nature ecology & evolution

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41559-019-1041-9

In the format provided by the authors and unedited.

The genomic footprint of sexual conflict

Ahmed Sayadi[1], Alvaro Martinez Barrio[2], Elina Immonen[3], Jacques Dainat[4], David Berger[1], Christian Tellgren-Roth[5], Björn Nystedt[2] and Göran Arnqvist[1*]

[1]Animal Ecology, Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden. [2]Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [3]Evolutionary Biology, Department of Ecology and Genetics, Uppsala University, Uppsala, Sweden. [4]Department of Medical Biochemistry and Microbiology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [5]Department of Immunology, Genetics and Pathology, National Genomics Infrastructure, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. *e-mail: Goran.Arnqvist@ebc.uu.se

Supplemental Information

The genomic footprint of sexual conflict

Ahmed Sayadi, Alvaro Martinez Barrio, Elina Immonen, Jacques Dainat, David Berger, Christian Tellgren-Roth, Björn Nystedt and Göran Arnqvist

Content

SI MATERIALS AND METHODS

Genome sequencing and assembly

For genome sequencing and assembly, we first subjected a line of SI (South India) to five subsequent generations of inbreeding by propagating a single female mated to a full sib brother (F > 0.67). This inbred line (SI4) was subsequently used.

For long-read sequencing, whole-body genomic high-qulity DNA was extracted using a salt-ethanol precipitation protocol. Beetles were first gently macerated and placed in preparation buffer (100 mM NaCl, 10 mM Tris-HCl, pH = 8.0, 0.5% SDS) together with proteinase K, vortexed and incubated at 50^{o}C overnight. Samples were then frozen overnight. To precipitate DNA, we added saturated NaCl several times before adding 95% ethanol, and then spun the DNA into a pellet. The DNA pellet was suspended in TE buffer (pH = 7.6). DNA quality and quantity was assessed using NanoDrop, Qubit and Bioanalyzer, followed by fragment length assessment on an agarose gel. To obtain enough input material, DNA from 12 male individuals was pooled, and PacBio sequencing libraries were generated as previously described (Pendleton *et al.* 2015). Sequencing data was obtained from 72 SMRT cells on a Pacific Biosciences RSII. The genome was then assembled using FALCON v 0.4.2 (https://github.com/PacificBiosciences/FALCON/) with default parameters, based on the PacBio read data.

Data type	Read length	Sex	Pooled individuals*	Amount of raw data	Coverage after quality filtering**
PacBio	9,011 bp (average)	Male	12	35 Gbp	32X
Illumina	2 x 100 bp	Male	2	159 Gbp	125X
Illumina	2 x 100 bp	Female	2	146 Gbp	109X

Table S 1. Summary of genomic data used for assembly, error-correction and identification of candidate sex-lined contigs.

* For PacBio, DNA from multiple individuals was pooled before sequencing. For Illumina, datasets from multiple individuals were pooled using an internal script after sequencing.

** Assuming a male genome size of 1.1 Gbp and a female genome size of 1.2 Mbp.

For short-read sequencing, we extracted whole-body genomic DNA from four individuals (two males and two females) and prepared Illumina TruSeq sequencing libraries with a 350-400 bp insert size that we sequenced separately for each individual on an Illumina HiSeq2000, generating 2 x 100 bp paired-end reads. Illumina reads were end-trimmed with Cutadapt 1.2.1 and quality-filtered with Trimmomatic 0.3 before use, removing start and end bases of reads with a PHRED score below Q20, requiring sliding windows of 4 bp along the read to average at least Q20, and removing all reads with a remaining read length below 50 bp (Cutadapt parameters: *-O 15 -n 2*, Trimmomatic parameters: *LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:50*).

Assembly error-correction

The assembly was error-corrected by a single round of Quiver (SMART portal

2.3) based on re-alignement by the full set of PacBio reads. In addition to single-nucleotide and indel corrections, 857 low-quality contigs were also removed from the assembly by Quiver and only contigs larger than 500bp retained. Contigs under 500bp corresponded to just ~0.006% of the total primary assembly.

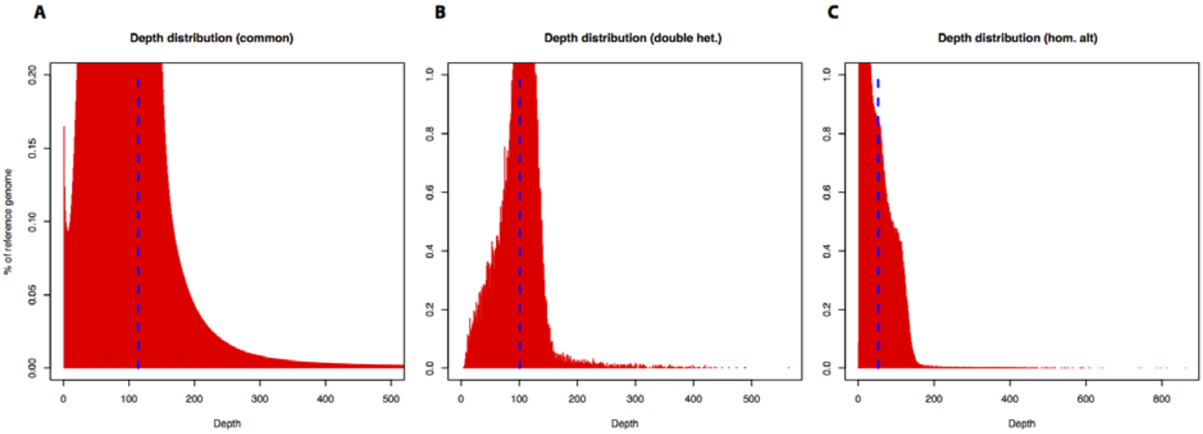

Figure S 1. (A) Depth distribution of all 6.85 million variants found by GATK Haplotype Caller. The spike on the left with a coverage below 6X likely corresponds to uncertain calls, while sites with a coverage above 175X are at high risk of representing uncertain calls at repeats. (B) Depth distribution of all ~36,000 double heterozygous sites. (C) Depth distribution all ~436,000 homozygous alternative sites. Blue dashed lines represent median coverages.

Further error-correction was done by alignment of deep (125X) quality-filtered Illumina male reads to the Quiver-corrected assembly, followed by identification of variants by GATK HaplotypeCaller as described here: after trimming, reads were aligned to the Quiver-corrected assembly using BWAmem 0.7.13, followed by indel realignment by GATK 3.3.0. A total of ~6.85 million variants (n=6,841,398; both SNPs and INDELs) were identified using GATK HaplotypeCaller with *sample_ploidy=2* (default), and the assembly was corrected using vcf-consensus from vcftools 0.1.14. Based on graphical inspection of the overall read coverage, we decided to not correct variants at sites with a read coverage below 6X and above 175X, to avoid false positive calls from insufficient coverage or low-mappability repeats (Figure S 1). In total, 393,703 homozygous alternative alleles were identified, and the assembly was corrected accordingly. In addition, 34,499 double heterozygous variants were identified and the assembly was corrected to the most commonly represented allele at each site. We note that among the homozygous alternative alleles, we found 66,698 deletions and 318,677 insertions, corresponding to the expected error distribution of PacBio data, with about 5-fold more insertions than deletions (Ross 2013, Bickhart 2017). Approximately 6.37 million variants were found in normal heterozygous positions and were left uncorrected.

Contamination screening

We screened the assembly to identify contigs potentially representing contaminating DNA in the sequencing sample. Blobtools (Kumar 2013) makes use of different databases to annotate high scoring matches to other organisms and presents the results in so-called "blob plots", identifying putative contamination contig sets based on homology, along with deviating patterns in read coverage and GC content. We used blobtools 0.9.19 with blastn 2.4.0+ against NCBI's nt database and Silva 5.0 for similarities against rRNA species. In addition, we used Diamond 0.7.12 to identify putative contaminant contigs based on Swissprot protein matches. Despite this extensive screen, no likely contaminations were identified in the assembly (Figure S 2), and no contigs were removed in this procedure.

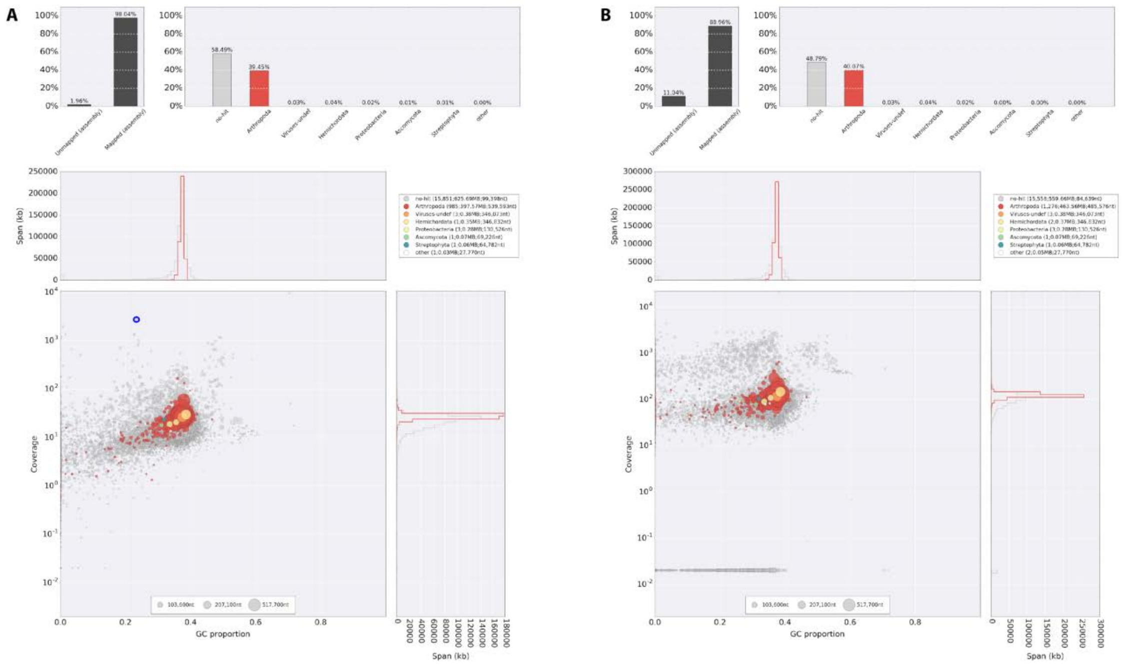

Figure S 2. Blobplots of all contigs in the Cmac assembly, with the GC-content (x-axis) plotted against the average read coverage in a log10 scale (y-axis). The size of the blobs represents the contig size, and the color denote the phylum matched in the NCBI taxonomy database. Size histograms of the two main categories are extracted per axis. The read coverage was computed either from the PacBio reads aligned against the primary, alternative and mtDNA Cmac assemblies (A), or from the Illumina reads aligned against the Cmac primary assembly (B). The blue circle in panel A indicates a Cmac mitochondrial contig.

Sex chromosome identification

Candidate sex-chromosome contigs were identified by comparing read coverage between male and female samples. Quality-filtered Illumina reads from male and female samples (see above) were mapped separately to the final assembly, and the median male and female read coverage was calculated, and normalized to represent 40X coverage of each (covX and covY, respectively). Manual inspection identified a vast majority of contigs to have an approximately 1-to-1 ratio between male and female coverage (Figure S 3). We used the following rationale to identify sex-linked contigs. A distinct subset of contigs showed approximately twice as high read coverage in female samples as in male samples, consistent with X-linkage (covY = 100/covX + covX/0.75). For Y-linked contigs, where a 50% male coverage and no female coverage is expected, fewer clearly identifiable contigs were present, and an *ad hoc* cutoff for very low female coverage was applied (covY = -3 + covX/2).

Using a combination of flow cytometry data for males and females (Arnqvist *et al*. 2015) and chromosome length determination from karyotype smears (Angus *et al*. 2011), we estimate the true size of the sex chromosomes to be X ≈ 93 Mb and Y ≈ 18 Mb. Here, we used the cut-off functions above to delineate candidate sex-linked contigs, which resulted in 1109 putative X-linked contigs with a total size of 27.6 Mbp and 167 putative Y-linked contigs with a total size of 3 Mbp (Figure S 3). We note that the sex-chromosomes are likely to be particularly rich in repeats and the list of contigs here regarded as candidate sex-linked contigs should be seen as an incomplete and preliminary representation of sex-linkage which may nevertheless be useful for investigating overall global features and trends.

Repeat annotation

RepeatMasker 4.0.5 was run on the assembly, identifying a very high fraction of repeats (64%) (Table SI 1). More than half of the repeated sequences (54%) could not be attributed to any specific repeat class by RepeatMasker, reflecting long evolutionary distances to previously known repeats.

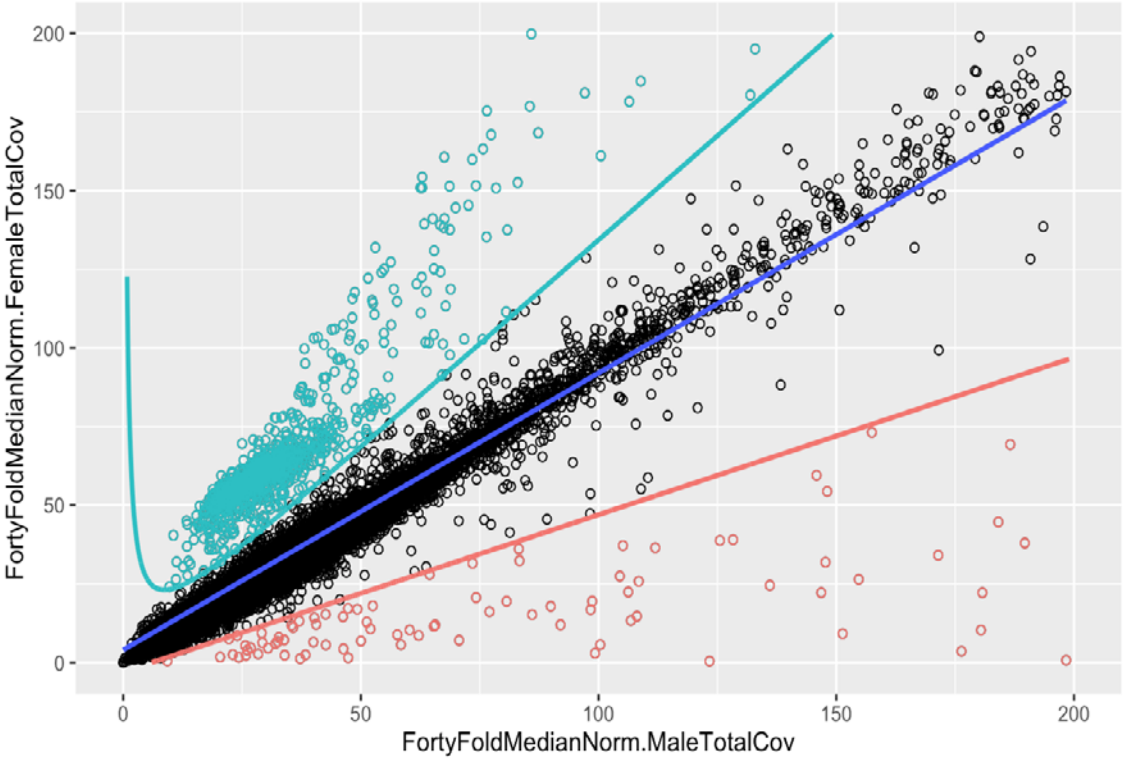

Figure S 3. Normalized median coverage of male (x-axis) and female (y-axis) samples per contig (circles). Black indicates putative automsomal contigs, magenta indicates putative X-linked contigs, and red indicates putative Y-linked contigs. A linear regression of the autosomal contigs is indicated in blue, and functions used to delineate candidate X- and Y-linked contigs are indicated in magenta and red, respectively. Here, the axis range has been restricted to 0-200x, to improve resolution at the lower part of the range of coverage for illustrative purposes.

Table S 2. Classification of annotated repeat content by RepeatMasker.

Repeat type	Number of elements	Total size (Kbp)	Fraction of genome (%)
DNA	128 138	93905	9.3%
LINE	215 257	153 343	15.2%
SINE	10 004	3 340	0.3%
LTR	18 188	19 864	2.0%
RC	7 183	4 214	0.4%
rRNA	764	983	0.1%
snRNA	687	438	0.0%
Satellite	3 834	3 516	0.3%
Simple	220 252	13 371	1.3%
Low-complexity	25 184	1 325	0.1%
Unknown	470 879	348 596	34.5%
TOTAL	**1 100 370**	**642 897**	**63.7%**

Gene annotation

The genome annotation service at the National Bioinformatics Infrastructure Sweden (www.nbis.se) carried out the genome annotation using MAKER3 (Holt and Yandell, 2011), as detailed below.

A species-specific repeat library was first built using RepeatModeler 1.0.8 (Smit and Hubley, 2010). Candidate repeats modelled by RepeatModeler were vetted against our protein set (excluding transposon proteins) to avoid masking nucleotide motifs stemming from low-complexity coding sequences. From the repeat library, identification of repeat sequences present in the genome was performed using RepeatMasker 4.0.5 (Smit *et al.*, 2010) and RepeatRunner (http://www.yandell-lab.org/software/repeatrunner.html). RepeatRunner is a program that integrates RepeatMasker with BLASTX, allowing the analysis of highly divergent repeats and divergent portions of repeats and identifying divergent protein coding portions of retro-elements and retroviruses not detected by RepeatMasker.

To guide the annotation with extant transcriptome evidence, ten paired-end strand specific libraries were individually assembled using Tophat2 2.0.9(Kim *et al.*, 2013) and Stringtie 1.2.2 (Mihaela Pertea, 2015) and a de novo transcriptome assembly of normalized merged samples was performed using Trinity (Grabherr *et al.*, 2011; Sayadi *et al.* 2016).

A first round of annotation was performed with MAKER3 using the following evidence data: i) Proteins from the Uniprot-Swissprot database; ii) transcripts from the reference-guided and *de novo* transcriptome assemblies (see above). The evidence-based gene build resulted in a first “release candidate” gene set (rc1) with 18,551 gene models and 32,349 mRNAs predicted. For each gene model, MAKER3 also assigned an Annotation Edit Distance (AED), quantifying the congruency between a gene annotation and its supporting evidence. The AED vary between 0.0 (fully supported by evidence data) and 1.0 (no evidence).

The evidence-based annotation is limited by the available sequence data, which can lead to fragmented gene models and missed genes. To prevent this from happening, we next performed an *ab initio* evidence-driven gene build, where protein and transcript evidence is used to help and guide *ab initio* tools during their prediction processes. From the first evidence-based gene build (rc1), we selected a high-confidence set of genes based on the following criteria: i) The genes have to be complete (i.e. start/stop codons mandatory), ii) the AED scores have to be below 0.3, iii) the genes have to be at a distance of at least 500 pb from each other, and iv) no similarity over 85% is allowed among the genes in the set. This filtering resulted in a set of 4,366 non-redundant high-confidence gene models, which were used to train the *ab initio* tools Augustus 2.7 (Stanke *et al.*, 2006) and Snap 2006-07-28 (Korf, 2004). We also trained GeneMark-ET 4.3 (Lomsadze *et al.*, 2014), which is a self-trained method integrating RNA-seq evidence using the *junctions.bed* file from Tophat. The *ab initio* evidence-driven annotation was performed with MAKER3, using both the output HMM-models from the trained *ab initio* tools (Augustus, Snap, and Genemark-ET), and the same evidence data as used previously. We also used EVidenceModeler (EVM) (Haas *et al.*, 2008), which allowed us to perform gene models based on the best possible set of exons produced by the other *ab initio* tools, and choose the most consistent according to the available evidence. The *ab initio* evidence-driven gene build (rc2) contained 20,564 gene models and 34,331 mRNAs.

Finally, all evidence-based gene models (from rc1) that mapped within an empty locus in the *ab initio* evidence-driven annotation (rc2), was added to rc2, to create a final build (rc3), containing 21,264 gene models and 35,160 mRNAs.

For the final gene build (rc3), we inferred putative functions for all coding mRNAs. To this end, we first predicted functional domains using InterProscan 5.7-48 (Jones *et al.*, 2014) to retrieve functional information from Interpro (Hunter *et al.*, 2012), PFAM (Finn *et al.*, 2014), GO (Ashburner *et al.*, 2000), MetaCyc (Caspi *et al.*, 2014), UniPathway (Morgat *et al.*, 2012), KEGG (Kanehisa *et al.*, 2014) and Reactome (Croft *et al.*, 2014). In order to assign protein and gene names to this dataset, we performed a BLASTp 2.2.28+ search with each of the predicted protein sequences against the Uniprot-Swissprot reference data set with an e-value cut-off at 1x10^{-6}. Functional annotations were assigned to 11,997 of the predicted genes and gene names were assigned to 11,127 of the predicted genes.

In addition to rc3, 6,948 tRNA genes were annotated through *tRNAscan* 1.3.1.

The annotated genome assembly, along with sequence data, is available from the European Nucleotide Archive (ENA) under accession PRJEB30475.

Assembly evaluations

Assembly and annotation evaluations were performed with a set of 248 universally conserved proteins as implemented in CEGMA v2.5 (Parra *et al.*, 2009) against the entire assembly, and a set of 2675 conserved arthropod proteins as implemented in BUSCO v1.1b1 (Simão *et al.*, 2015) against the gene build from the annotated assembly (Table S 3).

abdomen males and females (Log2FC values) (Immonen *et al*. 2017).

Table S 3. Assembly evaluation scores.

	Complete proteins	Partially complete proteins	Fraction of duplicated complete proteins
CEGMA (n=248)	212 (85%)	20 (8%)	15%
BUSCO (n=2675)	2027 (75%)	283 (11%)	33%

Gene sets

To gain information on the sex-specificity of gene expression, we leaned on the assembled transcriptome of *C. maculatus* (Sayadi *et al*. 2016) and used data from a previous study on transcript abundance in males and females (Immonen *et al*. 2017). We first blasted all transcriptome-assembled genes expressed in the abdomen of beetles against the longest CDSs annotated in the genome. We used strict parameters in the blast, such that only hits with a P-value < 10^{-6}, a sequence identity > 70% and a sequence coverage > 50% were retained. We blasted 12,412 expressed genes out of which 6,711 genes were recovered in the genome. This reduction is primarily due to our strict blast parameters but we also note that the transcriptome was assembled *de novo* and we therefore expect some dissimilarity between predicted genes in the genome and genes assembled from transcript data. We also discarded cases where several transcriptome genes mapped to the same CDS in the genome, as being ambiguous in terms of its expression. In the end, we retained 4,993 CDSs for further analysis. These represent CDSs where confident and unambiguous information on gene expression was available. Each of these CDSs was then associated with the corresponding degree of sex-biased expression based on transcript abundance in the

We also analyzed several distinct gene sets; enzymes involved in digestion of food in larval guts, male seminal fluid proteins, candidate female reproductive proteins, candidate Y-linked genes and candidate X-linked genes. Sex-linked genes were identified as CDSs residing on candidate sex-linked contigs (see above). There were 658 X-linked and 281 Y-linked genes.

Gene sequences annotated as digestive enzymes in Bruchid beetles were collected from several sources (Pauchet *et al*. 2010, Zhu-Salzman *et al*. 2003, Moon *et al*. 2004, Pedra *et al*. 2003, Chi *et al*. 2009, Guo *et al*. 2012, Wang *et al*. 2015). They were first manually checked and re-annotated using Blast2GO. We then removed redundancy from the collected sequences using CD-HIT at 100% sequence identity. In total we obtained 2137 gene sequences. We used this gene list to do a reciprocal blast against our genome to identify candidate digestive enzymes. Blast parameters were p-value < 10^{-6}, sequence identity > than 70% and sequence coverage > than 50%. This yielded a final list of 741 genes annotated as digestive enzymes.

Male seminal fluid proteins and candidate female reproductive proteins, the latter representing genes that are (1) expressed in the female reproductive tract and (2) upregulated there following mating, were identified using proteomic methods as reported in Bayram *et al*. (2017) and Bayram *et al*. (2019). We did a reciprocal blast against our genome to identify corresponding genes, using the following blast parameters: p-value < 10^{-6}, sequence identity > than 70% and sequence coverage > than 50%. In total, we were thus able to identify 185 genes for male seminal fluid proteins and 126 genes for candidate female reproductive proteins.

PoolSeq analyses

We extracted high-quality DNA samples from pools of individuals, using a salt-ethanol precipitation protocol. Beetles were first gently macerated and placed in preparation buffer (100 mM NaCl, 10 mM Tris-HCl, pH = 8.0, 0.5% SDS) together with proteinase K, vortexed and incubated at 50°C overnight. Samples were then frozen overnight. To precipitate DNA, we added saturated NaCl several times before adding 95% ethanol, and then spun the DNA into a pellet. The DNA pellet was suspended in TE buffer (pH = 7.6). DNA quality and quantity was assessed using NanoDrop, Qubit and Bioanalyzer, followed by fragment length assessment on an agarose gel.

We prepared two independent samples from each of the three populations (N = 6 samples), each sample consisting of a pool of N=100 males. Sequencing libraries were prepared from 1μg DNA for each sample, using the TruSeq PCRfree DNA sample preparation kit (cat# FC-121-3001/3002, Illumina Inc.) targeting an insert size of 350bp. The library preparation was performed according to the manufacturers' instructions (guide#15036187). Library preparation using TruSeq PCRfree DNA library preparation kit is an accredited method.

Libraries were then subjected to cluster generation and sequencing in 3 lanes using the Illumina HiSeq2500 system, paired-end 125bp read length and v4 sequencing chemistry. We sequenced on average some 300 million read pairs for each library, resulting in an average coverage per sample of about 62X and an average cover per population of about 125X.

The Popoolation and PoPoolation2 pipelines (Kofler *et al.* 2011a, 2011b) were then used to identify SNPs in our sequence data. The pipeline involved several stringent filtering steps, to avoid false SNPs. First, before identifying SNPs, read quality was assessed using FastQC software (Andrews 2015). Low quality reads with potential sequencing errors were removed. Bases with a phred quality threshold lower than 20 were trimmed. Reads shorter than 50 bp were discarded, and only reads with mates were used for the next step. Few reads were removed during the trimming step, which reflects the high quality of the sequence data. Trimming was done using the Popoolation script trim-fastq.pl. On average, we retained more than 280 million reads for each sample resulting in an average coverage of > 35x per sample after the trimming step.

Second, cleaned reads were then mapped to the reference genome using BWA aln (Li and Durbin 2009), with defaults parameters as recommended in the Popoolation pipeline (allowing gaps [12bp maximum length of insertion/deletion] and a maximum of 10% mismatches). On average, some 52% of the paired-end reads were properly mapped back to the genome (Table S 4).

Third, following the mapping step, mapped reads were cleaned by removing duplicated and ambiguously mapped reads. Duplicated reads represent errors that could be introduced by the Illumina technology. This step was done using Picard tools (http://broadinstitute.github.io/picard/). Ambiguously mapped reads can result in erroneous SNP prediction. Thus, only reads that mapped as proper pairs and with a mapping quality score >20 were retained. This was done using SAMtools (Li *et al*. 2009). As a result, we retained on average 37% of the total number of trimmed reads. This reflects the rigorous parameters used, aiming to avoid falsely mapped reads and as false prediction of SNPs.

Fourth, before calling SNPs, all aligned reads from all samples were first grouped to a single file using SAMtools mpileup utility (Li *et al*. 2009). The mpileup file provides a summary of all allele counts in all samples. Two scripts (identify-genomic-indel-regions.pl and filter-pileup-by-gtf.pl) provided by Popoolation were then used to identify SNPs and to discard SNPs surrounding indels, aiming to avoid false SNPs. Fifth, we only included SNPs that occurred in regions with 10X to 500X total coverage, that were present in at least 6 reads and that showed a base quality >20. In the end, this resulted in a total number of SNPs of 5,045,210, of which 167,168 were located within CDSs. The number of CDSs with ≥1 SNP was 12,136.

Autosomal and X/Y-linked regions differ in predicted coverage in our data, as males are hemizygous for sex-linked contigs. To assess whether differences in coverage affected the comparison of autosomal and sex-linked sites, we subsampled the mpileup file to 10X and to 20X coverage depth per site. To achieve this, the mpileup file was first converted to a sync file using the script 'mpileup2sync.jar' and then subsampled using the script provided by Popoolation2 software; 'subsample-synchronized.pl' with the option '- - method fraction'. This then allowed a conservative comparison between autosomal regions subsampled at 10X with sex-linked regions sampled at 20X.

We extracted estimates of Tajima's D, nucleotide diversity and p_N/p_S ratios using the scripts 'Variance-at-position.pl' and 'Syn-nonsyn-at-position.pl' respectively. Additional parts of the analyses (i.e., SNP density, polymorphic SNPs) were done using in-house Perl scripts.

All Pool-seq raw sequencing data have been deposited at the NCBI sequence read archive, under the accession number PRJNA503561.

Custom scripts have been published at GitHub where they are openly and freely available at: https://doi.org/10.5281/zenodo.3382061

Functional enrichment

To identify overrepresentation of Gene Ontology terms in the three different categories (Biological process, Cellular process and Molecular function) we used a hypergeometric test with a P-value cutoff < 0.05 implemented in the GOstats package v.2.46.0 (Falcon and Gentleman, 2007). Gene universes varied in different tests and are explicitly defined in Tables S 6 – 7.

SI RESULTS

Modelling SBG expression as a continuous variable revealed a pattern very similar indeed to that based on analyses of bins. For example, Tajima's D tended to describe a wave-shaped pattern when related to gene bias in expression, with weakly-intermediately FBGs showing overall positive values (Figure S 4).

The analyses of functional gene sets (Figure S 5) indicated that digestive enzyme, male reproductive protein and female reproductive protein genes generally showed a history of overall purifying selection (relatively low p_N/p_S), and there was little general evidence for current selection based on Tajima's D. A possible exception was the set of 185 seminal fluid protein genes in the C population, which showed a significantly higher Tajima's D than non-seminal fluid genes (separate variance *t*-tests: D_{NS}, $P = 0.007$, $P_{\text{boot}} = 0.009$; D_S, $P = 0.002$, $P_{\text{boot}} = 0.007$) and the average D_S for seminal fluid protein genes was significantly higher than zero (*t*-tests: P=0.003, $P_{\text{boot}} = 0.003$) in this population. This suggests that seminal fluid proteins genes are under balancing selection in at least one of the populations. Unsurprisingly, seminal fluid proteins show strong male-bias in expression (mean Log2FC = -5.723, SE = 0.481). None of these three gene sets, however, showed any obvious overrepresentation in terms of their contribution to shared intermediate frequency polymorphism (Figure S 7).

Genes located on candidate X- and Y-linked contigs showed the hallmarks of their lower effective population size and recombination rate: low SNP density, low nucleotide diversity and strong purifying selection (negative overall Tajima's D) (Figure S 5). This pattern was not the results of unequal coverage in sequence data, rendering rare variants on sex-chromosomes to be less likely to be represented in our pool of sequence reads, as subsampling X- and Y-linked sites to account to differences in coverage relative to autosomal sites generated the same basic pattern (Figure S 6). Sex-linked genes also had a markedly lower probability of showing shared intermediate frequency polymorphism across the three populations (Figure S 7).

Mean expression of X-linked genes was somewhat higher in females than in males, although not twice as high as would be expected in the absence of dosage compensation/inactivation (mean logFC = 0.61; N = 54) and the average degree of sex-bias was not significantly different in X-linked and autosomal loci (permutation test; P = 0.164). This strongly suggests that partial dosage compensation and/or female X-inactivation is occurring.

To further test for enrichment of SA loci on the X, we first asked whether the ratio of X-linked to autosomal genes was different across the 8 classes of sex-bias in genes expression. A test of this possibility showed no significant difference ($\chi^2_7 = 10.96$, P = 0.140). We then tested whether genes with male-limited (Log2FC < -5) or female-limited (Log2FC > 5) expression were significantly overrepresented on the X-chromosome relative to the autosomes. This was not the case for either male-limited (Fisher's exact test; P = 0.126) or female-limited (Fisher's exact test; P = 0.195) genes.

We note that X-linked genes were included in the overall analyses presented (e.g. Figure 1), but stress that they only made up 1% of all expressed genes analyzed here. Whether X-linked genes were included or not in our overall analyses had a very marginal quantitative effect indeed on our findings and had no qualitative effects whatsoever on our inferences (in terms of e.g. our ability or inability to reject null hypotheses).

To test the hypothesis that shared expression across tissues is different in the set of 149 candidate SA loci compared to the genome as a whole, we first derived the expected degree of shared expression among all genes expressed in the abdomen or the head and thorax of adult beetles from data in Immonen *et al.* (2017). For a gene to be regarded as expressed in a given tissue, it needed to be expressed at a level of >3 cpm in at least 3 different samples of that tissue. Using this criteria, 79% of all genes showed shared expression. The degree of shared expression among the 149 candidate SA loci was 92%, which is significantly higher than expected ($\chi^2_1 = 15.05$, $P = 0.0001$).

Because we lack a recombination map of the *C. maculatus* genome, it is not possible to compensate for variation in recombination rate across the genome in our analyses. To assess the potential impact of linked selection, we inspected the distribution of genes showing intermediate frequency polymorphism (hence IFP) across contigs. First, gene richness correlated well with presence of IFPs, as is expected if genes with IFP are randomly distributed across the genome. Across all 6,717 contigs, the total number of CDSs in each contig correlated (Goodman-Kruskal's rank correlation) well with both (1) the number of CDSs showing IFP in any of the three populations ($r_\gamma = 0.69$) and (2) the number of CDSs showing IFP in all three populations ($r_\gamma = 0.68$). When restricted to include only contigs with non-zero IFP, these association were $r_\gamma = 0.76$ ($N = 2339$) and $r_\gamma = 0.58$ ($N = 713$). Second, and more importantly, we assessed whether certain contigs were enriched with genes showing IFPs, as would be expected if linked selection affected our results. For each contig, we asked whether the ratio between the number of CDSs with IFP in that contig to the total number of CDSs with IFP in all contigs was different than the ratio between the number of CDSs in that contig to the total number of CDSs in all contigs, using Fisher's two-tailed exact tests. Linked selection would result in an over- or underrepresentation of CDSs with IFPs in a number of contigs, where linked selection would result in these two proportions differing. We then applied FDR correction with a permissive cutoff at 0.25. We found that none of the 6,717 contigs was significant (at $Q < 0.05$) for CDSs showing IFP in any of the three populations. Four out of 6,717 were significant (at $Q < 0.05$) for CDSs showing IFP in all three populations. These four contigs were all autosomal and were enriched with genes showing IFP (contig#, total number of CDSs harbored : number of CDSs showing IFP in all three populations: #3031, 10:8; #28, 26:10, #108, 23:9, and #247, 12:7). Hence, these analyses are consistent with linked selection having at most a marginal effect on the distribution of segregating SNPs, and only so in a very restricted part of the genome. We interpret this as strongly suggesting that linked selection is at most of minor importance for the genome wide patterns documented in our analyses.

To better characterize the properties of candidate SA genes, we performed Gene Ontology (GO) enrichment analyses of the 149 candidate loci (showing Log2FC > 1 and $D_{ns} > 0$ and $D_s > 0$ in all three populations) against all genes expressed in the female abdomen and against all female biased genes expressed in the female abdomen. These analyses showed significant enrichment for genes involved in a variety of (1) general metabolic processes, (2) organelle (e.g. mitochondrial) organization and (3) cell division and egg production (Table S 6). This was also reflected in a more stringent outlier detection, selecting a gene set that showed a signal of strong balancing selection in all three populations ($D_{ns} > 2$ or $D_s > 2$). There were 12 genes in this latter gene set, 10 of which showed significant homologies with annotated genes. Three represented genes involved in DNA repair and cell division/differentiation, key processes in oogenesis. Another four showed significant homologies with general metabolic genes. For example, one matched a LYR motif protein gene. LYR proteins interact with the oxidative phosphorylation (OXPHOS) core complexes in mitochondria, thus directly affecting ATP production. Yet another gene matched a nicotinate phosphoribosyltransferase and this gene resides on an X-linked contig in C. maculatus. This is a fundamental metabolic enzyme which is also involved in ATP production, as it catalyzes the rate-limiting step of nicotinamide adenine dinucleotide (NAD) biosynthesis, and its expression is known to affect life history traits such as life span in other species (Berger *et al*. 2004).

SI REFERENCES

Andrews, S. FastQC: A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).

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 (2011).

Arnqvist, G. *et al*. Genome size correlates with reproductive fitness in seed beetles. *Proc. Biol. Sci.* **282**, 20151421 (2015).

Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25-29 (2000).

Bayram, H., Sayadi, A., Goenaga, J., Immonen, E. & Arnqvist, G. Novel seminal fluid proteins in the seed beetle *Callosobruchus maculatus* identified by a proteomic and transcriptomic approach. *Insect Mol. Biol.* **26**, 58–73 (2016).

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

Berger, F., Ramírez-Hernández, M. H. & Ziegler, M. The new life of a centenarian: signalling functions of NAD (P). *Trends Biochem. Sci.* **29**, 111-118 (2004).

Bickhart, D. M. *et al*. Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643-650 (2017).

Caspi, R. *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44**, 471-480 (2015).

Chi, Y. H. *et al*. Cowpea bruchid midgut transcriptome response to a soybean cystatin - costs and benefits of counter-defence. *Insect Mol. Biol.* **18**, 97–110 (2009).

Chin, C. S. *et al*. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050-1054 (2016).

Croft, D. *et al*. The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, 472-477 (2013).

Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257-258 (2006).

Finn, R. D. *et al*. Pfam: the protein families database. *Nucleic Acids Res.* **42**, 222-230 (2013).

Grabherr, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

Guo, F. *et al*. Antagonistic Regulation, Yet Synergistic Defense: Effect of Bergapten and Protease Inhibitor on Development of Cowpea Bruchid *Callosobruchus maculatus*. *PLoS One* **7**, e41877 (2012).

Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).

Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).

Hunter, S. *et al*. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **40**, 306-312 (2011).

Immonen, E., Sayadi, A., Bayram, H. & Arnqvist, G. Mating Changes Sexually Dimorphic Gene Expression in the Seed Beetle *Callosobruchus maculatus*. *Genome Biol. Evol.* **9**, 677-699 (2017).

Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).

Kanehisa, M. *et al*. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, 199-205 (2013).

Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).

Kofler, R. *et al*. PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLoS One* **6**, e15925 (2011).

Kofler, R., Pandey, R. V. & Schlotterer, C. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27**, 3435-3436 (2011).

Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).

Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. & Blaxter, M. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front. Genet.* **4**, 237 (2013).

Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* **42**, e119 (2014).

Moon, J., Salzman, R. A., Ahn, J.-E., Koiwa, H. & Zhu-Salzman, K. Transcriptional regulation in cowpea bruchid guts during adaptation to a plant defence protease inhibitor. *Insect Mol. Biol.* **13**, 283-291 (2004).

Morgat, A. *et al*. UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res.* **40**, 761-769 (2011).

Parra, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. *Nucleic Acids Res.* **37**, 289-297 (2009).

Pauchet, Y., Wilkinson, P., Chauhan, R. & Ffrench-Constant, R. H. Diversity of Beetle Genes Encoding Novel Plant Cell Wall Degrading Enzymes. *PLoS One* **5**, e15635 (2010).

Pedra, J. H. F. *et al.* Transcriptome analysis of the cowpea weevil bruchid: identification of putative proteinases and alpha-amylases associated with food breakdown. *Insect Mol. Biol.* **12,** 405–412 (2003).

Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780-786 (2015).

Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290-295 (2015).

Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).

Sayadi, A., Immonen, E., Bayram, H. & Arnqvist, G. The De Novo Transcriptome and Its Functional Annotation in the Seed Beetle *Callosobruchus maculatus*. *PLoS One* **11**, e0158565 (2016).

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).

Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-3.0. http://www.repeatmasker.org/ (2010).

Smit, A. F. A., Hubley, R. & Green, P. RepeatModeler Open-1.0. http://www.repeatmasker.org/ (2015).

Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, 435-439 (2006).

Wang, L. *et al.* Transcriptomic response of cowpea bruchids to N-acetylglucosamine-specific lectins. *Insect Sci.* **22**, 83-94 (2014).

Zhu-Salzman, K., Koiwa, H., Salzman, R., Shade, R. & Ahn, J. Cowpea bruchid *Callosobruchus maculatus* uses a three-component strategy to overcome a plant defensive cysteine protease inhibitor. *Insect Mol. Biol.* **12**, 135-145 (2003).

Table S 4. Summary statistics of sequencing and mapping data (# of reads).

Sample	Total	Mapped	Mapping %	Mapped and cleaned	Mapping %
Bra1	332893986	185785844	55.81%	119453905	37.13%
Bra2	295030448	164559341	55.78%	103630468	36.40%
Ca1	289218514	100838452	34.87%	63163058	22.19%
Ca2	224072682	122540026	54.69%	75977662	34.90%
Yem1	300200932	168670544	56.19%	106086767	36.61%
Yem2	259985912	143300135	55.12%	89426442	35.62%

Table S 5. **(A)** Generalized linear model of the incidence of shared intermediate frequency polymorphism (0 or 1) (N = 4222 genes), using a binomial error distribution and a logit link function. “Mean” variables represent mean metric and “Difference in D_{NS}“ represents the difference between the largest and smallest estimate of D_{NS} (i.e., range) for a given gene, over the three populations. Given are also the sign of the covariation. The strong positive covariation between mean Tajima's D_{NS} and shared polymorphism is consistent with an enrichment of shared intermediate frequency polymorphism in genes that are consistently under balancing selection, such that genes with more positive values of D_{NS} in the three populations were much more likely to show shared polymorphism. In contrast, genes with more divergent estimates of D_{NS} in the three populations were less likely to show shared polymorphism. The fact that our estimates of p_N/p_S and nucleotide diversity both covaried positively with shared intermediate frequency polymorphism suggests that relaxed purifying selection also contributes to the likelihood of shared polymorphism, albeit to a lesser extent.
(B) General linear models of the effect of SBG expression on Tajima's D (based on synonymous and non-synonymous sites) in the three populations, when accounting for variation in overall gene expression, GC content and gene length by inclusion of these variables as covariates. Gene expression was here measured as normalized expression (FPKM) across all sampes reported in Immonen et al. (2017).

A: Source	**Wald χ^2**	**d.f.**	***P***	**Sign of effect**
Gene length	147.04	1	<0.001	+
Mean D_{NS}	278.4	1	<0.001	+
Difference in D_{NS}	22.93	1	<0.001	-
Mean p_N/p_S	35.56	1	<0.001	+
Mean π_S	88.4	1	<0.001	+
SBG category	9.43	7	0.223	

		Brazil D_S		**California D_S**		**Yemen D_S**		**Brazil D_{NS}**		**California D_{NS}**		**Yemen D_{NS}**	
B: Source	**d.f.**	***F***	***P***	***F***	***P***	***F***	***P***	***F***	***P***	***F***	***P***	***F***	***P***
SBG category	7	2.62	0.011	4.58	<0.001	5.30	<0.001	1.66	0.113	4.08	<0.001	1.36	0.216
FPKM	1	0.19	0.663	0.15	0.699	2.39	0.122	0.07	0.789	3.07	0.080	0.54	0.461
GC content	1	17.60	<0.001	10.46	0.001	10.50	0.001	2.49	0.115	8.24	0.004	7.30	0.007
Gene length	1	1.64	0.200	3.70	0.054	5.43	0.020	1.24	0.265	5.90	0.015	3.08	0.079

Table S 6. Functional enrichment of 149 genes showing D_S and D_{NS} > 0 in all three populations and a log2FC > 1, against a universe of all genes expressed in the female abdomen (sorted by counts).

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
54	GO:0044238	1.87E-02	2.118449	18.14325975	25	2033	primary metabolic process
86	GO:0071704	4.19E-02	1.888752	19.22310944	25	2154	organic substance metabolic process
14	GO:0043170	3.96E-03	2.51843	14.45749178	23	1620	macromolecule metabolic process
71	GO:0044237	2.83E-02	1.962829	14.73414749	21	1651	cellular metabolic process
20	GO:0044260	7.77E-03	2.338608	11.44997652	19	1283	cellular macromolecule metabolic process
63	GO:0006139	2.02E-02	2.212752	7.282292156	13	816	nucleobase-containing compound metabolic process
66	GO:0006725	2.45E-02	2.146343	7.460779709	13	836	cellular aromatic compound metabolic process
67	GO:0046483	2.61E-02	2.123855	7.523250352	13	843	heterocycle metabolic process
70	GO:1901360	2.73E-02	2.108024	7.56787224	13	848	organic cyclic compound metabolic process
28	GO:0090304	9.59E-03	2.548805	5.881164866	12	659	nucleic acid metabolic process
32	GO:0016070	1.47E-02	2.672924	3.998121184	9	448	RNA metabolic process
4	GO:0071840	7.76E-05	6.901062	1.472522311	8	165	cellular component organization or biogenesis
10	GO:0019222	2.61E-03	3.901235	2.480976984	8	278	regulation of metabolic process
1	GO:1902589	1.65E-07	22.462366	0.437294504	7	49	single-organism organelle organization
3	GO:0006996	1.47E-05	10.727104	0.838891498	7	94	organelle organization
6	GO:0016043	2.10E-04	6.832736	1.267261625	7	142	cellular component organization
17	GO:0080090	7.19E-03	3.540609	2.320338187	7	260	regulation of primary metabolic process
18	GO:0031323	7.19E-03	3.540609	2.320338187	7	260	regulation of cellular metabolic process
21	GO:0060255	7.80E-03	3.481988	2.356035698	7	264	regulation of macromolecule metabolic process
34	GO:0006355	1.72E-02	3.237825	2.115077501	6	237	regulation of transcription, DNA-templated
35	GO:0051252	1.72E-02	3.237825	2.115077501	6	237	regulation of RNA metabolic process
36	GO:2001141	1.72E-02	3.237825	2.115077501	6	237	regulation of RNA biosynthetic process
37	GO:1903506	1.72E-02	3.237825	2.115077501	6	237	regulation of nucleic acid-templated transcription
53	GO:0019219	1.85E-02	3.179521	2.150775012	6	241	regulation of nucleobase-containing compound metabolic process
55	GO:0031326	1.89E-02	3.165254	2.159699389	6	242	regulation of cellular biosynthetic process
56	GO:2000112	1.89E-02	3.165254	2.159699389	6	242	regulation of cellular macromolecule biosynthetic process
57	GO:0010556	1.89E-02	3.165254	2.159699389	6	242	regulation of macromolecule biosynthetic process
58	GO:0009889	1.89E-02	3.165254	2.159699389	6	242	regulation of biosynthetic process
61	GO:0051171	1.92E-02	3.151108	2.168623767	6	243	regulation of nitrogen compound metabolic process
62	GO:0010468	1.92E-02	3.151108	2.168623767	6	243	regulation of gene expression
94	GO:0006351	4.49E-02	2.540948	2.641615782	6	296	transcription, DNA-templated

95	GO:0097659	4.49E-02	2.540948	2.641615782	6	296	nucleic acid-templated transcription
96	GO:0032774	4.62E-02	2.52226	2.659464537	6	298	RNA biosynthetic process
2	GO:0000226	6.59E-07	99.176471	0.080319399	4	9	microtubule cytoskeleton organization
5	GO:0007010	1.09E-04	19.741176	0.258806952	4	29	cytoskeleton organization
7	GO:0034470	5.67E-04	12.294118	0.392672616	4	44	ncRNA processing
8	GO:0007017	1.42E-03	9.429864	0.499765148	4	56	microtubule-based process
16	GO:0034660	5.99E-03	6.166791	0.740723344	4	83	ncRNA metabolic process
29	GO:0006396	1.03E-02	5.220746	0.865664631	4	97	RNA processing
31	GO:0046907	1.35E-02	4.797903	0.937059652	4	105	intracellular transport
59	GO:0051649	1.89E-02	4.315126	1.035227806	4	116	establishment of localization in cell
65	GO:0051641	2.23E-02	4.089731	1.088774072	4	122	cellular localization
9	GO:0008033	2.04E-03	13.826374	0.258806952	3	29	tRNA processing
64	GO:0006399	2.21E-02	5.479121	0.60685768	3	68	tRNA metabolic process
11	GO:0033043	2.68E-03	33.436508	0.080319399	2	9	regulation of organelle organization
12	GO:0016569	3.34E-03	29.25	0.089243776	2	10	covalent chromatin modification
13	GO:0016570	3.34E-03	29.25	0.089243776	2	10	histone modification
15	GO:0051128	4.05E-03	25.993827	0.098168154	2	11	regulation of cellular component organization
19	GO:0016568	7.57E-03	17.978632	0.133865665	2	15	chromatin modification
30	GO:0043414	1.08E-02	14.597222	0.160638798	2	18	macromolecule methylation
33	GO:0006325	1.60E-02	11.666667	0.196336308	2	22	chromatin organization
60	GO:0032259	1.89E-02	10.60101	0.214185063	2	24	methylation
87	GO:0051276	4.25E-02	6.642857	0.330201973	2	37	chromosome organization
22	GO:0001682	8.92E-03	Inf	0.008924378	1	1	tRNA 5'-leader removal
23	GO:0032886	8.92E-03	Inf	0.008924378	1	1	regulation of microtubule-based process
24	GO:0070507	8.92E-03	Inf	0.008924378	1	1	regulation of microtubule cytoskeleton organization
25	GO:0031167	8.92E-03	Inf	0.008924378	1	1	rRNA methylation
26	GO:0031110	8.92E-03	Inf	0.008924378	1	1	regulation of microtubule polymerization or depolymerization
27	GO:0031109	8.92E-03	Inf	0.008924378	1	1	microtubule polymerization or depolymerization
38	GO:1902099	1.78E-02	114.027027	0.017848755	1	2	regulation of metaphase/anaphase transition of cell cycle
39	GO:0010965	1.78E-02	114.027027	0.017848755	1	2	regulation of mitotic sister chromatid separation
40	GO:0007091	1.78E-02	114.027027	0.017848755	1	2	metaphase/anaphase transition of mitotic cell cycle
41	GO:0030162	1.78E-02	114.027027	0.017848755	1	2	regulation of proteolysis
42	GO:1903050	1.78E-02	114.027027	0.017848755	1	2	regulation of proteolysis involved in cellular protein
43	GO:0033044	1.78E-02	114.027027	0.017848755	1	2	regulation of chromosome organization
44	GO:0033045	1.78E-02	114.027027	0.017848755	1	2	regulation of sister chromatid segregation

45	GO:0033047	1.78E-02	114.027027	0.017848755	1	2	regulation of mitotic sister chromatid segregation
46	GO:0051983	1.78E-02	114.027027	0.017848755	1	2	regulation of chromosome segregation
47	GO:0031145	1.78E-02	114.027027	0.017848755	1	2	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
48	GO:0061136	1.78E-02	114.027027	0.017848755	1	2	regulation of proteasomal protein catabolic process
49	GO:0030071	1.78E-02	114.027027	0.017848755	1	2	regulation of mitotic metaphase/anaphase transition
50	GO:0032434	1.78E-02	114.027027	0.017848755	1	2	regulation of proteasomal ubiquitin-dependent protein catabolic process
51	GO:0044784	1.78E-02	114.027027	0.017848755	1	2	metaphase/anaphase transition of cell cycle
52	GO:1903362	1.78E-02	114.027027	0.017848755	1	2	regulation of cellular protein catabolic process
68	GO:0001510	2.65E-02	57	0.026773133	1	3	RNA methylation
69	GO:0016575	2.65E-02	57	0.026773133	1	3	histone deacetylation
72	GO:0043161	3.52E-02	37.990991	0.035697511	1	4	proteasome-mediated ubiquitin-dependent protein catabolic process
73	GO:0007088	3.52E-02	37.990991	0.035697511	1	4	regulation of mitotic nuclear division
74	GO:0035601	3.52E-02	37.990991	0.035697511	1	4	protein deacylation
75	GO:0031329	3.52E-02	37.990991	0.035697511	1	4	regulation of cellular catabolic process
76	GO:0000154	3.52E-02	37.990991	0.035697511	1	4	rRNA modification
77	GO:0006476	3.52E-02	37.990991	0.035697511	1	4	protein deacetylation
78	GO:1901990	3.52E-02	37.990991	0.035697511	1	4	regulation of mitotic cell cycle phase transition
79	GO:0051783	3.52E-02	37.990991	0.035697511	1	4	regulation of nuclear division
80	GO:1901987	3.52E-02	37.990991	0.035697511	1	4	regulation of cell cycle phase transition
81	GO:0007346	3.52E-02	37.990991	0.035697511	1	4	regulation of mitotic cell cycle
82	GO:0098732	3.52E-02	37.990991	0.035697511	1	4	macromolecule deacylation
83	GO:0044770	3.52E-02	37.990991	0.035697511	1	4	cell cycle phase transition
84	GO:0044772	3.52E-02	37.990991	0.035697511	1	4	mitotic cell cycle phase transition
85	GO:0010498	3.52E-02	37.990991	0.035697511	1	4	proteasomal protein catabolic process
88	GO:0042176	4.39E-02	28.486486	0.044621888	1	5	regulation of protein catabolic process
89	GO:0018022	4.39E-02	28.486486	0.044621888	1	5	peptidyl-lysine methylation
90	GO:0051493	4.39E-02	28.486486	0.044621888	1	5	regulation of cytoskeleton organization
91	GO:0034968	4.39E-02	28.486486	0.044621888	1	5	histone lysine methylation
92	GO:0010564	4.39E-02	28.486486	0.044621888	1	5	regulation of cell cycle process
93	GO:0016571	4.39E-02	28.486486	0.044621888	1	5	histone methylation

	GOMFID	**Pvalue**	**OddsRatio**	**ExpCount**	**Count**	**Size**	**Term**
5	GO:0005488	8.10E-04	2.384858	54.54964235	68	4594	binding
2	GO:0043167	4.38E-04	2.207483	20.77968526	35	1750	ion binding
3	GO:0046872	4.43E-04	2.539285	10.46108727	22	881	metal ion binding

4	GO:0043169	5.18E-04	2.505914	10.57982833	22	891	cation binding
1	GO:0008270	6.74E-05	3.882875	4.23905579	14	357	zinc ion binding
6	GO:0046914	1.34E-03	2.817404	5.6758226	14	478	transition metal ion binding
18	GO:0004518	4.49E-02	4.00957	0.79556509	3	67	nuclease activity
7	GO:0003682	9.98E-03	15.479237	0.15436338	2	13	chromatin binding
8	GO:0008170	1.16E-02	14.187243	0.16623748	2	14	N-methyltransferase activity
10	GO:0034061	2.30E-02	9.449931	0.23748212	2	20	DNA polymerase activity
11	GO:0004197	3.00E-02	8.096414	0.27310443	2	23	cysteine-type endopeptidase activity
12	GO:0004386	3.50E-02	7.390231	0.29685265	2	25	helicase activity
17	GO:0008757	4.31E-02	6.534663	0.33247496	2	28	S-adenosylmethionine-dependent methyltransferase activity
9	GO:0003964	1.19E-02	Inf	0.01187411	1	1	RNA-directed DNA polymerase activity
13	GO:0033558	3.52E-02	42.103659	0.03562232	1	3	protein deacetylase activity
14	GO:0004826	3.52E-02	42.103659	0.03562232	1	3	phenylalanine-tRNA ligase activity
15	GO:0004407	3.52E-02	42.103659	0.03562232	1	3	histone deacetylase activity
16	GO:0004526	3.52E-02	42.103659	0.03562232	1	3	ribonuclease P activity
19	GO:0019213	4.67E-02	28.065041	0.04749642	1	4	deacetylase activity
20	GO:0017150	4.67E-02	28.065041	0.04749642	1	4	tRNA dihydrouridine synthase activity
21	GO:0016888	4.67E-02	28.065041	0.04749642	1	4	endodeoxyribonuclease activity, producing 5'-phosphomonoesters
22	GO:0000179	4.67E-02	28.065041	0.04749642	1	4	rRNA (adenine-N6,N6-)-dimethyltransferase activity

	GOCCID	**Pvalue**	**OddsRatio**	**ExpCount**	**Count**	**Size**	**Term**
1	GO:0005622	2.59E-07	15.926829	10.62005277	23	966	intracellular
3	GO:0044464	3.71E-07	15.469239	10.79595427	23	982	cell part
4	GO:0005623	3.71E-07	15.469239	10.79595427	23	982	cell
2	GO:0044424	2.93E-07	12.254157	9.49868074	22	864	intracellular part
7	GO:0043229	1.44E-05	6.204206	7.44283201	18	677	intracellular organelle
8	GO:0043226	1.47E-05	6.190909	7.45382586	18	678	organelle
10	GO:0043227	3.18E-05	5.587185	5.39797713	15	491	membrane-bounded organelle
11	GO:0043231	3.18E-05	5.587185	5.39797713	15	491	intracellular membrane-bounded organelle
9	GO:0005634	2.45E-05	5.999273	3.92480211	13	357	nucleus
5	GO:0005819	1.16E-05	153.204545	0.05496922	3	5	spindle
6	GO:0000922	1.16E-05	153.204545	0.05496922	3	5	spindle pole
12	GO:0005815	3.99E-05	76.534091	0.0769569	3	7	microtubule organizing center
13	GO:0015630	2.89E-03	12.642045	0.29683377	3	27	microtubule cytoskeleton
15	GO:0044430	2.22E-02	5.650086	0.61565523	3	56	cytoskeletal part

19	GO:0005856	4.40E-02	4.244805	0.80255057	3	73	cytoskeleton
14	GO:0033588	1.10E-02	Inf	0.01099384	1	1	Elongator holoenzyme complex
16	GO:0000152	3.26E-02	46.8125	0.03298153	1	3	nuclear ubiquitin ligase complex
17	GO:0005680	3.26E-02	46.8125	0.03298153	1	3	anaphase-promoting complex
18	GO:0005801	4.33E-02	31.194444	0.04397537	1	4	cis-Golgi network

Table S 6 continued. Functional enrichment of 149 genes showing D_S and $D_{NS} > 0$ in all three populations and a log2FC > 1, against a universe of all FBGs genes (log2FC > 1) expressed in the female abdomen (sorted by counts).

	GOBPID	**Pvalue**	**OddsRatio**	**ExpCount**	**Count**	**Size**	**Term**
1	GO:0071840	0.008350179	3.52	3.1908397	8	33	cellular component organization or biogenesis
2	GO:1902589	0.000274469	8.681004	1.5470738	7	16	single-organism organelle organization
3	GO:0006996	0.007543997	3.993209	2.5139949	7	26	organelle organization
4	GO:0016043	0.011708279	3.591398	2.7073791	7	28	cellular component organization
5	GO:0000226	0.000350691	41.647059	0.4834606	4	5	microtubule cytoskeleton organization
6	GO:0007010	0.000350691	41.647059	0.4834606	4	5	cytoskeleton organization
7	GO:0046907	0.015009768	5.848739	1.0636132	4	11	intracellular transport
8	GO:0051641	0.020954263	5.102941	1.1603053	4	12	cellular localization
9	GO:0051649	0.020954263	5.102941	1.1603053	4	12	establishment of localization in cell
10	GO:0007017	0.020954263	5.102941	1.1603053	4	12	microtubule-based process
11	GO:0033043	0.025699062	19.666667	0.2900763	2	3	regulation of organelle organization
12	GO:0051128	0.025699062	19.666667	0.2900763	2	3	regulation of cellular component organization

	GOMFID	**Pvalue**	**OddsRatio**	**ExpCount**	**Count**	**Size**	Term
4	GO:0008270	0.041012053	1.884058	8.7428181	14	77	zinc ion binding
1	GO:0004175	0.007589266	8.151899	0.9083447	4	8	endopeptidase activity
2	GO:0004197	0.03542821	15.975309	0.3406293	2	3	cysteine-type endopeptidase activity
3	GO:0016810	0.03542821	15.975309	0.3406293	2	3	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds

	GOCCID	**Pvalue**	**OddsRatio**	**ExpCount**	**Count**	**Size**	Term
1	GO:0005815	0.00131388	Inf	0.3409091	3	3	microtubule organizing center
2	GO:0005819	0.00131388	Inf	0.3409091	3	3	spindle
3	GO:0000922	0.00131388	Inf	0.3409091	3	3	spindle pole
4	GO:0015630	0.004855908	26.454545	0.4545455	3	4	microtubule cytoskeleton
5	GO:0005856	0.020737574	8.727273	0.6818182	3	6	cytoskeleton
6	GO:0044430	0.020737574	8.727273	0.6818182	3	6	cytoskeletal part

Table S 7. Functional enrichment of 15 genes showing shared intermediate frequency polymorphism across populations, signs of balancing selection within all populations ($D_{NS} > 1$) and female-biased expression (Log2FC >1) against a universe of all genes expressed in the female abdomen (sorted by count).

	GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
2	GO:1902589	0.003125656	34.63095	0.089279675	2	44	single-organism organelle organization
29	GO:0006996	0.009860633	18.66234	0.160297599	2	79	organelle organization
42	GO:0016043	0.021318373	12.21983	0.239431857	2	118	cellular component organization
45	GO:0071840	0.027885513	10.51119	0.27595536	2	136	cellular component organization or biogenesis
1	GO:0009186	0.002029084	Inf	0.002029084	1	1	deoxyribonucleoside diphosphate metabolic process
3	GO:1903362	0.004054735	590	0.004058167	1	2	regulation of cellular protein catabolic process
4	GO:0051983	0.004054735	590	0.004058167	1	2	regulation of chromosome segregation
5	GO:0061136	0.004054735	590	0.004058167	1	2	regulation of proteasomal protein catabolic process
6	GO:0030071	0.004054735	590	0.004058167	1	2	regulation of mitotic metaphase/anaphase transition
7	GO:1902099	0.004054735	590	0.004058167	1	2	regulation of metaphase/anaphase transition of cell cycle
8	GO:0031145	0.004054735	590	0.004058167	1	2	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
9	GO:0033044	0.004054735	590	0.004058167	1	2	regulation of chromosome organization
10	GO:0033045	0.004054735	590	0.004058167	1	2	regulation of sister chromatid segregation
11	GO:0033047	0.004054735	590	0.004058167	1	2	regulation of mitotic sister chromatid segregation
12	GO:0007091	0.004054735	590	0.004058167	1	2	metaphase/anaphase transition of mitotic cell cycle
13	GO:0032434	0.004054735	590	0.004058167	1	2	regulation of proteasomal ubiquitin-dependent protein catabolic process
14	GO:1903050	0.004054735	590	0.004058167	1	2	regulation of proteolysis involved in cellular protein catabolic process
15	GO:0010965	0.004054735	590	0.004058167	1	2	regulation of mitotic sister chromatid separation
16	GO:0044784	0.004054735	590	0.004058167	1	2	metaphase/anaphase transition of cell cycle
17	GO:0030162	0.004054735	590	0.004058167	1	2	regulation of proteolysis
18	GO:0043161	0.006076959	294.9	0.006087251	1	3	proteasome-mediated ubiquitin-dependent protein catabolic process
19	GO:0010498	0.006076959	294.9	0.006087251	1	3	proteasomal protein catabolic process
20	GO:0007346	0.00809576	196.53333	0.008116334	1	4	regulation of mitotic cell cycle
21	GO:0051783	0.00809576	196.53333	0.008116334	1	4	regulation of nuclear division
22	GO:0007088	0.00809576	196.53333	0.008116334	1	4	regulation of mitotic nuclear division
23	GO:1901990	0.00809576	196.53333	0.008116334	1	4	regulation of mitotic cell cycle phase transition
24	GO:0042176	0.00809576	196.53333	0.008116334	1	4	regulation of protein catabolic process
25	GO:1901987	0.00809576	196.53333	0.008116334	1	4	regulation of cell cycle phase transition
26	GO:0044770	0.00809576	196.53333	0.008116334	1	4	cell cycle phase transition
27	GO:0044772	0.00809576	196.53333	0.008116334	1	4	mitotic cell cycle phase transition
28	GO:0031329	0.00809576	196.53333	0.008116334	1	4	regulation of cellular catabolic process
30	GO:0010564	0.010111143	147.35	0.010145418	1	5	regulation of cell cycle process
31	GO:0009894	0.012123112	117.84	0.012174501	1	6	regulation of catabolic process

32	GO:0051302	0.012123112	117.84	0.012174501	1	6	regulation of cell division
33	GO:0051301	0.014131672	98.16667	0.014203585	1	7	cell division
34	GO:0007067	0.016136828	84.11429	0.016232668	1	8	mitotic nuclear division
35	GO:1903047	0.016136828	84.11429	0.016232668	1	8	mitotic cell cycle process
36	GO:0000280	0.016136828	84.11429	0.016232668	1	8	nuclear division
37	GO:0000278	0.016136828	84.11429	0.016232668	1	8	mitotic cell cycle
38	GO:0000226	0.016136828	84.11429	0.016232668	1	8	microtubule cytoskeleton organization
39	GO:0033043	0.018138584	73.575	0.018261752	1	9	regulation of organelle organization
40	GO:0009132	0.020136945	65.37778	0.020290835	1	10	nucleoside diphosphate metabolic process
41	GO:0051128	0.020136945	65.37778	0.020290835	1	10	regulation of cellular component organization
43	GO:0048285	0.022131916	58.82	0.022319919	1	11	organelle fission
44	GO:0022402	0.026111704	48.98333	0.026378086	1	13	cell cycle process
46	GO:0051726	0.032056072	39.14667	0.032465336	1	16	regulation of cell cycle
47	GO:0032268	0.034030796	36.6875	0.03449442	1	17	regulation of cellular protein metabolic process
48	GO:0051246	0.037970174	32.58889	0.038552587	1	19	regulation of protein metabolic process
49	GO:0007275	0.045808775	26.62727	0.046668921	1	23	multicellular organismal development

	GOMFID	**Pvalue**	**OddsRatio**	**ExpCount**	**Count**	**Size**	**Term**
1	GO:0008168	0.001484029	54.74713	0.06216328	2	60	methyltransferase activity
2	GO:0016741	0.001740309	50.34921	0.06734356	2	65	transferase activity, transferring one-carbon groups

	GOCCID	**Pvalue**	**OddsRatio**	**ExpCount**	**Count**	**Size**	**Term**
1	GO:0005680	0.007638416	260.33333	0.007653061	1	3	anaphase-promoting complex
2	GO:0000152	0.007638416	260.33333	0.007653061	1	3	nuclear ubiquitin ligase complex
3	GO:0005819	0.012706325	130	0.012755102	1	5	spindle
4	GO:0000922	0.012706325	130	0.012755102	1	5	spindle pole
5	GO:0005815	0.015232989	103.93333	0.015306122	1	6	microtubule organizing center
6	GO:0031461	0.022783887	64.83333	0.022959184	1	9	cullin-RING ubiquitin ligase complex
7	GO:0030131	0.027793638	51.8	0.028061224	1	11	clathrin adaptor complex
8	GO:0000151	0.027793638	51.8	0.028061224	1	11	ubiquitin ligase complex
9	GO:0030119	0.030291278	47.06061	0.030612245	1	12	AP-type membrane coat adaptor complex
10	GO:0015630	0.032784103	43.11111	0.033163265	1	13	microtubule cytoskeleton
11	GO:0030118	0.032784103	43.11111	0.033163265	1	13	clathrin coat

Table S 6 continued. Functional enrichment of 10 genes showing shared intermediate frequency polymorphism across populations, signs of balancing selection within all populations (D_{NS} > 1) and male-biased expression (Log2FC <-1) against a universe of all genes expressed in the male abdomen (sorted by count).

	GOBPID	**Pvalue**	**OddsRatio**	**ExpCount**	**Count**	**Size**	**Term**
1	GO:0044710	0.032397	11.12903	0.852117	3	654	single-organism metabolic process

	GOMFID	**Pvalue**	**OddsRatio**	**ExpCount**	**Count**	**Size**	**Term**
11	GO:0003824	0.031207	Inf	1.681919	4	2086	catalytic activity
1	GO:0016705	0.000878	83.01695	0.049184	2	61	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
2	GO:0005506	0.000967	78.95161	0.051602	2	64	iron ion binding
3	GO:0020037	0.001435	64.22368	0.062891	2	78	heme binding
4	GO:0046906	0.001472	63.37662	0.063697	2	79	tetrapyrrole binding
9	GO:0046914	0.029168	12.69337	0.293489	2	364	transition metal ion binding
10	GO:0016491	0.031021	12.25401	0.303165	2	376	oxidoreductase activity
5	GO:0016743	0.001612	1652	0.001613	1	2	carboxyl- or carbamoyltransferase activity
6	GO:0016597	0.002417	825.8333	0.002419	1	3	amino acid binding
7	GO:0031406	0.002417	825.8333	0.002419	1	3	carboxylic acid binding
8	GO:0043177	0.002417	825.8333	0.002419	1	3	organic acid binding
12	GO:0016741	0.049075	26.7541	0.04999	1	62	transferase activity, transferring one-carbon groups

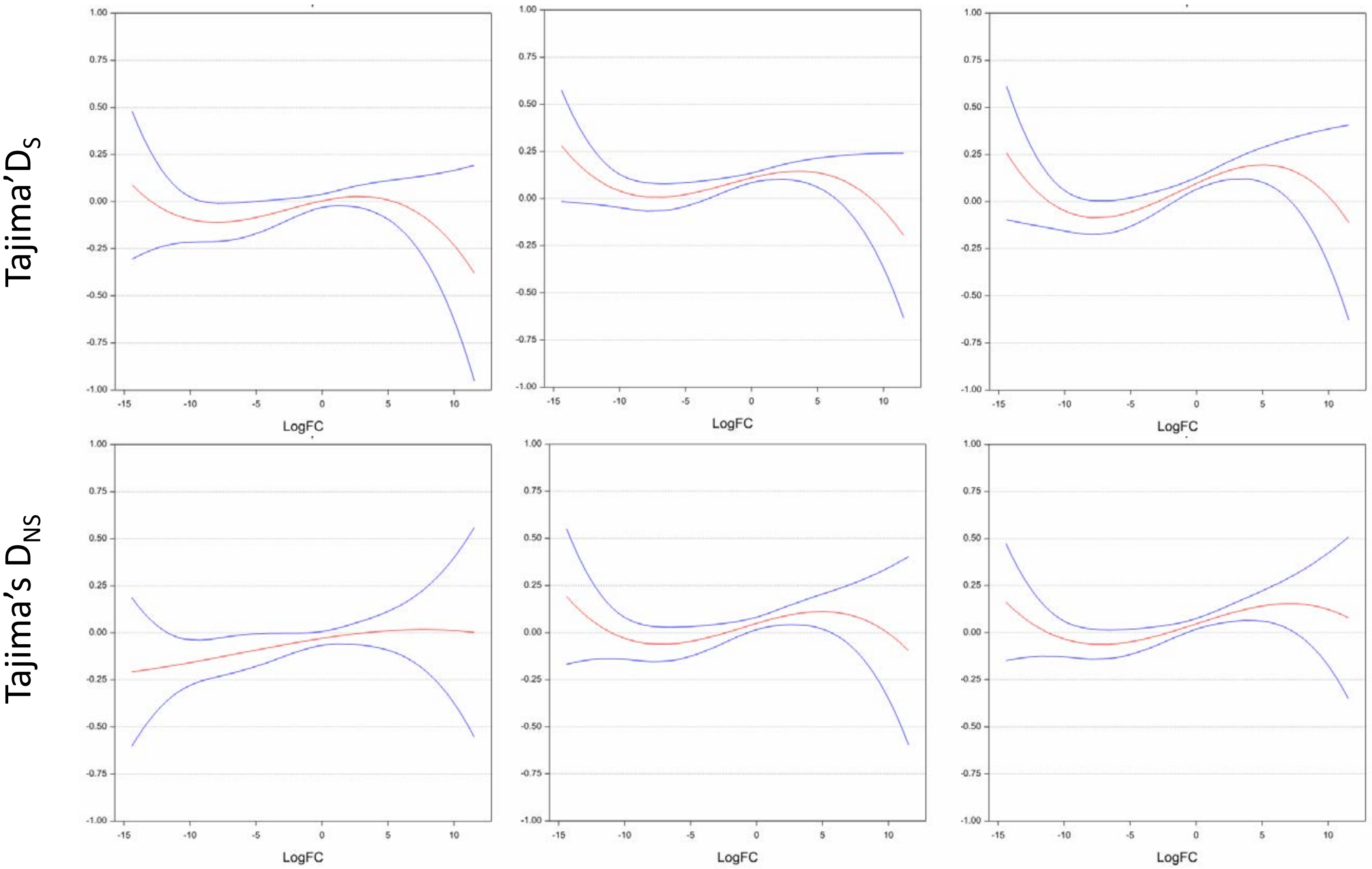

Figure S 4. Fitted cubic polynomial regression models (±95% CI) relating variation in Tajima's D across genes to the absolute level of sex-biassed gene expression (Log2FC), illustrating the wave-shaped pattern between these gene characteristics. The left column shows the B, the center the C and the right the Y population.

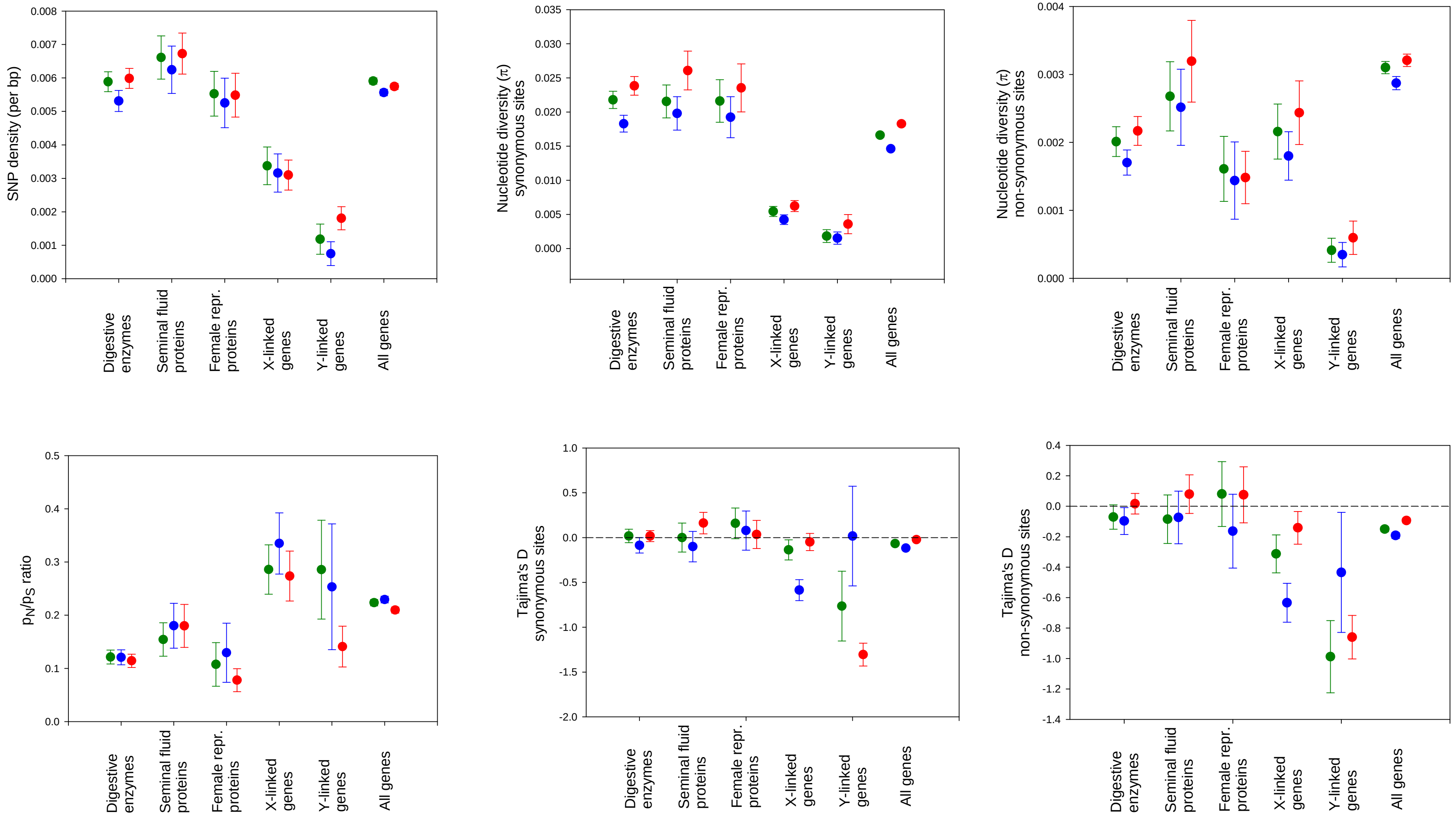

Figure S 5. Population genomic metrics (mean ±95% bootstrap CI) for the three populations studied (blue = B; red = C; green = Y) for different sets of genes.

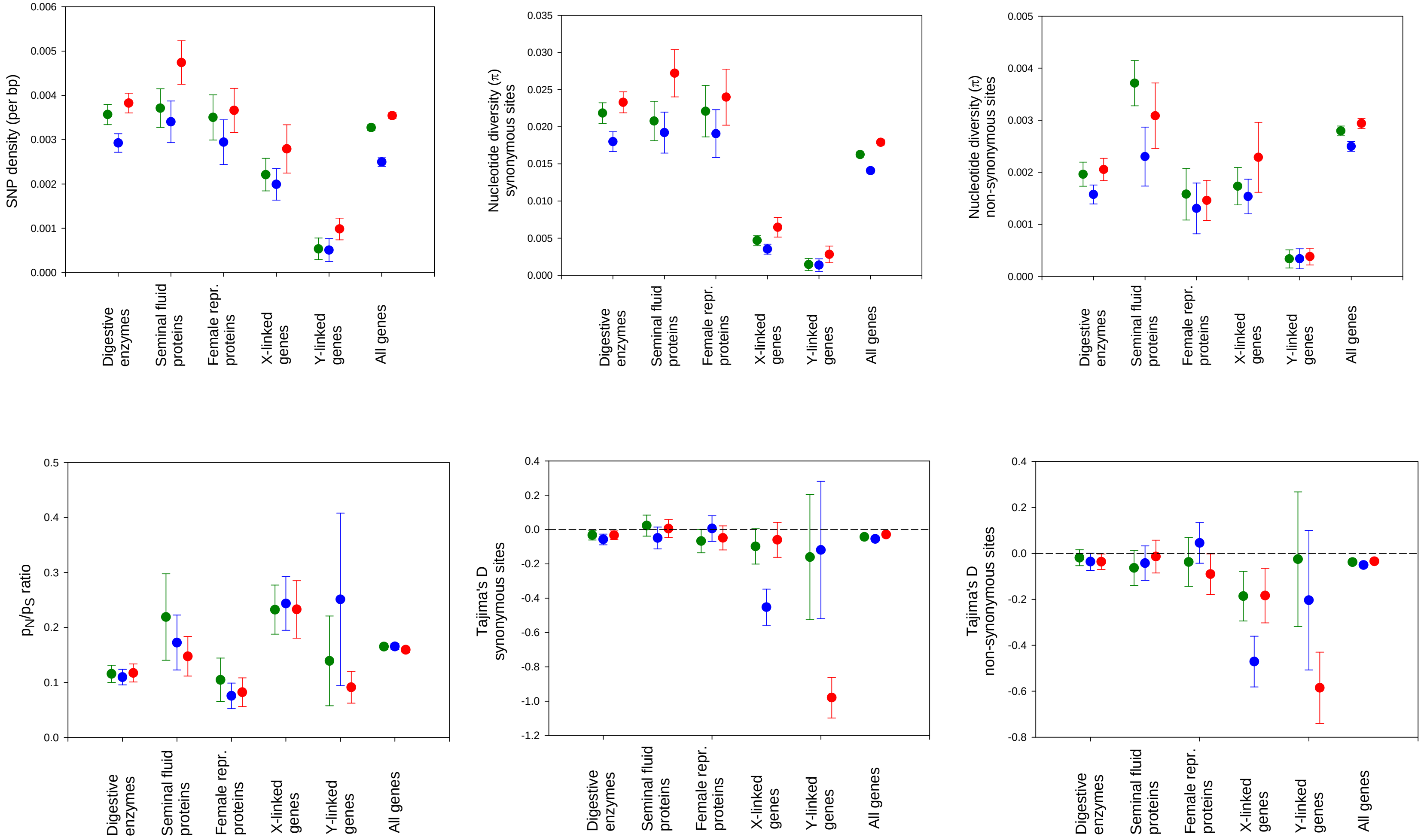

Figure S 6. Population genomic metrics (mean ±95% bootstrap CI) for the three populations studied (blue = B; red = C; green = Y) for different sets of genes based on subsampling reads from X- and Y-linked genes down to 20x and other sets down to 10x, to facilitate the comparisons between autosomal and sex-linked genes at equal coverage. A comparison with SI Figure 5 shows that the deviant behaviour of X- and, in particular, Y-linked genes is not caused by unequal sequencing depth but, presumably, by their lower effective population size and recombination rate.

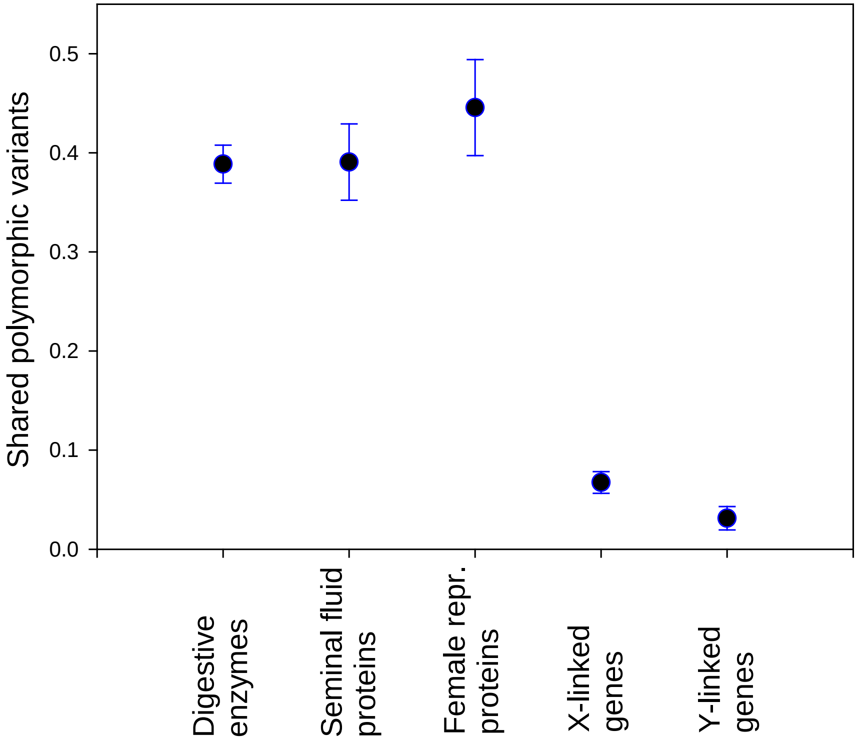

Figure S 7. The observed level of shared polymorphism for gene sets across the three populations. Predicted values (±SE) of the probability that a gene harbors ≥1 SNP showing intermediate frequency polymorphism in all three populations, from a generalized linear model (binomial errors and a logit link) accounting for the effects of gene length and SNP density.