Electronic Supplementary Materials

Repeated sex chromosome evolution in vertebrates supported by expanded avian sex chromosomes

Hanna Sigeman[a#], Suvi Ponnikas[a], Pallavi Chauhan[a], Elisa Dierickx[b], M. de L. Brooke[b], Bengt Hansson[a#]

[a] Department of Biology, Lund University, Ecology Building, 223 62 Lund, Sweden

[b] Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

[#] Correspondence: hanna.sigeman@biol.lu.se or bengt.hansson@biol.lu.se

Content:

Supplementary Methods S1

S1a. DNA extraction and sequencing

DNA from the blood samples was extracted from each sample (n = 8) using a phenol–chloroform protocol [1]. The extracted DNA was sequenced with Illumina HiSeqX (150 bp, paired-end) by SciLifeLab Sweden.

S1b. *De novo* assembly and mapping

To obtain reference genomes, we created *de novo* genome assemblies from the male sequence data for each of the study species. The resequencing data was trimmed using nesoni clip (https://github.com/Victorian-Bioinformatics-Consortium/nesoni) with a minimum read quality of 20 and minimum read length of 20 bp. The trimmed reads were then assembled with Spades v3.5.0 [2] using 5 different kmer lengths (21, 33, 55, 77 and 127) and the setting "careful". Scaffolds shorter than 1 kbp were discarded. Quality statistics from the assemblies were calculated using Quast v4.5.4 [3]. Samples from all species were aligned to their corresponding reference genome. However, based on results from downstream analyses (S1d; figure S2), two of the reference genomes (the horned lark and the bearded reedling) were discarded from the final analyses, and the samples from these species were instead analysed using the reference genome of their closest relative, the Raso lark. This decision was made as the level of heterozygosity in these two species influenced the relative mapping success between the female and male sample on a genome-wide scale, as the reference genome was based on the same male sample, thus not revealing the sex-linked genomic regions as clearly as when the samples were aligned to the reference genome of their relative, the Raso lark (see figure S2). This strategy meant keeping the assemblies of the Raso lark (N50=103 kbp, 28304 scaffolds) and the bearded reedling (N50=68 kbp, 36455 scaffolds), while the assemblies for the Eurasian skylark (N50=8 kbp, 256822 scaffolds) and horned lark (N50=22 kbp, 109088 scaffolds) were discarded. Genome assembly statistics are given in electronic supplementary material, table S1.

Reads from all of the samples (n = 8) were cleaned for adaptor sequences with Trimmomatic v.0.3.6 [4] using the adaptor file TruSeq3-PE and options seedMismatches: 2, palindromeClipThreshold: 30 and simpleClipThreshold: 10. Trimming of low-quality bases was done using a quality threshold of 15 from the leading end and 30 from the trailing end. The reads were further trimmed for a minimum quality of 20 over sliding windows of 4 bp. Lastly, any reads shorter than 90 bp were excluded from further analyses. The number of remaining reads in our samples ranged from 176 to 404 million. The samples were quality checked using fastqc v0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

The male and female samples of Raso lark and bearded reedling were aligned to their respective genome assemblies, while the Eurasian skylark and horned lark individuals were aligned to the genome assembly of their closest relative, the Raso lark. The alignment was done with bwa mem v0.7.17 [5], marking shorter split hits as secondary (option -M) for downstream compatibility. The aligned reads were sorted with samtools v1.7 [6], and duplicated reads were removed using picardtools v2.18.0 (http://broadinstitute.github.io/picard).

Assembly and alignment statistics are provided in electronic supplementary material, table S1 and table S2.

S1c. Chromosome anchoring

The scaffolds in the genome assemblies were grouped into different chromosomes and ordered into chromosome-level using the genome assembly of the zebra finch (*Taeniopygia guttata*). The zebra finch genome assembly (taeGut.3.2.4 [7]) was downloaded from Ensembl [8] and transformed into a database using the last v876 [9] program lastdb. The two genome assemblies, of the Raso lark and bearded reedling, were aligned to the zebra finch genome using the program lastal and converted to psl format using the script maf-convert, both from the same software suite last v876. From there, we extracted chromosome anchoring coordinates based on the longest match to the zebra finch genome for each 5 kbp window in the Raso lark and bearded reedling assemblies, with a minimum requirement of 500 matching base pairs per 5 kbp window. The assembly positions from the output files of the coverage and single nucleotide variant (SNV) analyses were then translated to the starting positions of the match to the zebra finch genome assembly (see section below). With this method, between 81 and 95% of the scaffolds larger than 5 kbp were anchored to the zebra finch reference genome, with the largest proportion aligning to the Raso lark (95%) and bearded reedling (94%). See statistics in table S3.

S1d. Identification of sex-linked genomic regions

We identified sex-linked regions using two different kinds of genomic signatures: (i) differential mapping success in males and females (i.e. sex-specific genome coverage), and (ii) an excess or deficit of female-specific genetic variation. Sex chromosomes almost invariably evolve recombination suppression in the heterogametic sex so that regions that have been sex-linked for a long time (such as the sex chromosomes that formed in the ancestor of all birds) will show pronounced sequence divergence and degeneration in the non-recombining chromosome (the W in birds [10]). The regions belonging to the ancestral sex chromosomes can thus be identified by lower female coverage, and fewer female-specific genetic variants compared to males. This is because reads from the female-specific W-chromosome will either not map to the male reference genome due to substantial differentiation between the Z and W or because of deletions on the W chromosome. More recently formed sex-linked regions may be identified by lower mapping success in females, although we expect a subtler difference as the W-linked genomic region may not have yet developed substantial differentiation from the Z-linked homologous region. Here, we also expect a higher amount of female-specific mutations compared to males, due to differentiation between the Z and W sex chromosome copies.

To uncover differential mapping success in males and females for each species, we calculated genome-wide coverage for 5 kbp windows with bedtools v2.71.1 [11] from alignment files with increasingly strict settings for maximum allowed mismatches between the samples and the reference genome, allowing a maximum of (i) 4, (ii) 3, (iii) 2, (iv) 1 and lastly (v) 0 mismatches. All genome coverage values were normalised between the female and the male sample by dividing the median female-to-male coverage ratio for each 5 kbp window by the genome-wide median female-to-male coverage ratio. We also excluded windows with a read count >1500 in either sample or a coverage ratio >2. Both of these thresholds were set to be well outside the normal distribution for all species and was done in order to filter out repeat regions which may have a large effect on the differences between the two samples. To identify sex-linked regions, we binned the female-to-male coverage ratio values for every 1 Mbp genomic region of chromosomes larger than 1 Mbp and extracted the mean value from each bin. Results from all maximum mismatches settings were visually inspected and we decided on a cut-off of maximum 2 mismatches, which best reveals the sex-linked regions in the data. We then binned the data into 0.1 Mbp windows and calculated the mean female-to-male coverage ratio within each bin.

To analyse female-specific variation, we called variants in the alignment files (with all mismatches allowed) with freebayes v1.1.0 for each species separately (n = 2 in each analysis) using freebayes-parallel [12] and parallel v20180322 [13]. The output was then parsed for any SNV that had been marked with a flag other than PASS (--remove-filtered-all), a minimum quality of 20 and minimum depth of 3x using vcftools v0.1.15 [14]. Private alleles (minor alleles occurring only in one sample in a heterozygous state) were extracted with vcftools using option --singletons. We calculated the difference between the number of female-specific private alleles and male-specific private alleles for each 5 kbp window and extracted the average difference across 1 Mbp and 0.1 Mbp windows.

Supplementary figure S2 shows the results from the 1 Mbp window analysis from the Eurasian skylark and horned lark aligned to the reference genome of the Raso lark, as well as to the reference genome of their respective species.

S1e. Extraction of Z–W sequences of gametologous genes

We used the whole-genome synteny aligner program SatsumaSynteny v. 2.0 [15] to align the Raso lark assembly to the zebra finch assembly (taeGut.3.2.4), and then used kraken [16] to make a lift-over of the zebra finch annotations to the Raso lark assembly. Of the 18204 transcripts and 17488 genes in the zebra finch annotation, 14466 transcripts (79%) from 13764 genes (79%) were annotated in the Raso lark. We used Freebayes v.1.1.0 [12] (--report-monomorphic) to call variants for every base pair within all exons in the Raso lark based on the genome coordinates from the lift-over.

We *in silico* extracted gametologous gene sequences from sex-linked regions using in-house scripts (code provided as Code S1; general methodology described in [17] based on the genotypes of the female and male samples). The scripts use sex-specific genotype information to phase the data into a Z and W sequence. We replaced a site by "N" if it had a quality score or sequence depth below 20 or if the genotype in either sample was not called. Any site without variants in either sample was extracted as such, and remaining variants between the male and female were phased based on sex-specific allele compositions provided in electronic supplementary material, table S4.

To confirm that both Z and W gametologs were present in the data, we calculated genome coverage values for every exon using bedtools v.2.27.1 multicov [11] and normalised the values between the two samples of each species. Exons with female coverage less than 75% of the male sample were masked with N:s, as this suggests absence of a W gametolog.

We used TransDecoder v3.0.1 (https://github.com/TransDecoder/TransDecoder/) to find the longest open reading frames for each of the sequences (--retain_pfam_hits, --retain_blastp_hits, --single-best-orf). Orthologous genes from the zebra finch, collared flycatcher (*Ficedula albicollis*) and chicken (*Gallus gallus*) were downloaded through BioMart (database: Ensembl Genes 93), and the longest transcript from the flycatcher and chicken corresponding to the zebra finch transcripts in the annotation was added as additional sequences. The gene sequences of the zebra finch were later used to analyse the relative evolutionary rate between Z and W gametologs. The sequences were codon-aware aligned using *Prank* v.150803 [18]. Each N in the sequences was transformed into "-", and all sites including this character in any sequence were then removed using gblocks v0.91b [19]. We calculated pairwise substitution rates between Z–W gametologs using codeml (estimated kappa and codon frequency F3X4) from the PAML v4.9 package [20]. Gene sequences longer than 500 bp, and where the Z and W sequences within a species had a synonymous substitution rate (dS) value above 0.01, were kept for further analysis. No comparisons between gene sequences had synonymous substitution values (dS) > 1, which may have been biasing the divergence estimates due to mutation saturation. The maximum dS value in our dataset was 0.4697 with > 90% having dS values < 0.1.

S1f. References for Supplementary Methods S1

1. Sambrock J, Russel DW. 2001 Molecular Cloning: A Laboratory Manual (3rd edition). Cold Spring Harbor Laboratory Press, New York.

2. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. 2012 SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477. (doi:10.1089/cmb.2012.0021)

3. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013 QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075. (doi:10.1093/bioinformatics/btt086)

4. Bolger AM, Lohse M, Usadel B. 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. (doi:10.1093/bioinformatics/btu170)

5. Li H, Durbin R. 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760. (doi:10.1093/bioinformatics/btp324)

6. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)

7. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, et al. 2010 The genome of a songbird. *Nature* **464**, 757–762. (doi:10.1038/nature08819)

8. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. 2015 Ensembl 2015. *Nucleic Acids Research* **43**, D662–9. (doi:10.1093/nar/gku1010)

9. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011 Adaptive seeds tame genomic sequence comparison. *Genome Research* **21**, 487–493. (doi:10.1101/gr.113985.110)

10. Zhou Q, Zhang J, Bachtrog D, An N, Huang Q, Jarvis ED, et al. 2014 Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346**, 1246338–1246338. (doi:10.1126/science.1246338)

11. Quinlan AR, Hall IM. 2010 BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842. (doi:10.1093/bioinformatics/btq033)

12. Garrison E, Marth, G. 2012 Haplotype-based variant detection from short-read sequencing. *arxiv.org*.

13. Tange O. 2018 *GNU Parallel 2018*. Lulu.com.

14. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. 2011 The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158. (doi:10.1093/bioinformatics/btr330)

15. Grabherr MG, Russell P, Meyer M, Mauceli E, Alföldi J, Di Palma F, et al. 2010 Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* **26**, 1145–1151. (doi:10.1093/bioinformatics/btq102)

16. Zamani N, Sundström G, Meadows JRS, Höppner MP, Dainat J, Lantz H, et al. 2014 A universal genomic coordinate translator for comparative genomics. *BMC Bioinformatics* **15**, 227. (doi:10.1186/1471-2105-15-227)

17. Sigeman H, Ponnikas S, Videvall E, Zhang H, Chauhan P, Naurin S, et al. 2018 Insights into avian incomplete dosage compensation: sex-biased gene expression coevolves with sex chromosome degeneration in the Common Whitethroat. *Genes* **9**, 373. (doi:10.3390/genes9080373)

18. Loytynoja A. 2014 Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.* **1079**, 155–170. (doi:10.1007/978-1-62703-646-7_10)

19. Castresana J. 2000 Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* **17**, 540–552. (doi:10.1093/oxfordjournals.molbev.a026334)

20. Yang Z. 2007 PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol.* **24**, 1586–1591. (doi:10.1093/molbev/msm088)

Supplementary Figures

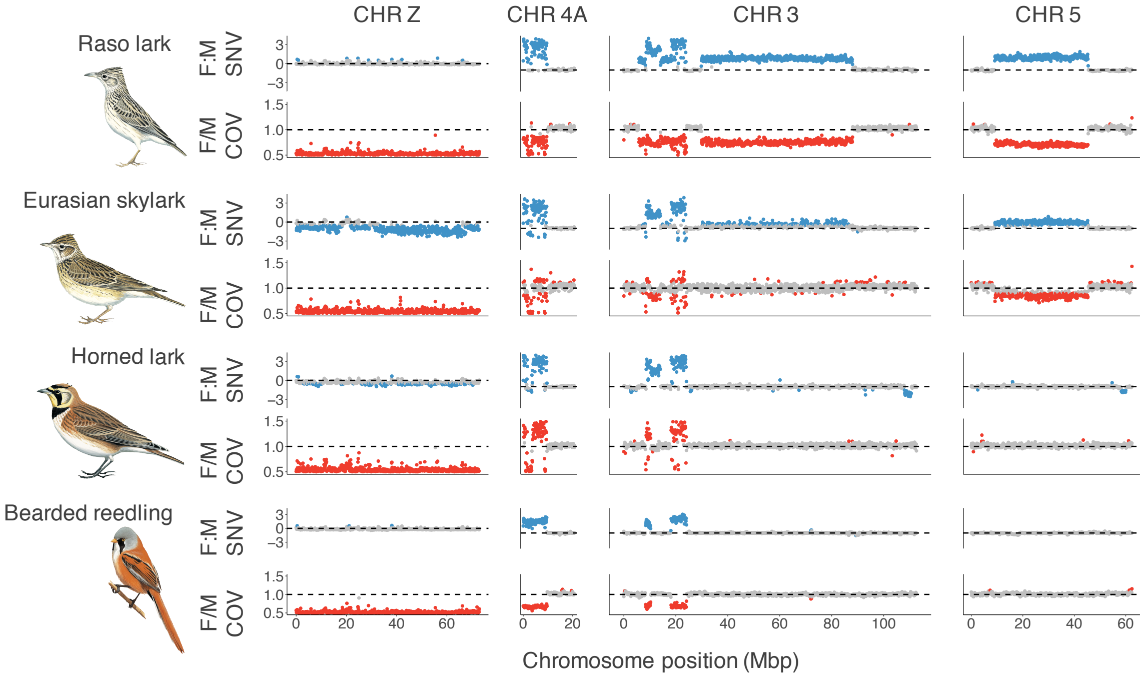

Figure S1. Distribution of female-to-male difference in number of private single nucleotide variants (SNVs), and female-to-male coverage ratio, in the four study species across chromosomes Z, 4A, 3 and 5. Chromosome names and positions follow the zebra finch genome structure (see Methods section). The upper panel for each chromosome shows sex-specific SNV differences (tick marks represent 3000, 0 and -3000, with values > 500 or < -500 in blue) and the lower panel shows female-to-male coverage ratios (with values > 0.9 or < 1.1 in red), Both measurements are means across 0.1 Mbp windows. Bird drawings ordered from HBW.

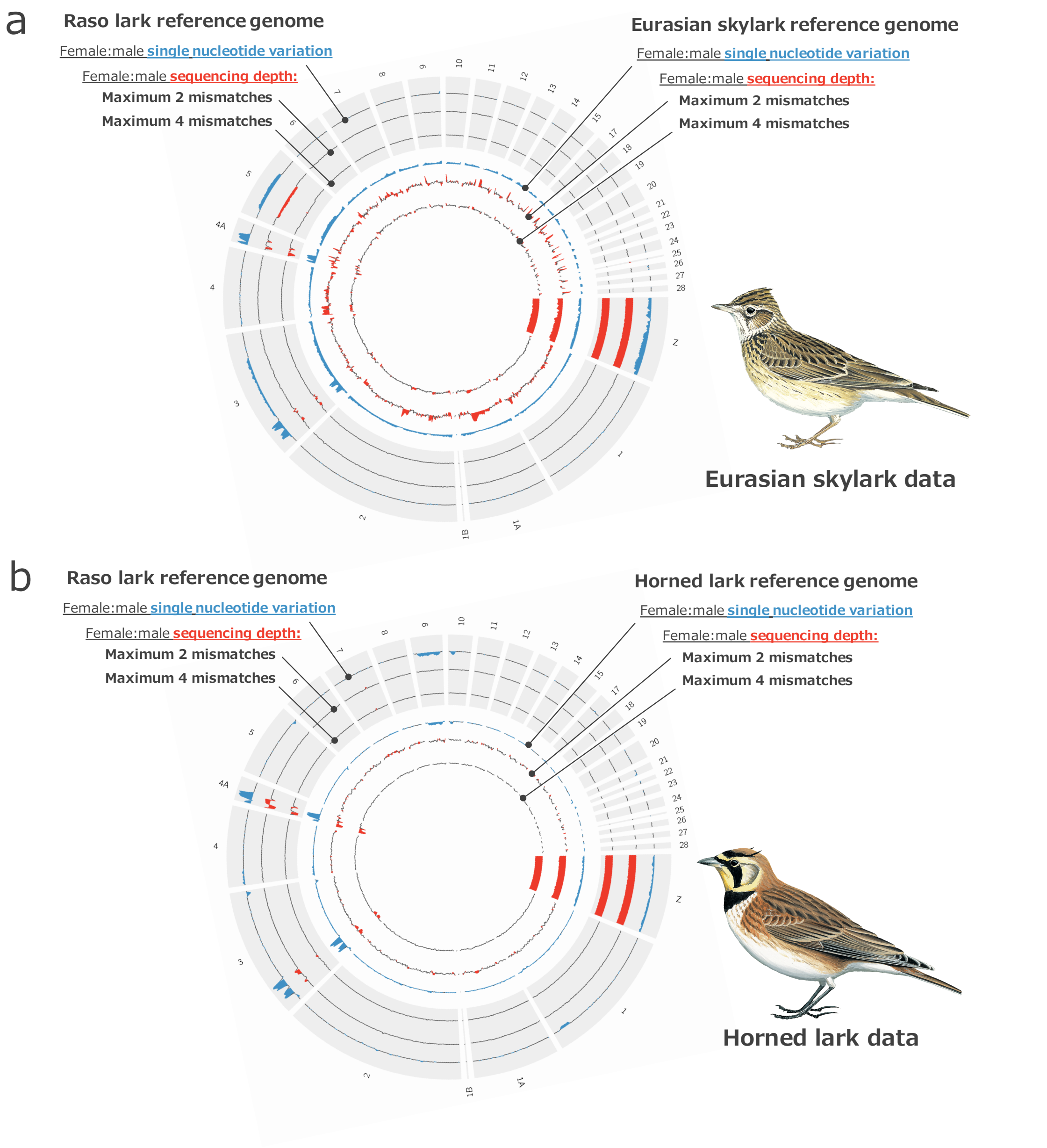

Figure S2. For each species, Eurasian skylark (a) and horned lark (b), the grey ring shows the same data that is presented for this species in Figure 1. The inner rings with white background show the results when aligned to its own reference genome, starting with single nucleotide variation data (in blue), followed by genome coverage data (in red) when parsed for 2 and 4 allowed mismatches. It is evident from the single nucleotide data for both species that the sex-linked regions are the same whether the reference genome is the Raso lark or their respective species. The genome coverage is however less clear, especially in the Eurasian skylark, which is why the Raso lark was chosen as a reference genome for these two species.

Supplementary Tables

Table S1. Genome assembly statistics (related to Supplementary Method S1b)

Statistics was computed using the software Quast version 4.5.4 [3].

Assembly	Raso lark	Eurasian skylark	Horned lark	Bearded reedling
N contigs (≥ 0 bp)	28304	256822	109088	36455
N contigs (≥ 1 kbp)	28304	256822	109088	36455
N contigs (≥ 5 kbp)	17977	79725	51289	24408
N contigs (≥ 10 kbp)	14382	31198	32587	19461
N contigs (≥ 25 kbp)	9767	3546	10545	12025
N contigs (≥ 50 kbp)	6145	609	2805	6436
Total length (≥ 0 bp)	1003896371	1276061662	1086584534	1020354129
Total length (≥ 1 kbp)	1003896371	1276061662	1086584534	1020354129
Total length (≥ 5 kbp)	979649475	887743951	962416262	991752640
Total length (≥ 10 kbp)	953912007	543408884	827105499	955927387
Total length (≥ 25 kbp)	878299503	145311035	479556955	833768162
Total length (≥ 50 kbp)	748288932	49632478	216102095	634399538
N contigs	28304	256822	109088	36455
Largest contig (bp)	851956	346983	447328	534950
Total length (bp)	1003896371	1276061662	1086584534	1020354129
GC (%)	42.31	42.59	42.17	42.14
N50	103210	8477	21578	68188
N75	49107	4130	10418	32980
L50	2726	41480	13291	4314
L75	6239	94969	31395	9642
# N's per 100 kbp	0.87	23.63	16.02	0.85

Table S2. Alignment statistics (related to Supplementary Method S1b)

Alignment statistics for the female and male sample from the four studied species. No. read pairs refer to the number of read pairs remaining after quality trimming. The number of properly aligning reads was calculated with samtools v1.7 flagstat [6]. Sample ID correspond to the raw data files uploaded to NCBI short read archive.

					Properly paired mapped reads (after deduplication)	
Sample ID	Species	Reference genome	Sex	No. read pairs	Before mismatch filtering	<=2 mismatches
QL-1681-95694_S52_L008	Raso lark	Raso lark	female	101453455	174881744	161376774
QL-1681-246_S51_L008	Raso lark	Raso lark	male	87822864	152776522	147354471
QL-1681-19_S46_L006	Eurasian skylark	Raso lark	female	151270768	224364208	111786956
QL-1681-21_S47_L006	Eurasian skylark	Raso lark	male	147067636	227283590	114762381
QF-1504-H-19_S8_L003	Horned lark	Raso lark	female	130242267	180857630	69507055
QF-1504-H-88_S7_L003	Horned lark	Raso lark	male	202099679	275199656	104740879
QF-1504-2KR32024_S2_L001	Bearded reedling	Bearded reedling	female	98173644	138219068	125403224
QF-1504-1ET92164_S3_L001	Bearded reedling	Bearded reedling	male	163911536	230747108	218069639

Table S3. Proportion of reference genomes aligned to the zebra finch genome (related to Supplementary Method S1c)

Statistics showing the length of each reference genome and the proportion that aligns to the zebra finch genome in the synteny analysis. The table show the total length of each genome in addition to the total length of each genome only counting scaffolds larger than 5 kbp. As only scaffolds larger than 5 kbp were kept in the synteny analysis, the proportion of the genome aligning to the zebra finch is based on this value.

Assembly	**Raso lark**	**Eurasian skylark**	**Horned lark**	**Bearded reedling**
Total length (>= 1000 bp)	1003896371	1276061662	1086584534	1020354129
Total length (>= 5000 bp)	979649475	887743951	962416262	991752640
Matching ZF genome	934820000	717725000	842675000	931690000
Prop. of >5kbp scaffolds aligned	0.95	0.81	0.88	0.94

Table S4. Gametolog extraction criteria (related to Supplementary Method S1e)

Z and W gametolog sequences were extracted based on the genotype distributions between the male and female sample.

Variant type	Nr of mutations	Male	Female	Extracted Z allele	Extracted W alle
biallelic	0	0/0	0/0	0	0
	1	0/1	0/0	N	N
	1	0/0	0/1	0	1
	2	0/1	0/1	N	N
	3	1/1	0/1	1	0
	3	0/1	1/1	N	N
	4	1/1	1/1	1	1
triallelic	1	1/2	1/1	N	1
	1	1/1	1/2	1	2
	2	1/2	1/2	N	N
	3	2/2	1/2	2	1
	3	1/2	2/2	N	N
	4	2/2	2/2	2	2

0/0 = no variation in relation to reference genome
0/1 = heterozygote genotype in relation to reference genome
1/1 = homozygote genotype with one allele not found in reference genome
1/2 = heterozygote genotype with two different variants not found in reference genome
2/2 = homozygote genotype not found in reference genome

Table S5. Signatures of sex-linkage across 1 Mbp windows (related to Figure 1)

The mean female-to-male coverage ratio (cov) and difference in number of female-specific mutations and male-specific mutations (snv) of all 1 Mbp windows per chromosome.

Chromosome	Raso lark		Eurasian skylark		Horned lark		Bearded reedling	
	coverage	snv	coverage	snv	coverage	snv	coverage	snv
1	1.03	-55.76	1.03	59.39	1.01	-428.35	1.00	-90.65
1A	1.03	-453.08	1.04	174.25	1.02	265.73	1.01	-193.37
1B	0.98	130.50	1.06	278.00	1.03	-83.00	1.06	37.00
2	1.03	129.38	1.03	91.55	1.02	46.71	1.00	-30.68
3	0.84	13459.84	1.00	4958.29	1.03	2775.37	0.98	1898.89
4	1.02	218.61	1.03	-121.41	1.01	503.89	1.00	-121.34
4A	0.88	13033.77	0.97	9762.91	1.05	10978.41	0.85	10497.82
5	0.83	11011.22	0.93	5102.22	1.02	-90.25	1.01	-122.97
6	1.03	-64.81	1.03	52.16	1.02	43.51	1.01	78.08
7	1.03	119.59	1.03	-65.32	1.02	-11.24	1.02	-64.37
8	1.02	-309.62	1.03	3.86	1.01	-6.00	1.01	-62.31
9	1.03	-727.96	1.03	409.54	1.01	-5353.43	1.02	21.68
10	1.02	950.95	1.03	55.23	1.02	-1323.82	1.02	-171.36
11	1.02	637.95	1.03	47.00	1.01	-18.09	1.02	-275.05
12	1.03	130.96	1.03	-15.91	1.02	74.39	1.02	-9.70
13	1.02	191.17	1.03	-6.50	1.02	-120.17	1.03	-367.50
14	1.02	107.88	1.04	132.82	1.02	99.76	1.03	99.94
15	1.02	-902.73	1.03	9.80	1.02	-8.47	1.03	-188.53
17	1.02	-112.08	1.03	80.69	1.02	140.23	1.04	-448.00
18	1.02	-666.17	1.03	-43.50	1.02	-71.42	1.03	467.58
19	1.02	102.77	1.03	97.54	1.02	64.31	1.04	38.54
20	1.02	-137.18	1.03	18.59	1.02	-19.94	1.04	-65.41
21	1.02	-880.43	1.03	52.14	1.02	-1986.71	1.03	-563.29
22	1.03	246.75	1.03	27.75	1.02	328.00	1.03	716.75
23	1.01	-780.00	1.03	76.00	1.03	-99.57	1.04	-20.29
24	1.03	-355.78	1.03	57.00	1.03	-78.22	1.04	285.00
25	1.03	-99.50	1.07	-681.50	1.04	231.00	1.06	90.50
26	1.02	-198.17	1.05	-1245.50	1.02	-95.50	1.05	103.00
27	1.01	336.33	1.04	93.33	1.03	-228.83	1.04	160.33
28	1.02	-52.50	1.04	187.83	1.03	134.33	1.05	-35.83
Z	0.53	-174.12	0.54	-9719.91	0.54	-3682.34	0.52	-1570.14

Table S6. Sex-linked regions in each of the four species (related to Figure 2).

Mean values for the female-to-male coverage ratio (coverage) and female-to-male difference in number of private SNVs across 0.1 Mbp windows for each stratum. The strata are numbered according to the most parsimonious order of emergence based on phylogenetic analyses (see Main text).

Stratum	Chromosome	Genomic region (Mb)	Stratum size (Mb)		Raso lark	Eurasian skylark	Horned lark	Bearded reedling
1	Z	0-72.9	72.9	Coverage	0.53	0.54	0.54	0.52
				SNV	-17.77	-992.10	-375.85	-160.48
2	4A	0-9.6	9.6	Coverage	0.74	0.91	1.07	0.66
				SNV	2957.35	2227.36	2518.46	2413.16
3	3	8.4-10.4,	8	Coverage	0.75	0.96	1.22	0.69
		18.1-24.1		SNV	3386.68	2435.65	3404.79	2787.17
4	3	10.4-14.0	3.6	Coverage	0.71	0.86	0.98	*NA (1.01)*#
				SNV	3019.06	2161.91	2512.71	*NA (-10.54)*#
5a	3	5.8-8.4,	64.9	Coverage	0.75	1.00	*NA (1.01)*#	*NA (1.00)*#
		14.0-18.1,		SNV	1785.39	451.48	*NA (1.73)*#	*NA (1.77)*#
		29.8-88.0						
5b	5	9.1-45.4	36.3	Coverage	0.71	0.87	*NA (1.01)*#	*NA (1.00)*#
				SNV	1952.04	879.14	*NA (20.19)*#	*NA (-37.92)*#

#Bearded reedling and horned lark have no sex-linkage for stratum 5a and 5b, and bearded reedling not for stratum 4, and are therefore marked with NA in addition to the corresponding values.

Table S7. Nucleotide substitution values for gametologous (Z–W) gene pairs (related to Figure 3)

Nucleotide substitution values for gametologous (Z-W) gene pairs from Raso lark and Eurasian skylark for each sex chromosome strata.

Stratum	Chromosome		No. genes	Median dS	Median dN	Median dN/dS
1	Z	Raso lark	9	0.224	0.030	0.097
		Eurasian skylark	10	0.213	0.030	0.082
2	4A	Raso lark	33	0.080	0.010	0.124
		Eurasian skylark	32	0.085	0.013	0.123
3	3	Raso lark	23	0.073	0.009	0.123
		Eurasian skylark	23	0.072	0.010	0.137
4	3	Raso lark	18	0.032	0.008	0.294
		Eurasian skylark	20	0.036	0.006	0.196
5a	3	Raso lark	168	0.018	0.002	0.116
		Eurasian skylark	186	0.018	0.002	0.115
5b	5	Raso lark	164	0.019	0.004	0.187
		Eurasian skylark	169	0.020	0.004	0.168

Table S8. Nucleotide differentiation between evolutionary strata (related to Figure 3)

P values from Kruskal-Wallis tests of nucleotide differences between pairs of evolutionary strata (for details, see methods).

		Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5a
Analysis of synonymous substitutions (dS)						
Raso lark	Stratum 2	< 0.001				
Eurasian skylark		0.002				
Raso lark	Stratum 3	< 0.001	0.167			
Eurasian skylark		0.001	0.029			
Raso lark	Stratum 4	< 0.001	< 0.001	< 0.001		
Eurasian skylark		0.001	< 0.001	< 0.001		
Raso lark	Stratum 5a	< 0.001	< 0.001	< 0.001	< 0.001	
Eurasian skylark		< 0.001	< 0.001	< 0.001	< 0.001	
Raso lark	Stratum 5b	< 0.001	< 0.001	< 0.001	< 0.001	0.185
Eurasian skylark		< 0.001	< 0.001	< 0.001	< 0.001	0.087
Analysis of non-synonymous substitutions (dN)						
Raso lark	Stratum 2	0.273				
Eurasian skylark		0.66				
Raso lark	Stratum 3	0. 273	0.887			
Eurasian skylark		0. 660	0.986			
Raso lark	Stratum 4	0.142	0.746	0.772		
Eurasian skylark		0.208	0.66	0.66		
Raso lark	Stratum 5a	< 0.001	< 0.001	< 0.001	< 0.001	
Eurasian skylark		0.001	< 0.001	< 0.001	< 0.001	
Raso lark	Stratum 5b	0.001	< 0.001	0.001	0.003	< 0.001
Eurasian skylark		0.007	< 0.001	0.002	0.01	0.002
Analysis of rate of evolution (dN/dS)						
Raso lark	Stratum 2	0.861				
Eurasian skylark		0.874				
Raso lark	Stratum 3	0. 861	0.861			
Eurasian skylark		0. 874	0.874			
Raso lark	Stratum 4	0.105	0.073	0.861		
Eurasian skylark		0.089	0.133	0.874		
Raso lark	Stratum 5a	0.861	0.861	0.861	0.005	
Eurasian skylark		0.874	0.874	0.874	0.125	
Raso lark	Stratum 5b	0.861	0.861	0.861	0.861	0.002
Eurasian skylark		0.535	0.758	0.874	0.874	0.032

Table S9. Rates of evolution between Z gametologs and zebra finch, and W gametologs and zebra finch (related to Results & Discussion 2b)

The table show median substitution rates between zebra finch and Z-linked gametologs (Zlinked) and between zebra finch and W-linked gametologs (Wlinked) for all evolutionary strata that have formed after the split with the zebra finch (i.e. strata 2-5b). P values from paired samples Wilcoxon test, and after Benjamini & Hochberg correction for five tests (one per stratum).

					Median substitution rate to zebra finch			
Stratum	**Chromosome**	**Species**	**Substitution type**	**No. Genes**	**Zlinked**	**Wlinked**	**paired samples Wilcoxon test**	
							p value	**adjusted p value**
2	4A	Raso lark	dN	33	0.005	0.015	0.000	**< 0.001**
		Raso lark	dS	33	0.125	0.115	0.437	0.501
		Eurasian skylark	dN	31	0.008	0.015	0.000	**< 0.001**
		Eurasian skylark	dS	31	0.113	0.119	0.184	0.307
3	3	Raso lark	dN	23	0.007	0.017	0.000	**< 0.001**
		Raso lark	dS	23	0.106	0.095	0.501	0.501
		Eurasian skylark	dN	23	0.007	0.016	0.000	**< 0.001**
		Eurasian skylark	dS	23	0.103	0.101	0.482	0.603
4	3	Raso lark	dN	19	0.012	0.014	0.005	**0.005**
		Raso lark	dS	19	0.108	0.087	0.169	0.281
		Eurasian skylark	dN	19	0.010	0.018	0.003	**0.003**
		Eurasian skylark	dS	19	0.089	0.094	0.768	0.768
5a	3	Raso lark	dN	209	0.014	0.015	0.000	**< 0.001**
		Raso lark	dS	209	0.117	0.117	0.009	**0.024**
		Eurasian skylark	dN	209	0.013	0.014	0.000	**< 0.001**
		Eurasian skylark	dS	209	0.112	0.118	0.000	**< 0.001**
5b	5	Raso lark	dN	184	0.011	0.013	0.000	**0.019**
		Raso lark	dS	184	0.134	0.129	0.004	**0.019**
		Eurasian skylark	dN	184	0.011	0.013	0.000	**< 0.001**
		Eurasian skylark	dS	184	0.128	0.132	0.000	**< 0.001**

Table S10. Tests for enrichment of sex-related genes (related to Figure 4)

Results from binominal tests for significant enrichment of genes involved in a range of sex-related functions (see Methods section 4.6 for details). “Sex-related genes within region” corresponds to the observed genes in Figure 4. Expected genes in Figure 4 was calculated as the total number of sex-related genes within the genome (n = 323) divided by “Proportion of genome”. Genes mentioned in the Discussion section are marked in red.

Chromosome	Region	Proportion of genome		Sex-related genes		P value	Adjusted p value
Strata level analysis		Observed	Estimated (CI)	Genome-wide	Within region		
Z	1	0.058	0.05 (0.029-0.079)	323	16[#]	0.633	1
4A	2	0.008	0.015 (0.005-0.036)	323	5[$]	0.105	0.419
3	3	0.006	0.015 (0.005-0.036)	323	5[&]	0.058	0.289
3	4	0.003	0 (0.000-0.011)	323	0	1	1
3	5a	0.052	0.087 (0.058-0.123)	323	28€	0.008	**0.047**
5	5b	0.029	0.015 (0.005-0.036)	323	5[!]	0.183	0.548
Total sex-linked region per chromosome analysis							
Z	1	0.058	0.05 (0.029-0.079)	323	16	0.633	0.633
4A	2	0.008	0.015 (0.005-0.036)	323	5	0.105	0.314
3	3 + 4 + 5a	0.061	0.1 (0.071-0.140)	323	33	0.005	**0.019**
5	5b	0.029	0.015 (0.005-0.036)	323	5	0.183	0.365

# B4GALT1, CRHBP, DDX4, DMRT1, DMRT3, DNAJA1, FANCC, FANCG, HEXB, KIF2A, LHFPL2, RAD23B, RPS6, SKP2, SPIN1, ZNF366

$ AR, DACH2, DIAPH2, MSN, SEPT6

& FSHR, LBH, MEA1, MSH2, LHCGR (NA in zebra finch)

€ AMD1, ARID4B, CCR6, CGA, CITED2, CYP1B1, ESR1, FOXO3, HSF2, LATS1, MAP3K4, MCM9, MEI4, NA (Uncharacterized protein), PACRG, PGM3, QKI, ROS1, RWDD1, SLC22A16, SRD5A2, STRN, TBPL1, TCF21, TCTE1, UBE2J1, UBR2, UFL1

! CELF1, EIF2B2, SLIRP, TTLL5, TYRO3

Phylopic credits

Turtle and lemur silhouettes: Roberto Díaz Sibaja. Figure available for reuse under the Creative Commons Attribution 3.0 Unported licence. http://creativecommons.org/licenses/by/3.0/

Lizard silhouette: Ghedo and T. Michael Keesey. Figure available for reuse under the Creative Commons Attribution 3.0 Unported licence. http://creativecommons.org/licenses/by/3.0/

Frog silhouette: Pedro de Siracusa. Figure available for reuse under the Creative Commons Attribution 3.0 Unported licence. http://creativecommons.org/licenses/by/3.0/