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

Ian Brettell

2022-09-11

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

Genetic loci associated with somite development periodicity

1.1 Background[1]

During the development of an embryo, somites are the earliest primitive segmental structures that form from presomitic 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 ??** 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

[1]This Chapter describes a project carried out in collaboration with Ali Seleit and Alexander Aulehla from the Aulehla Group at EMBL Heidelberg. Drs Seleit and Aulehla performed the experiments and gathered the data; my role was to carry out the bioinformatics involved in mapping the genetic variants associated with the phenotypes of interest.

2014; Matsuda et al. 2020). **Figure ??** shows the a series of time-stamped images of somite segmentation in medaka fish, generated by Ali Seleit.

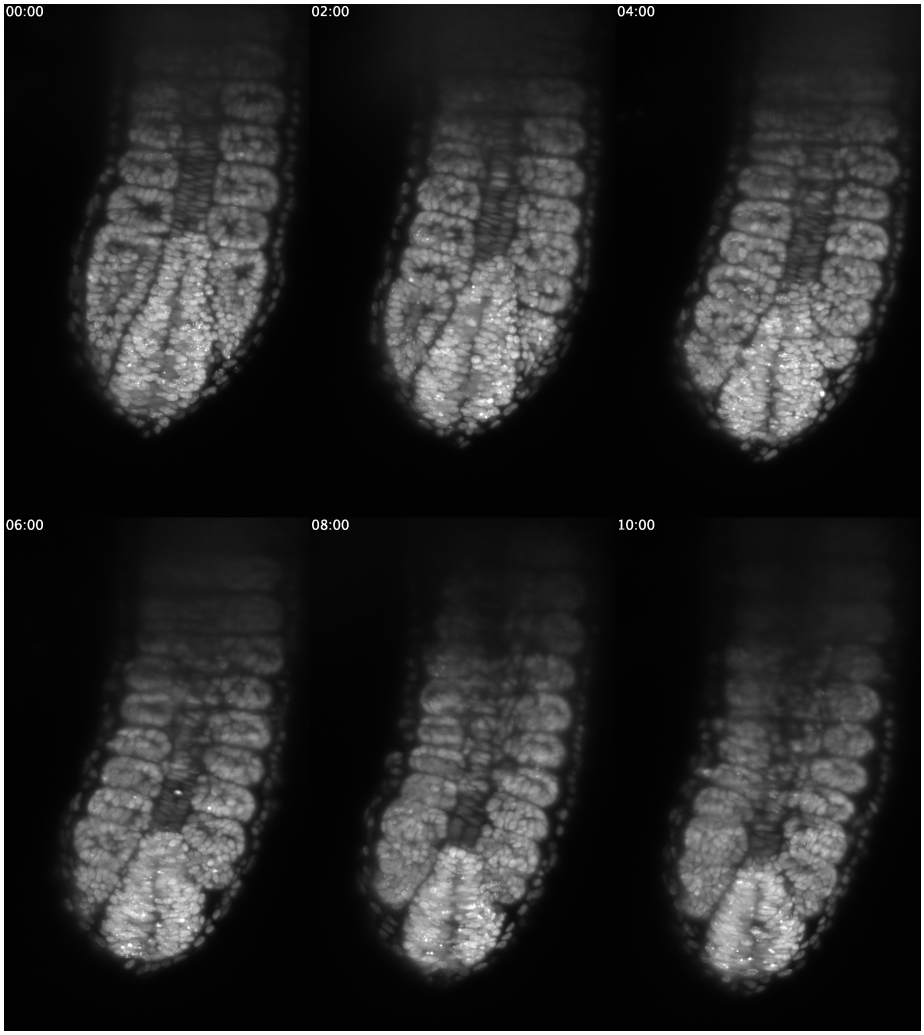

Figure 1.2: Time-stamped images of somite segmentation in medaka, generated by Ali Seleit.

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 at the tip of the tail 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).

Figure 1.1: Image of a mouse embryo at day 9.5 from Gridley (2006), showing somites in darker colours.

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 could be driven by biochemical reaction speeds – specifically, mRNA and protein degradation rates, transcription and translation delays, and intron and splicing delays. However, the authors did not provide a genetic explanation for why these biochemical reaction rates are different. Expanding on this work, our collaborators Ali Seleit and Alexander Aulehla at EMBL-Heidelberg are exploring the genetic basis of differences in segmentation clock periods. Carina Vibe of the Aulehla group used a CRISPR-Cas9 knock-in approach to establish a medaka *Cab* strain with an endogenous, fluorescing reporter gene (*Her7*-Venus, ~1.5 kb in length at the locus 16:28,706,898-28,708,417) for the oscillation signalling pathway.[2] 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 ??**). Our collaborators accordingly set up a one-way F2 cross experiment as

[2]This work is yet to be published, but the approach is similar to that described in Seleit, Aulehla, and Paix (2021).

described in Chapter ??, 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.

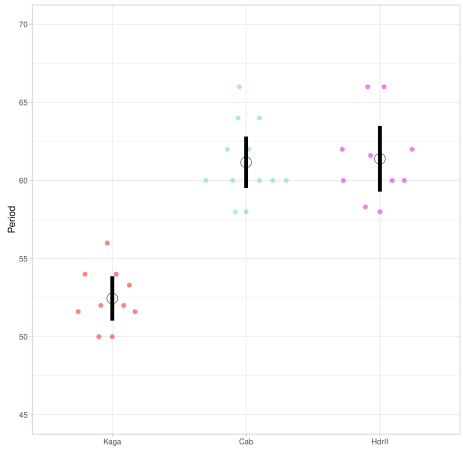

Figure 1.3: Comparison of period for three inbred medaka strains (*Cab*, *Kaga* and *HdrR*). Kaga's period is lower, and therefore it takes less time to form each somite than *Cab*. Figure generated by Ali Seleit.

They inter-crossed the hybrid F1 generation to create a sample of 622 F2 individuals (having selected for the F2 individuals carrying either one or two copies of the *Her7*-Venus reporter gene), 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 ??** 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.

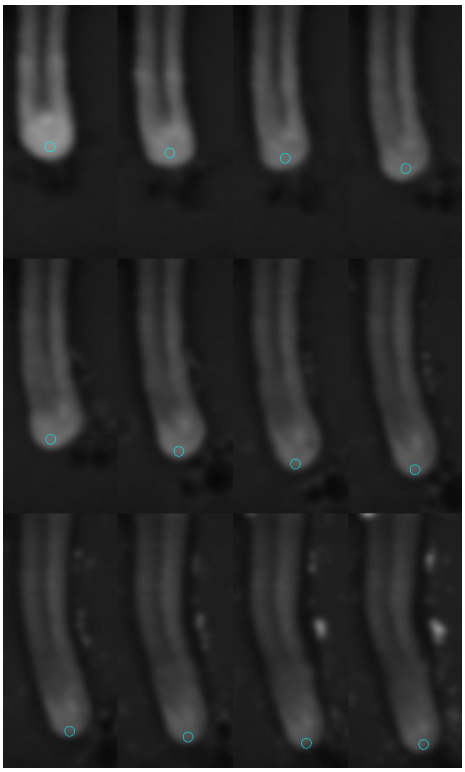

Figure 1.4: Screenshots of vertebral elongation in an F2 individual captured by Ali Seleit during imaging. The blue circle represents the point tracked by pyBOAT over time, generating the quantitative phenotype data on period development used in this study.

1.2 Phenotypes of interest

1.2.1 Somite development period

Figure ?? 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 ??**. I experimented with using the F2 individuals' mean period and period intercept as the phenotype of interest. The two measures are highly correlated ($Pearson's\ r = 0.84$, $p < 2.2 \text{ x } 10^{-16}$), so after displaying the distributions for both measures in Figure **??**, 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 **??**.

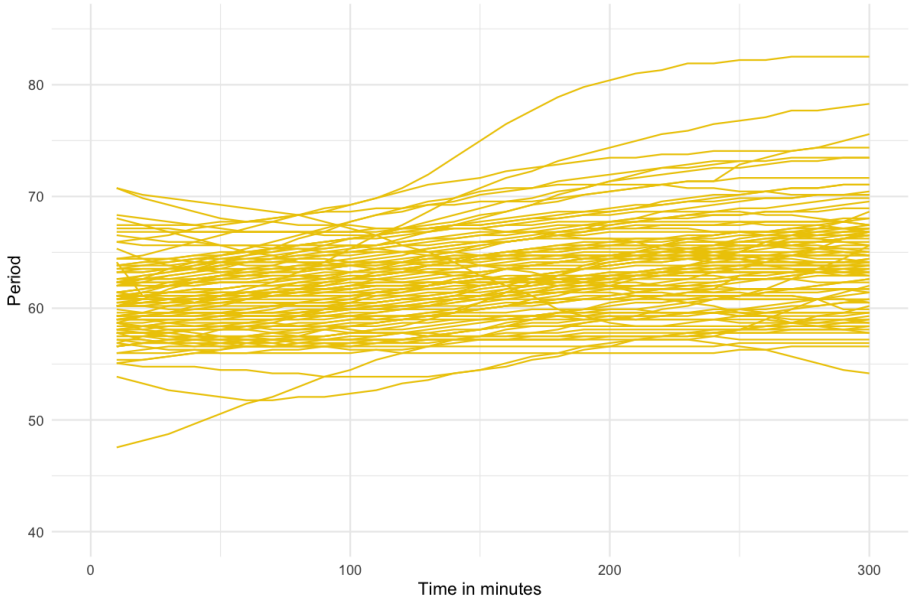

Figure 1.5: PyBOAT results for 100 illustrative F2 samples, showing the period length in minutes over the course of 300 minutes. Period tends to increase over time, meaning that as the embryo develops, each successive somite takes longer to form. Figure generated by Ali Seleit.

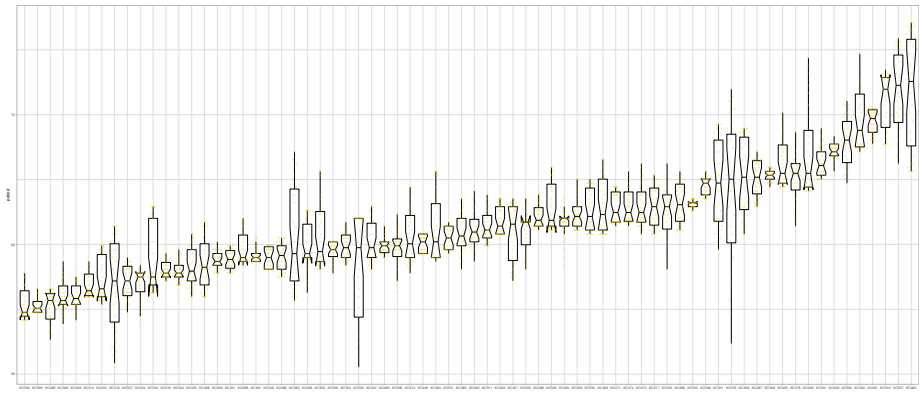

Figure 1.6: Period measurements for 70 F2 individuals displayed as boxplots with each individual's median and interquartile range. Figure generated by Ali Seleit.

1.2.2 Unsegmented presomitic mesoderm area (PSM)

In the proceeding analyses, I also included a second phenotype of interest: the total area of the presomitic mesoderm prior to the formation of any somite segments (**PSM area**). 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 the period phenotype. The measurements for PSM area comparing F0 *Cab* and *Kaga* strains are set out in **Figure ??**.

1.2.3 Comparisons between F0, F1 and F2 generations

The distributions across the F0, F1 and F2 generations are unexpected (**Figure ??**). I rather 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 ??), 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 = 1.34 x 10^{-40}), and a 3.5-minute difference in the F2 means for the period mean measure (Kruskal-Wallis = 141.79, p = 1.08 x 10^{-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 difference was found for the PSM area.

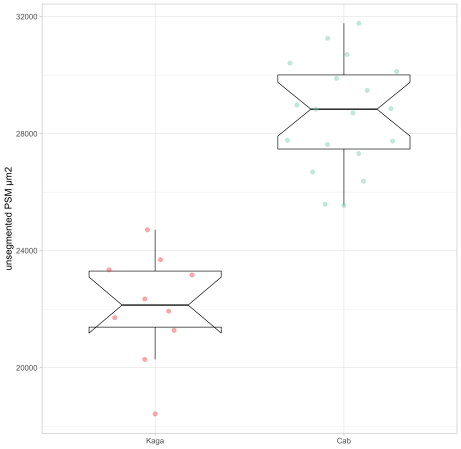

Figure 1.7: Measurements of unsegmented PSM area in pixels for the F0 individuals from the *Cab* strain ($N = 19$) and *Kaga* strain ($N = 10$). *Kaga* tends to have a smaller PSM than *Cab*. Figure generated by Ali Seleit.

1.2.3.1 Inverse-normalisation

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**

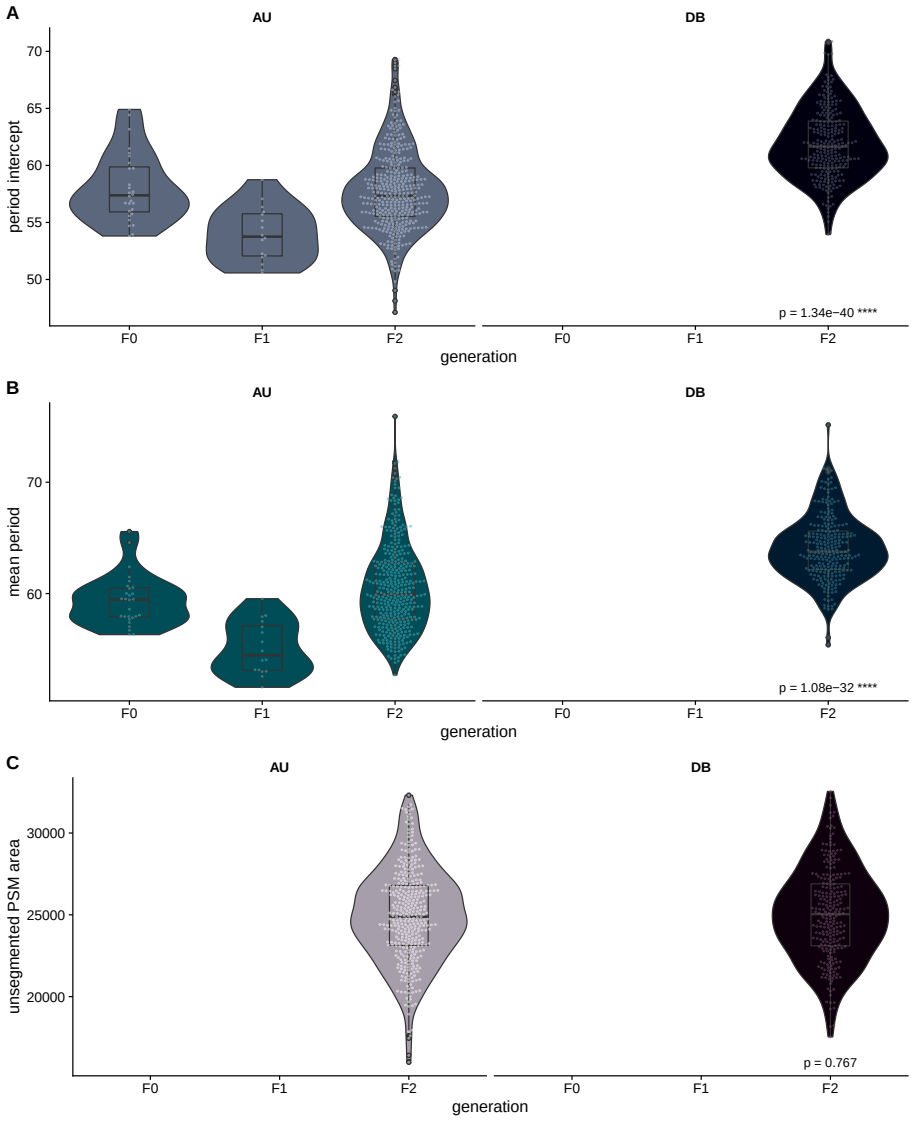

Figure 1.8: Comparisons between the F0, F1 and F2 generations for the three phenotypes of interest. Here, the F0 only includes *Cab* individuals. **A**: period intercept. **B**: period mean. **C**: unsegmented PSM area. P-values are derived from Kruskal-Wallis tests comparing the F2 individuals across microscopes.

??), and used this transformed phenotype for the downstream analysis. Inverse-normalisation is a rank-based normalisation 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 'qnorm()' returning 'Inf' for the last-
      ranked value
  res = qnorm(res/(length(res)+0.5))
  return(res)
}
```

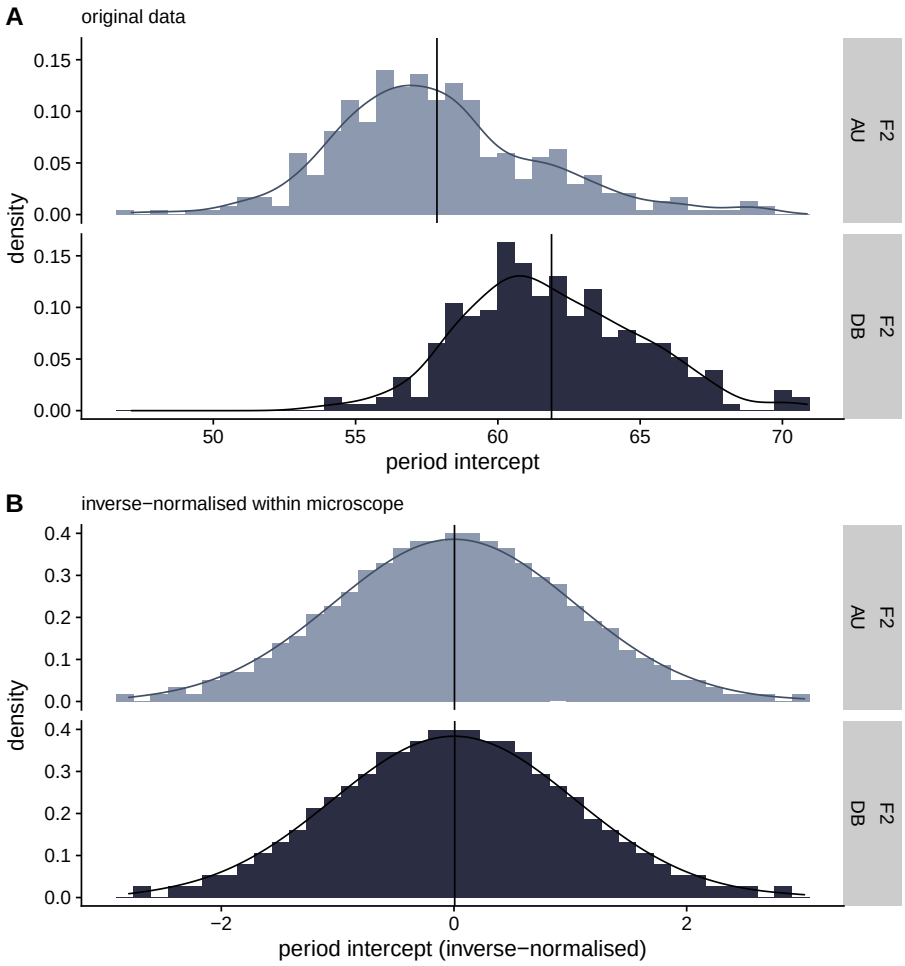

Figure 1.9: Comparison of the period intercept phenotype data for the F2 generation before (**A**) and after (**B**) inverse-normalisation, with vertical lines marking the mean of each group.

1.3 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 (Danecek et al. 2021). **Figure ??** 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 section ?? below for further details).

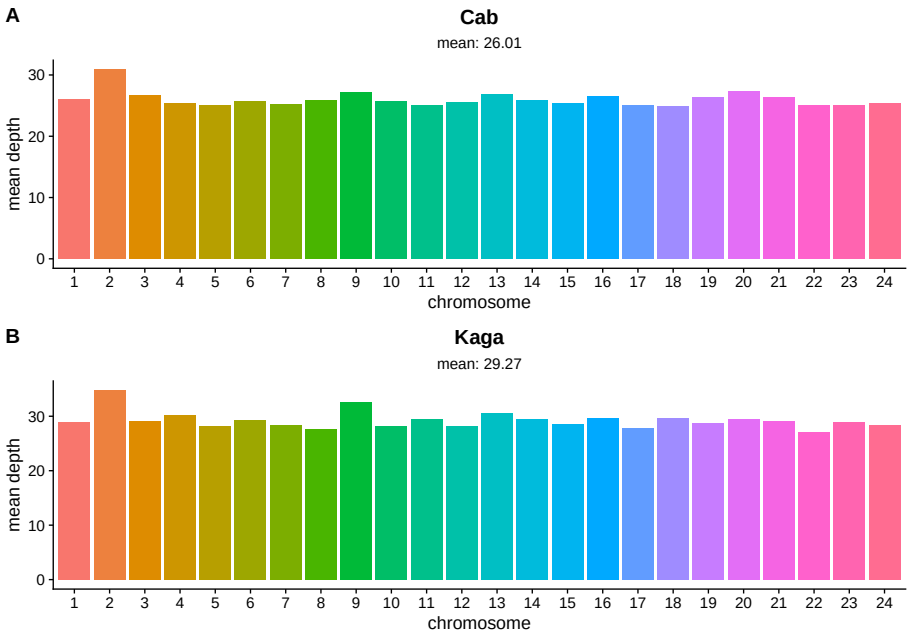

Figure 1.10: Mean sequencing depth per chromosome for *Cab* and *Kaga* F0 strains, with genome-wide mean depth across all chromosomes shown under the subtitles.

1.4 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 affect our ability to accurately call the F2 generation. That is to say, for regions where either F0 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* reference (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 ?? 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%.

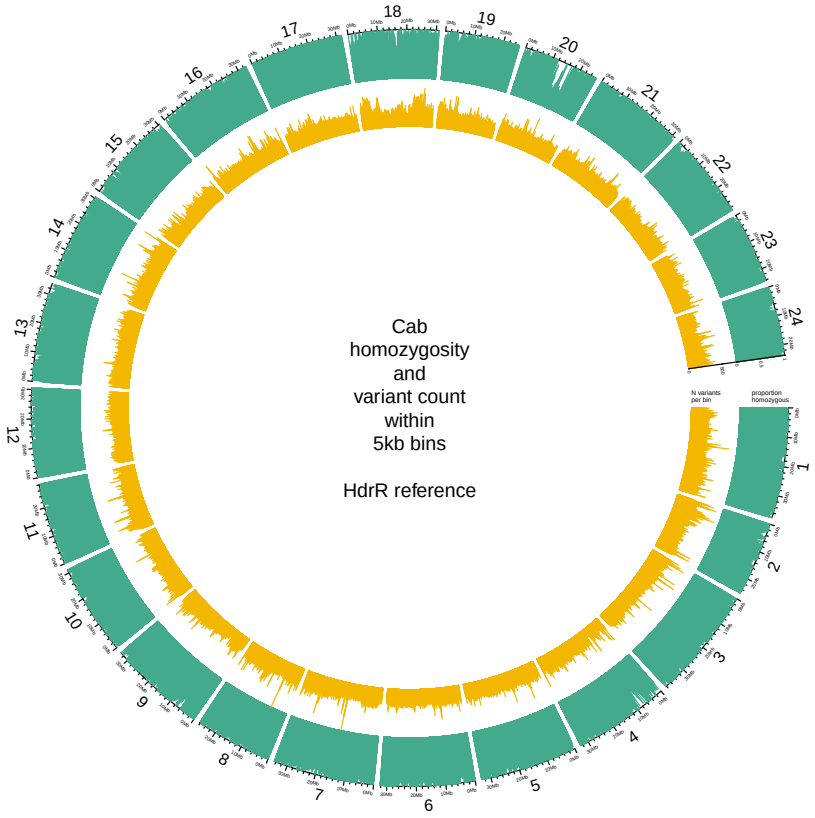

Figure 1.11: Proportion of homozygous SNPs within *5* kb bins in the *Cab* F0 generation genome (green), and number of SNPs in each bin (yellow).

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% (**Figure ??**). This was a surprise, as it is an established strain of [XXXX] generations, and we therefore expected the level of homozygosity to be commensurate with that observed in the *Cab* strain. An obvious exception is chr22, for which *Kaga* appears to be homozygous across its entire length.

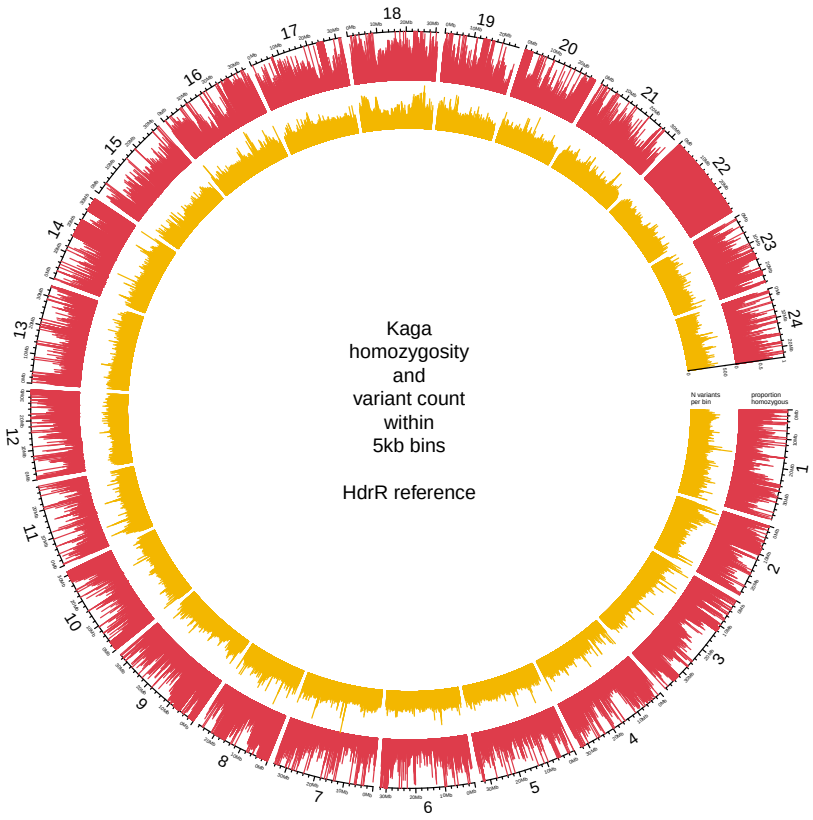

Figure 1.12: Proportion of homozygous SNPs within 5 kb bins in the *Kaga* F0 generation genome (red), and number of SNPs in each bin (yellow).

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 ??**. 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. I understand from our collaborators that the low homozygosity of this *Kaga* individual must have resulted from the strain having been contaminated at some stage by breeding with a different inbred strain.

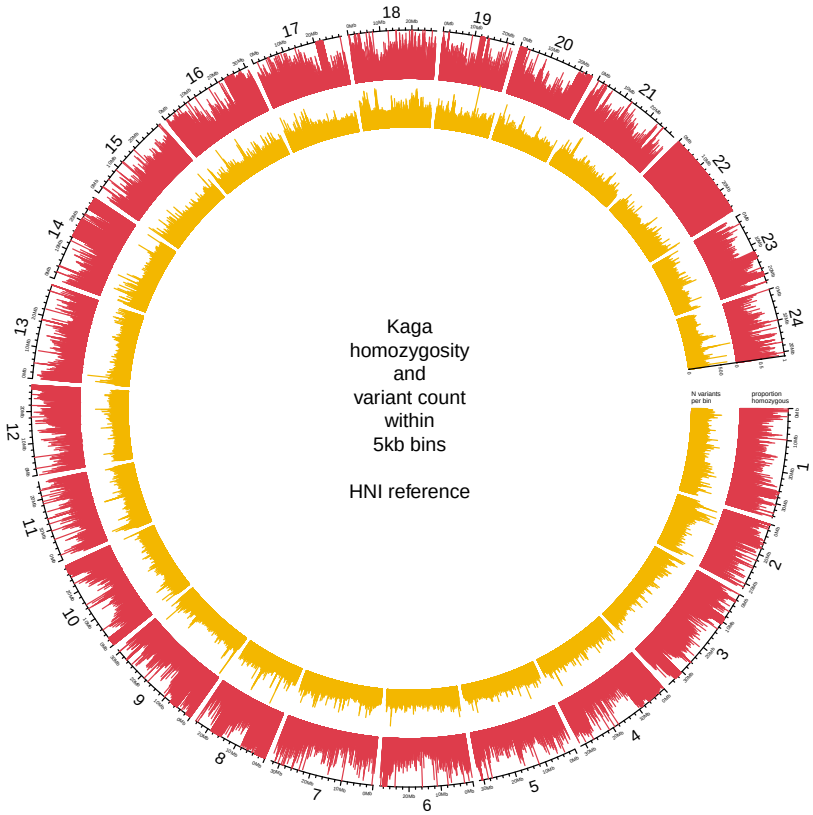

Figure 1.13: Proportion of homozygous SNPs within 5 kb bins in the *Kaga* F0 generation genome when aligned to the *HNI* reference (red), and number of SNPs in each bin (yellow).