

Chapter 2

Genomic variations in the MIKK panel

footnote

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:

- Tomas Fitzgerald et al.¹
- Adrien Leger et al.²

2.1 The Medaka Inbred Kiyosu-Karlsruhe (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".³ This is especially true for complex traits such as behaviour, which I investigate in Chapters 3 and 4.

¹"The Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel," *Genome Biology* 23, no. 1 (February 21, 2022): 59, <https://doi.org/10.1186/s13059-022-02623-z>.

²"Genomic Variations and Epigenomic Landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel," *Genome Biology* 23, no. 1 (February 21, 2022): 58, <https://doi.org/10.1186/s13059-022-02602-4>.

³Robert Plomin and Kathryn Asbury, "Nature and Nurture: Genetic and Environmental Influences on Behavior," *The ANNALS of the American Academy of Political and Social Science* 600, no. 1 (July 1, 2005): 86–98, <https://doi.org/10.1177/0002716205277184>.

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*),⁴ common bean (*Phaseolus vulgaris L.*),⁵ tomato (*Lycopersicon esculentum*),⁶ maize (*Zea mays*),⁷ nematode (*Caenorhabditis elegans*),⁸ fruit fly (*Drosophila melanogaster*),⁹ and mouse (*Mus musculus*).¹⁰

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),¹¹ Diversity Outcross (DO)¹² and B6-by-D2 (BXD)¹³ are derived from only a small number of individuals. As gene-environment studies seek to ultimately understand their

⁴Joy Bergelson and Fabrice Roux, “Towards Identifying Genes Underlying Ecologically Relevant Traits in *Arabidopsis Thaliana*,” *Nature Reviews Genetics* 11, no. 12, 12 (December 2010): 867–79, <https://doi.org/10.1038/nrg2896>.

⁵William C. Johnson and Paul Gepts, “Segregation for Performance in Recombinant Inbred Populations Resulting from Inter-Gene Pool Crosses of Common Bean (*Phaseolus Vulgaris L.*)”, *Euphytica* 106, no. 1 (March 1, 1999): 45–56, <https://doi.org/10.1023/A:1003541201923>.

⁶Vera Saliba-Colombani et al., “Efficiency of RFLP, RAPD, and AFLP Markers for the Construction of an Intraspecific Map of the Tomato Genome,” *Genome* 43, no. 1 (February 2000): 29–40, <https://doi.org/10.1139/g99-096>.

⁷Anis M. Liinami et al., “Genetic and Physiological Analysis of Germination Efficiency in Maize in Relation to Nitrogen Metabolism Reveals the Importance of Cytosolic Glutamine Synthetase,” *Plant Physiology* 130, no. 4 (December 1, 2002): 1860–70, <https://doi.org/10.1104/pp.009647>.

⁸Kathryn S. Evans et al., “From QTL to Gene: *C. Elegans* Facilitates Discoveries of the Genetic Mechanisms Underlying Natural Variation,” *Trends in Genetics* 37, no. 10 (October 1, 2021): 935–47, <https://doi.org/10.1016/j.tig.2021.06.005>.

⁹Trudy F. C. Mackay and Wen Huang, “Charting the Genotype–Phenotype Map: Lessons from the *Drosophila Melanogaster* Genetic Reference Panel,” *WIREs Developmental Biology* 7, no. 1 (2018): e289, <https://doi.org/10.1002/wdev.289>.

¹⁰Michael C. Saul et al., “High-Diversity Mouse Populations for Complex Traits,” *Trends in Genetics* 35, no. 7 (July 1, 2019): 501–14, <https://doi.org/10.1016/j.tig.2019.04.003>.

¹¹David W. Threadgill et al., “The Collaborative Cross: A Recombinant Inbred Mouse Population for the Systems Genetic Era,” *ILAR Journal* 52, no. 1 (January 1, 2011): 24–31, <https://doi.org/10.1093/ilar.52.1.24>.

¹²Karen L Svenson et al., “High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population,” *Genetics* 190, no. 2 (February 1, 2012): 437–47, <https://doi.org/10.1534/genetics.111.132597>.

¹³Jeremy L. Pearce et al., “A New Set of BXD Recombinant Inbred Lines from Advanced Intercross Populations in Mice,” *BMC Genetics* 5, no. 1 (April 29, 2004): 7, <https://doi.org/10.1186/1471-2156-5-7>.

Need ~~the~~ Chapter info –
discuss with Gran.

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,¹⁴ and is gaining recognition elsewhere as a powerful genetic model for vertebrates.¹⁵ 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”.¹⁶

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.¹⁷ and Leger et al.¹⁸ 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).

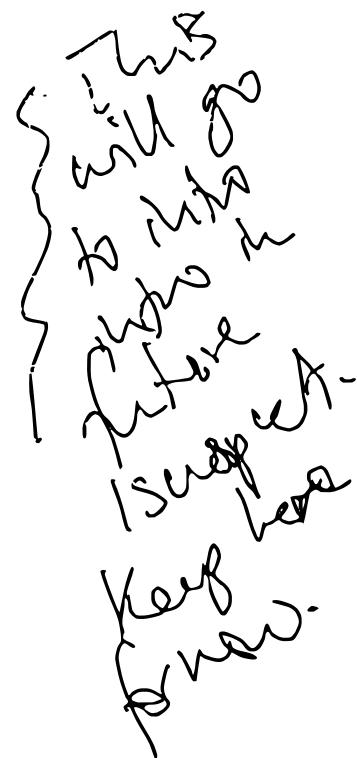
¹⁴ Joachim Wittbrodt, Akihiro Shima, and Manfred Schartl, “Medaka — a Model Organism from the Far East,” *Nature Reviews Genetics* 3, no. 1, 1 (January 2002): 53–64, <https://doi.org/10.1038/nrg704>.

¹⁵ Mikhail Spivakov et al., “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, no. 3 (March 1, 2014): 433–45, <https://doi.org/10.1534/g3.113.008722>.

¹⁶ Fitzgerald et al., “The Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel.”

¹⁷

¹⁸ “Genomic Variations and Epigenomic Landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel.”



*We sequenced
everything we could!*

2.2 Genomic characterisation of the MIKK panel

2.2.1 MIKK panel DNA sequence dataset

For the preparation of Fitzgerald et al.¹⁹ 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.²⁰ 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.²¹ 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

¹⁹“The Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel.”

²⁰“Genomic Variations and Epigenomic Landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel.”

²¹Spivakov et al., “Genomic and Phenotypic Characterization of a Wild Medaka Population.”

*we performed ~ pilot study
we sequenced everything we could!*

the MIKK panel. I used this data on relative eye diameters to generate **Fig. 2.1C**.

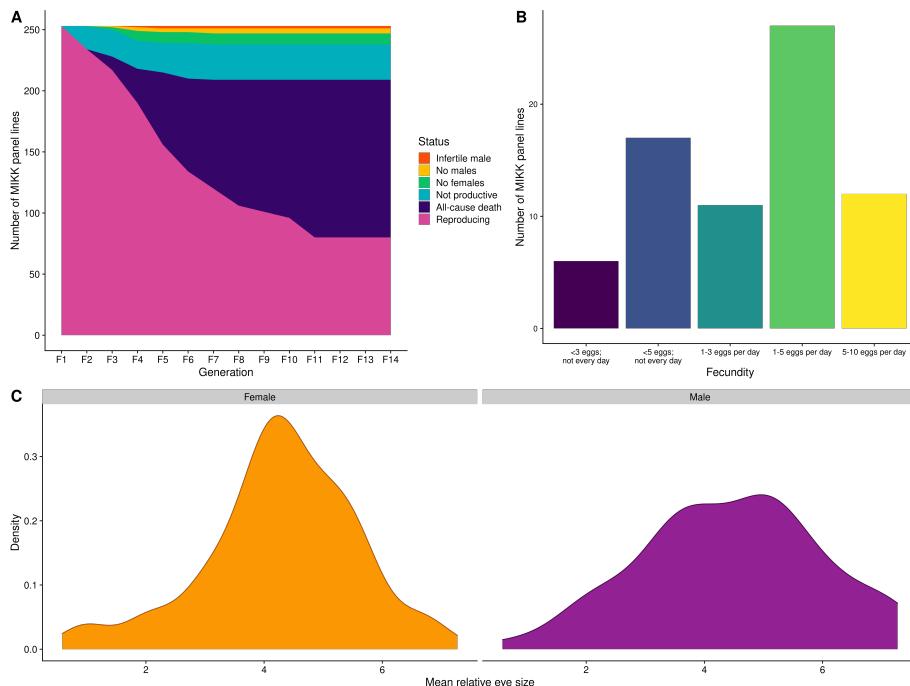


Figure 2.1: Inbreeding, fecundity and eye size in the MIKK panel lines. **A:** Status of all MIKK panel lines during the first 14 generations of inbreeding, showing cause of death for non-extant lines. **B:** Average fecundity of MIKK panel lines in generation F16, as measured during peak egg production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

as we can see...
← Explain me
... be distinct
of phenotypes...

2.2.3 Introgession 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.²² Using the

²²Index of /Pub/Release-102/Emf/Ensembl-Compara/Multiple_alignments/50_fish.epo/, accessed January 25, 2022, https://ftp.ensembl.org/pub/release-102/emf/ensembl-compara/multiple_alignments/50_fish.epo/.

introduce question
here - is there interbreeding North/South?
or earlier.

phylogenetic tree provided with the dataset, and the *ape* R package,²³ 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 then combined that dataset with the MIKK Illumina call set and variant calls for the southern Japanese *iCab* strain (see 2.2.1). *ancestor* .

I then carried out an ABBA BABA analysis to calculate a modified ‘admixture proportion’ statistic \hat{f}_d ²⁴ 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 Martin, Davey, and Jiggins²⁵ on their GitHub page.²⁶ *Explain* .

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.²⁷, 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.²⁸

slow surprise wrt to geography!
speciation

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

²³Emmanuel Paradis and Klaus Schliep, “Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R,” *Bioinformatics* 35, no. 3 (February 1, 2019): 526–28, <https://doi.org/10.1093/bioinformatics/bty633>.

²⁴Simon H. Martin, John W. Davey, and Chris D. Jiggins, “Evaluating the Use of ABBA-BABA Statistics to Locate Introgressed Loci,” *Molecular Biology and Evolution* 32, no. 1 (January 1, 2015): 244–57, <https://doi.org/10.1093/molbev/msu269>.

²⁵Simon martin, *Simonhmartin/Genomics_general*, 2022, https://github.com/simonhmartin/genomics_general.

²⁶“Genomic and Phenotypic Characterization of a Wild Medaka Population.”

²⁷Takafumi Katsumura et al., “Medaka Population Genome Structure and Demographic History Described via Genotyping-by-Sequencing,” *G3 Genes/Genomes/Genetics* 9, no. 1 (January 1, 2019): 217–28, <https://doi.org/10.1534/g3.118.200779>.

Fst?

Question being answered
“I am curious”
“What can we learn”
“What is the history”
“What are the characteristics”
“What is the relationship”

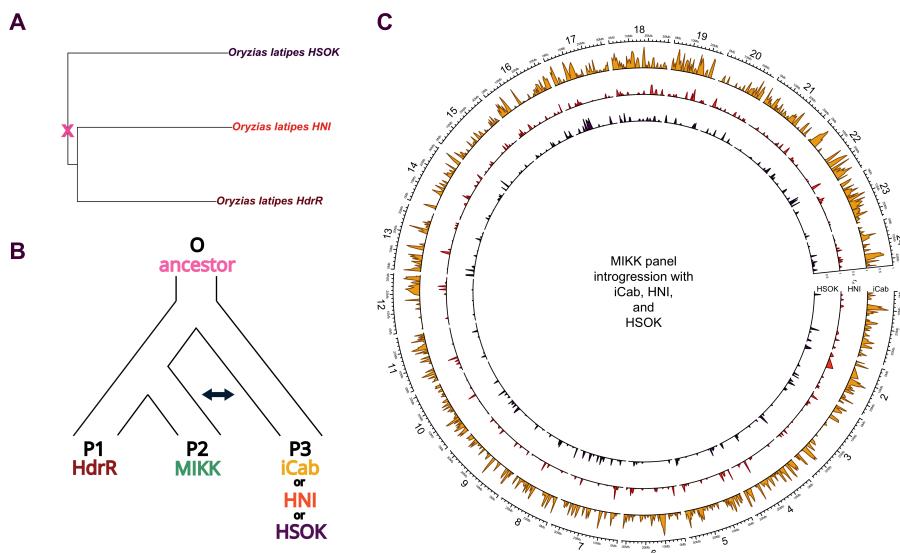


Figure 2.2: **Figure 2:** ABBA-BABA analysis. **A.** Phylogenetic tree generated from the Ensembl release 102 50-fish multiple alignment, showing only the medaka lines used in the ABBA-BABA analysis. **B.** Schema of the comparisons carried out in the ABBA-BABA analysis. **C.** Circos plot comparing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

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.

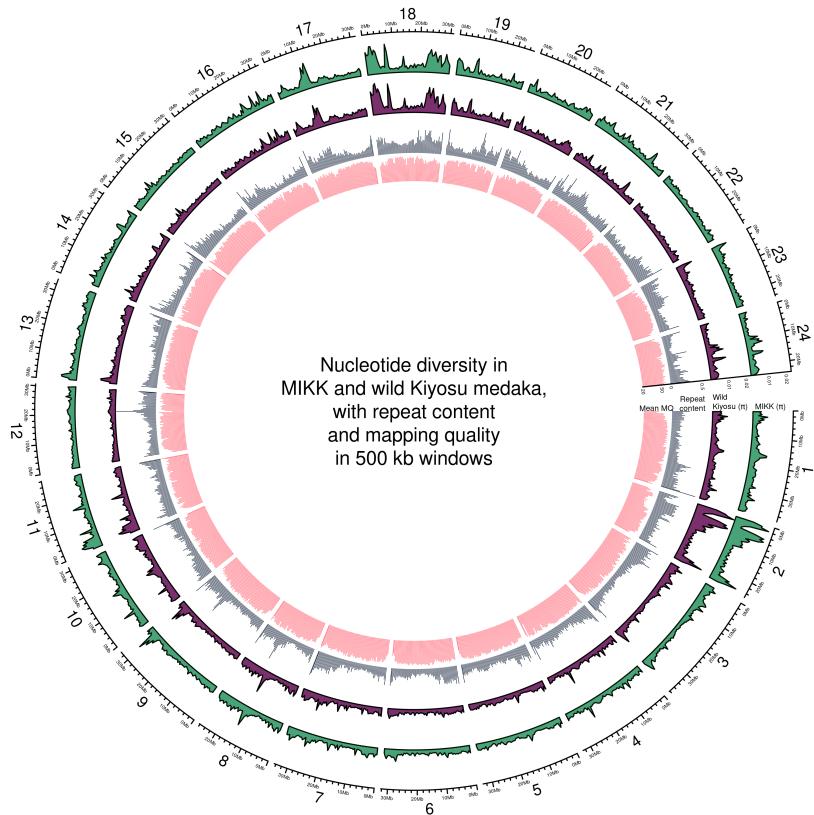


Figure 2.3: Circos plot with nucleotide diversity ($\hat{\pi}$) calculated within 500-kb non-overlapping windows for 63 non-sibling lines from the MIKK panel (green) and 7 wild Kiyosu medaka samples from the same originating population (purple); proportion of sequence classified as repeats by RepeatMasker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

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.²⁹ Inversions cause permanent heterozygosity,³⁰ and duplications and deletions may have increased the density of called SNPs in these regions,³¹ so the observed depressions in homozygosity 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.

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.³² 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.³³ I then used PLINK³⁴ to calculate the minor allele frequencies for all non-missing, biallelic SNPs in both the MIKK non-sibling and 1KG 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,

²⁹Leger et al., “Genomic Variations and Epigenomic Landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel.”

³⁰Ary A. Hoffmann, Carla M. Sgr'o, and Andrew R. Weeks, “Chromosomal Inversion Polymorphisms and Adaptation,” *Trends in Ecology & Evolution* 19, no. 9 (September 1, 2004): 482–88, <https://doi.org/10.1016/j.tree.2004.06.013>.

³¹David Fredman et al., “Complex SNP-related Sequence Variation in Segmental Genome Duplications,” *Nature Genetics* 36, no. 8, 8 (August 2004): 861–66, <https://doi.org/10.1038/ng1401>.

³²“A Global Reference for Human Genetic Variation,” *Nature* 526, no. 7571 (2015): 68–74, <https://doi.org/10.1038/nature15393>.

³³Aaron McKenna et al., “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data,” *Genome Research* 20, no. 9 (January 9, 2010): 1297–1303, <https://doi.org/10.1101/gr.107524.110>.

³⁴Christopher C Chang et al., “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets,” *GigaScience* 4, no. 1 (December 2015): 7, <https://doi.org/10.1186/s13742-015-0047-8>; Shaun M Purcell and Christopher C Chang, *PLINK 1.9*, n.d., www.cog-genomics.org/plink/1.9/.

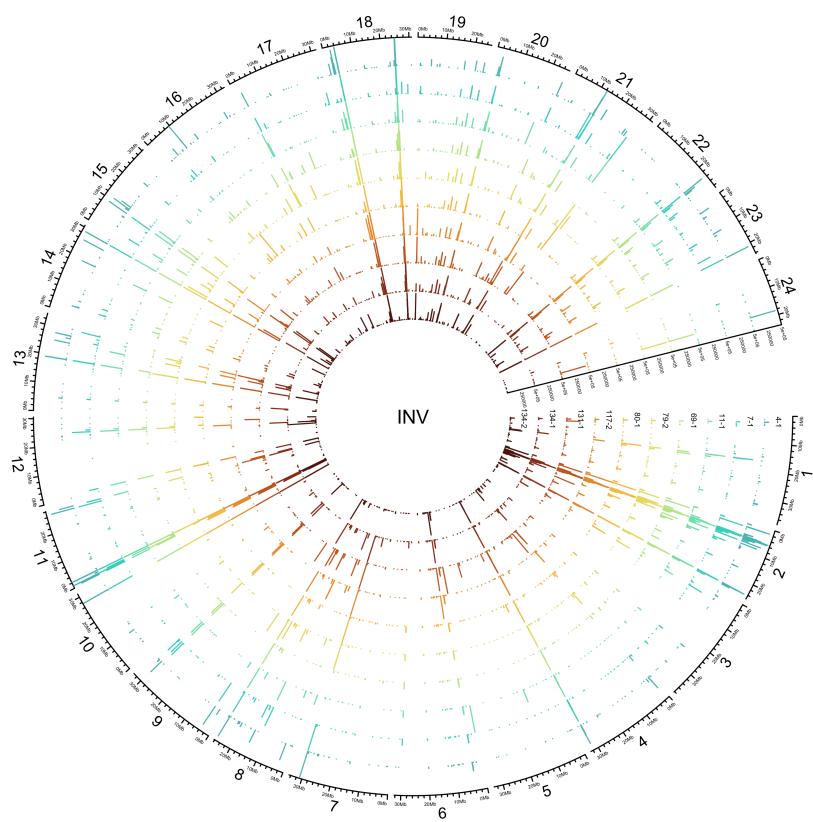
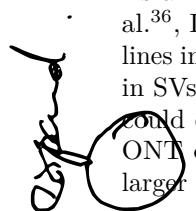


Figure 2.4: Inversions identified in 9 MIKK panel lines using a combination of Oxford Nanopore Technologies long-read and Illumina short-read sequences (see Chapter 2.3 below).

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 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 Kiyoshi Naruse et al.³⁵, 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.³⁶, 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.

³⁵“A Detailed Linkage Map of Medaka, *Oryzias latipes*: Comparative Genomics and Genome Evolution,” *Genetics* 154, no. 4 (April 1, 2000): 1773–84, <https://www.genetics.org/content/154/4/1773>.

³⁶“Genomic Variations and Epigenomic Landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel.”

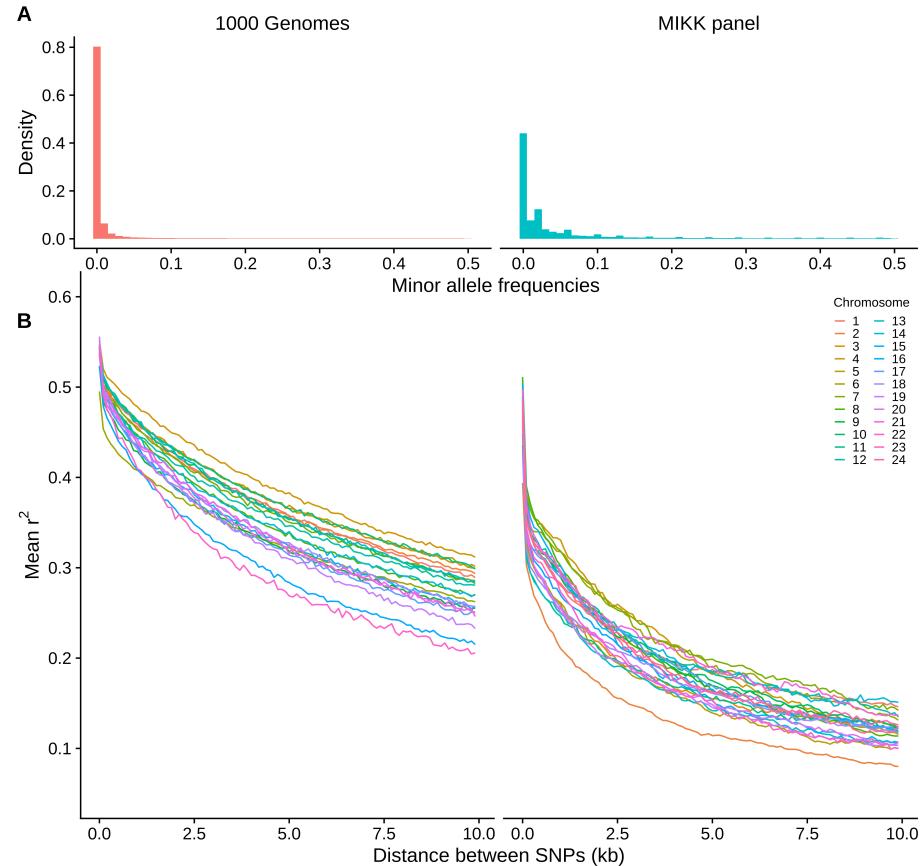


Figure 2.5: Minor allele frequency distributions and LD decay for biallelic, non-missing SNPs in the 1000 Genomes Phase 3 variant calls ($N = 2,504$) (1KG), and the MIKK panel Illumina-based calls excluding one of each pair of sibling lines ($N = 63$), across all autosomes (1KG: chrs 1-22; MIKK: chrs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

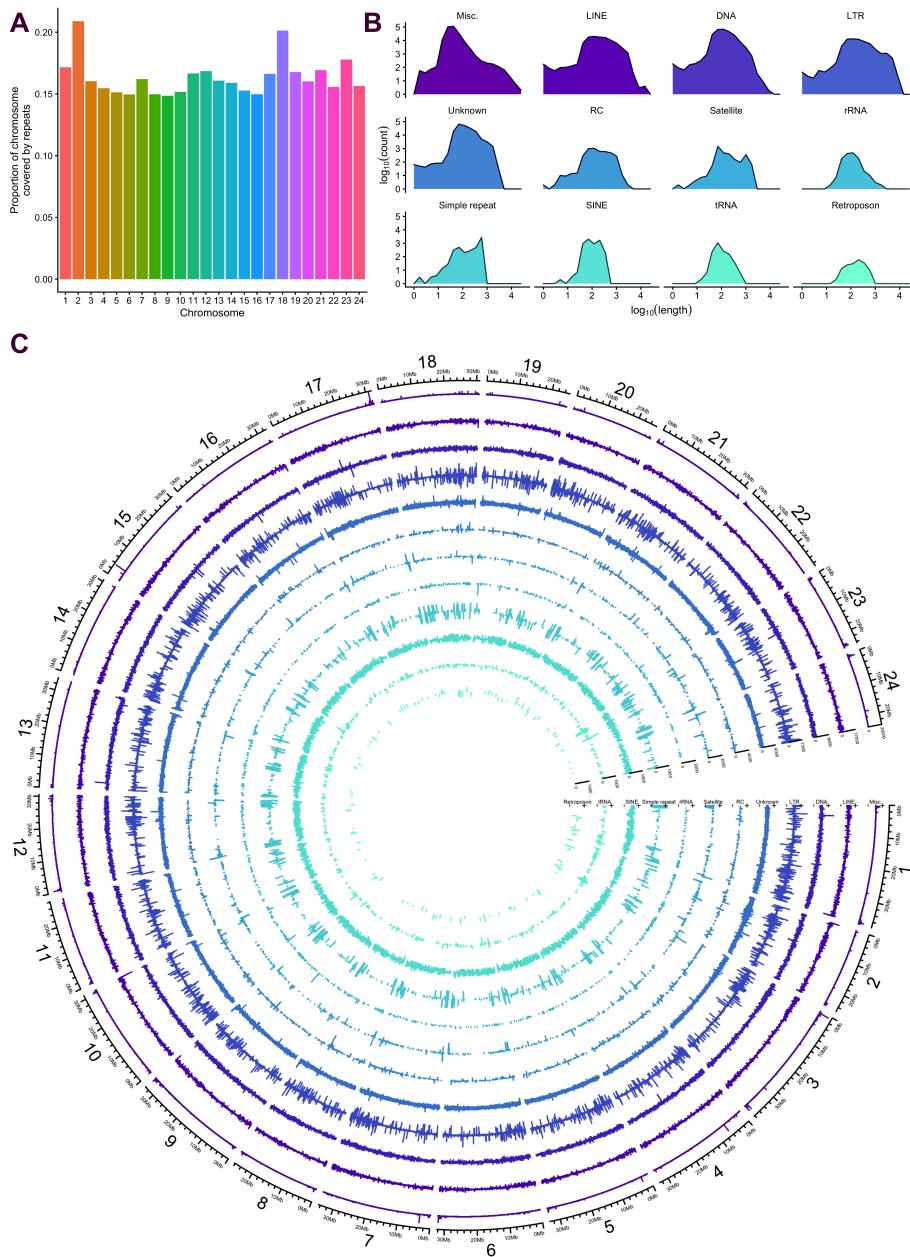


Figure 2.6: Repeat content in the *HdrR* genome based on RepeatMasker results obtained by Jack Monahan. **A.** Proportion of repeat content per-chromosome. **B.** log₁₀ of repeat lengths and counts per repeat class. “Misc” includes all repeats assigned to their own specific class, for example “(GAG)_n”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

The 9 “polished” samples contained a mean per-sample count of approximately 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)³⁷ 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.³⁸). 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 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

³⁷Monkol Lek et al., “Analysis of Protein-Coding Genetic Variation in 60,706 Humans,” *Nature* 536, no. 7616, 7616 (August 2016): 285–91, <https://doi.org/10.1038/nature19057>.

³⁸“Genomic Variations and Epigenomic Landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) Panel.”

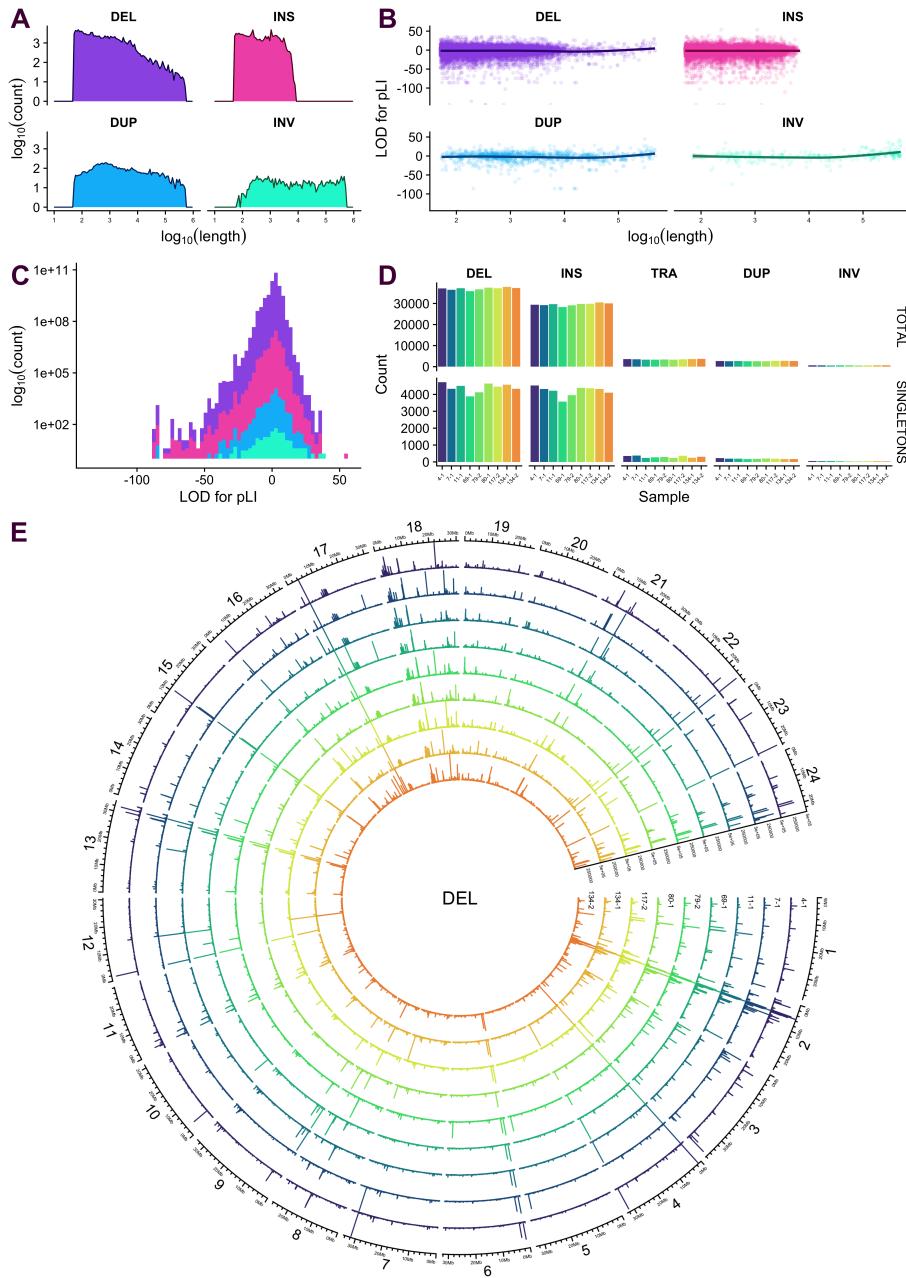


Figure 2.7: Polished SVs in 9 MIKK panel lines sequenced with ONT. DEL: deletion; INS: insertion; TRA: translocation; DUP: duplication; INV: inversion. **A.** Aggregate \log_{10} counts and lengths of distinct SVs by type, excluding TRA. **B.** pLI LOD scores in distinct SVs by SV type. **C.** Histogram 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.

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.

8nt
segregating

as expected

fst?