Fitzgerald *et al. Genome Biology* (2022) 23:59
https://doi.org/10.1186/s13059-022-02623-z

Genome Biology

RESEARCH Open Access

The Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel

Tomas Fitzgerald[1†], Ian Brettell[1†], Adrien Leger[1], Nadeshda Wolf[2], Natalja Kusminski[2], Jack Monahan[1], Carl Barton[1], Cathrin Herder[2], Narendar Aadepu[2,3], Jakob Gierten[3], Clara Becker[3], Omar T. Hammouda[3], Eva Hasel[3], Colin Lischik[3], Katharina Lust[3], Natalia Sokolova[3], Risa Suzuki[3], Erika Tsingos[3], Tinatini Tavhelidse[3], Thomas Thumberger[3], Philip Watson[3], Bettina Welz[3], Nadia Khouja[2], Kiyoshi Naruse[4], Ewan Birney[1], Joachim Wittbrodt[3] and Felix Loosli[2*]

*Correspondence: felix.loosli@kit.edu

[†]Tomas Fitzgerald and Ian Brettell are equal contributors.

[2] Institute of Biological and Chemical Systems, Biological Information Processing (IBCS-BIP), Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

Full list of author information is available at the end of the article

Abstract

Background: Unraveling the relationship between genetic variation and phenotypic traits remains a fundamental challenge in biology. Mapping variants underlying complex traits while controlling for confounding environmental factors is often problematic. To address this, we establish a vertebrate genetic resource specifically to allow for robust genotype-to-phenotype investigations. The teleost medaka (*Oryzias latipes*) is an established genetic model system with a long history of genetic research and a high tolerance to inbreeding from the wild.

Results: Here we present the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel: the first near-isogenic panel of 80 inbred lines in a vertebrate model derived from a wild founder population. Inbred lines provide fixed genomes that are a prerequisite for the replication of studies, studies which vary both the genetics and environment in a controlled manner, and functional testing. The MIKK panel will therefore enable phenotype-to-genotype association studies of complex genetic traits while allowing for careful control of interacting factors, with numerous applications in genetic research, human health, drug development, and fundamental biology.

Conclusions: Here we present a detailed characterization of the genetic variation across the MIKK panel, which provides a rich and unique genetic resource to the community by enabling large-scale experiments for mapping complex traits.

Keywords: Inbred panel, Medaka, Genetics, Quantitative traits, Genome sequencing, Population genetics, Copy number variation, eQTL

Background

The relationship between natural genetic variation and the variance of quantitative traits in different species is one of the founding questions in genetics [1, 2] and has become a very active field of research today. Experimentally, it has been addressed using plant and animal models [3–5] and has been studied in human populations [6–8]. In model organisms such as *Arabidopsis thaliana* and *Drosophila melanogaster*, genome-wide association studies (GWAS) and specific crosses have been used to examine complex genetic traits, bridging population association models to more traditional controlled-cross strategies [9, 10]. GWAS have had tremendous success in discovering genomic loci underlying human traits by leveraging observational outbred cohorts of individuals [11]. However, this outbred sampling strategy still leaves many questions in vertebrate genetics unanswered, including the importance of gene-by-environment interactions (GxE) and epistatic interactions, i.e., between two genetic loci (GxG)—for phenotypic extremes in particular—as well as the potential combination of both interactions (GxGxE) [12, 13]. To explore these questions, one must turn to laboratory vertebrates where one can vary and control both genetic and environmental sources of variance.

The mouse (*Mus musculus*) is an established vertebrate model for human genetics, and researchers have created panels of recombinant inbred lines (RILs), such as the Collaborative Cross (CC), Diversity Outbred cross (DO), and the BXD cross [14–17], in order to run GWAS in these managed populations [18] and have a controlled source of genetic variation. As mice are mammals, they have excellent orthologous organ systems and cell types to humans, and an unsurpassed repertoire of tools to control their precise genome, including large-scale genomic engineering [19, 20]. However, the genetic variation of all laboratory mice follows from the complicated history of the domestication of "fancy mice," originating from three separate species that were bred in captivity, and then undergoing complex domestication before the laboratory lines were established [21, 22]. As such, polymorphisms between laboratory strains and GxG (epistatic) interactions result from the non-natural creation of these lines. Furthermore, even large panels of RILs, such as the CC, DO, and BXD, come from a limited number of parents [23]. The unnatural genetic origin of laboratory mice and their limited parentage means that they have deficiencies in modelling outbred populations such as humans. It would be optimal to supplement the mouse RIL lines with other vertebrate species both to better capture outbred settings and to provide another window into vertebrate genetics that can be controlled in the laboratory. In addition, association studies often require high numbers of phenotyping experiments, so the ease of phenotyping and economical husbandry are also important features for a suitable vertebrate animal model.

The teleost medaka (*Oryzias latipes*) is a long-established model organism that combines a number of important features [24]. Economical husbandry, high fecundity, and transparency of embryos and largely also adults allow one to carry out a wide range of phenotyping on large numbers of individuals. Medaka has a long history in genetic research [2, 25] and comes with a wide range of established molecular genetic tools that allow in-depth analysis of gene function and phenotypes [26], including CRISPR/Cas-based homologous recombination protocols that permit the precise testing of genetic variants in different inbred strains [27]. Over 70% of human genes have teleost orthologs, and nearly all the major organ systems in humans have a teleost counterpart [28, 29], making it possible to translate many findings between species.

Also with respect to sex determination, medaka is a valuable genetic model. Sex determination and the evolution of sex chromosomes are very active fields of research and are studied in vertebrate and invertebrate model systems [30]. Medaka has a male heterogametic XX-XY sex determination system. The molecular differences between the X and the male Y chromosome are subtle, and not visible in chromosome spreads, leading to the X/Y chromosome assigned the label "Chromosome 1" in medaka genetics. The small Y-specific segment suggests that the XY heterogametic sex system in medaka is at an early stage of evolution [31]. This segment contains the male sex determination gene, the transcription factor *DMY* (also known as *dmrt1b*), which is sufficient and necessary to induce male sex determination [32]. Interestingly, DMY is the male sex determination gene in the closely related sister species *Oryzias curvinotus* but not in *Oryzias luzonesis.* Furthermore, other *Oryzias* species have a female heterogametic ZZ-ZW sex determination system [32]. Thus, sex determination in the genus *Oryzias* shows a high divergence between evolutionary closely related species. This renders the genus *Oryzias* ideal for comparative studies of the evolution of vertebrate sex determination and the genetic cascade of downstream genes.

Importantly, medaka is highly tolerant to inbreeding from the wild [33]. Isogenic inbred medaka strains have been established from natural populations, and a number of these laboratory strains have been inbred for over a hundred generations [34]. This indicates that medaka's tolerance to inbreeding is a species-specific trait, rather than depending on the starting population. These medaka isogenic inbred strains have been bred by single full-sibling-pair (brother-sister) crosses [35]. Such full-sibling-pair crosses are commonly used for inbreeding in medaka as they confer a strong genetic bottleneck in each generation without requiring the maintenance of more than one generation per cross, as opposed to "backcrosses" which involve crossing F1 individuals with their parents.

We have previously identified a polymorphic wild population, sampled from Kiyosu, Japan [36]. By single full-sibling-pair inbreeding for 9 generations, we have established a panel of 80 near-isogenic inbred lines from this wild population, known as the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel. In this paper, we describe this panel, outline the success of this inbreeding strategy, and provide a view of the lines based on their whole-genome sequences. We show that both molecular and organismal phenotypes are distinguishable in this panel and that molecular traits can be mapped to specific loci. We also comment on aspects of medaka genetics and genomic biology from the deep sequencing of the MIKK panel, from population genetics to loss of function alleles and structural variation. We show that the MIKK panel is an inbred near-isogenic resource with a good representation of wild alleles in the Kiyosu population, and is well suited for future genetic association and mapping studies.

Results

Inbreeding of the MIKK panel

We have previously reported the identification of a wild medaka population that satisfies critical conditions for the establishment of a vertebrate near-isogenic panel [36]. In July 2010, we selected a genetically diverse medaka population from the Kiyosu area near Toyohashi, Aichi Prefecture, and determined that it was free of significant population structure and introgression from aquarium populations. Segregation analysis of trios revealed advantageous linkage disequilibrium (LD) properties, in terms of relatively short blocks and sharp LD decay, with the LD estimates expressed as r^2 reaching a minimum of ~ 0.12 at a distance of approximately 12.5 kb.

To commence the inbreeding process, we first established founder families by setting up 115 random crosses of single mating pairs from the wild Kiyosu population. For each founder family, we then set up mostly two, but up to five, single full-sibling-pair inbreeding crosses. This resulted in a total of 253 single full-sibling-pair crosses for the F1 generation, which we used to initiate the inbreeding lines. Lines derived from the same founder family are referred to as "sibling lines." In the subsequent 8 generations of inbreeding, we only used one mating pair per line. In total, 19 of the F1 crosses did not result in productive mating, so we proceeded to inbreed the F2 generation with 234 productive single full-sibling-pair crosses.

During the first 9 generations of inbreeding (using only one mating pair per line), to avoid selecting for specific traits, we intentionally did not consider body size, generation time, fecundity, fertilization rate, survival rate, or sex ratio. We continued inbreeding with all lines independent of these traits up to the 9th generation, or until the strain became extinct (Fig. 1A). The only criterion we considered when selecting individuals for inbreeding crosses was average wild-type morphology (irrespective of size), so we did not use individuals with severe malformations, such as a strongly bent tail. When all fish in a given line showed the same morphological abnormalities, for example line 4-1 which has an unusually large abdomen, we continued inbreeding with these fish, assuming a genetic cause for the malformation.

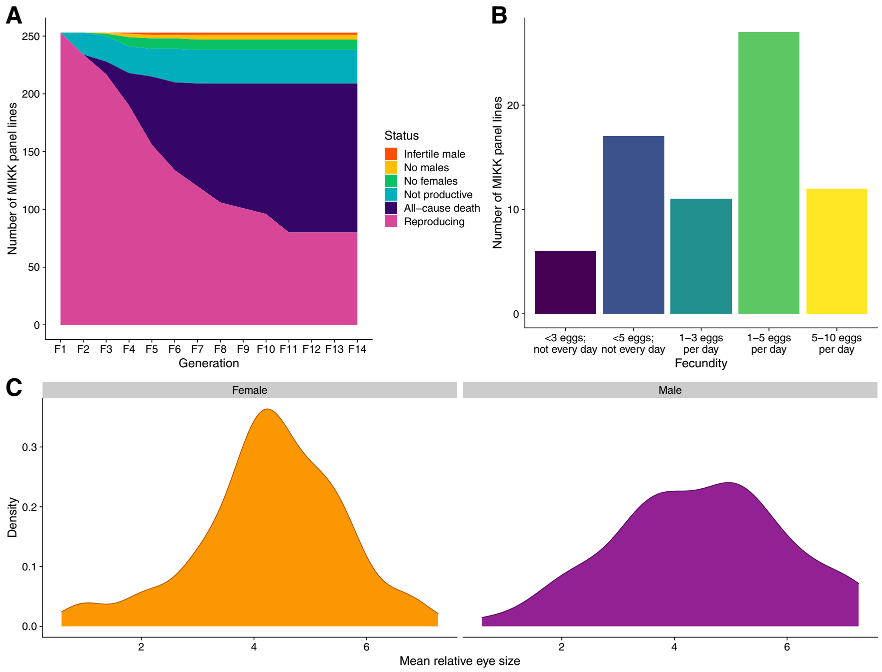

Fig. 1 Inbreeding, fecundity and eye size in the MIKK panel lines. **A** Status of all MIKK panel lines during the first 14 generations of inbreeding, showing cause of death for non-extant lines. **B** Average fecundity of MIKK panel lines in generation F16, as measured during peak egg production in July 2020. **C** Distribution of mean relative eye size for male and female medaka across all MIKK panel lines

The most frequent cause for extinction of a line during inbreeding was all-cause mortality (Fig. 1A). In the majority of cases, individuals died during the first 3 weeks after hatching due to unknown causes. The highest incidence of line extinction occurred while inbreeding generations 3 to 5. Dissection of dead fish revealed a severe infection of inner organs with *Flavobacterium columnare.* Antibiotic treatment stopped mortality within 24 h, indicating that the bacterial infection was the main cause of death. Outbred Kiyosu fish were not affected; we therefore assume that inbreeding led to a higher susceptibility to bacterial infections. In addition, 13 lines were lost due to extreme shifts in the sex ratio. We also observed unproductive crosses in 10 lines, and we terminated the lines when 4 alternative sibling crosses from the line were all unproductive. At the time of submission, the MIKK panel generation F18 comprised a total of 80 near-isogenic inbred lines (Column B, Additional file 1: Table S1).

Fecundity and morphological measures across the MIKK panel

We assessed the fecundity of the MIKK lines at generation F16. Over a period of 4 weeks, we monitored each line's egg production and assigned them to different classes of fecundity based on the approximate number of eggs produced per day. This assignment was facilitated by the fact that fertilized eggs remain attached to the females after mating. Under the constant summer conditions of the fish facility (14 h light/10 h dark), medaka mate every day [37]. This made it possible to semi-quantitatively measure fecundity by visually inspecting the number of females with fertilized eggs, and estimating the number of eggs per female that had mated. We observed a spectrum of mating frequencies and eggs per mating that ranged from occasional, irregular mating with few eggs per female, to daily mating with up to 10 eggs per female. For the majority of lines, we observed daily mating with 1–3 or more eggs per female (Fig. 1B and Additional file 2: Table S2). Thus, the overall fecundity of the majority of MIKK lines is sufficient to meet the demands for phenotyping quantitative traits.

Craniofacial variance in medaka has a genetic basis, and strain-specific variation of specific craniofacial traits has been shown [36, 38]. We chose relative eye diameter and relative distance between eyes as morphometric parameters to analyze whether MIKK lines show line-specific craniofacial variance. These parameters of the medaka head are conspicuous and therefore permit reliable quantification for the required statistical analysis of broad sense heritability. To quantify variance across the MIKK panel, we took high-resolution images of 77 lines for one male and two female fish aged between 6 and 9 months post-hatching, imaged in both lateral and dorsal orientations (https://www.ebi.ac.uk/birney-srv/medaka-ref-panel/). The resulting 462 images were used for quantitative phenotyping. We applied image segmentation using modern machine learning techniques (Methods) allowing the extraction of relative eye diameter (Fig. 1C) and relative distance between eyes of each fish. We then calculated heritability estimates across the MIKK panel. To exclude confounding factors such as differences in developmental stage, we only extracted information that would allow us to generate relative measurements within each fish. For most measurements, we used the overall number of pixels within the entire segmented body shape as the quantitative within-fish normalization value [39].

To measure the relative eye size across MIKK panel lines, we used the total number of pixels within the segmented eyes divided by the total length from nose to tail (in pixels), which produced a relative measure that corrects for the variance of body size. We performed one way analysis of variance (ANOVA) tests, calculating effect size (η^2) as a measure of broad sense heritability (H^2). The amount of variance explained by MIKK panel lines overall and thus broad sense heritability (H^2) is 0.68. We performed similar analyses for the relative distance between eyes and female abdominal size parameters, observing 0.51 and 0.39 levels of heritability for these traits respectively (Additional file 3: Table S3).

Overall, the between-line variance for these simple morphological measurements across the MIKK panel is high compared to the within-line variance, with broadly similar heritability estimates to human and mouse morphometric measurements [40, 41], suggesting that these traits are heritable in medaka and the MIKK panel captures phenotypic diversity present in the founding population as expected.

Homozygosity of the MIKK panel genomes

Medaka has a XX/XY sex determination system, similar to mammals [31, 42, 43]. The medaka male carries the heterogametic XY genotype. Therefore, to sequence both X and Y chromosomes, we extracted DNA from dissected whole male brains as donor tissue for whole-genome sequencing across the panel. The male donor medakas were 6 months old. The per-sample coverage was between 21 and 34×. As expected, over 97.3% of the reads aligned to the *HdrR* reference genome, and we called single-nucleotide polymorphisms (SNPs) to this reference using a standard pipeline (Methods). We calculated the level of homozygosity across the medaka genome for each MIKK panel line using the number of heterozygous SNP calls across the genome with a total of 6.3% of all SNP genotypes across all panel lines (excluding genotypes called as missing) being called as heterozygous. The medaka sex chromosome, LG1, was excluded in this heterozygosity estimation since heterogametic XY males were used for whole-genome sequencing. However, as expected, this heterozygosity is not distributed evenly across the genome in each line. To explore this, we split the genome up into 10 kilobase (kb) non-overlapping windows and counted the number of heterozygous SNPs in each window. We then trained a two-state Hidden Markov Model (HMM) on all 10-kb windows from all MIKK panel lines jointly, using the expectation maximization (EM) algorithm to separate out true homozygous regions from regions of residual heterozygosity (Methods). We then used the Viterbi algorithm to find the most likely sequence of hidden states (Viterbi path) for each panel strain separately. We performed a similar procedure on the classic inbred strains (*iCab* and *HO5*) and wild-caught medaka from the same founding population as the MIKK panel [36]. Below are summary plots showing the proportion of called homozygous regions comparing the MIKK panel to classical inbred strains and wild Kiyosu medaka.

The genome-wide homozygosity observed in the MIKK panel lines is similar to that of the classical laboratory inbred strains (*iCab* and *HO5*) which have been inbred for more than 70 generations. Both classical inbred strains and the MIKK panel show a clear difference in the number of homozygous regions genome-wide relative to wild Kiyosu samples (Fig. 2A). Over 70% of MIKK panel lines are

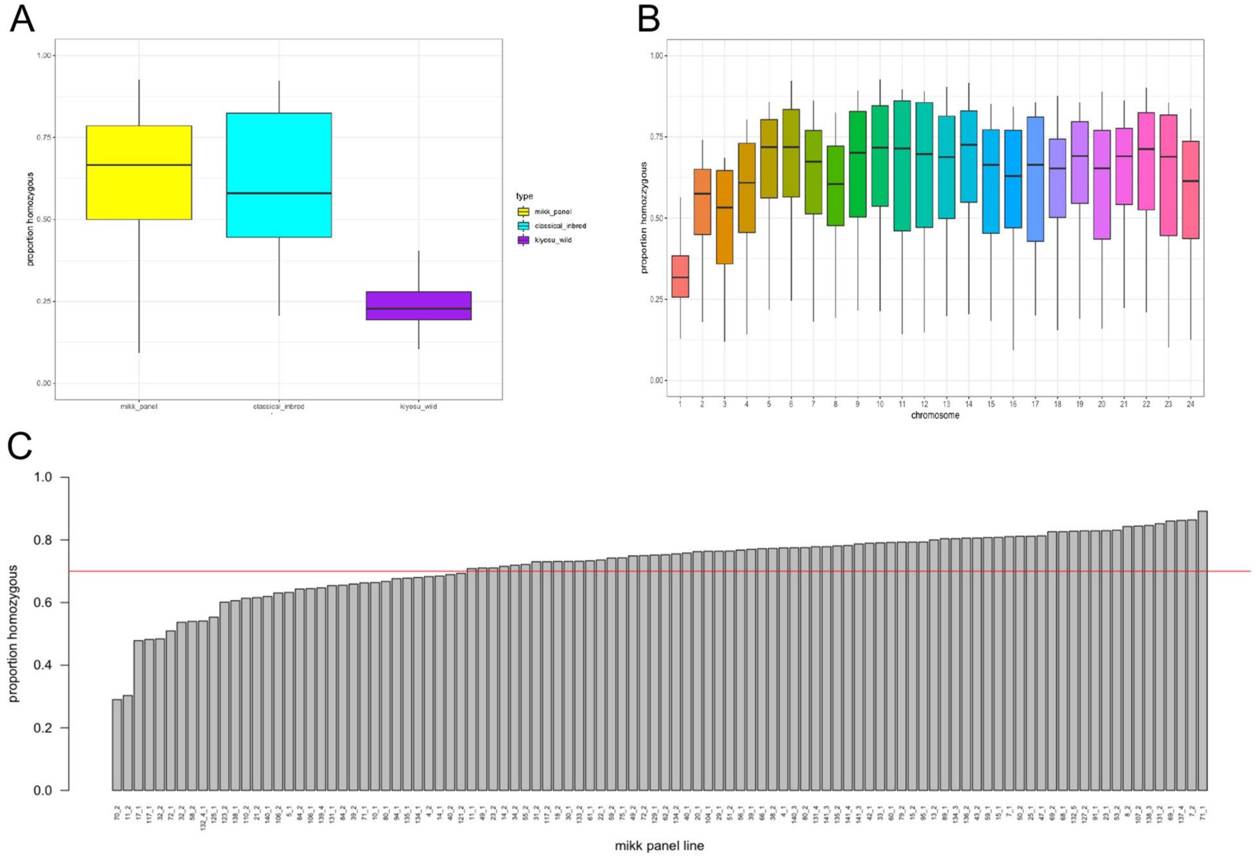

Fig. 2 Homozygosity of the MIKK panel. **A** The overall proportion of homozygous 10 kb genomic regions comparing the MIKK panel to classical inbred and wild Kiyosu fish. **B** The proportion of homozygous 10-kb genomic windows across the MIKK panel for each chromosome. **C** The proportion of homozygous 10-kb genomic windows (excluding chr1) for all MIKK panel lines

homozygous for greater than 70% of their entire genome and, as expected, when looking at the level of homozgosity across chromosomes, we observe a clear decrease on chromosome 1 across all MIKK panel lines (Fig. 2B), reflecting that the sex determination region must nearly always remain heterozygous in male medaka [43]. In addition, chromosomes 2 and 3 showed higher levels of heterozygosity compared to the other chromosomes, which in the case of chromosome 2 may be due to many large structural variants including inversions. However, in the case of chromosome 3, it is unclear what is causing this increased heterozygosity although we did detect two relatively small inversions at specific regions on chromosome 3 in some MIKK panel lines [44]. At the time of sequencing, all MIKK panel lines had undergone 9 generations of single full-sibling-pair inbreeding, and although the majority of lines have highly homozygous genomes, there is a subset of lines that show a considerably lower degree of homozygosity (i.e., 5 lines that show < 50% homozygosity; Fig. 2C). We hypothesize that the lower degree of homozygosity in this subset of lines is due to an accidental outcross with distantly related MIKK fish during the inbreeding process, leading to a significant drop in the homozygosity. To test our hypothesis, we are currently carrying out additional inbreeding crosses with these lines with the aim to test whether these crosses will increase the lower homozygosity.

As an additional means of assessing genetic diversity in the MIKK panel, we calculated nucleotide diversity $\left(\hat{\pi}\right)$ 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 $\hat{\pi}$: MIKK = 0.0038, wild = 0.0037; median $\hat{\pi}$: 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 ($r = 0.386$, $p < 0.001$) (Additional file 4: Fig. S1). We also calculated $\hat{\pi}$ for each line individually—levels of $\hat{\pi}$ around the sex determination region of 1:~16–17 Mb (see below) are elevated in all lines relative to the consistently low levels found in most other chromosomes, such as chromosome 22 (Additional file 4: Fig. S2).

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 (Additional file 4: Fig. S3). These regions are also enriched for large deletions and duplications [44]. Inversions cause permanent heterozygosity [45], and duplications and deletions may have increased the density of called SNPs in these regions [46], 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, the MIKK panel shows similar levels of homozygosity compared to classical laboratory inbred medaka strains and contains a strong increase in isogenic genotypes compared to wild medaka from the original wild population.

Enhanced definition of the medaka sex determination region

To further investigate the homozygosity levels on chromosome 1, we used a 4-state HMM trained on the number of heterozygous SNPs in 10-kb windows. This created a finer-grained view across all MIKK panel lines, allowing us to estimate the critical region containing important genes necessary for sex determination in medaka [42].

The sequenced male XY regions on chromosome 1 that are maintained in their heterozygous state across MIKK panel lines refine the critical region for sex determination in medaka (Fig. 3A). We determined the critical region using two different approaches. First, for all 10-kb windows, we calculated the proportion of lines of the MIKK panel that had greater than 5 heterozygous SNPs (fitted blue line in Fig. 3B). Second, we used the HMM states to create a weighted estimate where states with higher heterozygosity levels contributed more weight in the model (fitted red line in Fig. 3B). Overall, when using the SNP count-based estimate, we observe a large region across the middle of chromosome 1 that remains heterozygous in most MIKK panel lines. However, when incorporating further information from the HMM, we detected a 0.5 Megabase (Mb) region (chr1:15959156-16459156) that showed considerably higher levels of heterozygosity in most MIKK panel lines. This region includes 17 genes in the *HdrR* reference genome and could be of interest for further mapping studies to investigate the precise location of the duplicated copy of the *DMRT1* gene known to be critical in sex determination [47], and how the suppression of recombination creates hetereogametic sex determination. It has been suggested that this 0.5 Mb region with higher levels of heterozygosity corresponds to an early stage of Y-chromosome evolution [31] which could eventually result in a non-recombining mature sex chromosome [30]. We note that a small number of lines were homozygous across this region, consistent with occasional XX male sex determination seen in medaka (see "Discussion").

Potential functional impact of small genomic variation

We discovered a total of 3,001,493 variants (SNPs and INDELs) across the MIKK panel compared to the *HdrR* reference, of which more than 70% (2,248,228) were either synonymous or intergenic (Additional file 5: Table S4). We examined the potential functional impact of variation in protein-coding genes across the panel. There were a total of 644,509 non-synonymous variants, 36,444 splice site variants, and 82,312 potential Loss-of-Function (LoF) variants (stop codons or frameshifts). Similar to human studies [48], apparent LoF variation is enriched in regions where the genome assembly or gene predictions were poor, so we used a variety of computational screens (Methods) to select 35,154 high-confidence LoF variants in the MIKK panel (Additional file 4: Fig. S5).

As expected, these high-confidence LoF variants were more likely to be found in a heterozygous state (15% of the LoF SNPs were present in a heterozygous-called block compared to 9% on average), consistent with successful homozygosity being prevented in some cases due to a deleterious LoF allele. However, 63% had confident homozygous calls, with 34% of LoF SNPs showing no evidence of heterozygosity in any of the individual lines. Unsurprisingly, we see the expected increase in the number of LoF variants in heterozygous blocks for chromosome 1 due the sex determination region, and interestingly, both chromosomes 2 and 18 show consistently higher

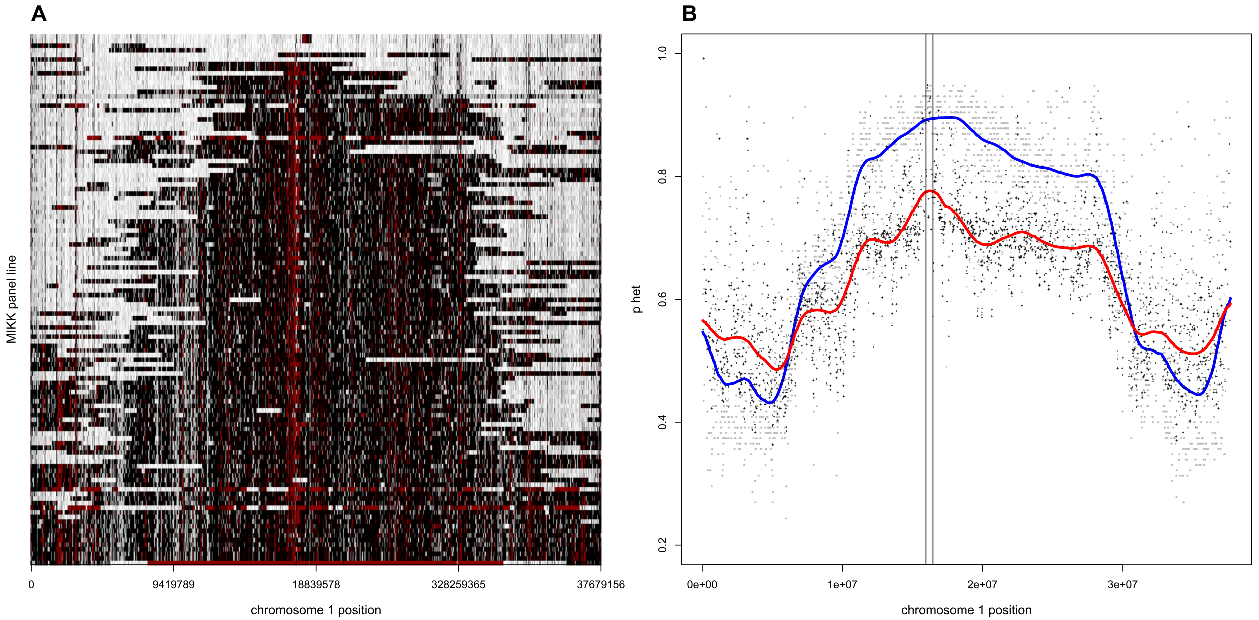

Fig. 3 Heterozygosity across chromosome 1 in the MIKK panel defines the sex-determining region. **A** Homozygosity state from the HMM for all 10-kb windows across all MIKK panel lines (the color ordering white, gray, black, and red represent increasing levels of heterozygosity). **B** Two different estimates of the proportion of MIKK panel lines that are classified as heterozygous across all 10-kb windows on chromosome 1 (blue: based on heterozygous SNP count; red: based on HMM state space)

numbers of LoF variation for all LoF variant classes (Additional file 4: Fig. 5A-D) which is likely a result of larger numbers of structural variation observed within these chromosomes [44].

Previous work in humans has characterized genes for their intolerance to mutations using the probability of LoF intolerance (pLI scores) which has been widely adopted in human studies [49]. For most LoF variant classes, there is a slight enrichment towards higher pLI scores of the orthologous human genes (Methods) inside heterozygous blocks compared to homozygous blocks (Additional file 4: Fig. S5E). This high-confidence, homozygous set of LoF variants were less present in medaka genes orthologous to human genes under stringent purifying selection, or disease-causing genes in humans (Additional file 4: Fig. S5E,F). There were 1441 cases with a homozygous high-confidence LoF in the MIKK panel orthologous to a human dosage-sensitive gene (pLi > 0.5), listed in Additional file 6: Table S5. These variants were found across 614 medaka genes with 38% (235) being annotated with the molecular function “binding” (GO:0005488) and 64% (393) with the biological process “cellular process” (GO:0009987) during gene ontology (GO) analysis using PANTHER [50]. For pathway focussed GO analysis at a false discovery rate (FDR) of 1%, there were zero significant pathways; however, the top two were the Wnt signaling pathway (P00057) at 2.9% (18 genes) and the Gonadotropin-releasing hormone receptor pathway (P06664) at 2.4% (15 genes). These individual variants, and the lines that carry them, may be of direct interest to researchers focused on these critical human genes.

Identity-by-descent (IBD) mapping

To assess the population structure of the MIKK panel, we performed an identity-by-descent (IBD) analysis across the panel, including both classical inbred strains (*iCab* and *HO5*) and MIKK panel lines. As in previous studies, overall IBD is tracked via the identity-by-state (IBS) of SNPs. We first sought to assess the overall relationship between the classical inbred strains and the MIKK panel using IBS measures (Fig. 4A). We observe a clear separation between the classical inbred strains and MIKK lines, reflecting comparably low recent shared ancestry.

Next we compared IBS measures for all MIKK panel lines that had a “sibling line” available (as described above). Therefore, these sibling lines can be expected to share 25% of the genome from their parents. At the time of sequencing, there were 44 MIKK panel lines with a remaining productive sibling line (22 sibling-line pairs) (Column C, Additional file 1: Table S1). When we cluster sibling-line pairs based on IBD (Fig. 4B), 77% (17 pairs) are directly adjacent, showing a higher degree of relatedness within the sibling pair than without. Two sibling-line pairs cluster as a quartet, suggesting that the founding Kiyosu individuals for these two pairs may have been closely related. As expected, we find that the majority of sibling-line pairs show a high degree of relatedness, which not only provides interesting population structure across the panel, but also shows that the careful and detailed tracking of the MIKK panel lines throughout the inbreeding process has been robust. The three “orphan” sibling-line pairs which do not cluster directly adjacent to each other may well have arisen from random segregation in the inbreeding producing more divergent sib lines.

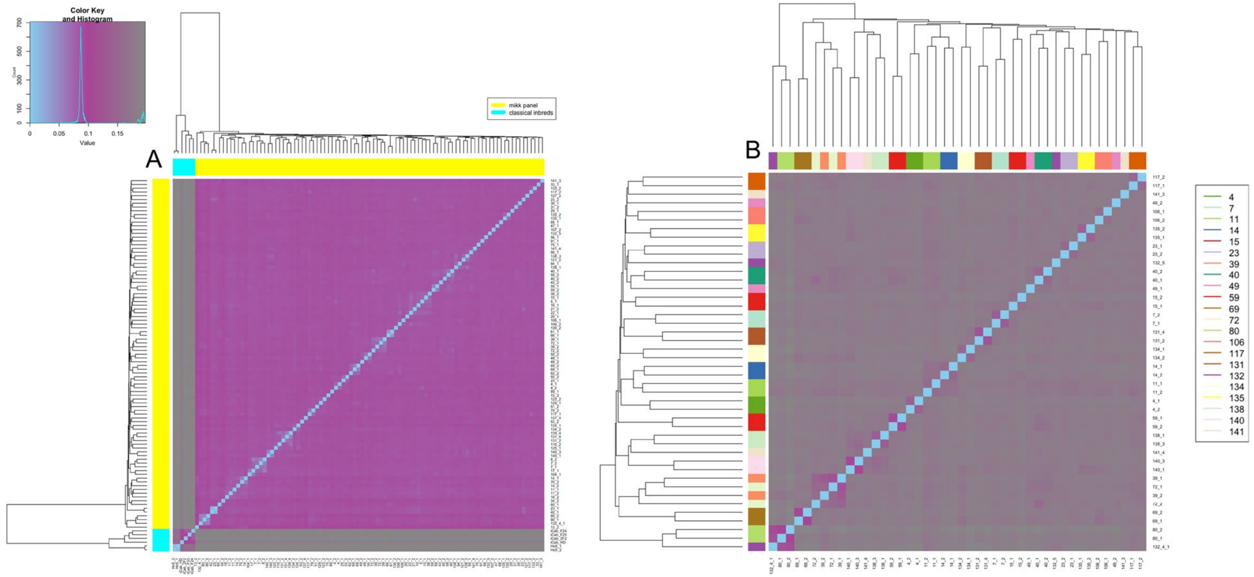

Fig. 4 Identity by state (IBS) across the MIKK panel and classical inbred strains. **A** 1-IBS from all medaka samples, including classical inbred strains. **B** 1-IBS for all MIKK panel lines that have an available "sibling line"

Genetic composition of the MIKK panel

The MIKK panel was created from a single southern Japanese population of wild medaka from Kiyosu, Japan. We sought to determine whether there had been any large-scale selection or allele skew on panel genotypes from the original wild founding population during the inbreeding process. To search for allele skew, we used the parents from eight medaka trio datasets collected from the same founding population [36] to measure population differentiation due to genetic structure with the fixation index (F_{ST}). F_{ST} is a well-established parameter of population differentiation [51]. It provides a measure of genetic structure based on the variance of allele frequencies or allelic skew between populations [52]; high F_{ST} indicates that allele distributions between populations are divergent, whereas low F_{ST} indicates they are similar.

We calculated F_{ST} across all shared genotype positions between 16 wild medaka sequences [36] and the MIKK panel. To look at F_{ST} genome-wide, we defined 10-kb non-overlapping windows and took the mean F_{ST} for each window across the medaka genome (Fig. 5A). Over 93% of 10-kb windows had greater than 50 SNPs available to include when calculating the mean F_{ST}, with a median of 115 sites per window (Fig. 5B). Overall, we see very little allele skew genome-wide, with no large regions showing high F_{ST} values. When looking at the 10-kb windows that display higher F_{ST} values, we see that the majority were based on windows with low numbers of available SNPs (Fig. 5C) and when excluding sites with less than 50 SNPs, the number of regions showing high F_{ST} values is reduced substantially (Additional file 4: Fig. S6). Reassuringly, we did not observe any significant differences in the total genetic variance genome-wide between the MIKK panel and wild medaka samples from the same founding population. This shows that the genetic diversity within the MIKK panel is a reasonable representation of the genetic diversity found in wild medaka from the Kiyosu region.

Introgression with northern Japanese and Korean medaka populations

To determine whether the MIKK panel's founder population showed signs of introgression with northern Japanese and Korean medaka populations, we ran an ABBA-BABA analysis [53–55]. We used the 50-fish multiple alignment in Ensembl's release 102 [56] to obtain the aligned genome sequences of the established inbred medaka strains *HdrR* (southern Japan), *HNI* (northern Japan), and *HSOK* (Korea), in order to orientate each SNP to its ancestral state (Fig. 6A). We then combined this data with the Illumina-based SNP calls for the MIKK panel and the established inbred medaka strain *iCab* (southern Japan). We used this combined dataset to calculate $\hat{f}_d$ [55], a modified version of the "admixture proportion" statistic $\hat{f}$ [53], to measure the proportion of shared genome in 500-kb sliding windows between the MIKK panel and either *iCab*, *HNI*, or *HSOK*. For this purpose, we used *HdrR* as the MIKK panel's most closely related P1 population, the MIKK panel as the P2 focal population, either *iCab*, *HNI*, or *HSOK* as the P3 introgressing population, and the most recent common ancestor as the outgroup (O) (Fig. 6B). $\hat{f}_d$ measures across the medaka genome are presented in Fig. 6C. 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

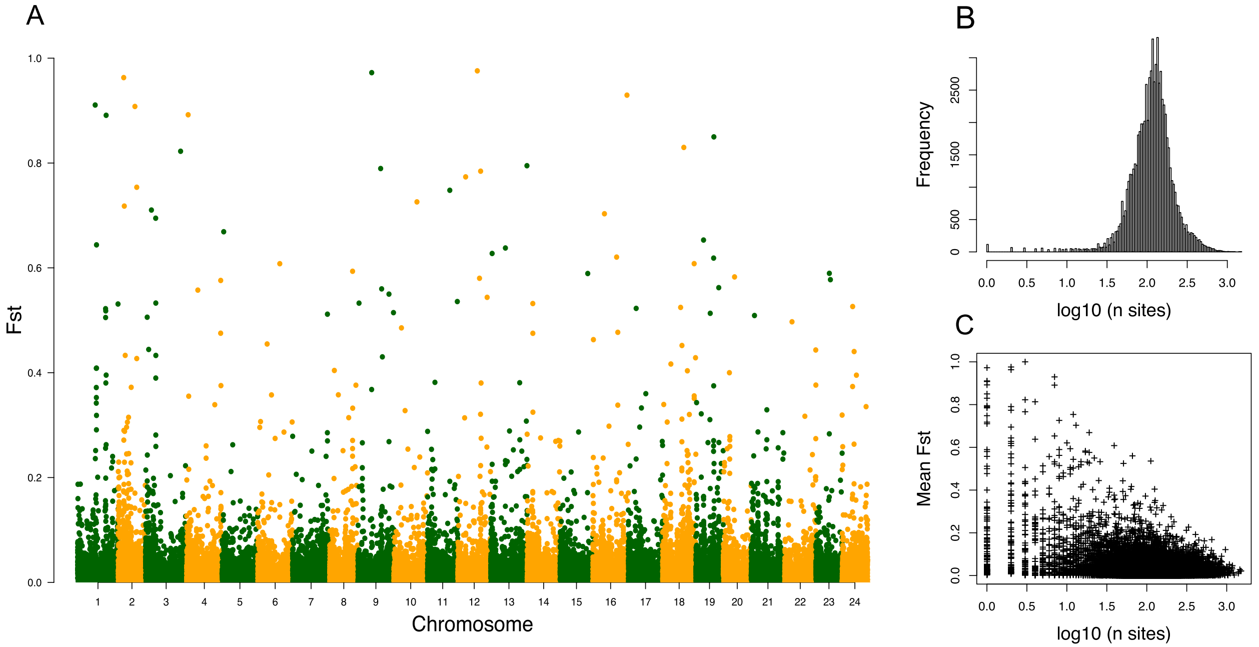

Fig. 5 Fixation Index (F_{ST}) comparing allele frequencies of the MIKK panel against wild Kiyosu medaka. **A** Mean F_{ST} in 10-kb windows across the medaka genome. **B** $\log_{10}$ of the number of SNPs per 10-kb window. **C** $\log_{10}$ of number of SNPs per window vs mean F_{ST}

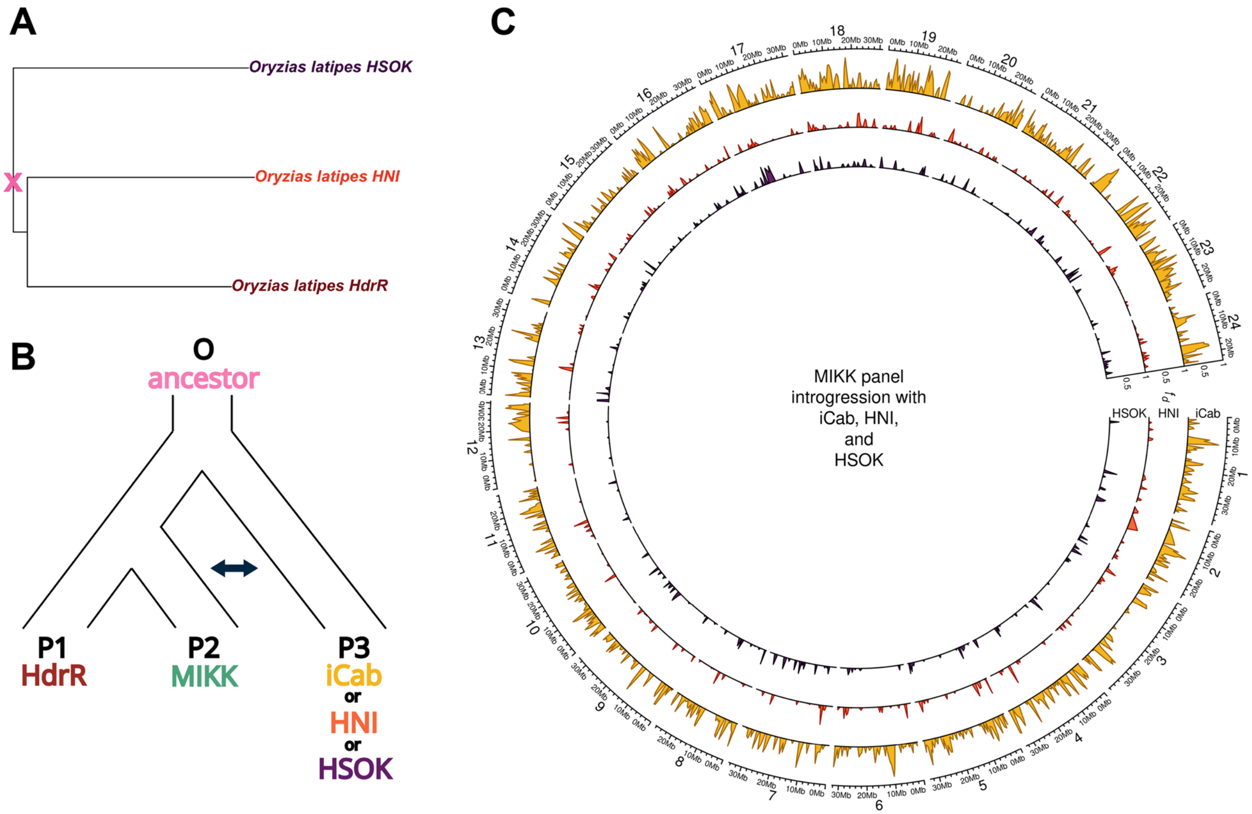

Fig. 6 ABBA-BABA analysis. **A** Phylogenetic tree generated from the Ensembl release 102 50-fish multiple alignment, showing only the medaka lines used in the ABBA-BABA analysis. **B** Schema of the comparisons carried out in the ABBA-BABA analysis. **C** Circos plot comparing introgression ($\hat{f}_d$) 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

with medaka from northern Japan. This supports the findings in [36], where the authors found little evidence of significant interbreeding between southern and northern Japanese medaka since the populations diverged.

LD decay

We analyzed 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. Figure 7A 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 ($N = 2504$) (1KG) in human [57]. As expected, the 1KG and MIKK panel

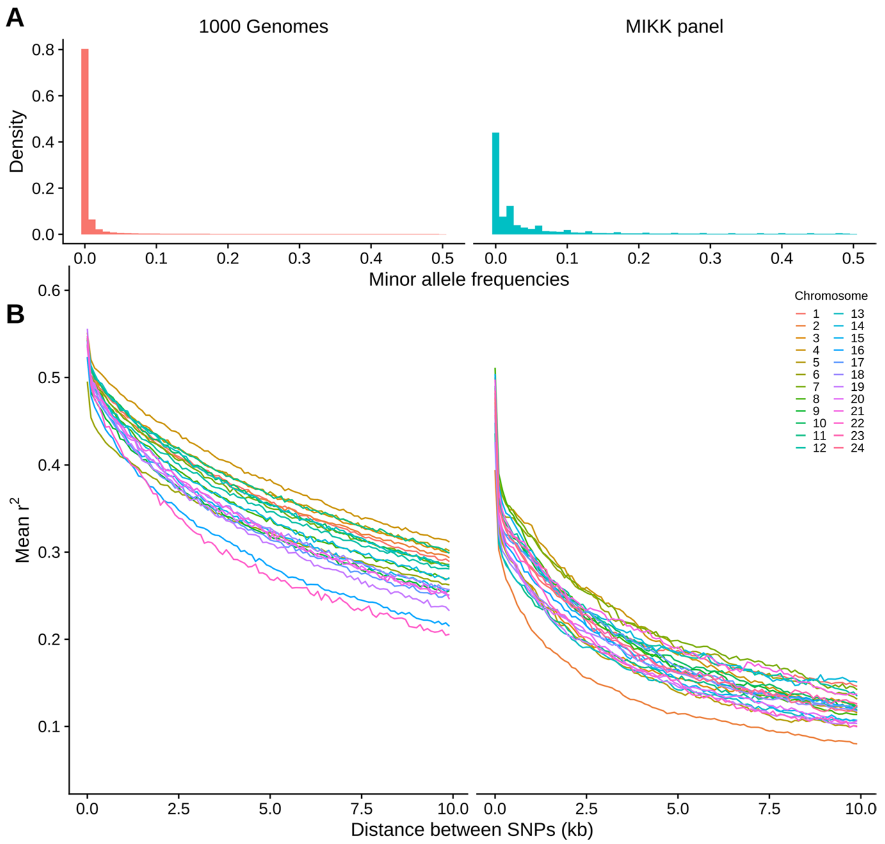

Fig. 7 Minor allele frequency distributions and LD decay for biallelic, non-missing SNPs in the 1000 Genomes Phase 3 variant calls ($N = 2504$) (1KG), and the MIKK panel Illumina-based calls excluding one of each pair of sibling lines ($N = 63$), across all autosomes (1KG: chrs 1-22; MIKK: chrs 1-24). **A** Density 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. Inset: mean r^2 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 [58]

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 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 r^2 between all pairs of biallelic SNPs with a minor allele frequency (MAF) greater than 0.10, within 10 kb of one another (Methods). We then grouped the paired SNPs by distance from one another into non-overlapping 100-bp windows and calculated the mean r^2 for each window in order to represent how LD decays with distance between loci. Figure 7B 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 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 [58] 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 (Additional file 4: Fig. S4).

Expression QTL analysis

To investigate the genetic diversity of the panel at the transcriptome level, we performed a proof of concept expression-based quantitative trait analysis (eQTL) using liver tissue from 12-month-old females of 50 MIKK panel lines that were kept under 14/10 LD summer conditions (Column G, Additional file 1: Table S1). We found 14,835 significant eQTL SNPs (*1% FDR*) in 3795 transcripts corresponding to 3543 genes (Fig. 8A, Additional file 7: Table S6 and Additional file 8: Fig. S9) and observed a highly consistent pattern of expression between all samples. Interestingly, most of the sibling lines included in the RNA analysis clustered closely together, suggesting that broad-scale expression differences, as a direct consequence of the genomic sequence in medaka, are heritable across multiple generations (Additional file 4: Fig. S7), and suggestive of broader trans-effects in expression. To further explore the genotypic pattern associated with expression profiles, we looked in detail at a stretch of 3 SNPs associated with an eQTL in an isoform of *cyp17a2* (ENSORLT00000002786.2)—a gene likely involved in progesterone metabolism (Fig. 8B,C). For all 3 SNPs, there is a strong correlation between the level of transcript expression and the genotype, with T genotype being associated with a much higher median expression, while A genotype has little to no expression.

Finally, we generated a metagene plot showing an aggregated distribution of the location of all significant eQTLs, showing a slight enrichment immediately downstream and upstream of genes and an enrichment peak inside gene bodies shortly after the transcription start site (TSS) (Fig. 8D). This is consistent with variation in and around the TSS being likely to impact gene expression. Although noisy due to the small number of associations from this limited sample size, the pattern is similar to the previously reported eQTL metagene plot in vertebrates [60]. Altogether, we showed that despite the small number of samples, we are able to identify transcripts associated with eQTLs

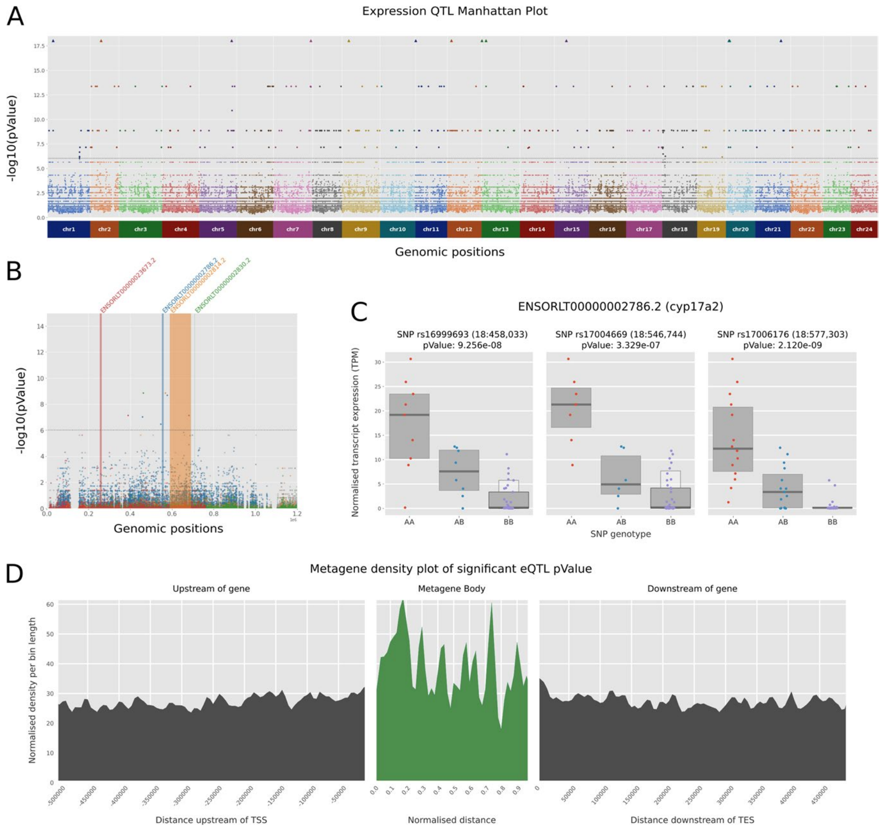

Fig. 8 eQTL analysis results. **A** Manhattan plot reporting only the 5 best *p*-values per transcript. For visual comfort, the *p*-value threshold (dashed line) was set to 1e−6. **B** Zoomed-in slice of the Manhattan plot, showing all *p*-values in the genomic window chr18:0-1200000. Transcripts associated with at least 1 *p*-value lower than 10e−6 are highlighted in color, with all SNPs associated shown in the same color. The width of the colored bands corresponds to the full length of the transcript. **C** Normalized TPM expression of MIKK lines for 3 significant eQTL SNPs associated with transcript ENSORLT00000002786.2 (cyp17a2 gene). Lines are sorted according to their genotypes (AA, AB, or BB). Each point represents one individual line, overlaid with letter-value plot boxes showing the quantile distribution [59]. **D** Metagene density plot with upstream and downstream 500k regions showing the density of significant eQTL hits according to their positions relative to a prototypic gene. Values are aggregated in 100 bins of 5000 bases for the adjacent regions and 50 bins dynamically computed for *p*-values falling inside gene bodies. Densities are weighted by bin length to allow comparisons across the different regions

in the MIKK panel with similar properties to human and other vertebrate eQTL studies [61, 62].

Discussion

We have established a panel of 80 near-isogenic inbred lines of medaka from a carefully selected wild population without prior genetic domestication of founder animals. In addition to the number of inbred lines, the MIKK panel is unique among vertebrate inbred panels as it has been derived from a wild population. It is well known that the domestication of wild animals often has severe consequences for the genetic diversity of a population [23, 63]. In the MIKK panel, genetic adaptation to husbandry conditions prior to inbreeding did not occur. This design allowed us to preserve much of the naturally occurring genetic variation in its wild founding population. We speculate that the tolerance to inbreeding from the wild is in part due to an ability of wild medaka to adapt to facility husbandry conditions. Consistent with this, medaka has a long history as a pet animal in Japan, demonstrating the natural propensity of this species to tolerate and reproduce in captivity [24]. The MIKK panel is therefore a unique tool to examine phenotype-genotype associations in a natural vertebrate population under precisely controlled laboratory conditions. Our estimates of fecundity levels indicate that egg production is sufficient to meet the demands for the phenotyping of complex traits that often show quantitative variation, and therefore require the measurements of large numbers of individuals. In the future, carefully controlled quantification of fecundity levels will address whether the observed variation between lines has a genetic component.

We confirmed that the panel has captured phenotypic variability by exploring two phenotypes. First, for a morphological trait by image analysis, we observed a broad sense heritability (H^2) measure of 0.68 for eye size as an extractable trait. Second, we performed RNA-Seq on female liver tissues. We detected 622 significant cis expression quantitative trait loci (cis-eQTLs) in 541 protein-coding genes, providing a proof of concept as the basis for future extended analyses. The location of cis-eQTLs with respect to gene structure shows the expected enrichment towards transcriptional start sites, although less pronounced than in previous studies involving larger sample sizes [61, 64]. In summary, this indicates that genetically determined phenotypic variation is well represented in the MIKK panel.

High-depth Illumina sequencing revealed a high degree of homozygosity across all panel genomes. We see a striking reduction in genome-wide heterozygosity with isogenic levels similar to that of classical inbred strains that have been inbred for more than 70 generations. The distribution of homozygosity is largely consistent across the genome, with the exception of the sex-determining locus on chromosome 1 and specific regions on several other chromosomes that are likely to have been caused by the presence of certain structural variations. As expected, we detected the expected drop in homozygosity around the sex-determining locus. Interestingly, 3 male fish showed no heterozygosity across this region in the X chromosome. This is consistent with the known XX males observed in medaka fish [65] and is also consistent with the recent evolution of genetic sex determination in *Oryzias latipes.* We therefore speculate that phenotypic males of XX genomic configuration were sequenced in these three cases. These lines continue to produce balanced ratios of males and females at the time of writing (generation F18) and will be interesting for studying aspects of the medaka sex determination system in the future.

The MIKK panel originates from a single population of wild medaka caught in Kiyosu, Japan. When comparing the genome-wide allele frequency distribution between the inbred lines and wild Kiyosu medaka, we observed similar allele frequencies and no large regions with distortions. This suggests that the genetic variation in the MIKK panel is a good representation of wild genetic variation, making the panel an excellent resource for genome-wide association studies (GWAS) using natural standing variation. To assess the resolution for genetic association mapping, we calculated the linkage disequilibrium (LD) across MIKK panel lines. We compared the LD decay to that observed in human cohorts (1000 Genomes Project Phase 3) and observed a marked increase in LD decay, indicating that when a causal variant is present in more than one MIKK panel line, the mapping resolution may be higher than in humans. We note the faster decay of LD in chromosome 2, which is consistent with the known high recombination rate seen across this chromosome relative to its physical distance (3.36 cM/Mb in chromosome 2 vs 1.02-2.76 cM/Mb for the other chromosomes) [58], possibly attributable to that chromosome's high repeat content (Additional file 4: Fig. S4A). There are potential analogies with the high recombination frequency present across chromosome 19 in humans [66].

One interesting aspect of the panel design is the fact that some lines have a "sibling" line (originating from the same founder F1 cross). This special genetic structure across the panel can be exploited for specific research applications, for example to explore the aggregate impact of epistatic (GxG) interactions without locus-level resolution. This genetic structure also acted as a definitive control for us during this analysis. When we look at measures of relatedness across the panel, the IBD analysis showed that "sibling" lines cluster closely together, indicating that they are indeed more related to each than other panel lines. This finding not only shows that the panel displays the favorable genetic structure that was designed from the outset, but also that all the crossing procedures during the inbreeding process were robust.

To place the MIKK panel in context of medaka population genetics around Japan, we performed ABBA-BABA analysis to compare the level of admixture between the MIKK panel and southern Japanese (*iCab*), northern Japanese (*HNI*), and Korean (*HSOK*) medaka strains. As expected, the MIKK panel shows the most admixture with the southern strain *iCab*, as Kiyosu is in southern Japan. However, there is a slight increase in admixture proportion for *HSOK* compared to *HNI*. Although the proportional difference is small, this 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 [67].

The majority of MIKK panel genomes are highly isogenic, and as expected, we observe low numbers of predicted loss-of-function (LoF) variants (< 0.3% of all variants). Small homozygous LoF variants (SNPs and INDELs) show a slight enrichment in regions of residual heterozygosity in the MIKK panel lines and on chromosomes that show higher levels of larger structural variation overall. This may indicate a purging of LoF variation in essential dosage-sensitive genes, although the observed enrichment is slight, which is likely due to the low chance of LoF variation occurring in essential genes across the 9 generations of inbreeding, given the very low rate of spontaneous mutations. There are, however, many rare genetic variants that occur in genes that are predicted to have a functional impact or an intolerance to dosage. This is, of course, not unexpected but it does highlight one of the powerful applications of genetic screening using the MIKK panel, where specific F2-cross designs can be used to balance the distribution of rare impactful variants across an F2 population, allowing a functional analysis of rare variants during trait association analyses. Exploration of the phenotypes of these cases, or mapping of potential modifier mutations that might compensate for such loss of function variation, will be a valuable line of research for both the medaka and human disease communities.

Within the vertebrate clade, teleosts including medaka diverged from the tetrapod lineage about 400 million years ago [68, 69]. Due to this divergence, orthology relationships between medaka and mammalian species may be obscured for certain genes.

Furthermore, an additional genome duplication in teleosts imposes further caveats for gene comparisons between teleosts and mammals [70]. Thus, for studies using the MIKK panel for comparison and evaluation of mammalian GWAS data, this needs to be taken into consideration. Direct comparisons between association studies with the MIKK panel and mammals need to consider divergence of genes as well as functional divergence of orthologous genes. Recent studies provided important tools for such comparisons, showing that genome data of basal fish such as the spotted gar can help to bridge mammal-teleost gene comparisons [71]. Thus, even though medaka is highly diverged from human and mouse, it will be a relevant model to support GWAS from these species. We expect that comparisons of Medaka derived phenotypes to either other teleost species as well as more distant vertebrates such as mammals will be useful both for understanding fundamental biology and exploring the function of orthologous genes between these species.

It is important to note that the value of medaka and the MIKK panel is not only the biomedical relevance to model human disease and complex genetic traits. This economical teleost combines several advantages that render it an important model for basic research and thus an important and valuable addition to existing vertebrate and invertebrate population genetic models for basic research. The MIKK panel of inbred lines was established directly from a wild population without prior domestication and thus represents the genetic variation of the original population in a reproducible form. Furthermore, the possibility to sample from the wild and validate experimental data obtained with the MIKK inbred panel provides options to study population genetics that are unique for a vertebrate laboratory model. The MIKK inbred panel will prove to be an important additional model to establish tools for association studies and validation including functional testing of genetic variants

Conclusions

Here we provide a detailed analysis of some of the sources of variation present across the MIKK panel lines, and put in place important resources to allow the MIKK panel to be fully exploited during further research studies. Further sequencing of the MIKK panel lines, using both long- and short-read technologies, is planned in the future as well as extensive molecular, organ function and organismal phenotyping across the panel. In a separate analysis [44], we looked at building line-specific genome reference datasets from a subset of the MIKK panel and show that by using non-linear reference alignment approaches (graph genomes), we are able to build a more complete representation of genetic variation in the MIKK panel, uncovering additional variation that is masked when using standard reference alignment approaches. Further detailed characterization of the MIKK panel in terms of genetic variation as well as molecular, organ, and organismal phenotyping will provide a wealth of information allowing us to leverage the isogenic properties of the panel to investigate phenotype-to-genotype interactions (including gene-to-environment effects, or GxE) at high genetic resolution. We invite the medaka, teleost, and broader vertebrate genetics community to make use of the resources presented here, and to contact the authors to explore the hosting of experiments across the MIKK panel in the style of a “research hotel.”

Methods

MIKK husbandry and inbreeding

An unstructured polymorphic wild population of medaka was sampled in the irrigation canals of Kiyosu near Toyohashi as described [36]. In total, 141 random single pair mating crosses of wild Kiyosu fish were set up at NIBB, Okazaki, and their F1 offspring shipped to the medaka facility of KIT. Of these 141 crosses, the surviving 115 founder F1 families were raised to adulthood and used for 9 generations of single full-sibling-pair inbreeding to establish the panel of near-isogenic lines now known as the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel.

Full-sibling-pair inbreeding crosses [72] were set up as follows: approximately 1 month after the fishes (siblings) of a line started productive mating, random single brother-sister crosses were set up in 2-l mating cages with water recirculation. In crosses resulting in only unfertilized eggs, the male was exchanged with a male of the same strain (i.e., brother). Eggs were collected from females after 4 days of successful mating. Unproductive crosses were aborted after 2 weeks and a new single brother-sister (full-sibling-pair inbreeding) cross of the same line was set up. Eggs were collected from a given mating cross until at least 50 hatchlings were obtained that survived for 2 weeks.

MIKK panel fish at the medaka facility of KIT were kept as described in [73], in Müller & Pfleger systems (40 6-l tanks per recirculation system; Müller & Pfleger, Rockenhausen; Germany) with the following modifications: eggs were raised at 22 °C. Under the constant summer conditions (14 h light/10 h dark: 14:10LD), medaka mate every day, mostly at the onset of the light period [37]. A *flavobacterium columnare* infection that occurred while inbreeding generations 4-6 was treated with Baytril (Baytril Enrofloxacine 10% injection solution, Bayer; 2 ml/100 l system water for 7 days followed by a 50% system water exchange). Thereafter, the inbreeding lines were kept at 23 °C to reduce the microbial load in the system water. The notebook for generating Fig. 1 can be found here: https://birneylab.github.io/MIKK_genome_main_paper/01_Fecundity.html.

Imaging of adults and image analysis

The 6- to 9-month-old adults were sacrificed by hypothermic shock. Lateral and dorsal images were taken with a Canon EOS 250D camera and a Sigma 18-250mm lens. Using a subset of the MIKK panel images, a number of machine learning models were trained to detect and measure phenotypes from the images of the entire panel of fish. The code for this was developed in Python using an existing implementation of semantic image segmentation in the Keras (https://github.com/divamgupta/image-segmentation-keras) machine learning framework with Tensorflow [74] used as the backend.

The network architecture used in these models was a semantic segmentation network known as SegNet. This is a deep fully convolutional neural network architecture for semantic pixel-wise segmentation. The SegNet architecture is based on a modification of the VGG16 network proposed which makes the network more lightweight and quicker to train. This architecture was chosen primarily as it has been shown to improve boundary detection and reduce the number of parameters that need optimisation, allowing for easier end-to-end network training. Additionally, this architecture has been shown to achieve high training accuracy when training with small datasets.

A subset of the panel images were segmented using the Labelbox platform [75] to generate the training data to train models to segment the fish body, eye, tail, and anal fin within the images. During training, the training dataset was augmented using the Python library imgaug [76]. The following affine transformation where randomly applied to the images during training:

- Scale images to 80–120% of their size,
- Translate images by − 20 to +20 relative to height/width (per axis),
- Rotate images by − 45 to + 45°,
- Shear images by − 16 to +16°.

After training the models, they were used to predict the chosen features in the remaining unseen images and the following parameters were extracted from these segmentations.

- Nose to tail length in pixels,
- Maximum width of fish in pixels,
- Fish area in pixels,
- Eye area in pixels,
- Length of abdominal region in pixels,
- Length of caudal region in pixels,
- Ratio between lengths of the abdominal and caudal region,
- Eye diameter in pixels,
- Distance from the top of the eye to the top of the head,
- Distance from the bottom of the eye to the bottom of the head.

The notebook for generating Fig. 1 can be found here: https://birneylab.github.io/MIKK_genome_main_paper/01_Fecundity.html (Fitzgerald et al.).

Tissue dissection

For whole-genome sequencing, brains were dissected from 6-month-old male medaka adults. Fish were sacrificed by hypothermic shock. The brain was dissected and shock frozen in liquid nitrogen. For RNA-Seq analysis, 12-month-old adults that were kept at either 14 light:10 dark (summer condition) or 10 light:14 dark (winter condition) light cycles respectively were sacrificed by hypothermic shock and the organs after dissection were shock frozen in liquid nitrogen. Fish were acclimatized for at least 1 month to the 10 light:14 dark (winter condition) prior to tissue dissection.

DNA and RNA extraction

DNA extraction and whole-genome sequencing of the MIKK panel

DNA was extracted from medaka brains in 2-ml Eppendorf tubes using the Qiasymphony DSP DNA Mini kit (Cat. No. 937236). Tissue_HC_200_v7 Protocol, 100 μl elution volume. Pre-treatment samples were homogenized with 5-mm stainless steel beads in 220 μl buffer ATL at 30 Hz for 20 s on the TissueLyser. Then, 20 μl proteinase K was added, and samples were incubated for 1 h at 56 °C, 900 rpm on the Qiacube (an alternative heater-shaker platform would be an Eppendorf Thermomixer). RNase treatment for 2 min was done at RT. Following DNA extraction from whole brain samples, each line of the MIKK panel was whole genome sequenced at greater than 30× coverage using Illumina X10 instrumentation at the Wellcome Trust Sanger Institute. Library preparation was performed following the standard PCR-free Illumina protocol [77]. In total, 1 µg was picked from DNA extraction plates within the Sanger sample logistics facility and passed into the Sanger sequencing pipeline from library preparation. Following successful preparation of sequencing libraries, the samples were QCed using Qubit and samples passing the facility quality control threshold were multiplexed sequenced in paired end mode with 5 samples per Illumina X10 flow cell.

RNA extraction and Illumina RNA sequencing

RNA extraction from 52 liver samples was performed on a Qiagen automated extraction platform using QIAsymphony RNA Kits, where polyA RNA was extracted from liver samples for paired end RNA-Seq analysis. Samples were prepared for Illumina RNA sequencing using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina and sequenced on a Hiseq 4000 sequencing platform following the manufacturer's instructions. We obtained data passing the facility quality control threshold for 50 samples out of the 52 initially processed.

Bioinformatic methods and data

Raw sequencing data can be retrieved from ENA linked to the following project ID set out in the "Availability of data and materials" section. All the scripts and metadata used for this study are extensively described in the associated github repository available at https://github.com/birneylab/MIKK_genome_main_paper [78].

DNA sequence alignment and SNP calling

After sequencing, the MIKK panel genomic data was transferred securely from Sanger compute storage to EBI where alignment and variant calling was performed. All whole-genome sequence (WGS) datasets were aligned against the latest medaka *HdrR* reference genome from ENSEMBL (release 94) using the Burrow-Wheeler Aligner for short-read alignment (BWA) [79]. For variant calling, we used the best practices approach for calling SNPs and INDELs using the GATK software [80]. Briefly, aligned BAM files were processed using the Picard software [81] to make duplicate reads and correct read group tags prior to the creation of gVCF files using the HaplotypeCaller command from GATK. Finally, gVCFs were combined and genotyped using GenotypeGVCFs from GATK. The result of this processing was a single multisample VCF file containing SNP and INDEL calls across all lines of the MIKK panel (MIKK Illumina callset).

RNA sequencing data processing and eQTL analysis

We developed a Snakemake pipeline [82] called NanoSnake [83] to run the entire analysis, including read trimming, mapping, quality control, and transcript abundance estimate. For this study, we ran NanoSnake v0.0.3.1 RNA_Illumina workflow (https://github.com/a-slide/pycoSnake/tree/0.0.3.1) [84]. All the tools and environment are version-controlled in individual conda environments. Briefly, reference genome, transcriptome, and annotations were obtained from ensembl Release 98 (Japanese medaka *HdrR* ASM223467v1, https://www.ensembl.org/Oryzias_latipes/Info/Index). For each of the 50 datasets (Column G, Additional file 1: Table S1), Illumina pair-end reads were cleaned up using Fastp v0.20.0 [85], aligned to the reference genome using STAR v2.7.3a [86] and obtained estimated counts per gene. In parallel, we also performed a transcriptome pseudo-alignment transcript level quantification using Salmon v1.1.0 [87]. The configuration file and run script used for NanoSnake are available at https://github.com/birneylab/MIKK_genome_main_paper/tree/master/eQTL/code/nanosnake [78].

We estimated transcript abundances with Salmon [87] and ran the eQTL analysis with Limix [88], looking for Cis-eQTLs within 500 kb of each transcript and using the genetic relatedness matrix between lines (Additional file 4: Fig. S8) as a covariate.

For the eQTL analysis, we used the transcript abundance estimates obtained from Salmon and applied a pre-filtering step to retain transcripts with a minimal count of 5 reads and a minimal TPM (transcript per million) of 1 in at least half of the samples (30/60). For valid transcripts (13,413 out of 36,777), we then normalized the TPM values using a quantile Gaussian normalization. SNP genotypes were cleaned up and preprocessed with PLINK [89]. We performed a light LD pruning to filter out highly correlated variants, retaining 45.7% of all variants. Using limix v3.0.4 [88], we first computed a genetic relatedness (kinship) matrix. Then, for each valid transcript, we ran a LMM univariate association test between the observed phenotypes (normalized TPM) and the genotypes of every SNPs within 500 kb upstream and downstream to the transcript, using the kinship matrix as a covariate (limix.qtl.scan). Finally, we used python scripting to extract significant hits and generate visualizations, including Manhattan and metagene plots. The entire analysis notebook and raw data are available at https://github.com/birneylab/MIKK_genome_main_paper/tree/master/eQTL [78], with the rendered notebook available to view here: https://birneylab.github.io/MIKK_genome_main_paper/eQTL/code/LIMIX_eQTL_analysis.html.

Definition of homozygous blocks across the medaka genome

To assess the level of homozygosity in the MIK panel, we applied a number of different approaches. Firstly, we used the PLINK software [89] to calculate the inbreeding coefficient (*F*) from the MIKK panel genotype calls. Additionally, we used PLINK to call runs of homozygosity across each of the MIKK panel lines across the medaka genome, as well as calculating identity by descent (IBD) of all panel lines. To estimate an accurate level of homozygosity and create a homozygous block call set across the entire medaka genome, we developed a Hidden Markov Model (HMM) to call the allelic status across all MIKK panel lines using 10 kilobase (kb) resolution across the medaka genome. Briefly, the number of heterozygous SNP genotype calls was counted across 10-kb non-overlapping windows in the medaka genome for each MIKK panel line separately. These genotype counts were used to train a 2-state HMM with a Poisson distribution using the expectation maximization (EM) algorithm, followed by the Viterbi (forward/backward) algorithm to define the most likely state space path. Following this process, for each MIKK panel line, we have 10-kb regions of the medaka genome that were called as either heterozygous or homozygous.

Nucleotide diversity and structural variation in the MIKK panel

We calculated nucleotide diversity in 500-kb windows across the MIKK panel based on 63 “non-sibling” MIKK panel lines (column H, Additional file 1: Table S1), and across 7 wild Kiyosu individuals, using [90, 91] for the variant processing and pipeline management, and [92] to calculate $\hat{\pi}$. The full code used for that section is set out here: https://birneylab.github.io/MIKK_genome_main_paper/03_Nucleotide_diversity.html [78]. In this section, we also incorporated additional analyses that we carried out separately on repeat content (see below) and structural variation [44]. The code relevant to the structural variation analysis used in this section can be found here: https://birneylab.github.io/MIKK_genome_main_paper/06_Structural_variation.html [78].

Medaka to human ortholog mapping

To map medaka genes to human orthologs, we used the PANTHER resource [93] which includes the complete set of protein-coding genes for 142 organisms including medaka (*Oryzias latipe*s) derived from the reference proteome project at Uniprot [94]. Briefly, we used the PANTHER inferred least diverged orthologs (LDOs), to map to HGNC and ENSEMBL gene identifiers resulting in direct mapping of medaka genes to human orthologs. We were able to assign human orthologs to over half (12,952 genes) of the annotated coding genes in the current ENSEMBL gene build for the HdrR reference genome (ASM223467v1). Over 90% (11,678) resulted in a direct one to one mapping of medaka to human ortholog, and once we obtained the human gene identifiers, we were able to further annotate these genes with additional information derived from human studies such as the intolerance to mutation predictions using the probability of loop of function (LoF) intolerance (pLI scores) from the Exome Aggregation Consortium (ExAC) database [49].

Population genetic analysis of the MIKK panel

First, we wanted to test the genotypes contained in the MIKK panel lines against the distribution of alleles from the original wild population of Kiyosu medaka. To do this, we calculated the fixation index (F_{ST}) using genotype information from 7 wild Kiyosu lines, comparing the allele frequencies against those in the MIKK panel.

Introgression with Northern and Korean medaka strains

In light of previous findings [36], we sought to compare the extent to which the MIKK panel showed evidence of introgression with established inbred medaka strains originating from southern Japan (*iCab*), northern Japan (*HNI*), and Korea (*HSOK*). We used the 50-fish multiple alignment from Ensembl release 102 [95] to obtain the aligned genome sequences for *HdrR*, *HNI*, and *HSOK*, as well as the most recent common ancestor of all three strains. Using the phylogenetic tree provided with the dataset, and the *ape* R package version 5.4.1 [96], we determined the most recent common ancestor of those three strains. For each locus with a non-missing base for *HdrR*, we 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 Illumina VCF containing SNPs called in the MIKK panel lines and *iCab* strain. We then calculated $\hat{f}_d$ in sliding windows as described in [55], using the scripts provided by the first author on their GitHub page (https://github.com/simonhmartin/genomics_general). Data processing and analyses were carried out with R version 4.0.2 [97], and the Tidyverse suite of R packages version 1.3.0 [98]. Figure 7 was generated using the R packages *ape* [96] and *circlize* [99]. All code used for this section can be found here: https://birneylab.github.io/MIKK_genome_main_paper/04_Introgression.html [78].

MAF and LD decay in the MIKK panel

The MIKK Illumina callset contained SNP and INDEL calls for 79 of the 80 extant MIKK panel lines (column D, Additional file 1: Table S1). In order to avoid allele frequency biases introduced by the 16 pairs/triplets of "sibling lines," we removed each pair's arbitrarily labelled second sibling line from the variant call set, leaving 63 MIKK panel lines for this analysis (column H, Additional file 1: Table S1) ("MIKK non-sibling calls"). To assess how accurately one may be able to map genetic variants using the MIKK panel relative to a human dataset, we compared the MIKK panel's minor allele frequency (MAF) distribution and LD structure against that of the 2504 humans in the 1KG Phase 3 release [57]. To prepare the "1KG calls," we first downloaded the VCFs 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 [80].

We used PLINK [100, 101] to calculate the minor allele frequencies for all non-missing, biallelic SNPs in both the MIKK non-sibling and IKG calls ($N_{\text{SNPs}} = 16{,}395{,}558$ and 81,042,381 respectively), and then used R [97] and the tidyverse suite of R packages [98] to produce Fig. 8A. To visualize the MIKK panel's LD structure, we first used PLINK to filter the MIKK non-sibling callset for non-missing, biallelic SNPs with MAF ≥ 0.10, leaving 2,968,786 SNPs. We then used PLINK to take a random sample of 3000 SNPs per chromosome and recode them for analysis in Haploview [102]. We used Haploview to generate LD plots covering the length of each of the 24 chromosomes in the default color scheme, showing r^2 values for each pair of SNPs within 1 Mb of each other (Additional file 9: Fig. S10). 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 calls and the 1KG calls, we 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). We then used R and the tidyverse suite of R packages to group 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, then calculated the mean r^2 in each of those bins and generated the plots in Fig. 8B using the mean r^2 and left boundary of each bin. All code used for this section can be found here: https://birneylab.github.io/MIKK_genome_main_paper/02_LD_decay.html [78].

Prediction and annotation of repetitive and transposable elements

The *RepeatModeler* pipeline (v2.0.0) [103] for the automated de novo identification of repetitive and transposable elements was run on all chromosomes in the *HdrR* genome assembly [104]. RepeatModeler was run with its default parameters and the additional long terminal repeat (LTR) structural discovery sub-pipeline that includes the *LTRharvest* [105] and *LTR_retriever* [106] tools.

The RepeatModeler library of repeats was filtered to remove non-TE protein-coding sequences by using a protein BLAST (Altschul *et al.*, 1990) to align (E-value $\leq$ 1e−5) the *Oryzias latipes* proteome (Ensembl v99) and *pfam* peptide database (v32) against the RepeatMasker peptide library. Finally, a nucleotide BLAST was used to remove any RepeatModeler repeats that aligned (E-value $\leq$ 1e−10) against the corresponding transcripts.

RepeatMasker (v4.1.0) [107] was used to align the chromosomes in the *HdrR* assembly against the filtered RepeatModeler library of consensus repeats and the existing RepeatMasker repeat families.

Additionally, *Exonerate* (Slater and Birney, 2005) was used to align the two subtypes of the *Teratorn* mobile element found in the *Oryzias latipes* genome against the *HdrR* reference (The *Teratorn* element being the result of a fusion between a *piggyBac* DNA transposon and a member of the *Alloherpesviridae* family [108]).

Additional code used to generate Additional file 4: Fig. S4 can be found here: https://birneylab.github.io/MIKK_genome_main_paper/05_Repeats.html [78].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02623-z.

Additional file 1: Table S1. *MIKK panel lines*. Table setting out all MIKK panel lines and the specific analyses each was included in.

Additional file 2: Table S2. *Fecundity data*. Table with MIKK panel fecundity data.

Additional file 3: Table S3. *Fish imaging parameters*. Table with MIKK panel morphometric data.

Additional file 4: Figures S1-S8. *Supplementary figures*. Various supplementary figures.

Additional file 5: Table S4. *Variant types*. Counts of short variants in each class based on predicted variant effect.

Additional file 6: Table S5. *High confidence LoF variants*. Table of homozygous high-confidence loss-of-function variants in the MIKK panel orthologous to a human dosage-sensitive gene.

Additional file 7: Table S6. *eQTL SNPs*. Significant eQTL SNPs with annotations.

Additional file 8: Figure S9. *Copenhagen plot*. Plot showing genomic positions of significant eQTL SNPs.

Additional file 9: Figure S10. *Haploview plots*. Linkage disequilibrium (LD) plots generated with Haploview for each of the 24 medaka chromosomes, showing r^2 values in the MIKK panel for each pair of SNPs within 1 Mb of one another.

Additional file 10. Review history.

Acknowledgements

The authors acknowledge Thomas Seitz, and Irina Oks and Marzena Majewski for supporting the fish husbandry, and Alicia Günthel, Rachel Müller, and Beate Wittbrodt for laboratory assistance. F.L. dedicates this paper to Sabina and Rolf Loosli-Walther.

Review history

The review history is available as Additional file 10.

Peer review information

Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

Conception of the project: E.B., T.F., F.L., K.N., J.W.; Project management and supervision: E.B., T.F., F.L., J.W.; Sampling of wild fish: K.N.; Inbreeding: N.K., F.L., N.W.; Sample preparation: N.A., C.B., J.G., O.T.H., E.H, C.H., C.L., F.L., K.L., N.S., R.S., E.T., T.Ta., T.Th., P.W., B.W., J.W.; Image acquisition: N.A., C.H., N.Kh.; Data analysis: C.B., E.B., I.B., T.F., A.L., F.L., J.M., J.W.; Manuscript writing: E.B., I.B., T.F., A.L., F.L., J.W. All author(s) read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. The work was funded by the Helmholtz funding program BIFTM to F. Loosli, N. Wolf, N. Kusminski, C. Herder, and N. Aadepu. E. Birney, T. Fitzgerald, A. Leger, C. Barton, J. Monahan, and I. Brettell were funded by the EMBL European Bioinformatics Institute (EMBL-EBI). This work was supported by Heidelberg University Core Funding to T. Tavhelidse, T. Thumberger, and J. Wittbrodt. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 810172), from the NIH UH-3338-03 (JW), the German Ministry for Research (BMBF: HIGH-life 05K19VH1, Code-Vita 05K16VH1, JW), and the German Center for Heart Diseases DZHK (JW, JG).

Availability of data and materials

The datasets supporting the conclusions of this article are available in the European Nucleotide Archive (ENA) hosted at the EBI: https://www.ebi.ac.uk/ena/browser/home.
The individual raw sequencing datasets are linked to the following project IDs:
• Nanopore DNA sequencing data (PRJEB43089): https://www.ebi.ac.uk/ena/browser/view/PRJEB43089 [109].
• Illumina DNA sequencing data (PRJEB17699): https://www.ebi.ac.uk/ena/browser/view/PRJEB17699 [110].
• Illumina RNA sequencing data (PRJEB43091): https://www.ebi.ac.uk/ena/browser/view/PRJEB43091 [111].
All the scripts and metadata used for this study are extensively described in the associated github repository under MIT License available at https://github.com/birneylab/MIKK_genome_main_paper [78] and https://doi.org/10.5281/zenodo.5779413 [112]. Accession numbers are included in the relevant methods sections.

Declarations

Ethics approval and consent to participate

Medaka (***Oryzias latipes***) fish were maintained at the medaka facility of the Institute of Biological and Chemical Systems, Biological Information Processing (IBCS-BIP). No animal experiments were carried out for this study. Animal husbandry was performed in accordance with local and European Union animal welfare standards (Tierschutzgesetz §11, Abs. 1, Nr. 1, AZ 35-9185.64/BH KIT). The facility is under the supervision of the local representative of the animal welfare agency.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [2]Institute of Biological and Chemical Systems, Biological Information Processing (IBCS-BIP), Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany. [3]Centre for Organismal Studies, Heidelberg University, Campus Im Neuenheimer Feld 230, 69120 Heidelberg, Germany. [4]National Institute for Basic Biology, Laboratory of Bioresources, Okazaki, Japan.

Received: 30 June 2021 Accepted: 31 January 2022
Published online: 21 February 2022

References

1. Fisher RA. XV.—The correlation between relatives on the supposition of Mendelian inheritance. Earth Environ Sci Trans R Soc Edinb. 1919;52(2):399–433.
2. Toyama K. On some Mendelian characters (in Japanese). Rep Jap Breed Soc. 1916;1:1–9.
3. Alonso-Blanco C, Aarts MGM, Bentsink L, Keurentjes JJB, Reymond M, Vreugdenhil D, et al. What has natural variation taught us about plant development, physiology, and adaptation? Plant Cell. 2009;21(7):1877–96.
4. Mackay TFC, Huang W. Charting the genotype-phenotype map: lessons from the Drosophila melanogaster Genetic Reference Panel. Wiley Interdiscip Rev Dev Biol. 2018 7(1). Available from: https://doi.org/10.1002/wdev.289
5. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The Drosophila melanogaster Genetic Reference Panel. Nature. 2012;482(7384):173–8.
6. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78.
7. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. Cell. 2016;167(5):1415–29.e19.
8. Ganna A, Verweij KJH, Nivard MG, Maier R, Wedow R, Busch AS, et al. Large-scale GWAS reveals insights into the genetic architecture of same-sex sexual behavior. Science. 2019;365(6456) Available from: https://doi.org/10.1126/science.aat7693.
9. Mackay TFC, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet. 2009;10(8):565–77.
10. Bergelson J, Roux F. Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. Nat Rev Genet. 2010;11(12):867–79.
11. Visscher PM, Naomi WR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS Discovery: biology, function, and translation. Am J Hum Genet. 2017;101(1):5–22.
12. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nat Rev Genet. 2019;20(8):467–84.
13. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013;9(1):1–9.
14. The Complex Trait Consortium. The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet. 2004;36(11):1133–7.

15. Ashbrook DG, Arends D, Prins P, Mulligan MK, Roy S, Williams EG, Lutz CM, Valenzuela A, Bohl CJ, Ingels JF, McCarty MS, Centeno AG, Hager R, Auwerx J, Lu L, Williams RW. A platform for experimental precision medicine: the extended BXD mouse family. Cell Syst. 2021;12(3):235–47.e9.
16. Svenson KL, Gatti DM, Valdar W, Welsh CE, Cheng R, Chesler EJ, et al. High-resolution genetic mapping using the Mouse Diversity outbred population. Genetics. 2012;190(2):437–47.
17. Saul MC, Philip VM, Reinholdt LG. Center for Systems Neurogenetics of Addiction, Chesler EJ. High-diversity mouse populations for complex traits. Trends Genet. 2019;35(7):501–14.
18. Threadgill DW, Miller DR, Churchill GA, de Villena FP-M. The collaborative cross: a recombinant inbred mouse population for the systems genetic era. ILAR J. 2011;52(1):24–31.
19. Rosenthal N, Brown S. The mouse ascending: perspectives for human-disease models. Nat Cell Biol. 2007 ;9(9):993–9.
20. Schofield PN, Hoehndorf R, Gkoutos GV. Mouse genetic and phenotypic resources for human genetics [Internet]. Hum Mutat. 2012;33:826–36 Available from: https://doi.org/10.1002/humu.22077.
21. Morse HC III. Origins of Inbred Mice. Elsevier. 2012:736.
22. Wade CM, Daly MJ. Genetic variation in laboratory mice. Nat Genet. 2005;37(11):1175–80.
23. Larson G, Burger J. A population genetics view of animal domestication. Trends Genet. 2013;29(4):197–205.
24. Takeda H, Shimada A. The art of medaka genetics and genomics: what makes them so unique? Annu Rev Genet. 2010;44:217–41.
25. Aida T. On the inheritance of color in a fresh-water fish, APLOCHEILUS LATIPES Temmick and Schlegel, with special reference to sex-linked inheritance. Genetics. 1921;6(6):554–73.
26. Kirchmaier S, Naruse K, Wittbrodt J, Loosli F. The genomic and genetic toolbox of the teleost medaka (Oryzias latipes). Genetics. 2015;199(4):905–18.
27. Gutierrez-Triana JA, Tavhelidse T, Thumberger T, Thomas I, Wittbrodt B, Kellner T, et al. Efficient single-copy HDR by 5′ modified long dsDNA donors. Elife. 2018;7 Available from: https://doi.org/10.7554/eLife.39468.
28. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2009;19(2):327–35.
29. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013;496(7446):498–503.
30. Bachtrog D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat Rev Genet. 2013;14(2):113–24.
31. Schartl M. A comparative view on sex determination in medaka. Mech Dev. 2004;121(7-8):639–45.
32. Matsuda M, Sakaizumi M. Evolution of the sex-determining gene in the teleostean genus Oryzias. Gen Comp Endocrinol. 2016;239:80–8.
33. Hyodo-Taguchi - Zool. Mag.(Tokyo) Y, 1980. Establishment of inbred strains of the teleost, Oryzias latipes. ci.nii.ac.jp [Internet]. 1980; Available from: http://ci.nii.ac.jp/naid/10005820467/
34. Murata K, Kinoshita M, Naruse K, Tanaka M, Kamei Y. Medaka: Biology, management, and experimental protocols. Hoboken: Wiley-Blackwell; 2019. p. 368.
35. Hyodo-Taguchi Y. Inbred strains of the medaka, Oryzias latipes (Development of Medaka Biology in Japan-Part I). Fish Biol J Medaka. 1996;8:11–4.
36. Spivakov M, Auer TO, Peravali R, Dunham I, Dolle D, Fujiyama A, et al. Genomic and phenotypic characterization of a wild medaka population: towards the establishment of an isogenic population genetic resource in fish. G3. 2014;4(3):433–45.
37. Koger CS, Teh SJ, Hinton DE. Variations of light and temperature regimes and resulting effects on reproductive parameters in medaka (Oryzias latipes). Biol Reprod. 1999;61(5):1287–93.
38. Kimura T, Shimada A, Sakai N, Mitani H, Naruse K, Takeda H, et al. Genetic analysis of craniofacial traits in the medaka. Genetics. 2007;177(4):2379–88.
39. Kimura T, Takehana Y, Naruse K. pnp4a is the causal gene of the medaka iridophore mutant guanineless [Internet]. G3. 2017;7:1357–63 Available from: https://doi.org/10.1534/g3.117.040675.
40. Weinberg SM, Cornell R, Leslie EJ. Craniofacial genetics: where have we been and where are we going? PLoS Genet. 2018;14(6):e1007438.
41. Claes P, Roosenboom J, White JD, Swigut T, Sero D, Li J, et al. Genome-wide mapping of global-to-local genetic effects on human facial shape. Nat Genet. 2018;50(3):414–23.
42. Matsuda M, Nagahama Y, Shinomiya A, Sato T, Matsuda C, Kobayashi T, et al. DMY is a Y-specific DM-domain gene required for male development in the medaka fish. Nature. 2002;417(6888):559–63.
43. Nanda I, Kondo M, Hornung U, Asakawa S, Winkler C, Shimizu A, et al. A duplicated copy of DMRT1 in the sex-determining region of the Y chromosome of the medaka, Oryzias latipes. Proc Natl Acad Sci U S A. 2002;99(18):11778–83.
44. Leger A, Brettell I, Monahan J, Barton C, Wolf N, Kusminski N, et al. Genomic variations and epigenomic landscape of the Medaka Inbred Kiyosu-Karlsruhe (MIKK) panel [Internet]. bioRxiv. 2021:2021.05.17.444424 [cited 2021 May 21]. Available from: https://www.biorxiv.org/content/10.1101/2021.05.17.444424v1.
45. Hoffmann AA, Sgrò CM, Weeks AR. Chromosomal inversion polymorphisms and adaptation. Trends Ecol Evol. 2004;19(9):482–8.
46. Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ. Complex SNP-related sequence variation in segmental genome duplications. Nat Genet. 2004;36(8):861–6.
47. Otake H, Shinomiya A, Kawaguchi A, Hamaguchi S, Sakaizumi M. The medaka sex-determining gene DMY acquired a novel temporal expression pattern after duplication of DMRT1. Genesis. 2008;46(12):719–23.
48. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
49. Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. Measuring intolerance to mutation in human genetics. Nat Genet. 2019;51(5):772–6.

50. Mi H, Thomas P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol Biol. 2009;563:123–40.
51. Nagylaki T. Fixation indices in subdivided populations. Genetics. 1998;148(3):1325–32.
52. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). Nat Rev Genet. 2009;10(9):639–50.
53. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710–22.
54. Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Mol Biol Evol. 2011;28(8):2239–52.
55. Martin SH, Davey JW, Jiggins CD. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. Mol Biol Evol. 2014;32(1):244–57.
56. Details on a Compara analysis [Internet]. [cited 2020 Oct 7]. Available from: http://apr2020.archive.ensembl.org/info/genome/compara/mlss.html?mlss=1828
57. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
58. Naruse K, Fukamachi S, Mitani H, Kondo M, Matsuoka T, Kondo S, et al. A detailed linkage map of medaka, Oryzias latipes: comparative genomics and genome evolution. Genetics. 2000;154(4):1773–84.
59. Hofmann H, Wickham H, Kafadar K. Letter-Value Plots: Boxplots for Large Data [Internet]. J Comput Graphical Stat. 2017;26:469–77 Available from: https://doi.org/10.1080/10618600.2017.1305277.
60. Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. 2008;4(10):e1000214.
61. Strunz T, Grassmann F, Gayán J, Nahkuri S, Souza-Costa D, Maugeais C, et al. A mega-analysis of expression quantitative trait loci (eQTL) provides insight into the regulatory architecture of gene expression variation in liver. Sci Rep. 2018;8(1):5865.
62. Mason VC, Schaefer RJ, McCue ME, Leeb T, Gerber V. eQTL discovery and their association with severe equine asthma in European Warmblood horses. BMC Genomics. 2018;19(1):581.
63. Geiger M, Sánchez-Villagra MR, Lindholm AK. A longitudinal study of phenotypic changes in early domestication of house mice. R Soc Open Sci. 2018;5(3):172099.
64. Ikeda D, Koyama H, Mizusawa N, Kan-No N, Tan E, Asakawa S, et al. Global gene expression analysis of the muscle tissues of medaka acclimated to low and high environmental temperatures. Comp Biochem Physiol Part D Genomics Proteomics. 2017;24:19–28.
65. Naruse K, Tanaka M, Takeda H. Medaka: a model for organogenesis, human disease, and evolution. Switzerland: Springer Nature; 2011. p. 387.
66. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, et al. A high-resolution recombination map of the human genome. Nat Genet. 2002;31(3):241–7.
67. Katsumura T, Oda S, Hiroshi M, Oota H. Medaka population genome structure and demographic history described via genotyping-by-sequencing [Internet]. Available from: https://doi.org/10.1101/233411.
68. Nelson JS, Grande TC, Wilson MVH. Fishes of the World. Hoboken: Wiley; 2016. p. 752.
69. Volff J-N. Genome evolution and biodiversity in teleost fish. Heredity. 2005;94(3):280–94.
70. Postlethwait J, Amores A, Cresko W, Singer A, Yan Y-L. Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends Genet. 2004;20(10):481–90.
71. Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, et al. The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. Nat Genet. 2016;48(4):427–37.
72. Hyodo-Taguchi Y. Establishment of inbred strains of the teleost, Oryzias latipes. Zool Mag (Tokyo). 1980;89:283–301.
73. Loosli F, Köster RW, Carl M, Kühnlein R, Henrich T, Mücke M, et al. A genetic screen for mutations affecting embryonic development in medaka fish (Oryzias latipes). Mech Dev. 2000;97(1-2):133–9.
74. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: a system for large-scale machine learning. In: 12th ${USENIX} symposium on operating systems design and implementation ({OSDI}$ 16); 2016. p. 265–83.
75. Labelbox: The leading training data platform for data labeling [Internet]. [cited 2021 May 6]. Available from: https://labelbox.com.
76. Jung AB, Wada K, Crall J, Tanaka S, Graving J, Yadav S, et al. Imgaug. San Francisco: GitHub; 2020.
77. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: overviews and challenges. Biotechniques. 2014;56(2):61–4 66, 68, passim.
78. Leger A, Brettell I. MIKK_genome_main_paper. Github. 2021; Available from: https://github.com/birneylab/MIKK_genome_main_paper/tree/v1.0.0.
79. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
80. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
81. Picard Tools - By Broad Institute [Internet]. [cited 2020 Mar 9]. Available from: http://broadinstitute.github.io/picard/
82. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics. 2012 ;28(19):2520–2.
83. Leger A. a-slide/NanoSnake: v0.0.3.1. 2020 [cited 2021 Feb 17]; Available from: https://zenodo.org/record/3630380
84. Leger A. a-slide/pycoSnake: v0.1a2 [Internet]. 2020. Available from: https://zenodo.org/record/4110611
85. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
86. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
87. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–9.

88. Casale FP, Rakitsch B, Lippert C, Stegle O. Efficient set tests for the genetic analysis of correlated traits. Nat Methods. 2015;12(8):755–8.
89. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
90. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021;10:33.
91. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.
92. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
93. Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, et al. Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat Protoc. 2019;14(3):703–21.
94. Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, et al. Standardized benchmarking in the quest for orthologs. Nat Methods. 2016;13(5):425–30.
95. [No title] [Internet]. [cited 2021 Apr 9]. Available from: ftp://ftp.ensembl.org/pub/release-102/emf/ensembl-compara/multiple_alignments/50_fish.epo/.
96. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019 ;35(3):526–528.
97. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna: R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/
98. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. J Open Source Soft. 2019;4(43):1686.
99. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics. 2014;30(19):2811–2.
100. Purcell S, Chang C. PLINK 1.9 [Internet]. [cited 2020 Aug 4]. Available from: http://www.cog-genomics.org/plink/1.9/.
101. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
102. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21(2):263–5.
103. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2: automated genomic discovery of transposable element families. Genomics. bioRxiv. 2019:378. https://doi.org/10.1101/856591.
104. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, et al. The medaka draft genome and insights into vertebrate genome evolution. Nature. 2007;447(7145):714–9.
105. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008;9:18.
106. Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018;176(2):1410–22.
107. Smit AFA, Hubley R, Green P. RepeatMasker home page [Internet]. 2010. Available from: http://www.Repeatmasker.org
108. Inoue Y, Saga T, Aikawa T, Kumagai M, Shimada A, Kawaguchi Y, et al. Complete fusion of a transposon and herpesvirus created the Teratorn mobile element in medaka fish. Nat Commun. 2017;8(1):551.
109. Fitzgerald L. [Brettel, Birney]. Nanopore DNA-seq of MIKK medaka brain samples. PRJEB43089. Gene Expression Omnibus. https://www.ebi.ac.uk/ena/browser/view/PRJEB43089.
110. Fitzgerald L. [Brettel, Birney]. Illumina DNA-seq of MIKK medaka brain samples (Medaka Kiyosu panel). PRJEB17699. Gene Expression Omnibus. https://www.ebi.ac.uk/ena/browser/view/PRJEB17699.
111. Fitzgerald L. [Brettel, Birney]. Illumina RNA-sequencing of MIKK medaka liver samples. PRJEB43091. Gene Expression Omnibus. https://www.ebi.ac.uk/ena/browser/view/PRJEB43091.
112. Leger A, Brettell I. birneylab/MIKK_genome_main_paper: Final-submission-snapshot [Internet]. Zenodo. 2021; Available from: https://zenodo.org/record/5779413.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.