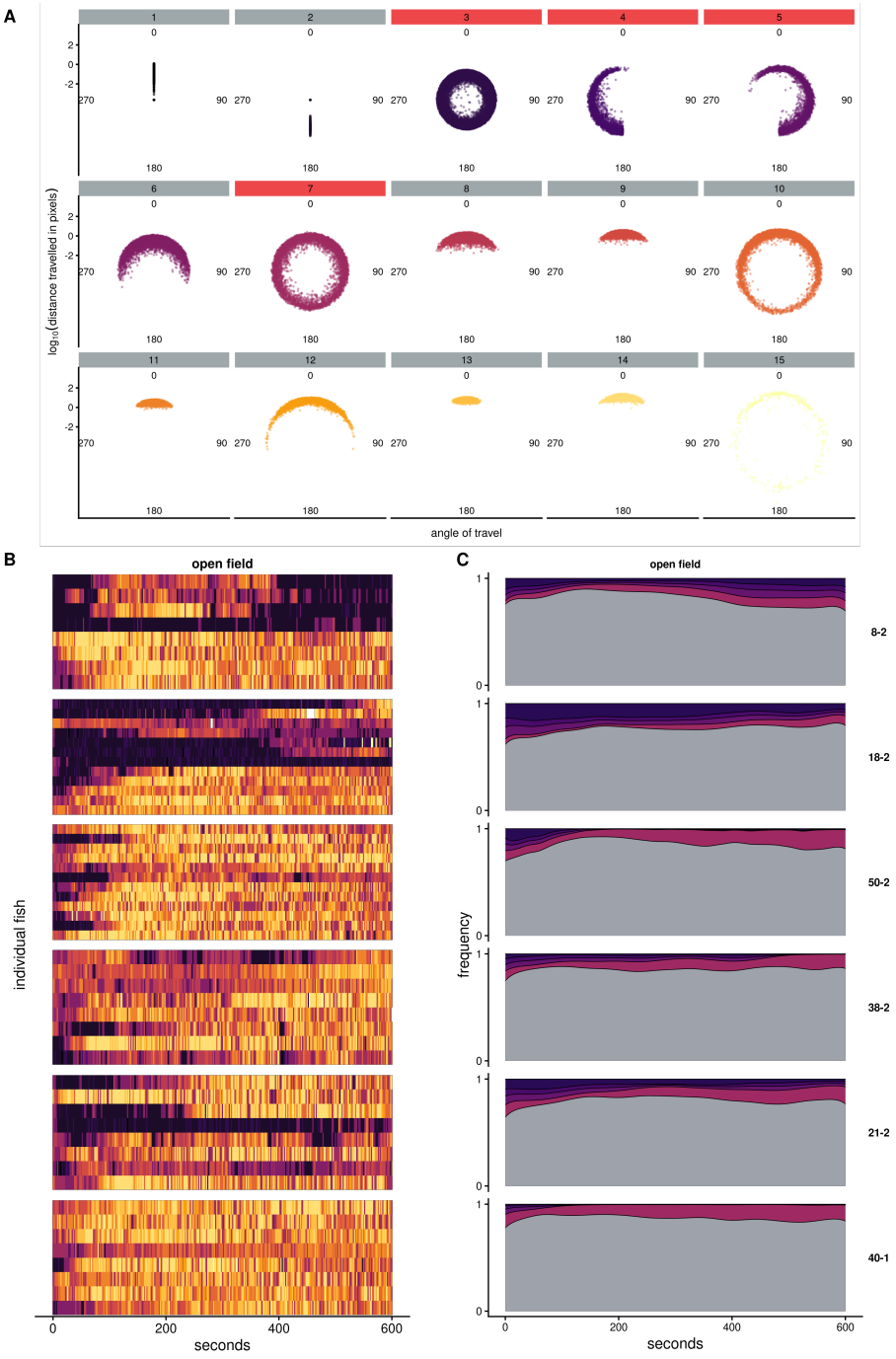


Figure 4.14: Differences between MIKK F0 lines in the HMM states they occupied during the open field assay component. **A**: 15 HMM states with panels coloured red to indicate significant differences between MIKK F0 lines in the proportion of time spent in those states. **B**: Transitions between HMM states across time for each individual test fish, grouped by MIKK line Tiles are coloured by the state most frequently occupied by each fish within 2-second intervals. **C**: Densities within each MIKK line for the occupation of states that significantly differed between strains (colour), with other states consolidated (grey).

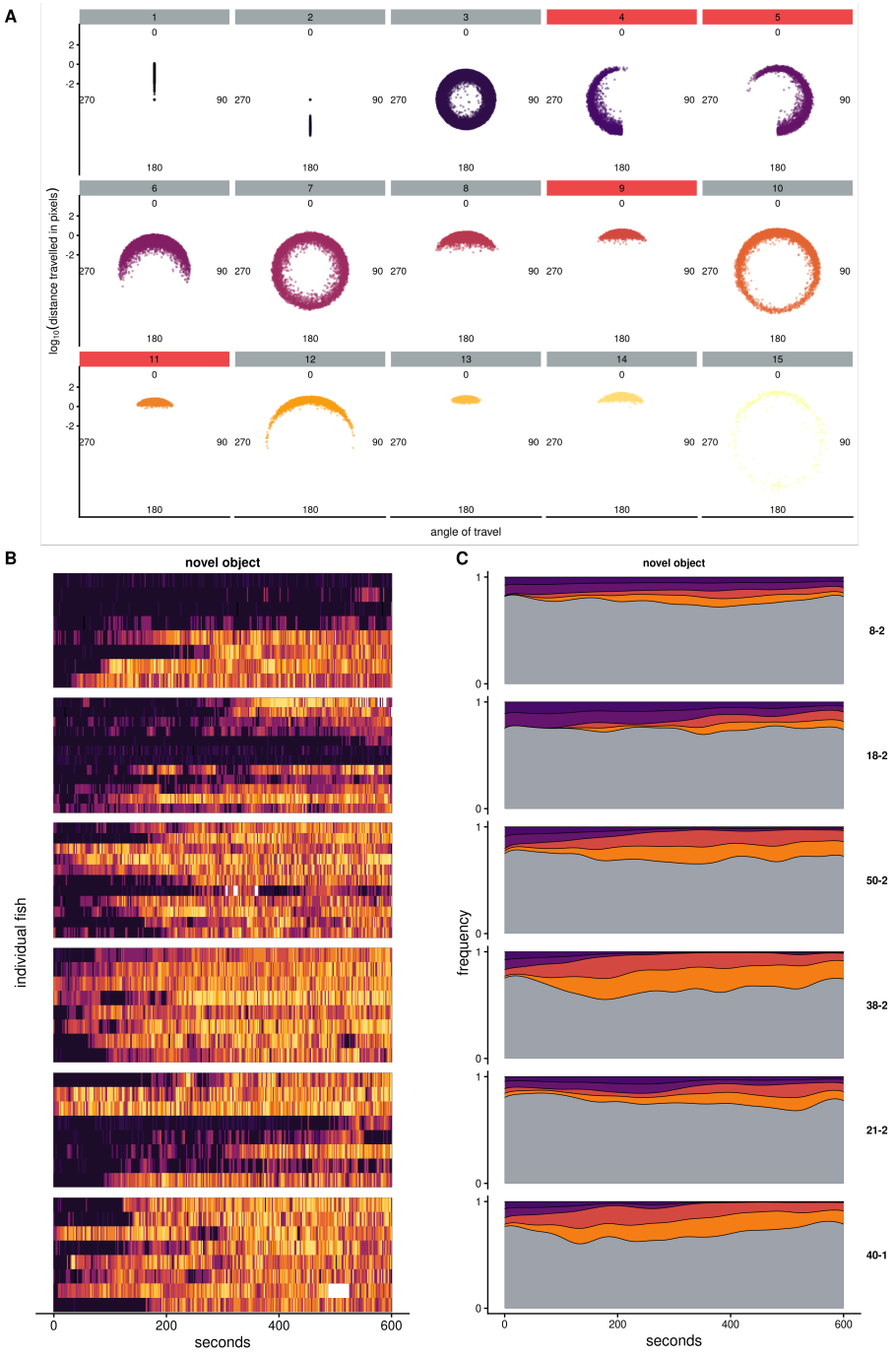


Figure 4.15: Differences between MIKK F0 lines in the HMM states they occupied during the novel object assay component. **A**: 15 HMM states with panels coloured red to indicate significant differences between MIKK F0 lines in the proportion of time spent in those states. **B**: Transitions between HMM states across time for each individual test fish, grouped by MIKK line Tiles are coloured by the state most frequently occupied by each fish within 2-second intervals. **C**: Densities within each MIKK line for the occupation of states that significantly differed between strains (colour), with other states consolidated (grey).

Table 4.3: Significant differences in the proportion of time spent in each HMM state by iCab reference fishes depending on the MIKK F0 line of their tank partner during the open field and novel object assay components.

paternal line	maternal line	count
21-2	40-1	60
38-2	40-1	57
38-2	18-2	35
8-2	40-1	24
50-2	18-2	23
38-2	21-2	19
8-2	38-2	15
50-2	38-2	12
18-2	21-2	7
21-2	50-2	7
40-1	50-2	6
50-2	8-2	6

counts for the 12 crosses used to generate these 271 test fishes are set out in **Table 4.3**.

I again used *idtrackerai* (Romero-Ferrero et al. 2019) to track the F2 individuals across frames. When splitting the 271 videos of pairs of fish into their open field and novel object assay components for a total of 542 videos, for 526 of them both fish were tracked across at least 85% of frames. I only used these 526 videos in the downstream analysis, but prior to publication I will seek to improve the tracking process so that these individuals can be included.

4.7.2 Behavioural phenotyping

To ensure that the predicted HMM states for behaviour were consistent across both F0 and F2 generations, I had included these F2 individuals for training and prediction in the models described above. I again calculated the proportions of time that every individual spent in each state ("**state frequency**"), then inverse-normalised the values within each combination of assay component (open field or novel object) and state (1 to 15). Inverse-normalisation is a rank-based nor-

malisation approach which involves replacing the values in the phenotype vector with their rank (where ties are averaged), then converting the ranks into a normal distribution with the quantile function (Wichura 1988). The inverse-normalisation function I used for this analysis is set out in the following R code:

```
invnorm = function(x) {
  res = rank(x)
  # The arbitrary 0.5 value is added to the denominator below to avoid 'qnor
  res = qnorm(res/(length(res)+0.5))
  return(res)
}
```

Figures 4.16 and **4.17** compares the phenotype pre- and post-transformation with this normalisation approach. For the higher-movement states there are increasing numbers of individuals who spent no time in those states, which are responsible for the apparently non-normal distributions observed for the skewed distributions post-transformation. States 1, 3, 6, 8, 9 and 11 already appear to have a large amount of variation between individuals, but this normalisation process will increase the variance for states where there is small, yet potentially meaningful, variation between individuals. One exception may perhaps be state 15, which involves very large distances of travel in all directions, and therefore may correspond to tracking errors.

4.7.3 Genotyping

After phenotyping the F2 samples, our collaborators in the Loosli Group at KIT took finclips from the F2 individuals, extracted their DNA, and arranged for them to be shallow-sequenced on the Illumina short-read platform at a coverage of ~1x per sample. Using a similar method to what I described in Chapter 5, I aligned these sequences to the *HdrR* reference with BWA-MEM2 (Vasimuddin et al. 2019), sorted the reads and marked duplicates with Picard ("Picard Toolkit" 2019), then indexed the resulting BAM files with samtools (Danecek et al. 2021). I then used *bam-readcount* (Khanna et al. 2022)

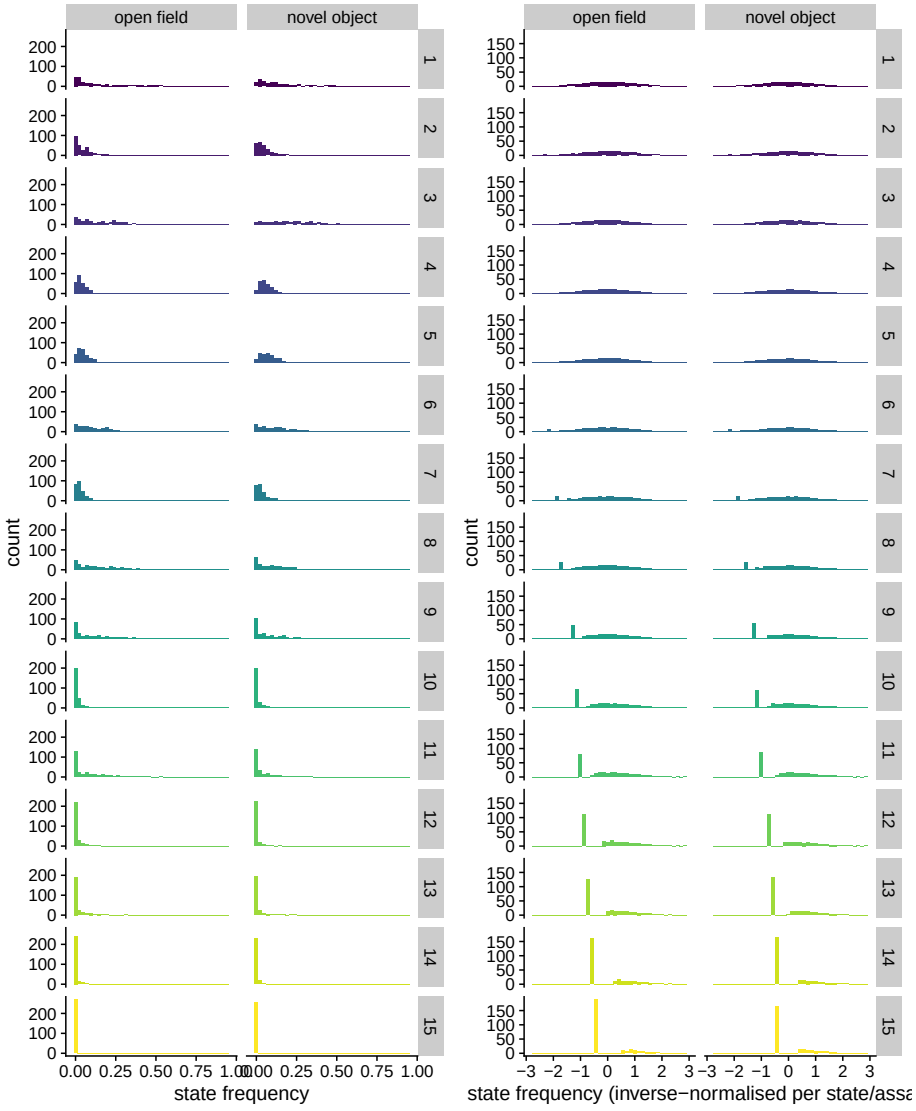


Figure 4.16: Effect of inverse-normalisation on the HMM state frequency of the F2 test fishes.

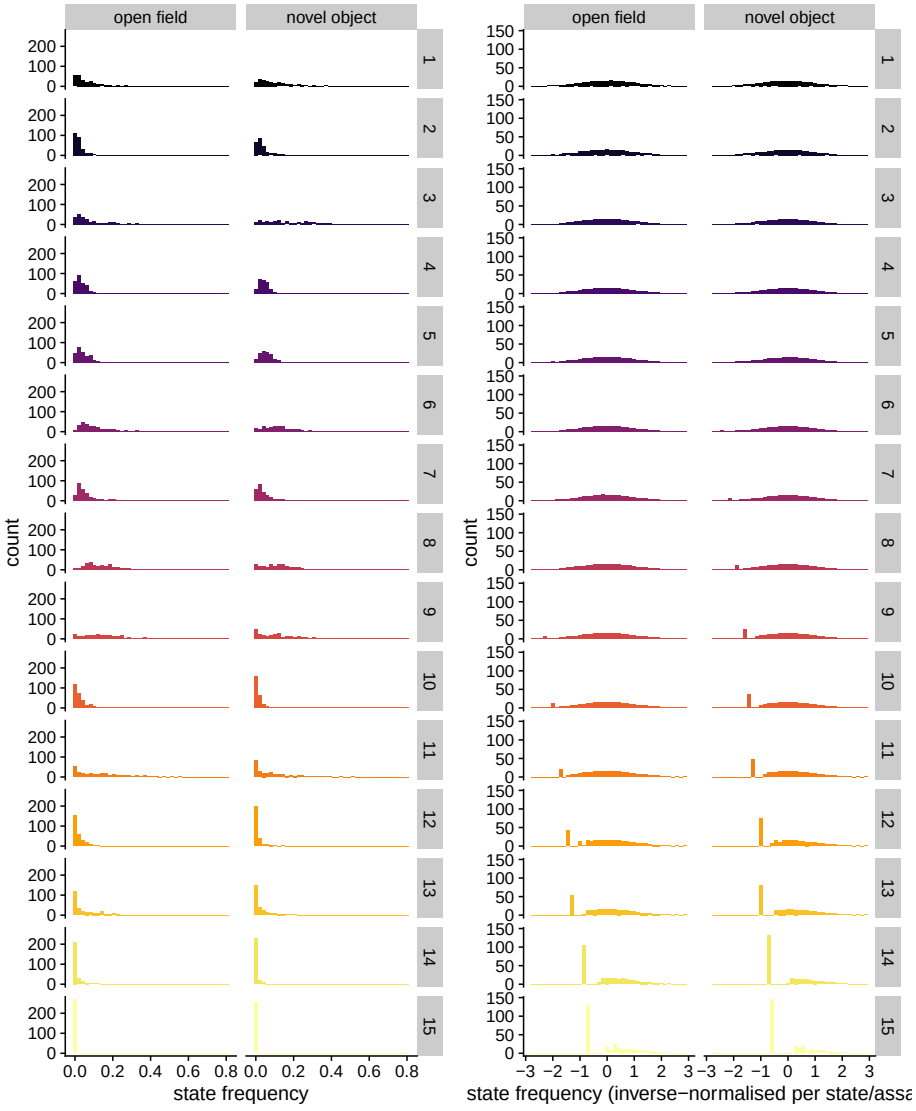


Figure 4.17: Effect of inverse-normalisation on the HMM state frequency of the F2 *iCab* reference fishes.

to count the reads that supported either the paternal or the maternal allele for all biallelic SNPs that were homozygous-divergent in the given sample's two parental strains (i.e. homozgyous reference in the paternal line, and homozygous alternative in the maternal line, or *vice versa*), summed the read counts within 5 kb blocks, and calculated the frequency of reads within each bin that supported the maternal allele. This generated a value for each bin between 0 and 1, where 0 signified that all reads within that bin supported the paternal allele, and 1 signified that all reads within that bin supported the maternal allele. Bins containing no reads were imputed with a value of 0.5.

I then used these values for all F2 individuals as the input to a Hidden Markov Model (HMM) with the software package *hmmlearn* (*Hmmlearn/Hmmlearn* [2014] 2022), which I applied to classify each bin as one of three states, with state 0 corresponding to homozygous for the paternal allele, 1 corresponding to heterozygous, and 2 corresponding to homozygous for the maternal allele. Across each chromosome of every sample, the output of the HMM was expected to produce a sequence of states. Based on previous analyses described in Chapter 5, I used the same HMM parameters as I did there, and used the HMM state outputs to generate the recombination blocks shown in **Figure 4.18**. Missing genotype state calls arose from a sample having insufficient reads within a bin, for which I imputed the missing state calls within each sample's chromosome based on the previous state call on that chromosome. A karyoplot retaining the missing blocks is provided in **Appendix A.6**.

I then took these HMM-generated haplotype block calls and used them to impute each sample's SNP-level genotypes using the homozygous biallelic SNP calls in the high-coverage .vcf file for the MIKK panel F0 lines described in Chapter 2. This set of variants included a total of ~20.7M SNPs, which I used to generate a Plink-format .bed file, forming the genotype input for the genetic linkage analysis.

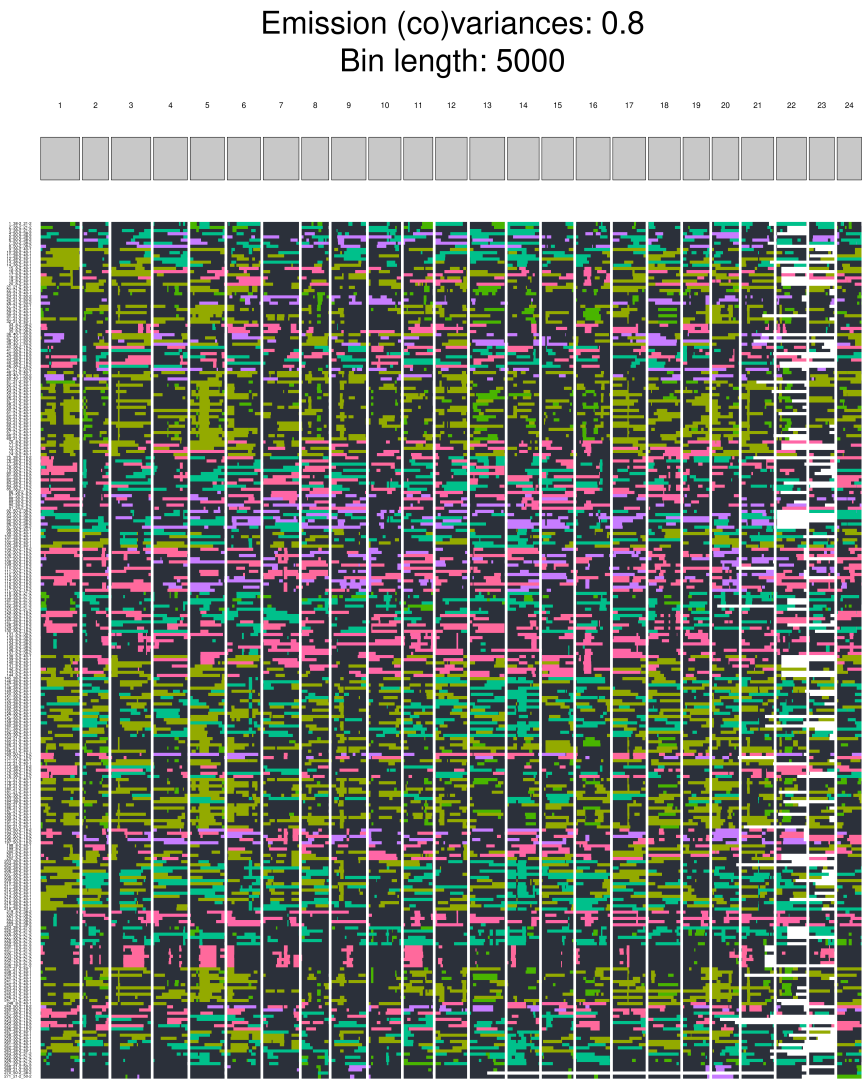


Figure 4.18: Karyoplot for F2 samples, coloured by genotype. Samples are sorted in the order in which they were phenotyped. Blocks are filled with the colour of the paternal F0 line for the homozygous paternal haplotype block, black for heterozygous, and the colour of the maternal F0 line for the homozygous maternal haplotype block. Missing calls were imputed based on the previous successful call on a given chromosome.

4.7.4 Genetic linkage analysis

For the purpose of using the *GCTA* software package (Yang et al. 2011) for the genetic linkage analysis, That software requires the construction of a genetic relationship matrix (**GRM**) $\mathbf{A} = \mathbf{WW}'/N$. $\mathbf{W}$ is a standardised genotype matrix with the ij^{th} element $w_{ij} = (x_{ij} - 2p_i)/\sqrt{(2p_i(1-p_i)}$, where x_{ij} is the number of copies of the reference allele for the i^{th} SNP of the j^{th} individual and p_i is the frequency of the reference allele (Yang et al. 2011).

For the GRM, I first filtered the .bed for SNPs that had no missing calls for any samples (M_{SNPs} = 44,360). I then used these SNPs to construct a "leave-one-chromosome-out" (**LOCO**) genetic relationship matrix for each chromosome – that is, if the "focal" chromosome was chr1, I would exclude the SNPs on that chromosome before constructing the GRM. To illustrate, **Figure 4.19** is a GRM constructed using all 44,360 non-missing SNPs. However, given the relatively small amount of non-missing SNPs on each chromosome, the number of SNPs on the focal chromosome that were excluded was small, resulting in GRMs that appear almost identical by eye. See the LOCO-GRM for chromosome 1 in **Appendix A.7**. The individuals from each cross neatly cluster together, and the individuals that share one parental strain cluster nearby.

I used these GRMs in the mixed linear model based association analysis (**MLMA**) implemented in GCTA (Yang et al. 2011), where the model was generally specified as follows:

$$y = a + bx + g + e$$

y was the phenotype, a was the mean term, b was the additive effect (fixed effect) of the candidate SNP tested for association, x was the SNP genotype indicator variable coded as 0, 1 or 2, g was the polygenic effect (random effect) i.e. the accumulated effect of all SNPs (as captured by the GRM calculated using all SNPs) and e was the residual (Yang et al. 2011). For y, I used the inverse-normalised state frequencies described above, and the LOCO-GRM as g for all SNPs on the focal chromosome. For example, for all SNPs on chr18, I used

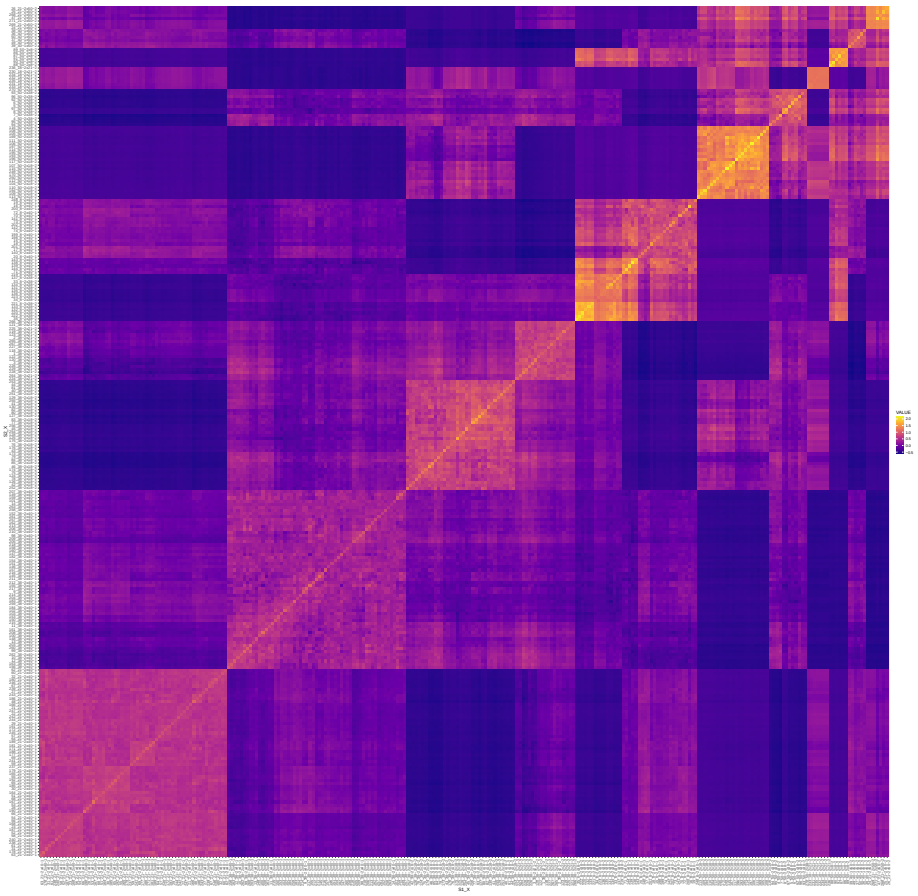


Figure 4.19: Genetic relationship matrix for 271 F2 samples based on 44,360 non-missing SNPs.

the LOCO-GRM that excluded all SNPs from that chromosome. In addition, I included the time of assay and the tank quadrant as covariates, which the software regresses out from the phenotype prior to running the MLMA. I excluded the date of assay and the tank side as covariates because individuals from the same cross tended to be assayed in the same test tank on the same day, and are therefore confounded with their genetics.

To set a significance threshold, for each combination of HMM state and assay, I ran 10 MLMA tests over the same dataset where I had permuted the phenotype and covariates using a different random seed. The logic behind this method is to determine the lowest p-value that one could expect when there is no true relationship between the indviduals' genetics and their phenotype. I then extracted the smallest p-value from all 10 results, and used this as the significance threshold. I additionally calculated the Bonferroni threshold as α/M, where α is set to 0.05 and M is the total number of SNPs in the dataset (2,726,797) = 1.83×10^{-8}.

4.7.4.1 Direct genetic effects

For the DGE phenotypes (comparing the state frequencies across test fishes), I sought to identify the genetic loci in the F2 individuals that were associated with differences in their own behaviour. For neither the open field nor novel object component did the p-values exceed the Bonferroni threshold, however a number of loci across several chromosomes exceeded the thresholds set by the permutations. I plotted the p-values in Manhattan plots for each combination of state and assay component, and provide them all in the Appendix. Here I showcase the Manhattan plot for state 3, containing a number of significant loci that I discuss further below.

4.7.4.2 Social genetic effects

For social genetic effects, I was attempting to discover genetic variants in the F2 individuals that caused differences in the behaviour

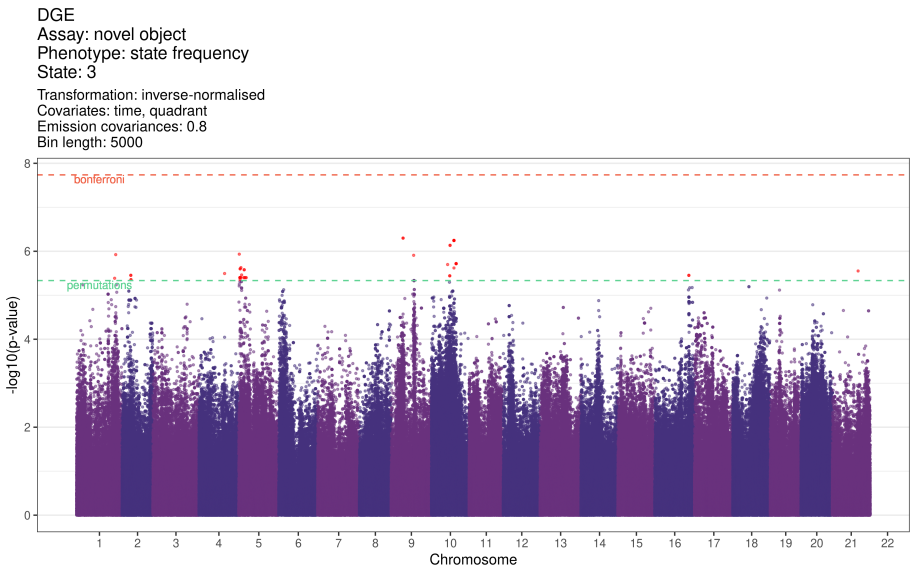


Figure 4.20: Manhattan plot for inverse-normalised HMM state 3 frequency of F2 test fishes during the novel object assay.

of their *iCab* tank partners. As expected, fewer loci reached significance for this transmitted indirect genetic effect than for the direct genetic effects, however I still found several significant loci based on the permutations threshold for states 4 (chr4), 7 (chr1), 12 (chr11), and 13 (chr12) during the open field assay component; but only one, barely significant SNP for state 5 (chr 13) during the novel object assay. This difference was somewhat surprising given our previous suppositions (DISCUSS) that the novel object component drew out stronger social genetic effects.

4.7.5 Candidate SNPs for CRISPR-knockouts

The next goal was to attempt to identify the SNPs that are mostly likely to be the causal variants, or closest in proximity to the causal variants, responsible for differences in the phenotypes of interest. I therefore proceeded to extract the significant SNPs and use Ensembl's Variant Effect Predictor (**VEP**) (McLaren et al. 2016) to identify the SNPs that are most likely to have a functional consequence,

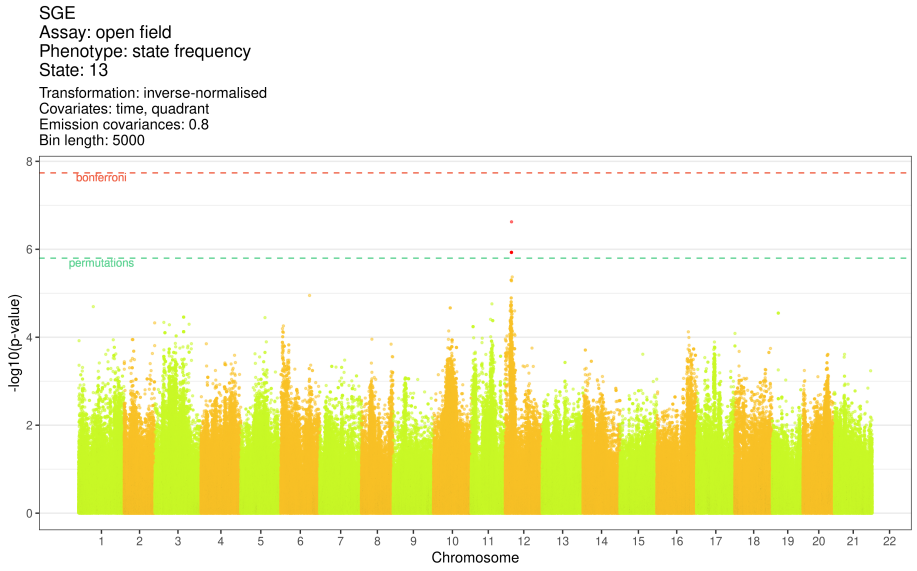


Figure 4.21: Manhattan plot for inverse-normalised HMM state 3 frequency of *iCab* reference fishes during the novel object assay component.

and determine whether the genes they reside in have already been identified as being involved in biological pathways that could be related to behaviour. **Table 4.4** shows the counts for the different potential consequences of the unique significant SNPs, based on Ensembl's annotation of the *HdrR* reference. I note that as multiple genes can overlap the same locus (including those transcribed from opposing strands), one SNP may map to multiple genes and therefore be counted as having more than one type of consequence.

As expected, most variants reside in non-coding regions [CITE], but in an attempt to isolate the most likely causative SNPs, I extracted those that are predicted to cause a missense mutation and therefore most likely to have a functional consequence (**Table @ref(tab:F2-sig-snps-missense**).

One SNP, 10:15,319,434, was significant for DGE for frequencies in both state 3 (**Figure 4.22**) and 10 during the novel object assay (see **Figure 4.20** above for state 3), and maps to two genes: ENSORLG00000029574 and ENSORLG00000024866. EN-

Table 4.4: Counts for consequences of significant SNPs.

consequence	count
intron_variant	349
intergenic_variant	178
upstream_gene_variant	92
downstream_gene_variant	75
synonymous_variant	9
3_prime_UTR_variant	6
missense_variant	6
5_prime_UTR_variant	3
splice_acceptor_variant	2
splice_region_variant,intron_variant	2
splice_region_variant,synonymous_variant	2
splice_donor_variant	1

Table 4.5: Missense SNPs.

Genetic effect	Assay	State	Chr	Pos	Ref	Alt	Allele	Gene
DGE	open field	1	1	22508288	A	C	C	ENSORLG0
DGE	open field	7	13	33714439	G	T	T	ENSORLG0
DGE	open field	10	3	28342550	A	G	G	ENSORLG0
DGE	novel object	3	1	31428698	G	A	A	ENSORLG0
DGE	novel object	3	10	15319434	C	T	T	ENSORLG0
DGE	novel object	3	10	15319434	C	T	T	ENSORLG0
DGE	novel object	10	10	15319434	C	T	T	ENSORLG0
DGE	novel object	10	10	15319434	C	T	T	ENSORLG0
SGE	open field	13	12	5136828	G	A	A	ENSORLG0

SORLG00000029574 is a novel gene on the forward strand with no recorded phenotypes. However, ENSORLG00000024866 is a gene on the reverse strand for protocadherin alpha-C2-like, a well-known protein involved in mammalian synapse formation (Junghans et al. 2008; Phillips et al. 2003).

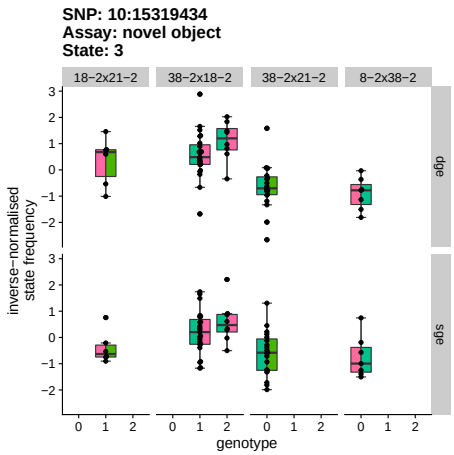


Figure 4.22: State 3 frequency during the novel object assay for counts of the alternative allele (T) at SNP 10:15,319,434. The boxes are coloured on the left by the paternal line, and on the right by the maternal line.

The second predicted missense variant of note was detected for SGE in the open field assay for frequencies in state 13. The locus 12:5,136,828 maps to gene ENSORLG00000002961, which is described as a complement *C9*. Discovered phenotypes for species orthologues include reduced prepulse inhibition and heart phenotypes in mice (Mouse Genome Database, Blake et al. 2021), and neurodevelopmental disorders in rats (Rat Genome Database, Smith et al. 2020). Here, in the F2 individuals, it appears to have an inconsistent effect in the behaviour of the test fishes, but apparently causes the reference fish to increase the proportion of time it spends in the fast-moving state 13 (**Figure 4.21**).

There are additional significant loci of note shown in **Figure 4.20** above that are not predicted to alter proteins, but nevertheless appear to affect genes related to neurological development (**Table 4.6**).

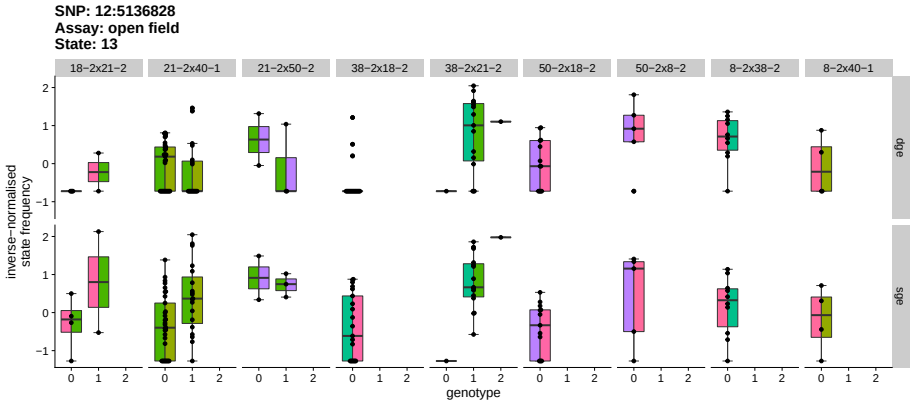


Figure 4.23: State 13 frequency during the open field assay for counts of the alternative allele (A) at SNP 12:5,136,828. The boxes are coloured on the left by the paternal line, and on the right by the maternal line.

Table 4.6: Significant loci on chromosome 9 for direct genetic effects on state 3 frequency during the novel object assay.

Genetic effect	Assay	State	Chr	Pos	Ref	Alt	Allele	Gene
DGE	novel object	3	9	9802754	C	A	A	ENSORLG0(
DGE	novel object	3	10	18537719	C	A	A	ENSORLG0(

The first is located around 9:9,802,754 (**Figure 4.24**), and maps to a gene for glutamate ionotropic receptor delta type subunit 2, an orthologue of the human gene *GRID2*. Deletions in this gene have been found to cause ataxia in humans (Hills et al. 2013; Utine et al. 2013), and other mutations in this gene have been found to cause various neurological disorders in mice (Mouse Genome Database, Blake et al. 2021).

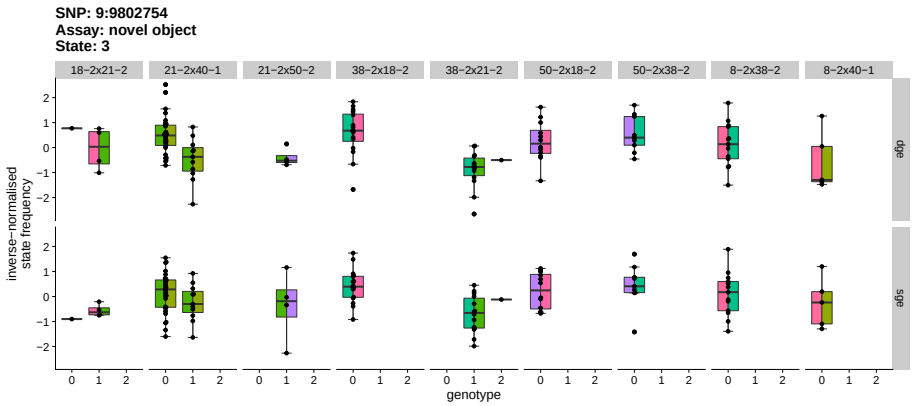


Figure 4.24: State 3 frequency during the novel object assay for counts of the alternative allele (A) at SNP 9:9,802,754. The boxes are coloured on the left by the paternal line, and on the right by the maternal line.

The second locus resides around 10:18,537,719 (**Figure 4.25**) within the neuroligin 3 gene, an orthologue of the human gene *NLGN3*, which has been linked with autism in humans (Jamain et al. 2003), and various neurological in mice and rats (Blake et al. 2021; Smith et al. 2020).

These results provide promising evidence that the assay and methods described above can identify genetic loci associated with differences in behaviours - and potentially difference in the transmission of behaviour onto social companions - with functional relevance to humans. However, before seeking to functionally validate these variants, there are a number of steps that we seek to take to refine the analysis and thereby increase our confidence in the variants we ultimately select for validation.

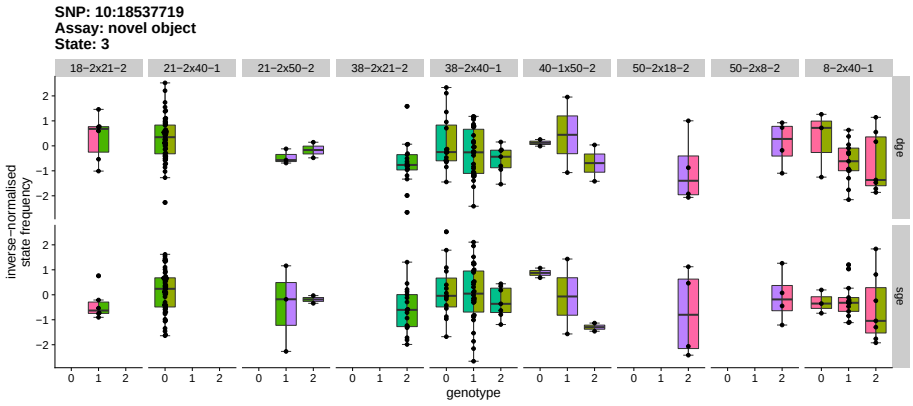


Figure 4.25: State 3 frequency during the novel object assay for counts of the alternative allele (A) at SNP 10:18,537,719. The boxes are coloured on the left by the paternal line, and on the right by the maternal line.

4.8 Discussion

With the benefit of hindsight, there are a several aspects of this analysis that I would have performed differently. The first relates to the choices of certain parental lines for the F2 cross, which were too heavily reliant on the aggregate measures of speed and charisma, without due regard to the level of behavioural variance observed within each line. As evident in **Figures 4.12** and **4.13**, lines 8-2 and 21-2 showed large within-line variances. For 8-2 in particular, the differences appear to correspond to the date on which they were assayed - that is to say, in the first run the four 8-2 individuals showed almost no movement, whereas in the second run they showed much higher levels of movement. Although the different within-line variances are an interesting phenotype to explore further, as they suggest that different lines show different degrees of sensitivity to environmental changes, high within-line variances make it difficult to ascertain the line's "true" phenotype, and therefore make them insuitable for selection based on our axes of interest for this study (bold vs shy; high vs low behavioural transmission).

An appropriate substitute for 8-2 may be line 139-4, which had low

within-line variance for mean speed, and the highest median state co-occupancy in the novel object assay component. From the analysis of Chapter 3, which concluded after the F2 cross lines were already selected, it became apparent that the novel object assay was particularly useful for revealing social genetic effects. I hypothesise that at times of higher stress or predation threat, the individual fish take more behavioural cues from their tank partners. Line 139-4 therefore would have been a good candidate for the slow-moving, high-charisma line.

Similarly, lines 13-2 or 94-1 would be preferred substitutes for line 21-2 as the fast-moving, high charisma line. They are faster than 21-2, but unlike the fastest line 10-1, they show a degree of habituation (where most individuals moved slowly at the beginning of the assay before eventually speeding up). They also showed lower within-line variance than 21-2, but comparable measures for the SGE measures of state co-occupancy and reference deviation.

One way of avoiding these issues would have been to used a more quantitative, objective approach to selecting the lines. For example, I could have used Mann-Whitney tests to quantify the levels of differences between pairs of lines for a trait of interest (as I did for **Figure 4.4**), and then select the lines that maximised that statistic. Such metrics would still perhaps need to be qualitatively weighted against other relevant traits, such as within-line variance. Nevertheless, my downstream analyses show that despite not having potentially selected the ideal lines for the F2 cross, the differences observed were sufficient to identify genetic loci associated with our phenotypes of interest, even with a relatively small sample size of 271 F2 individuals.

4.9 Future directions

Our collaborators are in the process of breeding more F2 individuals to increase the sample size for this analysis. I have suggested that we replace line 21-2 with either 13-2 or 23-1, which both show lower within-line variance in mean speed, with similarly high levels of speed and behavioural transmission (as measured by state co-

occupancy and *iCab* reference deviation). Any additional F2 individuals will still need to be phenotyped and sequenced, and the phenotype data collection (i.e. video recording the behavioural assay) can be relatively laborious, so this process will likely take some time.

In the meantime, I have 96 Kiyosu CC individuals whose videos have already been processed, and whose DNA is already shallow-sequenced (~1x). However, it was not feasible to include that data in this analysis due to time constraints. As they have been allowed to breed freely from the same Kiyosu population as the MIKK panel, they are not strongly homozygous, and they are likely to possess many additional genetic variants that are not represented in the MIKK panel. I am therefore not able to impute their genotypes from high-coverage parental strains, as I did with the F2 individuals. The alignments to the reference are likely to include more mapping errors due to the lack of sufficient coverage. The additional information they provide will therefore not be as useful as additional F2 individuals, but it should nevertheless improve the power to detect casual variants. The level of information we obtain from the Kiyosu CC individuals will also inform whether we include them in this way in future F2 crosses.

A second issue I plan to address is the failure of F2 reads to align to chromosomes 22 and 23 (see **Figure 4.18** above). As the issue occurs across most of the sample, and the coverage does not look skewed for those chromosomes (**Figure 4.26**), there is likely to a be an error in the bioinformatic pipeline that I constructed.

A third issue is the matter of selecting the optimal HMM parameters for calling the recombination blocks, as described in Chapter 5. For the analysis in this current Chapter I used the same HMM parameters as I had used there. In that Chapter, I used the p-values from the resulting genetic linkage analysis as a rough measure of genotyping quality (assuming that lower p-values were driving by more accurate genotypes), together with a qualitative assessment of how many crossovers I expected to observe in a particular chromosome for a subset of the sample. The issue can be resolved by obtaining "ground truth" genotype calls through high-coverage sequence data for a set of samples that have also been shallow-sequenced. This would allow

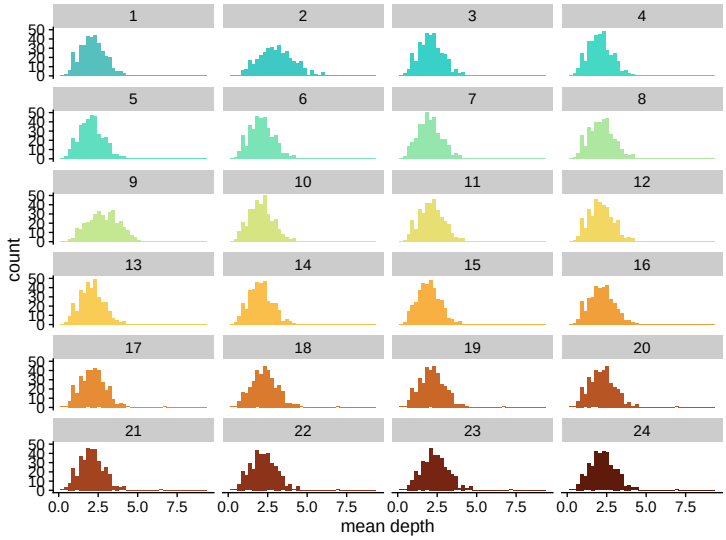


Figure 4.26: Mean depth of sequencing coverage per chromosome for the 271 F2 samples used in this analysis.

one to compare the HMM recombination block calls with the ground truth genotypes, and thereby train the model to optimise the accuracy of its predictions for the locations of the recombination sites (while validating the trained parameters with a test dataset). I expect the improved genotype calls will further improve the power to detect genetic variants.

With respect to the genetic linkage analysis, due to the method I used to calculate the genetic relationship matrix (**GRM**), I was compelled to exclude all SNPs that were missing calls in any of the samples, reducing the number of SNPs to merely ~40K out of the ~20M in the full dataset. Not only does this reduce the accuracy of the GRM as a means of capturing the relative relatedness between samples, but the problem will only be exacerbated with the inclusion of additional samples (as the likelihood of a SNP having at least one sample with a missing call will increase with the number of samples). I will therefore adapt my method to include SNPs with missing calls, which should further increase statistical power.

Finally, I applied a separate genetic linkage association test for the

state frequency of each HMM state, where it may be beneficial to apply a method that can combine the information across all states into a single phenotype (if not fewer phenotypes), and then run an association test on that reduced phenotype. There are a number of alternatives to consider, including principal components analysis (PCA) or MTAG (Multi-Trait Analysis of Genome-wide summary statistics) (Turley et al. 2018). Ultimately the hope is that by leveraging the information about the frequencies in which the fishes occupy each state, combined across *all* states, while appropriately taking into account the interdependence of those frequencies (e.g. the more time a fish spends in state 3, the less time it is likely to spend in state 13), it will further strengthen these results.

4.10 Lessons

- Describe using a different MIKK panel line as the reference, preferring a moderate-moving line to enable easier detection of social genetic effects from fast-moving lines, i.e. which lines cause *iCab* to move more boldly?
- Explain why I didn't choose 22-1 or 10-1.
- Add path plots for interesting lines
- Figure showing choices of crosses

Chapter 5

Genetic loci associated with somite development periodicity

5.1 Background

During the development of an embryo, somites are the earliest primitive segmental structures that form from presomatic mesoderm cells (**PSM**) (Kim et al. 2011). They later differentiate into vertebrae, ribs, and skeletal muscles, thereby establishing the body's anterior-posterior axis. **Figure 5.1** depicts a number of formed somites in a 9.5-day-old mouse embryo.

Somite formation occurs rhythmically and sequentially, with the time between the formation of each pair of somites referred to as the "period". The period of somite formation varies greatly between species: ~30 minutes for zebrafish, 90 minutes for chickens, 2-3 hours for mice, and 5-6 hours for humans (Hubaud and Pourqui'e 2014; Matsuda et al. 2020). **Figure 5.2** shows the a series of time-stamped images of somite segmentation in medaka fish, generated by Ali Seleit.

Figure 5.1: Image of a mouse embryo at day 9.5 from Gridley (2006), showing somites in darker colours.

The period of somite formation is controlled by a molecular oscillator, known as the 'segmentation clock', which drives waves of gene expression in the Notch, fibroblast growth factor (FGF), and Wnt pathways, forming a signalling gradient that regresses towards the tail in concert with axis elongation (Gomez et al. 2008). Over the course of elongation, the wave period increases (i.e. each somite takes longer to form), and the PSM progressively shrinks until it is exhausted, eventually terminating somite formation (Gomez et al. 2008).

It is not fully understood how the phase waves of the segmentation clock are intially established (Falk et al. 2022). Matsuda et al. (2020) found that period differences between mouse and human occur at the single-cell level (i.e. not due to intercellular communication), and are driven by biochemical reaction speeds - specifically, mRNA and protein degradation rates, transcription and translation delays, and intron and splicing delays. To identify the genetic basis of these biochemical differences, our collaborators Ali Seleit and Alexander Aulehla at EMBL-Heidelberg used a CRISPR-Case9 knock-in approach (Seleit, Aulehla, and Paix 2021) to establish

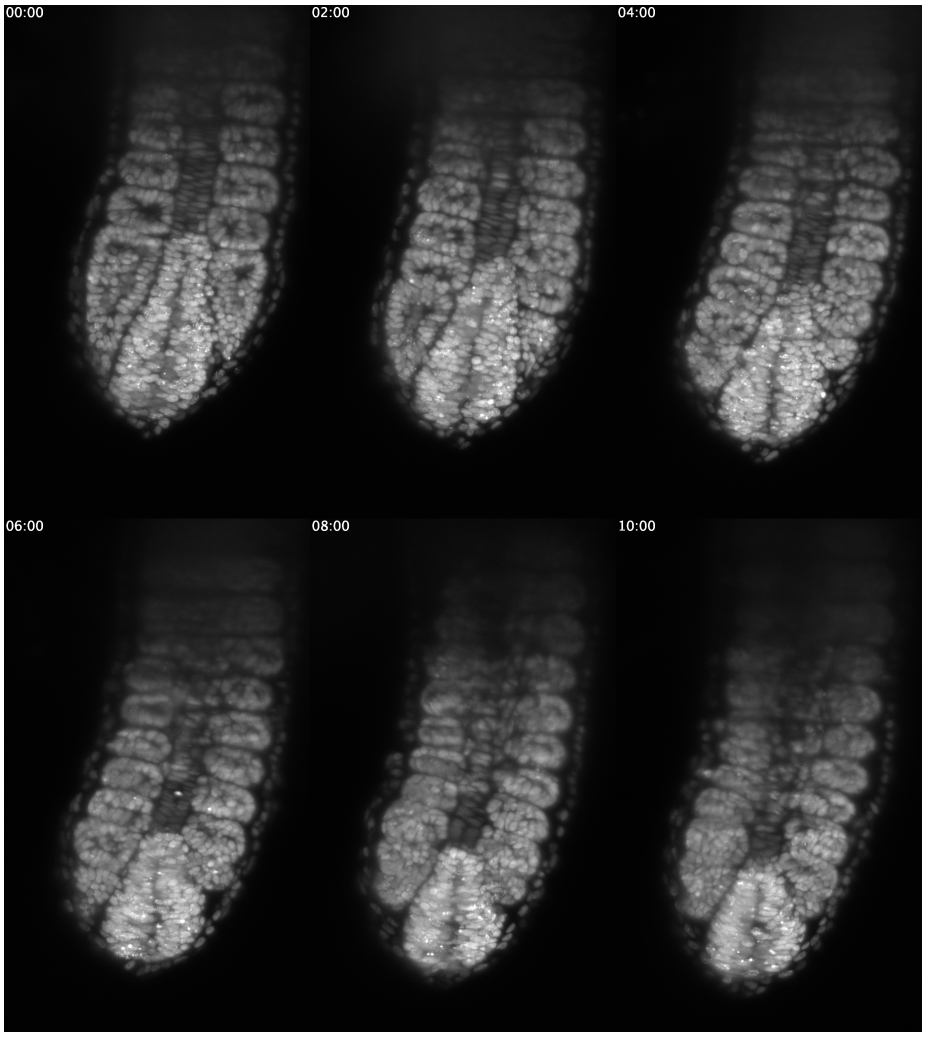


Figure 5.2: Time-stamped images of somite segmentation in medaka, generated by Ali Seleit.

a medaka *Cab* strain with an endogenous, fluorescing reporter gene (Her7-Venus) for the oscillation signalling pathway. This method allows them to image somite formation and extract quantitative measures for segmentation clock dynamics.

In medaka, it is known that the southern Japanese *Cab* strain and the northern Japanese *Kaga* strain have divergent somite periodicity, where *Kaga*'s tends to be faster, and *Cab*'s slower (**Figure 5.3**).

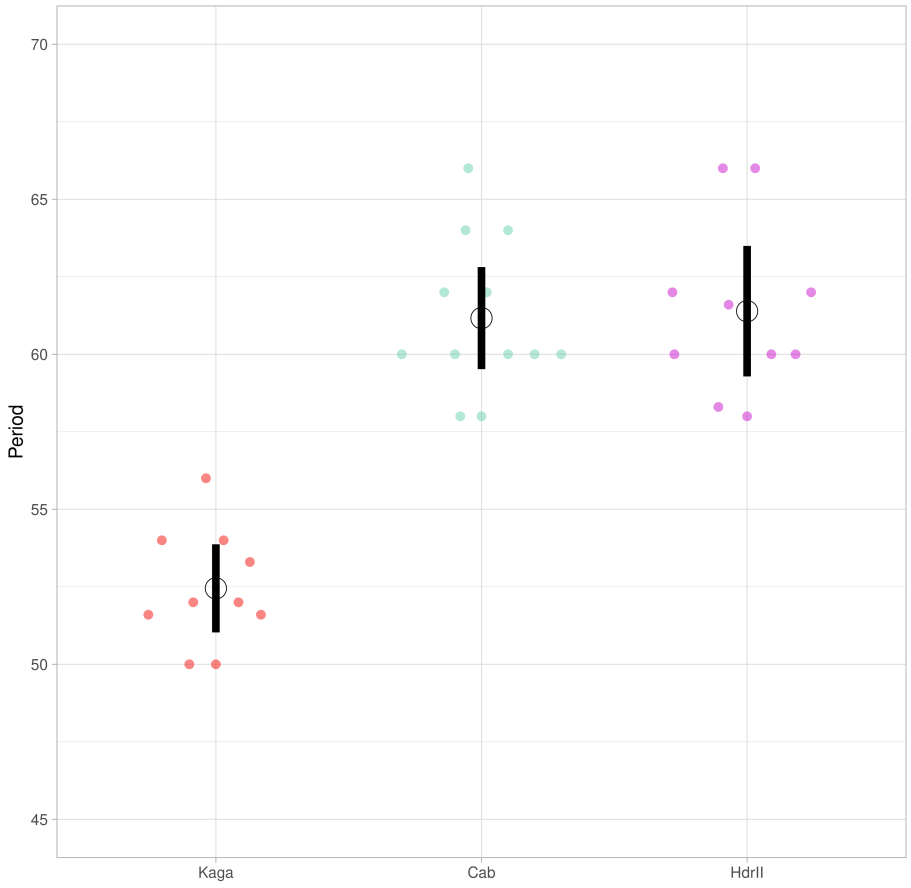


Figure 5.3: Comparison of period for three inbred medaka strains (*Cab*, *Kaga* and *HdrR*). Kaga's period is lower, and therefore it takes less time to form each somite than *Cab*. Figure generated by Ali Seleit.

Our collaborators accordingly set up a one-way F2 cross experiment

as described in Chapter 4.1, using the reporter-carrying *Cab* strain and the *Kaga* strain as the parental F0 strains, in order to identify genetic loci associated with these differences in clock dynamics. They inter-crossed the hybrid F1 generation to create a sample of 622 F2 individuals, imaged the developing embryos of these F2 samples, and used pyBOAT (Schmal, Mönke, and Granada 2022) to extract the oscillation features during somite development. **Figure 5.4** shows a series of raw images used by pyBOAT to track the elongation of a medaka tail during somitogenesis, with the identified posterior tip of the embryo labelled with a blue circle.

5.2 Phenotypes of interest

5.2.1 Somite development period

Figure 5.5 shows the period data generated by pyBOAT for this study, for 100 illustrative F2 samples over 300 minutes. The same data can be represented by boxplots as shown in **Figure 5.6**. I experimented with using the F2 individuals' mean period and period intercept as the phenotype of interest. The two measures are highly correlated ($Pearson's\ r$ = 0.84, p < 2.2x10-16), so after displaying the distributions for both measures in Figure 5.8, I proceed to only discuss the analysis of period intercept, as it would appear to potentially be more robust to the changes in slope that can be observed in Figure 5.5.

5.2.2 PSM area

In the proceeding analyses, I also included a second phenotype of interest: the total size of the PSM prior to the formation of any somite segments. As the measure is simply based on the total number of pixels covered by the embryo object, I considered it to be potentially more robust than the period measurements, and therefore included it as a type of positive control for the genetic association analyses on

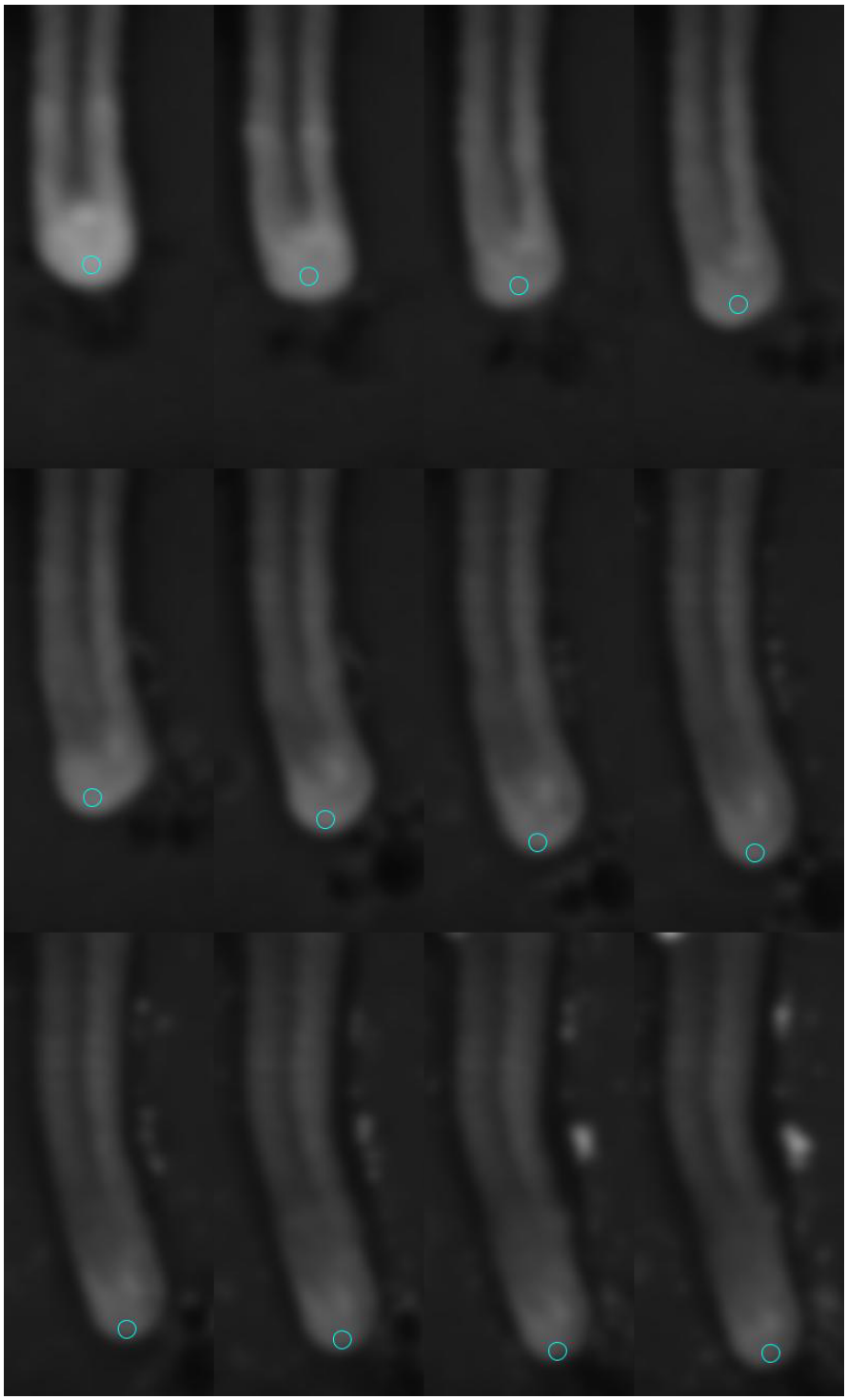


Figure 5.4: Screenshots of vertebral elongation in an F2 individual captured by Ali Seleit during imaging. The blue circle represents the point tracked by pyBOAT over time, generating the quantitative phenotype data on period development used in this study.

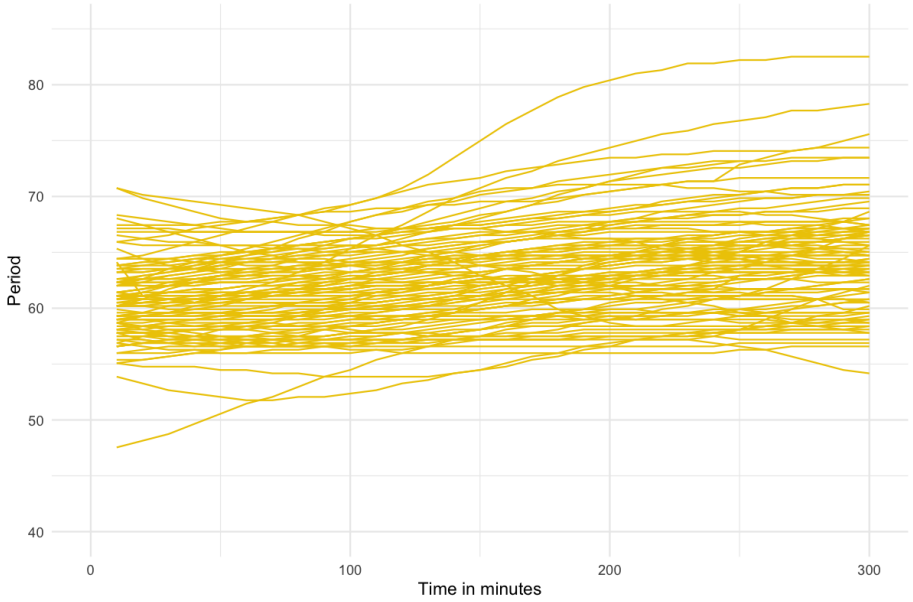


Figure 5.5: PyBOAT results for 100 illustrative F2 samples, showing the period length in minutes over the course of 300 minutes. Period tends to increase over time, meaning that as the embryo develops, each successive somite takes longer to form. Figure generated by Ali Seleit.

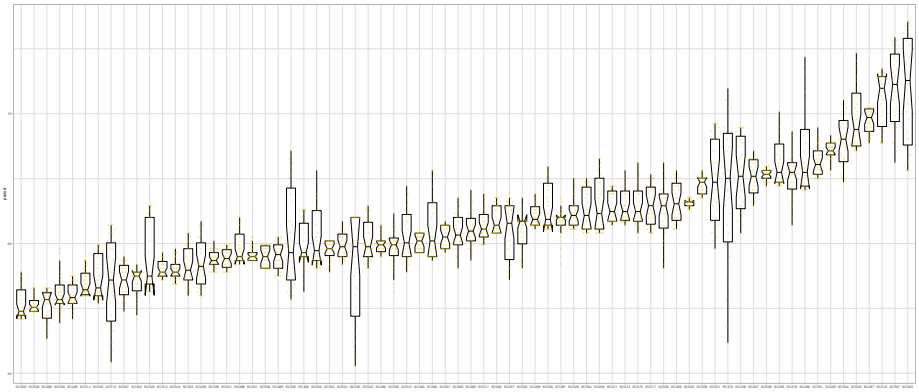


Figure 5.6: Period measurements for 70 F2 individuals displayed as boxplots with each individual's median and interquartile range. Figure generated by Ali Seleit.

the period phenotype. The measurements for PSM area comparing F0 *Cab* and *Kaga* strains are set out in **Figure 5.7**.

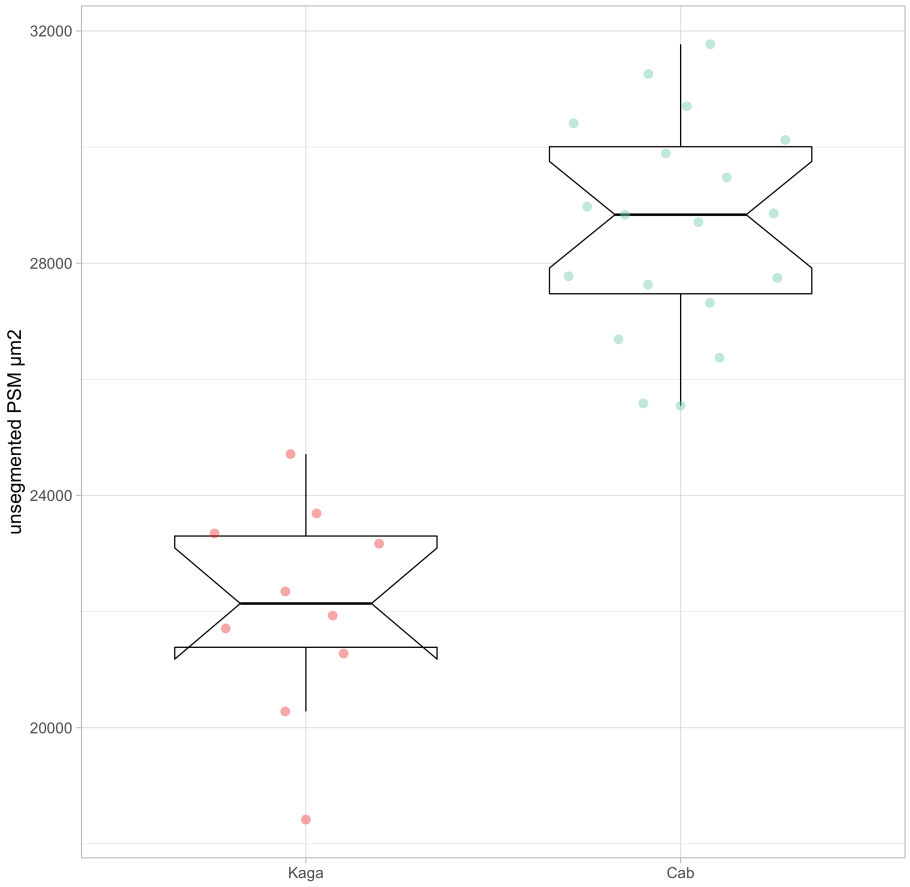


Figure 5.7: Measurements of unsegmented PSM area in pixels for the F0 individuals from the *Cab* strain ($N = 19$) and *Kaga* strain ($N = 10$). *Kaga* tends to have a smaller PSM than *Cab*. Figure generated by Ali Seleit.

5.2.3 Comparisons between F0, F1 and F2 generations

The distributions across the F0, F1 and F2 generations are curious (**Figure 5.8**). I expected to observe an F2 distribution with a similar median to the F1, and a variance that spanned across the extremes

of the F0 strains. Instead, I observed that for the period phenotypes, the F2 generation had a mean that was slightly higher than the median of the higher-period F0 *Cab* strain, and many F2 samples exceed the period values in those F0 samples. Our collaborators assured me that these observations were unlikely to be caused by technical issues. The *Cab* and *Kaga* strains originate from different Japanese medaka populations (southern and northern respectively) that are understood to be at the point of speciation (see Chapter 2), so this slower period may be driven by a biological incompatibility between their genomes in cases where they do not have a complete chromosome from each parent (as the F1 generation does). I nevertheless proceeded with the genetic analysis with a view to potentially discovering the reason for this unusual distribution.

Another important issue to note is that the F2 individuals were sequenced using different microscopes, denoted as 'AU' and 'DB'. Our collaborators noticed that there was a difference between the microscopes in their temperatures of 0.7-0.8°C, translating to a 4-minute difference in the F2 means for the period intercept measure (Kruskal-Wallis = 177.97, p = 1.34x10-40), and a 3.5-minute difference in the F2 means for the period mean measure (Kruskal-Wallis = 141.79, p = 1.08x10-32). This difference would need to be accounted for in the downstream analysis through either adjusting the phenotype prior to running the genetic association model, or by including microscope as a covariate in the model. No significant was found for the PSM area. To resolve this difference between microscopes for the period intercept data, I elected to transform it for the F2 generation by inverse-normalising the period intercept within each microscope (**Figure 5.9**). I then used this transformed data in the downstream analysis.

5.3 Genetic sequencing data

Our collaborators extracted DNA from the F0, F1, and F2, and sequenced the F0 and F1 samples with the Illumina platform at high coverage (~26x for *Cab* and ~29x for *Kaga*), as measured by samtools

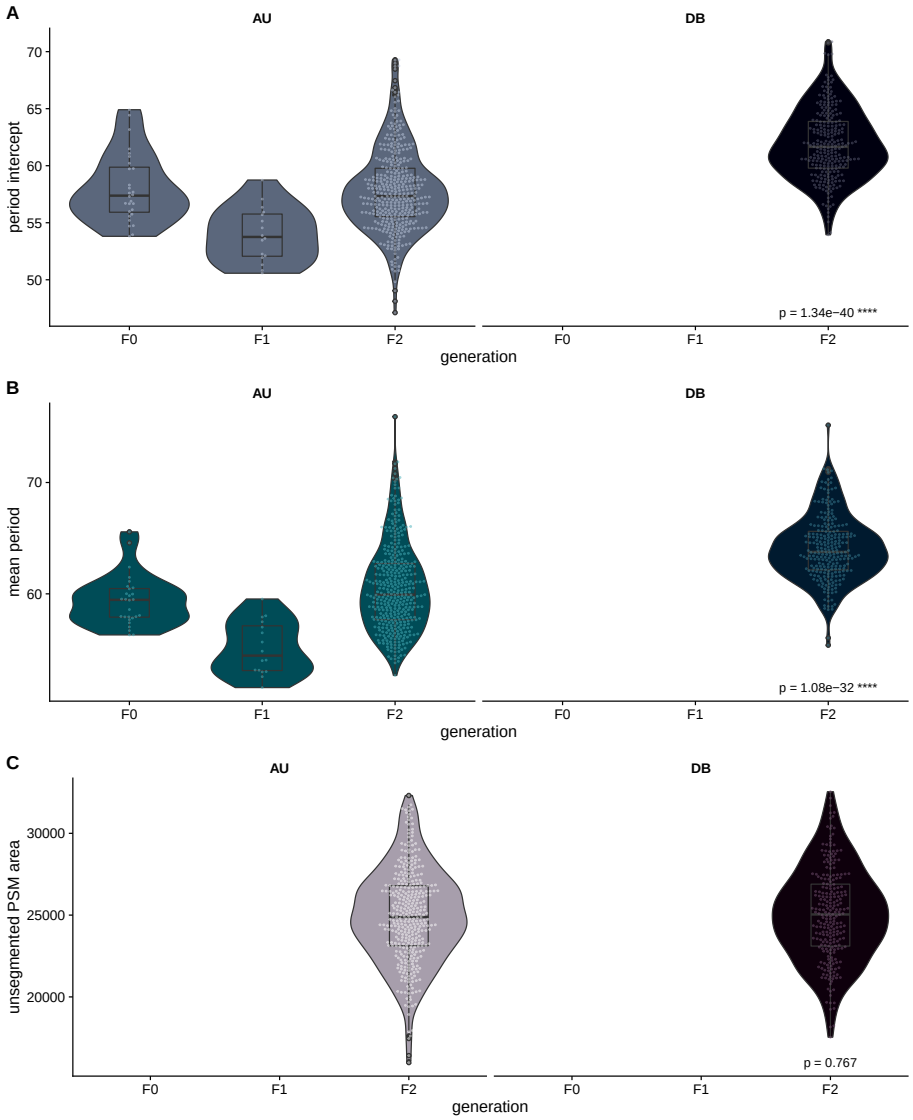


Figure 5.8: Comparisons between the F0, F1 and F2 generations for the three phenotypes of interest. Here, the F0 only includes *Cab* individuals. **A**: period intercept. **B**: period mean. **C**: unsegmented PSM area. P-values are derived from Kruskal-Wallis tests comparing the F2 individuals across microscopes.

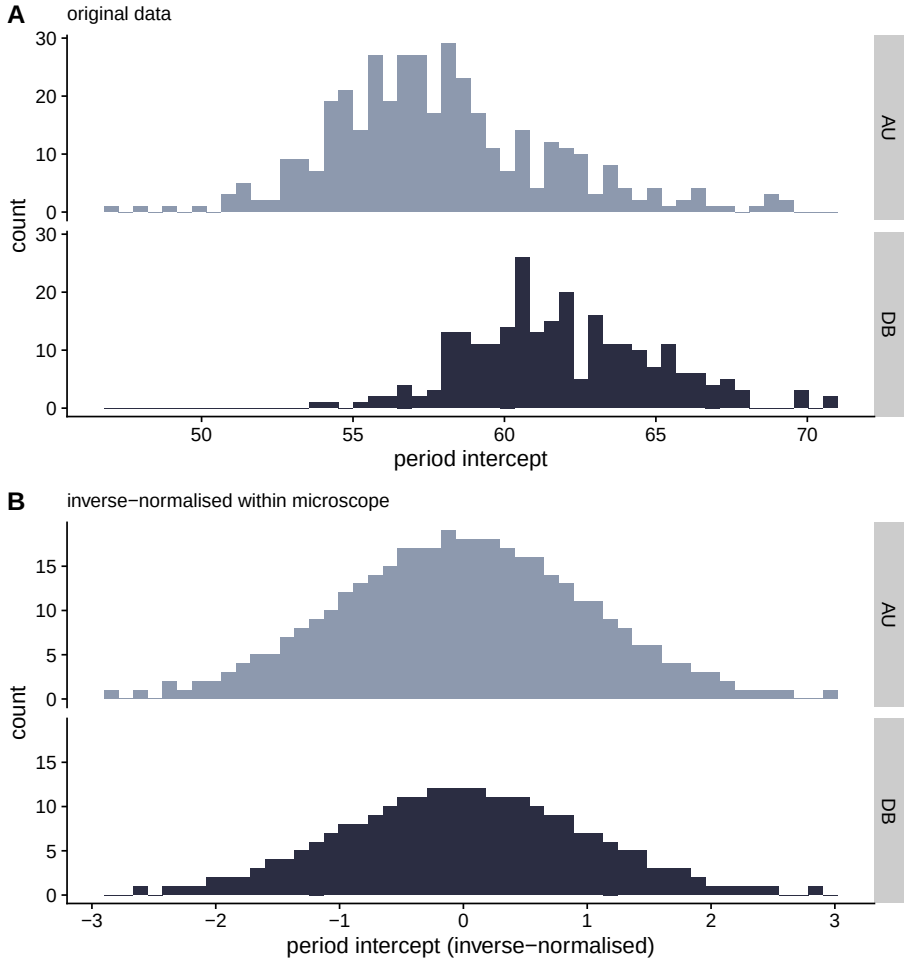


Figure 5.9: Comparison of the period intercept phenotype data for the F2 generation before (**A**) and after (**B**) inverse-normalisation.

(Danecek et al. 2021). **Figure 5.10** sets out the mean sequencing depth within each chromosome and across the whole genome for the *Cab* and *Kaga* F0 samples. Our collaborators then sequenced the F2 samples at low coverage (~1x), which would be sufficient to map their genotypes back to the genotypes of their parental strains (see Chapter 5.6 for further details).

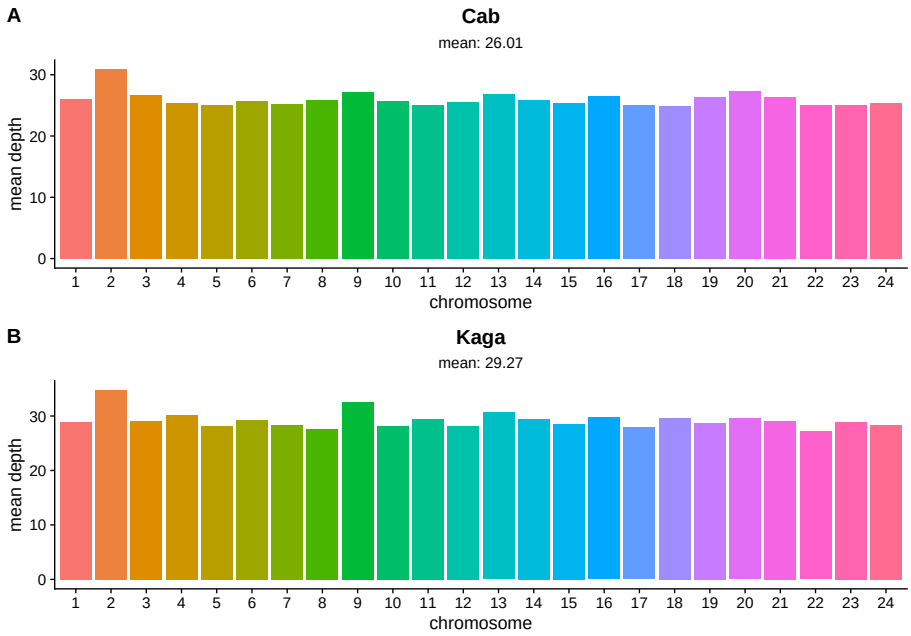


Figure 5.10: (ref:F0-coverage)

5.4 F0 homozygosity and F1 heterozygosity

Before proceeding to map the F2 sequences to the genotypes of the F0 generation, I first investigated the levels of homozygosity in the F0 *Cab* and *Kaga* strains, as this would have an effect on the resolution of our genetic mapping. That is to say, for regions where either parent is consistently heterozygous, it would be difficult to determine the parent from which a particular F2 individual derived its chromosomes at that locus. I therefore aligned the high-coverage sequencing data for the F0 *Cab* and *Kaga* strains to the medaka *HdrR* refer-

ence (Ensembl release 104, build ASM223467v1) using BWA-MEM2, sorted the aligned .sam files, marked duplicate reads, and merged the paired reads with picard ("Picard Toolkit" 2019), and indexed the .bam files with Samtools (Li et al. 2009).

To call variants, I followed the GATK best practices (to the extent they were applicable) (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera and O'Connor 2020) with GATK's HaplotypeCaller and GenotypeGVCFs tools (Poplin et al. 2018), then merged all calls into a single .vcf file with picard ("Picard Toolkit" 2019). Finally, I extracted the biallelic calls for *Cab* and *Kaga* with bcftools (Danecek et al. 2021), counted the number of SNPs within non-overlapping, 5-kb bins, and calculated the proportion of SNPs within each bin that were homozgyous.

Figure 5.11 is a circos plot generated with circlize (Gu et al. 2014) for the *Cab* F0 strain used in this experiment, featuring the proportion of homozygous SNPs per 5-kb bin (green), and the total number of SNPs in each bin (yellow). As expected for a strain that has been inbred for over 10 generations, the mean homozygosity for this strain is high, with a mean proportion of homozygosity across all bins of 83%.

However, the levels of homozygosity in the *Kaga* strain used in this experiment was far lower, with a mean homozygosity across all bins of only 31%. This was a surprise, as it is an established strain of X generations, and we therefore expected the level of homozygosity to be commensurate with that observed in the *Cab* strain.

To determine whether the low levels of observed homozygosity in Kaga was affected by its alignments to the southern Japanese *HdrR* reference, we also aligned the F0 samples to the northern Japanese *HNI* reference (**Figure 5.13**. This did not make differences to the levels of observed homozygosity in either sample, which gave us confidence that the low homozygosity observed in *Kaga* was not driven by reference bias.

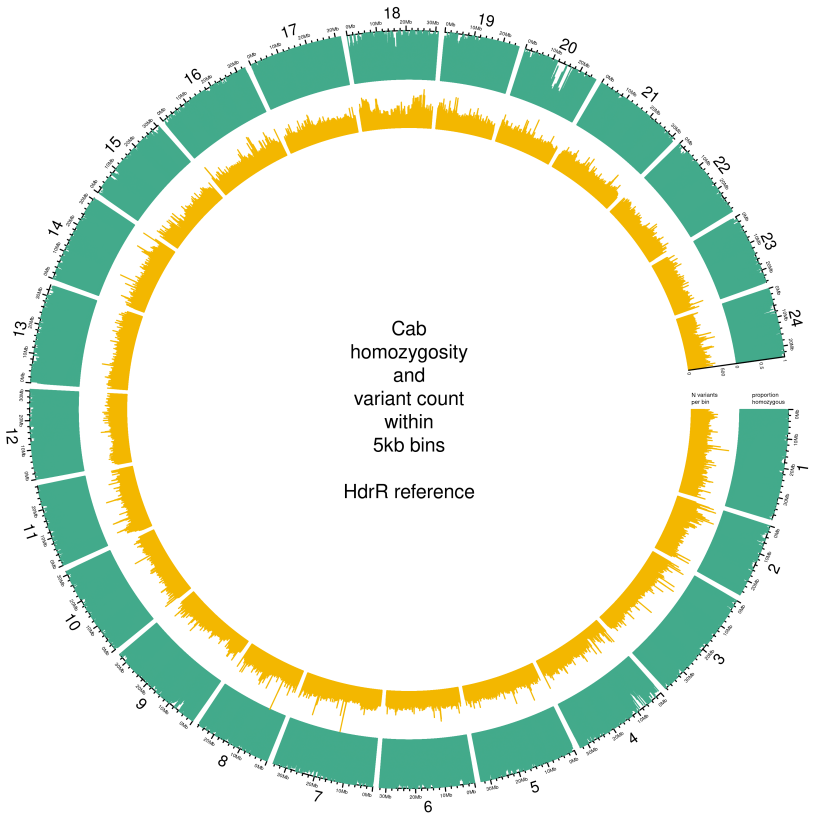


Figure 5.11: Proportion of homozygous SNPs within 5 kb bins in the *Cab* F0 generation genome (green), and number of SNPs in each bin (yellow).

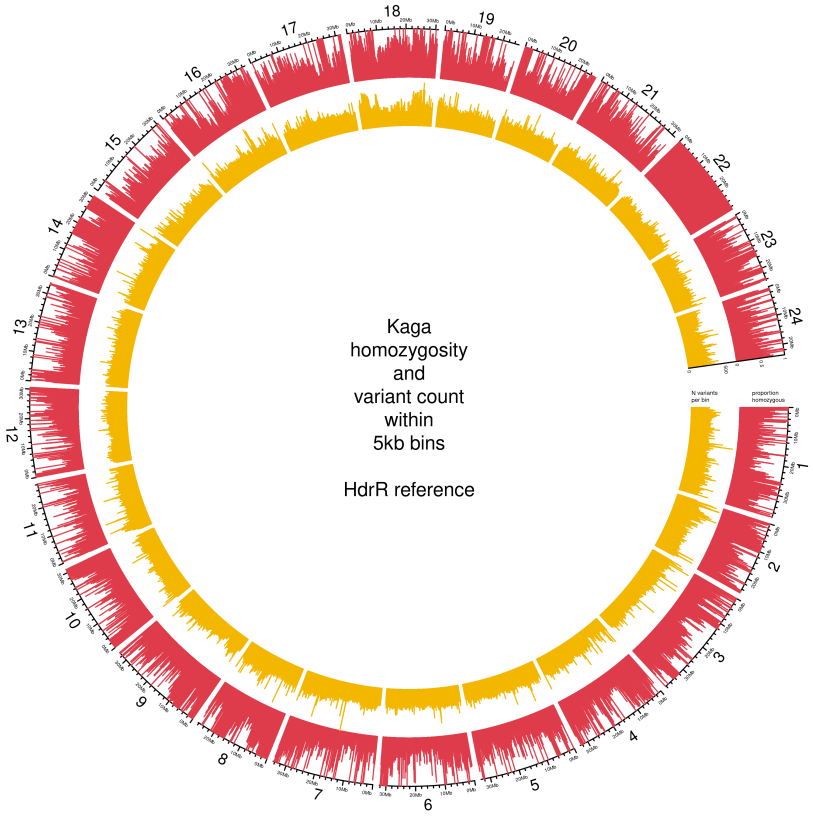


Figure 5.12: Proportion of homozygous SNPs within 5 kb bins in the *Kaga* F0 generation genome (red), and number of SNPs in each bin (yellow).

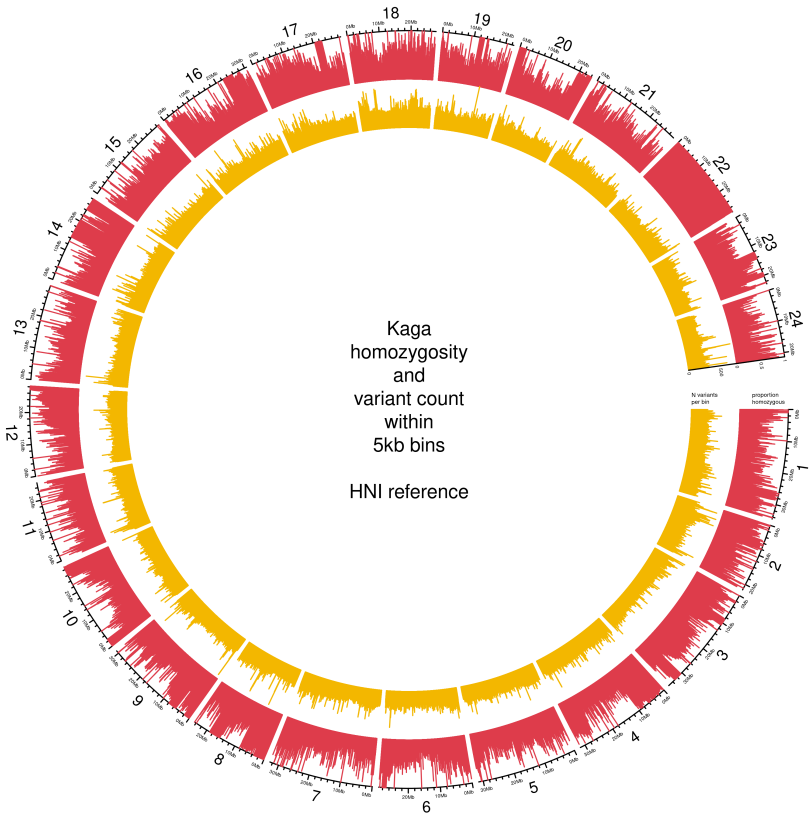


Figure 5.13: Proportion of homozygous SNPs within 5 kb bins in the *Kaga* F0 generation genome when aligned to the *HNI* reference (red), and number of SNPs in each bin (yellow).

5.5 F1 homozygosity

I next examined the level of heterozygosity in the F1 generation from the *Cab-Kaga* cross. **Figure 5.14** shows the level of heterozygosity across the genome of the F1 hybrid in brown measured by the proportion of heterozygous SNPs within 5-kb bins (brown), and the number of SNPs in each bin (yellow). Approximately half the chromosomes show inconsistent heterozygosity, with a mean heterozygosity across all bins of 67%. This lower level of heterozygosity than expected was likely caused by the low levels of heterozygosity in the *Kaga* F0 parent.

For the purpose of mapping the F2 sample sequences to the genomes of their parental strains, I selected only biallelic SNPs that were homozygous-divergent in the F0 generation (i.e. homozygous reference allele in *Cab* and homozygous alternative allele in *Kaga* or vice versa) *and* heterozygous in the F1 generation. The number of SNPs that met these criteria per chromosome are set out in **Figure 5.15**.

5.6 F2 genotyping

To maximise the efficiency of our sequencing runs, our collaborators "shallow-sequenced" the F2 generation with the short-read Illumina platform at a depth of ~1x. We then aligned these sequences to the *HdrR* reference with BWA-MEM2 (Vasimuddin et al. 2019), sorted the reads and marked duplicates with Picard ("Picard Toolkit" 2019), then indexed the resulting BAM files with samtools (Danecek et al. 2021). Genotyping these shallow sequences with the same method as used for the high-coverage sequences for the F0 and F1 generation would be inappropriate. We therefore used a different method whereby we used *bam-readcount* (Khanna et al. 2022) to count the reads that supported either the *Cab* or the *Kaga* allele for all SNPs that met the criteria described above in section 5.5, summed the read counts within 5 kb blocks, and calculated the frequency of reads within each bin that supported the *Kaga* allele. This generated a

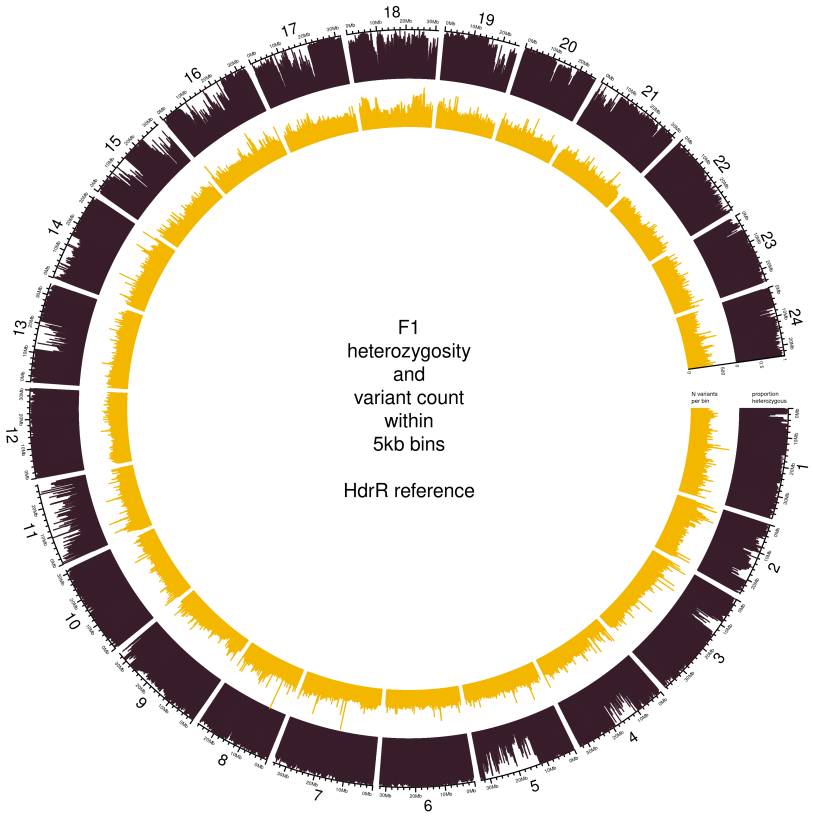


Figure 5.14: Proportion of heterozygous SNPs within 5 kb bins in the *Cab-Kaga* F1 cross (brown), and number of SNPs in each bin (yellow).

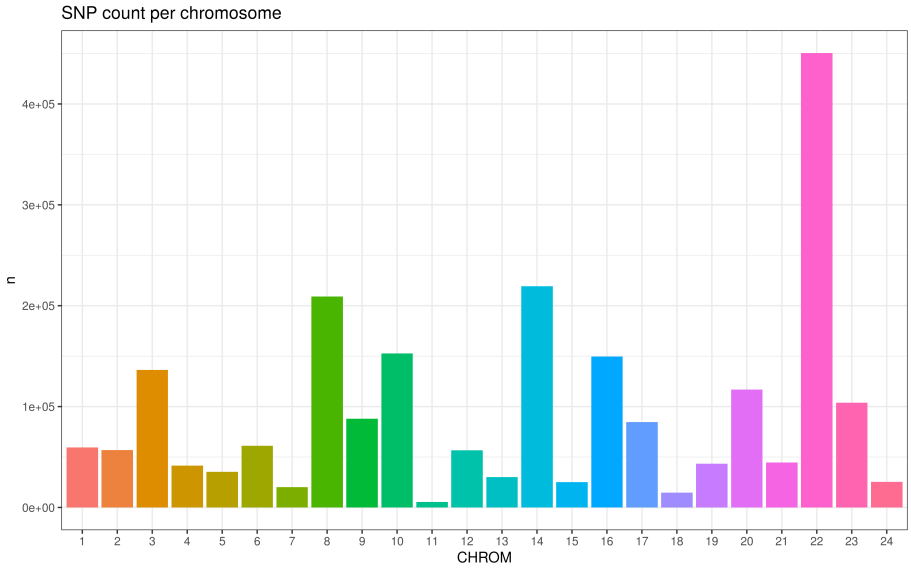


Figure 5.15: Number of SNPs per chromosome that were homozygous-divergent in the F0 *Cab* and *Kaga* generations, and heterozygous in the F1 generation.

value for each bin between 0 and 1, where 0 signified that all reads within that bin supported the *Cab* allele, and 1 signified that all reads within that bin supported the *Kaga* allele. Bins containing no reads were imputed with a value of 0.5.

I then used these values for all F2 individuals as the input to a Hidden Markov Model (HMM) with the software package *hmmlearn* (*Hmmlearn/Hmmlearn* [2014] 2022), which I applied to classify each bin as one of three states, with state 0 corresponding to homozygous-*Cab*, 1 corresponding to heterozygous, and 2 corresponding to homozygous-*Kaga*. Across each chromosome of every sample, the output of the HMM was expected to produce a sequence of states. Based on previous biological knowledge that crossover events occur on average less than once per chromosome (Haenel et al. 2018) (see **Figure 5.16** for the average crossover rates per chromosome in zebrafish), I expected to observe the same state persisting for long stretches of the chromosome, only changing to another state between 0 and 3 times, and rarely more.

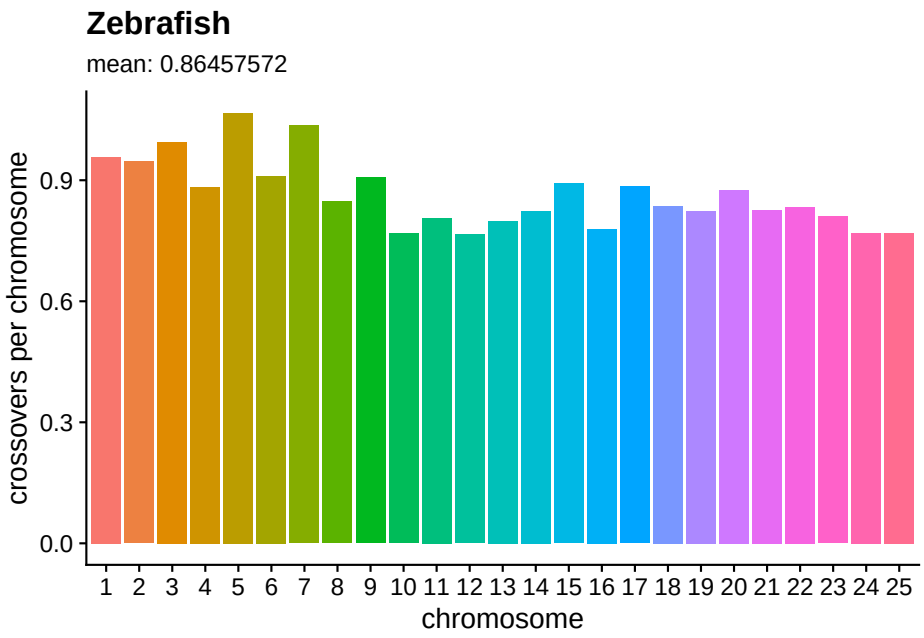


Figure 5.16: Crossovers per chromosome based on data provided in Haenel et al. (2018), where "crossovers per chromosome" for each chromosome c was calculated by $\frac{crossover\ rate_c(cM/Mb) \times length_c(Mb)}{100}$. The medaka genome is shorter in length than zebrafish genome (~800 Mb compared to ~1,300 Mb), which according to the authors would suggest that medaka likely has a higher average crossover rate than what is presented in this figure.

Figure @??fig:hmm-standard) shows how adjusting the HMM parameters changed the called genotypes for 10 F2 samples on chromosome 18. Allowing the HMM to train itself for the transition probabilities and emission variances, the HMM produced an apparently noisy output (**Figure @??fig:hmm-standard)A**). Fixing the transition probabilities to make it very likely for a state to transition back to itself rather than to another state.

I used these genotype-block calls to generate the recombination karyoplots shown in **Figures 5.18** and **5.19**.

Figure 5.20 shows the proportion of 5-kb bins called as either homozygous-*Cab*, heterozygous, or homozygous-*Kaga* within each F2 sample (points). The ordinary expectation for the ratios would be 0.25, 0.5, and 0.25 respectively. However, we observe a skew towards homozygous-*Cab* and away from homozygous *Kaga*. This may have been caused by the low homozygosity observed in *Kaga*.

5.7 Genome-wide linkage anlaysis

Finally, I used the called recombination blocks as pseudo-SNPs in a genetic linkage analysis. To detect associations between the pseudo-SNPs and the three phenotypes of interest, I used a mixed linear model (**MLM**) as implemented in GCTA (Yang et al. 2011). That paper describes the model as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Wu} + \epsilon var(\mathbf{y}) = \mathbf{V} = \mathbf{WW}'\sigma_u^2 + \mathbf{I}\sigma_\epsilon^2$$

Where **y** is a n x 1 vector of phenotypes with n being the sample size, **W** is a standardised genotype matrix, **u** is a vector of SNP effects, and ϵ is a vector of residual effects. I additionally used the leave-one-chromosome-out implementation of GCTA's MLM, with excludes the chromosome on which the candidate SNP is located when calculating the GRM.

As described above in 5.2, the microscope used to image the embryos (either AU or DB) differed by several degrees in heat, which

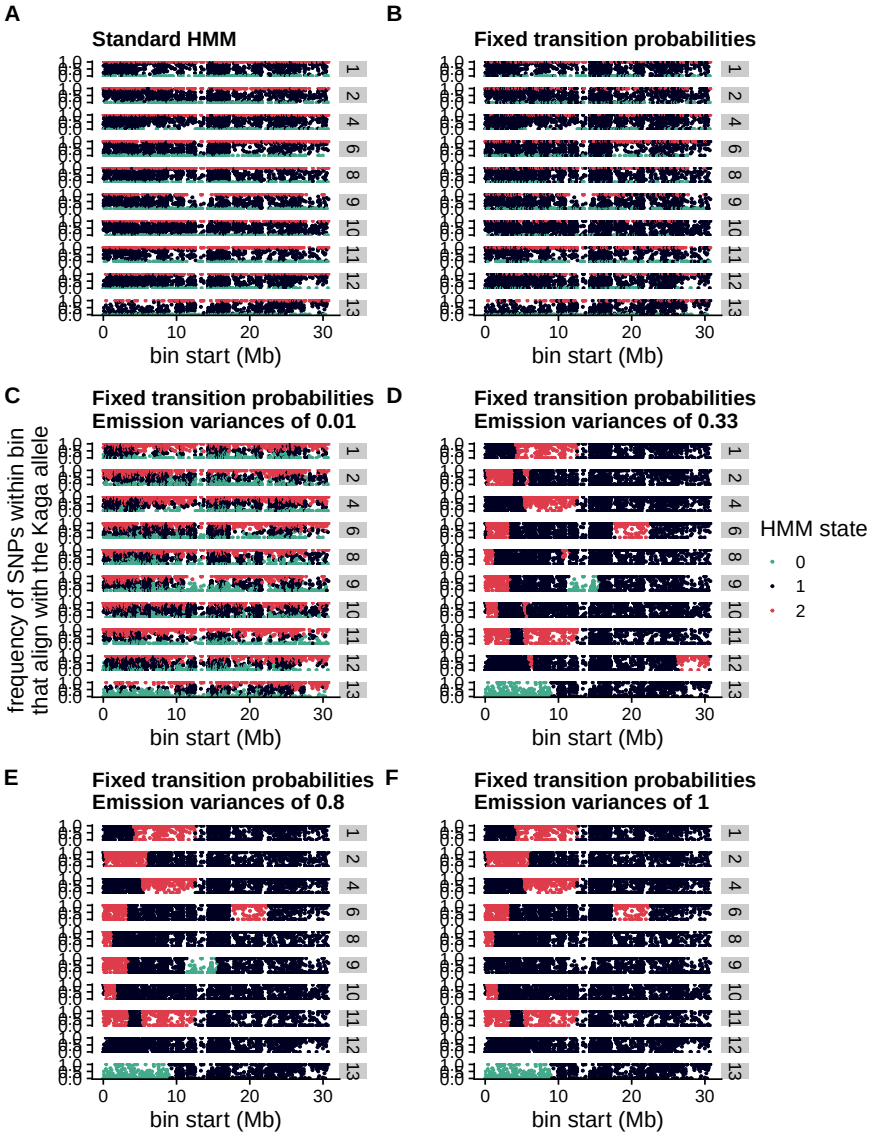


Figure 5.17: HMM states called for each bin across chr18 for 10 F2 samples. States 0, 1, and 2 correspond to homozygous *Cab*, heterozygous, and homozygous *Kaga*. Each point represents a 5-kb bin. Y-axis is the proportion of reads within each bin that align to the *Kaga* allele. X-axis is the bp location of the start of each bin. **A**: Standard HMM with all model parameters trained on the data. **B**. HMM with fixed transition probabilities of 0→0 or 1→1 or 2→2 = 0.999; 0→1 or 2→1 = 0.00066; 0→2 or 2→0 = 0.000333; 1→0 or 1→2 = 0.0005. **C**-**F** retain those transition probabilities but with different fixed emission variances of 0.01 (**C**), 0.33 (**D**), 0.8 (**E**), and 1 (**F**).

Figure 5.18: Recombination blocks in 622 F2 samples based on the ratio of reads mapping to either the *Cab* or *Kaga* allele within 5-kb bins, with homozygous-*Cab* blocks in green, heterozoygous blocks in navy blue, and homozygous *Kaga* blocks in red. Most blocks show 0-2 crossover events, as expected, with some regions showing higher numbers of crossovers interpreted as noise. Unfilled regions are those with no state called by the HMM due to a lack of reads mapping to SNPs within those 5-kb bins.

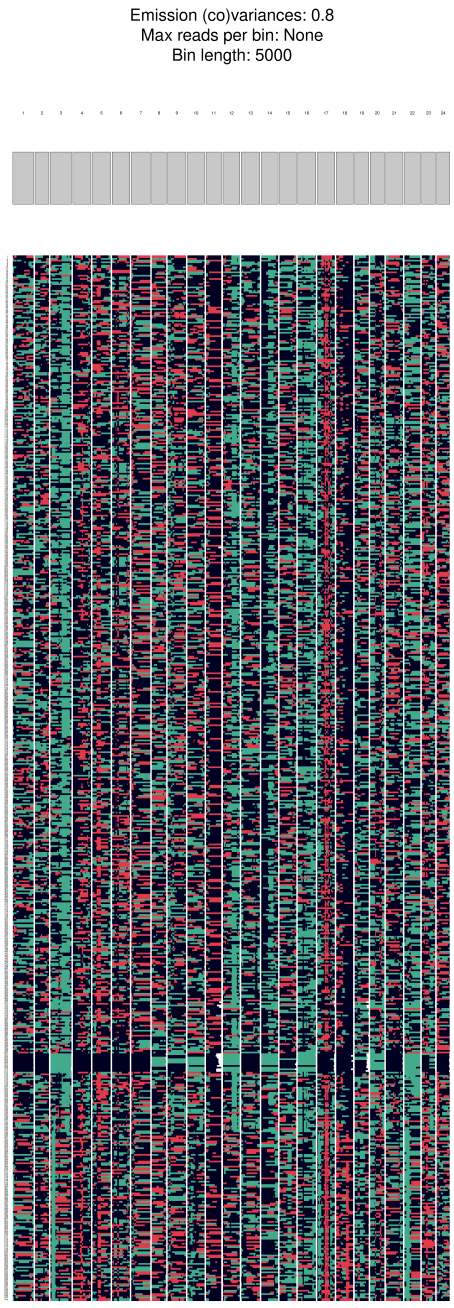


Figure 5.19: (ref:karyo-no-missing)

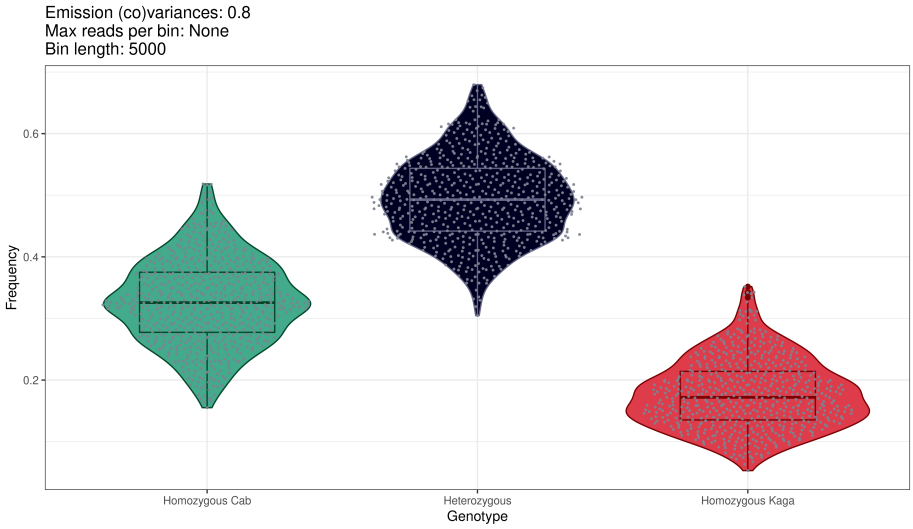


Figure 5.20: Proportions of 5-kb blocks called as either homozygous-*Cab*, heterozygous, or homozygous-*Kaga*.

likely caused differences in the measurements observed. We accordingly experimented with including microscope as a covariate, either alone or together with the genotype for the reporter locus (either homozygous or heterozygous), or excluding it altogether. In an attempt to avoid complications resulting from its inclusion, we also tried inverse-normalising the period phenotype within each microscope group, transforming the phenotype to fit a normal distribution across both microscopes.

To set the significance threshold, I permuted the phenotype across samples using 10 different random seeds, together with all covariates when included, and ran a separate linkage model for each permutation. I then set the lowest p-value from all 10 permutation as the significance threshold for the non-permuted model. I additionally applied a Bonferroni correction to our p-values by dividing α (0.05) by the number of pseudo-SNPs in the model, and set this as a secondary threshold.

Table 5.1: Significant 5-kb bin ranges for period intercept below the minimum p-value from 10 permutations.

Chromosome	Bin start	Bin end	Length (kb)
3	31880001	35420000	3540
4	18090001	18095000	5
10	2995001	3690000	695

5.7.1 Period intercept

Figure 5.21 is a Manhattan plot of the genetic linkage results for the period intercept phenotype, inverse-normalised across microscopes. The regions found to be significant based on the permutations' minimum p-value are set out in **Table 5.1**.

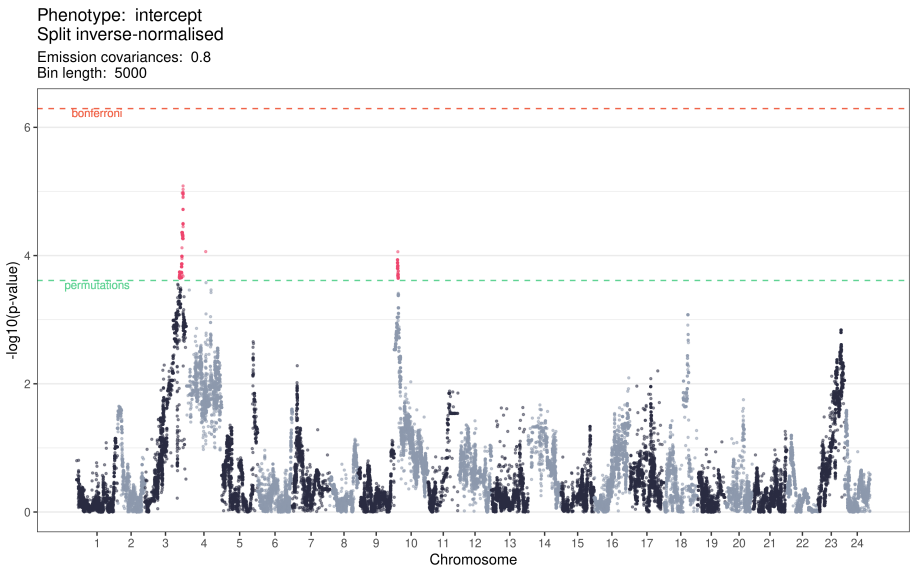


Figure 5.21: Manhattan plot of the genetic linkage results for the period intercept phenotype, inverse-normalised across microscopes. Pseudo-SNPs with p-values lower than the permutation significance threshold are highlighted in red.

These regions contained a total of 46,872 SNPs imputed from the genotype of the F0 parental strains. I ran Ensembl's Variant Effect Predictor (McLaren et al. 2016) over these SNPs to identify those

Table 5.2: Variant Effect Predictor results for SNPs in the bins.

Consequence	Count
intron variant	47211
intergenic variant	20045
upstream gene variant	7304
downstream gene variant	5229
3 prime UTR variant	1082
synonymous variant	694
missense variant	383
5 prime UTR variant	201
splice region variant,intron variant	126
missense variant,splice region variant	19
splice region variant,synonymous variant	17
stop gained	3
splice donor variant	1
start lost	1
stop lost	1
stop lost,splice region variant	1

that would be most likely to have functional consequences. The full counts of SNPs falling into each category of 'consequence' are set out in **Table 5.2**. From this process I identified 38 genes that included a missense variant, 1 that included a missense variant and a start lost (ENSORLG00000014616), ane 1 that included a missense variant and a stop lost (ENSORLG00000015149).

Our collaborators then combined these results with bulk RNA-seq that they had performed on F0 *Cab* and *Kaga* individuals, to determine which of these genes are expressed in the tail during embryogenesis. This allowed them to reduce to the list to 29 genes, and a gene ontology analysis of this found that the list of genes was enriched for body axis, somitogenesis, and segmentation (**Table 5.3**). For this list of genes, our collaborators are now in the process of knocking in the protein-altering *Cab* allele into *Kaga* embryos, and *vice versa*, to functionally validate these variants.

Table 5.3: Target genes for functional validation expressed in the unsegmented PSM and containing protein alterations. Table generated by Ali Seleit.

chromosome_name	ensembl_gene_id	description
3	ENSORLG00000014656	mesoderm posterior protein 2-like
3	ENSORLG00000014659	mesp
10	ENSORLG00000020474	protocadherin 10b
3	ENSORLG00000014616	NA
10	ENSORLG00000020551	FAT atypical cadherin 4
10	ENSORLG00000020531	neurogenin 1
3	ENSORLG00000015149	ADAM metallopeptidase with throm
3	ENSORLG00000015418	matrix metallopeptidase 15
10	ENSORLG00000020488	transforming growth factor beta indu
3	ENSORLG00000028055	NA
10	ENSORLG00000020481	ArfGAP with RhoGAP domain, anky
10	ENSORLG00000020525	TBC1 domain family member 2A-lik
3	ENSORLG00000015260	synapse associated protein 1
10	ENSORLG00000028553	NA
3	ENSORLG00000015460	dpy-19 like C-mannosyltransferase 3
10	ENSORLG00000022010	beta-1,4-galactosyltransferase 1
10	ENSORLG00000020498	protein NipSnap homolog 3A
10	ENSORLG00000020494	nitric oxide associated 1
10	ENSORLG00000020504	ATP-binding cassette sub-family A m
3	ENSORLG00000015118	phosphorylase kinase regulatory sub
10	ENSORLG00000020493	RE1 silencing transcription factor
3	ENSORLG00000015365	solute carrier family 7 member 6 opp
10	ENSORLG00000029052	exosome component 3
3	ENSORLG00000015278	adhesion G protein-coupled recepto
3	ENSORLG00000015287	adhesion G-protein coupled recepto
10	ENSORLG00000025674	matrin-3
10	ENSORLG00000023325	matrin-3
10	ENSORLG00000022388	poly(A) binding protein interacting p
3	ENSORLG00000015096	integrin alpha FG-GAP repeat contai

Table 5.4: Significant 5-kb bin range for PSM area below the minimum p-value from 10 permutations.

CHROM	Bin start	Bin end	Length (kb)
3	20375001	26285000	5910

5.7.2 PSM area

Figure 5.22 is a Manhattan plot of the genetic linkage results for the PSM area phenotype. The regions found to be significant based on the permutations' minimum p-value are set out in **Table 5.4**, although they exceed the Bonferroni correction threshold as well. I note that this ~6 Mb significant region on chromosome 3 does not overlap at all with the significant region discovered for the period intercept phenotype.

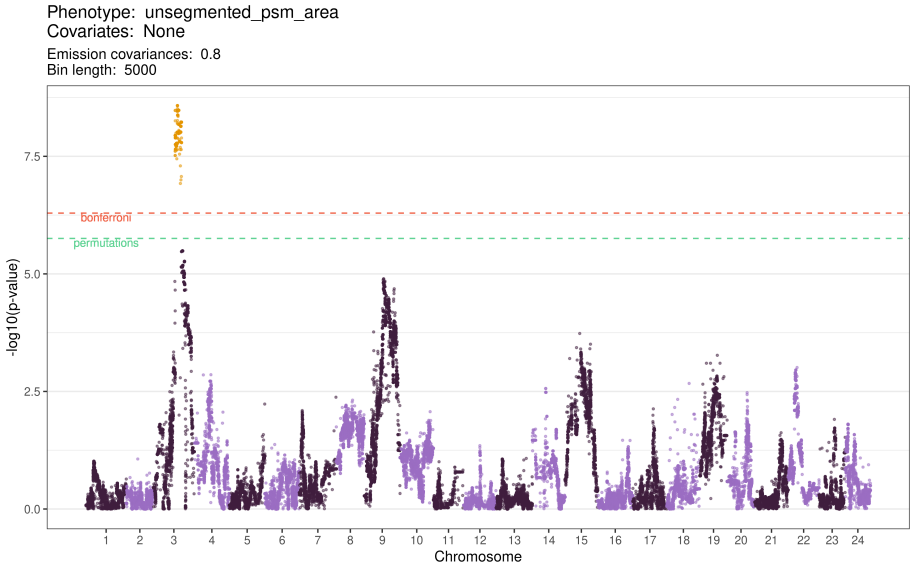


Figure 5.22: Manhattan plot of the genetic linkage results for the PSM area phenotype. Pseudo-SNPs with p-values lower than the permutation significance threshold are highlighted in yellow.

This region contained a total of 29,096 SNPs imputed from the genotype of the F0 parental strains. I ran Ensembl's Variant Effect Predictor (McLaren et al. 2016) over these SNPs to identify those that would

Table 5.5: Variant Effect Predictor results for SNPs in the bins.

Consequence	Count
intron variant	23189
intergenic variant	9171
downstream gene variant	8894
upstream gene variant	8491
3 prime UTR variant	2104
synonymous variant	1141
missense variant	716
5 prime UTR variant	433
splice region variant,intron variant	184
splice region variant,synonymous variant	18
missense variant,splice region variant	7
stop gained	2
splice donor variant	1
start lost	1

be most likely to have functional consequences. The full counts of SNPs falling into each category of 'consequence' are set out in **Table 5.5**. From this process I identified 114 genes that included a missense variant, and 1 that included a both a missense variant and a start lost (ENSORLG00000010863, a centriole, cilia and spindle-associated protein).

Our collaborators then combined these results with bulk RNA-seq that they had performed on F0 *Cab* and *Kaga* individuals, to determine which of these genes are expressed in the unsegmented tail during embryogenesis. This allowed them to reduce to the list to 96 genes, although they were not apparently associated with a specific gene ontology. As with the period intercept phenotype, our collaborators are now in the process of knocking in the *Cab* allele into *Kaga* embryos, and *vice versa*, to functionally validate these variants.

Chapter 6

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

6.1 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.

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

6.1.1.1 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).

6.1.1.2 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).

6.1.2 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.

6.1.2.1 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 6.1.2.4 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 6.1.2.4 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).

6.1.2.2 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; Jonathan K. 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 6.1.3) 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).

6.1.2.3 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 6.1.2.1), 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 6, 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.

6.1.2.4 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.

6.1.3 F_{ST}

The widely-used fixation index (F_{ST}) was introduced independently by Sewall Wright (S. 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 6.1 shows the location and F_{ST} value for all SNPs in the 1000 Genomes dataset, and Figure 6.2 shows their distribution across the range of F_{ST} values.

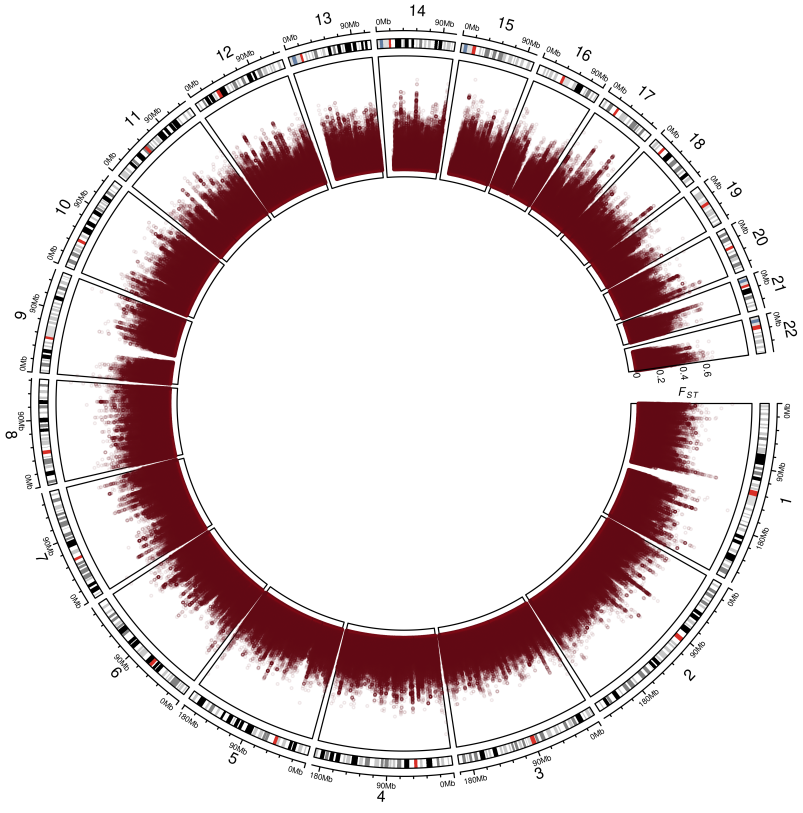


Figure 6.1: (ref:fst-circos)

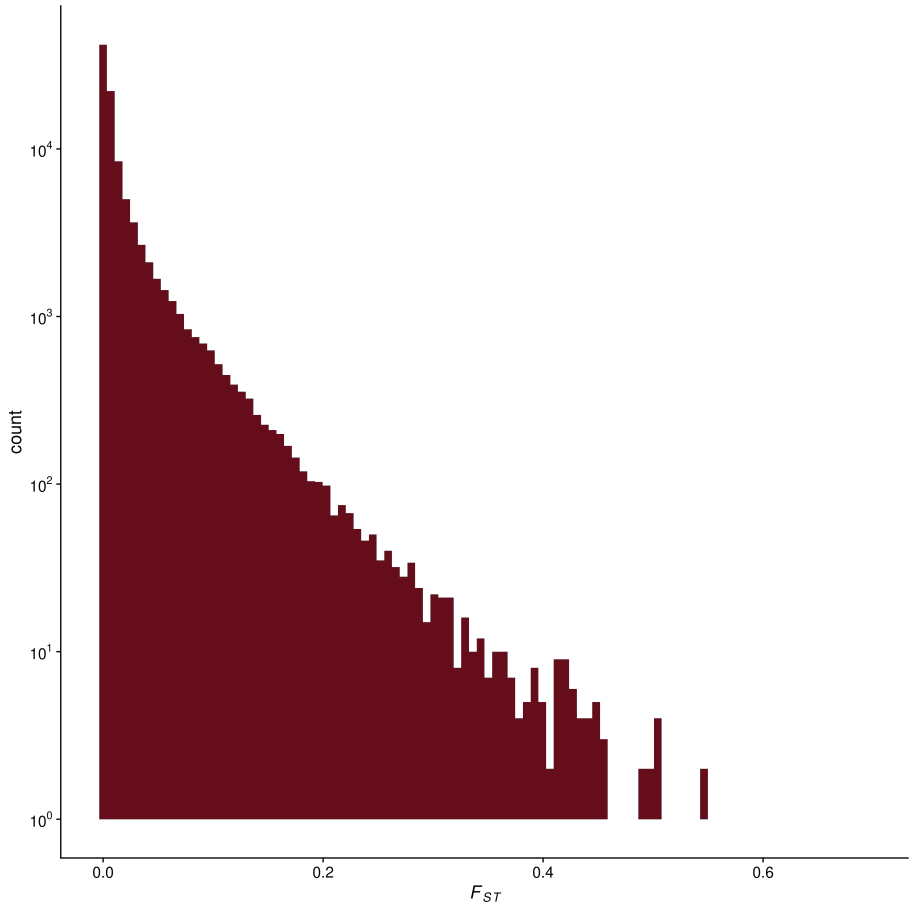


Figure 6.2: (ref:fst-histo)

Figure 6.1 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 6.2 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.

6.2 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 6.2). 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**).

6.2.0.1 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.

6.2.1 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 6.3** 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.

6.2.2 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.

6.2.3 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

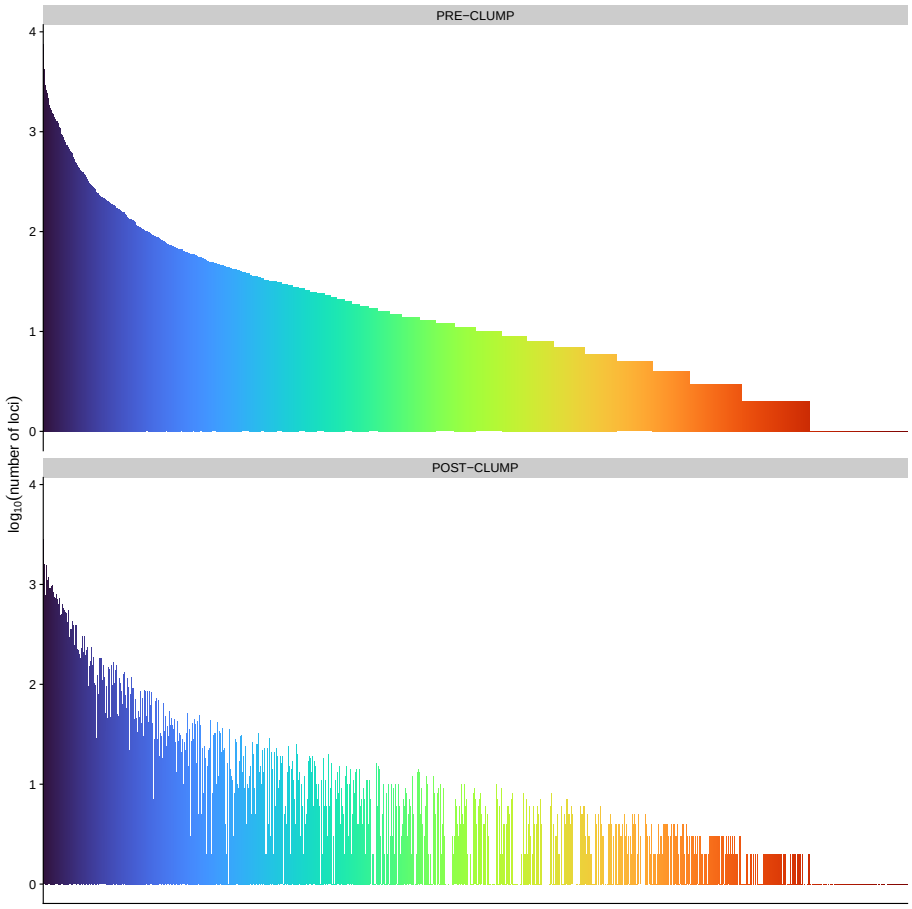


Figure 6.3: log10 counts of associated SNPs for each trait before and after the clumping process, which involved: a) excluding all SNPs with a p-value greater than 10^{-8}; and b) starting with the SNPs with the lowest p-values ("index" SNPs), excluding all other SNPs within a 1 Mb region of the index 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 6.4** 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 6.4A** shows the densities of SNPs as a function of F_{ST}, and **Figure 6.4B and C** show their empirical Cumulative Distribution Functions (eCDFs). **Figure 6.4B** includes the eCDFs of control SNPs in grey. eCDF figures for all 587 traits that passed our filters (Methods) are provided in the Appendix B.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 6.5. This figure shows that GWAS performed for the more comprehensively-studied phenotypes such as body height,

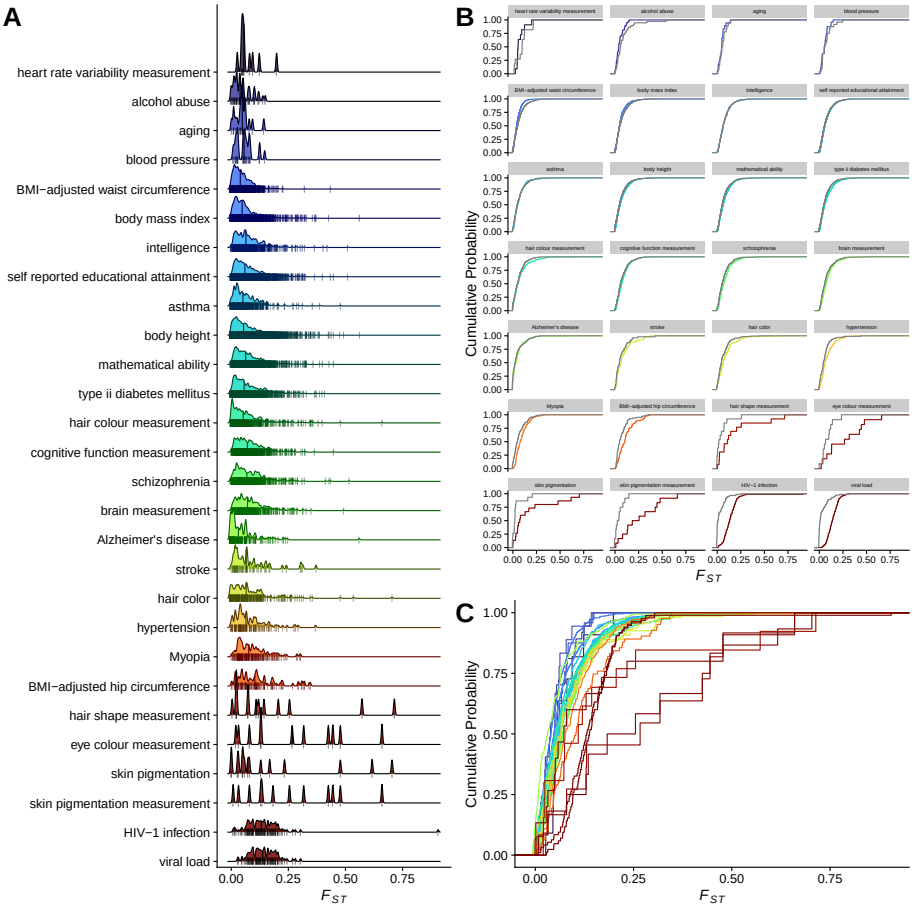


Figure 6.4: Distributions of F_{ST} across 28 illustrative human traits, ranked by signed-D (Kolmogorov-Smirnov test) comparing trait and control SNPs. **A**. F_{ST} density ridge plots with SNP markers. **B**. Empirical Cumulative Distribution Functions (eCDFs) of F_{ST} for trait-associated (colour) 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.

BMI, and educational attainment, included individuals from many 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 6.5.

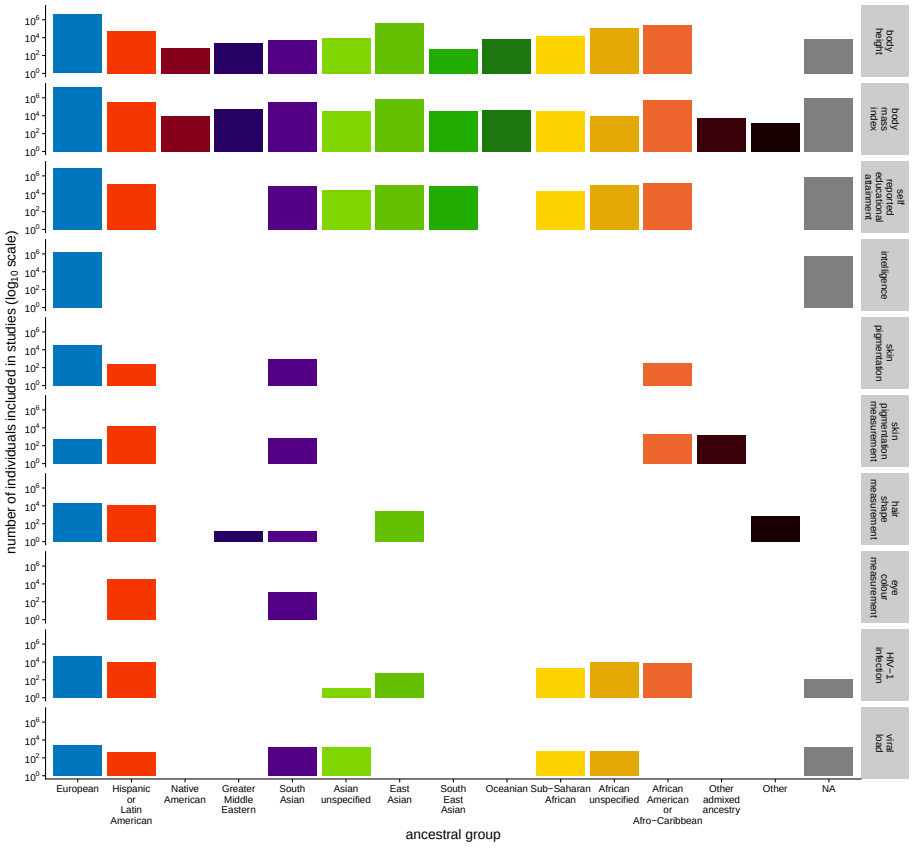


Figure 6.5: (ref:fst-n-indivs)

6.3 Implications

6.3.1 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.

6.3.2 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.

Abecasis, G. R., L. R. Cardon, and W. O. C. Cookson. 2000. "A General Test of Association for Quantitative Traits in Nuclear Families." *The American Journal of Human Genetics* 66 (1): 279–92. https://doi.org/10.1086/302698.

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.

Blake, Judith A, Richard Baldarelli, James A Kadin, Joel E Richardson, Cynthia L Smith, Carol J Bult, and the Mouse Genome Database Group. 2021. "Mouse Genome Database (MGD): Knowledgebase for Mouse–Human Comparative Biology." *Nucleic Acids Research* 49 (D1): D981–87. https://doi.org/10.1093/nar/gkaa1083.

Brown, C., F. Jones, and V. A. Braithwaite. 2007. "Correlation Between Boldness and Body Mass in Natural Populations of the Poeciliid Brachyrhaphis Episcopi." *Journal of Fish Biology* 71 (6): 1590–1601. https://doi.org/10.1111/j.1095-8649.2007.01627.x.

Brown, Culum, Fiona Burgess, and Victoria A. Braithwaite. 2007. "Heritable and Experiential Effects on Boldness in a Tropical Poeciliid." *Behavioral Ecology and Sociobiology* 62 (2): 237–43. https://doi.org/10.1007/s00265-007-0458-3.

Buss, David M. 1991. "Evolutionary Personality Psychology." *Annual Review of Psychology* 42: 459–91. https://doi.org/10.1146/annurev.ps.42.020191.002331.

Carola, Valeria, Olivier Mirabeau, and Cornelius T. Gross. 2011. "Hidden Markov Model Analysis of Maternal Behavior Patterns in Inbred and Reciprocal Hybrid Mice." *PLOS ONE* 6 (3): e14753. https://doi.org/10.1371/journal.pone.0014753.

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.

Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. "Twelve Years of SAMtools and BCFtools." *GigaScience* 10 (2): giab008. https://doi.org/10.1093/gigascience/giab008.

Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life*. J. Murray. https://books.google.com?id=gGUYt2PsDJcC.

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran V. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data." *Nature Genetics* 43 (5, 5): 491–98. https://doi.org/10.1038/ng.806.

Devlin, B., and Kathryn Roeder. 1999. "Genomic Control for Association Studies." *Biometrics 55* (4): 997–1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x.

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.

Ellen, Esther D., T. Bas Rodenburg, Gerard A. A. Albers, J. Elizabeth Bolhuis, Irene Camerlink, Naomi Duijvesteijn, Egbert F. Knol, et al. 2014. "The Prospects of Selection for Social Genetic Effects to Improve Welfare and Productivity in Livestock." *Frontiers in Genetics* 5. https://www.frontiersin.org/articles/10.3389/fgene.2014.00377.

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.

Ewens, W J, and R S Spielman. 1995. "The Transmission/Disequilibrium Test: History, Subdivision, and Admixture." *American Journal of Human Genetics* 57 (2): 455–64. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1801556/.

Falk, Henning J, Takehito Tomita, Gregor Mönke, Katie McDole, and Alexander Aulehla. 2022. "Imaging the Onset of Oscillatory Signaling Dynamics During Mouse Embryo Gastrulation." *Development (Cambridge, England)* 149 (13): dev200083. https://doi.org/10.1242/dev.200083.

Fisher, R. A. 1919. "XV.—The Correlation Between Relatives on the Supposition of Mendelian Inheritance." January. https://doi.org/10.1017/s0080456800012163.

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.

Galton, Francis, and S. Okamoto. 1874. "On Men of Science, Their Nature and Their Nurture." In. Royal Institution of Great Britain.

Gomez, C'eline, Ertuğrul M. Özbudak, Joshua Wunderlich, Diana Baumann, Julian Lewis, and Olivier Pourqui'e. 2008. "Control of Segment Number in Vertebrate Embryos." *Nature* 454 (7202, 7202): 335–39. https://doi.org/10.1038/nature07020.

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.

Gridley, Thomas. 2006. "The Long and Short of It: Somite Formation in Mice." *Developmental Dynamics* 235 (9): 2330–36. https://doi.org/10.1002/dvdy.20850.

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.

Gu, Zuguang, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. 2014. "Circlize Implements and Enhances Circular Visualization in R." *Bioinformatics* 30 (19): 2811–12. https://doi.org/10.

1093/bioinformatics/btu393.

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.

Haenel, Quiterie, Telma G. Laurentino, Marius Roesti, and Daniel Berner. 2018. "Meta-Analysis of Chromosome-Scale Crossover Rate Variation in Eukaryotes and Its Significance to Evolutionary Genomics." *Molecular Ecology* 27 (11): 2477–97. https://doi.org/10.1111/mec.14699.

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.

Hills, L. Benjamin, Amira Masri, Kotaro Konno, Wataru Kakegawa, Anh-Thu N. Lam, Elizabeth Lim-Melia, Nandini Chandy, et al. 2013. "Deletions in Grid2 Lead to a Recessive Syndrome of Cerebellar Ataxia and Tonic Upgaze in Humans." *Neurology* 81 (16): 1378–86. https://doi.org/10.1212/WNL.0b013e3182a841a3.

Hmmlearn/Hmmlearn. (2014) 2022. hmmlearn. https://github.com/hmmlearn/hmmlearn.

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.

Hubaud, Alexis, and Olivier Pourqui'e. 2014. "Signalling Dynamics in Vertebrate Segmentation." *Nature Reviews Molecular Cell Biology* 15 (11, 11): 709–21. https://doi.org/10.1038/nrm3891.

"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 /Vol1/Ftp/Data_collections/1000g_2504_high_coverage/Working/20201028_3202_phas n.d. Accessed March 24, 2022. http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/.

Jamain, St'ephane, H'el'ene Quach, Catalina Betancur, Maria Råstam, Catherine Colineaux, I. Carina Gillberg, Henrik Soderstrom, et al. 2003. "Mutations of the X-linked Genes Encoding Neuroligins Nlgn3 and Nlgn4 Are Associated with Autism." *Nature Genetics* 34 (1, 1): 27–29. https://doi.org/10.1038/ng1136.

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.

Junghans, Dirk, Matthias Heidenreich, Iris Hack, Verdon Taylor, Michael Frotscher, and Rolf Kemler. 2008. "Postsynaptic and Differential Localization to Neuronal Subtypes of Protocadherin []16 in the Mammalian Central Nervous System." *European Journal of Neuroscience* 27 (3): 559–71. https://doi.org/10.1111/j.1460-9568.2008.06052.x.

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.

Khanna, Ajay, David E. Larson, Sridhar Nonavinkere Srivatsan, Matthew Mosior, Travis E. Abbott, Susanna Kiwala, Timothy J. Ley, et al. 2022. "Bam-Readcount - Rapid Generation of Basepair-Resolution Sequence Metrics." *Journal of Open Source Software* 7 (69): 3722. https://doi.org/10.21105/joss.03722.

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.

Kim, Woong, Takaaki Matsui, Masataka Yamao, Makoto Ishibashi, Kota Tamada, Toru Takumi, Kenji Kohno, et al. 2011. "The Period of the Somite Segmentation Clock Is Sensitive to Notch Activity." *Molecular Biology of the Cell* 22 (18): 3541–49. https://doi.org/10.1091/mbc.e11-02-0139.

Kimura, Tetsuaki, Minori Shinya, and Kiyosi Naruse. 2012. "Genetic Analysis of Vertebral Regionalization and Number in Medaka (Oryzias Latipes) Inbred Lines." *G3 Genes|Genomes|Genetics* 2 (11): 1317–23. https://doi.org/10.1534/g3.112.003236.

Kirchmaier, Stephan, Kiyoshi Naruse, Joachim Wittbrodt, and Felix Loosli. 2015. "The Genomic and Genetic Toolbox of the Teleost Medaka (Oryzias Latipes)." *Genetics* 199 (4): 905–18. https://doi.org/10.1534/genetics.114.173849.

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.

L'opez-Olmeda, Jose Fernando, Haiyu Zhao, Markus Reischl, Christian Pylatiuk, Tyrone Lucon-Xiccato, Felix Loosli, and Nicholas S. Foulkes. 2021. "Long Photoperiod Impairs Learning in Male but Not Female Medaka." *iScience* 24 (7): 102784. https://doi.org/10.1016/j.isci.2021.102784.

Laland, Kevin, Jens Krause, and Culum Brown. 2011. *Fish Cognition and Behavior*. John Wiley & Sons. https://books.google.com?id=cI9gbVyH6lsC.

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.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map (SAM) Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Lima, Steven L., and Lawrence M. Dill. 1990. "Behavioral Decisions Made Under the Risk of Predation: A Review and Prospectus." *Canadian Journal of Zoology* 68 (4): 619–40. https://doi.org/10.1139/z90-092.

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.

Lucon-Xiccato, Tyrone, and Angelo Bisazza. 2017. "Individual Differences in Cognition Among Teleost Fishes." *Behavioural Processes*, The Cognition of Fish, 141 (August): 184–95. https://doi.org/10.1016/j.beproc.2017.01.015.

Lucon-Xiccato, Tyrone, Francesca Conti, Felix Loosli, Nicholas S. Foulkes, and Cristiano Bertolucci. 2020. "Development of Open-Field Behaviour in the Medaka, Oryzias Latipes." *Biology* 9 (11, 11): 389. https://doi.org/10.3390/biology9110389.

Lucon-Xiccato, Tyrone, Felix Loosli, Francesca Conti, Nicholas S. Foulkes, and Cristiano Bertolucci. 2022. "Comparison of Anxiety-Like and Social Behaviour in Medaka and Zebrafish." *Scientific Reports* 12 (1, 1): 10926. https://doi.org/10.1038/s41598-022-14978-1.

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.

Matsuda, Mitsuhiro, Hanako Hayashi, Jordi Garcia-Ojalvo, Kumiko Yoshioka-Kobayashi, Ryoichiro Kageyama, Yoshihiro Yamanaka, Makoto Ikeya, Junya Toguchida, Cantas Alev, and Miki Ebisuya. 2020. "Species-Specific Segmentation Clock Periods Are Due to Differential Biochemical Reaction Speeds." *Science* 369 (6510): 1450–55.

Matsunaga, Wataru, and Eiji Watanabe. 2010. "Habituation of Medaka (Oryzias Latipes) Demonstrated by Open-Field Testing." *Behavioural Processes* 85 (2): 142–50. https://doi.org/10.1016/j.beproc.2010.06.019.

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.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. "The Ensembl Variant Effect Predictor." *Genome Biology* 17 (1): 122. https://doi.org/10.1186/s13059-016-0974-4.

Members of the Complex Trait Consortium. 2003. "The Nature and Identification of Quantitative Trait Loci: A Community's View." *Nature Reviews Genetics* 4 (11, 11): 911–16. https://doi.org/10.1038/nrg1206.

Mitchell, Kevin J. 2007. "The Genetics of Brain Wiring: From Molecule to Mind." *PLOS Biology 5* (4): e113. https://doi.org/10.1371/journal.pbio.0050113.

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.

Phillips, Greg R., Hidekazu Tanaka, Marcus Frank, Alice Elste, Lazar Fidler, Deanna L. Benson, and David R. Colman. 2003. "[]-Protocadherins Are Targeted to Subsets of Synapses and Intracellular Organelles in Neurons." *Journal of Neuroscience* 23 (12): 5096–5104. https://doi.org/10.1523/JNEUROSCI.23-12-05096.2003.

"Picard Toolkit." 2019. *Broad Institute, GitHub Repository*. https://broadinstitute.github.io/picard/; Broad Institute.

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.

Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, et al. 2018. "Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples." bioRxiv. https://doi.org/10.1101/201178.

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.

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155 (2): 945–59. https://doi.org/10.1093/genetics/155.2.945.

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.

Raterman, S. T., J. R. Metz, Frank A. D. T. G. Wagener, and Johannes W. Von den Hoff. 2020. "Zebrafish Models of Craniofacial Malformations: Interactions of Environmental Factors." *Frontiers in Cell and Developmental Biology* 8.

https://www.frontiersin.org/articles/10.3389/fcell.2020.600926.

Romero-Ferrero, Francisco, Mattia G. Bergomi, Robert C. Hinz, Francisco J. H. Heras, and Gonzalo G. de Polavieja. 2019. "Idtracker.ai: Tracking All Individuals in Small or Large Collectives of Unmarked Animals." *Nature Methods* 16 (2, 2): 179–82. https://doi.org/10.1038/s41592-018-0295-5.

Ruxton, Graeme D. 2006. "The Unequal Variance t-Test Is an Underused Alternative to Student's t-Test and the Mann–Whitney U Test." *Behavioral Ecology* 17 (4): 688–90. https://doi.org/10.1093/beheco/ark016.

Ruzzante, D. E., and R. W. Doyle. 1990. "Behavioural and Growth Responses to the Intensity of Intraspecific Social Interaction Among Medaka, Oryzias Latipes (Temminck and Schlegel) (Pisces, Cyprinodontidae)." *Journal of Fish Biology* 37 (5): 663–73. https://doi.org/10.1111/j.1095-8649.1990.tb02531.x.

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.

Schjolden, Joachim, Argaudas Stoskhus, and Svante Winberg. 2005. "Does Individual Variation in Stress Responses and Agonistic Behavior Reflect Divergent Stress Coping Strategies in Juvenile Rainbow Trout?" *Physiological and Biochemical Zoology* 78 (5): 715–23. https://doi.org/10.1086/432153.

Schmal, Christoph, Gregor Mönke, and Adri'an E. Granada. 2022. "Analysis of Complex Circadian Time Series Data Using Wavelets." In *Circadian Regulation: Methods and Protocols*, edited by Guiomar Solanas and Patrick -Simon Welz, 35–54.

Methods in Molecular Biology. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-2249-0_3.

Schneider, Jonathan, Jade Atallah, and Joel D. Levine. 2017. "Social Structure and Indirect Genetic Effects: Genetics of Social Behaviour." *Biological Reviews* 92 (2): 1027–38. https://doi.org/10.1111/brv.12267.

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.

Seleit, Ali, Alexander Aulehla, and Alexandre Paix. 2021. "Endogenous Protein Tagging in Medaka Using a Simplified CRISPR/Cas9 Knock-in Approach." *eLife* 10 (December): e75050. https://doi.org/10.7554/elife.75050.

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.

Sloan Wilson, David, Anne B. Clark, Kristine Coleman, and Ted Dearstyne. 1994. "Shyness and Boldness in Humans and Other Animals." *Trends in Ecology & Evolution* 9 (11): 442–46. https://doi.org/10.1016/0169-5347(94)90134-1.

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.

Smith, Jennifer R, G Thomas Hayman, Shur-Jen Wang, Stanley J F Laulederkind, Matthew J Hoffman, Mary L Kaldunski, Monika Tutaj, et al. 2020. "The Year of the Rat: The Rat Genome Database at 20: A Multi-Species Knowledgebase and Analysis Platform." *Nucleic Acids Research* 48 (D1): D731–42. https://doi.org/10.1093/nar/gkz1041.

Snoek, Basten L., Rita J. M. Volkers, Harm Nijveen, Carola Petersen, Philipp Dirksen, Mark G. Sterken, Rania Nakad, et al. 2019. "A Multi-Parent Recombinant Inbred Line Population of C. Elegans Allows Identification of Novel QTLs for Complex Life History Traits." *BMC Biology* 17 (1): 24. https://doi.org/10.1186/s12915-019-0642-8.

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.

Svartberg, Kenth. 2002. "Shyness–Boldness Predicts Performance in Working Dogs." *Applied Animal Behaviour Science* 79 (2): 157–74. https://doi.org/10.1016/S0168-1591(02)00120-X.

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.

Turley, Patrick, Raymond K. Walters, Omeed Maghzian, Aysu Okbay, James J. Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, et al. 2018. "Multi-Trait Analysis of Genome-Wide Association Summary Statistics Using MTAG." *Nature Genetics* 50 (2, 2): 229–37. https://doi.org/10.1038/s41588-017-0009-4.

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.

Utine, G. Eda, Göknur Haliloğlu, Bilge Salancı, Arda Çetinkaya, P. Özlem Kiper, Yasemin Alanay, Dilek Aktaş, Koray Boduroğlu, and

Mehmet Alikaşifoğlu. 2013. "A Homozygous Deletion in Grid2 Causes a Human Phenotype With Cerebellar Ataxia and Atrophy." *Journal of Child Neurology* 28 (7): 926–32. https://doi.org/10.1177/0883073813484967.

Van der Auwera, Geraldine A., and Brian D. O'Connor. 2020. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media.

Vasimuddin, Md, Sanchit Misra, Heng Li, and Srinivas Aluru. 2019. "Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems." In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 314–24. IEEE.

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.

"What Was Megatherium?" n.d. Accessed August 17, 2022. https://www.nhm.ac.uk/discover/what-was-megatherium.html.

Wichura, Michael J. 1988. "Algorithm AS 241: The Percentage Points of the Normal Distribution." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 37 (3): 477–84. https://doi.org/10.2307/2347330.

Wilson, David S., Kristine Coleman, Anne B. Clark, and Laurence Biederman. 1993. "Shy-Bold Continuum in Pumpkinseed Sunfish (Lepomis Gibbosus): An Ecological Study of a Psychological Trait." *Journal of Comparative Psychology* 107 (3): 250–60. https://doi.org/10.1037/0735-7036.107.3.250.

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, Dominic, Roger K. Butlin, and Örjan Carlborg. 2006. "Epistatic Regulation of Behavioural and Morphological Traits in the Zebrafish (Danio Rerio)." *Behavior Genetics* 36 (6): 914–22. https://doi.org/10.1007/s10519-006-9080-9.

Wright, D., L. B. Rimmer, V. L. Pritchard, R. K. Butlin, and J. Krause. 2003. "Inter and Intra-Population Variation in Shoaling and Boldness in the Zebrafish (Danio Rerio)." *Journal of Fish Biology* 63 (s1): 258–59. https://doi.org/10.1111/j.1095-8649.2003.216bw.x.

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.

Yang, Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. 2011. "GCTA: A Tool for Genome-wide Complex Trait Analysis." *The American Journal of Human Genetics* 88 (1): 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

Yengo, Loic, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, An-

drew 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.

Young, Simon N. 2008. "The Neurobiology of Human Social Behaviour: An Important but Neglected Topic." *Journal of Psychiatry and Neuroscience* 33 (5): 391–92. https://www.jpn.ca/content/33/5/391.

Zhang, Zhiwu, Elhan Ersoz, Chao-Qiang Lai, Rory J. Todhunter, Hemant K. Tiwari, Michael A. Gore, Peter J. Bradbury, et al. 2010. "Mixed Linear Model Approach Adapted for Genome-Wide Association Studies." *Nature Genetics* 42 (4, 4): 355–60. https://doi.org/10.1038/ng.546.

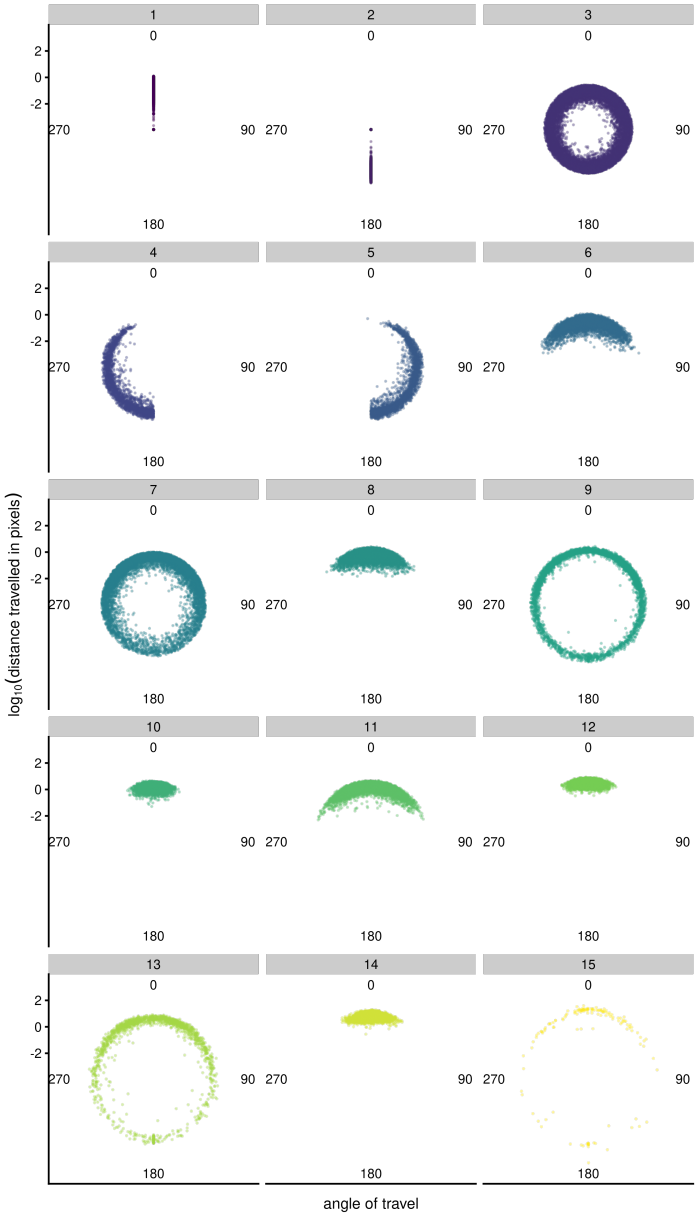


Figure A.1: HMM states predicted for the F0 dataset, using the combination of a 0.05-second interval between which the distance and angle variables were calculated.

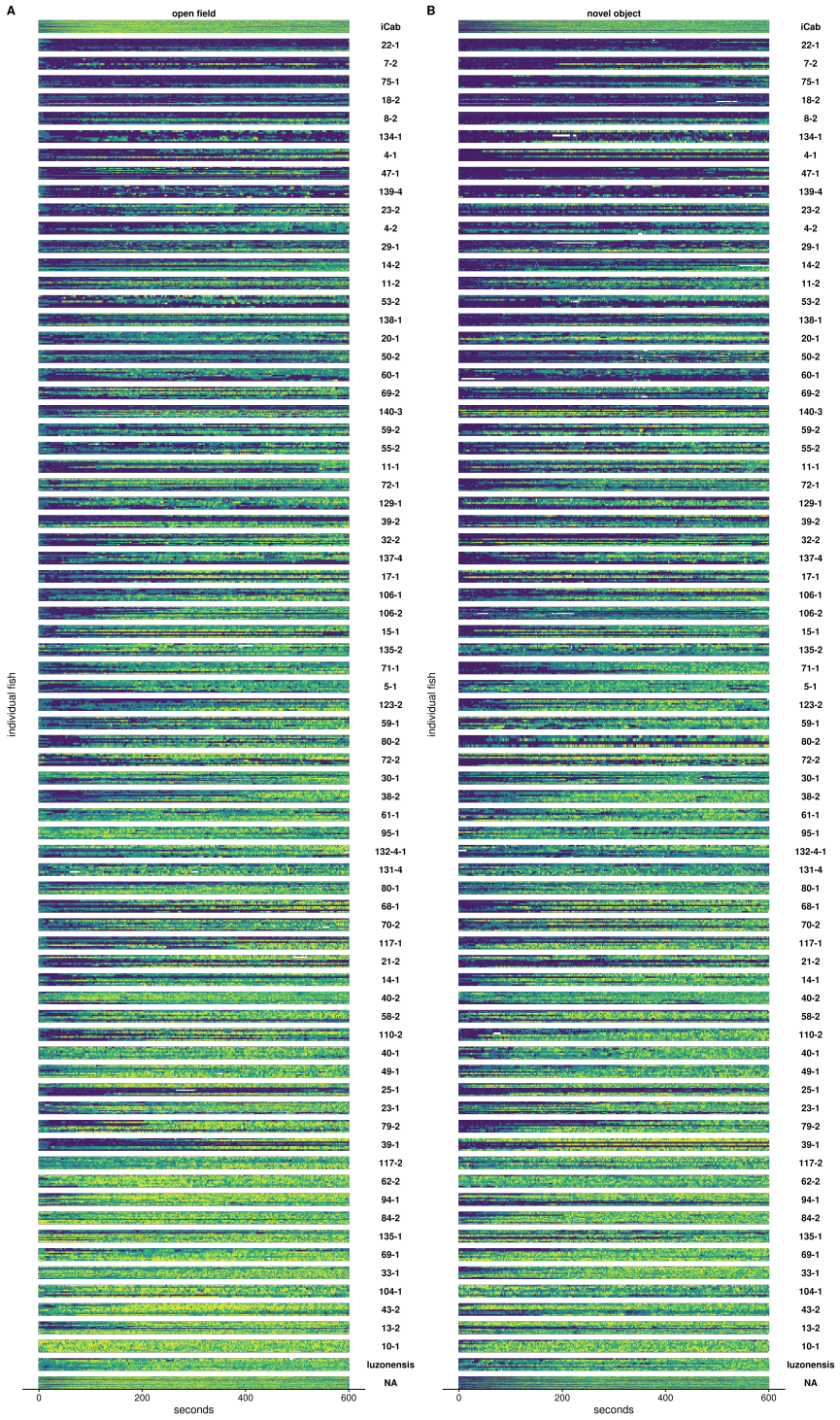


Figure A.2: Tile plot for all 1610 test fishes included in the MIKK panel behaviour analysis, ordered by each line's group median for individual mean speed over the course of the 20-minute video (open field and novel object combined). The order of lines is identical to that shown in **Figure 4.2**.

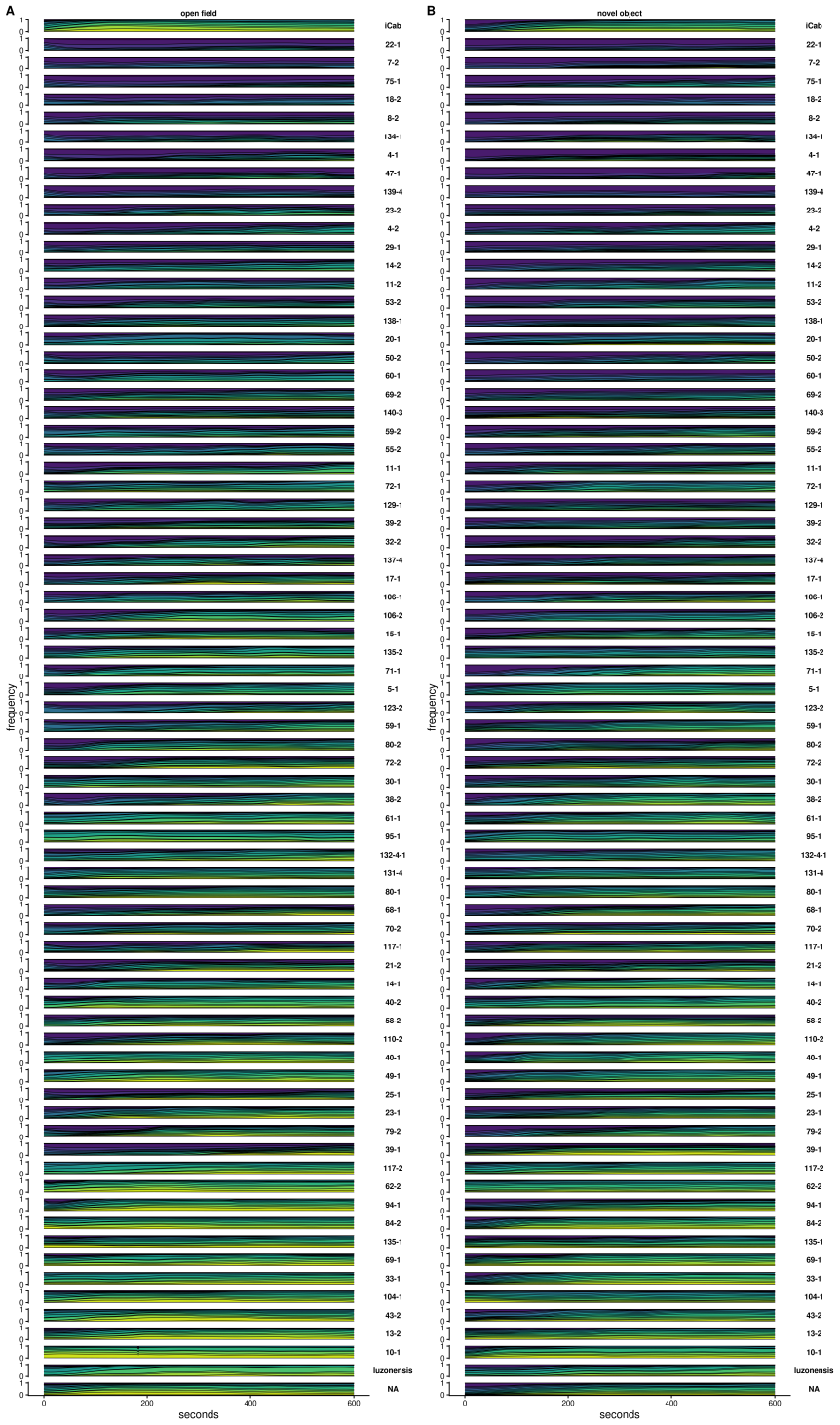


Figure A.3: Tile plot for all 1610 test fishes included in the MIKK panel behaviour analysis, ordered by each line's group median for individual mean speed over the course of the 20-minute video (open field and novel object combined). The order of lines is identical to that shown in **Figure 4.2**.

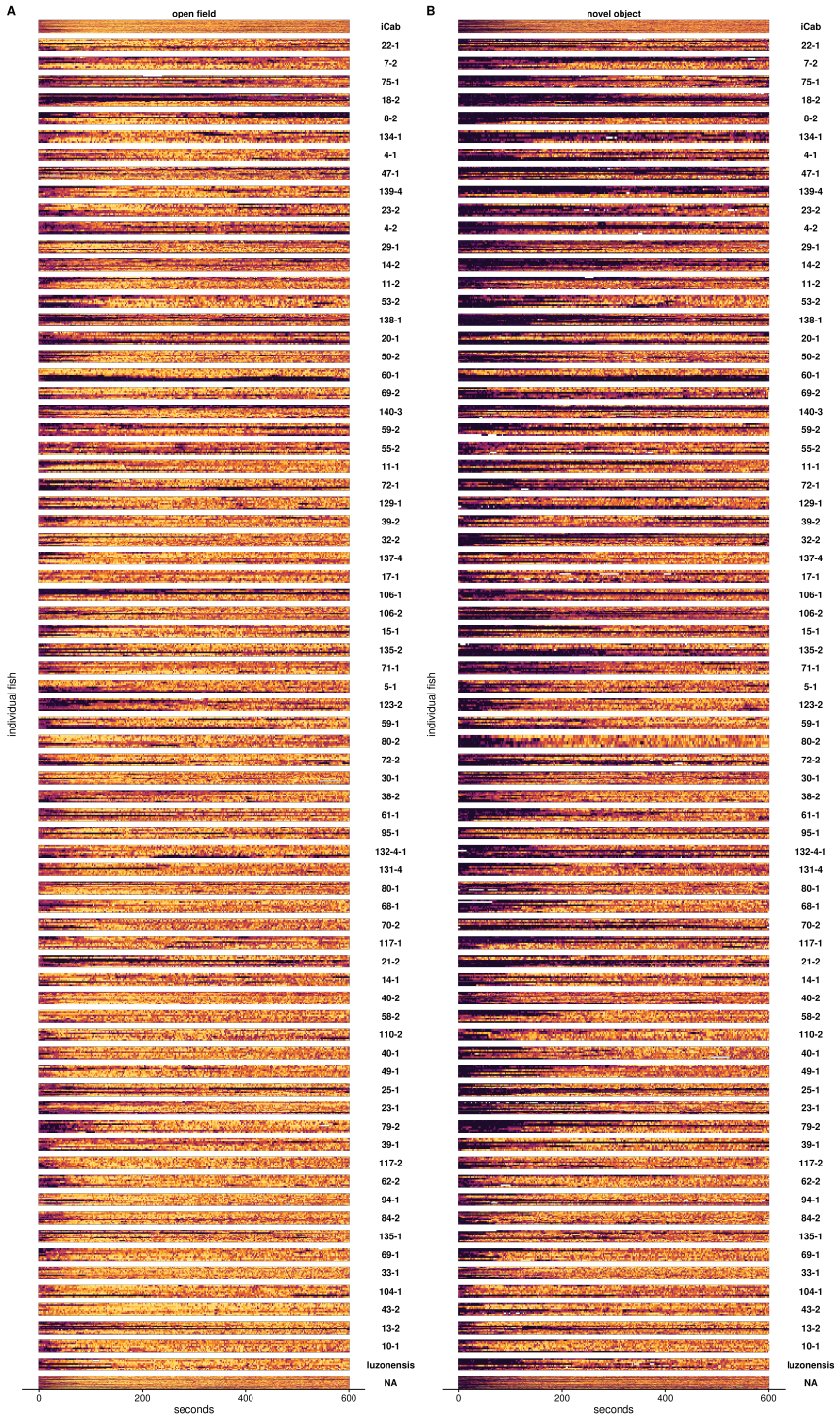


Figure A.4: Tile plot for all 1610 reference fishes included in the MIKK panel behaviour analysis, ordered by each line's group median for individual mean speed over the course of the 20-minute video (open field and novel object combined). The order of lines is identical to that shown in **Figure 4.2**.

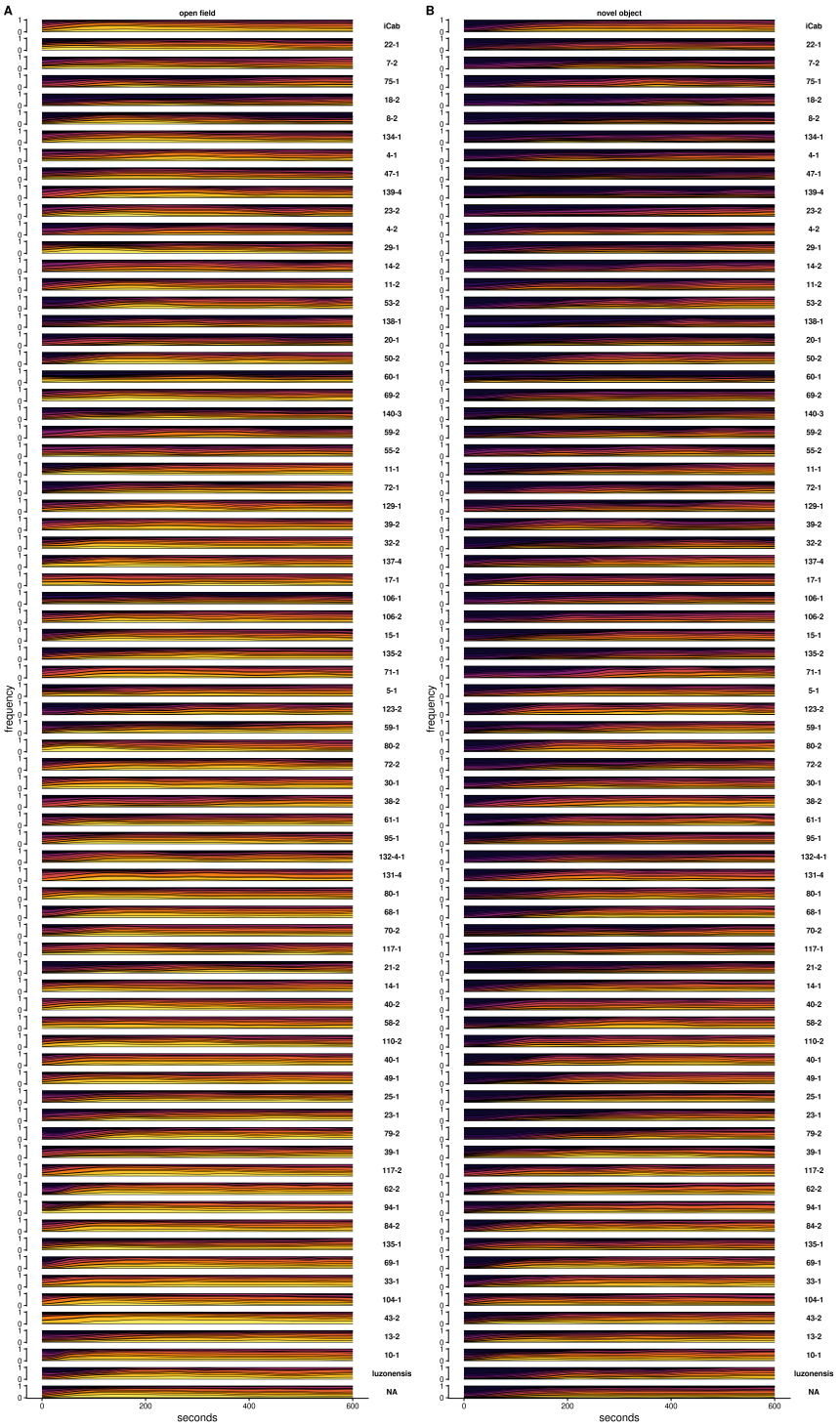


Figure A.5: Tile plot for all 1610 reference fishes included in the MIKK panel behaviour analysis, ordered by each line's group median for individual mean speed over the course of the 20-minute video (open field and novel object combined). The order of lines is identical to that shown in **Figure 4.2**.

Appendix A

Chapter 4: supplementary information

A.1 15 HMM states with 0.05 second interval

A.2 HMM state time dependence for all MIKK panel lines

A.2.1 Direct genetic effects

A.2.1.1 Tile plot

A.2.1.2 Density plot

A.2.2 Social genetic effects

A.2.2.1 Tile plot

A.2.2.2 Density plot

A.3 F2 recombination karyoplot with missing calls

A.4 LOCO GRM for chromosome 1

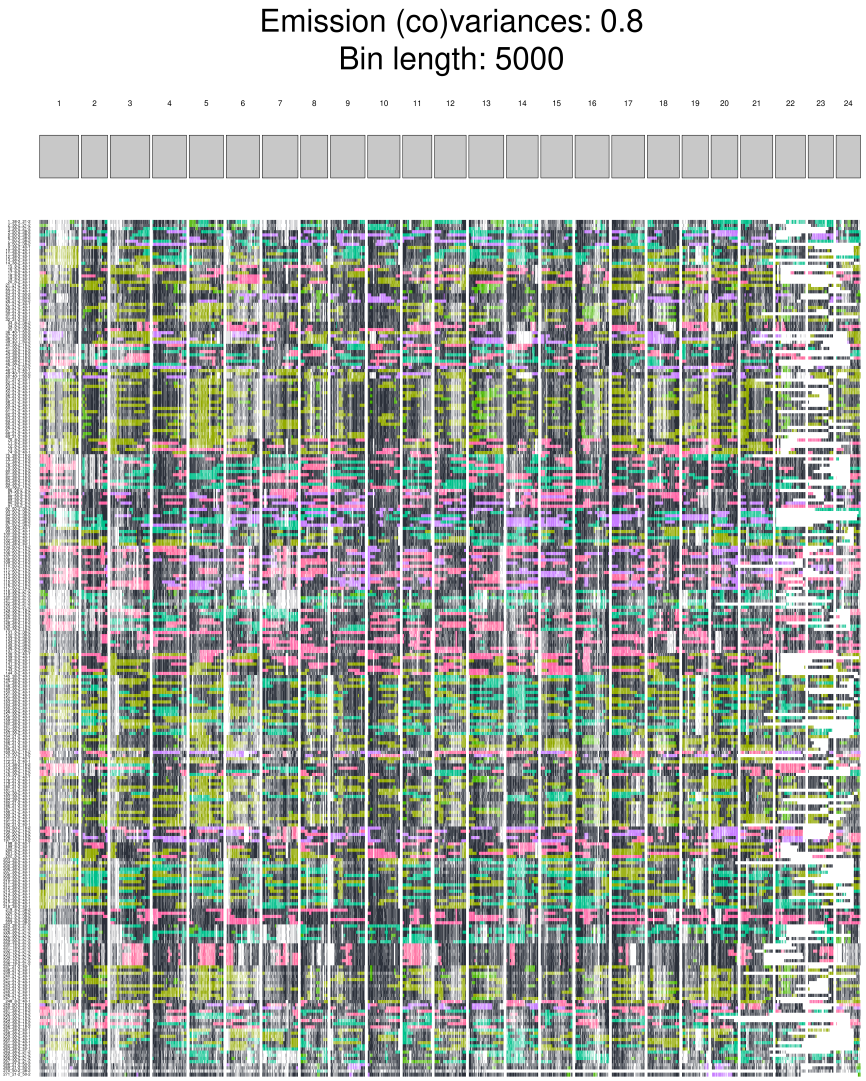


Figure A.6: Karyoplot for F2 samples, coloured by genotype. Samples are sorted in the order in which they were phenotyped. Blocks are filled with the colour of the paternal F0 line for the homozygous paternal haplotype block, black for heterozygous, and the colour of the maternal F0 line for the homozygous maternal haplotype block.

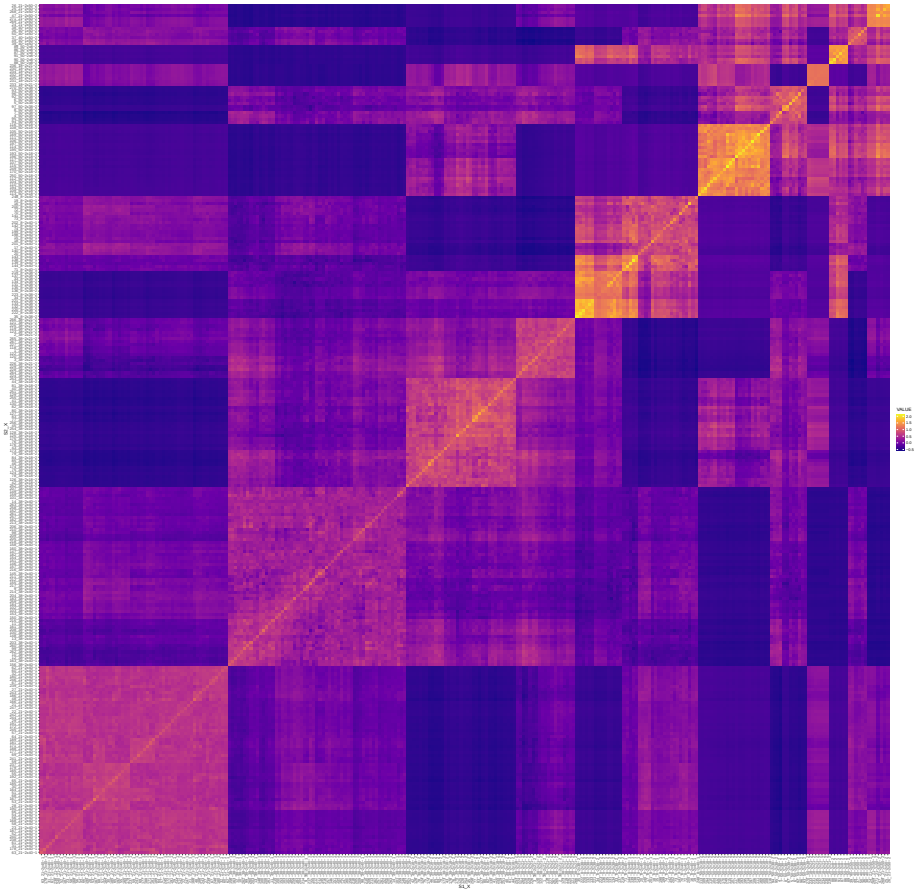


Figure A.7: "Leave-one-chromosome-out" genetic relationship matrix for 271 F2 samples based on 44,284 non-missing SNPs, having excluded 76 SNPs on chromosome 1.

Appendix B

Supplmentary information for Chapter 6

B.1 eCDF of all polygenic traits in the GWAS Catalog ranked by D_t^S

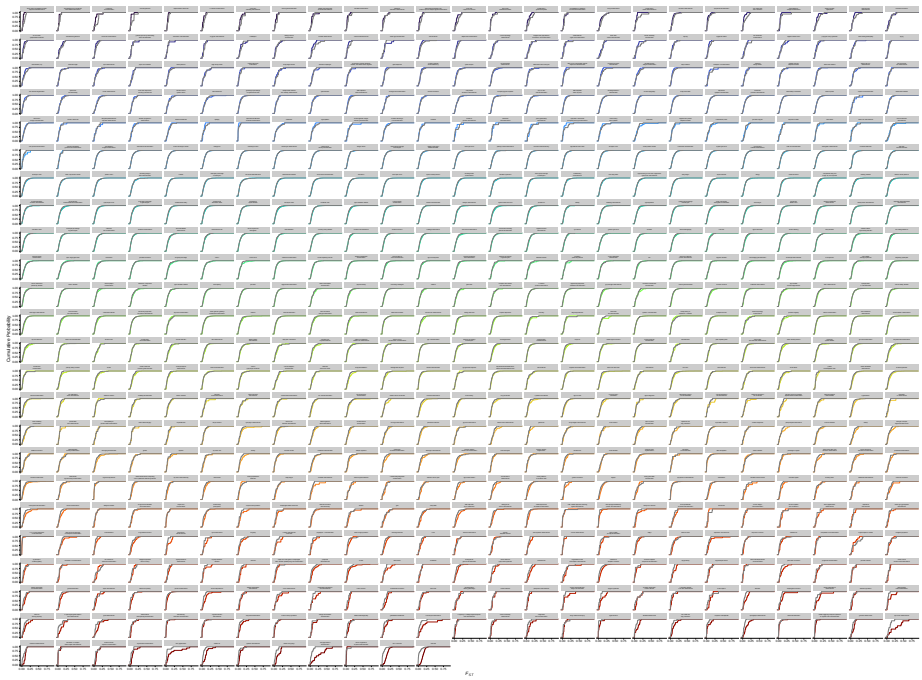


Figure B.1: 587 traits from the GWAS Catalog that passed our filters for polygenic traits, ranked by D_t^S.