MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2015) doi: 10.1111/1755-0998.12450

Characterization of the genome and transcriptome of the blue tit *Cyanistes caeruleus*: polymorphisms, sex-biased expression and selection signals

JAKOB C. MUELLER,*[1] HEINER KUHL,†[1] BERND TIMMERMANN† and BART KEMPENAERS*

**Department of Behavioural Ecology & Evolutionary Genetics, Max Planck Institute for Ornithology, 82319 Seewiesen, Germany, †Sequencing Core Facility, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany*

Abstract

Decoding genomic sequences and determining their variation within populations has potential to reveal adaptive processes and unravel the genetic basis of ecologically relevant trait variation within a species. The blue tit *Cyanistes caeruleus* – a long-time ecological model species – has been used to investigate fitness consequences of variation in mating and reproductive behaviour. However, very little is known about the underlying genetic changes due to natural and sexual selection in the genome of this songbird. As a step to bridge this gap, we assembled the first draft genome of a single blue tit, mapped the transcriptome of five females and five males to this reference, identified genomewide variants and performed sex-differential expression analysis in the gonads, brain and other tissues. In the gonads, we found a high number of sex-biased genes, and of those, a similar proportion were sex-limited (genes only expressed in one sex) in males and females. However, in the brain, the proportion of female-limited genes within the female-biased gene category (82%) was substantially higher than the proportion of male-limited genes within the male-biased category (6%). This suggests a predominant on-off switching mechanism for the female-limited genes. In addition, most male-biased genes were located on the Z-chromosome, indicating incomplete dosage compensation for the male-biased genes. We called more than 500 000 SNPs from the RNA-seq data. Heterozygote detection in the single reference individual was highly congruent between DNA-seq and RNA-seq calling. Using information from these polymorphisms, we identified potential selection signals in the genome. We list candidate genes which can be used for further sequencing and detailed selection studies, including genes potentially related to meiotic drive evolution. A public genome browser of the blue tit with the described information is available at http://public-genomes-ngs.molgen.mpg.de.

Keywords: blue tit, brain/gonad transcriptome, *Cyanistes caeruleus*, *de novo* genome, meiotic drive, RNA-seq, sex-biased expression, signatures of selection, SNP calling

Received 7 April 2015; revision received 22 July 2015; accepted 22 July 2015

Introduction

A major goal in evolutionary ecology is to identify and understand adaptive processes from the perspective of genes, individuals and populations with a focus on the individual phenotypic and genomic level (Springer *et al.* 2011; Diz *et al.* 2012). Whereas the phenotypic approach can be used to analyse and detect ongoing processes of adaptation and potential constraints on adaptation by pleiotropy and plasticity, the genomic approach has the potential to reveal adaptive processes over a longer timescale including past selective episodes in the population history (Van Oers & Mueller 2010). The combined investigation of the maintenance and dynamics of phenotypic and genomic variation in populations is important for understanding the potential for trait evolution. However, for many species, the knowledge about genomic evolution substantially lags behind the knowledge about phenotypic evolution.

The blue tit (*Cyanistes caeruleus*) and the closely related great tit (*Parus major*) have been established as ecological model species for more than 60 years (Kluyver 1951; Lack 1964). Since then, fitness consequences of variation in many life history and behavioural traits have been described in the wild (Merilä & Sheldon 1999; Ellegren & Sheldon 2008; Clutton-Brock & Sheldon

Correspondence: Jakob C. Mueller, Fax: ++49-8157-932-400; E-mail: mueller@orn.mpg.de

[1] Shared first authorship.

2010). In blue tits, a specific focus has been on extra-pair mating behaviour (for recent publications see Vedder *et al.* 2011; Schlicht & Kempenaers 2013; Arct *et al.* 2013). However, only a few studies of genotype–phenotype or genotype–fitness relationships have been performed (Olano-Marin *et al.* 2011a,b; Kluen *et al.* 2012 and Steinmeyer *et al.* 2012). This is mainly due to the paucity of genomic information in this species; so far, only random genomic markers – although large sets – have been developed (e.g. Olano-Marin *et al.* 2010). Genome and transcriptome sequencing will help to bridge this gap.

High-throughput (next-generation) sequencing techniques enable the generation of new avian genome databases (Zhang *et al.* 2014), which can then be improved with the help of 'reference' genome assemblies such as that of the chicken (*Gallus gallus*) and zebra finch (*Taeniopygia guttata*). These reference genomes have a chromosomal scale and have been assembled with the help of genetic linkage maps (ICGSC 2004; Warren *et al.* 2010). Reference species facilitate the assembly of new bird genomes, particularly because chromosomal organization appears to be highly conserved among avian taxa (Ellegren 2010; Frankl-Vilches *et al.* 2015). Recently, bird genomes that have attracted particular interest are those of species with clear categories of domesticated breeds (Dalloul *et al.* 2010), morphotypes (Shapiro *et al.* 2013) or subspecies (Ellegren *et al.* 2012; Poelstra *et al.* 2014), or that are adapted to extreme environments or lifestyles (Cai *et al.* 2013; Huang *et al.* 2013; Zhan *et al.* 2013). However, the publication of genome information on ecological model species should be a helpful long-lasting resource for micro-evolutionary studies in the field.

The general aim of this study was to provide a detailed description of the blue tit genome, transcriptome and genetic polymorphisms in order to promote genetic studies on ecologically relevant traits in the field. Here, we report on (i) the assembly of the blue tit genome, assisted by the Tibetan ground tit (*Pseudopodoces/Parus humilis*) and chicken genome, and (ii) the characterization of the blue tit transcriptome, based on sequencing of different tissue samples of five males and five females. We analysed gene expression in the female and male gonads and brain in comparison with the pooled remaining tissues to (iii) describe general patterns of sex-differential gene expression in brain and gonads and to (iv) list potential candidate genes for sexual conflict and/or rapid evolution (Parsch & Ellegren 2013). After identifying genomic variants in the transcribed regions, we used their population frequencies, patterns of linkage disequilibrium and functional annotation to (v) reveal potential loci under selection.

Materials and methods

Sampling

We collected 10 adult blue tits (five males and five females) under licence (Regierung von Oberbayern, Az. 55.1-8642.3-45-2011) from a wild population in a natural forest reserve in southern Germany (Westerholz, 48°08′N, 10°53′E; for details of the area see Steinmeyer *et al.* 2012). We caught all individuals inside a nestbox at the beginning of the breeding season, between 28 March and 3 April 2012, while they were inspecting a box or roosting inside. Individuals were brought to the laboratory, sacrificed and directly dissected to extract brain, gonads, skin, pectorals, heart, lung, liver, gizzard, spleen, pancreas, intestine and kidney. Halves of each organ were directly frozen or stabilized in RNAlater (Qiagen). A blood sample of one of the male birds (reference bird BT333_1) was stored in Queen's lysis buffer (Seutin *et al.* 1991) for DNA extraction.

DNA/RNA extraction and sequencing

DNA was extracted from the blood sample with the Nucleospin Blood QuickPure kit (Macherey-Nagel) and RNA was extracted from the organ samples with RNeasy Mini kits (Qiagen) according to the manufacturer's recommendations. In each sample, we measured RNA concentration with a Bioanalyzer (Agilent Technologies). RNA samples were pooled into three tissue groups per individual for sequencing: total brain, gonads and all other remaining organs (equal RNA amounts).

The genome of the single male reference blue tit (BT333_1) was sequenced by the Illumina HiSeq 2000 technique using a 500-bp and a 1000-bp insert paired-end library with 101-bp read length. The transcriptome samples of the 10 blue tit individuals (three tissue samples per individual) were separately prepared and filtered for RNAs with poly-A tails using Illumina's TruSeq RNA Sample Preparation Kits before sequencing by the Illumina HiSeq paired-end technique with read lengths of 50 or 101 bp (for the reference individual BT333_1) and an insert size of about 200 bp. For sequencing details including barcodes for each sample and SRA accession numbers see Table S1 (Supporting information).

Genome assembly and gene annotation

We filtered and trimmed the whole genome shotgun raw sequences of blue tit BT333_1 to obtain high-quality nonduplicate read pairs. This was achieved by a custom script that extracted only the largest parts of reads that were longer than 32 bp and contained no base with

Phred quality score lower than 11 (Frankl-Vilches *et al.* 2015). We compared the first 32 bp of each read in a pair to those of other read pairs and kept only one pair if the same sequence was found more than once (deduplication). The reads were assembled by a hybrid strategy which involved (i) the assembly of all short reads by a 'de Bruijn graph' assembler using IDBA-UD v1.1.1 (Peng *et al.* 2012), and (ii) assembling the resulting contigs and a subset of the short read data by a 'overlap layout consensus' long read assembler using CELERA ASSEMBLER v7 (Myers *et al.* 2000). For more details regarding parameters applied see Appendix S1 (Supporting information). Some putative interchromosomal missassemblies were identified by whole genome alignment using LAST v266 (Kielbasa *et al.* 2011; Appendix S1, Supporting information) with the Tibetan ground tit (*Pseudopodoces/Parus humilis*; Cai *et al.* 2013; GenBank Accession no. ANZD00000000). First, the Tibetan ground tit scaffold identifiers were assigned to chicken chromosomes by whole genome alignment. The so identified missassemblies were removed by splitting the corresponding blue tit scaffolds. Superscaffolds were subsequently established by colinearity with the Tibetan ground tit genome as described in Frankl-Vilches *et al.* (2015). The final blue tit assembly was compared with the chicken assembly using LAST to assign putative chromosome IDs to the superscaffolds. Additionally, we assigned the blue tit superscaffolds to zebra finch chromosomes (Table S2, Supporting information) in order to take into account the characteristic chromosome splits in the passerine lineage (e.g. on chromosomes 1 and 4).

Protein-coding sequences in the blue tit genome were annotated by aligning all avian protein sequences of the NCBI and Ensembl protein database (reference date: 19 May 2014) using SPALN v2.1.2 (Gotoh 2008; Appendix S1, Supporting information; see track 'Aligned Proteins' in the genome browser). A single best protein match for each cluster of gene predictions in the genome was selected according to the best SPALN alignment score (which summarizes alignment length and identity of a match, thus preferring more complete predictions from less diverged species). We used two steps to remove redundant transcripts that had lower SPALN scores based on genomic location. First, we removed transcripts with strand-specific exact overlap of single exons between two or more transcript models. Second, we removed residual redundant transcripts that had strand-specific, but not exact, overlaps (see track 'Best Protein Match' in the genome browser). Gene names were assigned by these alignments. We further refined the coding gene annotation – in particular the UTR annotation – by merging SPALN predictions (which had been converted to gtf format) with the RNA-seq/Cufflinks-derived transcript models of the male BT333_1 using Cuffmerge (Trapnell *et al.* 2010). As coding sequence annotation is lost during this step, we used TransDecoder r20131117 (https://transdecoder.github.io/) to reassign ORF information for protein translation (see 'Protein Coding Genes' track in the genome browser).

Transcriptome mapping and assembly

The transcriptome sequence reads (RNA-seq) of each tissue sample from each individual were mapped to the genome sequence using TOPHAT2 V2.0.9 (Kim *et al.* 2011; Appendix S1, Supporting information), with support of a splice-junction database from earlier versions of genome annotations (we used predicted exon/intron boundaries from the SPALN protein alignments as well as exon/intron boundaries from the first round of TOPHAT using the 101-bp RNA-seq reads from BT333_1). Parsimonious sets of transcripts for each tissue type were assembled using CUFFLINKS v2.1.1 (Trapnell *et al.* 2010; Appendix S1, Supporting information). This step provides a complete picture of the transcriptome, because it takes the untranslated regions of genes into account, and considers differential splicing and poly-adenylated long noncoding RNA and miRNA precursors (Guttman *et al.* 2009; Necsulea & Kaessmann 2014; Philippe *et al.* 2014). The Cufflinks-predicted exons mapped to the genome were first merged across the three tissues within each individual using Cuffmerge and then intersected across groups of individuals using BEDTOOLS V2.19.1 (Quinlan & Hall 2010) to reveal the following four groups: (i) exons common to all males for the comparison of RNA-seq and DNA-seq variants in the single reference male, (ii) exons common to all individuals, (iii) exons common to all females (but not found in all males) and (iv) exons common to all males (but not found in all females). According to the three last groups, we categorized SNPs identified from the transcriptome as common to both sexes or as detected exclusively in the female- or male-specific transcriptome.

Sex-specific expression analysis

We tested for sex-differential expression of the 67 359 Cufflinks-predicted genes in the three tissue samples using Cuffdiff (Trapnell *et al.* 2013; Appendix S1, Supporting information). This algorithm estimates expression as FPKM (fragments per kilobase of transcript per million mapped reads) at transcript/isoform-level resolution and controls for variability across replicates (five males and five females). We allowed a false discovery rate of 0.05. Among the genes that showed sex-biased expression, we discriminated between those that were expressed in both sexes and those that were only expressed in either males or females (sex-limited expression; FPKM <= 0.0001 in one sex). The proportion of annotated genes in the sex-biased and unbiased category was estimated by the relative number of Cufflinks-predicted genes, which showed positional overlap with best protein hit regions, using BEDTOOLS.

Variant calling and annotation

After removing duplicate reads (using PICARD V1.122, Appendix S1, Supporting information), we called SNPs and indels (insertions/deletions) and assigned individual genotypes from both the single-individual DNA-seq data and the 10-individual RNA-seq data using the HaplotypeCaller of GATK v3.2 (DePristo *et al.* 2011; Appendix S1, Supporting information). The genotypes of the RNA-seq data were determined for each of the 10 individuals in a single step following GATK best practices (Van der Auwera *et al.* 2013; Appendix S1, Supporting information). The variant sets of both the DNA-seq and RNA-seq data were then filtered according to fixed rules from the GATK best practices recommendations (Van der Auwera *et al.* 2013; Appendix S1, Supporting information) and to coverage (included if read depth per individual > 6). We also excluded high-coverage variants (read depth > 99th percentile of base coverage) from the genomic set to avoid variant calls from duplicated genome regions. To filter out potential false-positive calls due to mapping errors, we allowed a maximal indel length of 10 bp.

Variants were functionally annotated according to the gene position using SNPEFF v3.5 (Cingolani *et al.* 2012). Here, gene position was defined by the refined protein annotation ('Protein Coding Genes' track in genome browser).

Scan for selection signatures

We applied three different methods for the detection of potential selection signals in the RNA-seq SNP data. Our focus was on methods which are able to detect ongoing selection in small genomic regions, but also included a standard selection scan of allele frequency spectra. Only SNPs that were scored in all 10 individuals were used, because allele and haplotype frequencies need to be estimated, and SNPs located in the homologous region of chromosome Z were excluded due to its hemizygotic nature in females. In all tests, we applied a window size of 25 kb, which is close to the average size of avian genes (see e.g. chicken genome at genome.ucsc.edu). The genes identified as candidates for selection were tested for functional enrichment in gene ontology (GO) categories using g:Profiler (Reimand *et al.* 2011) with the 11 827 annotated *G. gallus* Ensemble genes as background/null distribution (Ensemble genomes 22). *P*-values were corrected for multiple testing by the native method implemented in g:Profiler.

1. Tajima's *D* based on unphased genotypes was calculated for all adjacent 25-kb windows within each chromosome/scaffold using VCFTOOLS v0.1.11 (Danecek *et al.* 2011). Clusters of a minimum of two adjacent windows with $|D| > 1.96$ were considered interesting. This test is mostly sensitive to historical selective sweeps and balancing selection.
2. All SNPs were phased and imputed within each chromosome/scaffold based on the physical sequence reads and population frequency information using SHAPEIT v2.790 (Delaneau *et al.* 2013). This method is based on the Gibbs' sampling scheme in which each individual's haplotypes are sampled conditionally on both the current estimates of all other individuals and the phase-informative sequencing reads of the individual (Delaneau *et al.* 2013). The latter increases the reliability of the estimated short-distance haplotypes. The estimation of long-distance haplotypes, however, is more error-prone given our small sample of 10 individuals, particularly across regions with no SNPs. After phasing, we calculated the standardized log-ratio of the integrated extended haplotype homozygosities of both alleles at each SNP (iHS according to Voight *et al.* 2006) using the R package REHH V1.11 (Gautier & Vitalis 2012). The iHS test statistic should be mostly independent of local variation in recombination rates, because both the reference and the alternate allele are located at the same genomic position. It is standardized within the frequency classes of the minor allele in 5% bins between 10% and 50%. SNPs with minor allele frequency <10% were excluded. The iHS statistic has an approximately standard normal distribution and provides a measure of how unusual the haplotype extensions around a given SNP are, relative to the haplotype frequencies. As simulations have shown that selection produces clusters of extreme iHS values (Voight *et al.* 2006), we considered a minimum of 10 absolute iHS values > 3.09 (0.1% level) within any window of 25 kb indicative for a partial selective sweep.
3. For the annotated gene regions, we also considered clusters of common nonsynonymous SNPs (NS-SNPs) as interesting; that is, we determined all cases of at least seven NS-SNPs with minor allele frequency > 0.35 within any 25 kb window. Such cases might indicate balancing selection maintaining high structural protein variation or ongoing (partial) selective sweeps on haplotypes with strong impact on structural variation.

All these selection scans are restricted to genomic regions transcribed in both sexes. However, the detection power might still depend on the local transcriptome coverage and therefore SNP density in these regions. To assess the severity of this potentially confounding factor, we tested the strength of the correlations between Tajima's *D* or mean absolute iHS values and transcript coverage across all neighbouring 25-kb windows. Transcript coverage and SNP density in these windows were correlated ($r = 0.42$; 95% CI: 0.41–0.44).

Results

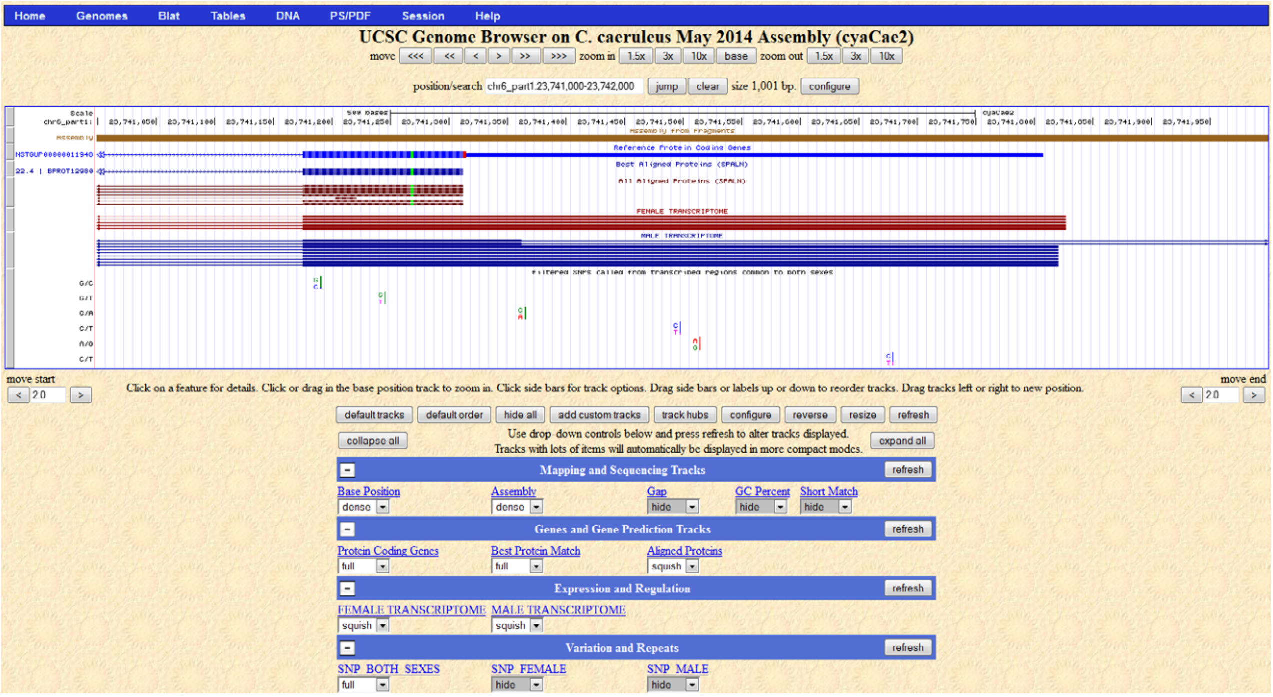
Genome

DNA sequencing yielded more than 54 Gb in 543 million high-quality reads. The resulting draft assembly comprised about 1099 Mb in 54 108 contigs with an N50 length of 85.7 kb, or – after reference-assisted scaffolding – in 29 311 scaffolds with an N50 length of 16.8 Mb. Median base coverage of the assembly was 47× (5th percentile: 15×; 95th percentile: 68×).

Ninety-one per cent of the blue tit assembly (394 superscaffolds) could be mapped to chicken chromosomes or linkage groups. All chromosomes or linkage groups of the assembled chicken genome aligned with the blue tit genome except the smaller chromosome 32 and the sex chromosome W. The latter was not sequenced with our single male blue tit.

We found 21 434 best protein gene hits on the genome, which comprised 37% of the genome size (including introns). Alignment sites of these hits (180 549 protein-coding exons) covered 2.6% of the genome size (for comparison see e.g. chicken genome at genome.ucsc.edu). Using our transcriptome data, we obtained complete gene annotations including UTRs for 17 062 of the coding genes defined above, which increased the annotated genome size to 43% (6% in exons).

The public blue tit genome browser can be found at http://public-genomes-ngs.molgen.mpg.de. Figure 1 shows an example screen shot of the browser at the gene end of MKI67. Known protein-coding genes are annotated in the 'Best Protein Match' and 'Aligned Proteins' track and transcribed regions can be inferred from the 'TRANSCRIPTOME' tracks. By combining the information of these tracks, as done in the 'Protein Coding Genes' track, one can identify UTRs. The 'MAF2' track shows the alignments with the ground tit, medium ground finch (*Geospiza fortis*) and zebra finch. The current chromosome naming refers to chicken chromosomes. Assignments to zebra finch chromosomes can be found in Table S2 (Supporting information). Identified SNPs are visible in the 'SNP' tracks. The browser has the characteristic functionalities of a UCSC browser (genome.ucsc.edu) including BLAT searches, DNA sequence extraction and table browser. Use the help button for further information.



UCSC Genome Browser on C. caeruleus May 2014 Assembly (cyaCae2)

position/search chr6_part1:23,741,000-23,742,000 size 1,001 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing Tracks
Base Position: dense
Assembly: dense
Gap: hide
GC Percent: hide
Short Match: hide

Genes and Gene Prediction Tracks
Protein Coding Genes: full
Best Protein Match: full
Aligned Proteins: squish

Expression and Regulation
FEMALE TRANSCRIPTOME: squish
MALE TRANSCRIPTOME: squish

Variation and Repeats
SNP_BOTH_SEXES: full
SNP_FEMALE: hide
SNP_MALE: hide

Fig. 1 Example screen shot of the blue tit browser showing the 3′ end of gene MKI67.

Transcriptome

RNA sequencing yielded between 91 and 195 million high-quality reads per individual and tissue sample (brain, gonads and pooled other organs). Merging all Cufflinks transcript predictions of all individuals and tissues revealed 67 359 predicted genes/transcribed regions (including coding and noncoding 'genes', see Necsulea & Kaessmann 2014) and 209 472 isoforms. Most (82%) of the best protein hits overlapped one of the Cufflinks-predicted genes. However, only 33% of the Cufflinks-predicted genes overlapped with best protein hits, leaving a substantial number of transcripts nonannotated and/or nonprotein coding.

In addition, we report the overlapping/intersecting genomic regions of Cufflinks-predicted exons common to both sexes, specific to females and specific to males. This is useful for filtering SNPs for which genotype information is available for all 10 individuals or only for a single sex. A total of 155 340 exons (58.5 Mb) were common to all individuals, 18 813 exons (7.5 Mb) were specific to females, and 37 477 exons (17.2 Mb) were specific to males.

Sex-specific expression

As expected, a high number of genes were differentially expressed between testis and ovaries: 22 389 of 42 128 expressed genes showed a significant sex bias (53%, including sex-limited genes; Table 1; Fig. 2). More genes showed a higher or exclusive expression in males than in females (Table 1). The proportion of Z-chromosomal vs. autosomal location was larger for the male-biased (8%) than for the female-biased (4%) genes (χ^2 = 176.8, d.f. = 1, P < 0.0001). A relatively low proportion of the genes expressed in gonads overlapped best protein hits in both the unbiased (52%) and sex-biased genes (56%), indicating that a high number of genes expressed in the gonads remain nonannotated.

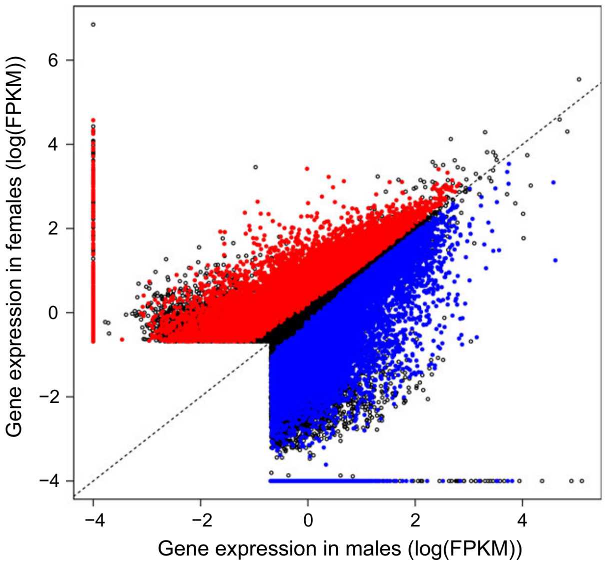


Fig. 2 Sex-specific expression of 42 128 genes in the gonads of blue tits. Genes with low numbers of read alignments in both sexes (<10 fragment alignments) were not tested (lower left corner in plot). The lowest expression in a single sex was set to 0.0001 FPKM [=−4 log(FPKM), fragments per kilobase of transcript per million mapped reads] and indicates genes with sex-limited expression. Significant male-biased expression is shown in blue and significant female-biased expression in red. The dotted line indicates equal expression. Nonsignificance can also be due to large variance in expression between individuals of one sex.

Table 1 Number of differentially expressed genes in the three tissue samples of blue tits. Percentages within the respective higher level group are given in brackets

	Gonads	Brain	Pooled other tissues
Expressed genes (total)	42 128	24 431	22 761
Male-biased	12 071 (29%)	325 (1.3%)	84 (0.4%)
Male-limited	2183 (18%)	20 (6%)	19 (23%)
Z-chromosomal	926 (8%)	206 (63%)	8 (10%)
Female-biased	10 318 (24%)	278 (1.1%)	202 (0.9%)
Female-limited	1025 (10%)	227 (82%)	141 (70%)
Z-chromosomal	363 (4%)	23 (8%)	13 (6%)

In the brain, we found 603 of 24 431 tested genes that were expressed differentially between the sexes (2.5%; Table 1; Fig. 3). Although there was no difference in the number of genes with male-biased vs. female-biased expression (χ^2 = 3.55, d.f. = 1, P = 0.059), the number of genes that were expressed only in females (227) was significantly larger than the number of genes exclusively expressed in males (20; χ^2 = 350.0, d.f. = 1, P < 0.0001). Thus, most of the genes with male-biased expression belong to the group of genes which are expressed in both sexes (see Fig. 3). Of the 603 genes that showed sex-biased expression in the brain, 229 were found on the Z-chromosome. The majority of these genes (206) showed a male-biased expression (sex difference: χ^2 = 344.4, d.f. = 1, P < 0.0001). Those genes that showed a sex-biased expression were less likely to be annotated with best protein hits (52%) than the unbiased genes (71%). This difference was due to the very low annotation rate among genes with female-biased expression (14% vs. 85% for male-biased genes).

In the pooled samples from the other organs, as expected, only few genes (286 of 22 761 tested genes)

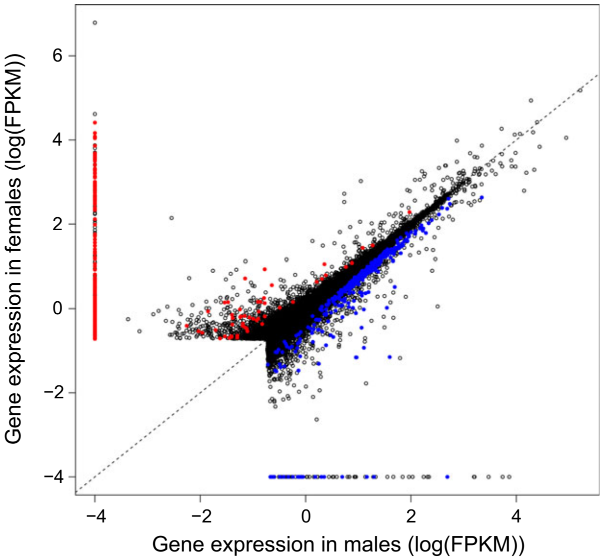


Fig. 3 Sex-specific expression of 24 431 genes in the blue tit brain. Genes with low numbers of read alignments in both sexes (<10 fragment alignments) were not tested (lower left corner in plot). The lowest expression in a single sex was set to 0.0001 FPKM [=−4 log(FPKM), fragments per kilobase of transcript per million mapped reads] and indicates genes with sex-limited expression. Significant male-biased expression is shown in blue and significant female-biased expression in red. The dotted line indicates equal expression. Nonsignificance can also be due to large variance in expression between individuals of one sex.

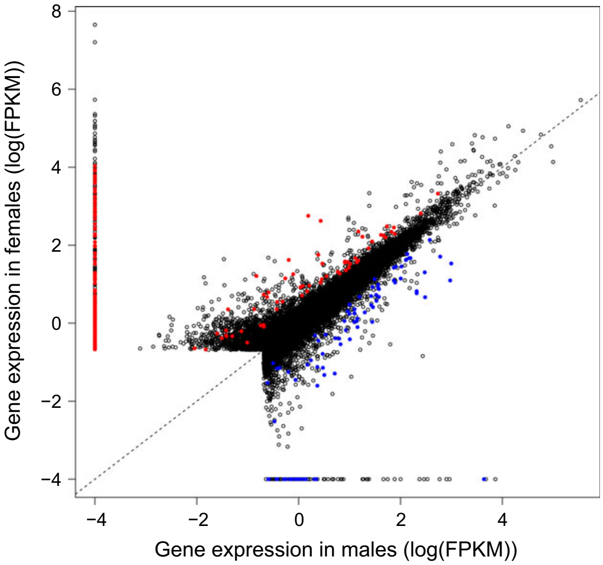


Fig. 4 Sex-specific expression of 22 761 genes in the pooled sample of other organs of blue tits. Genes with low numbers of read alignments in both sexes (<10 fragment alignments) were not tested (lower left corner in plot). The lowest expression in a single sex was set to 0.0001 FPKM (= −4 log(FPKM), fragments per kilobase of transcript per million mapped reads) and indicates genes with sex-limited expression. Significant male-biased expression is shown in blue and significant female-biased expression in red. The dotted line indicates equal expression. Nonsignificance can also be due to large variance in expression between individuals of one sex.

showed sex-differential expression (1.3%; Table 1; Fig. 4). Among these, genes with female-biased expression were more common than genes with male-biased expression (χ^2 = 48.2, d.f. = 1, P < 0.0001). There is no evidence for enrichment of Z-chromosomal genes within the male-biased gene group in comparison with the female-biased one (χ^2 = 0.8, d.f. = 1, P = 0.36). Again, genes that showed sex-biased expression were less likely to be annotated with best protein hits (39%) than the unbiased genes (75%). And again, this difference was mainly driven by the low annotation rate among genes with female-biased expression (27% vs. 68% for male-biased genes).

Characterization of variants

After filtering, we identified 3 964 922 heterozygous sites of single nucleotide changes (SNPs) from the genomic sequences in the single male blue tit BT333_1. This translates to one SNP every 277 base pairs. As expected from the general gene size and structure in birds (UCSC browser chicken), 43% of the polymorphisms were located in genic regions (exons and introns; Table 2). There are more 3′ UTR than 5′ UTR SNPs, as expected because of the larger size of 3′ UTRs. The nonsynonymous SNPs, with potentially strong functional impacts, sum up to 13 973 SNPs.

Table 2 Genomic variants determined from DNA-seq data of the reference blue tit according to their location and functional type. Variants can appear in multiple exonic subcategories due to overlapping and nested gene structure. Percentages of the main categories are given in brackets

Variant type	SNPs	Indels
Intergenic	2 243 928 (57%)	308 270 (58%)
Exon	167 551 (4%)	19 251 (4%)
Nonsynonymous	13 973	
Synonymous	44 553	
Frameshift		693
Nonframeshift		386
5′ UTR	24 184	2987
3′ UTR	103 150	15 601
Intron (rest)	1 553 443 (39%)	202 979 (38%)

We identified 530 500 heterozygous occurrences of insertions or deletions (indels) based on the genomic sequence in the reference bird (Table 2). The distribution across exons, introns and intergenic regions appears to be similar to that of the SNPs, except that the number of indels in coding regions is relatively low (only 6% of exonic indels). This is probably due to selection against indels with strong impact on the protein structure (local codon changes and/or frame shifts).

Based on the transcriptomic sequences, we detected 435 143 SNPs in the Cufflinks regions which are common to both sexes. The majority of these SNPs were located in annotated exons and most of these are in the 3′ UTR, in agreement with relatively large 3′ UTRs (Table 3). The high proportion of annotated SNPs was also due to the fact that annotated genes were often larger than nonannotated ones. Still, a substantial number of SNPs (70 204) were located distally to known exons, that is in currently nonannotated transcripts. A total of 46 741 and 20 387 SNPs were detected in the male- and female-specific Cufflinks regions, respectively. Notably, the majority of sex-specific SNPs were located in nonannotated transcribed regions. Also, within the annotated regions, the proportion of nonsynonymous SNPs is relatively high among the sex-specific SNPs (Table 3).

Comparison between genomic and transcriptomic variant calling

We identified variants (heterozygotes) for the reference male bird (BT333_1) both from DNA and RNA sequences and were thus able to compare the calling success from the two different sources. We identified 235 006 SNPs from the DNA sequence data in comparison with 211 269 SNPs from the RNA sequence data in the common male Cufflinks regions (187 370 exons across 82.9 Mb). Eighty per cent of the genomic SNPs were also found in the transcriptomic SNP set and 89% of the transcriptomic SNPs were validated by genomic SNPs. The called genotypes were identical in 99.99% of the overlapping SNP set, indicating a high mapping congruence between RNA-seq and DNA-seq data.

We identified 27 586 indels in the common male Cufflinks regions from the genomic data and 17 913 indels from the transcriptomic data. Forty-one per cent of the genomic indels were also found in the transcriptomic indel set and 63% of the transcriptomic indels were detected in the genomic set. Indels identified in both sequence sets had identical genotypes in 92% of the cases. Due to this low cross-validation rate between genomic and transcriptomic indels and the relatively low number of transcriptomic indels in comparison with the SNPs, we excluded indels from further selection analyses.

Selection signatures in the transcriptomic regions

Mean Tajima's *D* across all 16,489 informative 25-kb windows (with > 2 SNPs) was −0.53, indicating an overall excess of low-frequency variants (Fig. S1, Supporting information). This most likely represents a signal of purifying selection in the transcriptomic regions, but in theory could also be due to population expansion. There was no association of Tajima's *D* with transcript coverage ($r = -0.01$; 95% confidence interval CI: −0.02 to 0.004) across 25-kb windows. Only four clusters of strong negative Tajima's *D* (2 adjacent windows with $D < -1.96$) were identified as potential regions for historical selective sweeps: chr2_9:750 000–775 000 (gene SNRNP48), chr4_6:3 775 000–3 800 000 (genes PHKA1, HDAC8), chr14_1:16 775 000–16 800 000 (gene KCTD5) and chr21_1:4 625 000–4 650 000 (genes SLC35E2B, CDC2L1, MMP23B). This gene list is not enriched for any GO category. No clusters of strong positive Tajima's *D*, indicative of balancing selection, were detected.

Relative extended haplotype homozygosities measured as iHS values are shown across all informative chromosomes and scaffolds (with > 4 SNPs) in Fig. S2 (Supporting information). There was a weak, but significant, negative association of iHS means with transcript coverage ($r = -0.05$; 95% CI: −0.06 to −0.03) across all 25-kb windows, indicative of a weak confounding effect. Based on our definition of clusters of extreme iHS values (see Materials and methods), we identified 15 regions with unusual haplotype structure (Table 4). The most extreme SNPs within these clusters are mostly common 3′ UTR SNPs or SNPs from nonannotated regions, except for two common nonsynonymous SNPs in the MKI67 and DNMT3A gene. The haplotype breakdowns of the

Table 3 Transcriptomic SNPs in blue tits according to their location and functional type in the reference protein annotation for the three Cufflinks categories (transcripts expressed in both sexes, expressed in males only and expressed in females only). Percentages of the main categories are given in brackets

SNP type	Both sexes	Male-specific	Female-specific
Exon	343 649 (79%)	15 576 (33%)	5002 (25%)
Nonsynonymous	35 920	3247	1020
Synonymous	91 896	4565	1827
5′ UTR	14 722	3052	405
3′ UTR	201 111	4712	1750
Nonannotated	91 494 (21%)	31 165 (67%)	15 385 (75%)
Up/downstream (5 kb off exon)	21 290	5653	3041
Non-up-/downstream	70 204	25 512	12 344

Table 4 Genomic windows with clusters of strong iHS selection signals identified from SNPs in the transcripts common to both sexes. Entries are ordered according to max |iHS|. Although the search for high iHS clusters was performed across all possible 25-kb windows, the size indicated here is the restricted window in which all SNPs with iHS > 3.09 were located

Position	Size (bp)	Genes	Proportion of SNPs with |iHS| > 3.09	Max |iHS| (position)	Frequency of major allele
Chr2_2 (3 267 143–3 269 466)	2323	–	12/34	6.94 (3 267 266)	0.60
Chr2_5 (11 793 224–11 798 603)	5379	FAM135B	17/85	−6.41 (11 798 568, 3′ UTR)	0.55
Chr2_1 (18 761 958–18 763 148)	1190	TPK1	12/27	6.04 (18 761 958, 3′ UTR)	0.60
Chr2_6 (10 958 063–10 961 897)	3834	–	10/24	5.94 (10 961 620)	0.60
Chr6_1 (23 739 440–23 741 191)	1751	MKI67	11/19	−5.66 (23 739 768, CDS, NS)	0.65
Chr1_1 (4 124 737–4 128 617)	3880	Hypothetical protein; micro RNA cluster	20/34	−5.42 (4 127 210, intron; flanking)	0.70
Chr2_3 (4 397 557–4 400 396)	2839	C2H8ORF37	13/34	−5.25 (4 400 333, 3′ UTR)	0.55
Chr1_9 (1 000 981–1 002 289)	1308	ETNK1	14/21	−5.22 (1 001 189, 3′ UTR)	0.55
Chr1_4 (19 825 405–19 828 229)	2824	PRKX	14/34	−5.15 (19 825 991, 3′ UTR)	0.75
Chr5_1 (24 999 583–25 000 364)	781	DNMT3A	11/18	−5.01 (25 000 298, CDS, NS)	0.55
Chr14_1 (4 254 439–4 255 759)	1320	–	10/21	4.74 (4 255 757)	0.85
Chr2_1 (32 895 288–32 898 683)	3395	SLC35B3	13/38	−4.71 (32 895 723, 3′ UTR)	0.70
Chr1_3 (376 895–378 471)	1576	TMEM139	10/13	−4.63 (376 895, 3′ UTR)	0.50
Chr1_3 (26 867 588–26 871 222)	3634	N6AMT1	11/14	4.54 (26 867 588, intron or 3′ UTR)	0.55
Chr5_1 (14 687 758–14 690 266)	2508	–	13/33	−4.30 (14 689 424)	0.70

two alleles at the five SNPs with the highest iHS values are shown in Figs S3–S7 (Supporting information). The major or relatively common minor allele of these SNPs extends in an unusually frequent haplotype (see extended haplotype homozygosity (EHH) plots and bifurcation plots S3–S7). The genes listed in Table 4 are not enriched for any GO category.

We found 2695 common nonsynonymous SNPs with a minor allele frequency of more than 0.35. Some of these common NS-SNPs formed clusters of 7 SNPs within 25 kb (Table 5). It appears that the clusters show a general enrichment in NS-SNPs (common and noncommon). Interestingly, one of the clusters with a high density of NS-SNPs overlaps with one of the extreme iHS clusters and represents the MKI67 gene (Fig. 1). The genes associated with common NS-SNP clusters are enriched for the GO category 'centrosome localization' ($P = 0.027$; genes ASPM and SYNE2). All NS-SNPs within the three clusters of the mentioned genes (MKI67, ASPM, SYNE2) do not show heterozygote excess, are in regions of average genomic coverage and in regions of perfect alignment with known proteins, indicating a high reliability of annotation.

Discussion

We characterized the genome sequence and transcriptome of a common songbird, the blue tit. The data – available via a genome browser – can be used for studies on the genetic architecture and evolutionary potential of ecologically relevant traits. Information and patterns solely based on sequences can be utilized to identify genomic regions of interest, which might then be linked to the evolution of ecologically relevant traits via trait mapping or the analysis of functional pathways. Detecting signals of ongoing selection is of particular interest for micro-evolutionary processes on an ecological timescale (Vitti *et al.* 2013). An additional approach for the detection of interesting genes in the context of sexual selection and sex-specific behaviour is the analysis of sex-biased expression, because genes that show sex-biased expression are known to evolve rapidly (Parsch & Ellegren 2013).

Genome and sex-specific transcriptome characterization

In a first step, we assembled the blue tit genome with high sequence continuity, that is with large superscaffolds (N50 of 16.8 Mb). This was possible because a high-quality draft genome of the related ground tit was available (Cai *et al.* 2013; Qu *et al.* 2013). Similar to the ground tit, a high proportion of the blue tit genome (91%) could be mapped to all chicken chromosomes except two smaller chromosomes. This is not unexpected given the general high synteny observed among bird species (Ellegren 2013).

In a second step, the transcriptome of five female and five male birds was assembled and mapped to the reference genome. Whereas the majority (52–75%) of the predicted genes that expressed at similar levels in both

Table 5 Genomic windows with clusters of common (>0.35) nonsynonymous SNPs (NS-SNPs) identified from all SNPs in the transcripts common to both sexes. Entries are ordered according to number of common NS-SNPs. Although the search for common NS-SNP clusters was performed across all possible 25-kb windows, the size indicated here is the restricted window in which all common NS-SNPs were located

Position	Size (bp)	Genes	Number of common NS-SNPs/ noncommon NS-SNPs/ S-SNPs
Chr3_5 (1 020 415–1 034 835)	14 420	GVINP1	44/124/122
Scaffold1812_1 (4334–8810)	4476	LOC101806664	30/22/50
Chr17_1 (8 312 834–8 320 044)	7210	LOC101816715	13/48/21
Chr2_4 (11 718 875–11 754 418)	35 543	CEP192	11/31/46
Chr6_1 (23 729 867–23 739 961)	10 094	MKI67	10/42/24
Chr14_1 (2 639 521–2 648 307)	8786	SLX4	9/12/6
Chr8_1 (8 994 539–9 012 440)	17 901	ASPM	9/52/35
Chr1_1 (41 521 487–41 527 470)	5983	DDIAS	8/23/16
Chr2_1 (22 790 125–22 800 493)	10 368	KIAA0947	8/33/11
Chr3_2 (8 800 604–8 821 307	20 703	MIA3	8/11/24
Chr7_3 (290 285–294 282)	3997	LOC101820543	8/7/6
Chr1_12 (1 793 558–1 800 896)	7338	SON or LOC101920289	7/48/57
Chr11_6 (1873–5970)	4097	(HYDIN)	7/2/2
Chr19_1 (5 727 003–5 734 228)	7225	SPAG5	7/21/13
Chr2_3 (2 390 388–2 391 459)	1071	FBXO43	7/1/4
Chr3_1 (5 467 340–5 483 714)	16 374	MDN1	7/30/21
Chr5_1 (7 033 045–7 045 444)	12 399	SYNE2	7/16/21
Chr6_1 (5 727 574 –5 732 587)	5013	CYP2W1	7/8/10

sexes could be mapped to annotated protein-coding genes, only 39–56% of the genes showing sex-biased expression were annotated with aligned protein sequences. This could indicate that sex-biased expression often involves noncoding, regulatory elements of the genome, such as long noncoding RNAs (Reinius *et al.* 2010; Buckberry *et al.* 2014) or microRNAs (Marco 2014). If this is true, the vast majority of studies on sex-biased gene expression, which only focus on protein-coding genes (Ellegren & Parsch 2007), may represent only a part of the sex-biased transcriptome. Given that some differentially spliced isoforms might have been missed during the annotation process, it could also mean that some of the sex-biased transcripts represent alternatively spliced exons. Such sex differences in splicing are widespread in the adult human brain (Trabzuni *et al.* 2013). The regulatory system behind sex-specific expression certainly represents an open field for further studies (Mank *et al.* 2013).

In contrast to the pooled organ sample which showed only 1.3% of genes with sex-biased expression, a high proportion of genes expressed in testes and ovaries showed a significant sex bias (53%). Similarly high percentages of genes with sex-biased expression have been observed in the adult reproductive tissues of other organisms, including chicken (Ellegren & Parsch 2007; Mank *et al.* 2010). Whereas there was little difference in the number of genes with male- vs. female-biased expression across all autosomes, male-biased genes were significantly enriched on the Z-chromosome. This finding is consistent with incomplete dosage compensation in the homogametic (ZZ) males (Storchova & Divina 2006; Parsch & Ellegren 2013). However, the majority of genes with sex-biased expression appear to be located on the autosomes and hence require a different mechanism for sex-specific expression.

The brain transcriptome also showed a higher proportion of genes with sex-biased expression (2.5%) than the transcriptome from the pooled other organs. This proportion is similar to what has been found in the zebra finch and common whitethroat (*Sylvia communis*) brain (Naurin *et al.* 2011). Among the 325 genes with male-biased expression, which comprise about half of the genes with sex-biased expression, the majority (63%) was linked to the Z-chromosome. In contrast, only a small number of genes with female-biased expression was Z-linked (8%). Interestingly, most of the genes with female-biased expression in the brain were female-limited (expressed exclusively in females), whereas the majority of genes with male-biased expression were expressed in both sexes. This pattern across both autosomes and the Z-chromosome shows some similarity with the expression pattern on the Z-chromosome found in the zebra finch and the common whitethroat brain, where Z-linked male bias in gene expression increased with mean expression level across the sexes (Naurin *et al.* 2012). Male bias can positively correlate with overall expression level when female-limited genes (i.e. with a negative male bias) are included. We thus hypothesize that the sexual dimorphism in gene expression in the avian brain is based on two distinct components: a qualitative female-limited expression of mostly ncRNA genes and a quantitative male-biased expression of mostly coding genes. The male-biased expression is Z-linked and most likely represents chromosomal dosage differences. The exclusive expression of genes in females could provide an explanation for their higher evolutionary rates in comparison with genes with male-biased expression as observed in the chicken brain (Mank *et al.* 2007).

Selection signals in polymorphism data

Identifying genomic variants from RNA-seq data is seen as an important add-on to transcriptome assembly and expression analysis, but it also poses a challenge because of the intrinsic complexity in the transcriptome (Piskol *et al.* 2013; Quinn *et al.* 2013). We called the RNA-seq variants according to the best practice suggestions of the genome analysis toolkit GATK (Van der Auwera *et al.* 2013 and associated web pages) and were able to evaluate them with the variants identified from DNA-seq data on the same individual. The cross-validation of our RNA-seq and DNA-seq SNP calling suggests that the reliability was sufficiently high (more than 80% of SNPs called from one data set were also called in the other data set and the identified genotypes were nearly identical in the overlapping SNP set). The difference in called SNPs from RNA-seq and DNA-seq data can be due to different procedures in mapping, duplicate removal, splice-site alignment and variant filtering, but can also result from differences in read coverage between RNA-seq and DNA-seq data. We assume that allele-specific expression bias is of minor importance in individually called genotypes and does not influence allele frequency estimates. However, the indels identified by RNA-seq and DNA-seq data were quite different. Indel calling is sensitive to local realignment steps, and the guidelines for indel calling from RNA-seq data are still preliminary. We thus disregarded the indels for further analyses on selection signals.

We utilized different polymorphism information to identify selection signals in the genome: allele frequency spectra, EHH and common nonsynonymous coding SNPs. The test statistics (Tajima's *D* and iHS) appeared to be only weakly confounded by local genomic features such as the density of transcribed regions. However, because iHS is partly based on error-prone long-distance haplotype estimations, we suggest that the identified genomic regions and genes only serve as candidates for further evaluating studies on recent or ongoing adaptive processes in blue tit populations. The gene lists obtained through the first two approaches were not enriched for any specific functional annotation compared to the complete gene composition. However, the list of genes with clusters of common nonsynonymous SNPs was functionally enriched for the term 'centrosome localization'. The products of the identified genes *ASPM* and *SYNE2* play important roles in the establishment and maintenance of centrosome localization in the cytosol during mitosis and meiosis. The dynamic evolution of meiotic drive of selfish gene elements predicts rapid evolution of proteins associated with chromosome movements during meiosis (Henikoff *et al.* 2001; Axelsson *et al.* 2010). The enrichment of common nonsynonymous SNPs could thus indicate an ongoing or partial selective sweep of structural variants in genes associated with chromosome separation, for which centrosome positioning plays a role. The observation of incomplete selective sweeps is not unlikely if repeated selective events related to the evolution of meiotic drive are predicted. However, it could also indicate balancing selection among different structural variants of the genes. For the *ASPM* gene, which shows positive selection during primate evolution (Mekel-Bobrov *et al.* 2005), functions like the regulation of brain and gonad development have been reported (Pulvers *et al.* 2010).

One gene (*MKI67*) was identified in two different approaches and showed clusters of extreme EHH at common nonsynonymous SNPs. It is therefore a good candidate for a partial selective sweep on structural variants. The gene product of *MKI67* is a well-known marker for cell proliferation in cancer medicine, but little is known about its molecular function (Scholzen & Gerdes 2000). Because of its centromere localization during specific stages in the mitotic and meiotic cell cycle (Traut *et al.* 2002), it might also have an important function during meiotic chromosome movements, similar to *ASPM* and *SYNE2*. It therefore represents another candidate for repeated meiotic drive evolution.

Acknowledgements

We thank Lisa Trost, Agnes Tuerk, Antje Girndt, Silke Laucht, Kim Teltscher and Christine Baumgartner for help in the field and laboratory, and the editor Anna Santure and three anonymous reviewers for helpful comments on the manuscript. This work was funded by the Max Planck Society.

References

Arct A, Drobniak SM, Podmokła E, Gustafson L, Cichoń M (2013) Benefits of extra-pair mating may depend on environmental conditions – an experimental study in the blue tit (*Cyanistes caeruleus*). *Behavioral Ecology and Sociobiology*, **67**, 1809–1815.

Axelsson E, Albrechtsen A, van AP *et al.* (2010) Segregation distortion in chicken and the evolutionary consequences of female meiotic drive in birds. *Heredity*, **105**, 290–298.

Buckberry S, Bianco-Miotto T, Bent SJ, Dekker GA, Roberts CT (2014) Integrative transcriptome metaanalysis reveals widespread sex-biased gene expression at the human fetal–maternal interface. *Molecular Human Reproduction*, **20**, 810–819.

Cai Q, Qian X, Lang Y *et al.* (2013) Genome sequence of ground tit *Pseudopodoces humilis* and its adaptation to high altitude. *Genome Biology*, **14**, R29.

Cingolani P, Platts A, Wang LL *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, **6**, 1–13.

Clutton-Brock TH, Sheldon BC (2010) Individuals and populations: the role of long-term, individual-based studies in ecology and evolutionary biology. *Trends in Ecology and Evolution*, **25**, 562–573.

Dalloul RA, Long JA, Zimin AV *et al.* (2010) Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biology*, **8**, e1000475.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**:2156–2158.

Delaneau O, Howie B, Cox A, Zagury JF, Marchini J (2013) Haplotype estimation using sequence reads. *American Journal of Human Genetics*, **93**, 787–796.

DePristo M, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Diz AP, Martinez-Fernandez M, Rolan-Alvarez E (2012) Proteomics in evolutionary ecology: linking the genotype with the phenotype. *Molecular Ecology*, **21**, 1060–1080.

Ellegren H (2010) Evolutionary stasis: the stable chromosomes of birds. *TREE*, **25**, 283–291.

Ellegren H (2013) The evolutionary genomics of birds. *Annual Review of Ecology and Systematics*, **44**, 239–259.

Ellegren H, Parsch J (2007) The evolution of sex-biased genes and sex-biased gene expression. *Nature Reviews Genetics*, **8**, 689–698.

Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in wild populations. *Nature*, **452**, 169–175.

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, **491**, 756–760.

Frankl-Vilches C, Kuhl H, Werber M *et al.* (2015) Using the canary genome to decipher the evolution of hormone-sensitive gene regulation in seasonal singing birds. *Genome Biology*, **16**, 19. doi:10.1186/s13059-014-0578-9.

Gautier M, Vitalis R (2012) rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics*, **28**, 1176–1177.

Gotoh O (2008) A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Research*, **36**, 2630–2638.

Guttman M, Amit I, Garber M *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **12**, 223–227.

Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science*, **293**, 1098–1102.

Huang Y, Li Y, Burt DW *et al.* (2013) The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nature Genetics*, **45**, 776–783.

International Chicken Genome Sequencing Consortium (ICGSC) (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**, 695–716.

Kielbasa SM, Wan R, Sato K, Horton P, Frith MC (2011) Adaptive seeds tame genomic sequence comparison. *Genome Research*, **21**, 487.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2011) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, **14**, R36.

Kluen E, Kuhn S, Kempenaers B, Brommer JE (2012) A simple cage test captures intrinsic differences in aspects of personality across individuals in a passerine bird. *Animal Behaviour*, **84**, 279–287.

Kluyver HN (1951) The population ecology of the Great Tit, *Parus m. major* L. *Ardea*, **39**, 1–135.

Lack D (1964) A long-term study of the great tit (*Parus major*). *Journal of Animal Ecology*, **33**, 159–173.

Mank JE, Hultin-Rosenberg L, Axelsson E, Ellegren H (2007) Rapid evolution of female-biased, but not male-biased, genes expressed in the avian brain. *Molecular Biology and Evolution*, **24**, 2698–2706.

Mank JE, Nam K, Brunstroem B, Ellegren H (2010) Ontogenetic complexity of sexual dimorphism and sex-specific selection. *Molecular Biology and Evolution*, **27**, 1570–1578.

Mank JE, Wedell N, Hosken DJ (2013) Polyandry and sex-specific gene expression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **368**, 20120047.

Marco A (2014) Sex-biased expression of microRNAs in *Drosophila melanogaster*. *Open Biology*, **4**, 140024.

Mekel-Bobrov N, Gilbert SL, Evans PD *et al.* (2005) Ongoing adaptive evolution of ASPM, a brain size determinant in *Homo sapiens*. *Science*, **309**, 1720–1722.

Merilä J, Sheldon BC (1999) Genetic architecture of fitness and nonfitness traits: empirical patterns and development of ideas. *Heredity*, **83**, 103–109.

Myers EW, Sutton GG, Delcher AL *et al.* (2000) A whole-genome assembly of *Drosophila*. *Science*, **287**, 2196–2204.

Naurin S, Hansson B, Hasselquist D, Kim Y-H, Bensch S (2011) The sex-biased brain: sexual dimorphism in gene expression in two species of songbirds. *BMC Genomics*, **12**, 37.

Naurin S, Hasselquist D, Bensch S, Hansson B (2012) Sex-biased gene expression on the avian Z chromosome: highly expressed genes show higher male-biased expression. *PLoS One*, **7**, e46854.

Necsulea A, Kaessmann H (2014) Evolutionary dynamics of coding and non-coding transcriptomes. *Nature Review Genetics*, **15**, 734–748.

Olano-Marin J, Dawson DA, Girg A *et al.* (2010) A genome-wide set of 106 microsatellite markers for the blue tit (*Cyanistes caeruleus*). *Molecular Ecology Resources*, **10**, 516–532.

Olano-Marin J, Mueller JC, Kempenaers B (2011a) Heterozygosity and survival in blue tits (*Cyanistes caeruleus*): contrasting effects of presumably functional and neutral loci. *Molecular Ecology*, **20**, 4028–4041.

Olano-Marin J, Mueller JC, Kempenaers B (2011b) Correlations between heterozygosity and reproductive success in the blue tit (*Cyanistes caeruleus*): an analysis of inbreeding and single locus effects. *Evolution*, **65**, 3175–3194.

Parsch J, Ellegren H (2013) The evolutionary causes and consequences of sex-biased gene expression. *Nature Reviews Genetics*, **14**, 83–87.

Peng Y, Leung HC, Yiu SM, Chin FY (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420–1428.

Philippe N, Samra EB, Boureux A *et al.* (2014) Combining DGE and RNA-sequencing data to identify new polyA+ non-coding transcripts in the human genome. *Nucleic Acids Research*, **42**, 2820–2832.

Piskol R, Ramaswami G, Li JB (2013) Reliable identification of genomic variants from RNA-seq data. *American Journal of Human Genetics*, **93**, 641–651.

Poelstra JW, Vijay N, Bossu CM *et al.* (2014) The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, **344**, 1410–1414.

Pulvers JN, Bryk J, Fish JL *et al.* (2010) Mutations in mouse Aspm (abnormal spindle-like microcephaly associated) cause not only microcephaly but also major defects in the germline. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16595–16600.

Qu Y, Zhao H, Han N *et al.* (2013) Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau. *Nature Communications*, **4**, 2071.

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

Quinn EM, Cormican P, Kenny EM *et al.* (2013) Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. *PLoS One*, **8**, e58815.

Reimand J, Arak T, Vilo J (2011) g:Profiler – a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research*, **39**, W307–W315.

Reinius B, Shi C, Hengshuo L *et al.* (2010) Female-biased expression of long non-coding RNAs in domains that escape X-inactivation in mouse. *BMC Genomics*, **11**, 614.

Schlicht E, Kempenaers B (2013) Effects of social and extra-pair mating on sexual selection in blue tits (*Cyanistes caeruleus*). *Evolution*, **67**, 1420–1434.

Scholzen T, Gerdes J (2000) The Ki-67 protein: from the known and the unknown. *Journal of Cellular Physiology*, **182**, 311–322.

Seutin G, White BN, Boag PT (1991) Preservation of avian blood and tissue samples for DNA analysis. *Canadian Journal of Zoology*, **69**, 82–90.

Shapiro MD, Kronenberg Z, Li C *et al.* (2013) Genomic diversity and evolution of the head crest in the rock pigeon. *Science*, **339**, 1063–1067.

Springer SA, Crespi BJ, Swanson WJ (2011) Beyond the phenotypic gambit: molecular behavioural ecology and the evolution of genetic architecture. *Molecular Ecology*, **20**, 2240–2257.

Steinmeyer C, Kempenaers B, Mueller JC (2012) Testing for associations between candidate genes for circadian rhythms and individual variation in sleep behaviour in blue tits. *Genetica*, **140**, 219–228.

Storchova R, Divina P (2006) Nonrandom representation of sex-biased genes on chicken Z chromosome. *Journal of Molecular Evolution*, **63**, 676–681.

Trabzuni D, Ramasamy A, Imran S *et al.* (2013) Widespread sex differences in gene expression and splicing in the adult human brain. *Nature Communications*, **4**, 2771.

Trapnell C, Williams BA, Pertea G *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**, 511–515.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, **31**, 46–53.

Traut W, Endl E, Scholzen T, Gerdes J, Winking H (2002) The temporal and spatial distribution of the proliferation associated Ki-67 protein during female and male meiosis. *Chromosoma*, **111**, 156–164.

Van der Auwera GA, Carneiro M, Hartl C *et al.* (2013) From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, **43**, 11.10.1–11.10.33.

Van Oers K, Mueller JC (2010) Evolutionary genomics of animal personality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **365**, 3991–4000.

Vedder O, Komdeur J, van der Velde M, Schut E, Magrath MJL (2011) Polygyny and extra-pair paternity enhance the opportunity for sexual selection in blue tits. *Behavioral Ecology and Sociobiology*, **65**, 741–752.

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annual Review of Genetics*, **47**, 97–120.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biology*, **4**, e72.

Warren WC, Clayton DF, Ellegren H *et al.* (2010) The genome of a songbird. *Nature*, **464**, 757–762.

Zhan X, Pan S, Wang J *et al.* (2013) Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nature Genetics*, **45**, 563–568.

Zhang G, Li C, Li Q *et al.* (2014) Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, **346**, 1311–1320.

J.C.M. conceived of the study, participated in the study design and coordination, performed the differential expression, selection and statistical analyses and drafted the manuscript. H.K. participated in the study design and coordination, assembled and annotated the genome, mapped the transcriptome and helped draft the manuscript. B.T. participated in the study design and coordination. B.K. conceived of the study, participated in the study design and coordination and revised the manuscript. All authors read and approved the final manuscript.

Data Accessibility

The blue tit genome browser is available at http://public-genomes-ngs.molgen.mpg.de/. The raw sequences are deposited at the NCBI/SRA archive under bioproject no. PRJNA284903. Individual accession numbers are listed in Table S1. Files of annotated SNPs, sex-differential expression results and selection scans can be found under Dryad doi:10.5061/dryad.dc56b.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Command lines with parameters used for assembly, mapping and variant calling/selection.

Fig. S1 Histogram of all Tajima's *D* values.

Fig. S2 Absolute iHS values across all informative chromosomes/scaffolds (different colours; only larger chromosomes named).

Fig. S3 (A) Breakdown of EHH over physical distance for both alleles of the focal SNP chr2_2:3 267 266 (dashed line). (B) Bifurcation plots showing the haplotype-breakdown on both sides of the focal SNP: major allele (60%) in blue, minor allele (40%) in red.

Fig. S4 (A) Breakdown of EHH over physical distance for both allels of the focal SNP chr2_5:11 798 568 in gene FAM135B (dashed line). (B) Bifurcation plots showing the haplotype-breakdown on both sides of the focal SNP: major allele (55%) in blue, minor allele (45%) in red.

Fig. S5 (A) Breakdown of EHH over physical distance for both allels of the focal SNP chr2_1:18 761 958 in gene TPK1 (dashed line). (B) Bifurcation plots showing the haplotype-breakdown on both sides of the focal SNP: major allele (60%) in blue, minor allele (40%) in red.

Fig. S6 (A) Breakdown of EHH over physical distance for both allels of the focal SNP chr2_6:10 961 620 (dashed line). (B) Bifurcation plots showing the haplotype-breakdown on both sides of the focal SNP: major allele (60%) in blue, minor allele (40%) in red.

Fig. S7 (A) Breakdown of EHH over physical distance for both allels of the focal SNP chr6_1:23 739 768 in gene MKI67 (dashed line). (B) Bifurcation plots showing the haplotype-breakdown on both sides of the focal SNP: major allele (65%) in blue, minor allele (35%) in red.

Table S1 Sequencing details, barcodes and accession numbers.

Table S2 Assignment of blue tit superscaffolds to zebra finch chromosomes.