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

Ian Brettell

2022-09-09

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 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 1.1** depicts a number of formed somites in a 9.5-day-old mouse embryo.

presomitic

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 1.2** shows the a series of time-stamped images of somite segmentation in medaka fish, generated by Ali Seleit.

Figure 1.1: 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).

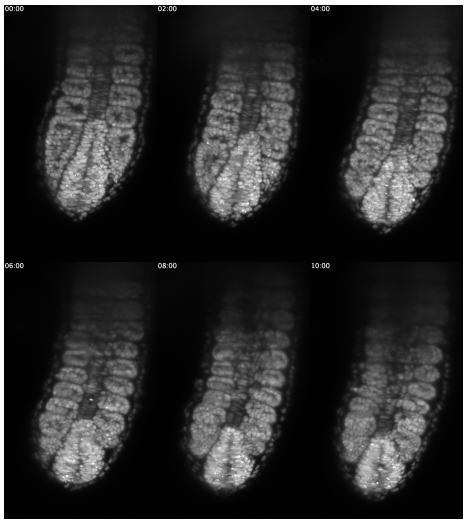

Figure 1.2: Time-stamped images of somite segmentation in medaka, generated by Ali Seleit.

It is not fully understood how the phase waves of the segmentation clock are intially established (Falk et al. 2022). Matsuda et al.

Here it is important to distinguish between 2 things A) there is an inherent period gradient in every tail where the posterior most period (at tip of tail) is the fastest and more anterior cells slow down (slowest where next somites will form), this actually creates the impression of the traveling wave of gene expression along the A-P axis and B) the fact that even posterior most period (tip of tail) slows down over time: this phenomenon B is true in medaka but i do not know if it is also true in mouse and zebrafish

Matsuda provides some evidence to back the claim that differences in biochemical reaction speeds COULD explain the differences in segmengation rate between mouse and men, but they do not provide a genetic explanation….or in other words how and why are the biochemical reaction rates different? they just say we measured them and it looks like they are different for some genes with the same difference in period we observe

I wouldnt say we neccesarily are proving or disproving the hypothesis that biochemical reaction speeds play a role in setting tempo of segmentation clock….we are just looking for genetic variants, whatever they may be, that modulate clock tempo

(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 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. This method allows them to image somite formation and extract quantitative measures for segmentation clock dynamics.

This her7venus line was generated by Carina Vibe a PhD student in ALexander's lab before i joined (no preprint yet of that work)

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 1.3**).

this was not known before, we found that out

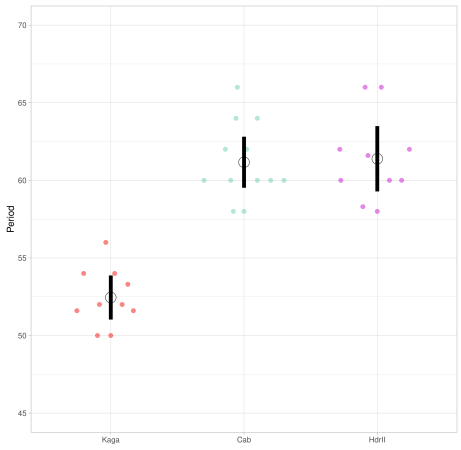

Figure 1.3: Comparison of period for three inbred medaka strains (*Cab*, *Kaga* and *HdrR*). Kaga's period is lower, and therefore it takes less time to form each somite than *Cab*. Figure generated by Ali Seleit.

Our collaborators accordingly set up a one-way F2 cross experiment as 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. They intercrossed 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 1.4** 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.

1.2 Phenotypes of interest

1.2.1 Somite development period

Figure 1.5 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 1.6**. I experimented with using the F2 individuals' mean period and period intercept as the phenotype of interest. The two measures are highly correlated ($Pearson's\ r = 0.84$, $p < 2.2 \times 10^{-16}$), so after displaying the distributions for both measures in Figure 1.8, 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 1.5.

presomitic

1.2.2 Unsegmented presomatic mesoderm area (PSM)

actually it is the area of the unsegmented tissue (PSM) at the 10-11 somite stage, so 10 somites have already formed by that stage

In the proceeding analyses, I also included a second phenotype of interest: the total area of the presomatic 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 1.7**.

presomitic

1.2.3 Comparisons between F0, F1 and F2 generations

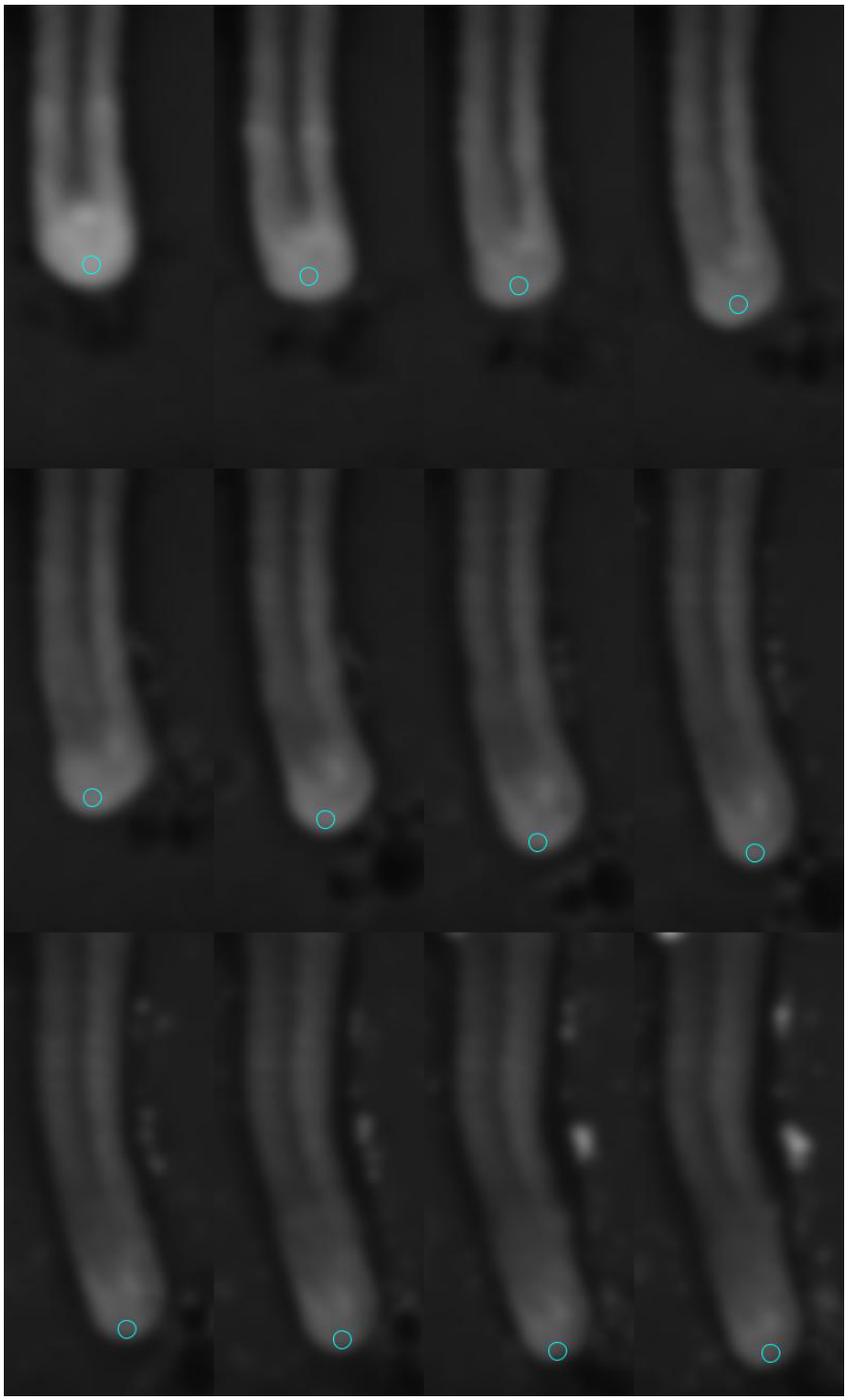

Figure 1.4: Screenshots of vertebral elongation in an F2 individual captured by Ali Seleit during imaging. The blue circle represents the point tracked by pyBOAT over time, generating the quantitative phenotype data on period development used in this study.

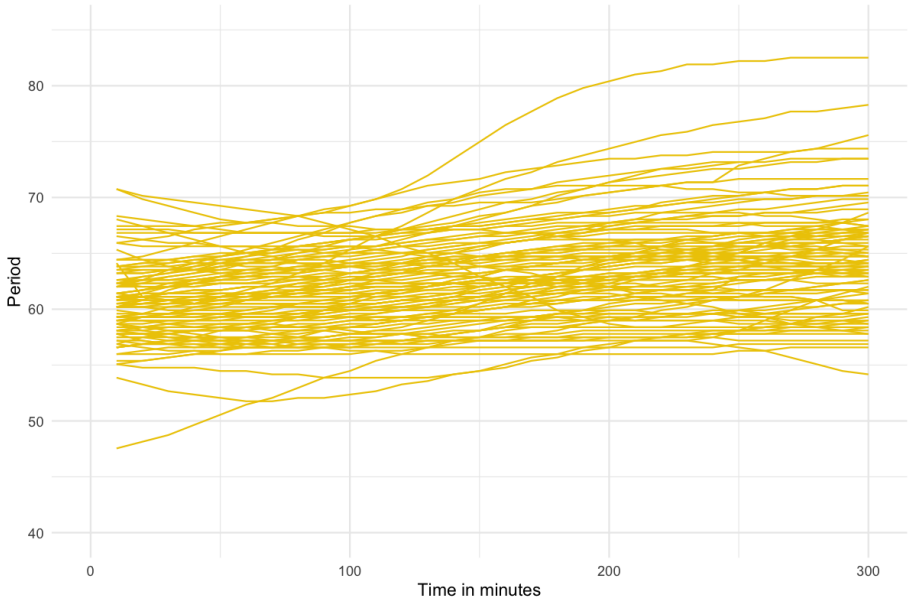

Figure 1.5: PyBOAT results for 100 illustrative F2 samples, showing the period length in minutes over the course of 300 minutes. Period tends to increase over time, meaning that as the embryo develops, each successive somite takes longer to form. Figure generated by Ali Seleit.

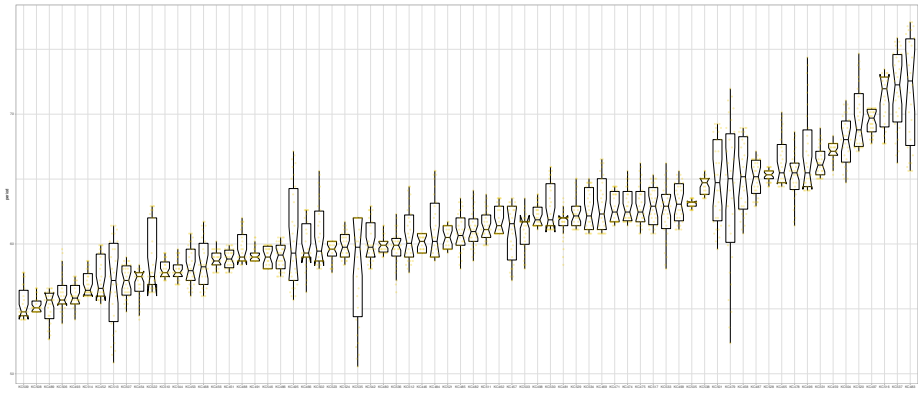

Figure 1.6: Period measurements for 70 F2 individuals displayed as boxplots with each individual's median and interquartile range. Figure generated by Ali Seleit.

The distributions across the F0, F1 and F2 generations are unexpected (**Figure 1.8**). 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.

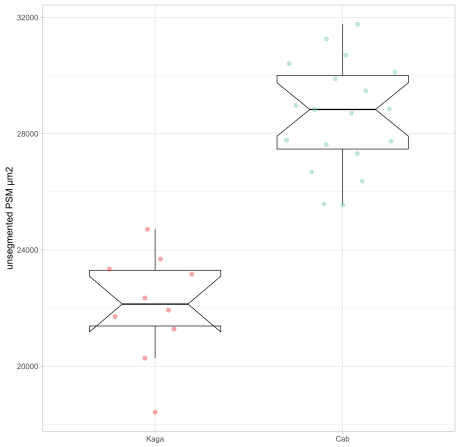

Figure 1.7: Measurements of unsegmented PSM area in pixels for the F0 individuals from the *Cab* strain ($N = 19$) and *Kaga* strain ($N = 10$). *Kaga* tends to have a smaller PSM than *Cab*. Figure generated by Ali Seleit.

I think a possible biological explanation of this is that there are way more genetic combinations that slow down the clock than there are to speed it up….this observation is backed up with data from medaka mouse and zebrafish where there are many ways to slow down the clock but it is considerably more difficult to speed it up….so distribution of F2 could reflect that fact

possible, but a simpler explanation would be that some novel genetic combinations lead to novel phenotypes that are not observed in F0s or F1s

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

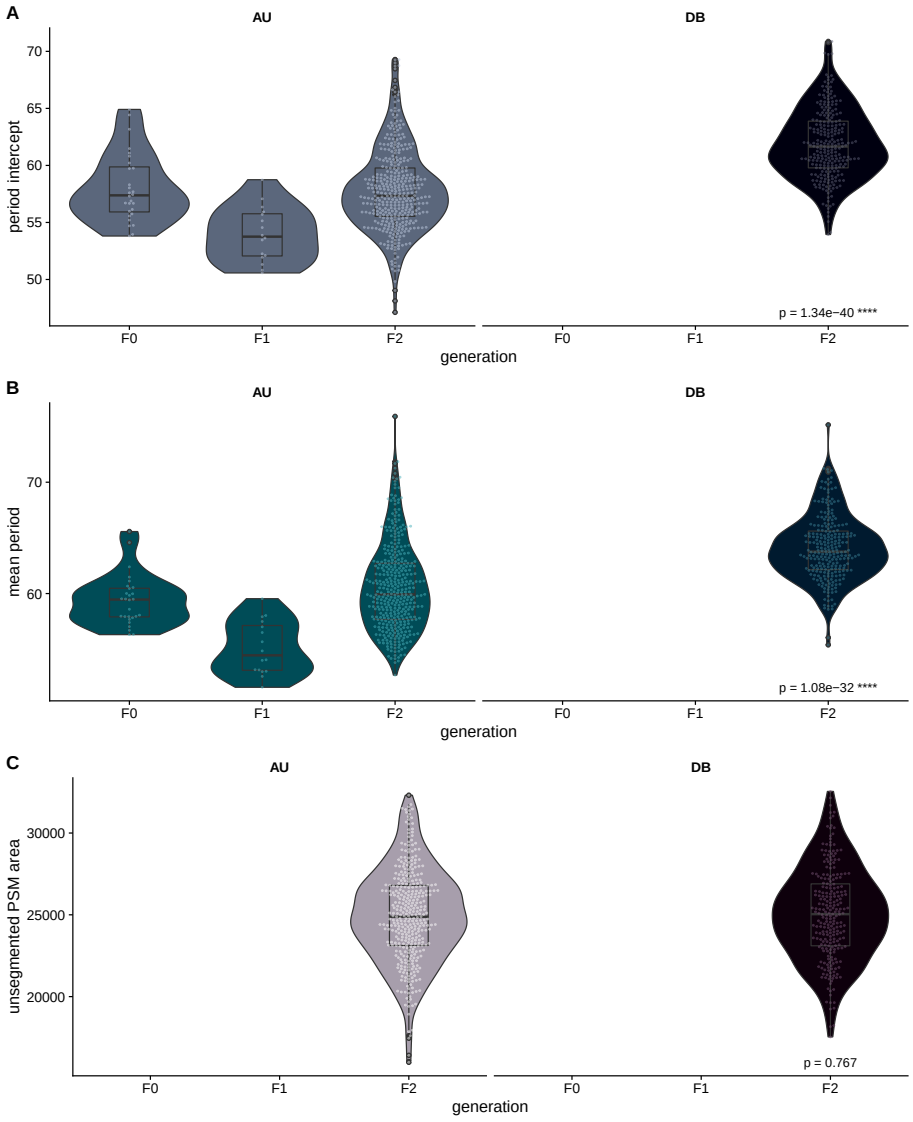

Figure 1.8: Comparisons between the F0, F1 and F2 generations for the three phenotypes of interest. Here, the F0 only includes *Cab* individuals. **A**: period intercept. **B**: period mean. **C**: unsegmented PSM area. P-values are derived from Kruskal-Wallis tests comparing the F2 individuals across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

by including microscope as a covariate in the model. No significant difference was found for the PSM area.

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 1.9**), 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)
}
```

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 1.10** 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

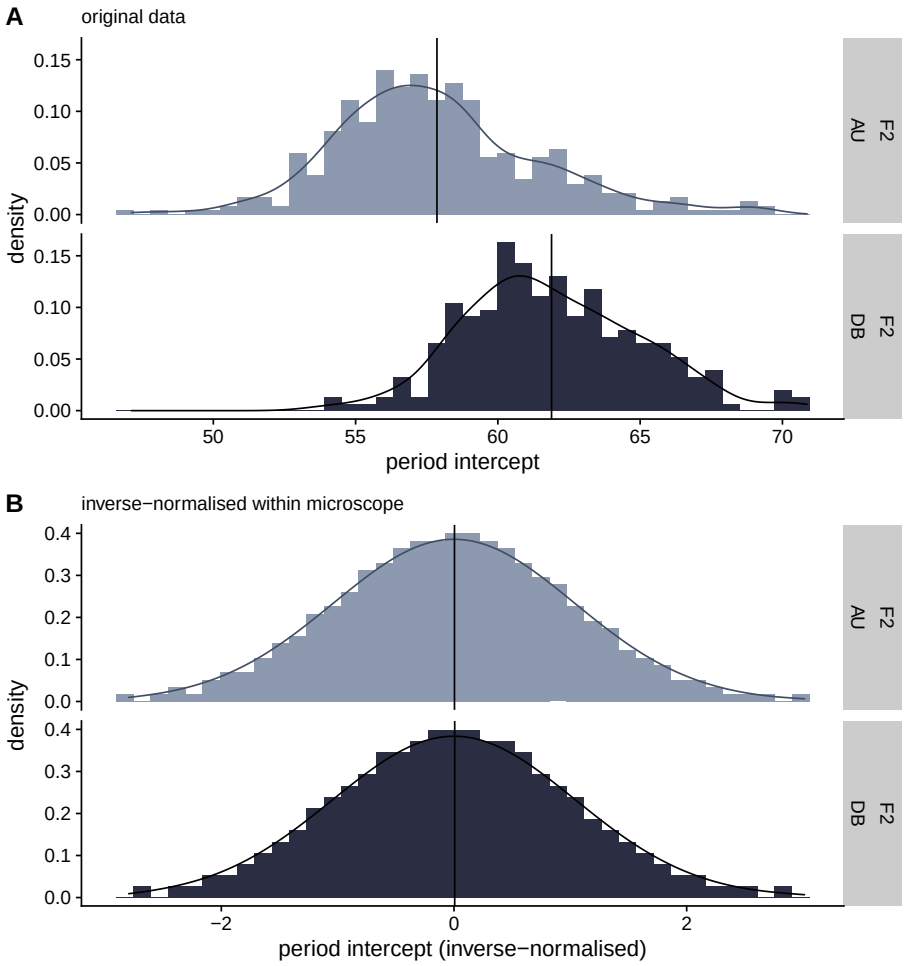

Figure 1.9: Comparison of the period intercept phenotype data for the F2 generation before (**A**) and after (**B**) inverse-normalisation, with vertical lines marking the mean of each group.

genotypes back to the genotypes of their parental strains (see section 1.6 below for further details).

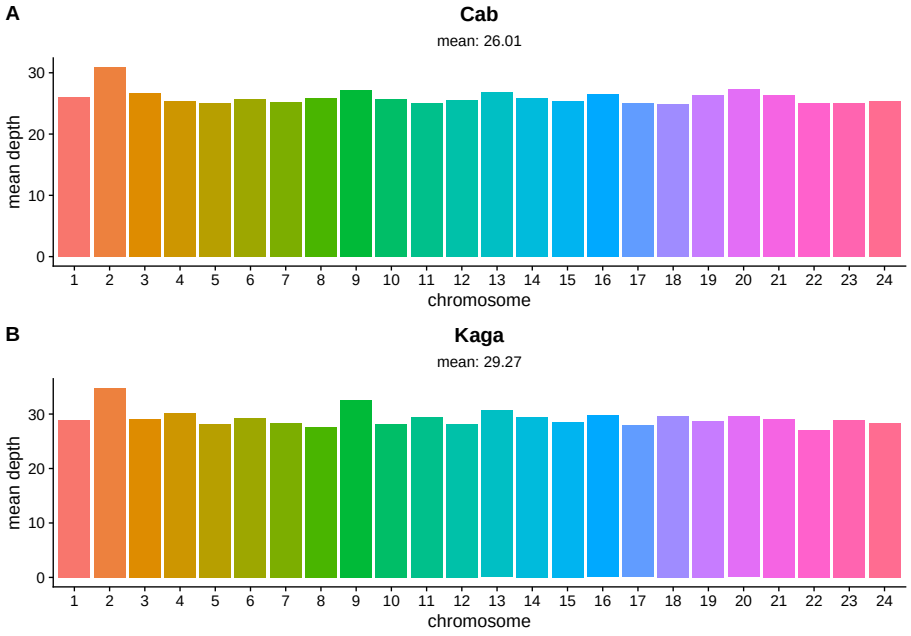

Figure 1.10: Mean sequencing depth per chromosome for *Cab* and *Kaga* F0 strains, with genome-wide mean depth across all chromosomes shown under the subtitles.

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 1.11 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%.

Better ask Felix how long they inbred Kaga… we got those fish from him at KIT

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 1.12**). 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.

I am not sure this is the case Ian, again those fish we got from Felix Loosli and there was no contamination from our side…

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 1.13**. 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. 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.

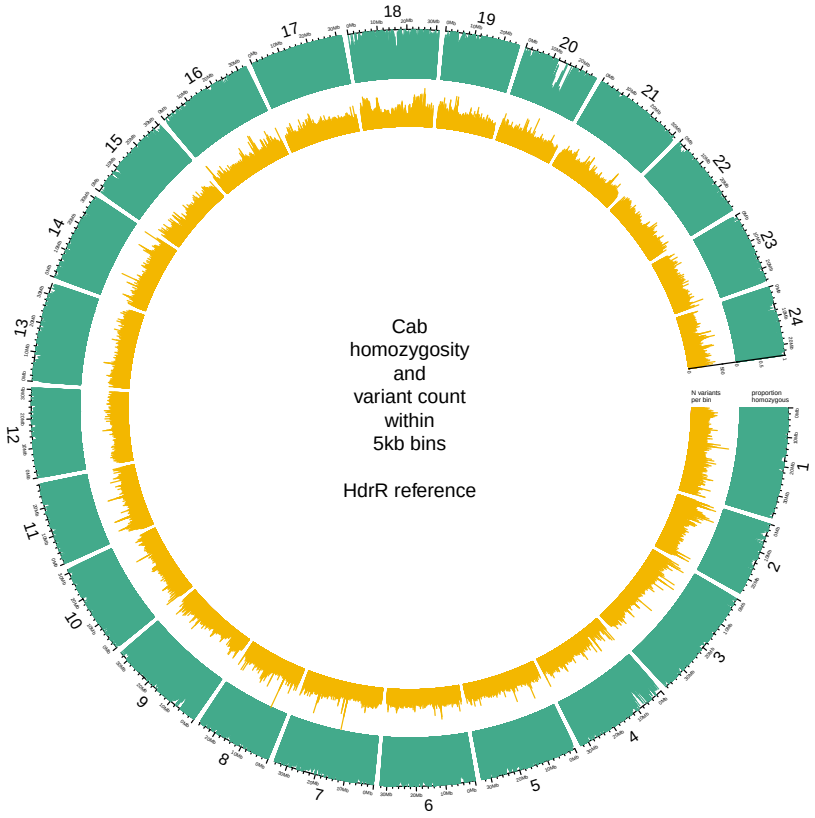

Figure 1.11: Proportion of homozygous SNPs within *5* kb bins in the *Cab* F0 generation genome (green), and number of SNPs in each bin (yellow).

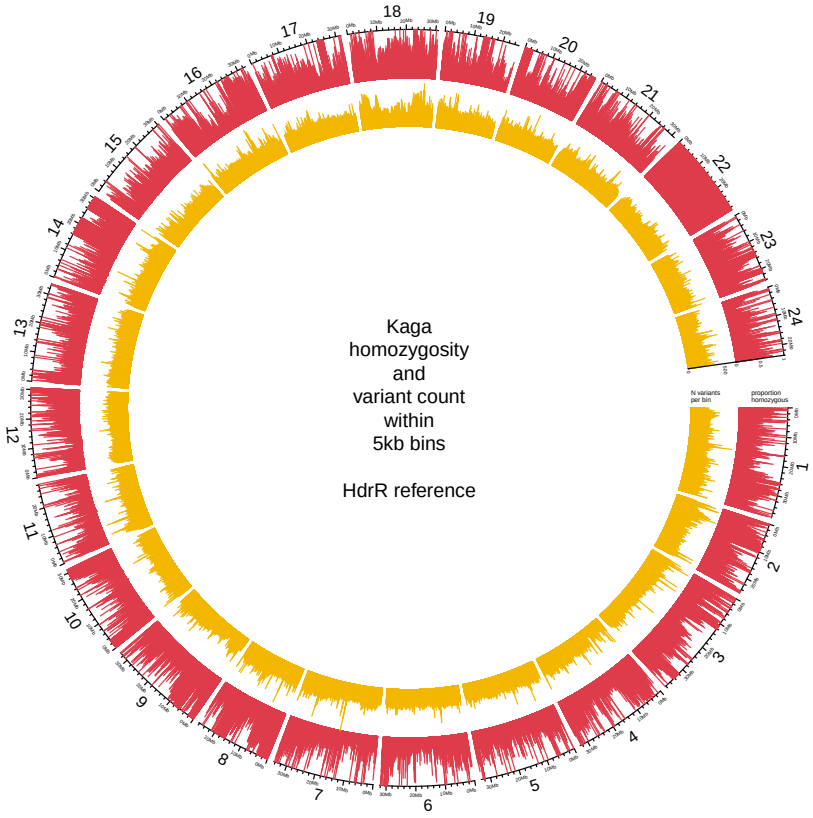

Figure 1.12: Proportion of homozygous SNPs within 5 kb bins in the *Kaga* F0 generation genome (red), and number of SNPs in each bin (yellow).

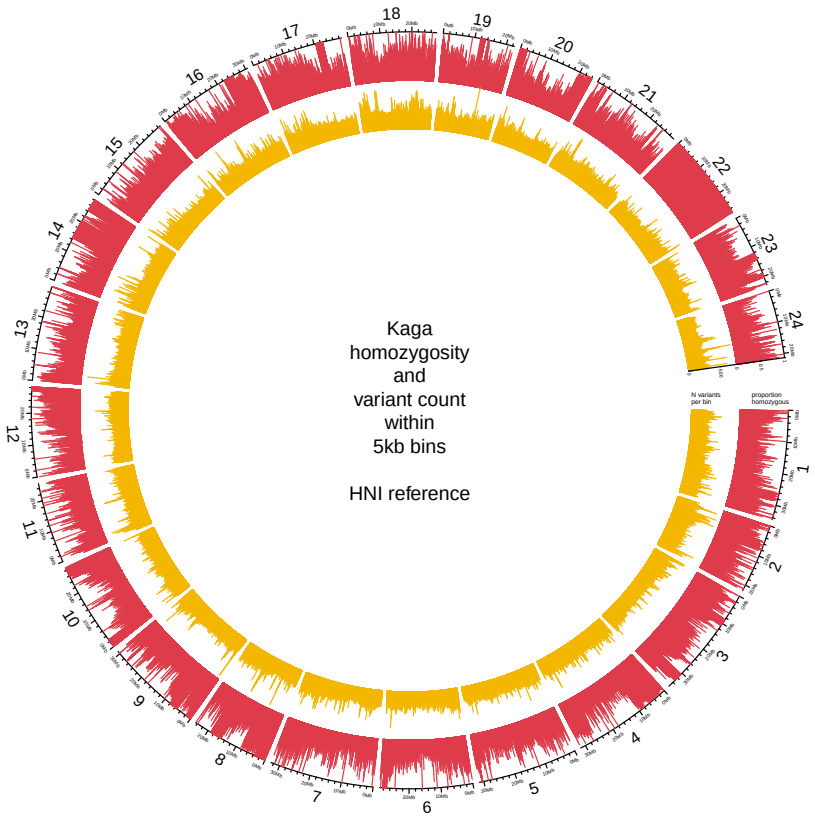

Figure 1.13: Proportion of homozygous SNPs within 5 kb bins in the *Kaga* F0 generation genome when aligned to the *HNI* reference (red), and number of SNPs in each bin (yellow).

1.5 F1 homozygosity

I next examined the level of heterozygosity in the F1 generation from the *Cab-Kaga* cross. **Figure 1.14** 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 apparent heterozygosity than expected was likely caused by the low levels of homozygosity 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 1.15**. The strong homozygosity of *Kaga* on chr22 is likely responsible for the much greater number of loci on that chromosome that can be used for calling genoytpes in the F2 generation, and highlights the importance of the parental strains being highly homozygous when used in experimental crosses such as this.

1.6 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

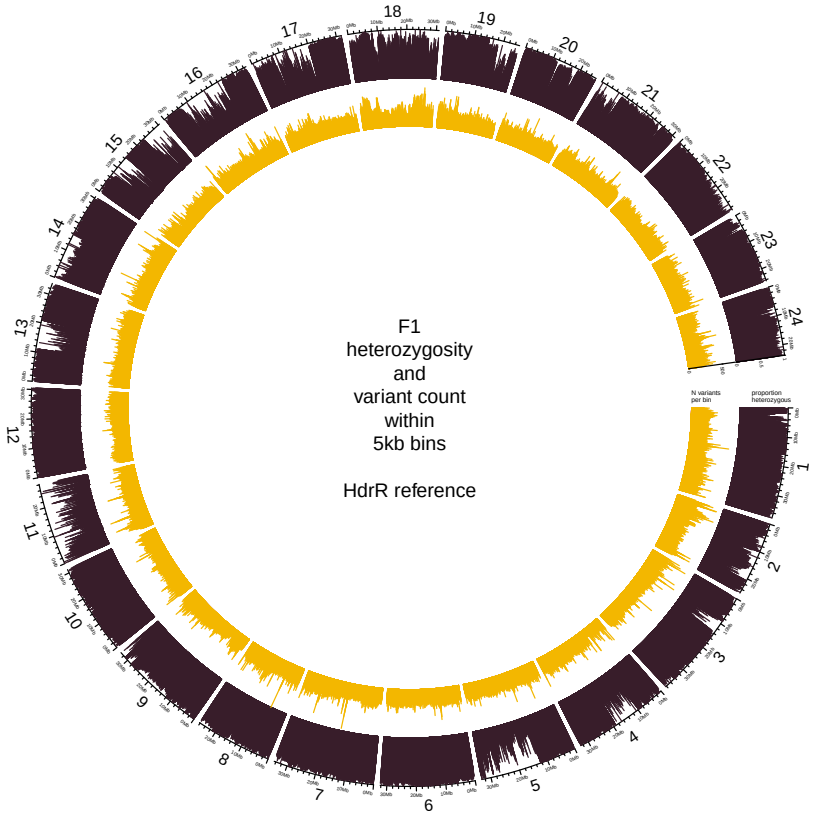

Figure 1.14: Proportion of heterozygous SNPs within 5 kb bins in the *Cab-Kaga* F1 cross (brown), and number of SNPs in each bin (yellow).

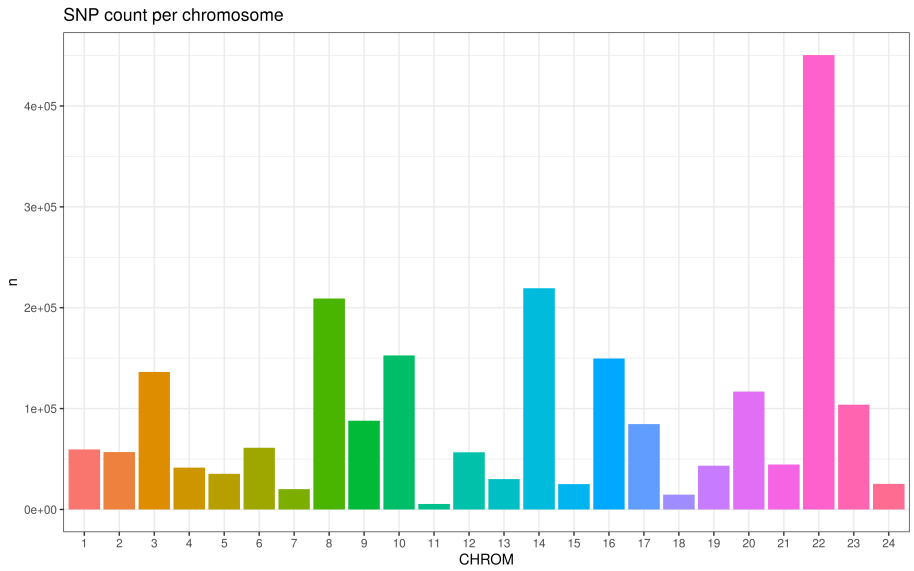

Figure 1.15: Number of SNPs per chromosome that were homozygous-divergent in the F0 *Cab* and *Kaga* generations, and heterozygous in the F1 generation.

count the reads that supported either the *Cab* or the *Kaga* allele for all SNPs that met the criteria described above in section 1.5, 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 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 1.16** 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.

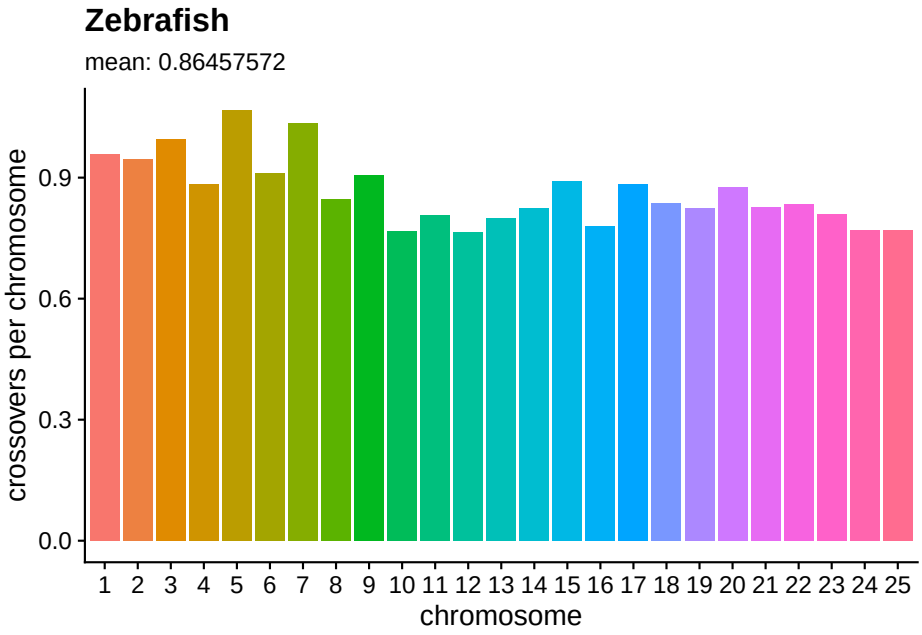

Figure 1.16: Crossovers per chromosome based on data provided in Haenel et al. (2018), where "crossovers per chromosome" for each chromosome c was calculated by $\frac{crossover\ rate_c(cM/Mb) \times length_c(Mb)}{100}$. The medaka genome is shorter in length than zebrafish genome (~800 Mb compared to ~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 1.17 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-scat)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 karyoplot shown in **Figure 1.18**, with homozygous-*Cab* blocks in green, heterozoygous blocks in navy blue, and homozygous *Kaga* blocks in red. Missing calls are blank, where the vertical blank lines

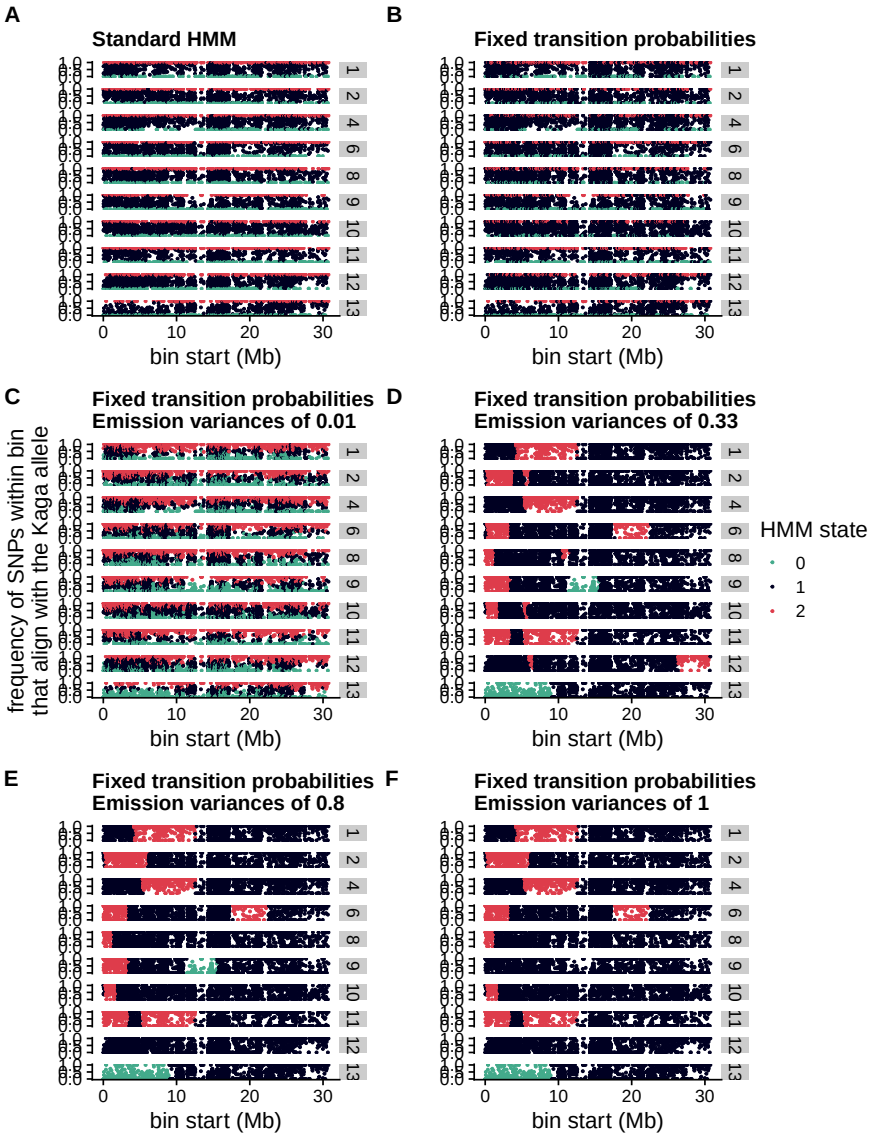

Figure 1.17: HMM states called for each bin across chr18 for 10 F2 samples. States 0, 1, and 2 correspond to homozygous *Cab*, heterozygous, and homozygous *Kaga*. Each point represents a 5-kb bin. Y-axis is the proportion of reads within each bin that align to the *Kaga* allele. X-axis is the bp location 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**).

indicate that the region could not be called for any F2 individuals, likely due to an insufficient number of informative SNPs residing in those 5-kb blocks; and horizontal blank lines indicate that the sample could not be called, likely due to low sequencing coverage for that sample.

In the downstream analysis, I excluded the 22 samples that showed poor coverage across the genome. For the remaining samples, I "filled" the bins with missing genotypes based on the call of the previous called bin, or if unavailable (e.g. the missing bin was at the start of the chromosome), then the next called bin (**Figure 1.19**; note that this figure retains the low-coverage samples that were excluded from further analysis to allow for a direct comparison with **Figure 1.18**). I used these filled genotype calls for the association tests described below in section **1.7**. These karyoplots show interesting recombination patterns for several chromosomes. The reporter gene resides on chr16 ~28.7Mb, so given the F2 individuals were selected for the reporter gene, as expected, all but one F2 individual are called as either homozygous-*Cab* or heterozygous at that locus, with that single exception being a genotyping error. Moreover, the selection for homozygous-*Cab* or heterozygous genotypes at that locus has caused a strong skew towards those genotypes across the whole chromosome.

I think another important validation point here is that when we compared the genotype calling from the HMM with the phenotype observed from the imaging (where i can distinguish between HET and HOM states based on intensit y of signal of her7venus) there was an 86% concurrence of the genotype called by HMM and phenotype called by me on each of the F2 samples….which means at least at that particular locus the HMM is quite accurate

On chr3, most samples are homozygous-*Cab* for the second half of the chromosome, with a consistent breakpoint around ~22 Mb. However, the final fifth of samples which show a different recombination pattern. The samples are sorted based on the order that they were phenotyped and sequenced, so this difference could have been caused by their being generated from different F1 individuals with distinct haplotypes [CHECK WITH ALI].

This is correct, they were from different F1 individuals

Figure 1.20 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 was likely caused by the hybrid incompatibility between *Cab* and *Kaga*, given the two strains were derived from populations that are

Figure 1.18: 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.

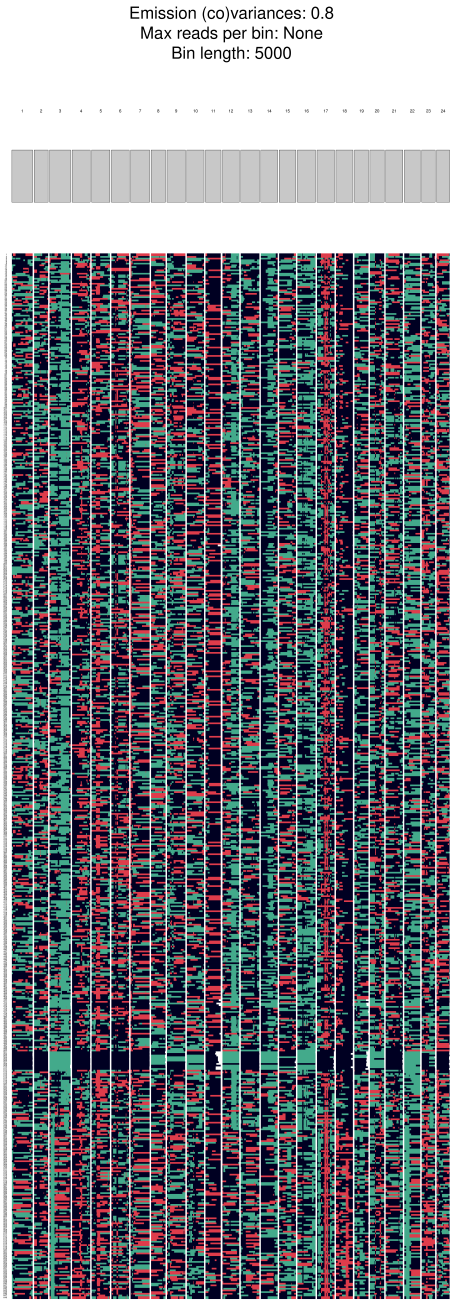

Figure 1.19: Recombination blocks in 622 F2 samples based on the ratio of reads mapping to either the *Cab* or *Kaga* allele within 5-kb bins, with homozygous-*Cab* blocks in green, heterozoygous blocks in navy blue, and homozygous *Kaga* blocks in red. Most blocks show 0-2 crossover events, as expected, with some regions showing higher numbers of crossovers interpreted as noise. Bins with missing genotypes were "filled" based on the call of the previous called bin, or if unavailable (e.g. the missing bin was at the start of the chromosome), then the next called bin.

thought to be at the point of speciation (see the previous analysis in **Chapter ??**, section **??**).

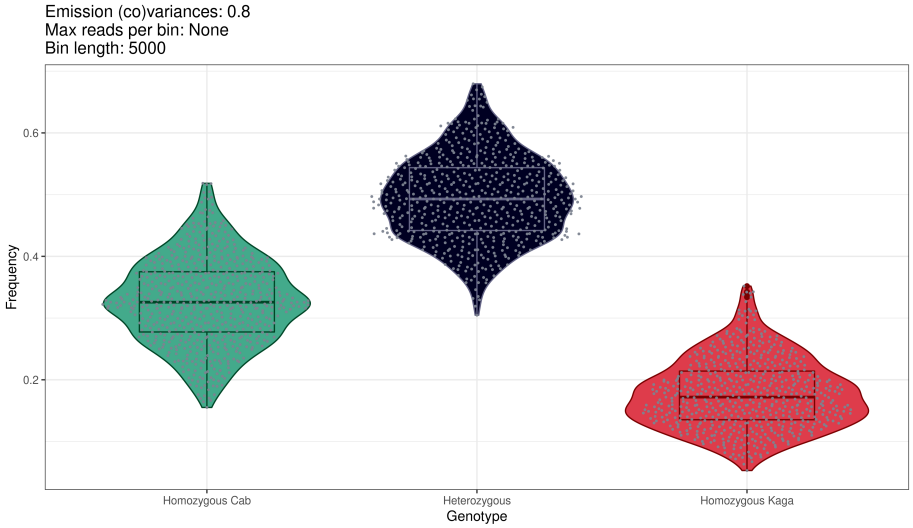

Figure 1.20: Proportions of 5-kb blocks called as either homozygous-*Cab*, heterozygous, or homozygous-*Kaga*.

1.7 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:

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

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.

	Chromosome	Bin start	Bin end	Length (kb)
1	3.00	31880001.00	35420000.00	3540.00
2	4.00	18090001.00	18095000.00	5.00
3	10.00	2995001.00	3690000.00	695.00

Table 1.1: Significant 5-kb bin ranges for period intercept below the minimum p-value from 10 permutations.

Both scopes are confocal Zeiss LSM 780s but they have different temp control units and incubator boxes AU scope denotes the Aulehla lab scope and DB scope denotes the Developmental Biology Unit scope…they are just a way for us to distinguish begween them during booking

difference is about 0.8 degrees celcius between both scopes

As described above in **section 1.2**, the microscope used to image the embryos (either AU or DB [WHAT DO THESE ACRONYMS STAND FOR?]) differed by several degrees in heat, which 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, Id 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.

1.7.1 Period intercept

Figure 1.21 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 ??**.

These regions contained a total of 46,872 SNPs imputed from the genotype of the F0 parental strains. I ran Ensembl's Variant Effect

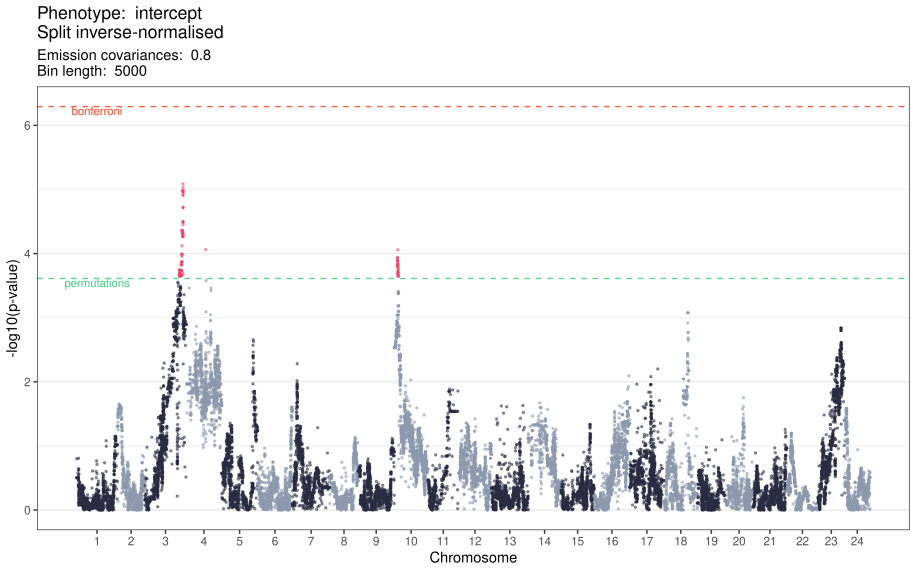

Figure 1.21: Manhattan plot of the genetic linkage results for the period intercept phenotype, inverse-normalised across microscopes. Pseudo-SNPs with p-values lower than the permutation significance threshold are highlighted in red.

Predictor (McLaren et al. 2016) over these SNPs to identify those 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 1.2**. 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 1.3**). 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.

Actually we are doing CRISPR knock outs not knock-ins to see if any of those genes plays a role in setting tempo

Knock ins are complicated by the fact that for many genes there are multiple SNPs, and for many we also identified multiple SNPS falling in non-coding regions containing ATAC seq peaks suggesting they are likely regulatory in nature….and thirdly KIs are limited by 1kb insertds and efficiency tends to go down significantly beyond 1 kb

Table 1.2: 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

Table 1.3: 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	Role
3	ENSORLG00000014656	mesoderm posterior protein 2-like	Somitogenesis
3	ENSORLG00000014659	mesp	Somitogenesis
10	ENSORLG00000020474	protocadherin 10b	Somitogenesis
3	ENSORLG00000014616	NA	Possible role in Somitogenesis (MAP-kinase)
10	ENSORLG00000020551	FAT atypical cadherin 4	Possible role in Somitogenesis (PCP, Yap1 regulator)
10	ENSORLG00000020531	neurogenin 1	Possible role in Somitogenesis (BHLH-TF regulation of Wnt)
3	ENSORLG00000015149	ADAM metallopeptidase with thrombospondin type 1 motif 18	Extracellular matrix
3	ENSORLG00000015418	matrix metallopeptidase 15	Extracellular matrix
10	ENSORLG00000020488	transforming growth factor beta induced	Extracellular matrix
3	ENSORLG00000028055	NA	Signal trasduction (Rho)
10	ENSORLG00000020481	ArfGAP with RhoGAP domain, ankyrin repeat and PH domain 3	Signal transduction (GTPase activator RhoGap)
10	ENSORLG00000020525	TBC1 domain family member 2A-like	Signal transduction Gtpase activator, cadherin recycling
3	ENSORLG00000015260	synapse associated protein 1	TOR acitivity
10	ENSORLG00000028553	NA	NA
3	ENSORLG00000015460	dpy-19 like C-mannosyltransferase 3	glycosylation
10	ENSORLG00000022010	beta-1,4-galactosyltransferase 1	glycosylation
10	ENSORLG00000020498	protein NipSnap homolog 3A	Mitochondria
10	ENSORLG00000020494	nitric oxide associated 1	Mitochondria
10	ENSORLG00000020504	ATP-binding cassette sub-family A member 1	Metabolism (cholestrol efflux)
3	ENSORLG00000015118	phosphorylase kinase regulatory subunit beta	Metabolism
10	ENSORLG00000020493	RE1 silencing transcription factor	RNA pol II (transcritpion regulation
3	ENSORLG00000015365	solute carrier family 7 member 6 opposite strand	RNA pol II (nuclear export)
10	ENSORLG00000029052	exosome component 3	rRNA + RNA binding
3	ENSORLG00000015278	adhesion G protein-coupled receptor G3	G-protein coupled receptor
3	ENSORLG00000015287	adhesion G-protein coupled receptor G5	G-protein coupled receptor
10	ENSORLG00000025674	matrin-3	inner nuclear protein (chromatin architecture
10	ENSORLG00000023325	matrin-3	inner nuclear protein (chromatin architecture
10	ENSORLG00000022388	poly(A) binding protein interacting protein 2	Translation repressor
3	ENSORLG00000015096	integrin alpha FG-GAP repeat containing 1	NA possibly T cell activation

Table 1.4: 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

1.7.2 PSM area

Figure 1.22 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 1.4**, 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.

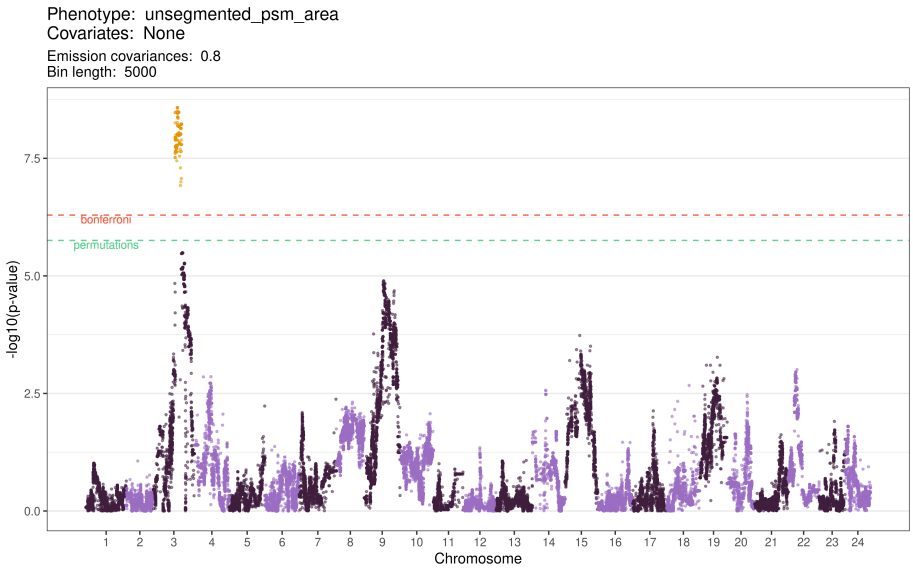

Figure 1.22: Manhattan plot of the genetic linkage results for the PSM area phenotype. Pseudo-SNPs with p-values lower than the permutation significance threshold are highlighted in yellow.

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 1.5: 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 1.5**. 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.

see above

References

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.