Genetic analysis of quantitative traits in medaka fish and humans

Ian Brettell

2022-05-26

Contents

Chapter 1

About

Code to render PDF:

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

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

1.1 Usage

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

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

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

1.2 Render book

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

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

Or build the book from the R console:

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

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

1.3 Preview book

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

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

Chapter 2

Genomic variations in the MIKK panel

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

This chapter will set out my contributions to the the following pair of papers published in *Genome Biology*:

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

2.1 Genomic characterisation of the Medaka Inbred Karlsruhe-Kiyosu (MIKK) panel

2.1.1 Assessing the inbreeding trajectory of the MIKK panel

The Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel was bred from a wild population of medaka found in the Kiyosu area near Toyohashi, Aichi Prefecture, in southern Japan. Mikhail Spivakov et al.[3] From this wild population, the

[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.

[3] "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.

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

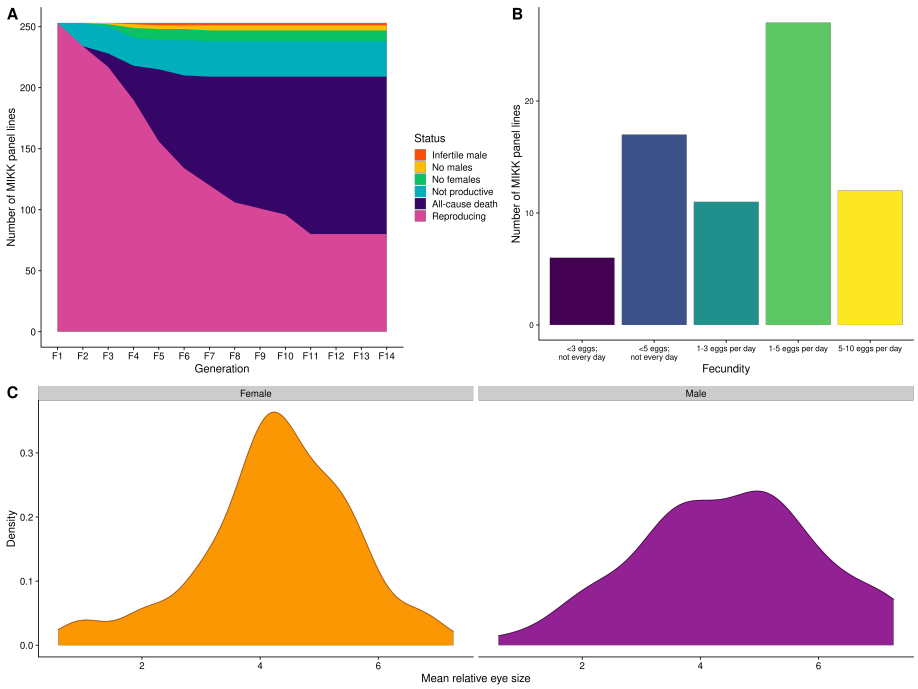

Figure 2.1: Inbreeding,fecundity and eye size in the MIKK panel lines. **A**: Status of all MIKK panel lines during the first 14 generations of inbreeding, showing cause of death for non-extant lines. **B**: Average fecundity of MIKK panel lines in generation F16, as measured during peak egg production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

2.1.2 Introgression with northern Japanese and Korean medaka populations

To explore the evolutionary history of the MIKK panel's founding population, we sought to determine whether there was evidence of introgression between that southern Japanese population, and northern Japanese and Korean medaka populations. [Refer to background on northern medaka speciating away from southern medaka.] 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.(2) I then carried out an ABBA BABA analysis to calculate a modified 'admixture proportion' statistic fd (3) 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**).

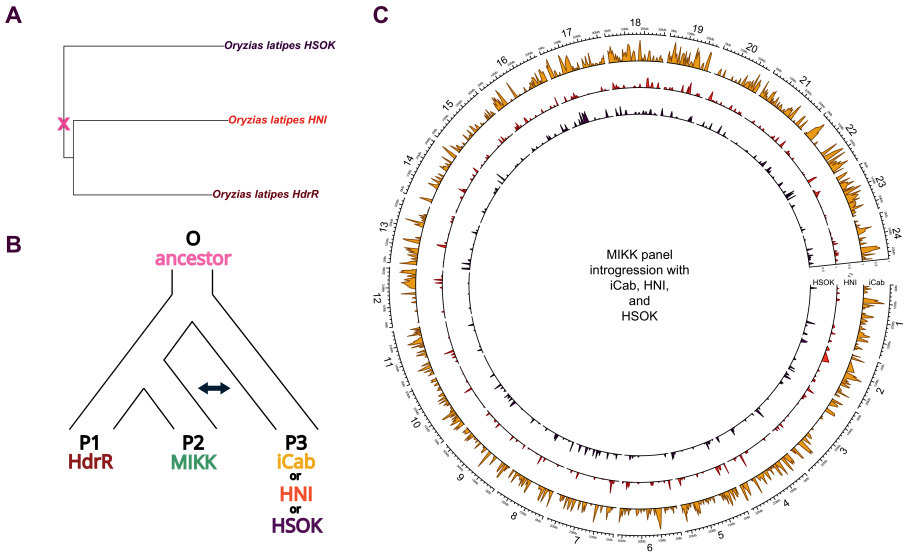

Figure 2.2: **Figure 2**: ABBA-BABA analysis. **A**. Phylogenetic tree generated from the Ensembl release 102 50-fish multiple alignment, showing only the medaka lines used in the ABBA-BABA analysis. B. Schema of the comparisons carried out in the ABBA-BABA analysis. C. Circos plot comparing introgression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

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

introgressed with medaka from Korea than with medaka from northern Japan. This supports the findings in Spivakov et al.[4], where the authors found little evidence of significant interbreeding between southern and northern Japanese medaka since the populations diverged.

2.1.3 Nucleotide diversity

As an additional means of assessing genetic diversity in the MIKK panel, I calculated nucleotide diversity (π) within 500-kb non-overlapping windows across the genome of 63 of the 80 MIKK panel lines (having excluded one line from each pair of sibling lines), and compared this to the nucleotide diversity in 7 wild medaka from the same Kiyosu population from which the MIKK panel was derived. Mean and median nucleotide diversity in both the MIKK panel and wild Kiyosu medaka were close to 0, and slightly higher in the MIKK panel (mean : MIKK = 0.0038, wild = 0.0037; median : MIKK = 0.0033, wild = 0.0031). The patterns of varying nucleotide diversity across the genome are shared between the MIKK panel and wild Kiyosu medaka, where regions with high levels of repeat content tend to have higher nucleotide diversity ($r = 0.386$, $p < 0.001$) 2.3.

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 2.4. These regions are also enriched for large deletions and duplications [cite companion paper]. Inversions cause permanent heterozygosity,[5] and duplications and deletions may have increased the density of called SNPs in these regions (5), 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.

2.1.4 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, we first filtered the Illumina-based variant calls to include only one inbred line from each pair of sibling lines, leaving 63 non-sibling inbred lines. Fig. 5A compares the allele frequency distribution for the 16.4M biallelic SNPs identified in those filtered calls, with the allele frequency distribution of the 81M biallelic SNPs in the 1000 Genomes Project Phase 3 release (N = 2,504) (1KG) in human. 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.

---

[4]

[5]Ary A. Hoffmann, Carla M. Sgr'o, and Andrew R. Weeks, "Chromosomal Inversion Polymorphisms and Adaptation," *Trends in Ecology & Evolution* 19, no. 9 (September 1,

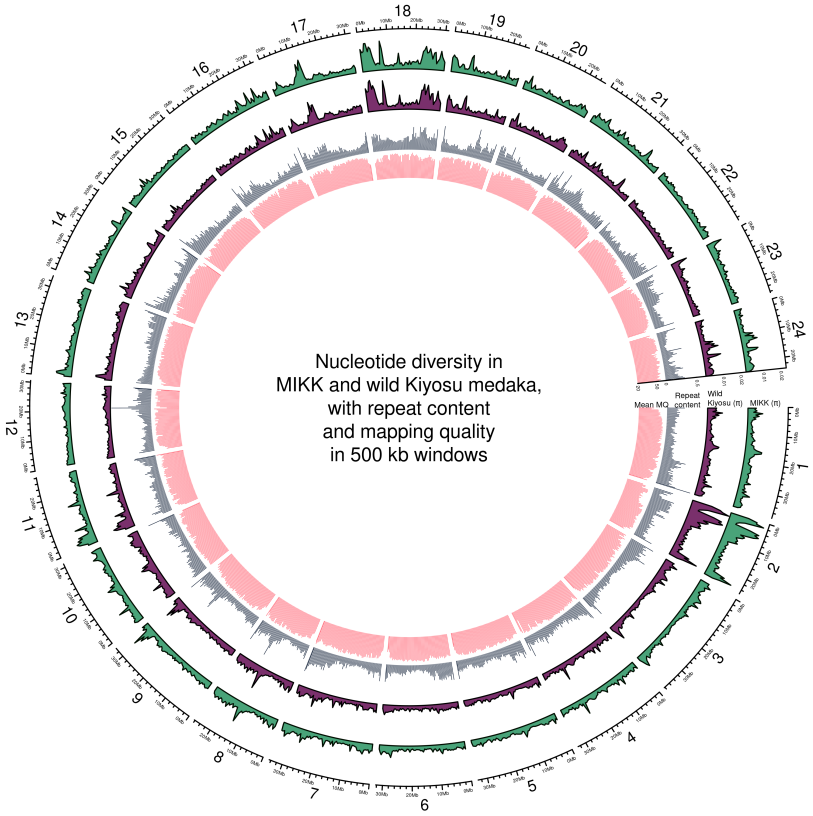

Figure 2.3: Circos plot with nucleotide diversity (π) calculated within 500-kb non-overlapping windows for 63 non-sibling lines from the MIKK panel (*green*) and 7 wild Kiyosu medaka samples from the same originating population (*purple*); proportion of sequence classified as repeats by RepeatMasker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

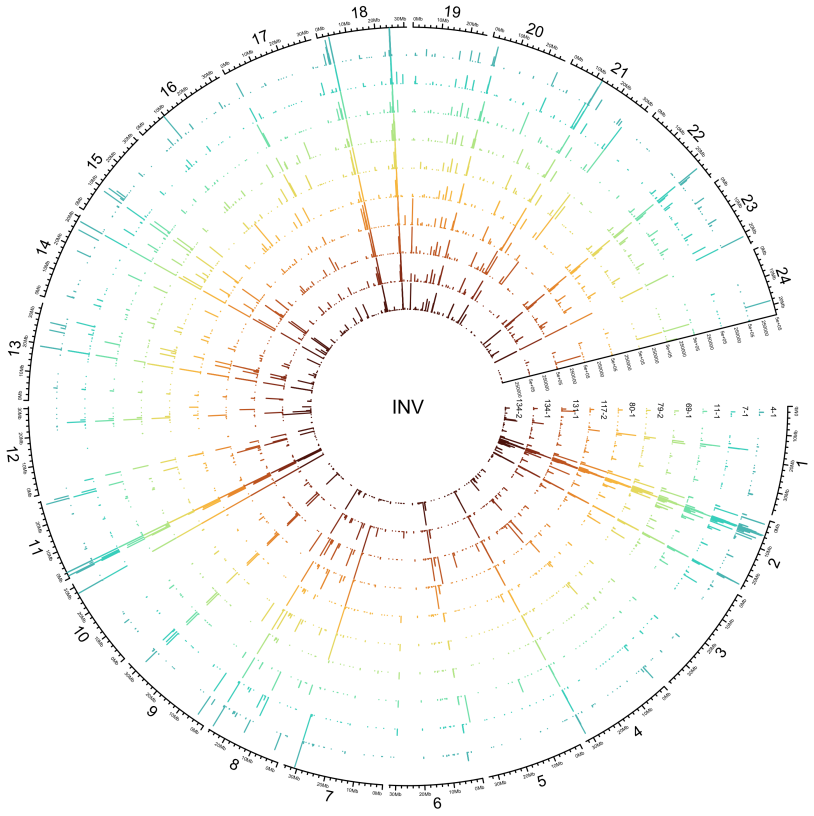

Figure 2.4: Inversions identified in 9 MIKK panel lines using a combination of Oxford Nanopore Technologies long-read and Illumina short-read sequences (see Genomic variations and epigenetic landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel below).

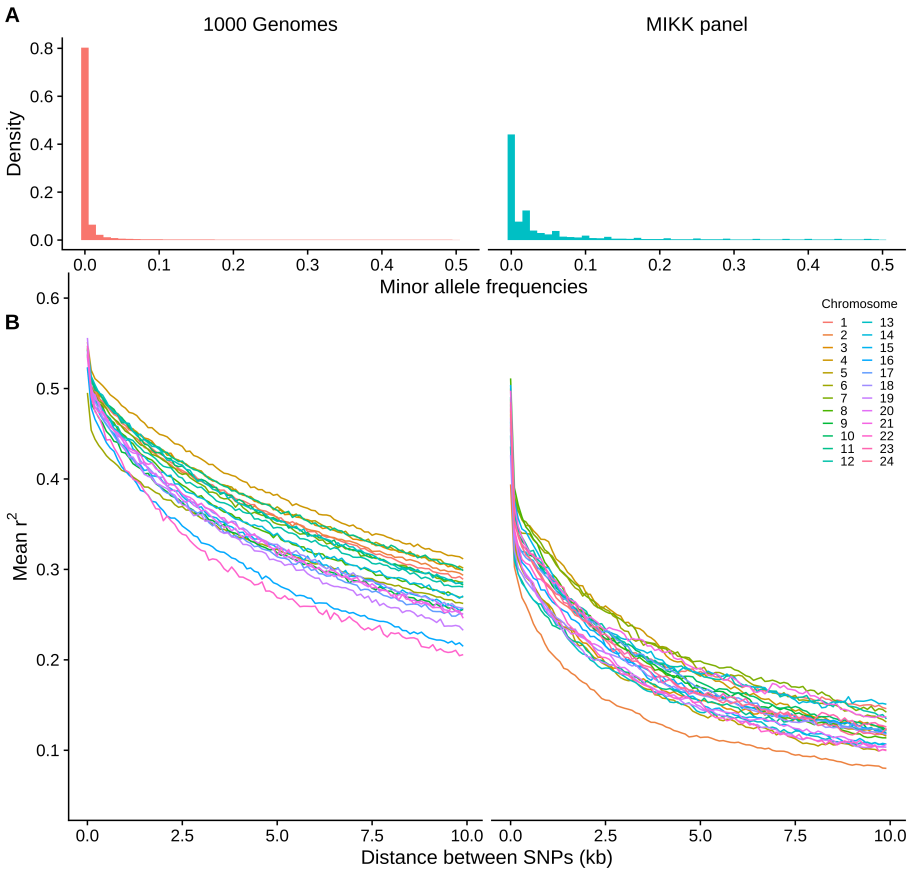

Figure 2.5: Linkage disequilibrium on chr 17, and LD decay in the MIKK panel and 1000 Genomes human dataset. **A**: Minor allele frequency distributions and LD decay for biallelic, non-missing SNPs in the 1000 Genomes Phase 3 variant calls (N = 2,504) (1KG), and the MIKK panel Illumina-based calls excluding one of each pair of sibling lines (N = 63), across all autosomes (1KG: chrs 1-22; MIKK: chrs 1-24); Density of allele frequencies in the 1KG and MIKK panel calls. **B**: LD decay for each autosome, calculated by taking the mean r2 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. **Inset**: mean r2 within 100 bp windows, up to a maximum of 1 kb. LD decays faster on chromosome 2 for the MIKK panel due to its higher recombination rate, consistent with the genetic linkage map described in (6). B: Heatmap showing the genotypes of 63 non-sibling MIKK panel lines across the region of high LD on chr17. **C**: Figure generated by Haploview [cite] of a large region on chr17 with high LD.

To review the MIKK panel's LD structure, for each autosome in humans (chromosomes 1-22) and each chromosome in medaka (chromosomes 1-24), we calculated r2 between all pairs of biallelic SNPs with a minor allele frequency (MAF) greater than 0.10, within 10 kb of one another [elaborate from Methods]. We then grouped the paired SNPs by distance from one another into non-overlapping 100 bp windows, and calculated the mean r2 for each window in order to represent how LD decays with distance between loci. **Fig. 5B** compares the LD decay between the 1KG and MIKK panel SNPs. Based on the 1KG calls under these parameters, LD decays in humans to a mean r2 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 r2 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 (6) 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. 6**).

2.2 Genomic variations and epigenetic landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel

As an alternative to the variation pangenome approach described in this paper [cite companion paper], I explored structural variation (SVs) 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 Nanopore sequencing 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. The Birney Group first called structural variants using only the ONT long reads, producing a set of structural variants classified into five types: deletions (DEL), insertions (INS), translocations (TRA), duplications (DUP) and inversions (INV). I then "polished" the called DEL and INS variants with Illumina short reads to improve their accuracy. The polishing process filtered out 7.4% of DEL and 12.8% of INS variants, and adjusted the breakpoints (i.e. start and end positions) for 75-77% of DEL and INS variants in each sample by a mean of 23 bp for the start position, and 33 bp for the end position. This process produced a total of 143,326 filtered SVs.

The 9 "polished" samples contained a mean per-sample count of approximately 37K DEL variants (12% singletons), 29.5K INS variants (14%), 3.5K TRA

2004): 482–88, https://doi.org/10.1016/j.tree.2004.06.013.

variants (9%), 2.5K DUP (7%) and 600 INV (7%) (Fig. 4D). 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. 4A). Fig. 4E 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. 6A).

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) (39) by calculating the logarithm of odds (LOD) for the pLI scores of all genes overlapping the variant (Fig. 4B,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. 4B).

We compared these polished INS and DEL calls with the high-quality graph-based alternative paths and large-scale deletions, respectively. We found that 2 of the 19 regions covered by graph-based alternative paths [cite], and 4 of the 16 regions covered by graph-based deletions [cite], 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 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.

Fitzgerald, Tomas, Ian Brettell, Adrien Leger, Nadeshda Wolf, Natalja Kusminski, Jack Monahan, Carl Barton, et al. "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.

Hoffmann, Ary A., 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.

Leger, Adrien, Ian Brettell, Jack Monahan, Carl Barton, Nadeshda Wolf, Natalja Kusminski, Cathrin Herder, et al. "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.

Spivakov, Mikhail, Thomas O Auer, Ravindra Peravali, Ian Dunham, Dirk Dolle, Asao Fujiyama, Atsushi Toyoda, 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.