Male colour patterns of the Trinidadian guppy (Poecilia reticulata) are typified by extreme variation governed by both natural and sexual selection. Since guppy colour patterns are often inherited faithfully from fathers to sons, it has been hypothesised that many of the colour trait genes must be physically linked to sex determining loci as a ‘supergene’ on the sex chromosome. Here, we phenotype and genotype four guppy ‘Iso-Y lines’, where colour was inherited along the patriline for 40 generations. Using an unbiased phenotyping method, we confirm the breeding design was successful in creating four distinct colour patterns. We find that genetic differentiation among the Iso-Y lines is repeatedly associated with a diverse haplotype on an autosome (LG1), not the sex chromosome (LG12). Moreover, the LG1 haplotype exhibits elevated linkage disequilibrium and evidence of sex-specific diversity in the natural source population. We hypothesise that colour pattern polymorphism is driven by Y-autosome epistasis.
Understanding the genetic architecture of highly diverse and ecologically-important traits is a fundamental problem in evolutionary biology. This is particularly true for traits where the underlying genes are predicted to be on the Y chromosome. Because the Y-chromosome does not recombine, it is expected to degrade over time, whilst its unique inheritance from fathers to sons will select for genes that increase male fitness. However, this model neglects important factors such as pleiotropy and epistasis. Sex-linked colour polymorphism provides a tractable trait for exploring the evolutionary and ecological drivers of balancing selection and sex-chromosome evolution, and genome sequencing methods have hugely enhanced our ability to detect the genetic basis of colour traits. Using a unique breeding strategy designed to delineate regions of the genome related to colour, we analyse whole-genome sequencing data to uncover the genetic basis of sex-linked colour polymorphism in the Trinidadian guppy (Poecilia reticulata).

Guppy colour traits have fascinated biologists for a hundred years, and present an exciting system for testing predictions of sex-linked polymorphic traits. Males display a mosaic of complex and diverse colouration patterns, varying in colour, number, shape, size, and position of spots, while females are a drab and uniform tan colour. Guppy colour patterns exhibit high levels of standing genetic variation, despite evidence that mate choice and predation impose directional selection. Considerable evidence suggests that genetic diversity is maintained by negative frequency dependent selection (NFDS), driven by female mate preference for rare or novel morphs and also frequency-dependent survival. This great diversity in colour patterns, and our understanding of the evolutionary processes maintaining it, the underlying genetic architecture remains largely unknown.

It has long been hypothesised that colour patterning genes and the sex determining locus (SDL) form a Y-linked supergene in the guppy. The supergene-hypothesis originates from early pedigree studies that found patrilineal inheritance for many guppy colour pattern elements. While clearly demonstrating the importance of the Y-chromosome for colour, this early work also reported among-population variability in the strength of Y-linkage, indicating a more complex genetic architecture. Indeed, although there is an enrichment of pigment genes on male-specific contigs, recent studies have also highlighted the importance of X-linked and autosomal inheritance of colour traits. A QTMT mapping study found that only 13% of colour trait loci mapped to the sex chromosome (LG1) and a pedigree analysis of colour pattern inheritance showed that ornaments are not completely Y-linked, hypothesising a potential role for Y-autosome epistasis. Combined, this work suggests that a non-recombining, male-specific region on LG12 plays an important, but not exclusive role in guppy colour patterning.

Recent genomic studies have therefore focussed on identifying the boundaries and genetic content of the non-recombining Y-region. Using male specific diversity across multiple populations, the Y-specific region was concluded to be small, possibly only a single gene, occurring near the distal end of LG13. Indeed, sex has been mapped to the distal end of LG12 in multiple mapping and pedigree crossing studies. Other studies however, conclude that the non-recombining region extends beyond this distal region in some populations based on similar sex-specific genomic diversity measures. Intriguingly, the distal end of LG12 has been found to be highly diverse among males with many segregating male-specific variants, indicative of multiple Y haplotypes, as would be predicted under NFDS for Y-linked colour traits. Within this candidate region, however, no gene associated with colour or sex has been found. Clearly, more work is needed to delineate the Y-linked genomic architecture of this important trait.

In this work, we use an innovative approach to identify genomic regions associated with highly variable, sex-linked, colour polymorphism in guppies. We phenotype and genotype four Iso-Y lines, which originated from a natural population, and show strong Y-linked parenteral heritability in colour pattern. Each Iso-Y line was founded by one male showing a distinct colour pattern. This colour pattern was introgressed onto a randomised genetic background by mating males that resembled the founder, to females from a randomly-mating stock population. Outcrossing to stock population females occurred every 2–3 generations over the 40 generations that the lines were propagated. Given that each backcrossed generation theoretically reduces the parental genome by half, we expect more than 99.99% of the genome to be homogenised through this approach. This experimental design allows us to delineate regions of the genome related to colour pattern, as it should homogenise regions unrelated to the colour differences among lines. Using a Pool-seq approach on each Iso-Y line, we find that the lines are consistently different on an autosome (LG1), not the sex chromosome (LG12). In order to examine the LG1 colour-linked candidates in a natural population, we then conduct an analysis of whole-genome sequencing (WGS) data from the source population, finding that a large and diverse autosomal haplotype is maintained in nature. We hypothesise that colour pattern variation is driven by Y-autosome epistasis. This genetic architecture could help to explain how high levels of Y-linked diversity is maintained in guppy colour patterns.

Results

Iso-Y lines are distinct across multiple dimensions of colour.

Male guppies have an ultraviolet component to their colour patterns, and guppies can detect and adjust their social behaviour based on ultraviolet colouration. Consequently, we used multispectral digital photography to capture human-visible and ultraviolet images of males from each Iso-Y line. We used geometric morphometrics to correct for individual differences in body size and shape among fish so that colour patterns could be measured as though they existed on identical male bodies. We then used the Colormesh pipeline to extract colour measurements from these images at sampling locations across the body and caudal fin.

Discriminant analysis of principal components (DAPC) revealed that males from the different Iso-Y lines were well-separated based on colour pattern. Discriminant functions (DF) 1, 2 and 3 accounted for 51.9%, 36.8%, and 11.4% of the variation among the Iso-Y lines, respectively (Fig. 1a), and distinguished Iso-Y9 from the other Iso-Y lines along axis 1. Axis 2 predominately separated Iso-Y10, and axis 3 subtly differentiated the lines (Supplementary Fig. 1). Colour variation on the caudal peduncle and in the anal region were most important for differentiating the Iso-Y lines (Fig. 1b). Variation in the red colour channel associated with DF1 captured the orange colour spot observed on the anal region in Iso-Y9 (Fig. 1c).

The Iso-Y lines also showed robust phenotypic differences when we examined mean colour measures using permutational MANOVA. Here, we reduced the dimensionality of the colour pattern data using PCA, with 17 PC’s explaining 59.3% of the total variation in colour measurements (see Supplementary Fig. 2 for PC distributions of each of the Iso-Y lines). The omnibus test indicated significant overall differences among Iso-Y lines (df = 3169, pseudo-F = 23.66, P < 0.001). Post-hoc pairwise tests revealed significant differences in colour pattern among all pairs (all P < 0.001; see Supplementary Table 1). Based on centroid distances, the greatest phenotypic differences were between Iso-
Y6 and Iso-Y10 and the smallest phenotypic differences between Iso-Y6 and Iso-Y8. Using permuted t-tests to determine whether the Iso-Y lines differ in phenotypic variance, we found that Iso-Y9 was significantly more variable (Supplementary Table 2; Supplementary Fig. 3).

Iso-Y lines are consistently different on an autosome. Using a pooled whole-genome sequencing (Pool-seq) approach and summarising all pairwise comparisons with a PCA, we were able to identify where along the genome the Iso-Y lines were consistently different (Fig. 2). Pool-seq of each of the four Iso-Y lines (n_per line = 48) resulted in a final dataset of 3,995,905 SNPs. Mean pairwise \( F_{ST} \) between the Iso-Y lines was high overall (\( F_{ST} = 0.091 \)). We first used pairwise \( F_{ST} \) values among the four Iso-Y lines, and then used a multivariate approach by normalising \( F_{ST} \) to \( Z_{F_{ST}} \) and summarising these with PCA \( F_{ST} \). The aim here was to summarise covariance of \( F_{ST} \) among comparisons and assess whether certain regions were consistently differentiated among the Iso-Y lines. \( Z_{F_{ST}} \) PC1 accounted for 37% of the total variance and reflected positive covariance in all pairwise \( F_{ST} \) comparisons (Supplementary Table 3).

LG1 and LG12 showed excess divergence among the Iso-Y lines and were clear outliers compared to the rest of the genome (Fig. 2a). LG1 and LG12 both had the highest chromosome-wide average \( Z_{F_{ST}} \) PC1 scores (LG1 = 1.48; LG12 = 1.62) and the highest percentage of SNPs with a \( Z_{F_{ST}} \) PC1 score above an upper 95% quantile of 3 (LG1 = 24%; LG12 = 29%). The remaining 47% of high-scoring SNPs were distributed across the genome, with other individual chromosomes or scaffolds accounting for <7% of high-scoring SNPs (according to an upper 95% quantile of 3; Supplementary Fig. 4). Other chromosomes showed inconsistent regions of localised differentiation (Supplementary Fig. 5) and we found no evidence indicative of a genome-wide relationship between recombination and \( Z_{F_{ST}} \) PC1 (Supplementary Fig. 6). Thus, LG1 and LG12 became our focus for investigating differentiation among the Iso-Y lines. Interestingly, the patterns of differentiation were different for these two focal chromosomes. \( Z_{F_{ST}} \) PC1 indicated three regions of high differentiation on LG1 (Fig. 2b). On LG12, these scores were elevated consistently along the entire chromosome (Fig. 2d).

By performing an additional PCA on LG1, we found good agreement between the areas of differentiation amongst the Iso-Y lines. \( Z_{F_{ST-LG1}} \) PC1 (52% of the total variance) showed high positive loadings among five of the six pairwise comparisons (Iso-Y6–Iso-Y8 being the exception; Supplementary Table 4), which was also reflected in the per-SNP pairwise \( F_{ST} \) comparisons, where Iso-Y6 and Iso-Y8 exhibited the lowest differentiation (Supplementary Fig. 7; Supplementary Table 5). The Iso-Y6–Iso-Y8 comparison loaded positively onto PC2 (17% of total variance) and \( Z_{F_{ST-LG12}} \) PC2 (and to some extent, PC3) highlighted the same regions of differentiation as PC1 (Supplementary Fig. 8). We found no relationship between \( Z_{F_{ST-LG12}} \) PC1 scores and the recombination landscape on LG1 (Supplementary Fig. 9).

In contrast, PC axis loadings for \( Z_{F_{ST-LG1}} \) on LG12 were different among the Iso-Y lines (Supplementary Table 6). This was also apparent in per-SNP pairwise \( F_{ST} \) values (Supplementary Fig. 10). All comparisons with Iso-Y9 loaded strongly onto PC1 (PC1 captured 37% of the total variance), suggesting Iso-Y9 is the most differentiated line on LG12. This was consistent with pairwise \( F_{ST-LG1} \), where Iso-Y9 had the highest mean pairwise \( F_{ST} \) SNPs (>0.2) and thus high \( Z_{F_{ST-LG12}} \) PC1 scores reflected the haplotype associated with the Iso-Y9 phenotype. \( Z_{F_{ST-LG12}} \) PC2 (PC2 captured 30% of total variance) loadings were high for the remaining two comparisons with Iso-Y6. This suggests SNPs with high \( Z_{F_{ST-LG12}} \) PC2 scores reflect the haplotype associated with the Iso-Y6 phenotype. \( Z_{F_{ST-LG12}} \) PC3 reflected the final comparison, Iso-Y8–Iso-Y10. \( Z_{F_{ST-LG12}} \) PC2 and PC3 also reflected Iso-Y line-specific loadings (Supplementary Fig. 11). Taken together, this demonstrates that Iso-Y specific haplotypes co-occur at the same regions on LG1, whereas on LG12, the Iso-Y specific haplotypes occur in different regions.

We were able to further identify areas of consistent divergence in three regions on LG1 using change point detection (CPD) on
the allele frequencies (AFs—polarised to the major allele of Iso-Y9) and \( Z_{ST} \) PC1 (Fig. 2c). Region 1 encompassed ~1.9 Mb (coordinates: 4,079,988–5,984,584 bp). Region 2 encompassed ~7.4 Mb (coordinates: 9,627,619–17,074,870 bp). Region 3 encompassed ~3 Mb (coordinates: 21,944,840–24,959,750 bp). Using the same CPD methods, no regions were consistently differentiated between the Iso-Y lines on LG12 (Fig. 2e).

Next, to explore diversity within the Iso-Y lines, we calculated \( \pi \) across 10 kb windows for our two focal chromosomes (LG1 and LG12) and performed the same multivariate approach as above. On LG1, we found that regions of differentiation could be explained by a shared chromosomal landscape of diversity among all Iso-Y lines with \( Z_{ST} \)-LG1 PC1 (PC1 captured 74% of the variance) accounting for the majority of variance in diversity (Supplementary Fig. 12; Supplementary Table 7). We found no evidence of increased coverage associated with increased diversity in the regions of LG1 (Supplementary Fig. 13). \( Z_{ST} \)-LG12 PC1 accounted for most of the variance (PC1 captured 88% of the variance) and showed a significant (>99% upper quantile of 4.6) peak in diversity at 24.27 Mb (Supplementary Fig. 14; Supplementary Table 8). \( Z_{ST} \)-LG12 PC2 (PC2 captured 7% of the variance) showed a high level of residual variance in \( \pi \), associated with the variance of Iso-Y9, that was not otherwise explained by the general chromosomal landscape. The diversity hotspot (peak at 24.28 Mb, >99% upper quantile of −1.4) represented by PC2 is thus unique diversity within Iso-Y9. Both LG12 peaks: 24.27 Mb and 24.28 Mb (Fig. 2e).

![Fig. 2 Genetic differentiation between the four Iso-Y lines: Iso-Y6; Iso-Y8; Iso-Y9; Iso-Y10.](https://example.com/fig2.png)

**a** Density plots of the SNP distribution of \( Z_{ST} \) PC1 for each chromosome of the guppy genome (23 chromosomes). The colour scale represents the \( Z_{ST} \) scores as depicted in the legend (yellow: low-scoring SNPs; indigo: high-scoring SNPs). Black lines within each density curve mark the median value of each chromosome. Dashed x-axis intersect marks the upper 95% quantile of 3. **b** Per-SNP \( Z_{ST} \) PC1 scores for LG1; yellow line represents a smoothed spline of the data. **c** Allele frequency (AF) plots of LG1 for each Iso-Y line (Iso-Y6—red; Iso-Y8—blue; Iso-Y9—purple; Iso-Y10—green). AFs were polarised to the major allele of Iso-Y9. Changes in the AFs are represented by the colours depicted in the legend. On LG1, Change Point Detection (CPD) pinpointed three regions of consistent differentiation, which are indicated at the top by black rectangles: Region 1 (4–5.9 Mb); Region 2 (9.6–17 Mb); Region 3 (21.9–24.9 Mb). **d** Per-SNP \( Z_{ST} \) PC1 scores for LG12; yellow line represents a smoothed spline of the data. **e** Allele frequency (AF) plots of LG12 for each Iso-Y line. AFs were polarised to the major allele of Iso-Y9 and changes in the AFs are represented by the colours depicted in the legend. On LG12, Change Point Detection (CPD) did not uncover any consistent regions of differentiation between the Iso-Y lines. X ticks are displayed in Megabases (Mb). Source data underlying all components of Fig. 2 are provided as Source data files.
Evidence of multiple haplotypes and candidate colour genes on LG1. In Region 2 (9.6–17 Mb), two bands of $F_{ST}$ were apparent in analysis of Z-$F_{ST}$-LG1 PC1, and in pairwise $F_{ST}$ (Fig. 2b; Supplementary Fig. 7). To explore this further, we assessed the segregation of the AFs within each of the three identified regions in more detail (Supplementary Fig. 15). Corresponding to the double-banding of $F_{ST}$, Region 2 showed unusual AF patterns within the lines. Iso-Y9 showed fixation, but the other three Iso-Y lines showed multiple bands of AFs, which taken together did not sum to 1. Additionally, in Region 3, Iso-Y6 also showed two sets of distinct bands of AFs. Assessment of the AF density distributions showed clear patterns of bimodality (tridomality in Iso-Y6) in Region 2 (Supplementary Fig. 16) and bimodality in Iso-Y6 in Region 3 (Supplementary Fig. 17). The distinct patterning corresponding to the different bands of AFs could suggest strong linkage between the SNPs segregating in each band, indicative of multiple maintained haplotypes. We reasoned that a complex haplotype structure exists, in which alleles associated with more recently derived haplotypes are nested within an older haplotype (Fig. 3). Bimodal AF distributions parsimoniously represent AFs associated with the older (larger density peak) and younger, derived (smaller density peak) haplotypes.

We found several strong candidates for colour, male-specific fitness and vision in the three differentiated regions on LG1 (see Supplementary Data 1 for a full list of gene annotations).

Region 1 (4–5.9 Mb) contained 62 predicted genes. Of interest was $D$ll, with a role in caudal fin, dorsoventral patterning in $D$. rerio$^{23}$, pcm1 involved in spermatogenesis$^{14}$, and $pptn13$ which is $Y$-linked in humans$^{45}$ and in some fish due to its physical linkage with $gsdf$, a gene which is highly conserved in fish sex differentiation pathways$^{46,47}$. Region 2 (9.6–17 Mb) contained 291 predicted genes. Genes with a potential role in colour included $xpa$, involved in pigmentation and photosensitivity to UV light$^{46}$, $pcahn10a$ involved in melanocyte migration, $crebbpa$, which has been identified as a candidate for plumage colouration in chickens$^{49}$, and $sho2$, which causes pigment abnormalities$^{50}$, as well as five keratin genes, which have a role in pigmentation$^{21}$. We also identified five retinal genes ($slc24a3$,$stru$, $plmpda$,$cab2a$ and $nxn1$), and three genes involved in spermatogenesis or sperm motility ($ttd7d2a$,$nanos3$ and $tekt4$)$^{39}$. Region 3 (21.9–24.9 Mb) contained 94 predicted genes. This region also contained several promising candidates including two paralogs annotated as $kita$, previously identified as a key gene involved in pigment pattern formation in guppy strains$^{50}$ and zebrailfish$^{51}$, $sox10a$, a sex determining region $Y$-box that regulates the expression of the $mitf$ gene, which is the master regulator of melanophore–melanocyte differentiation in teleosts$^{52}$, and is also responsible for colouration in rock pigeons$^{53}$, $mchr1$, a melanin-concentrating hormone receptor, and a $TRYP$, located in melanocytes and involved in the production of melanin$^{54}$. We also performed an assessment of Gene Ontology (GO) enrichment and KEGG mapping for the three regions on LG1 (see Supplementary Data 2 for results). Analysis using gene annotation information combined across all three LG1 regions showed a significant KEGG mapping to Lysozyme (11 genes). This is of potential interest given that the melanosome is a lysosome-related organelle.

Evidence for the LG1 haplotype in the natural population. We next examined natural data by performing whole-genome sequencing of 26 wild-caught guppies from the source population ($n_{\text{males}} = 16$, $n_{\text{males}} = 10$, average coverage ≥13×, final dataset = 1,021,495 SNPs). Using multiple lines of evidence, we found a large haplotype on LG1 (11.1–15.9 Mb) segregating in the natural source population, which lies within ‘Region 2’ identified in the Iso-Y analysis (9.6–17 Mb). We term this area ‘Region 2-NP’ (Region 2 Natural Population) (Fig. 4).

Analysis of linkage disequilibrium (LD) on LG1 revealed an area of high linkage between 11,114,772 and 15,890,374 bp, where it was apparent that several overlapping linkage blocks exist (Fig. 4a). We also analysed patterns of LD for males and females separately and found that males exhibited at least two distinct neighbouring linkage blocks, but females showed high linkage across the entire region (Supplementary Fig. 18). We then assessed shifts in local ancestry by performing a local PCA approach$^{35}$, which recapitulated the identified areas of high LD (Fig. 4b), with a pattern of significant differentiation (saturated eigenvalues > 0.01) represented by three overlapping multidimensional scales (MDS) on LG1: MDS1: 11,216,906–15,375,083 bp; MDS2: 11,430,012–14,570,259 bp; MDS3: 12,326,328–15,308,055 bp. This demonstrates that subsets of correlated SNPs within Region 2-NP exhibit ancestry relationships that deviate from those observed across the rest of the chromosome; a pattern indicative of inversions, long haplotypes under balancing selection, reduced recombination, or changes in gene density$^{65}$.

Intersex $F_{ST}$ showed differences between the sexes within Region 2-NP (Fig. 4c); in particular, an area of high density of elevated intersex $F_{ST}$ between 12.2 Mb to 13.1 Mb. We found that elevated intersex $F_{ST}$ was driven by a reduction in male-specific diversity, as measured by intersex $D_{XY}$ - female π (Fig. 4d). Overall, these results suggest that selection is operating differently between the sexes in the candidate region.

As recombination history is known to affect the maintenance of tightly linked genetic architectures$^{66}$, we also examined the genome features of Region 2-NP. We did not observe any differences in the proportion of GC content (Supplementary Fig. 19), nor repeat elements (Supplementary Fig. 20) in our candidate region. SNP density within Region 2-NP was considerably higher compared to the rest of the chromosome (Supplementary Fig. 21), but there was no evidence of extremes in read depth indicative of duplication or copy-number variation (Supplementary Fig. 22). Gene density showed a moderate number of genes across Region 2-NP, and a drop in gene density at ~13.2 Mb (Supplementary Fig. 23).

We performed similar analyses on LG12, confirming previously identified differences between the sexes$^{30,31,35,67}$. LD analysis showed two main regions of increased linkage; a fragmented region extending from 4.6 to 6.7 Mb, and another more clearly defined region near the terminal end of the chromosome (23.8–25.3 Mb) (Fig. 4e). A local PCA of LG12 identified two significant MDS axes (saturated eigenvalues >0.01) approximately corresponding with the areas of high linkage (Fig. 4f). MDS1 outliers were observed at the terminal end of the chromosome (coordinates: 21,913,633–25,664,533 bp). MDS2 depicted an area encompassing the latter part of the high linkage block, extending past it in the latter coordinates (coordinates: 5,608,703–7,063,435 bp; Fig. 4e).

Overlapping with MDS2, a significant elevation in intersex $F_{ST}$ was observed (Fig. 4g) and in the latter part of the region, a narrow peak in significantly increased male diversity was detected at 5.67 Mb (Fig. 4h). This second region overlaps with a previously identified $D_{XY}$ outlier window in an analysis of six natural populations (6.94–6.95 Mb and 7.00–7.01 Mb), where male-specific $D_{XY}$ was notably higher in high-predation (HP), compared to low-predation (LP) populations$^{31}$. Moreover, the LD in this region extends across 4.6–6 Mb which includes another previously identified male-biased candidate region (LG12: (π shared by all Iso-Y lines); and 24.28 Mb (π unique to Iso-Y9) overlap with previously recorded high male diversity at the putative non-recombining Y$^{31}$.

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-022-28895-4 | www.nature.com/naturecommunications
An assessment of the gene annotations within the MDS2 coordinates identified only a few candidates of interest (Supplementary Data 3): \textit{dmgdh}, a gene that affects sperm trait variation and is part of a sex-supergene in songbirds\textsuperscript{68} and \textit{bhmt}, a folate-related gene that shows an association with skin pigmentation in humans\textsuperscript{69}.

It is predicted that the terminal region of LG12 contains the sex-determining locus (SDL)\textsuperscript{28,70,71} and a recent multiple population genomics survey identified a sex-linked region between 24.5 and 25.4 Mb in LG12, which overlaps with the region of high LD and MDS1 region identified here\textsuperscript{31}. Moreover, analysis of intersex $F_{ST}$ showed high differentiation between the sexes in the region (Fig. 4g), and analysis of intersex $D_{xy}$ revealed that this was driven by an excess of male diversity (Fig. 4h). In contrast to LG1, which showed that elevated $F_{ST}$ is driven by reduced male diversity, the high diversity observed here is consistent with a hypothesis of multiple diverse Y-haplotypes and NFDS on the Y\textsuperscript{35}. Previous investigation of the gene content within the terminal region found it to be relatively gene-poor, containing multiple repeated copies of NLRP1-related genes\textsuperscript{31}, which form part of the inflammasome\textsuperscript{72} and have no known role in sex determination. Our analysis of gene content for the slightly expanded region found in this population identified a single potential colour pattern candidate (\textit{vldlr}), which is responsible for caudal fin patterning in \textit{Danio rerio}\textsuperscript{73}, and two male-trait candidates: \textit{AIG-1} (androgen-induced protein-1) family, and \textit{spag16} (sperm-associated antigen 16) (Supplementary Data 3).

**Pinpointing informative breakpoints within the LG1 haplotype.** Combining both the natural population data and our experimental Iso-Y lines, we found evidence for the maintenance of a large, highly diverse haplotype on LG1 associated with male-specific colour. In the Iso-Y lines, Region 2 encompassed multiple bands of tightly associated allele frequencies, suggesting multiple derived haplotypes. Analysis of the natural population data revealed that Region 2-NP is characterised by elevated linkage, and is defined by linkage patterns indicative of not just one, but multiple derived local ancestries. Comparisons of male and female sequences in the Iso-Y data revealed high diversity of male haplotypes, which were largely fixed for different derived haplotypes, consistent with the hypothesis of multiple derived male-specific haplotypes.
female diversity within the population suggests selection may operate differently between the sexes in Region 2-NP; in particular, between 12.2 Mb and 13.1 Mb, which showed high intersex $F_{ST}$ and a reduction in male diversity. Structural variant (SV) analyses using short-read and long-read data did not show support for SVs in Region 2 nor Region 2-NP (Supplementary Data 4). Our prediction is that the region encompasses a large and diverse haplotype, and that within this haplotype, genetic rearrangements have resulted in several derived haplotype segments. To interrogate this further, we performed an in-depth analysis.

Individual genotypes of the Iso-Y lines and the natural population data across LG1 revealed long stretches of maintained genotype states within Region 2-NP (Fig. 5a). In the Iso-Y line data, Iso-Y9 showed distinct blocks of fixed HOM ALT genotypes, whilst the other Iso-Y line genotypes were entirely heterozygous (HET). This observation is congruous with the multiple bands of allele frequencies observed in the Iso-Y line data. Of individuals from the natural population ($n_{total} = 26$), 17 ($n_{females} = 10$, $n_{males} = 7$) shared the same overall genotype signature of HOM ALT genotype blocks as seen in Iso-Y9. Seven individuals ($n_{females} = 4$, $n_{males} = 3$) showed genotype blocks alternating between areas with the Iso-Y9 HOM ALT genotype signature, and areas of heterozygosity. Lastly, only two individuals showed a signature of extended blocks of HOM REF genotypes: NAT08 and NAT16 ($n_{females} = 2$, $n_{males} = 0$), yet HOM REF genotypes were only maintained for a proportion of the entire region.

By phasing the heterozygous individuals ($n_{total} = 9$), we found evidence of two large divergent haplotypes encompassing the entirety of Region 2-NP with subsequent recombination at conserved breakpoints (Fig. 5b). This indicates that the whole high linkage block is not one large inversion. We identified six repeated breakpoints (i.e. occurring in ≥ 2 individuals): BP1: 11.6 Mb ($n_{females} = 2$); BP2: 11.7 Mb ($n_{females} = 3$); BP3: 12.2 Mb ($n_{females} = 3$, $n_{males} = 1$); BP4: 13.1 Mb ($n_{females} = 2$, $n_{males} = 2$); BP5: 14.4 Mb ($n_{females} = 2$); BP6: 15.4 Mb ($n_{females} = 2$).

These analyses allowed us to pinpoint breakpoints contributing to patterns of differential selection and linkage between the sexes. From the start of Region 2-NP to BP3 (12.2 Mb), there was a distinct difference in zygosity between males and females (Females: 11 HOM ALT, 4 HET, 1 HOM REF; Males: 10 HOM ALT). Therefore, males are out of HWE; based on the female AFs, we would expect 5 HOM ALT, 3 HET and 2 HOM REF in males. The region delineated by BP3 to BP4 overlaps with a high density of SNPs showing signatures indicative of sex-differential selection (12.2 Mb to 13.1 Mb). The only breakpoint consistent between males and females (BP4) also distinguishes the break in LD in males, the drop in gene density, and a break in derived ancestry estimates of 100 bp windows (Supplementary Figs. 18, 23, 24).
Fig. 5 LG1 Region 2 haplotype structure. a Genotype plot for Iso-Y lines and natural data combined showing the genotype of each SNP depicted as homozygous reference (HOM REF: yellow), heterozygous (HET: pink) or homozygous alternative (HOM ALT: purple); each individual is coloured by Iso-Y line, or by sex (females in purple, males in cyan). The bracket shows the heterozygous individuals used in the haplotype analysis in panel (b). b Haplotype plot of phased data for natural-derived heterozygous samples ($n_{total} = 9$, $n_{males} = 3$, $n_{females} = 6$) polarised to the Iso-Y9 haplotype (purple) and alternative haplotype (yellow). Symbols next to each of the individuals represent sex (females in purple, males in cyan). Breakpoints (BP) in the haplotype are identified when phases switch between purple and yellow. Dashed lines mark conserved BPs ($\geq 2$ individuals): BP1: 11.6 Mb ($n_{females} = 2$); BP2: 11.7 Mb ($n_{females} = 3$, $n_{males} = 1$); BP4: 13.1 Mb ($n_{females} = 2$, $n_{males} = 2$); BP5: 14.4 Mb ($n_{females} = 2$); BP6: 15.4 Mb ($n_{females} = 2$). Arrows at the bottom highlight the location of gene annotations for the region. Source data underlying all components of Fig. 5 are provided as Source data files.
Taken together, this localises our candidate region from the beginning of Region 2-NP to BP4 as both selected in males in a natural population and associated with colour in our breeding design.

Discussion

Overall, our results reveal a surprising genetic architecture for colour pattern diversity in guppies. Even though the differences in colour pattern between the Iso-Y lines are Y-linked (i.e. inherited faithfully from father to son), consistent genetic differences are most pronounced on an autosome, LG1. Analysis of the phenotype data showed that the breeding design was successful in producing four distinct Iso-Y lines with marked differences in colour pattern. Using these unique Iso-Y lines, we delineated three regions on LG1 that were consistently different among the differently coloured lines, and within these regions we identified many genes known to be linked to colour in model species. Then, using extensive WGS from the source population, we further highlight just one of these regions (Region 2-NP: 11.1–15.9 Mb), as having strong linkage and significant local ancestry, finding that a large and variable haplotype is maintained in nature.

We had hypothesised that the Y-linked colour pattern genes in our Iso-Y lines would be located on the sex chromosome, LG12. If colour pattern traits are fully Y-linked, and each Iso-Y line’s unique Y haplotype is fully inherited through the patriline, then with this breeding design we would predict a consistent pattern of high differentiation in the Y-linked region on LG12 between all Iso-Y line comparisons. Instead, we see that overall, the chromosome shows moderate divergence, driven by elevated differentiation in comparisons to Iso-Y9 and localised regions of differentiation in comparisons with Iso-Y6. Whilst elevated \( F_{ST} \) across the chromosome is expected due to sex linkage, we found no evidence of a consistently differentiated region on LG12. Moreover, all guppy chromosomes are acrocentric and show evidence of male heterochiasmy\(^{30,67,74}\); yet recombination events produce diversity patterns similar to those caused by positive selection\(^{85}\), and simulations suggest that balancing selection alone can maintain high-LD, and in particular, high divergence between colour phenotypes\(^{83}\). Taken together, our natural source population data reveals interesting signals in this region but determining the selective forces responsible for its maintenance requires a larger sample size, and colour pattern phenotype data from the natural source population, particularly for recombinant males.

The patterns described here, including the prominent association between the LG1 Region 2-NP haplotype and colour, the high differentiation of this candidate region, and the underlying genetic diversity, suggest that this region is likely to be under strong selection for coloured traits. This is supported by the observed linkage disequilibrium, which is evidence of ongoing selection acting on this region. Furthermore, the strong linkage and significant local ancestry found in this region suggest that this region is maintained in nature, and therefore likely to be under strong selection for coloured traits.

It has been argued that selection must be extremely strong, or unrealistic, for differences in allele frequencies to be maintained between the sexes on an autosome\(^{84,85}\). On the other hand, it is also contended that sex differences on autosomes are artefacts caused by duplications or translocations onto the Y chromosome\(^{86}\). Read coverage was not increased in the LG1 region, and we found no evidence of reduced mapping quality, which would be expected if this region was the result of a duplication or translocation event. Nor do we find evidence that our candidate region on LG1 is misassembled; we previously mined the selective forces responsible for its maintenance requires a larger sample size, and colour pattern phenotype data from the natural source population, particularly for recombinant males.

However, it has also been found that males in Paria exhibit strong patrilineal inheritance of colour pattern, suggesting the importance of the Y\(^{36}\). Given the success of our breeding design, where introgressed Iso-Y lines were significantly different in colour, we also found strong evidence for Y-linkage for colour. Ongoing work is directly testing the roles of Y-autosome epistasis versus Mendelian inheritance on guppy colour pattern by conducting crosses between Iso-Y lines with different LG1 haplotypes. Furthermore, it would be interesting to quantify the colour patterns of testosterone-treated females derived from these lines to more fully explore inheritance patterns in females.

By further exploring our natural source population data, we can begin to hypothesise how this epistatic interaction might operate and how it may lead to the maintenance of Y-linked variation in colour. We recorded differences between the sexes in LG1 Region 2-NP, although this region did not appear in a previous analysis examining consistent differences between the sexes across multiple populations\(^{31}\). We found that this difference is caused by a reduction in expected diversity in males, where males were not in HWE in the region from 11.1 Mb to 13.1 Mb (start of Region 2-NP to BP4), although females were in HWE. This could be indicative of a deleterious Y-autosome interaction, where certain Y haplotypes are incompatible with the ‘REF’ LG1 haplotype. Guppy adult sex-ratios, are indeed, generally female-biased\(^{78-81}\), and show a stronger female-bias in upstream low-predation (LP) environments, like Paria, albeit with sampling variability\(^{78,81}\). We found no evidence for large structural rearrangements (such as inversions) underlying the maintenance of our candidate region, LG1 Region 2-NP, and did not detect any relationship with our predicted conserved recombination points, nor increased GC or TE content. Therefore, exactly which molecular traits are driving this haplotype structure is unclear. By examining phased heterozygous individuals, it is however, apparent that crossing-over is restricted to key points along the haplotype and overall linkage is maintained in the population. Models of balancing selection show that balanced variants produce diversity patterns similar to those caused by positive selection\(^{85}\), and simulations suggest that balancing selection alone can maintain high-LD, and in particular, high divergence between colour phenotypes\(^{83}\). Taken together, our natural source population data reveals interesting signals in this region but determining the selective forces responsible for its maintenance requires a larger sample size, and colour pattern phenotype data from the natural source population, particularly for recombinant males.
do however recognise that evolution of sex determination mechanisms and sex chromosomes have occurred in laboratory-adapted populations in a similar number of generations, in both *Danio rerio* and *Xiphophorus*.

Early research found that the linkage between colour pattern traits and the Y was under selection, with increased Y-linkage for colour traits in downstream, high-predation (HP) environments and X-linkage in the upstream, low-predation (LP) environments. Whether Y-linkage varies by predation environment has been indirectly studied across Trinidad, where females treated with testosterone exhibited more colour patterns in LP populations compared to their HP counterparts. Such observations are also consistent with an autosomal genetic component of colour pattern traits that are under weakened selection in LP environments. Other studies suggest that the size of different strata on LG12 vary between HP and LP, with increased Y-linkage in LP (see also) but these results were not repeated across other HP-LP pairs. Our candidate region on LG1 detect differences in Y-linkage between predation ecotypes. Specifically, we identified a region between 12.2 Mb to 13.1 Mb with reduced diversity in males, compared to females, which also corresponded with a break in male patterns of LD. Previous analysis of molecular convergence between HP and LP populations identified two outlier windows in this area (12.17–12.18 Mb and 12.21–12.22 Mb), indicating that this particular region may be under divergent selection for HP-LP phenotypes. We further compared WGS data from other populations across Trinidad at LG1, and detected a strong signal of HP-LP association between 12.2 Mb to 13.1 Mb, but the large haplotype structure is unique to the population studied here (Supplementary Fig. 26). This suggests that this region may be involved in the differential selection of colour phenotypes depending on predation regime.

Based on our Iso-Y lines and natural population data, we hypothesise a Y-autosome epistatic genetic architecture for guppy colour traits. This architecture may be particularly well-suited to a Y-linked trait under NFDS, such as guppy colour pattern. Models suggest that sex-linked polymorphism can only be maintained by natural selection in unusual genetic systems, where the maintenance of Y-linked variation in Y-autonomous models involves frequency-dependent selection, or interactions with other chromosomes. Recognition of the importance of epistasis in response to selection is growing, and epistasis may be responsible for an increased, or non-linear rate, of adaptation in natural populations or artificial selected lines. Additionally, having loci under NFDS not physically linked to the Y can shield them from increased drift experienced on the Y, allowing for higher levels of variation to be maintained. Finally, for NFDS to operate, we would hypothesise that there should be a balance between keeping coadapted alleles together and breaking them apart to create variation. Together, this suggests that Y-autosome epistasis acting on a diverse autosomal haplotype presents a feasible hypothesis for the maintenance of colour pattern traits. The Iso-Y lines offer a unique resource to explore interactions between LG1 and LG12, and also to further distinguish the role of the three different regions on LG1 by performing focussed crosses between Iso-Y lines. Future studies should also aim to fully characterise the diversity of LG1, including additional long-read sequencing of multiple individuals from both LP and HP environments in order to determine the reservoir of variable haplotypes present in this region.

**Methods**

**Generation of the Iso-Y lines.** All procedures involving live animals were reviewed and approved by the Florida State University Animal Care and Use Committee (protocol no. 1442 and no. 1740). The Iso-Y lines were kindly provided by AE Houde, who established them by choosing male lineages in which colour pattern on the body was strongly Y-linked. Each line was founded by a single male drawn from the ‘Houde’ tributary of the Paria River in Trinidad (Trinidad National Grid System: PS 896886). Males from this tributary are known to show strong Y-linkage. Each Iso-Y line was maintained by breeding males with colour patterns similar to that of each line’s founder; hence, the colour patterns of the Iso-Y lines are ecologically relevant. Every generation, males were mated to females sired from the same tributary and the same line, or crossed into the stock population derived from the same Houde tributary. The lines have been maintained at Florida State University since 2012.

**Colour pattern phenotype analysis.**

*Photography.* We used multispectral digital photography (Sony A7 with full-spectrum conversion; Nikon 105mm f/2.8 El-Nikkor Enlarging Lenses) to capture human-visible and ultraviolet images of males from each Iso-Y line (Iso-Y*, n = 41; Iso-Y9n, n = 48; Iso-Y9n, n = 42; Iso-Y10n, n = 42). Fish were lightly anaesthetised by immersion in a Tricaine mesylate (Pentair) solution and placed on a clear petri dish above a grey background with the left side of the body facing upwards. A soft tip brush was used to raise the dorsal fin, flare the caudal fin, and lower the gonopodium so that these appendages were visible. Fish were illuminated by four metal halide lights (Hamilton, 6500 K bulbs) that simulate the natural photic environment. A size standard and two full-spectrum colour standards (grey = 20% reflectance, white = 99% reflectance; Labsphere) were included beside the fish, to ensure accurate reflectance measurements. Fish were also imaged using a custom system where the body was imaged in reflected light (using a combination of polychromatic lights) and a digital camera. For each fish in the visible spectrum, we measured the size and standard, and to place landmarks around the perimeter of the fish. Following Valvo et al.95, we placed seven traditional landmarks at the tip of the snout, the anterior dorsal fin attachment, the posterior dorsal fin attachment, the dorsal caudal fin attachment, the ventral caudal fin attachment, the posterior gonopodium attachment and the anterior gonopodium attachment. We then placed 55 semi-landmarks at approximately even intervals between the traditional landmarks. We used sliding of semi-landmarks to minimise any shape variation resulting from unequal distribution of semi-landmark placement. We used tpsSuper v2.03 to generate a consensus shape representing the average shape of the males in all 345 photos (173 males * n* = 2). Images of each individual were warped to this consensus shape, thereby mapping every pixel from the original image to an analogous location on the consensus shape.

*Colour measurement.* Analysis using Colormorph v2.07 was performed in R v4.0.2. We first performed Delaunay triangulation, which is used to reconstruct a complex shape (i.e. the shape of the unwrapped fish) using a concise number of points distributed across the surface of that shape. We then measured the average colour of pixels in a radius around each of these sample points. We used cross-validation to determine the optimal number of Delaunay triangulations (more triangulations result in more granular colour pattern data) and sample circle radius (see below). At each sample point, we extracted linear colour measurements for four colour channels: R (red), G (green), and B (blue) and UV. We accounted for any minor fluctuations in the lighting environment by calibrating the colour values for each channel by subtracting the average deviation of colour measured in the photo on the white and grey colour standards from the known reflectance values of those standards.

**Colour pattern differences among Iso-Y lines.** We used Discriminant Analysis of Principal Components (DAPC) to describe the properties of colour pattern as this analysis is recommended for characterising the properties of groups using high dimensional data sets. We used the dapc function in adegenet v2.1.39 to perform a DAPC on the Iso-Y lines following a modified scheme to define the linear combinations of PC scores that minimise within and maximise between group variances. We used DAPC in a cross-validation framework to determine the scheme for capturing colour data that allowed us to best discriminate among the Iso-Y lines.95 We used cross-validation to determine the optimal number of PC to retain, number of Delaunay triangulations to perform, and sample circle radius. Using the xvalDapc function, we performed DAPC on training and validation data, and examined the average proportion of successful assignments for a varying number of retained PC’s. We defined the training and validation populations each as 50% of the individuals from each line. We set the maximum number of PC’s to retain for cross-validation as npca = 173 (the number of fish), with cross-validations performed at 17 different retention levels in increments of 10 PC’s (up to 170). We performed 100 replicates per PC retention level. The average proportion of successful placements was maximised (at 98.8% successful) by retaining 10 PC’s, using 1000 Delaunay triangulations and a sample circle radius of 0.7. Consequently, we used this sampling scheme to measure colour pattern for all phenotypic analyses, retaining 10 PC’s for DAPC. This scheme resulted in 9904
colour measurements. We then compared PC scores among lines using the adonis multivariate analyses. These PC keeping the ratio of observations to variables greater than 10:1 for our subsequent multiple analyses. These PC's together explained 59.3% of the total variation in colour measurements. We then compared PC scores among lines using the adonis function in vegan v2.5-698, which computes the pairwise distances among observations and then performs a permutation test (randomly assigning labels among factors) to partition the distance matrix among sources of variation. P-values were then calculated for each whole-fish colour pattern, thereby accounting for the fact that different colour measures on the same fish are not independent of one another. We created 10 000 permuted samples per test. We additionally visualised the differences in phenotype among Iso-Y lines by generating density plots of each Iso-Y line for all 17 PC's used in permutation MANOVA (Supplementary Fig. 2).

Comparing colour pattern variance among Iso-Y lines. To quantify total within-line variance in colour pattern, we calculated the trace of the variance-covariance matrix among fish within a given line (across all colour channels and sampling locations). We then used permutation t-tests to determine whether phenotypic variance differed between each pair of lines, using the sample function. We permuted the line labels for each whole-fish colour pattern, thereby accounting for the fact that different colour measures on the same fish are not independent of one another. We created 10 000 permuted samples per test and performed two-tailed tests.

Genomic library preparation and variant genotyping Pool-seq of the Iso-Y lines. Genomic DNA was extracted from 48 males of each Iso-Y line using an ammonium acetate extraction method from caudal peduncle99. DNA quality was assessed using gel electrophoresis and DNA concentration calculated using Quant-iT PicoGreen dsDNA reagent (Invitrogen) on a Glomax Explorer Microplate reader (Promega). DNA of each individual was diluted to 18-22 ng/µl and 4 µl of each added to a pooled sample per Iso-Y line. Pools were cleaned using a NEB Monarch clean up kit, with final concentration and quality of each confirmed using a Nanodrop ND-1000 (Thermo Fisher) and a Bioanalyser (Agilent). Library preparation and sequencing was performed at the University of Exeter Sequencing Service, using a NextFlex RAPID PCR-free library preparation protocol. Each of the libraries were sequenced across multiple lanes on a HiSeq 2500 in standard mode, with a 125 bp paired-end metric (Supplementary Table 9).

Raw reads were cleaned using cutadapt v1.1380. Reads were aligned to the guppy genome31 (https://www.ebi.ac.uk/ena/browser/view/GCA_904066995) with bwa mem v0.7.17101 and converted to sorted bam alignment format with samtools (v1.3.1). Genotyping was performed using freebayes v1.3.1120 with GNU parallel120 by chunking the alignment files into 100 bp windows on coverage using samtools (v1.3.1). Freebayes was run with the options: --use-best-n-alleles 4 -g 1000. The raw variant output was filtered for bi-allelic SNP variants at QUAL > 30 and DP > 10, followed by a maximum missing filter of 80% applied to each pool separately followed by application of the minor allele frequency (maf) of 25% applied across all pools using vcftools v1.9077. The VCF file was used as input to ploidy v1.280 where the variants were filtered to exclude sites at <30% minimum coverage and <500x maximum coverage per pool, with a minimum read count per allele of 10 (Final SNP set = 3,995,905 SNPs). We used the curated liftover for LG12 for all analyses and plots of LG12109.

Whole-genome sequencing (WGS) of wild-caught individuals. For the WGS of 26 wild-caught guppies, we sequenced individuals from the Paria river (n = 9) and used these with previously available data31 (Supplementary Data 5). Genomic DNA was extracted using the Qiagen DNeasy Blood and Tissue kit (QIAGEN, Cat No./ID: 69506). DNA concentrations ≥35 ng/µl were normalised to 50 ng in 50 µl and were prepared as Low Input Transposase Enabled (LITE) DNA libraries at The Earlham Institute, Norwich. LITE libraries were sequenced on an Illumina HiSeq4000 with a 150 bp paired-end metric and a target insert size of 300 bp, and were pooled across several lanes so as to avoid technical bias with a sequencing coverage target of ≥10x per sample. Data from the LP Marianne river was previously generated (n = 17101). Although the Paria and LP Marianne guppies are sampled from different rivers, there is strong evidence that gene flow exists between the populations occupying the upper reaches of these rivers106,111. We found that genome-wide mean FST calculated between populations in this dataset was 0. Importantly, there was no effect of river on haplotype structure (Supplementary Data 5). Data processing followed previous methods using GATK4 v4.1.8.1112. Final filtered VCF file included filtering for bi-allelic SNPs at a minimum depth of 5 and a maximum depth of 200, removing 50% missing data and application of a 10%maf filter (Final SNP set = 1,021,495). Variants were phased individually with Beagle v5.0.113, and then phased again using Shapeit v2.r304.113,115 making use of phase-informative reads (PIR)116. We used the curated liftover for LG12 for all analyses and plots of LG12109.

Analysis of Iso-Y-line genomic data. Analyses were conducted in R v4.0.1. Pairwise Fst1 was calculated using the Anova Fst2, method implemented in poolfastst v1,2100. Per SNP Fst1 values were compared pairwise between each Iso-Y line. To identify chromosomes with the highest mean variance in Fst2 differentiation across our dataset, mean Z-scores of each PC were summarised for each chromosome117. These were calculated using the prcomp function using stats v3.6.2, centering and scaling the results.

To delineate line-specific blocks apparent from the distinct patterns of differentiation observed in the Fst2 analysis, we evaluated the allele frequencies (AF) at each locus across the Iso-Y line. We used the curated liftover for LG12 for all analyses and plots of LG12109. To assess diversity among the Iso-Y lines, π was calculated from the Iso-Y line allele frequencies on a per base pair basis121. To summarise the among-Iso-Y-line variation, per SNP π values were used as input to PCA, calculated using the prcomp function.

For functional gene annotation, we extracted the regions of interest from the guppy genome (https://www.ebi.ac.uk/ena/browser/view/GCA_904066995) using samtools faidx (v1.9)102 and aligned them to the previous guppy genome assembly (https://www.ebi.ac.uk/ena/browser/view/GCA_900636155.2) using minimap2 v2.17117 and pulled the uniprot gene IDs from the annotation using biomaRt v2.44.122. GO enrichment and KEGG mapping were performed using clusterGO v1.2.2119. For further visualisation, we used LDheatmap v1.04126.

Analysis of the natural population WGS data. WGS data comprising 26 wild-caught individuals was used to explore the identified Iso-Y regions in natural guppy populations. Linkage disequilibrium (LD) was calculated among polymorphic SNPs for LG1 and LG12. Invariant positions were removed using bcftools v1.8124. Variants were thinned at 5 kb intervals and r2 values were calculated in plink v1.9125. We used the curated liftover for LG12 for all analyses and plots of LG12109.

Population genetics statistics were calculated along LG1 and LG12 using PopGenome v2.7.3127. Intra-population Fst1, Dxy, and π were calculated in non-overlapping 1 kb windows. Inter-population Dxy was calculated in non-overlapping 1 kb windows. π was calculated in non-overlapping 1 kb windows.
Analysis of structural variants and coverage. For the medium-coverage WGS data derived from wild-caught individuals we used BreakDancer v1.4.5\textsuperscript{12} and Lumpy-based smoove v0.2.1\textsuperscript{19}. For the long-read sequencing data from one Iso-Y line individual (Iso-Y6) using snfifi the phased haplotypes to Iso-Y9. Breakpoints between the phases were quantified by identifying the location (in bp) where phase0 switched to phase1, and vice versa. Switchpoints were quantified in males and females and a switchpoint was considered to be conserved when the locations (within 50 kb) overlapped in two or more individuals. Inter-chromosomal linkage was calculated as above for intra-chromosomal linkage calculating r2 values between LG1 and LG12 using the --inter-chr flag in plink v1.9, outputting a normal pairwise matrix. Results were plotted in R using custom scripts\textsuperscript{120}.

Analyses of multiple haplotypes on LG1. Genotype and haplotype plots were created using Genotype Plot in Galaxy\textsuperscript{19}. For the haplotype-based analysis, we polarised the phased haplotypes to Iso-Y9. Breakpoints between the phases were quantified by identifying the location (in bp) where phase0 switched to phase1, and vice versa. Switchpoints were quantified in males and females and a switchpoint was considered to be conserved when the locations (within 50 kb) overlapped in two or more individuals. Inter-chromosomal linkage was calculated as above for intra-chromosomal linkage calculating r2 values between LG1 and LG12 using the --inter-chr flag in plink v1.9, outputting a normal pairwise matrix. Results were plotted in R using custom scripts\textsuperscript{120}.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The DNA sequencing data generated in this study have been deposited in the European Nucleotide Archive (ENA) under the Study Accession PRJEB36506 with the following codes: SAMEA6512722-SAMEA6512725 (Pool-iso Y1-Y10 data); SAMEA8750557-SAMEA8750565 (whole-genome sequencing data for Pairia); SAMEA8795870-SAMEA8795872 (long-read pacbio data for Iso-Y6). The DNA sequencing data generated for Upper Marianne individuals have been deposited in the ENA under the accession codes: SAMEA494973-SAMEA3649973. The male guppy reference genome can be accessed at the ENA under the accession codes: SAMEA904066995. Reassessment data were provided through personal communication with the permission of the authors.

Code availability
All code, scripts and additional data related to analysis are available on GitHub (https://github.com/josieparis/guppy-colour-polymorphism; https://doi.org/10.5281/zenodo.5903522) and https://github.com/jimWhitting1/gynotype_plot; https://doi.org/10.5281/zenodo.5913504).

Received: 26 April 2021; Accepted: 16 February 2022; Published online: 09 March 2022

References
1. Bachtrog, D. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat. Rev. Genet. 14, 113–124 (2013).
2. Williams, T. M. & Carroll, S. R. Genetic and molecular insights into the development and evolution of sexual dimorphism. Nat. Rev. Genet. 10, 797–804 (2009).
3. Connallon, T. & Clark, A. G. Sex linkage, sex-specific selection, and the role of recombination in the evolution of sexually dimorphic gene expression. Evolution 64, 3417–3422 (2010).
4. McKinnon, J. S. & Pierotti, M. E. R. Colour polymorphism and correlated characters: genetic mechanisms and evolution. Mol. Ecol. 19, 5101–5125 (2010).
5. Svensson, E. I. Back to basics: using colour polymorphisms to study evolutionary processes. Mol. Ecol. 26, 2204–2211 (2017).
6. Orteu, A. & Jiggins, C. D. The genetics of coloration provides insights into adaptive evolution. Nat. Rev. Genet. https://doi.org/10.1038/s41576-020-0234-x (2020).
7. Endler, J. A. Natural Selection on Color Patterns in Poecilia reticulata. Evolution 34, 76–91 (1980).
8. Houde, A. Sex, Color, and Mate Choice in Guppies (Princeton University Press, 1997).
9. Brooks, R. & Endler, J. A. Direct and indirect sexual selection and quantitative genetics of male traits in guppies (Poecilia reticulata). Evolution 55, 1002–1015 (2001).
10. Blows, M. W., Brooks, R. & Kraft, P. G. Exploring complex fitness surfaces: multiple ornamentation and polymorphism in male guppies. Evolution 57, 1622–1630 (2003).
11. Hughes, K. A., Rodd, F. H. & Reznick, D. N. Genetic and environmental effects on secondary sex traits in guppies (Poecilia reticulata). J. Evol. Biol. 18, 35–45 (2005).
12. Houde, A. E. & Endler, J. A. Correlated evolution of female mating preferences and male color patterns in the Guppy Poecilia reticulata. Science 248, 1405–1408 (1990).
13. Reznick, D. N., Shaw, F. H., Rodd, F. H. & Shaw, R. G. Evaluation of the rate of evolution in natural populations of Guppies (Poecilia reticulata). Science 275, 1934–1937 (1997).
14. Godin, J.-G. J. & McDonough, H. E. Predator preference for brightly colored males in the guppy: a viability cost for a sexually selected trait. Behav. Ecol. 14, 194–200 (2003).
15. Long, K. D. & Houde, A. E. Orange spots as a visual cue for female mate choice in the Guppy (Poecilia reticulata). Ethology 82, 316–324 (2010).
16. Farr, J. A. Male rarity or novelty, female choice behavior, and sexual selection in the Guppy, Poecilia reticulata Peters (Poeciliidae). Evolution 31, 162–168 (1977).
17. Hughes, K. A., Du, L., Rodd, F. H. & Reznick, D. N. Familiarity leads to female mate preference for novel males in the guppy, Poecilia reticulata. Anim. Behav. 58, 907–916 (1999).
18. Hughes, K. A., Houde, A. E., Price, A. C. & Rodd, F. H. Matting advantage for rare males in wild guppy populations. Nature 503, 108–110 (2013).
19. Graber, R. E., Senagolage, M., Ross, E., Houde, A. E. & Hughes, K. A. Mate preference for novel phenotypes: a fresh face matters. Ethology 121, 17–25 (2015).
20. Daniel, M. J., Koffinas, L. & Hughes, K. A. Mating preference for novel phenotypes can be explained by general neophilia in female Guppies. Am. Nat. 196, 414–428 (2020).
21. Olenendorf, R. et al. Frequency-dependent survival in natural guppy populations. Nature 441, 633–636 (2006).
22. Fraser, B. A., Hughes, K. A., Tosh, D. N. & Rodd, F. H. The role of learning by a predator, Rivulus hartii, in the rare-morph survival advantage in guppies. J. Evol. Biol. 26, 2587–2605 (2013).
23. Lindholm, A. & Breden, F. Sex chromosomes and sexual selection in poeciliid fishes. Am. Nat. 160, S214–S224 (2002). Suppl 6.
24. Kottler, V. A. & Schartl, M. The colorful sex chromosomes of teleost fish. Genes 9, 233 (2018).
25. Winge, O. A peculiar mode of inheritance and its cytological explanation. J. Genet. 12, 137–144 (1922).
26. Winge, O. The location of eighteen genes in Lebiasin reticulatus. J. Genet. 18, 1–43 (1927).
27. Morris, J., Darolti, I., Bloch, N. I., Wright, A. E. & Mank, J. E. Shared and species-specific patterns of nascent Y chromosome evolution in two Guppy species. Genes 9, 238 (2018).
28. Tripathi, N. et al. Genetic linkage map of the guppy, Poecilia reticulata, and quantitative trait loci analysis of male size and colour variation. Proc. Biol. Sci. 276, 2195–2208 (2009).
29. Morris, J., Darolti, I., van der Bijl, W. & Mank, J. E. High-resolution characterization of male ornamentation and re-evaluation of sex linkage in guppies. Proc. Biol. Sci. 287, 20200177 (2020).
30. Bergroth, R., Gardner, J., Bider, B., Yong, L. & Charlesworth, D. Exaggerated heterochromia in a fish with sex-linked male coloration polymorphisms. Proc. Natl Acad. Sci. USA 116, 6924–6931 (2019).
31. Fraser, B. A. et al. Improved reference genome uncovers novel sex-linked regions in the Guppy (Poecilia reticulata). Genome Biol. Evol. 12, 1789–1805 (2020).
32. Kirkpatrick, M. et al. Evolution of the canonical sex chromosomes of the guppy and its relatives. G3 https://doi.org/10.1093/g3journal/kjab435 (2021).

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-022-28895-4 | www.nature.com/naturecommunications
91. Wright, A. E. et al. On the power to detect rare recombination events. Proc. Natl Acad. Sci. USA 116, 12607–12608 (2019).
92. Whiting, J. R. et al. Drainage-structuring of ancestral variation and a common functional pathway shape limited genomic convergence in natural high- and low-predation guppies. PLoS Genet. 17, e1009566 (2021).
93. Clark, A. G. Natural selection and Y-linked polymorphism. Genetics 115, 569–577 (1987).
94. Hansen, T. F. Why epistasis is important for selection and adaptation. Evolution 67, 3501–3511 (2013).
95. Postma, E., Spyrou, N., Rollins, L. A. & Brooks, R. C. Sex-dependent selection differentially shapes genetic variation on and off the guppy Y chromosome. Evolution 65, 2145–2156 (2011).
96. Rohlf, F. J. The tps series of software. Hystrix 26, 9–12 (2015).
97. Rohlf, F. J. tpsDig2 version 2.30. TpsSeries. Stony Brook: SUNY, Department of Ecology and Evolution (2017).
98. Oksanen, J. et al. vegan: Community Ecology Package. R package version 2.5-5. 2020.
99. Nicholls, J. A., Double, M. C., Rowell, D. M. & Magrath, R. D. The evolution of cooperative and pair breeding in thornbills Acanthiza (Pardalotidae). J. Avian Biol. 31, 165–176 (2000).
100. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBO J. 17, 10–12 (2011).
101. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997. Preprint at https://arxiv.org/abs/1303.3997 (2013).
102. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
103. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced genome alignment visualization. Bioinformatics 32, 292–294 (2016).
104. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907. Preprint at https://arxiv.org/abs/1207.3907 (2012).
105. Tange, O. Gnu parallel–the command-line power tool. (2018).
106. Tarasov, A., Viéille, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31, 2023–2034 (2015).
107. Danischek, P. et al. The variant call format and VCFtools. Bioinformatics 27, 2156–2158 (2011).
108. Hiervet, V., Lebois, R., Petit, E. J., Gautier, M. & Vitalis, R. Measuring genetic differentiation from Pool-seq data. Genetics 210, 315–330 (2018).
109. Fraser, B. A., Whiting, J. R., Paris, J. R. & Bemm, F. Guppy genome: V1.0.0 - male guppy genome assembly. https://doi.org/10.5281/ZENODO.4020899 (Zenodo, 2020).
110. Willing, E.-M. et al. Genome-wide single nucleotide polymorphisms reveal population history and adaptive divergence in wild guppies. Mol. Ecol. 19, 968–984 (2010).
111. Blondel, L. et al. Evidence for contemporary and historical gene flow between guppy populations in different watersheds, with a test for associations with adaptive traits. Ecol. Evol. 9, 4504–4517 (2019).
112. Paris, J. R., Whiting, J. R. & Fraser, B. A. joisepiaris/gatk-snp-calling gatk-snp-calling. https://doi.org/10.5281/ZENODO.5903522 (Zenodo, 2022).
113. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. Am. J. Hum. Genet. 103, 338–348 (2018).
114. Delanoe, O., Zajgur, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods 10, 5–6 (2013).
115. Delanoe, O., Howie, B., Cox, A. J., Zajgur, J.-F. & Marchini, J. Haplotype estimation using sequencing reads. Am. J. Hum. Genet. 93, 687–696 (2013).
116. Malinsky, M. et al. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat. Ecol. Evol. 2, 1940–1955 (2018).
117. Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018).
118. Martin, M. et al. WhaSHap: fast and accurate read-based phasing. bioRxiv 085050. Preprint at https://doi.org/10.1101/085050 (2016).
119. Killick, R. & Eckley, I. changepoint: An R package for changepoint analysis. J. Stat. Softw. 58, 1–19 (2014).
120. Paris, J. R. joisepiaris/guppy-colour-polymorphism: Guppy sex-linked polymorphism. https://doi.org/10.5281/ZENODO.3036659 (Zenodo, 2022).
121. Begun, D. J. et al. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol. 5, e310 (2007).
122. Kasprzyk, A. BioMart: driving a paradigm change in biological data management. Database 2011, ba049 (2011).
123. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16, 284–287 (2012).
