Rapid gene content turnover on the germline-restricted chromosome in songbirds

Stephen Schlebusch (stephen.schlebusch@gmail.com)
Charles University https://orcid.org/0000-0003-2355-2652

Jakub Rídl
Czech Academy of Sciences

Manon Poignet
Charles University https://orcid.org/0000-0003-2445-830X

Francisco Ruiz-Ruano
Uppsala University https://orcid.org/0000-0002-5391-301X

Jiri Reif
Charles University in Prague; Palacký University Olomouc

Jan Pačes
Institute of Molecular Genetics of the Czech Academy of Sciences https://orcid.org/0000-0003-3059-6127

Tomas Albrecht
Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic https://orcid.org/0000-0002-9213-0034

Alexander Suh
Department of Evolutionary Biology, Uppsala University https://orcid.org/0000-0002-8979-9992

Radka Reifová
Charles University https://orcid.org/0000-0001-5852-5174

Article

Keywords:

Posted Date: March 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1359388/v1

Version of Record: A version of this preprint was published at Nature Communications on July 29th, 2023. See the published version at https://doi.org/10.1038/s41467-023-40308-8.

Abstract

The germline-restricted chromosome (GRC) of songbirds represents an extraordinary and taxonomically widespread example of programmed DNA elimination. Despite its apparent indispensability in songbirds, we still know very little about the GRC's genetic composition, function, and evolutionary significance. Here we assembled the GRC in two closely related species, the common and thrush nightingale. We identified 585 and 406 genes on the GRC of each species, respectively, many of them present in multiple copies. Interestingly, the GRC gene content differed dramatically between the two species, with only 192 genes being shared despite only 1.8 million years of species divergence. The chromosome appears to be under little selective pressure, with most GRC genes being present in pseudogenized fragments. Only one gene, *cpeb1*, had a complete coding region in all examined individuals of the two species and showed no copy number variation. The addition of this gene to the GRC corresponds with the earliest estimates of the GRC origin, making it a good candidate for the functional indispensability of the GRC in songbirds.

Introduction

In multicellular organisms, all cells of an individual normally contain the same genetic information. There are exceptions, however, where certain sequences are eliminated from all or some of the somatic cells during development, leaving the original genetic information to be maintained in the germ cells (Wang and Davis, 2014; Suh and Dion-Côté, 2021). An interesting example of this programmed DNA elimination has been described in songbirds, where a whole chromosome is lost from somatic cells early on in embryo development. The aptly named germline-restricted chromosome (GRC) was described for the first time in the zebra finch (*Taeniopygia guttata*) (Pigozzi and Solari 1998), with recent studies suggesting that it likely occurs in all songbirds (order Passeriformes, suborder Oscines) (Torgasheva, et al. 2019; Kinsella, et al. 2019). Songbirds diverged from the rest of the birds approximately 47 mya (Oliveros, et al. 2019) and comprise approximately 50% of all modern bird species, making them the largest taxonomic group with obligatory programmed DNA elimination. Despite the relatively wide distribution of the GRC, we still know very little about its genetic composition, evolutionary significance and function for birds.

Besides its exclusive presence in the germline, there is little that is consistent about this chromosome. The GRC is normally maternally inherited, but paternal inheritance has been shown to be possible (Pei, et al. 2022). It occurs in a single copy in male germ cells, which is excluded from the nucleus during meiosis, and in two copies in female germ cells (Pigozzi and Solari 2005), although again, there are exceptions (Malinovskaya, et al. 2020; Torgasheva, et al. 2021). The chromosome size varies dramatically, from the largest macrochromosome in the cell (macro-GRC) to a small microchromosome (micro-GRC), with no apparent phylogenetic pattern (Torgasheva, et al. 2019). This lack of conservation is in clear contrast to the apparent ubiquity of the GRC in songbirds.

Part of the reason why there is still so much unknown about the genetic composition of the GRC is that this chromosome is hard to sequence effectively. The GRC sequence is largely composed of recently diverged paralogous sequences from the other chromosomes in the cell (hereafter referred to as A

chromosomes), and as such can be hard to differentiate in a sequencing library (Itoh et al. 2009; Biederman et al. 2018; Kinsella, et al. 2019; Asalone, et al. 2021). In addition, gonads are composed of both somatic and germ cells, so this chromosome is only found in a subset of testis cells and a minimal proportion of ovary cells (Kinsella, et al. 2019). Thus, GRC sequences are underrepresented in the sequencing libraries from these tissues. Assembled sequence information from the GRC is scarce and currently limited to *T. guttata* (Itoh et al. 2009, Biedermann et al. 2018, Kinsella et al. 2019 and Pei, et al, 2022). Thus far, analyses of tissue-specific single-nucleotide polymporphisms (SNPs) and germline/somatic coverage differences have identified 269 putative genes as well as many high copy number regions on the *T. guttata* macro-GRC (Kinsella, et al. 2019; Asalone, et al. 2021). However, the total assembled length of GRC-linked sequences is 1.24 Mb (Kinsella, et al. 2019) plus 468 kb (Asalone, et al. 2021), which is approximately 1% of the expected 150 Mb *T. guttata* macro-GRC.

In this paper, we sequenced and assembled the GRC in two closely related songbird species, the common nightingale (*Luscinia megarhynchos*) and thrush nightingale (*L. luscinia*), both of which possess a micro-GRC (Poignet et al., 2021). These species from the Muscicapidae family diverged approximately 1.8 mya (Storchová et al. 2010) and still hybridize in a secondary contact zone (Reifová, et al. 2011; Mořkovský, et al. 2018; Albrecht, et al. 2019). Using a novel method to identify GRC reads from germline sequencing libraries, we assembled the majority of the GRCs for both species. Our results show rapid gene content turnover with significant differences observed not only between species but even among individuals of the same species. The vast majority of genes on the GRC were only partially present and presumably non-functional. The gene *cpeb1* was the only entire gene present in all individuals with no copy number variation. We show that this gene belongs to the oldest genes on the GRC, making it the standout candidate gene with an essential function on the GRC, which might be preventing the loss of the GRC in songbirds.

Results

GRC size estimation using meiotic spreads

We visualized the pachytene chromosomes in testis cells using antibodies against the synaptonemal complex (anti-SYCP3) and centromere (CREST) (see Fig. 1). These antibodies enable the identification of the unpaired, univalent GRC (del Priore and Pigozzi 2014; Torgasheva, et al. 2019). In addition, we immunostained the eliminated GRC from the secondary spermatocytes in the form of a micronucleus (see Supplementary Fig. 1) using an antibody against histone H3 lysine 9 methylation (H3K9me) (del Priore and Pigozzi 2014).

Both species had a GRC comparable in size with a microchromosome (i.e. a micro-GRC) as was described in Poignet et al. (2021). Consistent with this, the GRC micronucleus for both nightingale species was much smaller than in species with a macro-GRC (Supplementary Fig. 1; see del Priore and Pigozzi 2014 for visualization of the GRC micronucleus in *T. guttata* with a macro-GRC).

The length of the GRC was estimated by measuring the size of the 22 largest chromosomes, as well as the GRC, in the pachytene cells and comparing the sizes with assembled chromosome lengths (in bp) in collared flycatcher (*Ficedula albicollis*), a songbird species that diverged from nightingales 15 mya (Jetz, et al. 2012). Given the conservation of bird karyotypes (Kawakami et al. 2014), we assumed that chromosome lengths would be similar between *F. albicollis* and nightingales (see Supplementary Fig. 2). Using this approach, we estimated the GRC size to be 9.6 Mbp for *L. megarhynchos* and 7.5 Mbp for *L. luscinia*.

GRC assembly

To identify GRC-derived sequences, we sequenced and compared somatic and germline genomes in three individuals of each species. One individual from each species was sequenced with 10x Chromium linked-read sequencing and two individuals with standard Illumina technology. The GRC was assembled using (i) 10x linked reads that aligned in a germline-specific way to the germline genome assembly, (ii) reads that contained germline-specific SNPs, (iii) reads containing germline-specific repetitive elements, which was only applicable for *L. megarhynchos*, and (iv) any 10x linked reads that shared their 10x barcodes with reads selected in previous steps (Supplementary Fig. 3). Approximately 23 thousand 10x barcodes were identified, resulting in 5.6 million read pairs to assemble the *L. megarhynchos* GRC. In comparison, only 13 thousand 10x barcodes were identified in *L. luscinia*, which resulted in 3 million read pairs. Despite having fewer reads, the *L. luscinia* GRC assembly was longer (5.6 Mbp) and of higher quality (see Table 1) than the *L. megarhynchos* assembly (3.5 Mbp). While the GRC assemblies were highly fragmented, their cumulative length suggests that a large proportion of each GRC was assembled (36–75% of the estimated size). This number is however probably an underestimate, as it does not take recent within-GRC duplicated sequences into account (see “Recent copy number variation within the GRC” below).

Table 1
GRC assembly metrics for each nightingale species.

	L. megarhynchos	*L. luscinia*
Total Length	3.5 Mbp	5.6 Mbp
Largest Scaffold	110 Kbp	370 Kbp
Number of Scaffolds	1400	750
Scaffold N50	2.8 Kbp	46 Kbp
Contig N50	2.5 Kbp	24 Kbp
%N	3.3%	6.4%

Recent copy number variation within the GRC

The coverage across the GRC was calculated for the three individuals from each species. This was done in order to identify regions of the GRC that had been duplicated within the GRC but have not diverged sufficiently from their GRC paralogs to be differentiated by the assembly process. *L. megarhynchos* showed more near-identical duplications merged in the assembly process compared to *L. luscinia*. This was reflected by the higher average normalised GRC copy number in *L. megarhynchos* (3.0x) than in *L. luscinia* (1.6x). Importantly, this mostly accounts for recent within-GRC duplications and repetitive elements on the GRC, which the genome assembler was unable to differentiate. Older within-GRC duplications which have diverged in sequence are not captured in this metric and are expected to be assembled into separate paralogous sequences.

The lower proportion of near-identical duplications in the *L. luscinia* GRC at least partially explains the higher assembly quality in this species. This is supported by the fact that the longest scaffolds from *L. luscinia* consistently had low copy number (see Fig. 2). It is also interesting to note that there is considerable variation in copy number among individuals. This is especially noticeable in *L. megarhynchos* but is also present in the *L. luscinia* scaffolds with higher copy numbers (see Fig. 2). This suggests that there is substantial variation in recent within-GRC duplications between the GRCs, even within species.

The normalised copy number calculation allows for the estimation of the GRC size, accounting for near-identical duplicated sequences erroneously collapsed in the assembly, for each individual. Once these duplications are taken into account, the size of the GRC assembly is 1.3x-2.9x times larger than the original assembly suggested (see Fig. 3). The GRC assembly of the individuals sequenced by 10x linked reads was 10.2 Mbp long after this correction for *L. megarhynchos* and 7.0 Mbp for *L. luscinia*. These numbers are similar to the GRC size estimates from the meiotic spreads (9.6 Mbp and 7.5 Mbp, respectively), suggesting that we assembled the vast majority of the GRC in both species. And again, the variation among individuals within the same species is noticeable, with the *L. megarhynchos* GRC ranging in size from 8.8 Mbp to 12.5 Mbp and the *L. luscinia* GRC ranging from 7.0 Mbp to 11.6 Mbp. This variation is especially impressive considering that two of the three individuals in each species are not assembled, instead using the third individual as a reference, and will therefore underestimate any sequence that does not have a homologous region in the reference GRC.

Genetic content of the GRC

The variation in GRC size between the species and individuals may reflect highly different genetic content in the GRCs. To explore this possibility, we aligned the two GRC assemblies against each other as well as to the *L. megarhynchos* genome. Surprisingly, approximately only 1 Mbp of each GRC assembly aligned to the other GRC (before accounting for coverage; see Fig. 3), the rest being species-specific. Consistent with this, the A-chromosomal origins of each species' GRC sequences are strikingly different (see Fig. 4). Curiously, the part of the GRC which is orthologous between the two species was disproportionately repetitive, with the average normalised copy number in these orthologous regions being 5.2x and 3.1x for

L. megarhynchos and *L. luscinia* respectively, while the non-homologous regions had an average respective copy number of 2.1x and 1.4x.

The *L. luscinia* GRC has a large proportion paralogous to Chromosome 2, which is absent in the *L. megarhynchos* GRC assembly. Most of this Chromosome 2 derived sequence comes from a single region of Chromosome 2. This paralogous region is visible in many of the largest scaffolds from the *L. luscinia* assembly (see Supplementary Fig. 4). Despite being present in large blocks, and likely originating from a single A-to-GRC duplication, it is no longer continuous, presumably as a result of internal rearrangements and within-GRC duplications or else later A-to-GRC duplications. For example, there are two large scaffolds with Chromosome 2 ancestry that are clearly within-GRC duplicates (Scaffolds 9 and 10, Supplementary Fig. 4), but because this duplication happened long enough ago, their sequences have sufficiently diverged for the assembly process to distinguish them (see Fig. 2).

Divergence from A Chromosomes

The proportion of mismatches between homologous regions of each GRC and the A chromosomes of each species was calculated to determine if the GRC sequences originated before or after the divergence of the two nightingale species (see Fig. 5). Interestingly, a large proportion of the *L. megarhynchos* GRC appears to have originated after speciation, with approximately 1 Mbp of the assembly aligning better to its A-chromosomal paralogs than to the *L. luscinia* genome. In comparison, virtually all of the *L. luscinia* GRC appears to predate the divergence of the species, with no noticeable difference in mismatches between the GRC and the A-chromosomal paralogs from both species.

Gene Annotation

Genes were annotated on each of the GRC assemblies using *F. albicollis* protein-coding genes (FicAlb1.5), resulting in 585 identified genes in *L. megarhynchos* and 406 in *L. luscinia*. As might be expected given the largely different sequence origins of the two GRCs, the majority of genes were not shared between the two species, with only 192 genes being present in both species. Genes were assessed for their completeness (see Table 2) and copy number within the GRC (Supplementary Table 1). Notably, the vast majority of identified genes were both duplicated (with the average gene having a corrected copy number of 6.6x in *L. megarhynchos* and 3.7x in *L. luscinia*) and only partially present (with only 23 genes in *L. megarhynchos* and 18 genes in *L. luscinia* having more than 95% of the coding region present in the assembly). This observed gene fragmentation was measured after correcting for a possible lack of sequence conservation between the nightingales and *F. albicollis*, which used the percentage of the gene that was found on the A chromosomes as a baseline of expected conservation (see Materials and Methods; Supplementary Table 1 contains gene-specific details). Ten of the genes with 95% of their coding sequence present were shared between the two nightingale species, from which six were previously reported on the *T. guttata* GRC (Kinsella, et al, 2019).

Table 2

Number of GRC genes at varying levels of fragmentation identified in each nightingale species. A gene was counted as being shared between the two nightingales at a particular level of fragmentation if the proportion that was present was sufficient in both species. For genes shared between nightingales and *T. guttata*, the level of fragmentation had to be met for both nightingale species, and the gene had to be reported as putatively on the *T. guttata* GRC (Kinsella, et al. 2019).

Proportion of coding region found	Number of genes in *L. megarhynchos*	Number of genes in *L. luscinia*	Number of shared genes	Number of genes shared between nightingales and *T. guttata*
> 0%	585	406	192	25
> 25%	150 (26%)	99 (24%)	57 (30%)	13 (52%)
> 50%	81 (14%)	63 (16%)	39 (20%)	11 (44%)
> 75%	49 (8%)	36 (9%)	24 (13%)	9 (36%)
> 95%	22 (4%)	17 (4%)	10 (5%)	6 (24%)

While most genes on the GRC were species-specific, the genes that were shared had a similar percentage of their coding regions present in the two GRCs (see Fig. 6). The portion of these genes that was present on the two GRCs was also often the same portion (see Supplementary Fig. 5). This consistency suggests that the presence of these shared genes is the result of their presence on an ancestral GRC, rather than independent additions in each lineage. The consistency in the degree of fragmentation of these shared genes (see Fig. 6), as well as the parts of each gene which were missing (see Supplementary Fig. 5), also suggests that the observed gene fragmentation is largely a real phenomenon and not the result of incomplete and low quality GRC assemblies.

Among the 10 genes which had more than 95% of their coding region present on both species' GRCs, three were characterised as being homologs of endogenous retrovirus-derived proteins (*ervk* genes in Fig. 7). This includes three of the six genes that were found to be shared with *T. guttata.* The genes also include three uncharacterized genes and two homologs of Hydrocephalus-inducing proteins (*hydin* genes in Fig. 7). The two remaining genes are a zinc finger protein (*znf239* in Fig. 7) and a homolog of Cytoplasmic Polyadenylation Element Binding protein 1 (*cpeb1*). With the notable exception of *cpeb1*, these shared genes were often duplicated and differed in their copy number on the GRC, both between the species and between individuals of the same species (see Fig. 7). This means that while these genes are likely of ancient GRC linkage and potentially functional, they are still actively undergoing within-GRC duplication and deletion; and would be susceptible to all of the associated dosage changes. The *cpeb1* homolog, on the other hand, maintained its single copy number and open reading frame despite having been present on the GRC for a long time, diverging from the A chromosomal version before the common ancestor of all oscines and suboscines, early in passerine evolution (see Fig. 8).

Discussion

The GRC is an unusual chromosome. On the one hand, the apparent universal presence among songbirds (Torgasheva, et al. 2019) suggests that the GRC is not just a parasitic supernumerary B chromosome, as has been previously suggested (Camacho et al. 2000, Johnson Pokorná and Reifová 2021), but has some important function for these birds which prevents its loss. On the other hand, our data revealed the extremely dynamic nature of this chromosome, with a lack of conservation not only between closely related sister species that diverged merely 1.8 million years ago (Storchová et al. 2010), but even within species. Moreover, our results question the functionality of the majority of genes identified on the GRC.

The GRC represents a challenge to assemble. The chromosome is only found in germ cells, which represent a small subset of cells, even when harvesting the testes or ovaries specifically. This, combined with the fact that the GRC only occurs as a single copy in males, means that sequencing coverage of regions not duplicated within the GRC is low (about 20% of the A chromosome coverage in our data). Additionally, the GRC sequence is often very similar or indistinguishable from the sequence of A chromosomes (Kinsella, et al. 2019). For these reasons, previous attempts to identify GRC sequences, which relied on using highly repetitive GRC regions (having high germline coverage compared to somatic coverage) and germline specific SNPs, were unable to assemble regions that have a low copy number and are not highly differentiated from the Achromosomal sequence.

The method used in this paper to assemble the GRCs of the two nightingale species is able to assemble regions with low coverage and low divergence, as long as they have regions suitably nearby that are identifiable as GRC in origin, which can enable the classification of overlapping 10x linked barcodes. This is exemplified by the successful assembly of almost the whole GRC sequence in both nightingale species, despite the low copy number in *L. luscinia*. When near-identical duplications that are merged in the assembly are taken into account, our assemblies have a cumulative length that closely matches the chromosome sizes estimated from the cytogenetic visualisation, suggesting that we assembled the vast majority of the two nightingale GRCs. These assemblies thus represent the most complete and high quality GRC assemblies achieved to date.

Using the coding genes of *F. albicollis*, we identified 799 different coding genes on the two GRCs combined. Interestingly, most of them were species-specific, with only 192 of them occurring in both nightingale species. This is striking, given the recent divergence of the two nightingale species (1.8 Mya; Storchová, et al. 2010) and low genetic divergence of the A-chromosomes (Storchová, et al. 2010, Mořkovský, et al. 2018, Janoušek, et al. 2019). The comparison of the nightingale micro-GRC gene content with that of the *T. guttata* macro-GRC shows a similar lack of conservation. Of the 269 known GRC genes from *T. guttata* (Kinsella, et al. 2019), only 42 were found in either of the two nightingale species, with 25 being present in both.

The GRCs of the two nightingale species are also surprisingly different in sequence origin, as well as the proportion of within-GRC duplicated sequences. One stark difference is the large paralogous region of chromosome 2 on the GRC of *L. luscinia*, making up approximately half of the *L. luscinia* GRC.

Interestingly, analysis of the GRC and A chromosome divergence revealed that addition of this sequence to the GRC predates the divergence of the two species, suggesting the loss of the chromosome 2 paralogous region in *L. megarhynchos* rather than its recent addition in *L. luscinia*. On the other hand, we revealed that relatively large parts of the GRC in *L. megarhynchos* were added to the chromosome after the divergence of the two species. This suggests multiple frequent additions and deletions occurring on the GRC in a relatively short time span. Variation in copy number and the proportion of duplicated sequences was also high, even among individuals of the same species, suggesting that even within populations this chromosome is not well conserved.

The striking divergence of the GRC sequence and gene content, even between such closely related species, allows for the intriguing possibility that this chromosome might be involved in speciation in songbirds. Songbirds have a higher diversification rate compared to other bird taxa and comprise more than half of all modern bird species, despite only being one of many present lineages (Prum, et al. 2015; Oliveros, et al. 2019). We identified 29 species-specific GRC genes with a complete coding region on the two GRCs, which might be theoretically involved in the reproductive isolation of the two nightingale species mediated by female-limited hybrid sterility and possibly divergence of sperm morphology (Reifová et al. 2011; Mořkovský et al. 2018; Albrecht et al. 2019).

Most genes identified on the nightingale GRCs were only partially present, however, with approximately 75% of genes having less than a quarter of their coding region present and only 4% having their whole coding region present. This suggests that the vast majority of “genes” on the GRC are actually fragmented, non-functional pseudogenes. When this is combined with the observed rapid divergence of the GRC between species, it suggests that the GRC is largely non-functional and frequently acquires (and subsequently loses) sequences from the A-chromosomes. This, however, does not mean that at least a small proportion of the GRC is not important in function. Kinsella et al. (2019) found protein products for five genes on the *T. guttata* GRC, as well as signatures of selection on 10 GRC-linked genes, suggesting the functionality of at least some of the hundreds of GRC-linked genes. These genes could play important roles, for example in germline determination, oogenesis or spermatogenesis, although evolution of spermatogenesis functionality might be limited by the maternal inheritance of the GRC.

In an attempt to identify conserved genes on the GRC that may be preventing its loss from the songbird germline genome, we searched for genes with a complete coding sequence that were present in both nightingale species and *T. guttata*. However, 5 of the 6 genes identified this way represent genes such as endogenous retroviral homologs and uncharacterized, or poorly characterized genes, which despite being present in all three species, also show high variation in copy number within species. It thus seems unlikely that they represent indispensable GRC sequences.

The remaining complete gene that is present in both nightingale species as well as *T. guttata* is a paralog of *cpeb1*, cytoplasmic polyadenylation element binding protein 1. In addition to the normal Achromosomal version, this gene has a single copy on the GRC present in all 3 individuals of both species and does not feature any stop codons along its entire length. We estimated *cpeb1* to have

diverged from the A-chromosomal version early on in passerine evolution (before the divergence of suboscines and oscines, but after the split of Acanthisittidae). This makes this gene one of the oldest genes identified on the GRC so far and suggests that the GRC might be present not only in all songbirds but all passerine birds except for a small group of Acanthisittidae. Previous analysis of this gene on the zebra finch GRC found that it is under long-term purifying selection, further supporting the functionality of this gene (Kinsella, et al. 2019). *cpeb1* is known to play a role in transcript modification during oocyte maturation (Hake and Richter, 1994). Together, these findings make *cpeb1* the best candidate for a functionally important gene which may be preventing the loss of the GRC from the songbird germline.

The picture emerging of the GRC is that it is a tumultuous chromosome, where large stretches of DNA can be added and subtracted rapidly, seemingly without consequence. Once on the GRC, any sequence is liable to be duplicated on the chromosome multiple times. The pace and scale of these changes does not seem conducive to the fine scale refinement normally associated with natural selection. This, combined with the fact that most genes are fragmented, suggests that the vast majority of the chromosome is non-functional, with presumably a small ancestral region harbouring a gene or genes that are driving the continued existence of the chromosome in the songbird lineage.

The seemingly contradictory picture of the GRC highlights how programmed DNA elimination can change the evolutionary landscape of genetic sequences. The fact that the GRC is eliminated from somatic cells means that there are much less pleiotropic constraints on this chromosome compared to A-chromosomes, which may lead to less selection pressure acting on this chromosome. As a consequence, many genetic changes, which would have large negative consequences for an individual if they occurred on an Achromosome, are effectively silenced on the GRC.

Conclusion

This work represents the first comparison of GRCs between closely related songbirds, demonstrating the speed with which the GRC undergoes change. It also represents the most complete, albeit fragmented, GRC assemblies produced to date. Our results emphasise how rapidly this chromosome evolves, with large variation being observed between the two species on almost every metric we measured, and moderate variation being observed even within each species. This contrasts starkly with the normally conserved bird karyotype and makes the GRC the fastest evolving chromosome in the genome. We also show that most genes that are present on the GRC are present in a fragmented, presumably non-functional, state. While the ubiquity of the GRC within the songbird clade does suggest an important role for the chromosome, it is still unclear what that role is. The chromosome appears to be under uniquely relaxed evolutionary pressure, presumably as a result of its elimination from somatic cells, and it seems likely that its main function is limited to a few consequential genes, one of which seems to be *cpeb1*.

Materials And Methods

Nightingale Sampling

Three male individuals from each nightingale species were sampled for whole genome sequencing in allopatric regions (North-Eastern Poland for *L. luscinia* and South-Western Poland for *L. megarhynchos*). From each individual, somatic tissue (kidney) and gonadal tissue (testis) were dissected and either used immediately for DNA isolation or frozen in liquid nitrogen and stored in 80°C prior to DNA isolation. The work was approved by the General Directorate for Environmental Protection, Poland (permission no. DZP-WG.6401.03.123.2017.dl.3).

Preparation of meiotic spreads and estimation of GRC size

Measurements were made of immunostained synaptonemal complexes of pachytene chromosomes from Poignet et al. (2021). Chromosomes were immunostained with anti-SYCP3 antibody recognizing the lateral elements of the synaptonemal complex, and human anticentromere serum (CREST, 15–234, Antibodies Incorporated) binding kinetochores (see Poignet et al. 2021 for details). The GRC can be recognized from other chromosomes on these slides by its relatively weaker staining by anti-SYCP3 antibody and the CREST signal which covers the whole chromosome, instead of just the centromere.

The lengths of the 22 largest chromosomes and the GRCs were measured in high-quality cells for each species using ImageJ software (ImageJ 1.50i, Rueden et al. 2017). This resulted in 15 *L. megarhynchos* cells (12 from one individual and 3 from another) and 16 cells from *L. luscinia* (9 from one individual and 7 from the other) being used. The measured length of the GRC was divided by 1.5, due to a measurement discrepancy caused by its univalent nature (Malinovskaya et al. 2020), before the size was calculated using a linear regression (see Supplementary Fig. 2). This used the relationship between the logarithmic values of the 22 longest chromosomes lengths and the logarithmic size in base pairs of the 22 largest chromosomes from the *F. albicollis* genome, FicAlb1.5 (R^2 = 0.97 in *L. megarhynchos* and R^2 = 0.98 in *L. luscinia*). The approximate size was checked against the size of the eliminated GRC micronucleus using rabbit monoclonal anti-H3K9me3 antibody (ab8898, Abcam) (dilution 1:200).

Sequencing of somatic and germline genomes

We sequenced DNA from somatic (kidney) and gonadal (testis) tissues from three individuals of each species. One individual from each species had DNA from both tissues sequenced using 10x linked (Zheng, et al. 2016) Illumina sequencing technology, while the other two individuals had DNA samples sequenced with standard paired-end Illumina sequencing.

For 10x linked sequencing, high molecular weight DNA was extracted from frozen testis and liver samples using a phenol-chloroform methodology (Pajer, et al. 2006). The DNA was sent to SEQme (Dobris, Czech Republic) for 10x linked sequencing library construction and 2x150 bp paired-end sequencing using the NovaSeq 6000 (Illumina). For the standard Illumina sequencing, DNA was extracted from frozen tissue samples using MagAttract HMW DNA Kit (Qiagen) and sent to the Institute of Applied Biotechnologies (Prague, Czech Republic) where the sequencing libraries were prepared using NebNext Ultra II DNA Kit (New England Biolabs) and sequenced with the NovaSeq 6000 (Illuimna) using 2x150 bp paired-end mode.

Testis samples were sequenced to higher depth (105-150x) than the kidney samples (45-120x) to ensure sufficient coverage over the GRC (see Supplementary Table 2).

Identification and assembly of GRC reads

Linked 10x reads that originated from the GRC were identified using the following methods before being assembled using Supernova (Weisenfeld, et al. 2017) and the “megabubbles” option (method visualised in Supplementary Fig. 3):

1) The testis samples were assembled using the 10x linked reads and Supernova (Weisenfeld, et al. 2017). The 10x reads from the testes and kidneys were then processed using Long Ranger v2.2.2, trimmed and checked for adapters using Trimmomatic v0.39 (Bolger, et al. 2014), before being aligned to their respective genome using bwa v0.7.17 (Li and Durbin 2010). Regions of the testis genome assembly were identified using Samtools v1.14 (Li, et al. 2009) which were fully covered by reads from the testis dataset while having no reads align to them from the somatic dataset and that were at least 500 bp long. Reads from testis samples that overlapped these regions by at least 10 bp were used to identify 10x barcodes and their associated reads as originating from the GRC.

2) Sequencing reads from 10x libraries were processed using Long Ranger v2.2.2 before all reads were trimmed and checked for adapters using Trimmomatic v0.39 (Bolger, et al. 2014). These reads were then aligned to draft somatic genomes from their respective species created using Oxford nanopore data and Illumina reads (Rídl, et al. unpublished). SNP variants were identified using GATK v4.1.7.0 (Poplin et al, 2017). If a variant was present in all three testis samples from a species, but in none of the kidney samples from that species, it was considered to be a GRC variant. These SNP variants were used to create 29 bp kmer sequences (i.e. each variant resulted in 29 kmers). Any kmer that was present in the 10x kidney reads was removed and the remaining kmers were used to identify reads from the 10x testis dataset that had a matching sequence. The barcodes from these reads were used to identify any associated reads.

3) A sample of 100 000 reads from each of the 10x datasets was used to identify repetitive elements that might be unique to the GRC of each species using RepeatExplorer (Novak, et al. 2010). While no such repeats were found in *L. luscinia*, a candidate repeat was found in the testis dataset of *L. megarhynchos*. This result was confirmed using all the 10x reads from *L. megarhynchos* and Blastn (Altschul, et al. 1990), with a word size of 8, an e-value of 1e-5, and a max hsps of 1. Any read that matched with at least 100 bp and greater than 90% identity to the repetitive element was selected. Once again, all reads with the associated barcodes were designated as having originated from the GRC.

Gene Annotation

The first frame of the *F. albicollis* transcriptome coding sequences (v FicAlb1.5) were aligned independently to the two assembled GRCs using Tblastx (Altschul, et al. 1990) and an e-value cut-off of 1e-6. Overlapping alignments on the same strand of the GRC were merged. The resultant regions were

aligned back to the *F. albicollis* coding sequences and the top hit in the positive strand selected to identify which gene (and which portion of the gene) the exon represented.

In order to account for a possible lack of conservation, when calculating what proportion of a gene was present on the GRC, the fraction of a gene that was found in the GRC was normalised by the fraction of the gene that was found in the *L. megarhynchos* genome, to a minimum of 0.75. In other words, if 80% of a gene was found in the genome, and 80% was found in the GRC, that gene was treated as being 100% present in the GRC. The cut-off of 0.75 fully corrects for 2/3rds of all genes identified on the genome. This correction resulted in an average increase in the measured proportion of the found gene of 8.5% in *L. megarhynchos* and 8.9% in *L. luscinia*.

Coverage of GRC scaffolds

In order to identify scaffolds that represent near-identical duplicates and/or erroneous sequences in the GRC assembly, the assembled GRC sequence for each species was combined with the corresponding draft somatic genome assembly (Rídl et al. unpublished). The sequencing reads from both tissue types for all three individuals were aligned to the combined genome and GRC assembly using bwa v0.7.17 (Li and Durbin 2010) for each species. For each individual, regions of the GRC which had zero read coverage from the kidney dataset were identified. The modal testis coverage value of these regions for each individual was used as an estimate of the expected coverage for single copy GRC regions. The ratio of the expected GRC coverage to the modal genomic coverage was used as a proxy for the “expected” GRC coverage in the kidney samples. These expected coverage values were used to normalise the observed coverage for each sample. The coverage of the kidney sample was then subtracted from the coverage of the testis sample to control for A chromosomal reads misaligning to the GRC sequence. Finally, the average GRC coverage was calculated across 1 kb windows along the two GRCs, with a minimum value of zero.

GRC Scaffold Origin and Conservation

Given that GRC sequences appear to originate from A chromosomes, the GRC scaffolds were aligned to the draft *L. megarhynchos* and *L. luscinia* somatic genomes (Rídl et al. unpublished) using Blastn and an e-value cutoff of 1e-6 (Altschul, et al. 1990). Additionally, the two GRCs were aligned to each other using Blastn. The A chromosomal origin of the GRC sequences were determined by the top hit in the *L. megarhynchos* genome, since it was the higher quality genome assembly. The genomic scaffold was then aligned to the *F. albicolli* genome assembly to determine the chromosome identity using Nucmer from the MUMmer package (Kurtz et al. 2004).

cpeb1 Evolution

The *cpeb1* gene sequence was determined from the Tblastx results with the XP_005051706.1 transcript from the two GRCs and their respective genomes. The GRC sequence was used to identify homologous sequences in related species using the ‘nr’ database on NCBI and Tblastx. For each species, the top hit

was selected. These species included: *Acanthisitta chloris*, *Atrichornis clamosus*, *Calyptomena viridis*, *Corvus cornix cornix*, *Gallus gallus*, *Lonchura striata domestica*, *Sapayoa aenigma*, *Serilophus lunatus*, *Serinus canaria*, *Struthio camelus australis*, *Taeniopygia guttata* and *Tyrannus savana*. The sequences were aligned using ClustalW (Thompson, et al. 1994) and a maximum likelihood tree drawn with MegaX (Kumar, et al. 2018).

Declarations

Acknowledgements

This research was funded by the Czech Science Foundation (grant 20-23794S to R.R. and T.A.), the Charles University grant PRIMUS/19/SCI/008 to R.R. and the Grant Agency of Charles University (grant 1169420 to M.P.). F.J.R.R. was supported by a postdoctoral fellowship from Sven och Lilly Lawskis fond and a Marie Curie Individual Fellowship (875732). Computational analysis was mostly done using the Institute of Molecular Genetics (Czech Academy of Sciences, Prague, Czech Republic) computers. Additional computational resources were provided by the ELIXIR-CZ project (LM2018131), part of the international ELIXIR infrastructure.

Contributions

The project was conceptualized by R.R., S.A.S., A.S., T.A. and J.P.; Samples were collected by J.Re., T.A., R.R., and M.P.; Cytogenetic analysis was performed by M.P.; DNA extraction was done by J.Rí. Bioinformatic analyses were done by S.A.S.; Manuscript was written by S.A.S. and R.R. and all authors contributed to text editing.

References

1. Albrecht, T., et al. 2019. Sperm divergence in a passerine contact zone: Indication of reinforcement at the gametic level. Evolution **73**: 202–213. https://doi.org/10.1111/evo.13677
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. 1990. Basic local alignment search tool. J Mol Biol **215**: 403–410
3. Asalone, K. C., Takkar, A. K., Saldanha, C. J. and Bracht, J. R. 2021. A transcriptomic pipeline adapted for genomic sequence discovery of germline restricted sequence in zebra finch, *Taeniopygia guttata*. Genome Biol Evol **13**: **6**: evab088. https://doi.org/10.1093/gbe/evab088
4. Biederman, M. K., Nelson, M. M., Asalone, K. C., Pedersen, A. L., Saldanha, C. J. and Bracht, J. R. 2018. Discovery of the First Germline-Restricted Gene by Subtractive Transcriptomic Analysis in the Zebra Finch, *Taeniopygia guttata*. Curr Biol **28**: **10**: 1620–1627.e5. doi: 10.1016/j.cub.2018.03.067
5. Bolger, A. M., Lohse, M., and Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* btu170

6. Camacho, J. P., Sharbel, T. F. and Beukeboom, L. W. 2000. B-chromosome evolution. Philos Trans R Soc Lond B Biol Sci **355**: **1394**: 163–178. doi: 10.1098/rstb.2000.0556
7. del Priore, L. and Pigozzi, M. I. 2014. Histone modifications related to chromosome silencing and elimination during male meiosis in Bengalese finch. Chromosoma **123**: **3**: 293–302. doi: 10.1007/s00412-014-0451-3
8. Hake, L. E. and Richter, J. D. 1994. CPEB is a specificity factor that mediates cytoplasmic polyadenylation during Xenopus oocyte maturation. Cell **79**: **4**: 617–627. doi: 10.1016/0092-8674(94)90547-9
9. Itoh, Y., Kampf, K., Pigozzi, M. I., and Arnold, A. P. 2009. Molecular cloning and characterization of the germline-restricted chromosome sequence in the zebra finch. Chromosoma **118**: **4**: 527–536. https://doi.org/10.1007/s00412-009-0216-6
10. Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., Mooers, A. O. 2012. The global diversity of birds in space and time. Nature **491**: 444–448
11. Janoušek, V., et al. 2019. Postcopulatory sexual selection reduces Z-linked genetic variation and might contribute to the large Z effect in passerine birds. Heredity **122**: 622–635. https://doi.org/10.1038/s41437-018-0161-3
12. Johnson Pokorná, M. and Reifová, R. 2021. Evolution of B Chromosomes: From Dispensable Parasitic Chromosomes to Essential Genomic Players. Front Genet **12**: 727570. doi: 10.3389/fgene.2021.727570
13. Kawakami, T., et al. 2014. A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. Mol Ecol **23**: 4035–4058. https://doi.org/10.1111/mec.12810
14. Kinsella, C. M., et al. 2019. Programmed DNA elimination of germline development genes in songbirds. Nat Commun **10**: 5468. https://doi.org/10.1038/s41467-019-13427-4
15. Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution **35**: 1547–1549
16. Kurtz, S., et al. 2004. Versatile and open software for comparing large genomes. Genome Biology **5**: R12
17. Li, H., Durbin, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26**: **5**: 589–595. doi: 10.1093/bioinformatics/btp698
18. Li, H., et al. 2009. 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**: **16**: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
19. Malinovskaya, L. P., et al. 2020. Germline-restricted chromosome (GRC) in the sand martin and the pale martin (Hirundinidae, Aves): synapsis, recombination and copy number variation. Sci Rep **10**: 1058. https://doi.org/10.1038/s41598-020-58032-4
20. Mořkovský, L., et al. 2018. Genomic islands of differentiation in two songbird species reveal candidate genes for hybrid female sterility. Mol Ecol **27**: **4**: 949–958.

https://doi.org/10.1111/mec.14479

21. Novak, P., Neumann, P., Macas, J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics **11**: 378
22. Oliveros, C. H., et al. 2019. Earth history and the passerine superradiation. *Proceedings of the National Academy of Sciences* **116**: **16**: 7916–7925. doi: 10.1073/pnas.1813206116
23. Pajer, P., et al. 2006. Identification of Potential Human Oncogenes by Mapping the Common Viral Integration Sites in Avian Nephroblastoma. Cancer Research **66**: 78–86
24. Pei, Y., et al. 2022. Occasional paternal inheritance of the germline-restricted chromosome in songbirds. Proc Natl Acad Sci **119**: **4**: e2103960119. doi: 10.1073/pnas.2103960119
25. Pigozzi, M. I. and Solari, A. J. 1998. Germ cell restriction and regular transmission of an accessory chromosome that mimics a sex body in the zebra finch, *Taeniopygia guttata*. Chromosome Res **6**: **2**: 105–113. doi: 10.1023/a:1009234912307
26. Pigozzi, M. I. and Solari, A. J. 2005. The germ-line-restricted chromosome in the zebra finch: recombination in females and elimination in males. Chromosoma **114**: **6**: 403–409. doi: 10.1007/s00412-005-0025-5
27. Poignet, M., et al. 2021. Comparison of karyotypes in two hybridizing passerine species: conserved chromosomal structure but divergence in centromeric repeats. Front Genet **12**: 768987. doi: 10.3389/fgene.2021.768987
28. Poplin, R., et al. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* 201178. doi: 10.1101/201178
29. Prum, R., et al. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526: 569–573. https://doi.org/10.1038/nature15697
30. Reifová, R., Kverek, P. and Reif, J. 2011. The first record of a female hybrid between the Common Nightingale (*Luscinia megarhynchos*) and the Thrush Nightingale (*Luscinia luscinia*) in nature. J Ornithol **152**: 1063–1068. doi: 10.1007/s10336-011-0700-7
31. Rueden, C. T., et al. 2017. ImageJ2: ImageJ for the next generation of scientific image data. BMC Bioinformatics **18**: 529. https://doi.org/10.1186/s12859-017-1934-z
32. Storchová, R., Reif, J. and Nachman, M. W. 2010. Female heterogamety and speciation: reduced introgression of the z chromosome between two species of nightingales. Evolution **64**: 456–471. https://doi.org/10.1111/j.1558-5646.2009.00841.x
33. Suh, A. and Dion-Côté, A-M. 2021. New Perspectives on the Evolution of Within-Individual Genome Variation and Germline/Soma Distinction. *Genome Biol Evol* **13**: **6**: evab095. https://doi.org/10.1093/gbe/evab095
34. Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22**: **22**: 4673–4680. https://doi.org/10.1093/nar/22.22.4673

35. Torgasheva A. A., et al. 2019. Germline-restricted chromosome (GRC) is widespread among songbirds. Proc Natl Acad Sci U S A **116**: **24**: 11845–11850. doi: 10.1073/pnas.1817373116
36. Torgasheva, A., Malinovskaya, L., Zadesenets, K., Shnaider, E., Rubtsov, N. and Borodin, P. 2021. Germline-Restricted Chromosome (GRC) in Female and Male Meiosis of the Great Tit (Parus major, Linnaeus, 1758). Front Genet **12**: 768056. doi: 10.3389/fgene.2021.768056
37. Wang, J. and Davis, R. E. 2014. Programmed DNA elimination in multicellular organisms. Curr Opin Genet Dev Aug: **27**: 26–34. doi: 10.1016/j.gde.2014.03.012
38. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. and Jaffe, D. B. 2017. Direct determination of diploid genome sequences. Genome Res **27**: 757–767
39. Zheng, G., Lau, B., Schnall-Levin, M. et al. 2016. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol 34: 303–311. https://doi.org/10.1038/nbt.3432

Figures

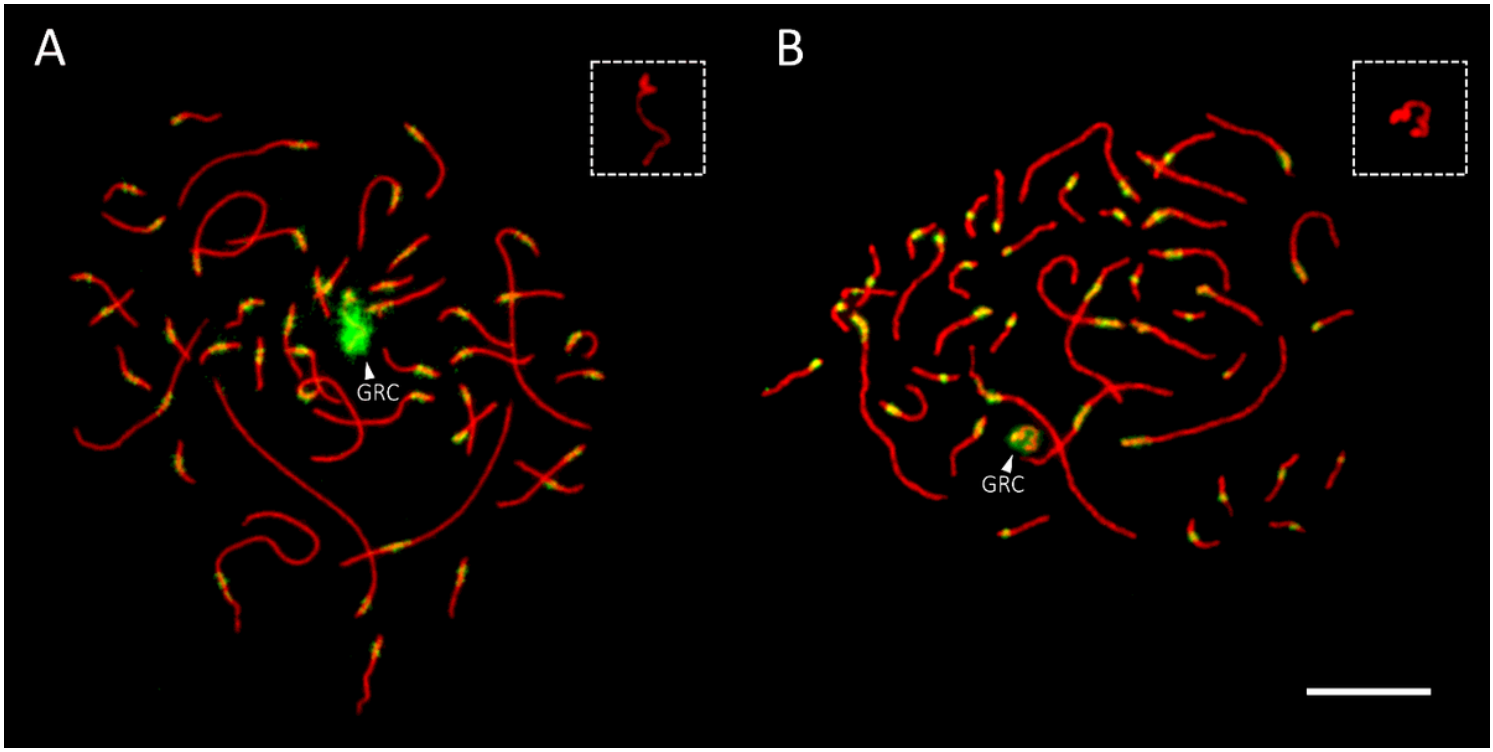

Figure 1

Pachytene chromosomes in *Luscinia megarhynchos* (A) and *L. luscinia* (B) immunostained with SYCP3 antibody against the synaptonemal complex (red) and CREST antibody against centromere (green). Arrowheads indicate the GRCs. The box in the top right corner shows the GRC in more detail (1.5x magnification) without the CREST signal. The scale bar represents 10 μm.

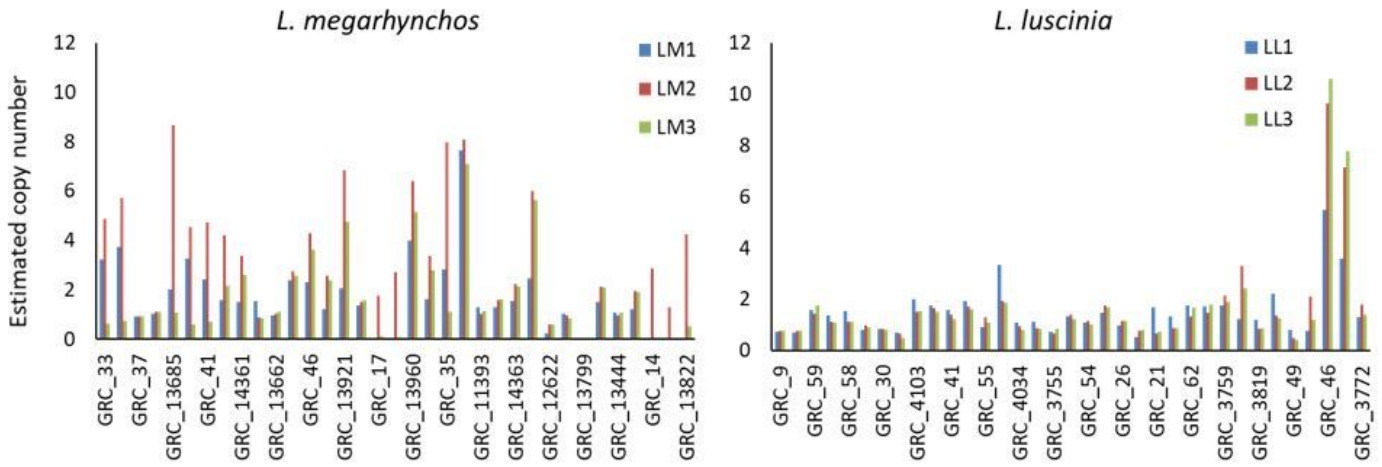

Figure 2

Average normalised copy number of the longest scaffolds in each nightingale individual.

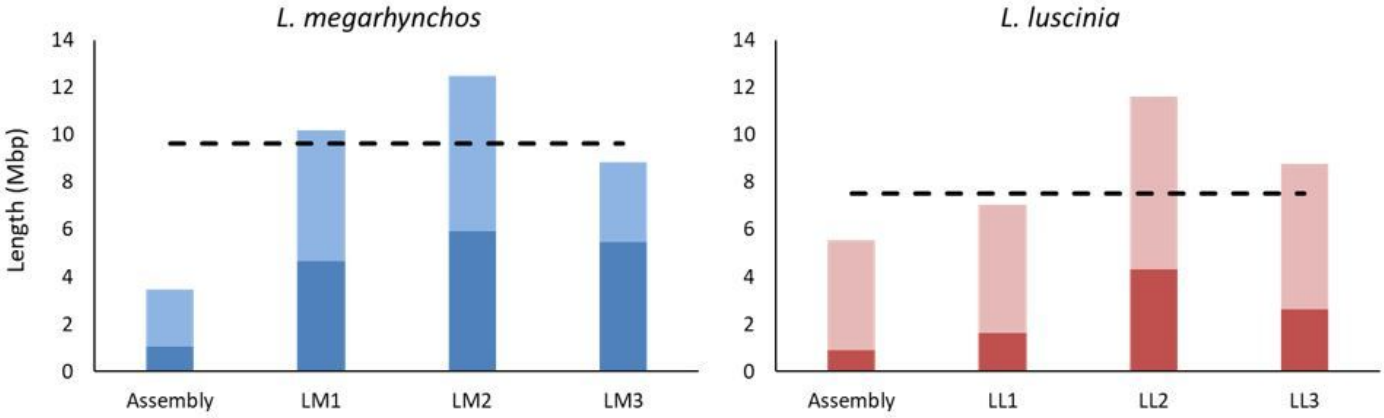

Figure 3

GRC size estimates from genomic data for the three nightingale individuals of each species, taking near-identical duplications merged in the assembly into account. The proportion of the GRC that is shared between the species is shown in a darker colour, while the proportion of the GRC that is species-specific is in lighter colour. Assemblies are based on the individuals LM1 and LL1 respectively. The estimated GRC size for each species from meiotic spreads is shown by the horizontal dashed lines.

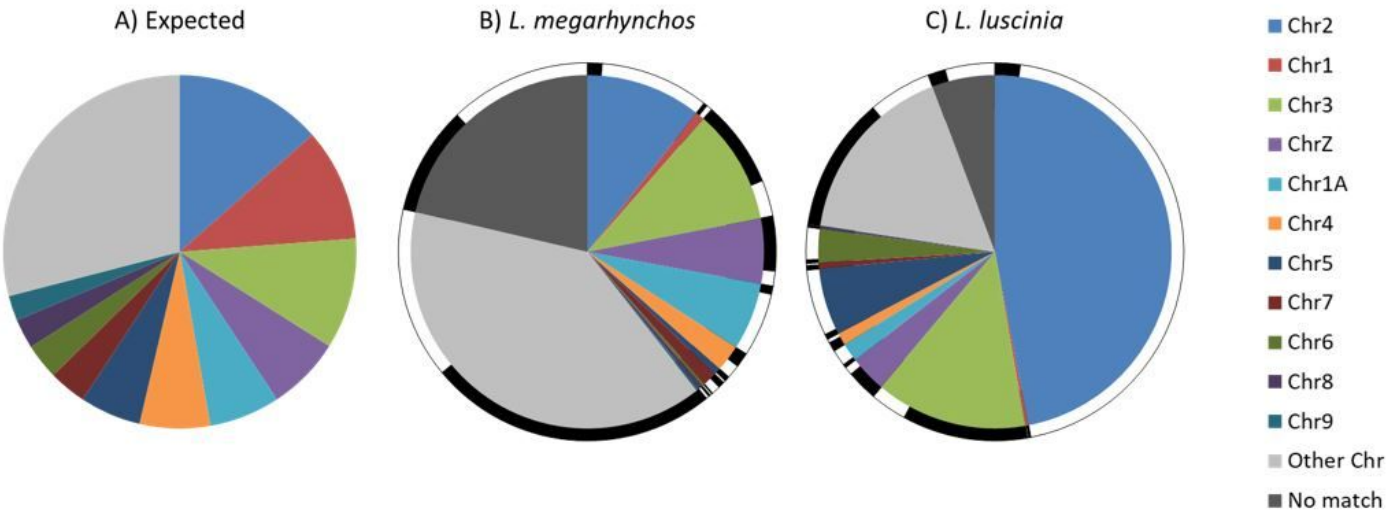

Figure 4

Origin of GRC sequence. A) Expected contribution from the A chromosomes to the GRC if each chromosome was contributing equally according to its size. B) Average observed contribution of A chromosomes to the *L. megarhynchos* GRC. C) Average observed contribution of A Chromosomes to the *L. luscinia* GRC. The black and white outer circles show the proportion of each sequence that has a homologous sequence in the other species. The chromosome order is from largest to smallest. The proportions have been corrected for coverage, which accounts for near-identical duplications collapsed in the assembly as well as assembly errors.

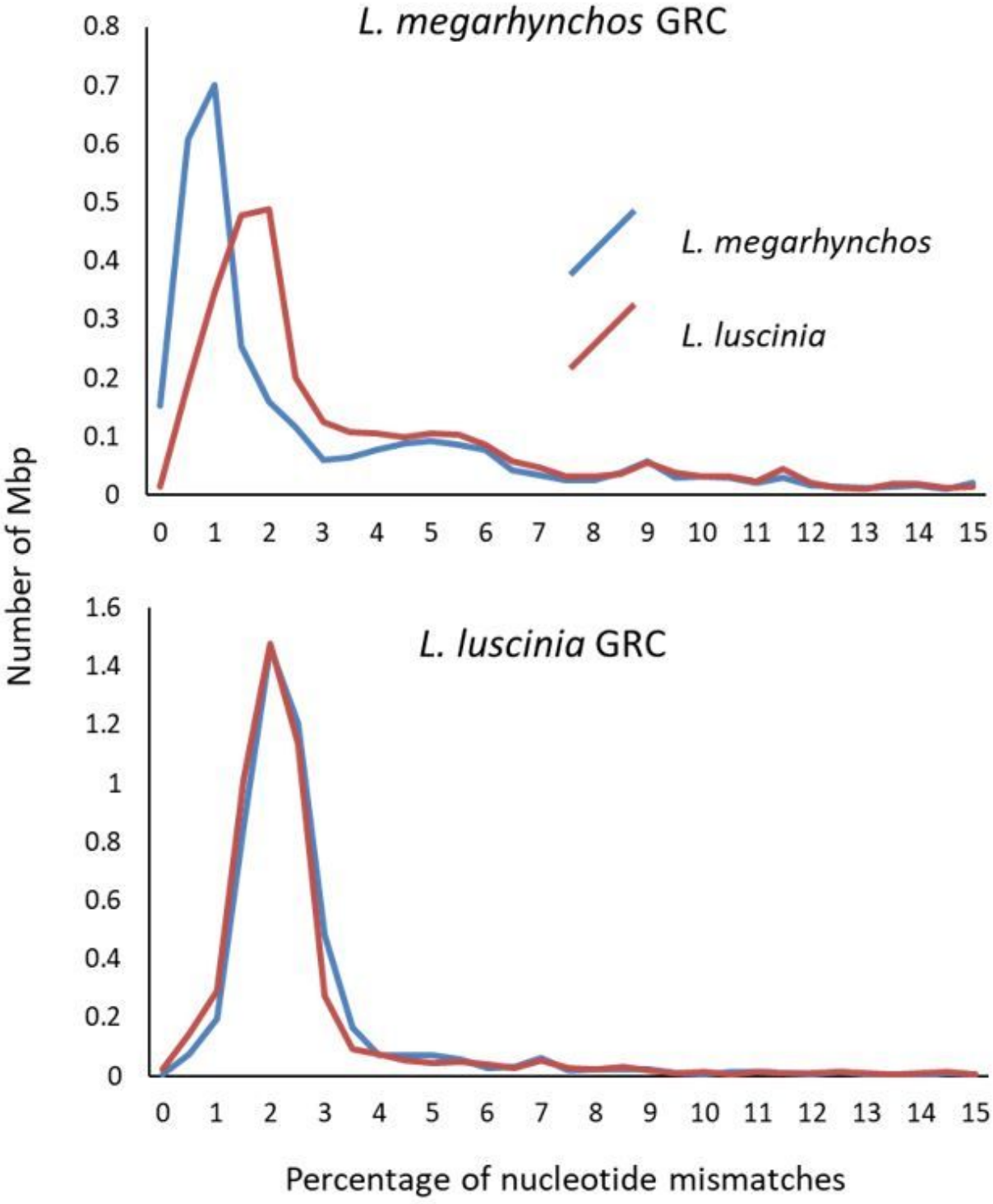

Figure 5

Divergence of each GRC assembly from A-chromosomal sequences in *L. megarhynchos* and *L. luscinia.*

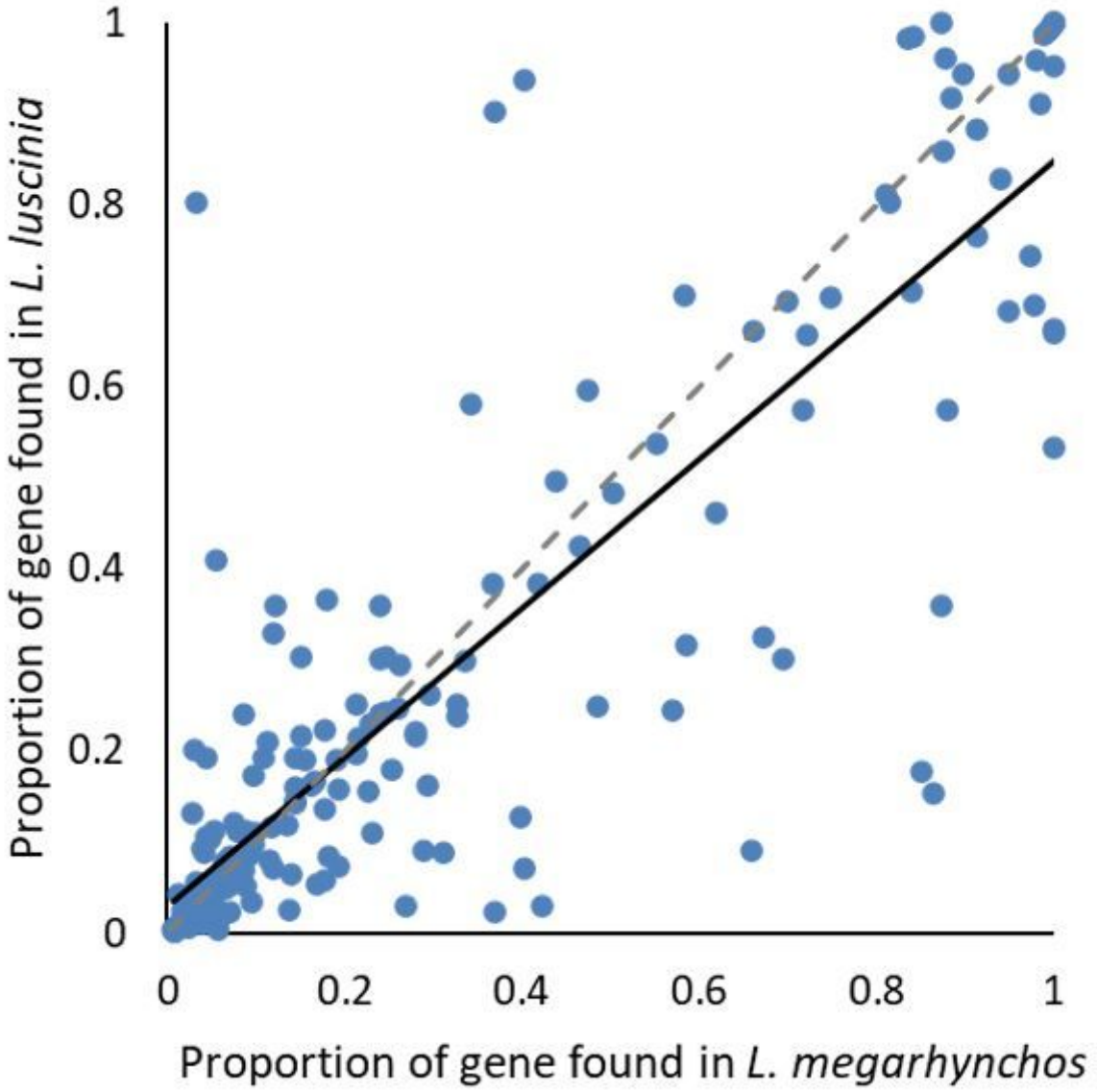

Figure 6

Completeness of genes found on the GRC of both *L. megarhynchos* and *L. luscinia*. The grey dashed line represents a 1:1 ratio. The black line represents the linear fit of the data (R^2 = 0.76).

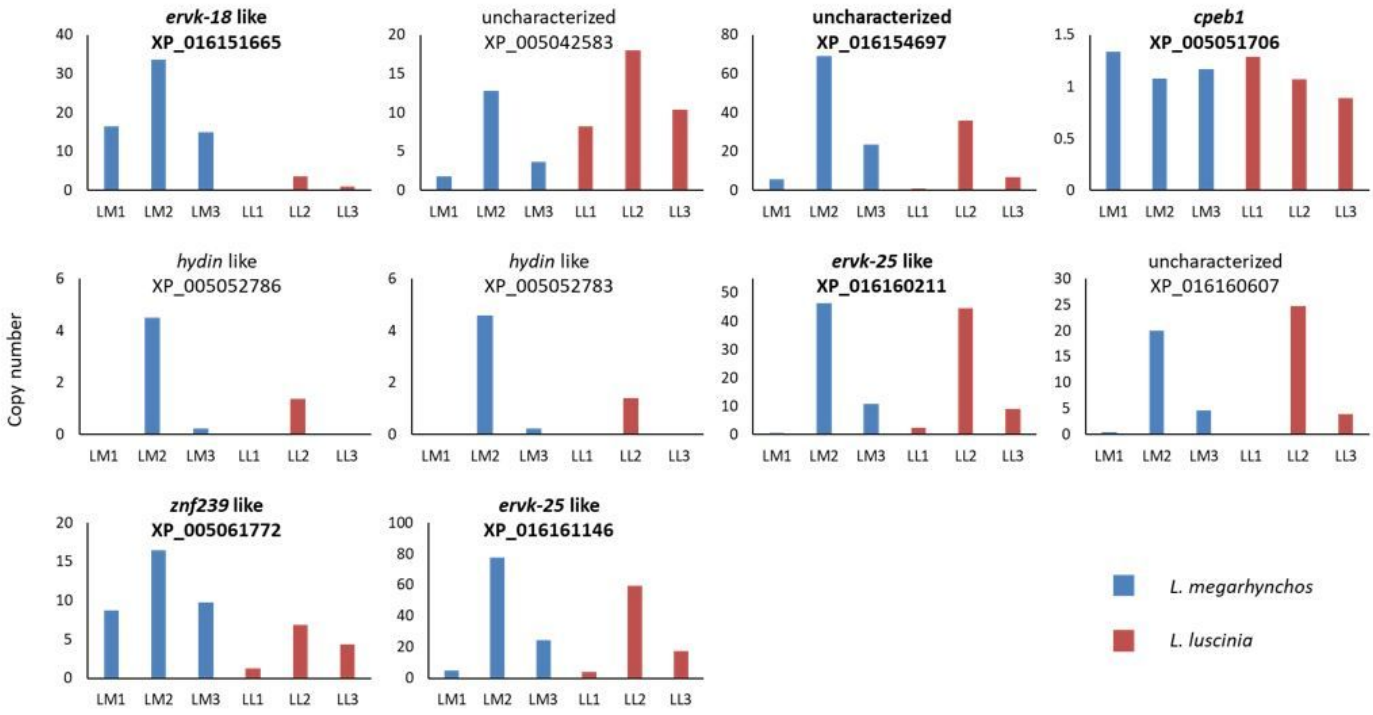

Figure 7

Estimated copy number of the 10 GRC genes with at least 95% of their coding region found in both nightingale species. Genes with bold names were also found in *T. guttata*. The copy number estimate for each individual is based off the average normalised coverage of each scaffold that the genes were present on.

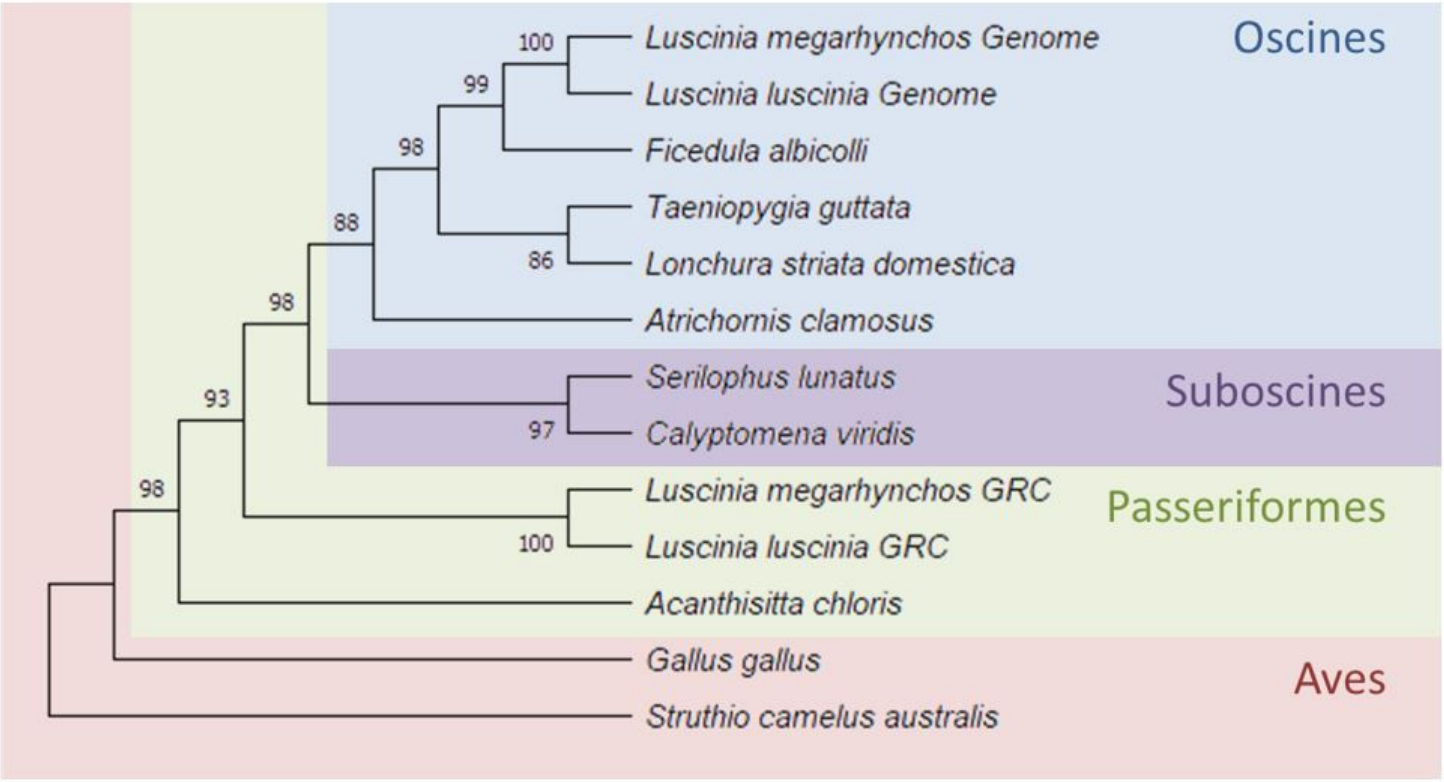

Figure 8

Divergence of the *cpeb1* GRC paralog from the A chromosomal version. The GRC paralogue diverges from the A-chromosomal version before the Oscine/Suboscine divergence. Branch values represent bootstrap support.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesv2.7.xlsx
- SupplementaryInformationv2.7.docx
- rs.pdf