Y-linked copy number polymorphism of target of rapamycin (TOR) is associated with sexual size dimorphism in seed beetles

Philipp Kaufmann[1*], R. Axel W. Wiberg[1,5], Konstantinos Papachristos[2], Douglas G. Scofield[3], Christian Tellgren-Roth[4], Elina Immonen[1*]

[1]Department of Ecology and Genetics (Evolutionary Biology program), Uppsala University, Norbyvägen 18D, 75234 Uppsala, Sweden
[2]Department of Cell and Molecular Biology, Molecular Evolution, Uppsala University, Sweden
[3]Uppsala Multidisciplinary Center for Advanced Computational Science, Uppsala University, Uppsala, Sweden
[4] National Genomics Infrastructure, Uppsala Genome Center, SciLifeLab, BioMedical Centre, Uppsala University, 751 08, Uppsala Sweden
[5] Ecology Division, Department of Zoology, Stockholm University, Svante Arrhenius väg 18B, 106 91, Stockholm, Sweden

*Authors of correspondence: philipp.kaufmann@ebc.uu.se, elina.immonen@ebc.uu.se

Abstract

The Y chromosome is theorized to facilitate evolution of sexual dimorphism by accumulating sexually antagonistic loci, but empirical support is scarce. Due to the lack of recombination Y chromosomes are prone to degenerative processes, which poses a constraint on their adaptive potential. Yet, in the seed beetle *Callosobruchus maculatus* segregating Y linked variation affects male body size and thereby sexual size dimorphism (SSD). Here we assemble *C. maculatus* sex chromosome sequences and identify molecular differences associated with Y-linked SSD variation. The assembled Y chromosome is largely euchromatic and contains over 400 genes, many of which are ampliconic with a mixed autosomal and X chromosome ancestry. Functional annotation suggests that the Y chromosome plays important roles in males beyond primary reproductive functions. Crucially, we find that, besides an autosomal copy of the gene *target of rapamycin* (*TOR*), males carry an additional *TOR* copy on the Y chromosome. TOR is a conserved regulator of growth across taxa, and our results suggest that a Y-linked *TOR* provides a male specific opportunity to alter body size. A comparison of Y haplotypes associated with male size difference uncovers a copy number variation for *TOR*, where the haplotype associated with decreased male size, and thereby increased sexual dimorphism, has two additional *TOR* copies. This suggests that sexual conflict over growth has been mitigated by autosome to Y translocation of *TOR* followed by gene duplications. Our results reveal that despite of suppressed recombination, the Y chromosome can harbour adaptive potential as a male-limited supergene.

Keywords: Y chromosome, Y polymorphism, sex chromosome, sexual dimorphism, Callosobruchus maculatus, sexual conflict, Target of rapamycin, TOR

Introduction

Driven by differences in reproductive strategies[1,2] males and females commonly experience sexually antagonistic selection on traits present in both sexes [e.g.3,4]. When the sexes share genetic variation in homologous traits[5], an allele beneficial to females may be deleterious for males and *vice versa*. This genetic dependency can hinder sex-specific adaptations, causing intralocus sexual conflict[6]. The Y chromosome is limited to males, and linkage of sexually antagonistic loci on the Y would represent a straightforward solution to this genetic conflict[7–10]. A male limited pathway to alter

the expression of a sexually antagonistic trait disconnects the genetic basis between the sexes and can reduce gender load by enabling males but also females to reach their fitness optima.

Yet, the role of Y chromosomes in the evolution of sexual dimorphism has historically been neglected[7]. This is because the Y typically degenerates quickly after the loss of recombination with the X, which is expected to constrain its adaptive potential[11]. The Y chromosome degenerates due to reduced efficacy of selection and mutation accumulation (reviewed in[12]), selective haploidisation driven by regulatory evolution[13] or a combination thereof. As a consequence of degeneration, the Y chromosomes are often heteromorphic from the X, void of genes and rich in repetitive elements. Y chromosome sequences are therefore difficult to assemble, and has been done only in a handful of mammalian species such as human[14], other primates[15,16], mouse[17] and bull[18]. There are also a few examples of near complete assemblies such as for malaria mosquitos[19], some fish[20,21] and neo-Y in *Drosophila miranda*[22]. Genes remaining on degenerate Y chromosomes are frequently translocated from the autosomes, allowing their male specialisation. Y linked genes show typically testis specific expression [e.g.14,16,17,20], which suggests they are associated with male primary reproductive functions. But recently, heteromorphic Y chromosomes have gotten more attention in regards to sexually dimorphic, non-reproductive traits in humans[23], drosophila[24], colour and behavioural variation in fish[21], and body size in seed beetles[25].

The order of Coleoptera (beetles) is the most species rich group of animals on the earth. Yet only a handful of beetle species have been sequenced thus far[26], and their Y chromosomes remain largely uncharted. Coleoptera has XY sex determination but the exact mechanism is unknown. A study of karyotypes of over 4000 beetle species shows occasional loss of Y chromosomes[27], suggesting that Y is not essential for sex determination[28]. Interestingly, most studied species in the largest Coleopteran suborder Polyphaga do not have an obligate XY chiasmata formation, and therefore lack a pseudo autosomal region (PAR) and XY recombination altogether[27]. One such species is *Callosobruchus maculatus* seed beetles that harbours a heteromorphic Y chromosome estimated to

represent less than 2% of the genome[29]. Based on the karyotype, the *C. maculatus* Y is euchromatic despite its small size[30].

We recently discovered segregating Y-linked genetic variance that alters male body size[25]. Female and male *C. maculatus* have different fitness optima for body size[31–33], which is closely connected to many other life history traits[34]. A combination of quantitative genetic analysis, artificial selection and isolating the effect of Y linked genetic variance by introgressing the putative Y haplotypes onto a common genetic background, revealed two phenotypically distinct male morphs associated with different Y haplotypes (Y_L and Y_S for large and small size, respectively). We could demonstrate that directional selection on males can deplete Y haplotype variation quickly[35], and that carrying either one of the Y haplotype alters the male size (weight), and consequently the level sexual size dimorphism, by 30%[25]. Body size is a classic quantitative trait, and while quantitative genetic evidence suggests that even the autosomal genetic architecture of body size in *C. maculatus* consist of a combination of few major effect loci and many small effect loci[35], finding that a substantial part of its architecture in males is controlled by a non-recombining 'supergene' is surprising. But a Y-linked element has also been implicated in male body size variation in humans[36] and fish[21,37] suggesting that Y-linkage may be a common way to mitigate sexual conflict over growth across diverse taxa.

In this study we assembled previously un-characterised X and Y chromosome sequences of *C. maculatus*, and identified molecular differences between the two Y haplotypes (Y_S & Y_L) associated with small or large male body size and with major effects on sexual size dimorphism, shedding light on the underlying molecular mechanisms. To do this we took advantage of comparing genomes of the Y_S and Y_L introgression lines[25] (hereafter referred to as S and L, respectively) that share inbred autosomal and X chromosomes and only differ in their respective Y haplotype. Here, we first identified non-recombining X and Y contigs in both S and L genomes, by comparing sequence coverage difference between males and females, and verified male specificity of the longest Y contig (8.4Mb) with PCR. To further confirm the identity of the Y contigs and characterise Y variation, we

compared the genomes of S and L lines that share variants in all other chromosomes except on the Y. We identified and functionally characterised protein coding genes and repeat structure of the Y contigs, analysed the origins of the Y genes as X gametologs or autosomal paralogs, studied gene duplications within the Y to identify putative ampliconic genes, and examined expression of Y-linked genes.

Results

Genome assembly

First, we separately assembled the genomes of the S & L lines, and annotated the S genome (carrying the ancestrally more frequent Y_S haplotype), which was subsequently used as the reference genome in this study. The assembled S genome has 938 contigs that yield a total genome length of 1.246 Gb. The N50 value of the assembly is 9.45 Mb with the longest contig being 37.4 Mb in length, the L50 is 38 and the BUSCO completeness scores over 98% (insecta_odb10: complete: 98.1% [single copy: 87.6%, double copy: 10.5%], fragmented: 0.2%, missing: 1.7%, n=1367), demonstrating that the assembly is of high quality and further improves the previously published genome for this species (*C. maculatus* reference genome; N50 of 0.15 Mb and total genome size of 1.01 Gb[29]). 597 contigs were shorter than 100 kb (consisting of 24.9 Mb length, 2.00% of the total S assembly length) and were not considered in the downstream sex chromosome identification analysis because their chromosome type (autosome, X or Y) could not be determined with confidence due to their short length. The L line assembly is also of high quality with 1323 contigs and a total length of 1.224 Gb, the longest contig being 30.2 Mb in length, with an N50 of 9.86 Mb, L50 of 38 and a BUSCO completeness score of over 97% (insecta_odb10: complete: 97.8% [single copy: 87.5%, double copy: 10.3%], fragmented: 0.3%, missing: 1.9%, n=1367). 945 contigs were shorter than 100 kb (consisting of 41.0 Mb, 3.35% of the total L assembly length).

After repeat-masking, 72.1% of the reference genome was soft-masked, of which 21.1% was identified as retroelements (primarily LINEs: 17.8%) and 21.0% as DNA transposons. A total of 24.8% interspersed repeats remained unclassified. Various low-complexity repeats formed the

118 remainder of the soft-masked content. The final set of annotated gene models include 35,865 genes
119 (68% increase compared to the original assembly[29]), 39,983 transcripts (3451 two-transcript gene
120 models and 297 with more than two transcripts; 14% increase in the total number of transcripts[29]). A
121 total of 25,651 transcripts received functional annotation.

122 **Identification of Y and X contigs**

123 We performed a coverage comparison analysis (with SATC[38]) using Illumina short-read
124 sequencing data from samples of both sexes[29] to identify novel sex chromosome sequences. SATC
125 correctly identified the sexes of the samples. In the S genome assembly four Y (total of 10.1 Mb) and
126 eight X contigs (total of 58.6 Mb) were detected based on significant coverage differences, while in
127 the slightly more fragmented L genome we identified five Y contigs (total of 4.89 Mb) and ten X
128 contigs (total of 64.2 Mb) (Table S1-S2). Importantly, the identified Y contigs from both assemblies
129 map to each other, demonstrating that SATC identified homologous sequences in both assemblies
130 (Fig. S1-S2). The same contigs were identified whether using unfiltered data or when using repeat
131 masked contigs, with minor exceptions (see Table S1-S2 and Fig. S1-S2). Note that the SATC
132 analysis identified several additional contigs to show significant coverage difference between the
133 sexes (see online Table O1), however, the relative coverage difference in these cases was below 10%
134 and we therefore took a more conservative approach and only considered sex chromosome contigs
135 above this threshold in our downstream analysis. It is possible however that these (or other
136 unidentified) contigs are still sex linked but less diverged between X and Y.
137 Gene ontology (GO) enrichment of Y linked transcripts, as compared to all identified sex-linked
138 transcripts, shows that Y is functionally different from the X (Fig. 1). The significantly enriched
139 processes on the Y include cell proliferation, regulation of development, cell death and apoptosis,
140 response to stress/external stimulus such as response to starvation, immune response, RNA
141 processing and regulation of posttranscriptional gene expression as well as protein modification
142 (ubiquitination), and various metabolic processes (Fig. 1 & online Table O2). GO enrichment of Y
143 transcripts for molecular function are presented in Fig. S7.

The identified Y contigs, in either the Y_S or the Y_L haplotype, do not seem to show a higher number of repeats per length, nor a higher percentage of repeat content, compared to the X or autosomal contigs (Fig. S4-S5). However, the composition of repeat content on the Y is somewhat unique, where repeats identified as DNA/Maverick and LINE/Penelope are overrepresented, as a proportion of all repeat elements, on the Y compared to the X or autosomal contigs (Fig. S6).

Gametologs, paralogs and ampliconic genes

We detected 437 transcripts on the Y_S, of which 202 transcripts are ampliconic, and have between 1 and 13 additional nearly identical copies on the Y (>99.9 % nucleotide similarity), forming 67 ampliconic groups. Hence, we identified 302 unique Y-linked transcripts, of which 235 are non-ampliconic transcripts and 67 form ampliconic groups. 424 transcripts have at least one gametolog or paralog in the genome (when searching for Y protein sequences against all *C. maculatus* proteins, using blastp[39] (>50% query coverage) and filtering for >80% sequence identity and E-value threshold = 1e-20). With these criteria we detected in total 281 unique autosomal paralogs, 214 unique X-linked gametologs and 359 unique Y-linked homologous transcripts (Fig. S9). When lowering the sequence similarity threshold, we find homologs even for the remaining 12 Y-linked transcripts, although the best hit protein sequence similarity drops to below 40% for some of them. To identify autosomal or X ancestry for each Y transcript, we categorized them as exclusively autosomal paralogs (n=157), exclusively gametologs on the X (n=99) or homologs on both (n=73) (Fig. 2), while excluding genes that have homologs on uncategorized contigs (<100 kb, n=51 genes). Y shared significantly more exclusive X gametologs than exclusive autosomal paralogs, when accounting for size difference (two-tailed Fisher's exact test: 95%CI (9.56, 16.07), $p < 0.0001$) or difference in the overall number of transcripts (two-tailed Fisher's exact test: 95%CI (8.32, 14.13), $p < 0.0001$). Gametologs that have been maintained on the Y are functionally enriched for RNA processing, particularly genes involved in RNA splicing, development as well as metabolic processes. Y transcripts with paralogs on the autosomes are functionally enriched for response to stimulus, developmental processes, protein translation and post-translation modification (Fig. S8). Interestingly, this elevated sequence similarity

between the X and the Y is not reflected in pronounced sequence synteny blocks between X and Y contigs, neither in nucleotide sequence nor gametolog synteny (Fig. S10).

Characterisation of the Y variation associated with the body size difference

Variant calling

The patterns of shared (hereafter SNP) and fixed (hereafter fixed SNV) single nucleotide variants in the S & L genomes are well aligned with the expectations considering how the lines were created, and further confirm the identity of the detected X and Y sequences. The majority of single nucleotide variants (2,823,154) are shared polymorphisms in both genomes (2,812,361 SNPs, 99.6%) and there are only few fixed SNV differences between the two introgression lines (10,793, 0.38%) (Table 1). Autosomal contigs have significantly more shared SNP/bp than contigs identified as the sex chromosomes (two-tailed Fisher's exact test: 95%CI (28.8, 30.5), $p < 0.0001$ and 95%CI (83.2, 109.6), $p < 0.0001$, for the X and Y respectively) (Fig. S11), and the X contigs have significantly more shared SNP/bp than the Y contigs (two-tailed Fisher's exact test: 95%CI (2.80, 3.71), $p < 0.0001$). In contrast, the Y contigs have significantly more fixed SNV differences/bp than autosomal contigs (two-tailed Fisher's exact test: 95%CI (43.4, 47.3), $p < 0.0001$), and there are no fixed SNV differences on the X contigs (Fig. S12).

We identified a total of 137 genes with fixed SNV differences within the gene region or in close proximity to genes (i.e. ±2 kb as an approximation of cis-regulatory up and downstream area) between the two Y-haplotypes (Fig. 3). For 12 out of these 137 genes, we could identify unique *D. melanogaster* orthologs and annotate their function via FlyBase[40] including DNA/RNA binding, mRNA splicing, regulation of cell proliferation and protein ubiquitination (Table S3, further details in online Table O3).

Y-linked TOR amplicon

One of the annotated Y-linked genes indicated strong homology to the gene *target of rapamycin (TOR),* a highly conserved member of the IIS/TOR pathway[41]. Mapping a consensus *TOR* protein to our assembly via exonerate[42] identified one autosomal *TOR* gene, detected in both Y_S and Y_L

196 genomes (see supplementary methods 'genome annotation' for full details). In addition, there is one
197 gene on the opposite strand that matches to adenosine deaminase 2 in several taxa (also involved in
198 cell proliferation).

199 We further discovered three consecutive copies of the Y *TOR* on the Y_S haplotype that causes the
200 small body size morph in males (Fig. 4A), but there is only a single Y-linked *TOR* in the Y_L haplotype
201 associated with the large body size morph in males (Fig. 4B), revealing Y-linked copy number
202 variation (CNV) of the *TOR* region. Aligning Y_S and Y_L contigs (Fig. S13) shows that the *TOR* region
203 CNV is located in the middle of an otherwise continuous alignment between two haplotypes, showing
204 that CNV is not an artefact caused by a broken Y_L haplotype contig.

205 A closer comparison of the autosomal and the Y linked *TOR* region reveals that all Y-linked *TOR*
206 copies (in both Y_S and Y_L haplotypes) lack the initial five 5' CDS compared to the autosomal *TOR*.
207 A maximum likelihood tree (Fig 5A) of concatenated CDS comparing all *TOR* regions (i.e. the
208 autosomal *TOR* of each assembly (A_S & A_L), three Y-linked Y_S *TOR* copies and one Y-linked Y_L
209 *TOR*) shows the following; First, the autosomal A_S & A_L *TOR* regions are isogenic, as expected.
210 Second, all Y-linked *TOR* sequences cluster together with high bootstrap confidence, indicating that
211 the *TOR* transposition from the autosome to the Y predates the two Y haplotypes. Also, DeepVariant
212 SNP calling did not detect any SNV differences within the *TOR* region between the Y_S and Y_L
213 haplotypes, which might be due to lower coverage when mapping the two Y haplotypes against each
214 other (Fig. 3). We find high confidence that $Y_S^{[a]}$ & $Y_S^{[b]}$ *TOR* copies are most similar to each other
215 but the remaining clustering of Y-linked *TOR* sequences has low support. Importantly, while the
216 exons align with high similarity across all *TOR* regions, non-coding sequences have diverged more
217 and show lower sequence similarities (Fig. 5B). This is particularly apparent when comparing
218 autosomal and Y-linked *TOR* regions, where non-coding sequences frequently do not align, indicating
219 structural differences between them (Fig. 5B, Fig. S13).

220 In addition to the copy number variation of the *TOR* locus, aligning the Y_S and Y_L haplotypes to
221 each other suggests that there are further structural differences between the haplotypes (Fig. S14).

Moreover, there are similarities between the classes and distribution of repetitive elements between the autosomal and the different copies of the Y-linked *TOR* region, indicating homology also at the level of *TOR*-associated repetitive elements (Fig. S15).

Discussion

Here we assembled two new *C. maculatus* genomes. We successfully identified 10 Mb of the Y chromosome, using a set of complementary methods, and analysed the gene content of the Y sequences. We discovered that despite having lost most of its sequence since divergence from the X, the *C. maculatus* Y is rich in genes, many of which are expressed, with diverse functions from metabolism, development and stimulus response to regulation of gene expression and translation. This is in line with karyotype information suggesting that the Y is mostly euchromatic[30], in spite of not recombining. We characterized molecular differences between the two Y haplotypes that were previously inferred from patterns of male limited inheritance of body size[25]. We identified a Y-linked copy of a functionally well conserved autosomal growth factor gene *target of rapamycin* (*TOR*), highlighting the potential for male specific growth regulation via the Y chromosome. Furthermore, we detected copy number variation of *TOR* between the two Y haplotypes underlying body size variation. Together our results indicate a central role of the Y chromosome in the evolution of sexual dimorphism in *C. maculatus* via male-specific evolution of *TOR*.

XY identification

Degenerate sex chromosomes are notoriously difficult to assemble[43] and in *C. maculatus* the entire genome has a high repeat content[29]. Despite these challenges, long read sequencing yielded high quality and contiguous genome assemblies for both Y introgression lines that allowed us to identify large parts of both the X and the Y chromosomes, greatly extending and curating previously identified sex chromosome-linked portions of the genome. Flow cytometry experiments in *C. maculatus* have previously given estimates of ~18 and 93Mb for the size of Y and X chromosomes, respectively, although there is great uncertainty in the size estimate for the Y[44]. We could assemble four Y contigs

yielding 10.1 Mb, ~56% of the current length estimate. By comparison, in *D. melanogaster* ~10% of the Y has been assembled thus far[45], and to our knowledge no Y assemblies exist yet for other beetle species. Additionally, we assembled a total of 58.6 Mb of the X chromosome, corresponding to approximately 63% of the estimated size. Sex chromosome contig identification depends on the quality and length of contigs, accuracy and completeness of the genome, and, importantly, on the state of differentiation between the X and Y chromosomes. Here we rely on a suite of complementary methods including coverage difference between male and female reads, as well as genomic patterns of single nucleotide variants (SNVs) between our Y introgression lines that share the genome apart from the non-recombining parts of the Y chromosome. Mapping of previously described putative X and Y contigs[29] also co-localize to our identified X and Y contigs in our assemblies, as expected. Furthermore, transcript expression patterns across the identified contig groups are in line with the unequal distribution of sex chromosomes between the sexes and show exclusive expression, or significantly higher male bias, of Y linked transcripts, and significantly less male bias for X-linked transcripts, as compared to the autosomal background. A large proportion of Y-linked transcripts show high sequence similarly with X-linked transcripts, in line with their common ancestry. Crucially, we confirmed the largest identified Y contig (8.4Mb) – carrying the identified Y linked *TOR* region – via male limited PCR amplification. The designed Y-specific primers work for both identified Y haplotypes and enables molecular sexing, a method that has thus far been lacking for *C. maculatus*.

Gene content on the Y

We find 437 Y-linked transcripts (417 genes) and the identified Y contigs show high gene density, which is similar to the gene rich Y chromosomes characterized in mouse[17] and bull[18] but in contrast with the general expectation that Y-chromosomes should be low in gene content[7], and other characterized insect Y chromosomes (*Drosophila*[46] and mosquito[19]) that harbour only few genes. Genes retained on the Y, despite its degeneration, can broadly be categorized into two classes, genes that are dosage sensitive X gametologs[47,48] and genes that are male beneficial, which may have been

recruited to the Y via autosomal translocations. Many (often ampliconic) Y genes show testis limited expression and are likely central to the male reproductive function[14,15,17,20,49,50]. However, testis specific expression of highly amplified genes may also be indicative of meiotic drive[17,18], in which case the involved genes need not be male beneficial but Y beneficial instead. Y is predicted to accumulate sexually antagonistic genes only beneficial to males[8], but there is still little direct evidence to support Y-linkage of traits that are also present in females. Metabolic rate, body size, longevity[31,33,51], immunity and gene expression[29,52,53] have all been implicated to be under sexually antagonistic selection in *C. maculatus.* We find functional enrichment of Y-linked genes that reflects these phenotypes remarkably well, including metabolic processes, immune response, response to stimulus and development, cell organisation, growth & cell apoptosis (Fig. 1). This suggests that the Y-linked genes affect sexually dimorphic phenotypes beyond the body size[25], and may offer a resolution to sexual conflict more broadly in *C. maculatus*.

Roughly 50% of the Y genes have either exclusively X gametologs (99 genes, 58.6 Mb), or autosomal paralogs (157 genes, 1,153 Mb) with >80% protein sequence similarity (Fig. 2), indicating different origins for these genes. Notably the number of identified X and autosomal homologs is positively correlated (Pearson correlation: $t = 20.55$, df 113, $p < 0.001$) for 115 Y-linked transcripts that have homologs on both regions (Fig. S16, note that this includes also transcripts that additionally have homologs on uncharacterized contigs), suggesting their coupling to transposable elements. The higher absolute number of genes acquired from the autosomes follows the pattern seen in *D. melanogaster*[54] and humans[14], where all, or a substantial proportion of functional genes originate from autosomes, respectively. In contrast to known XY systems, in female heterogametic ZW taxa the W chromosomes mainly harbor Z gametologs[e.g.55]. The difference is explained by sexual selection favoring transpositions to the Y in male heterogametic taxa[14,16,54]. The finding that *C. maculatus* Y contains a mix of male-specific but also seemingly functional X gametologs, with high sequence similarity retained to the X, suggests that the ancestral X genes have still important roles in males and may evolve under purifying selection. These X gametologs are enriched for functions related to

development and metabolic processes as well as RNA processing/splicing (Fig. S8). Sex-specific RNA splicing allows expression of alternative transcripts in the sexes and can facilitate sexual dimorphism[56,57]. However, whether the X gametologs are dosage-sensitive and sexually concordant or sexually antagonistic and on the Y specialized for male beneficial functions remain to be tested.

The male-specific Y genes acquired via transposition from autosomes are functionally enriched for a broader range of terms than the gametologs, and account much of the general enrichment patterns observed across all Y genes (Fig. S8). An important avenue by which the Y chromosome can affect male phenotypes is by modulating gene expression throughout the genome, an effect described in *D. melanogaster*[58] and for the *SRY* locus in mammals[59,60]. In line with such a mechanism, Y-autosome epistatic effects have also been associated with sexually antagonistic coloration in guppies[61]. *C. maculatus* Y shows significant enrichment for DNA binding molecular function, a category that *SRY* also falls into (Fig. S7), which is consistent with the idea that the Y chromosome has a wider regulatory role. The autosomal paralogs on the Y are enriched for genes involved in protein translation & post-translational modification. While transcriptional sex differences are the commonly evoked explanation for how sexual conflict can be resolved[62], this suggests that the Y chromosome may allow for translational modification to alter male beneficial phenotypic expression.

The Y contigs also contain a large number of sequences with strong homology to other Y loci, suggesting frequent gene duplication events, which is commonly observed in Y chromosomes[14,15,17,18,20,50,63,64]. The level of gene amplification we see is similar to the stickleback Y[20], but less pronounced than in well characterized mammalian Y chromosomes[14,15,17,18,50,64]. Amplification of genes on the Y may be fuelled by sexual conflict over associated traits[17]. Conservation of ampliconic genes is also observed in mammals[64] and suggests that male specific amplified genes may have a large evolutionary advantage that withstand degenerative processes, such as Muller's ratchet[64]. Ampliconic genes tend to be expressed in the testis and enriched for male-specific reproductive functions in mammals[14,15,17,50] and in stickleback[20]. But here we could not yet

detect any genes with obvious reproductive functions in males, based on GO terms or previously described *C. maculatus* seminal fluid proteins (N=185)[65].

X and Y do not form a chiasmata in *C. maculatus* and hence lack recombination by crossing-over and the PAR altogether. All species belonging to the *Callosobruchus* genus lack PAR based on their meiotic karyotypes[27], suggesting that XY divergence predates the genus. Long evolutionary history without recombination with the X could explain why our analysis of sequence similarity between X and Y does not indicate synteny between them (Fig. S10), nor do we find indications for inversion blocks that could have led to recombination suppression between X and Y. Ampliconic regions are prone to structural rearrangements[14,66], which may cause further amplification and can also contribute to the lack of synteny between the X and Y.

Molecular differences between the Y_S and Y_L haplotypes associated with body size variation

We detected substantial molecular differences between the sequences of the Y_S and Y_L haplotypes associated with two distinct male limited body size morphs[25]. As expected for a non-recombining, hemizygous genetic region, nearly all identified point mutation differences are fixed between the two Y haplotypes. The few detected SNPs could indicate real segregating polymorphisms (as the sample for sequencing consisted of multiple males) but are more likely artefacts caused by repetitive elements that add difficulty in genome assembly, accurate mapping, and variant calling. The 137 genes with fixed differences between the Y_S and Y_L haplotypes either in their coding or potential cis-regulatory regions are candidates to explain the phenotypic differences. Independent of body size, males with the Y_S haplotype also sire more offspring (*manuscript in preparation*). This could suggest that the Y haplotypes could cause differences in regulatory pathways associated with seminal fluid production or spermatogenesis, although we did not find any previously described *C. maculatus* seminal fluid proteins to be Y-linked. We identified *Drosophila* orthologs for 14 genes with fixed SNVs, and while we do not find any obvious causal candidate to explain the difference in reproductive capacity we find two notable orthologs: *male-specific lethal* 2 & *ubiquitin specific protease 47*. *Male-specific lethal* 2 is a well-known regulator of dosage compensation between the X and Y in *Drosophila*

melanogaster[67], which opens for differences in regulation of X-linked gene expression as a possible avenue to contribute to the phenotypic differences between the two Y haplotypes. Further, *ubiquitin specific protease 47* is known to interact with insulin/insulin-like signalling pathway (IIS) in *Drosophila*[68], an important pathway that connects nutrient levels to metabolism, growth, development and longevity.

Remarkably, the longest *C. maculatus* Y chromosome contig also contains a *TOR* gene ortholog, a strong candidate to explain the Y linked size variation between the sexes as well as in males. The TOR signalling pathway is highly conserved and present in organisms from bacteria and plants to animals, and is one of the most ancient nutrient-sensing pathways[41]. The TOR pathway regulates growth and lifespan by coupling the growth factor signaling with nutrient sensing[41]. It is centrally involved in controlling cell metabolism, growth, proliferation and apoptosis. Together with IIS, TOR has also been implicated in differential gene expression between the sexes[69] and more specifically sexual size dimorphism in *D. melanogaster*[57]. To our knowledge it has never been detected on a sex chromosome before. But the potential for Y specific regulation of the TOR pathway has recently been implicated in the male polymorphic Poecilid fish *P. parae*, where an inhibitor of the TOR pathway has been detected segregating on the Y[21]. To understand whether the *C. maculatus TOR* gene on the Y represents ancestral homology with the X, or has occurred by a transposition from the autosomes after X-Y divergence, we searched the genome for any *TOR* copies. We could not detect the *TOR* on the X, but only in one of the autosomal contigs, supporting its origin on the Y by translocation. We detected multiple transposable element sequences flanking the *TOR* sequence in all of the Y copies as well as the autosomal one (Fig. S15), which could have played a role in the translocation and should be subject to further investigations.

Importantly, we detected *TOR* copy number variation between the Y_S and Y_L haplotypes, which makes this gene the most likely candidate for the striking difference in male body size between the two haplotypes, and thereby sexual size dimorphism[25]. The elevated intronic sequence divergence between the autosomal and Y-linked *TOR* copies suggests that the translocation of the *TOR* has

happened before the split of the two Y haplotypes, and also that sufficient time has passed since the translocation, to accumulate such differences in the introns. The exonic sequences have diverged only little, suggesting purifying selection on all copies to retain Y *TOR* regions as functional.

How the Y-linked *TOR* functions and putatively interacts with the autosomal *TOR* pathway presents a novel and exciting area for future research. The Y-linked, exonic *TOR* sequence is nearly identical to its autosomal paralog apart from five missing exons on the 5'end. The N-terminal of TOR proteins consist of HEAT repeats[70] that mediate protein-protein interactions[71]. Empirical studies in *D. melanogaster*[72] and amoeba *Dictyostelium discoideum*[73] have demonstrated that overexpressing TOR inhibits cell growth and proliferation similar to loss of function mutants[72,73]. Even a truncated extra copy of TOR is enough to reduce growth[72,74]. Gene copy number can correlate positively with gene expression[75], and *TOR* expression in adults in our data is overall male-biased (Fig. 3). Paralog interference has been suggested as one possible consequence of gene duplications[76], whereby the paralogs can mutually exclude each other from binding with potential partners. The Y TOR could therefore affect male growth by interfering with the autosomal TOR pathway. The more common Y_S haplotype that makes males roughly 30% smaller than the Y_L haplotype has two additional copies of *TOR* on the Y, suggesting that the additional copies lead to further growth inhibition. Future work can establish how each of the copies in the two haplotypes may be expressed and function in regulating growth.

Conclusions

Recombination is a key mechanism generating and maintaining allelic diversity across loci. Finding substantial diversity in the absence of recombination is therefore unexpected, but echoes a similar recent finding in a Y-polymorphic fish[21,37]. Genetic variation should be rapidly fixed by selection and drift on the Y chromosome. Finding segregating Y polymorphism in both copy number and SNV differences in proximity to over 100 genes therefore suggests that processes such as frequency dependent selection likely have actively maintained these Y haplotypes in the population

for a longer evolutionary time. Our identification of over 400 genes on the *C. maculatus* Y, despite the evidence of its degeneration, suggests that males can harbor substantial evolutionary potential through their Y chromosomes. We find that males with different body size morphs vary in the number of copies of a conserved growth factor gene that our data suggests has translocated from an autosome. The Y chromosome thus enables decoupling of the genetic response in the body size evolution between the sexes, positing the *TOR* pathway as the central regulator of sexual size dimorphism in *C. maculatus*.

Methods

Study organism and generation of the Y-lines

As a model organism to study sexual conflict, much is known about sexual antagonism in the seed beetle *Callosobruchus maculatus*. Aphagous adult *C. maculatus* females oviposit directly onto legume seed pods, within which larvae will develop for about 3 weeks, which allows for large scale experiments across multiple generations. The populations used in this study all stem from originally field caught (2010) individuals from Lomé, Togo (more details in[77]) and have been kept in the lab for ~200 generations as 41 isofemale lines.

The creation of Y_S and Y_L haplotype introgression lines is described in detail in[25]. Briefly, replicated bi-directional selection was applied on male body size for 10 generations, giving rise to large (L) and small (S) males. We then crossed S and L males from these selection lines with females from a single inbred line (originating from the same Lomé base population as the selection lines, inbred for >20 generations[78]), respectively. At each subsequent generation, sons were backcrossed to females from the maternal line, for a total of 11 generations, after which the lines were sequenced. The backcrossing scheme replaced the original autosomes and the X chromosome with those from the inbred line while maintaining the non-recombining parts of the Y chromosome of the founding S or L males. Subsequent analysis of the body sizes confirmed the presence of two distinct male body

size classes, while there was no difference in female body size[25]. We chose a single line representing each of the Y_S and Y_L haplotypes for sequencing.

Sequencing and genome assembly

DNA extraction and library preparation

For the extraction of high molecular weight (HMW) DNA, we flash froze adult virgin males (within 24 h after emergence) in liquid nitrogen. Individual abdomens were dissected on ice to avoid thawing of the tissue by removing head, thorax and the elytra. 5 male abdomens were pooled and ground into fine powder with liquid nitrogen and a precooled pestle. The QIAGEN Genomic-Tip 20/g kit was used to extract HMW DNA, following the manufacturer's protocol, with an over-night incubation time for cell lysis (i.e. 12 h). To achieve the required amounts of HMW DNA, we pooled 2 samples (total of 10 males per Y introgression line). 5 µg of genomic DNA were sheared on a Megaruptor3 instrument (Diagenode, Seraing, Belgium) to a fragment size of about 13-16 kb. The SMRTbell library was prepared according to Pacbio's Procedure & Checklist – Preparing HiFi Libraries from low DNA input using SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA). The SMRTbells were sequenced on a Sequel IIe instrument, using the Sequel II sequencing plate 2.0, binding kit 2.2 on three Sequel® II SMRT® Cell 8M per introgression line, with a movie time of 30 hours and a pre-extension time of 2 hours.

The genomes of the Y_S and the Y_L introgression lines were assembled individually using hifiasm (v. 0.7-dirty-r25)[79] with default settings, yielding the S & L genomes. Haplotypes in the resulting assemblies were subsequently separated with purge_dups (v. 1.2.5, default parameters)[80] and genome assembly completeness was assessed via BUSCO[81], using default parameters on the insecta_odb10 database. We then chose the S genome (containing the ancestrally more frequently occurring Y_S haplotype) to be the reference genome, that we subsequently soft-masked for repetitive content[82] using RepeatMasker (v. 4.1.2)[83] and fully annotated using the BRAKER (v. 2.1.6)[84] /TSEBRA (v. 1.0.3)[85] pipeline. Detailed annotation methods are provided in the supplementary information: genome annotation.

Identification of Y and X contigs

Mapping of putative X and Y contigs

We used gmap (v. 2021-03-08)[86] default setting to quantify the number of hits of putative X and Y contigs, previously identified in *C. maculatus*[29], to our assembled contigs in both genomes. Mapping to contigs shorter than 100 kb were excluded.

Sex assignment through coverage (SATC)

To independently identify putative sex-linked contigs in each of our assemblies, we employed SATC[38], that uses normalised coverage information across contigs to first identify XX and XY individuals from sets of male and female samples[38]. Informed by a t-test, SATC compares normalised coverage at each contig between XX and XY individuals to find contigs with significantly different coverage. At sex-linked chromosomes, specific XY:XX coverage ratios are expected for X-linked contigs (0.5:1) and Y-linked chromosomes (1:0)[38]. In practice, coverage can be highly variable resulting in deviations from the strict expectation. Here we used the SATC approach to identify any contigs that had significantly different coverage between XX and XY individuals. We collected Illumina short-read sequencing data from Sayadi et al. (2019; ENA accession numbers: ERR3053159, ERR3053160, ERR3053163, ERR3053164, ERR3053161, ERR3053162, ERR3053165, ERR3053166)[29]. To increase certainty, we limited ourselves to the analysis of contigs > 100 kb in length. Contigs that are shorter than this account for a total of 24'939'325 bp and make up only 2.00% of the reference genome (1'246'713'675 bp). For more details see supplementary information: sex assignment through coverage (SATC).

PCR

We confirmed male-specificity of the longest identified Y contig with a multiplexed PCR, using two primers pairs, one that is Y specific and amplifies a 297 bp on utg000322l_1 product and a primer pair that amplifies an autosomal product of 189 bp length on utg000177l_1 as a positive control. For more detail see supplementary information: Molecular sexing.

Gene expression analysis

To assess how the identified sex-linked genes may be expressed we used gene expression data from virgin adult males (n=29) and females (n=32)[87]. The expression data was collected from reproductive tissues of the abdomen of virgin individuals from different inbred lines that originate from the same Lomé base population as the Y lines. We used splice variant aware mapping of transcript via STAR (v. 2.7.2b) with default settings[88]. We then used picard (v. 2.23.4)[89] to mark duplicates and subread (v. 2.0.0) featurecount[90] to summarize exons by gene IDs allowing for multimappers due to high gene duplications on the Y. Additionally, we also summarized exons by gene IDs using default setting (i.e. no multimapping). We then used DESeq2[91] to analyse the gene expression patterns in males and females. We split the dataset into genes on the autosomes, and the identified X and the Y contigs.

GO enrichment on the Y vs the X

We examined how the Y chromosome is functionally diverged from the X using gene ontology enrichment analysis. For this we used the topGO R package[92], with nodeSize parameter of 10, comparing the frequency of terms among the Y linked transcripts to those among all transcripts on the sex chromosomes (X and Y). Visualisation and clustering of the gene ontology enrichment analysis was done using ViSEAGO R package[93], clustering GO terms based on Wang's semantic similarity distance and ward.D2. Further aggregating of semantic similarity GO clusters was done with best-match average (BMA) method, as implemented in the ViSEAGO package.

Identification of Y homologs on the X and the autosomes

To identify whether the genes on the Y represent X gametologs or paralogs translocated from the autosomes, we used blastp (v.2.12.0)[39] to compare Y proteins against all proteins in our reference assembly, requiring a minimum Y protein coverage of >50% and at least >80% AA identity, excluding self matches. To test whether X contigs show elevated number of Y gametologs due to shared XY ancestry, we compared the number of Y transcripts for which we exclusively find

gametologs on the X, to the number of Y transcripts with exclusively paralogs on the autosomes, while accounting for difference in length or number of transcripts between autosomes and the X.

Y amplicons

The mammalian Y chromosomes contain large ampliconic regions enriched with high-identity segmental duplications e.g.[14]. Given that there seem to be a lot of duplicated genes also on the *C. maculatus* Y, we used blastn (v. 2.12.0)[39] to blast all concatenated CDS for each transcript on the Y contigs against each other with stringent requirements of >95% query coverage and >99.9% sequence identity and excluding self matches, to detect amplicon groups on the Y.

Characterisation of the Y_S and Y_L haplotypes

Variant calling

We used minimap2 (v. 2.18-r1015)[94] to align reads to the reference allowing for up to 20% sequence divergence (asm20), as was recommended for HiFi reads. Duplicates have been marked with picard (v. 2.23.4)[89]. We then used DeepVariant (v. 1.3.0, default settings)[95] with model type PACBIO for HiFi reads to call variants and GLnexus (v. 1.4.1)[96] to merge the variant calling files for both Y haplotype lines. We used vcftools[97] to filter the VCF files to only get single nucleotide variants (SNV) with a minimum depth of 5 and an upper depth cut-off of 65. SNV were categorized into single nucleotide polymorphisms that are shared between both Y introgression lines (SNP) and single nucleotide differences that are fixed between the two Y-introgression lines (fixed SNV).

TOR candidate gene analysis

A copy of a conserved gene coding for *target of rapamycin* (TOR) was discovered in a putative Y contig identified in the study that sequenced *C. maculatus* genome for the first time[29]. TOR is well known for its role in regulating growth across taxa (reviewed in[41]) and is therefore a prime candidate to explain the male body size difference between the Y_S and Y_L haplotypes. We identified and curated TOR locations in the genome (see supplementary methods 'genome annotation' for full details). To compare the different TOR regions we used MAFFT (v. 7.407, with --ep 0 --genafpair parameters)[98] to align concatenated *TOR* CDS of each identified region in both assembled genomes and create a

guided gene tree. To assess whether there are differences in exonic and non-coding sequences divergence across the different *TOR* regions we used fully factorial pairwise alignment of exonic and non-coding sequences separately. For the non-coding alignment, we masked identified exonic regions with bedtools maskfasta and then aligned the non-coding sequences with masked exons via Mummer nucmer (v. 4.0.0 with --maxmatch --c 100 parameters). For the exonic alignment we used the same procedure but masking the non-coding regions instead. We visualized the exonic and non-coding alignments with the R package circlize[99].

Data availability

Data and code will be made available upon acceptance on Dryad/Github repository.

Acknowledgements

We thank Johanna Liljestrand-Rönn for help and advice with DNA extraction, James M. Howie for help to develop the molecular sexing protocol and Madeline Chase for computational advice and feedback on figures. We thank Martyna Zwoinska and Göran Arnqvist for valuable discussions. The authors would like to acknowledge support of the National Genomics Infrastructure (NGI) /Uppsala Genome Center and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure. Work performed at NGI / Uppsala Genome Center has been funded by RFI / VR and Science for Life Laboratory, Sweden. Computations were enabled by resources in projects SNIC 2022/5-83 & SNIC 2022/22-176 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973. This work was funded by the grants from the Swedish Research Council (grant no. 2019-05038) and Carl Trygger Foundation (grant no. CTS-18:163) to E.I. and a research grant from Nilsson-Ehle Endowments (grant no 146700723) to P.K.

Author contribution:

The study idea and general design was conceived by E.I. SATC, Repeat analysis and identifying *Drosophila* orthologs of Y-linked genes (incl. associated tables & figures) was done by R.A.W.W. Molecular sexing protocol was designed by K.P and E.I and optimized by K.P. Genome was assembled by C.T.R. and annotated by D.S. All other laboratory work, data analyses and figures were done by P.K. with input from E.I. The manuscript was written by P.K and E.I. with contributions from R.A.W.W. and D.S.

Declaration of no conflicting interests

The authors declare no competing interests.

References:

1. Schärer, L., Rowe, L. & Arnqvist, G. Anisogamy, chance and the evolution of sex roles. *Trends Ecol. Evol.* **27**, 260–264 (2012).
2. Parker, G. A. Sexual conflict over mating and fertilization: An overview. *Philos. Trans. R. Soc. B Biol. Sci.* **361**, 235–259 (2006).
3. Andersson, M. *Sexual seleciton*. (Princeton University Press, 1994).
4. Arnqvist, G. & Rowe, L. *Sexual Conflict*. (Princeton University Press, 2005).
5. Lande, R. Sexual Dimorphism , Sexual Selection , and Adaptation in Polygenic Characters. *Soc. Study Evol.* **34**, 292–305 (1980).
6. Bonduriansky, R. & Chenoweth, S. F. Intralocus sexual conflict. *Trends Ecol. Evol.* **24**, 280–288 (2009).
7. Mank, J. E. Small but mighty: The evolutionary dynamics of W and y sex chromosomes. *Chromosom. Res.* **20**, 21–33 (2012).
8. Rice, W. R. Sex chromosomes and the evolution of sexual dimoprhism. *Evolution (N. Y).* **38**, 735–742 (1984).
9. Charlesworth, D. Plant sex determination and sex chromosomes. *Heredity (Edinb).* **88**, 94–101 (2002).
10. Charlesworth, D. & Charlesworth, B. Sex differences in fitness and selection for centric fusions

between sex-chromosomes and autosomes. *Genet. Res.* **35**, 205–214 (1980).

11. Bachtrog, D. Y chromosome evolution: emerging insights into processes of Y chromosome degeneration. *Nat Rev* **14**, 113–124 (2013).

12. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **355**, 1563–1572 (2000).

13. Lenormand, T., Fyon, F., Sun, E. & Roze, D. Sex Chromosome Degeneration by Regulatory Evolution. *Curr. Biol.* **30**, 3001-3006.e5 (2020).

14. Skaletsky, H. *et al*. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).

15. Hughes, J. F. *et al*. Chimpanzee and human y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536–539 (2010).

16. Hughes, J. F. *et al*. Strict evolutionary conservation followed rapid gene loss on human and rhesus y chromosomes. *Nature* **483**, 82–87 (2012).

17. Soh, Y. Q. S. *et al*. Sequencing the mouse y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159**, 800–813 (2014).

18. Hughes, J. F. *et al*. Sequence analysis in Bos taurus reveals pervasiveness of X–Y arms races in mammalian lineages. *Genome Res.* **31**, 1716–1726 (2020).

19. Hall, A. B. *et al*. Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E2114–E2123 (2016).

20. Peichel, C. L. *et al*. Assembly of the threespine stickleback y chromosome reveals convergent signatures of sex chromosome evolution. *Genome Biol.* **21**, 1–31 (2020).

21. Sandkam, B. A. *et al*. Extreme Y chromosome polymorphism corresponds to five male reproductive morphs of a freshwater fish. *Nat. Ecol. Evol.* **5**, 939–948 (2021).

22. Mahajan, S., Wei, K. H., Nalley, M. J., Gibilisco, L. & Bachtrog, D. De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture. 1–28 (2018).

23. Prokop, J. W. & Deschepper, C. F. Chromosome Y genetic variants: Impact in animal models and on human disease. *Physiol. Genomics* **47**, 525–537 (2015).

24. Nielsen, T. M., Baldwin, J. & Fedorka, K. M. Gene-poor Y-chromosomes substantially impact male trait heritabilities and may help shape sexually dimorphic evolution. *Heredity (Edinb)*. 1–6 (2023). doi:10.1038/s41437-023-00596-8

25. Kaufmann, P., Wolak, M. E., Husby, A. & Immonen, E. Rapid evolution of sexual size dimorphism facilitated by Y-linked genetic variance. *Nat. Ecol. Evol.* **5**, 1394–1402 (2021).

26. McKenna, D. D. *et al*. The evolution and genomic basis of beetle diversity. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 24729–24737 (2019).

27. Blackmon, H. & Demuth, J. P. Estimating tempo and mode of Y chromosome turnover: Explaining Y chromosome loss with the fragile Y hypothesis. *Genetics* **197**, 561–572 (2014).

28. Blackmon, H., Ross, L. & Bachtrog, D. Sex determination, sex chromosomes, and karyotype evolution in insects. *J. Hered.* **108**, 78–93 (2017).

29. Sayadi, A. *et al*. The genomic footprint of sexual conflict. *Nat. Ecol. Evol.* **3**, 1725–1730 (2019).

30. Angus, R. B., Dellow, J., Winder, C. & Credland, P. F. Karyotype differences among four species of Callosobruchus Pic (Coleoptera: Bruchidae). *J. Stored Prod. Res.* **47**, 76–81 (2010).

31. Berg, E. C. & Maklakov, A. A. Sexes suffer from suboptimal lifespan because of genetic conflict in a seed beetle. *Proc. R. Soc. B Biol. Sci.* **279**, 4296–4302 (2012).

32. Berger, D., Berg, E. C., Widegren, W., Arnqvist, G. & Maklakov, A. A. Multivariate intralocus sexual conflict in seed beetles. *Evolution (N. Y).* **68**, 3457–3469 (2014).

33. Berger, D. *et al*. Intralocus Sexual Conflict and the Tragedy of the Commons in Seed Beetles. *Am. Nat.* **188**, E98–E112 (2016).

34. Arnqvist, G., Goenaga, J., Liljestrand-Rönn, J. & Immonen, E. Concerted evolution of meta-bolic rate, economics of mating, ecology and pace-of-life across seed beetles. *Proc. Natl. Acad. Sci.* **Accepted M**, (2022).

35. Kaufmann, P., Howie, J. M. & Imm. Sexually antagonistic selection maintains genetic variance when sexual dimorphism evolves. *Proc. R. Soc. B* **accepted**, (2023).

36. Ellis, J. A., Stebbing, M. & Harrap, S. B. Significant population variation in adult male height associated with the Y chromosome and the aromatase gene. *J. Clin. Endocrinol. Metab.* **86**, 4147–4150 (2001).

37. Singh, P., Taborsky, M., Peichel, C. L. & Sturmbauer, C. Genomic basis of Y-linked dwarfism in

633 cichlids pursuing alternative reproductive tactics. *Mol. Ecol.* (2023). doi:10.1111/mec.16839

634 38. Nursyifa, C., Brüniche-Olsen, A., Garcia-Erill, G., Heller, R. & Albrechtsen, A. Joint identification of
635 sex and sex-linked scaffolds in non-model organisms using low depth sequencing data. *Mol. Ecol.*
636 *Resour.* **22**, 458–467 (2022).

637 39. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).

638 40. Larkin, A. *et al.* FlyBase: Updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Res.*
639 **49**, D899–D907 (2021).

640 41. Wullschleger, S., Loewith, R. & Hall, M. N. TOR signaling in growth and metabolism. *Cell* **124**, 471–
641 484 (2006).

642 42. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison.
643 *BMC Bioinformatics* **6**, 1–11 (2005).

644 43. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *bioRxiv* 2022.12.01.518724 (2022).
645 doi:10.1101/2022.12.01.518724

646 44. Arnqvist, G. *et al.* Genome size correlates with reproductive fitness in seed beetles. *Proc. R. Soc. B*
647 *Biol. Sci.* **282**, 11–14 (2015).

648 45. Hoskins, R. A. *et al.* The Release 6 reference sequence of the Drosophila melanogaster genome.
649 *Genome Res.* **25**, 445–458 (2015).

650 46. Zhang, J., Luo, J., Chen, J., Dai, J. & Montell, C. The role of y chromosome genes in male fertility in
651 drosophila melanogaster. *Genetics* **215**, 623–633 (2020).

652 47. Bellott, D. W. *et al.* Sensitive Regulators. *Nature* **508**, 494–499 (2014).

653 48. Bellott, D. W. *et al.* Avian W and mammalian Y chromosomes convergently retained dosage-sensitive
654 regulators. *Nat Genet.* **49**, 387–394 (2017).

655 49. Murphy, W. J. *et al.* Novel gene acquisition on carnivore Y chromosomes. *PLoS Genet.* **2**, 0353–0363
656 (2006).

657 50. Janečka, J. E. *et al.* Horse Y chromosome assembly displays unique evolutionary features and putative
658 stallion fertility genes. *Nat. Commun.* **9**, (2018).

659 51. Berger, D. *et al.* Intralocus sexual conflict and environmental stress. *Evolution (N. Y).* **68**, 2184–2196
660 (2014).

661 52. Immonen, E., Sayadi, A., Bayram, H. & Arnqvist, G. Mating changes sexually dimorphic gene

662 expression in the seed beetle Callosobruchus maculatus. *Genome Biol. Evol.* **9**, 677–699 (2017).

663 53. Bagchi, B. *et al.* Sexual conflict drives micro- and macroevolution of sexual dimorphism in immunity.
664 *BMC Biol.* **19**, 1–19 (2021).

665 54. Carvalho, A. B. Origin and evolution of the Drosophila Y chromosome. *Curr. Opin. Genet. Dev.* **12**,
666 664–668 (2002).

667 55. Smeds, L. *et al.* Evolutionary analysis of the female-specific avian W chromosome. *Nat. Commun.* **6**,
668 (2015).

669 56. Rogers, T. F., Palmer, D. H. & Wright, A. E. Sex-Specific Selection Drives the Evolution of Alternative
670 Splicing in Birds. *Mol. Biol. Evol.* **38**, 519–530 (2021).

671 57. Rideout, E. J., Narsaiya, M. S. & Grewal, S. S. The Sex Determination Gene transformer Regulates
672 Male-Female Differences in Drosophila Body Size. *PLoS Genet.* **11**, 1–23 (2015).

673 58. Lemos, B., Araripe, L. O. & Hartl, D. L. Polymorphic Y chromosomes harbor cryptic variation with
674 manifold functional consequences. *Science (80-.).* **319**, 91–93 (2008).

675 59. Berta, P. *et al.* Genetic evidence equating SRY and the testis-determining factor. *Nature* **348**, 448–450
676 (1990).

677 60. Goodfellow, P. N. & Lovell-Badge, R. SRY and sex determinaton in mammals. *Annu. Rev. Genet* **27**,
678 71–92 (1993).

679 61. Paris, J. R. *et al.* A large and diverse autosomal haplotype is associated with sex-linked colour
680 polymorphism in the guppy. *Nat. Commun.* **13**, 1–15 (2022).

681 62. Ellegren, H. & Parsch, J. The evolution of sex-biased genes and sex-biased gene expression. *Nat. Rev.*
682 *Genet.* **8**, 689–698 (2007).

683 63. Ellison, C. & Bachtrog, D. Recurrent gene co-amplification on Drosophila X and Y chromosomes.
684 *PLoS Genet.* **15**, 1–23 (2019).

685 64. Brashear, W. A., Raudsepp, T. & Murphy, W. J. Evolutionary conservation of Y Chromosome
686 ampliconic gene families despite extensive structural variation. *Genome Res.* **28**, 1826–1840 (2018).

687 65. Bayram, H., Sayadi, A., Immonen, E. & Arnqvist, G. Identification of novel ejaculate proteins in a seed
688 beetle and division of labour across male accessory reproductive glands. *Insect Biochem. Mol. Biol.*
689 **104**, 50–57 (2019).

690 66. Jackson, J. A. & Fink, G. R. Gene conversion between duplicated genetic elements in yeast. *Nature*

292, 306–311 (1981).

67. Zhou, S. *et al*. Male-specific lethal 2, a dosage compensation gene of Drosophila, undergoes sex-specific regulation and encodes a protein with a RING finger and a metalllothionein-like cysteine cluster. *EMBO J*. **14**, 2884–2895 (1995).

68. Ashton-Beaucage, D. *et al*. The Deubiquitinase USP47 Stabilizes MAPK by Counteracting the Function of the N-end Rule ligase POE/UBR4 in Drosophila. *PLoS Biol*. **14**, 1–27 (2016).

69. Graze, R. M., Tzeng, R. Y., Howard, T. S. & Arbeitman, M. N. Perturbation of IIS/TOR signaling alters the landscape of sex-differential gene expression in Drosophila. *BMC Genomics* **19**, 1–16 (2018).

70. Perry, J. & Kleckner, N. The ATRs, ATMs, and TORs are giant HEAT repeat proteins. *Cell* **112**, 151–155 (2003).

71. Baretić, D., Berndt, A., Ohashi, Y., Johnson, C. M. & Williams, R. L. Tor forms a dimer through an N-terminal helical solenoid with a complex topology. *Nat. Commun*. **7**, 1–9 (2016).

72. Hennig, K. M. & Neufeld, T. P. Inhibition of cellular growth and proliferation by dTOR overexpression in Drosophila. *Genesis* **34**, 107–110 (2002).

73. Swer, P. B., Mishra, H., Lohia, R. & Saran, S. Overexpression of TOR (target of rapamycin) inhibits cell proliferation in Dictyostelium discoideum. *J. Basic Microbiol*. **56**, 510–519 (2016).

74. Vilella-Bach, M., Nuzzi, P., Fang, Y. & Chen, J. The FKBP12-rapamycin-binding domain is required for FKBP12-rapamycin- associated protein kinase activity and G1 progression. *J. Biol. Chem*. **274**, 4266–4272 (1999).

75. Shao, X. *et al*. Copy number variation is highly correlated with differential gene expression: A pan-cancer study. *BMC Med. Genet*. **20**, 1–14 (2019).

76. Baker, C. R., Hanson-Smith, V. & Johnson, A. D. Following Gene Duplication, Paralog Interference Constrains Transcriptional Circuit Evolution. *Science (80-.)*. **342**, 104–108 (2013).

77. Berger, D., Berg, E. C., Widegren, W., Arnqvist, G. & Maklakov, A. A. Multivariate intralocus sexual conflict in seed beetles. *Evolution (N. Y)*. **68**, 3457–3469 (2014).

78. Grieshop, K. & Arnqvist, G. Sex-specific dominance reversal of genetic variation for fitness. *PLoS Biol*. **16**, 1–21 (2018).

79. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

720 80. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies.
721 *Bioinformatics* **36**, 2896–2898 (2020).
722 81. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and
723 Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of
724 Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
725 82. Arnqvist, G. *et al*. A near chromosome-level assembly of the seed beetle Callosobruchus maculatus
726 genome with annotation of its repetitive elements. *Submitt. MS* (2023).
727 83. Smit, A., Hubley, R. & Green, P. RepeatMasker 4.0. (2015).
728 84. Hoff, K., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-genome annotation with BREAKER. in
729 *Gene Prediction* (ed. Kollmar, M.) 65–95 (Humana Press, 2019). doi:10.1007/978-1-4939-9173-0
730 85. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for
731 BRAKER. *BMC Bioinformatics* **22**, 1–12 (2021).
732 86. Wu, T. D. & Watanabe, C. K. GMAP: A genomic mapping and alignment program for mRNA and
733 EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
734 87. Kaufmann, P. *The evolution of sexual dimorphism and its genetic underpinnings*. *ACTA*
735 *UNIVERSITATIS UPSALIENSIS UPPSALA* (ACTA UNIVERSITATIS UPSALIENSIS UPPSALA,
736 2022).
737 88. Dobin, A. *et al*. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
738 89. Broad Institute. Picard Toolkit. *GitHub Repository* (2019). Available at:
739 http://broadinstitute.github.io/picard/.
740 90. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning
741 sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
742 91. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq
743 data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
744 92. Alexa, A. & Rahnenführer, J. topGO: Enrichment analysis for gene ontology. *Bioconductor*
745 *Improvments* (2022). doi:10.18129/B9.bioc.topGO
746 93. Brionne, A., Juanchich, A. & Hennequet-Antier, C. ViSEAGO: A Bioconductor package for clustering
747 biological functions using Gene Ontology and semantic similarity. *BioData Min.* **12**, 1–13 (2019).
748 94. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

749 95. Poplin, R. *et al*. A universal snp and small-indel variant caller using deep neural networks. *Nat.*
750 *Biotechnol*. **36**, 983 (2018).
751 96. Lin, M. F. *et al*. GLnexus: joint variant calling for large cohort sequencing. *bioRxiv* 343970 (2018).
752 doi:10.1101/343970
753 97. Danecek, P. *et al*. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
754 98. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: A novel method for rapid multiple sequence
755 alignment based on fast Fourier transform. *Nucleic Acids Res*. **30**, 3059–3066 (2002).
756 99. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular
757 visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
758

Figures & Tables:

Table 1| The number of SNP and fixed differences between the two Y introgression lines split by chromosome type. Note that contigs shorter than 100 kb are not categorized as either Y, X or A.

	SNP	fixed SNV	Total length Y_S [bp]	Length Y_S-Y_L coverage [bp]
A	2,802,495	7,737	1,153,104,983	1,143,254,074
X	4,806	0	58,557,095	58,489,645
Y	212	2,622	10,112,272	8,276,151
<100 kb	4,571	185	24,939,325	22,280,823

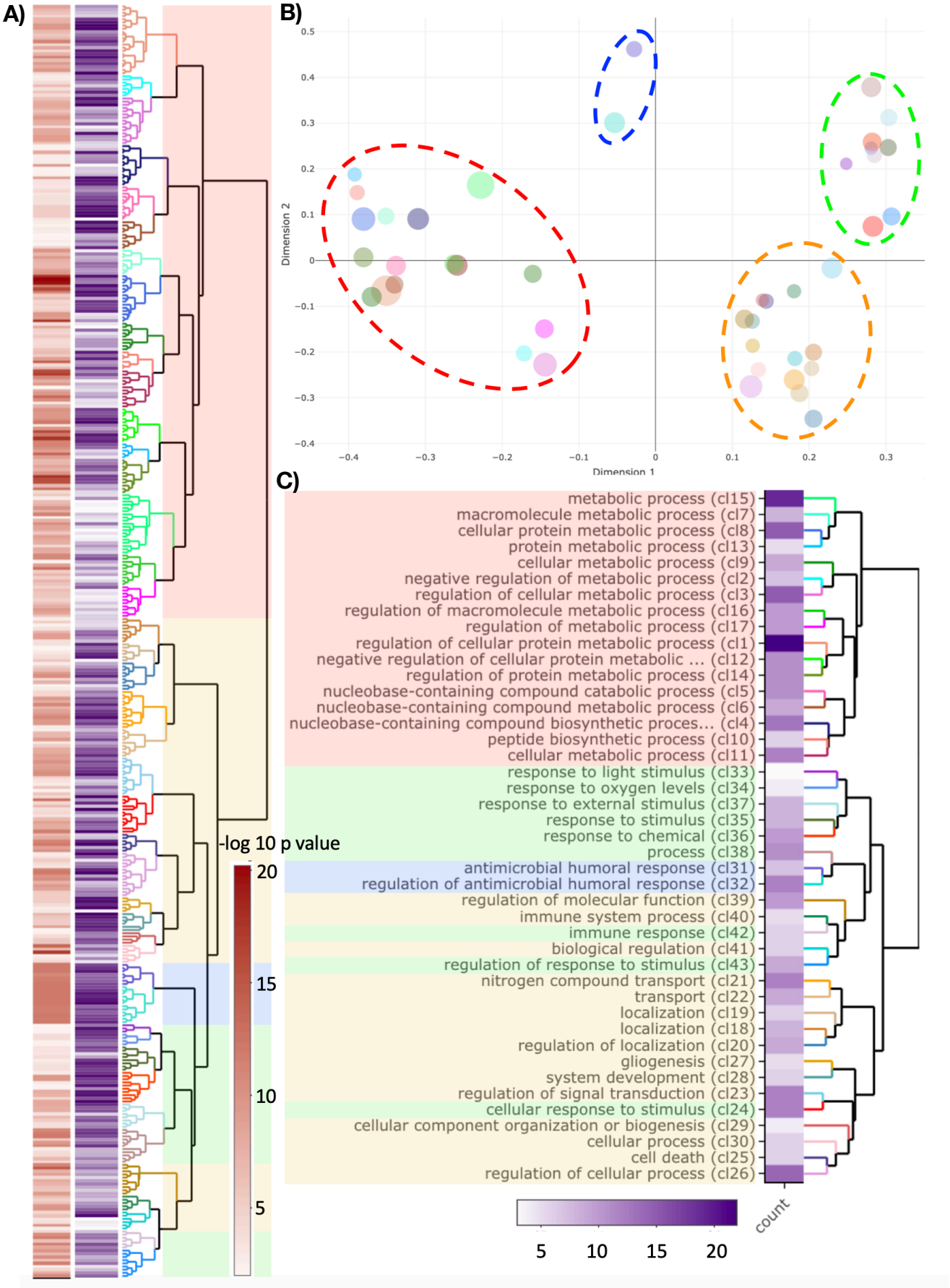


Fig. 1| Gene ontology enrichment for Y-linked genes with 'biological process' annotations. A) Dendrogram of GO terms based on Wang's semantic similarity distance, heatmap (red) indicates statistical significance as $-\log_{10}$ p-values (i.e. higher $-\log_{10}$ p-values have higher statistical support) and information content (purple). **B)** Multi-Dimensional Scaling (MDS) plot based on Best-Match Average (BMA) distance, representing the proximities of dendrogram clusters in (**A**). Dot size indicates the number of GO terms within each cluster. We highlight the four major functional groups with dashed ellipsoids (red ≅ metabolic processes; blue ≅ antimicrobial response; green ≅ response to stimulus; yellow ≅ development, cell organisation, growth & cell apoptosis) **C)** Dendrogram representation of clusters from (**B**) with GO term description of the first common GO ancestor and heatmap for the number of GO terms within each cluster.

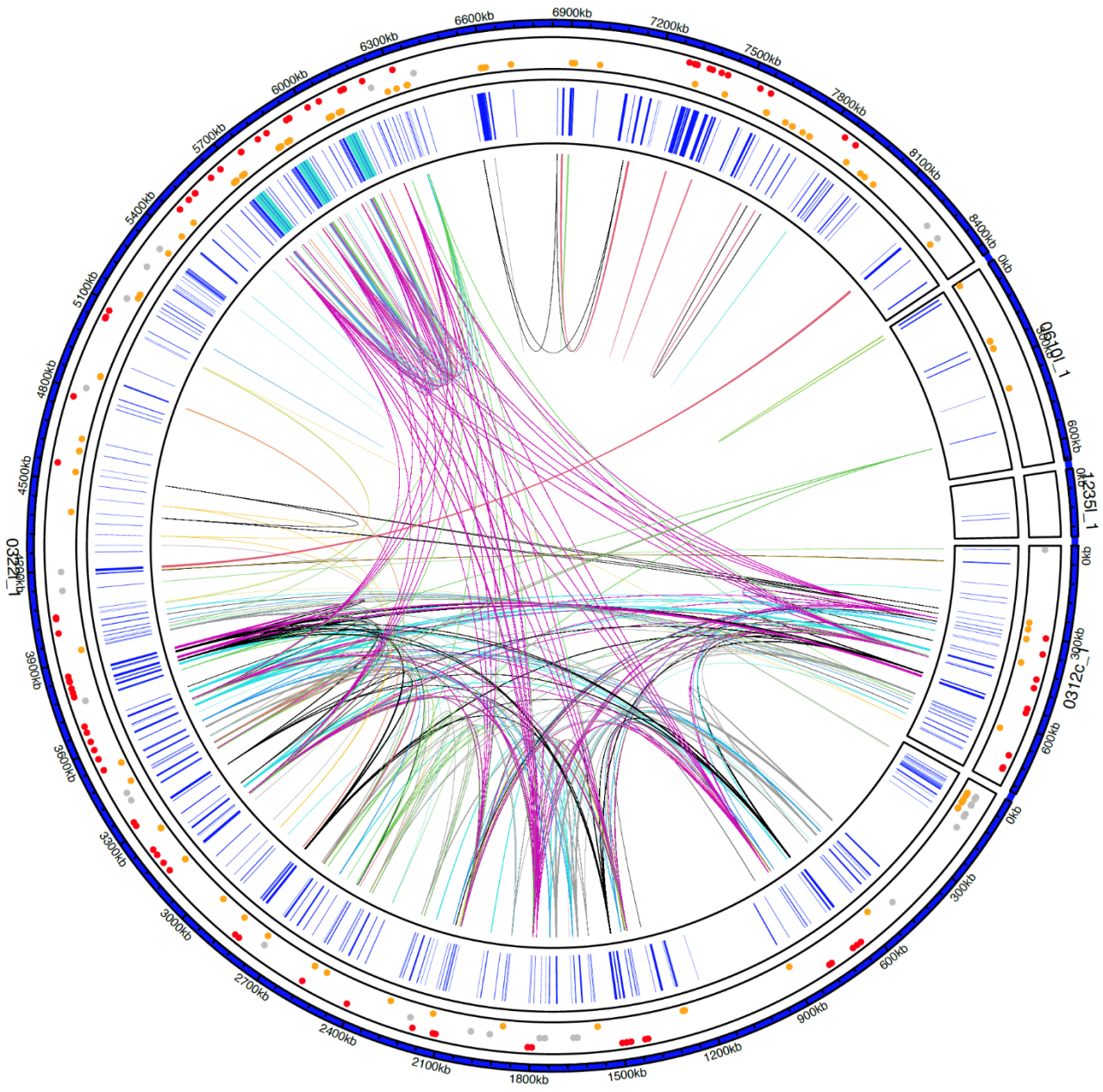


Fig. 2| Overview of Y genes. Each amplicon group (i.e. genes with >99.9% nucleotide sequence identity and >95% query coverage) is highlighted with a genomic link. 219 out of a total of 437 Y genes have at least one additional copy on Y. Gene positions of all Y-linked genes are shown in inner track in blue, regions containing *TOR* are highlighted in turquoise (see Fig. 4A for more details). The outer track indicates whether a gene has exclusively gametologs on the X (red, n=99), paralogs on the autosomes (yellow, n=157) or both (grey, n=73), dots are scattered by homolog type.

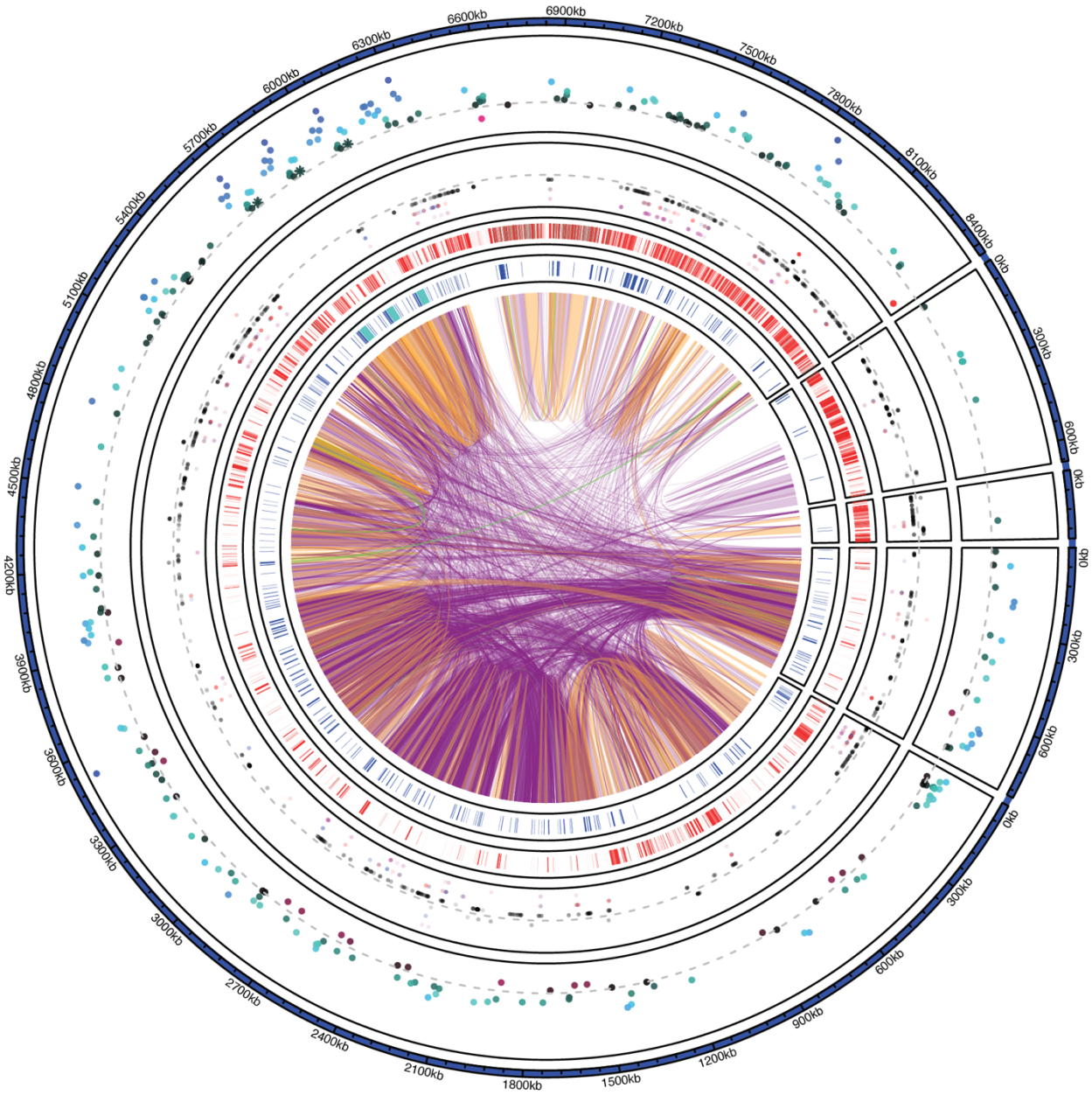
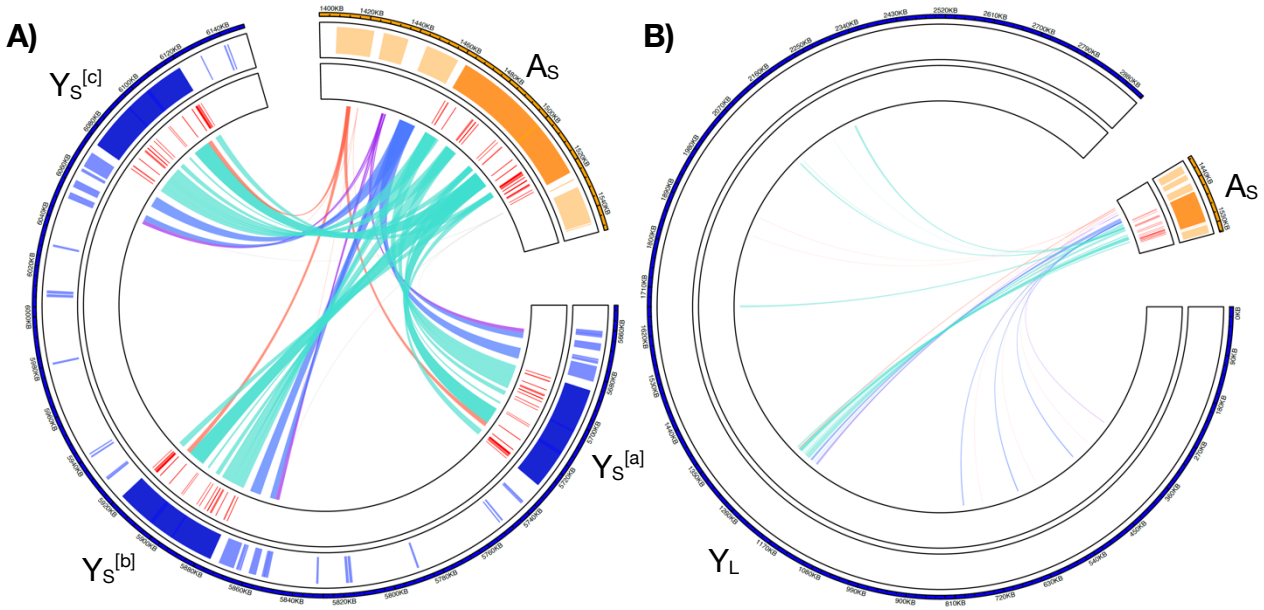


Fig. 3| Overview of Y contigs in the Y_S genome. Color coded genomic links show high level of nucleotide sequence similarity among and within Y contigs (purple > 99%, orange > 99.9%, green >99.99% nucleotide similarity matches > 1 kb in size). The inner track shows the position of annotated genes in blue, with the *TOR* regions highlighted in turquoise (Fig. 4). The second inner track highlights areas of low mapping coverage between the two Y haplotypes: areas shown in red indicate coverage lower than 17x (i.e. regions that may lack coverage for reliable variant calling via DeepVariant). The third track shows the position of single nucleotide variants (SNV), single nucleotide polymorphisms (SNPs) (outside of the dashed grey line) and fixed SNV (inside the dashed grey line). SNV in black (closest to the dashed grey line) are outside of gene regions, SNV in magenta (farthest from the dashed grey line) are within gene regions. SNV in blue and red are in 2 kb upstream or 2 kb downstream (proxy for cis-regulatory region of a gene) of a gene. The outer track shows

798 differential gene expression between males and females. Genes in blue, towards the outside are male
799 biased, genes in red, towards the center, are female biased (grey dashed line is shown as a reference
800 to indicate no difference in gene expression). *TOR* gene expressions are highlighted as asterisks and
801 are significantly male biased.

802



803

804 **Fig. 4| Zoom into the *TOR* region on the Y (blue) and on the autosome (yellow).** The inner track
805 shows the position of identified exons in red. Note that all Y linked *TOR* copies lack the exons 1-5
806 (from the 5' end), have one partial exon 6, all the other *TOR* exons (7-25) are present and complete.
807 **A)** Y_S haplotype, showing the three *TOR* regions (denoted with $^{[a]}$, $^{[b]}$ & $^{[c]}$). Note that we here only
808 present the relevant region of the whole utg0003221_1 Y_S contig (blue, see highlighted region in Fig.
809 2&3). **B)** Y_L contig containing the *TOR* region in the Y_L haplotype (blue, showing the whole contig)
810 and the autosomal *TOR* region (yellow, from the annotated Y_S assembly). In the Y_L haplotype we
811 identify only one Y-linked copy.

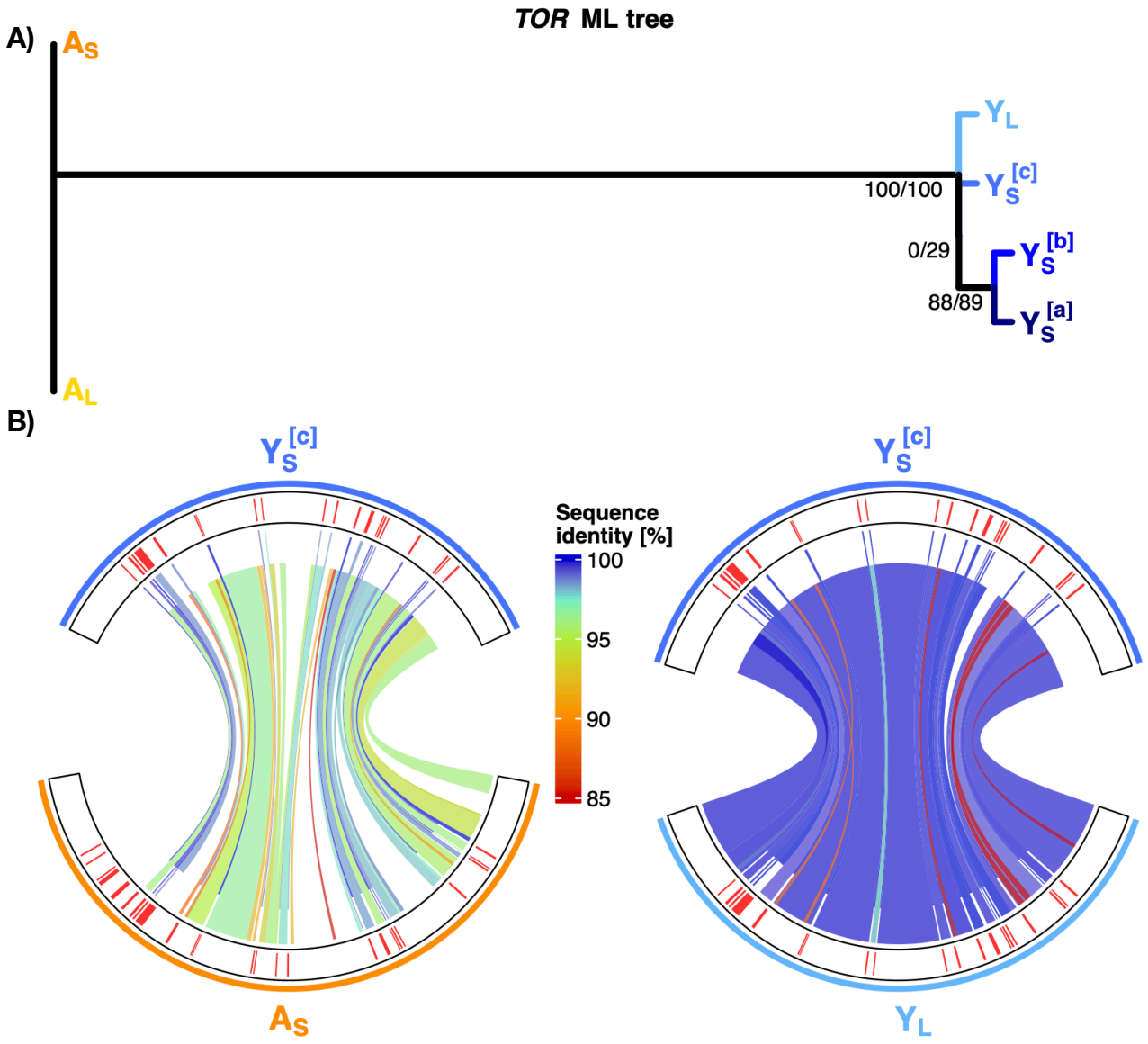


Fig. 5| *TOR* region genealogy and comparison. A) Maximum likelihood gene tree of exonic *TOR* sequences. Note that the A_S and A_L (highlighted in orange shades) have identical exonic sequences. High bootstrap values (100/100) indicate that Y-linked *TOR* (highlighted in blue shades) are more similar to each other than to the autosomal *TOR* and that $Y_S^{[a]}$ & $Y_S^{[b]}$ are sister sequences to each other (88/89). However, the clustering of the remaining Y-linked *TOR* sequences has low support (0/29). **B)** Nucleotide alignment with Mummer of exonic and non-coding regions separately. The outer track shows the *TOR* exon positions in red. Genomic links show nucleotide similarity, exonic links are shifted upwards, while non-coding alignments are shifted down to guide easier distinction between the two regions. Fully factorial pairwise comparison of all *TOR* regions is presented in Fig. S13. Left side: Comparison of the autosomal and the $Y_S^{[c]}$ *TOR* region. While there are structural

differences (gaps) and low sequence identity matches in the non-coding regions, the exonic regions seem conserved between the autosomal and the Y_S *TOR* copy. Note that all Y *TOR* regions lack the exon 1-5 (from the 5' end) and exon 6 is only partially present. Exon 7-25 are present in all Y *TOR* regions. Right side: Comparison of $Y_S^{[c]}$ *TOR* and the *TOR* region on the other Y haplotype (Y_L). We find high nucleotide similarity in both non-coding and exonic regions alike, with one structural difference in a non-coding region.